University of Valladolid
Computer Engineering School of Valladolid

Master in Computer Engineering
Specialty Big Data

The Blind Oracle

eXplainable Artificial Intelligence (XAI) and human agency

Valladolid, July 2020

Author:
Raúl Fernández Álvarez
Supervisor:
Fernando Díaz Gómez

Abstract

An explainable machine learning model is a requirement for trust. Without it the human operator cannot form a correct mental model and will distrust and reject the machine learning model. Nobody will ever trust a system which exhibit an apparent erratic behaviour.

The development of eXplainable AI (XAI) techniques try to uncover how a model works internally and the reasons why they make some predictions and not others. But the ultimate objective is to use these techniques to guide the training and deployment of fair automated decision systems that support human agency and are beneficial to humanity.

In addition, automated decision systems based on Machine Learning models are being used for an increasingly number of purposes. However, the use of black-box models and massive quantities of data to train them make the deployed models inscrutable. Consequently, predictions made by systems integrating these models might provoke rejection by their users when they made seemingly arbitrary predictions. Moreover, the risk is compounded by the use of models in high-risk environments or in situations when the predictions might have serious consequences.

Keywords— eXplainable Artififical Intelligence (XAI), Human-Caused Fires (HCFs)

Dedicated to all the people who risk their lives fighting wildfires.

Contents

List of Figures

List of Tables

Chapter 1

Objectives and structure

1.1 Objective

The objective of this work is the appraisal of different XAI techniques using a case of study.

The case of study consists in building predictive models of the burnt area caused by Human-Caused Fires (HCFs) in continental Portugal. I have chose this case of study as its presents a real problem with severe repercussions. Although this increases greatly the difficulty of the present work the reward is also larger, even if it ends in failure.

I have focused the case of study on the mainland territory of the Republic of Portugal, one of the most ravaged regions by HCFs. Moreover, climate change spells a grim outlook for the future with an increase in the occurrence and the scale of devastation of HCFs.

To explore the XAI techniques I will train two regression models. For one of the models I am going to use a black-box algorithm, Gradient Boosting Machines (GBM) [Fri00]. Whereas, for the other I am going to use a glass-box model, Generalised Linear Models (GLM) [Cra14]. Training two models allows me use XAI in their comparison.

To build the HCFs model I am going to use explanatory variables from the following categories:

- Weather factors
- Physiography variables
- Fuel risk factors
- Human factors

Weather factors and physiography variables shape the spatio-temporal patterns influencing HCFs. while, fuel risk factors and human factors serve

as proxy for human activity, and HCFs are caused directly or indirectly by people.

Also, I am going to use Major Habitat Types (MHTs) to stratify the data as factors influencing the occurrence of wildfires vary between MHTs.

1.2 Project lifecycle

The high-level stages and its composing tasks are inspired by the CRISP-DM and Team Data Science Process (TDSP) (especially the later):

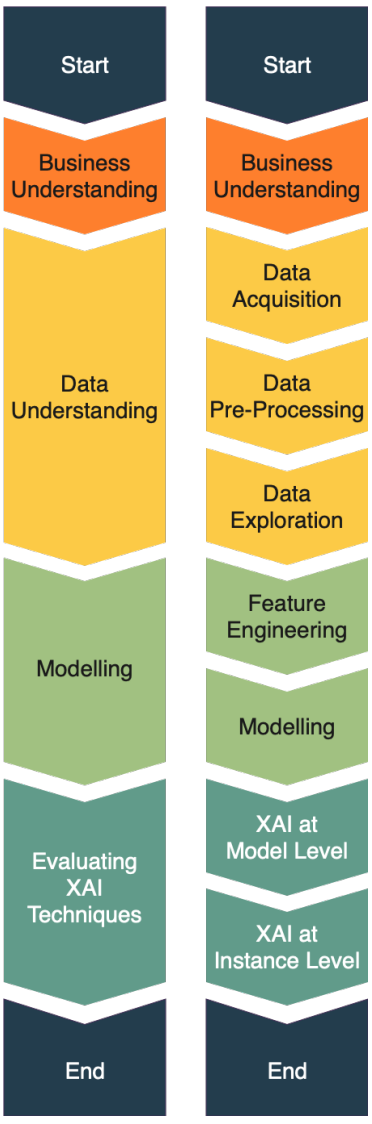


Figure 1.1: Project lifecycle high-level stages and its composing tasks

1.3 Document structure

This work is composed by the following parts and chapters:

Chapter 1: Objectives and structure The present chapter. It is a description of the objectives of this work and the use, the different sections composing it, the outline of the plan followed, and the personal involved and tools used in its development.

Part I – State of the art of XAI Review of recent development in the field of XAI

Chapter 2: On eXplainable AI (XAI) Introduction to the field of XAI, the problems it tries to address and approaches to tackle them. I also introduce some definitions commonly used in the field.

Part II – Case of study: Predicting burnt area caused by HCFs Case of study development from business understanding to modelling

Chapter 3: Human-Caused Fires (HCFs) Introduction to the use case that I have chosen as vehicle to survey the field of XAI by applying some of its proposed techniques to it. I begin by presenting a general vision of the problem to end by focusing on the zone of study, the mainland of the Republic of Portugal.

Chapter 4: Data acquisition Description of the different data sets indicating for each one the area of the problem that try to cover, its source, and the process of acquisition.

Chapter 5: Data pre-processing Description of the process employed to prepare the raw data so it can be easily consumed in later stages. Some of the actions I take in this phase are: extracting the part of the raw data relevant to the use case (as most of the datasets cover extensions bigger the zone of study), putting the data in a format easily consumed ()as tabular data in CSV files or in a GIS database), conversion raw variables, normalisation of GIS data into a common Spatial Reference System, etc.

Chapter 6: Data exploration Summary of the data exploration process for each of the data sets selected as candidates to be used in the building of the different models. I use statistical and graphical techniques to profile the different data set.

Chapter 7: Feature engineering Creation of the features from the raw data by handling outliers that could be present in the data, dealing with missing data, assigning the correct data types to

the variables, and applying transformation to convert raw data variables in features that would be used in the building of the machine-learning models.

Chapter 8: Modelling This chapter contains the presentation of the feature selection strategy and process that takes as input all the features I created in the previous chapter and the building and tunning of the machine-learning models.

Part III – Application of Explainable AI techniques to the case of study Appraisal of some XAI techniques in the context of the case of study.

Chapter 9: Explainable AI at model level Application of different XAI techniques used to explain a model as a whole by analysing its performance and quality of its predictions, and how the different features influences model's predictions.

Chapter 10: Explainable AI at instance level Application of different XAI techniques used to explain individual predictions to the models built trying to understand how they yield a prediction for a specific observation trying to understand the effects and influence of the features on the prediction.

Part IV – Conclusions conclusions I have reached while working in this project and possible future directions.

Chapter 11: Conclusions, and a look to the future Presentation of the conclusions I have reached after the development of this project and possible future directions that could be taken taking all I have learnt.

1.4 Personnel and roles

It is not meaningful to describe the personnel working in this project since this is a one-man effort. However, I am going to detail the different roles needed to carry out this project.

The list of roles is:

- Business owner, defines the business problem to be solved
- Business analyst, defines the feasibility of the project and setting its requirements
- Solution architect, organises the development
- Data analyst, acquires the needed data and interprets the data
- Data scientist, modelling

The lack of a domain expert is a big handicap in any project. To partly alleviate the problem, I will take advantage of the large number of resources, such as articles, available about the selected use case.

1.5 Plan

The plan and its phases are built by adapting two data science methodologies:

- CRoss-InduStry Process for Data Mining (CRISP-DM) [Sma20]
- Team Data Science Process (TDSP) by Microsoft [Mic20]

A short description of both methods can be found in Appendix A.

Both models allow the customisation of its different phases and activities to a project. In our case the main phases and their goals are:

1. Business understanding
   (a) Objectives definition
   (b) Case study exploration
2. Data acquisition
   (a) Data sources identification
   (b) Data acquisition
   (c) Data pre-processing
3. Data exploration
4. Modelling
   (a) Feature engineering
   (b) Model training and tuning
5. Evaluation
   (a) Evaluation using XAI techniques at model level
   (b) Evaluation using XAI techniques at instance level
6. Delivery
   (a) Project hand-off

Due to the iterative nature of both methodologies the different phases overlap with each other.

1.6 Infrastructure and resources

The list of hardware and software resources used to carry out the project is:

- Hardware: MacBook Pro MacBook Pro (2017 13-inch)
  - 3.1 GHz Intel Core i5
  - 8 GB 2133 MHz LPDDR3)
  - macOS 10.14.6
- Data storage
  - PostgreSQL 12.1 [Pos19b] with the PostGIS 3.0.0 [Pos19a] external extension
- GIS software
  - QGIS 3.10 [QGI20]
  - GDAL/OGR 2.4.4 [GDA]
  - PROJ 6.3.0 [PRO]
- Model building and evaluation
  - RStudio 1.2.5033 [RSt19]
  - R 3.6.1 [R C19]
  - H2O 3.30.0.1 [H2O20]
  - DALEX 1.2.0 [Bie20a]
- Documentation
  - Bookdown [Xie20a]
  - knitr [Xie20b]
  - R Markdown [Xie20c]
  - MacTeX [Mac20]
- R packages
  - data.table 1.12.8 [Dow19]
  - funModeling 1.9.3 [Cas19]
  - lubridate 1.7.8 [Spi20]
  - magrittr 1.5 [Mil14]
  - sf 0.9-2 [Peb20]
  - tidync 0.2.3 [Sum19b]

  - ncmeta 0.2.0 [Sum19a]
  - tidyr 1.0.2 [Wic20d]
  - dplyr 0.8.5 [Wic20a]
  - raster 3.1-5 [Hij20]
  - ggplot2 3.3.0 [Wic20b]
  - scales 1.1.0 [Wic19]
  - patchwork 1.0.0 [Lin19]
  - MASS 7.3-51.5 [Rip19]
  - colorspace 1.4-1 [Zei19]
  - reshape2 1.4.4 [Wic20c]
  - ggcorrplot 0.1.3 [Kas19]
  - minerva 1.5.8 [Fil19]
  - ingredients 1.2.0 [Bie20b]
- Support software
  - Git 2.21 [Git20]

Part I
State of the art of XAI

In this part I am going to review the recent ideas and methods of XAI. It is a field currently ongoing a fast evolution with a large number of articles and new methods appearing in recent times due to the increasing use of automated decision system.

Chapter 2

On eXplainable AI (XAI)

2.1 The need of XAI

The widespread adoption of automated decision-making systems across large parts of society is raising questions on human agency and liability [Wag19]. We cannot be comfortable by blindly accepting the decision of an automated system without any understanding of its rationale. To hold accountable and trust automated decision-making systems detailed explanations of its decisions are necessary.

The ability to explain the why behind our decisions to another person is an important aspect of social interaction between humans. Moreover, it is seen as a prerequisite for establishing a relationship based on trust.

Hence, for a human to trust an automated decision-making system it must be explainable, and to be explainable it must be interpretable so a human can understand why the system makes a decision, and not another.

[SWM17] cites as the most important arguments in favour of eXplainable Artificial Intelligence (XAI) the following:

- Verification of the predictions made by a system
- Location and improvement of the weakness of a system
- Distil knowledge from the system
- Legal compliance

2.2 Human agency over automated systems

One of the key factors necessary to a human-centric approach to AI is ensuring that a human is part of the decision and control loop [She00]. According to [RM16], human agency over an automated decision-making system is enhanced when:

- The system is predictable, reliable, and transparent
- A human can assess whether the system is fulfilling its goals
- A human can act in a timely fashion
- There is accountability of actions (and inactions)

Predictability and reliability are the primary metrics by which human measure whether a system is working correctly. However, predictability of the decisions made by any system depends on understanding the context where it operates.

Transparency means that a human can question the system whether it is working towards its goal inside the constraints placed upon it. Hence, the system must provide enough feedback to its human operator, so the latter is sufficiently aware of its internal processes.

Predictability, reliability, and transparency augment human agency over an automated system but they do not guarantee control over it, which depends on understanding the outcomes being sought and the environment where the system operates. Furthermore, the goals and environment are produced by and exists in a large socio-political system, and we must recognise that the degree of autonomy of a system is not only technical product but also a political one.

Even when an automated system has been designed to operate faster than human capacity, the human operator has still been able to take timely action instead of becoming only a witness.

Accountability serves as the basis upon which build the framework of expectations and responsibilities regulating the operation of an automated system by a human operator. The framework should also incorporate the ethical standards and sanctions guiding the goals and operation of the system by its human operator.

From these characteristics that enhance human agency over an automated system emerge the following limitations [Sch18]:

- Cognitive limitations
- Epistemic limitations
- Temporal limitations

Cognitive limitations arise from the two types of reasoning used by humans to make decisions:

- Deliberate (conscious and slow) reasoning, used for decisions of considerable weight
- Automatic (unconscious and fast) reasoning, used to routine events

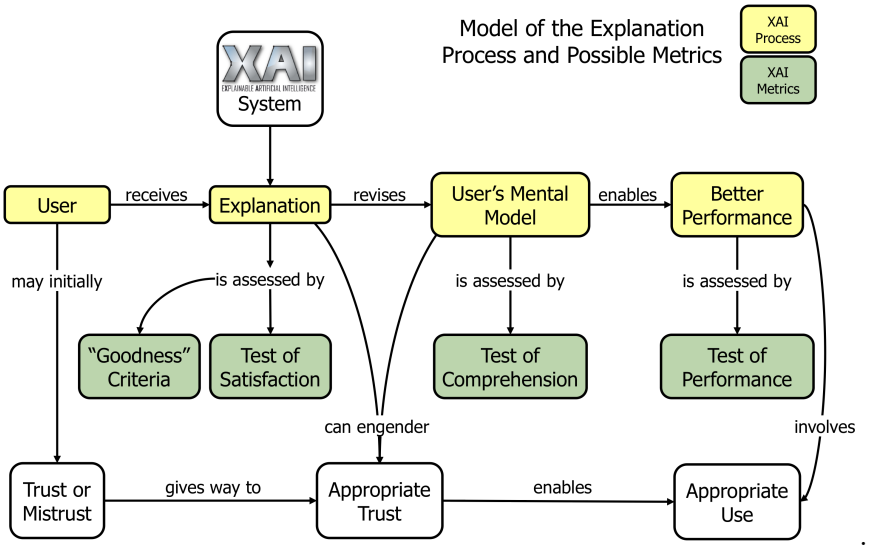


Figure 2.1: XAI Psychological Model of Explanation [Gun19]

Automatic reasoning is the default option, but it gives way to deliberate reasoning in novel situations and when a task requires active attention. It is also the default option when a human interacts with an automated system. Furthermore, the threshold to override automatic reasoning and the trust put on the automated system will increase with the speed and autonomy exhibit by it. As a result, human agency is diminished.

Some of the characteristics of automated reasoning makes it dangerous when we must make important decisions [Sha18]:

- Neglects ambiguity and suppresses doubt, it jumps to conclusions without searching for alternative interpretations
- Infers and invent causes and intentions, it links fragments of information to invent causal stories
- Is biased to believe and confirm, it favours uncritical acceptance bias
- Focuses on existing evidence and ignores absent evidence, it builds stories without consideration of missing information creating a false feeling of confidence

The reliance on systems trained on massive and highly processed quantities of data obscures data provenance and diminishes the epistemological understanding of the human operator. This problem is compounded by the use of intrinsically impenetrable black-box Machine Learning (ML) models.

Speed is one of the main allures of autonomous systems. However, as the time horizon available to the human operator to act shrinks it also does human agency over the system.

These limitations together with regulatory grey areas where laws regulating automated systems are not applied if there is a human implicated have produced "quasi-automated" system where the operator only task is to bear witness to the operation of the system without possibility of intervention.

[Wag19] suggest a series of criteria to define when is more likely the "quasi automation" of a system:

- Low amount of time assigned to the operator, so there is little or no time for intervention
- Low degree of qualification of the operator, so they cannot exert any meaningful control
- Great amount of legal liability assigned to the operator, so they serve as a scapegoat
- Low level of support received by the operator, so there is no need to support operators taking high-stakes decisions
- Great level of adaptation by the operator to the system, since the system is not designed with an operator in mind
- Small volume of information available to the operator, since they do not have to make the decision
- Little authority by the operator to change the outcome of the system, so they cannot change the decision of the system

However, regulation must acknowledge that neither human nor systems are autonomous agents and consider the socio-technological environment in which they developed their activity.

Placing all the blame in the operator or the system promotes the use of humans as scapegoats put in charge of fully automated system to evade limitations and safeguards imposed by the legal system.

Instead, regulators should focus on organisational behaviours and procedures that led to breaking laws and violating rights.

2.3 Responsible AI

The widespread use of AI has raised concerns about potential problems related to discrimination, interpretability, transparency, liability, and malicious use.

[Benjamins2019] presents a series of AI principles on how it should be developed and used so it respects human rights and makes society more inclusive. These principles are:

- Fair AI ensuring fair results and avoiding discriminating against people based on race, ethnic origin, religion, gender, sexual orientation, disability, or any other personal condition.

- Transparent AI explicitly describing the personal and non-personal data used and its purpose.

- Explainable AI enabling a certain level of understanding of the decisions of an AI system by generating explanations on how it reached that decisions and no others.

- Human-centric AI meaning that it should be at the service of society and always under human control.

- Privacy and security by design during the whole life cycle.

The General Data Protection Regulation (GDPR) [Eur16] deals with how data is collected and stored. However, "Article 22: Automated individual decision-making, including profiling" aims to enshrine these AI principles into law:

> Article 22
>
> Automated individual decision-making, including profiling
>
> 1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
> 2. Paragraph 1 shall not apply if the decision:
>   (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
>   (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
>   (c) is based on the data subject's explicit consent.
> 3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

The GDPR in Article 4 defines the following key terms:

- Personal data is "any information relating to an identified or identifiable natural person"
- Data subject is the natural person to whom data relates
- Processing is "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means"
- Profiling is "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person"

Thus, profiling is the subset of processing when this is automated, and its purpose is evaluation.

The regulation pays attention to profiling aimed at analysing or predicting a natural person behaviour, movements or location, economic situation, health, performance at work, personal preferences, and interests. It prohibits any decision based solely on automated processing that can significantly affect a natural person (albeit with some exceptions).

Although some claim that big data is neutral, it depends on data collected that might show traces of inequality, exclusion, or other kinds of discrimination as it reflects the population that generates it. Furthermore, it does not matter how objective is the ML algorithm or how accurate is the model if the data contains patterns of discrimination so it the decisions of the model.

The legislation puts the onus on the data processor to guarantee that the data using to build models is not discriminatory. However, this poses a challenge as data sets become increasing complex, large, and processed.

The GDPR in Articles 13 and 14 state that a person has the right to "meaningful information about the logic involved" in profiling of their data. This demand for transparency in automated decision-making faces three barriers [Bur16]:

- Intentional concealment by the data processor
- Gaps in technical literacy by the data subject
- No meaningful human-readable explanations

2.4 Explainability and interpretability

Explainable models output produce both predictions and insights about what caused their decisions. All explainable models are interpretable, but not all interpretable models are explainable [Gil+19].

Interpretability is a passive characteristic of a model that describes how intelligible it is to a human. Explainability is an active characteristic of a model that describes the actions taken to clarify its internal process [Bar+19].

An explanation can be evaluated according to its interpretability or to its completeness. Interpretability measures the degree to which humans can understand the internals of an automated system. It depends on the cognition, knowledge, and biases of humans. Completeness measures the accuracy of the description of how a system operates.

It is difficult to achieve interpretability and completeness at the same time. The most complete (i.e., accurate) explanations are usually hard to interpret. And, the most interpretable explanations are usually incomplete (i.e., not accurate).

However, and causing confusion, interpretability and explainability are often used interchangeably or conflated with another concepts. Some common terms included in the literature of XAI are [Bar+19]:

- Understandability (or also intelligibility) is the characteristic of a model that makes it understandable to a human without the need to explain its internal processes and structures. In other works that a human has a functional understanding of a model without needed to posses a low-level mechanistic or algorithmic understanding of it.

- Comprehensibility is the ability of an ML algorithm to represent knowledge in a way that makes it understandable by human. However, different stakeholders (those who do the comprehending) might have different requirements and abilities, so representations are often specific to certain stakeholders. Also, comprehensibility is not a goal but a tool used to reach one and decide whether it is appropriate to the problem.

- Interpretability is the ability to explain or provide meaning understandable by humans.

- Explainability is the characteristic of a model that makes it functioning clear and easy to understand.

- Transparency is the characteristic of a model of being understandable. In other words, understanding the variables included in the model and the interaction between them.

This lack of agreement on the terms and definitions in the XAI field is one of the obstacles in the search for a unifying theory of explainability that serves as a guide for every XAI system.

Another challenge lies in the fairness and ethics of the models where XAI could be used as a diagnostic tool to assess their internal process. Furthermore, it can also help with the reproducibility of models beyond the simple sharing of data and result.

However, the ultimate challenge faced by XAI is to provide explanations accessible for non-technical stakeholders such as police-makers, legislators, and the public in general. This scenario is the consequence of the increasing presence of automated systems and a necessity with the introduction of future regulations anticipated by the "right to explanation" laws proposed by the European Union [Ham+20].

And, involving non-technical people from other fields would allow to create multi disciplinary teams improving the development and appraisal of models.

2.5 Approaches to explainability

[DSB17] proposes the following classification of automated decision-making systems based on explainability:

- Opaque systems: the mapping of inputs to outputs is invisible such as models based on black-box algorithms.
- Interpretable systems: the mapping of inputs to outputs is visible so a human can examine the model such as linear regression where the importance of each feature can be interpreted by comparing the covariate weights.
- Comprehensible systems: they emit symbols (words or graphics) that allow a human to interpret and comprehend the mapping of inputs to outputs.

The difference between interpretable systems and comprehensible systems is that the former require the transparency of the underlying algorithms (i.e., glass-box models), and the later can be opaque (i.e., black-box models) but emit symbols that a human can reason over.

As stated in [Rud19], the often repeated and in many times unchallenged claim that more complex models are more accurate is false. Even though, it is true that more complex models are usually more flexible, thus allowing the approximation of more complex functions. Furthermore, it is when the function is complex and the data badly structured or of lower quality when there is a trade-off between interpretability and performance.

[Hol+17] and [Lip17] propose the following classification of approaches to explainability:

- Ante-hoc (or transparent) models, that intrinsically incorporate explainability. This approach is limited to algorithms with low complexity such as linear regression or decision trees, in contrast to others like artificial neural networks whose complexity renders them opaque.
- Post-hoc explanations, that explain predictions (e.g., Local Interpretable Model-Agnostic Explanations, LIME). This approach extracts information from already trained model and does not depend on how the algorithm used to build the model works, treating it as a black-box.

Models are transparent when:

- A human comprehends the whole model at the same time. However, given the cognitive capacity of human beings a sufficiently high-dimensional linear model or unwieldy decision rule list cannot be considered simple.
- Each of its parts (i.e., inputs, hyperparameters and calculations) admit an intuitive explanation. This requirement disqualifies models with heavily pre-processed or anonymous features.

Post-hoc explanations extracts information from trained models so they can be used to interpret opaque models.

Part II
Case of study: Predicting burnt area caused by human-caused fires

The case of study involves developing a regression model to predict the burnt area caused by human-caused fires (HCFs) in continental Portugal. The case of study consists on the following high-level stages:

- Business understanding
- Data understanding
- Modelling

Chapter 3

Human-Caused Fires (HCFs)

3.1 Extent and causes

Forest disturbances are the environmental fluctuations and destructive events that disturb forest health and/or structure and/or change the resources or the physical environment at any spatial or temporal scale [FU06]. Healthy forests disturbances caused by agents such as fire, insect pests and diseases are an integral part of the ecosystem.

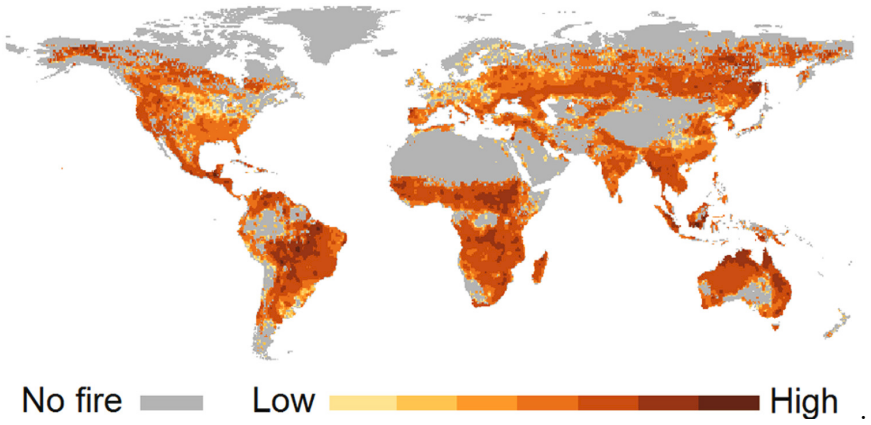


Figure 3.1: Count of observed fire occurrence readings from combined MODIS and ATSR remote sensing products from 1996 to 2007 [Mor+12]

However, the growth of human population leads to a conversion of natural vegetation to agricultural and pastoral systems, alongside the development of supporting infrastructure (e.g., roads). This transformation totally changes the fire regime, the natural frequency and seasonality of fire.

Moreover, catastrophic disturbances can cause profound impacts on forest ecosystems and adversely affect biodiversity, livelihoods, and climate. Therefore, an accurate assessment of the size and scope of forest distur-

bances is critical for monitoring the ecosystem, especially in the face of changing climatic conditions.

The Food and Agriculture Organization of the United Nations (FAO) aggregates information on forest disturbances in its periodic Global Forest Resources Assessment (FRA). Data from its report of 2015 covering the ten-year period from 2003 to 2012 showed that $> 90\%$ of the total land area and $> 99\%$ of the total forest area suffered fire disturbances [Lie+15]. An average of 341 million hectares (2.6%) of all land area was burned annually with 67 million hectares (1.7%) corresponded to forests. Moreover, total land and forest area burned over this period exhibit a decreasing trend.

By climatic domain, the largest area of land and forest burned was in the tropics, where more than 290 million hectares of land burned annually (between 79% and 91%) with over 53 million hectares were in forest land. This is cause of great concern, as tropical forests are home to more than half the biological diversity on Earth and support the livelihoods of hundreds of millions of people. Furthermore, tropical forests are extremely important for maintaining local and global weather and climate.

The largest area of land burned was in Africa with over 213 million hectares annually. And, of this amount, nearly 17 million hectares were in forest land. Also, nearly all of the land burned in North and Central America (total 5 million hectares) and Europe (total 3 million hectares) was forest land. The largest area of land burned was in Africa with over 213 million hectares annually. And, of this amount, nearly 17 million hectares were in forest land. Also, nearly all of the land burned in North and Central America (total 5 million hectares) and Europe (total 3 million hectares) was forest land.

By income category, indicate that the smallest area of land burned annually corresponds to high-income countries, 38 million hectares which approximately 13 million hectares were forest land.

Worldwide, people causes most fires, in the Mediterranean region the estimated proportion is 95% [FU07]. These fires are usually termed as "human-caused fires" (HCFs) and they encompass intentional and unintentional human actions, power lines and machinery. In opposition of "natural" fires that are caused by lightning or local phenomena such as volcanic eruptions or earthquakes.

The list of human-induced causes includes land clearing and other agricultural activities, maintenance of grasslands for livestock management, extraction of non-wood forest products, industrial development, resettlement, hunting, negligence, and arson [FU07]. The movement in recent years of people from rural areas to cities have resulted in worse fire management and increased fuel levels in rural areas, and consequently an increase in the risk and severity of fires.

The damage caused by fires common to most regions worldwide include environmental damages (e.g., forest degradation, soil erosion, loss of biolog-

ical diversity, ...), effects on the climate through the emission of greenhouse gases, the loss of lives and the impact on livelihoods, especially in rural and poor regions.

3.2 Fire Management

Fire prevention depends on the country but usually it targets people since they are the main cause of fires. Also, most countries have laws regulating the how and when setting fires, and fire prevention programmes. But, few of them have the capacity to enforce the laws or effectively manage the programmes.

Many countries, especially in the Mediterranean regions, implement measures to reduce fuel levels through controlled burnings or grazing. And, early warning systems are increasingly being used to anticipate periods of high risk. However, in other countries there is a neglect of prevention measures where most of the budget goes to suppression activities.

The suppression of fires starts with their detection through early warning systems or increasingly the use of satellite and aerial surveillance. Once a fire its detected, it is extinguished primarily by ground-based forces often reinforced by aerial units used for suppression and transport of ground forces.

Fire management responsibility varies from country to country often divided between two trends in institutional arrangements:

- The forest service is the sole responsible for fire prevention and control
- The forest service is only responsible for fire prevention and the fire service takes over the suppression role

The second arrangement represent a focus more on crisis-response than on prevention and management. A cause is the increasing urbanization of certain regions and consequently a lower community-participation. Also, the bigger number of agencies involved in fire activities might become a constraint due to a higher cost of integration and coordination of their efforts.

At the transnational level, a main area of collaboration is the research and development of technologies for fire management. Furthermore, the development of early warning systems has become especially urgent in preparation to deal with the effects of climate change on fire regimes across the world.

3.3 Fire patterns and risk factors

Historical knowledge of the conditions during wildfires is important in the case of HCFs that show identifiable spatial and temporal patterns [Cos17].

Spatial patterns are caused by physiography and socio-economic factors and temporal ones are caused by the climate. The identification of these patterns jump-started the interest in developing models, both binary occurrence and numerical, of HCFs occurrence since 1950. Fire occurrence modelling tries to identify the biotic and abiotic factors contributing to fire ignition.

Even though fire occurrence may differ seasonally there is an identifiable seasonal pattern. Whereas in some regions there is a high peak of fire occurrence in summer, others show two peaks in early winter and summer. After ignition fires grows being its size constrained by topography, available fuels, wind, and the suppression efforts.

Fire risk depends on the presence of ignition sources and environmental conditions. The later can be divided into two groups:

- Temporal factors: based on weather
- Spatial factors: derived from physiography, land cover / use, human-derived (e.g., distance to nearest settlement)

The weather factors are included in models through variables such as high and mean temperatures, precipitation, relative humidity, evapotranspiration, and insolation.

It is common to use indexes that estimate the moisture content caused by weather on diverse types of fuels and in the soil layers. The most-well known is the Canadian Forest Fire Weather Index (FWI) System [Nat]. It consists of six indicators accounting for the effects of fuel moisture and wind on fire behaviour. Its indicators are divided in the fuel moisture codes and the fire behaviour indices. The former rate the moisture content of different types of fuel, and the later rate different characteristics of fire.

Physiography variables considered are elevation and slope. HCFs risk increases with decreases in elevation and slope as they tend to occur in lowlands and gentle slopes where population tends to live.

Human factors are included in models since HCFs are caused directly or indirectly by people. Moreover, landscape structure is the result of the interaction between human and natural processes. Socio-economic activities influence in the number and distribution of human presence. Accessibility through roads or tracks is associated with an increase in fire occurrence, as is proximity to high populated areas. Also, proximity to agricultural activities or outdoor recreational areas increases the risk of fire.

Major Habitat Types (MHTs) [Ols+01] are used in many works to geographically stratify the area of study. MHTs are characterised by climate and influence the importance of other risk factors since the type and distribution of vegetation is characteristic of each MHTs.

3.4 Wildfires in Portugal

Portugal is a country especially impacted by wildfires. It is one of the countries with one of the highest fire risks in Europe. In the decadal average increased from under 75,000 ha during the 1980s', to 100,000 ha in the 1990s', to over 150,000 ha since 2000 [BH18]. The upwards trend continued until 2017 when severe drought, heat waves, massive oceans of flammable forests and scrublands, and the Hurricane Ophelia in mid-October came together in a "perfect storm" situation when 520,000 ha burned, and 120 lives were lost [Tur+19].

The causes of why Portugal has found itself in this situation are [BH18]:

- A high percentage of unmanaged forest lands
- An increase in the amount and extent of fuel loads
- A high number of human-caused ignitions during high risk periods
- An increase of periods of hot and dry weather that both lengthen and increase the severity of critical periods for extreme fire caused by climate change

Particular to Portugal is a highly variable annual burn area pattern denominated by [Per+05] as the "asymmetric nature of fire size distribution" with alternating years of higher highs and lower lows that puts an extreme stress on the environment.

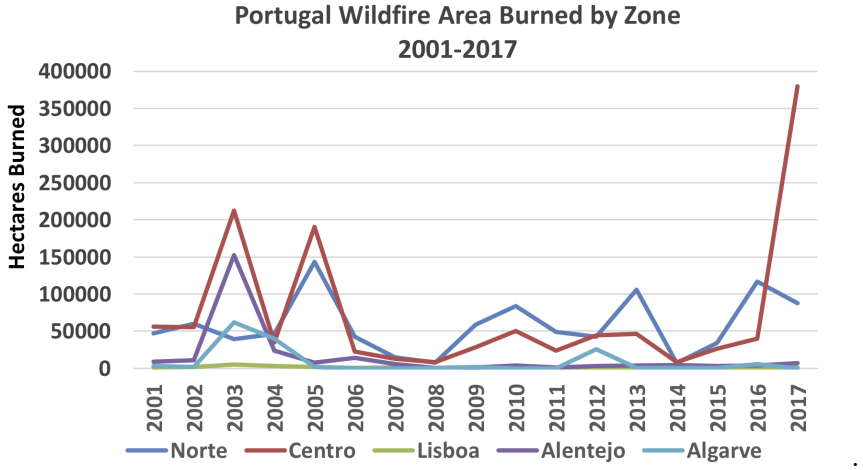


Figure 3.2: Inter-annual variability of area burnt by NUTS II region in continental Portugal

Climate models paint a grim situation in the coming years, particularly for Portugal and other southern European countries. Hotter years with lower

precipitations have become normal and this trend shows no sign of abating in the short-term future contributing to the increase in number and severity of wildfires.

Chapter 4

Data acquisition

In this section I am going to describe the different datasets and their sources that I am going to use in the case study. Although I am going to only include a short description of each dataset, the complete metadata can be found in the Appendix B.

The decision to use these specific datasets and not other was driven primarily by availability of sources that covered the different aspects deemed important by the literature in modelling human-caused fires (HCFs).

I am going include a summary of each dataset that I am going to use and the motivation of its inclusion:

- Human-Caused Fires (HCFs) in Portugal from 2011 to 2015. Fires have been a scourge for the Mediterranean countries, especially for Portugal and Spain with a climate that contributes to its occurrence.
- Major Habitat Types (MHTs). Used in many studies to stratify the data since the factors contributing to the occurrence of wildfires vary from one habitat to another.
- Weather factors. A major factor influencing the occurrence of wildfires together with the physiography.
  - Meteorological data from 2011 to 2015. As climate patterns shape the temporal patterns exhibit by wildfires.
  - Forest Fire Weather Index (FWI) data from 2011 to 2015. Derived from meteorological data, it is widely used to model the risk of wildfires and in predictive models.
- Physiography. A major factor influencing the occurrence of wildfires together with weather factors. Different elevations and slopes determine the vegetation that exists in one are and, as a result, the fuel load available for wildfires.

  - Elevation
  - Slope

- Fuel risk factors through the land cover in continental Portugal. One of the factors used as proxy for human activity that influences the occurrence of human-caused fires (HCFs).

- Human factors. Factors used to model locations favoured by people that deliberately or not ignite fires.

  - Distance to nearest road
  - Distance to nearest build-up area

4.1 Human-Caused Fires (HCFs)

To cover the explanatory variable, that is Human-Caused Fires in Portugal, I am going to use the historical fire data from the website Central de Dados[1] covering the period from 1980 to 2015. The website also mirrors all its data in its repository on GitHub[2].

The format of the downloadable files is Comma-Separated Values (CSV) with one file per year. Even though not all files contain the same variables the necessary subset I need for this project is common to the files belonging to the period of study selected, 2011–2015.

The original source of the data is the Instituto da Conversão da Natureza e Florestas (ICFN) [Rep20] of the Portugal Republic government. The following changes have been made to the original data set:

- Use of the ISO-8601[3] (YYYY-MM-DD) for the dates
- Harmonisation of column names
- Deletion of unnecessary columns
- Deletion of quote characters
- Deletion of empty hour field from the dates
- Unification of hour and minute columns into a single column
- Deletion of NULL values
- Unification of end-of-line format

[1] `http://centraldedados.pt`
[2] `https://github.com/centraldedados/`
[3] `https://www.iso.org/iso-8601-date-and-time-format.html`

- Use of UTF-8 to encode the files

From all the variables included in the dataset I am going to select the following:

- x: easting coordinate expressed using EPSG:20790 [Kloa] in meters
- y: northing coordinate expressed using EPSG:20790 [Kloa] in meters
- data_alerta: date of fire detection
- area_total: total area burned (in hectares)

Additionally, I am going to use to filter the raw data set two more variables to discriminate between natural and human-caused fires:

- falso_alarme: whether the fire was a false alarm
- tipo_causa: classification of the source of the fire

4.2 Major Habitat Types (MHTs)

Many studies use MHTs to stratify the data [Cos17]. The source of the MHTs data is the Global Map of Terrestrial Ecoregions determining Major Habitat Types (MHTs) [Olson2001], which can be downloaded from the web of the World Wildlife Fund (WWF)[4].

Ecoregions can serve as a framework for analysing biodiversity patterns, assessing conservation priorities, and directing effort and support.

The originator of the data set is the WWF and in its 2.0 version from 2004 (this is an update to version 1.0 which was completed in 2001).

The data is publicly available as a compressed file (in the ZIP format) containing one vector data set encoded using ESRI Shapefiles format. It contains the ecoregions and biomes covering the entire world.

The data set divides the world into 827 terrestrial ecoregions nested within two higher-order classifications: 14 biomes and 8 biogeographic realms. Together, these nested classification levels provide a framework for comparison among units.

Furthermore, it has been demonstrated that the fire risk factors and their importance change between MHTs [Cos17].

The ecoregions are categorised within 14 biomes and eight biogeographic realms:

[4] `https://www.worldwildlife.org/publications/terrestrial-ecoregions-of-the-world`

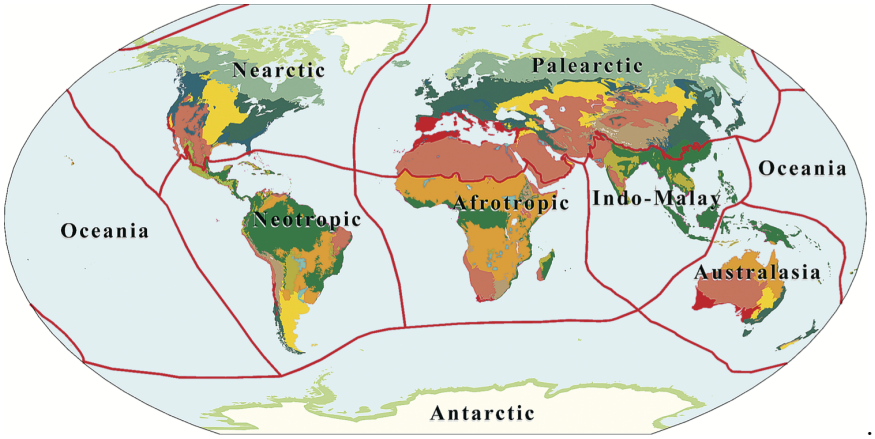


Figure 4.1: Biomes and biogeographic realms of the world

Nested inside the biomes are subdivided into 867 distinct terrestrial ecoregions:

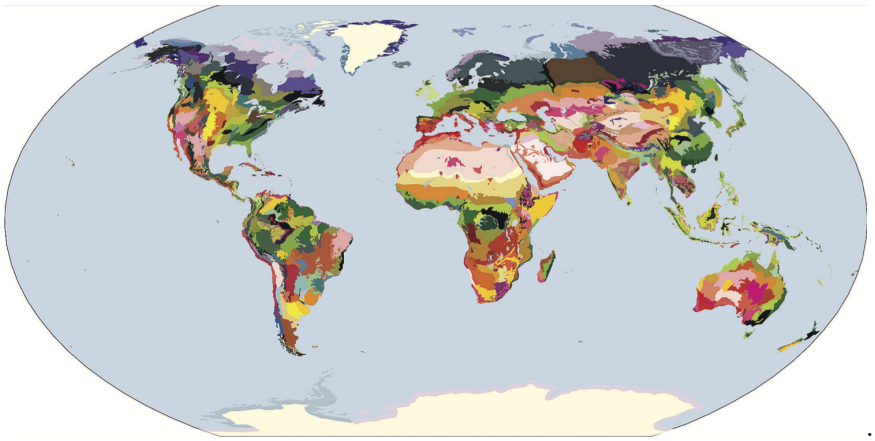


Figure 4.2: Terrestial ecoregions of the world

From all the variables included in the dataset I am going to use the biomes to stratify the data. The biomes are represented by two variables:

- a geometry, in this case multipolygons
- a number encoding to which one of the 14 biomes belongs the geometry

4.3 Weather factors

4.3.1 Meteorological data

Meteorological variables influence physical combustion requirements. High mean and maximum temperatures, low precipitation, and low relative hu-

midity are some of the meteorological conditions favouring the occurrence of fire. However, annual (or event larger) weather patterns also influence the number of fires.

I have chosen as the source for the meteorological data the gridded agro-meteorological data published by the Monitoring Agricultural Resources (MARS) [Eur15] project for data gathering used in the implementation of the EU's Common Agricultural Policy (CAP). Weather monitoring is part of the crop monitoring and yield forecasting activity within the AGRI4CAST / MARS4CAST (i.e., Agricultura forecast / MARS forecast) project.

The data is published as part of the AGRI4CAST data by the MARS4CAST project through its data portal[5]. It is published as files split by geographical cover as files using the CSV format.

It contains meteorological parameters from weather stations interpolated on a 25×25 km grid. Meteorological data are available on a daily basis from 1975 to the last calendar year completed, covering the EU Member States, neighbouring European countries, and the Mediterranean countries.

Figure 4.3: AGRI4CAST grid with location of observations

I am going to include all the variables in the data set:

- TEMPERATURE_MAX: Maximum air temperature (°C)
- TEMPERATURE_AVG: Mean air temperature (°C)
- WINDSPEED: Mean daily wind speed at 10m (m/s)

[5] `https://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx`

- VAPOURPRESSURE: Vapour pressure (hPa)
- PRECIPITATION: Sum of precipitation (mm/day)
- RADIATION: Total global radiation (KJ/m^2/day)

To combine the variables with the other data I am going to use the following variables form the data set:

- LONGITUDE: easting coordinate expressed using EPSG:4326 [Klob] (decimal degrees)
- LATITUDE: northing coordinate expressed using EPSG:4326 [Klob] (decimal degrees)
- DAY: date of the observation

To access and download the data you need to create and account in the service. Once you have created and logged in, the data can be selected and downloaded as a single file using the CSV format.

4.3.2 Canadian Forest Fire Weather Index (FWI)

The Canadian Forest Fire Weather Index (FWI) System rates fire danger. It consists of six indicators accounting for the effects of fuel moisture and wind on fire behaviour [Nat].

The FWI indicators are divided in the fuel moisture codes and the fire behaviour indices. The former rate the moisture content of different types of fuel, and the later rate different characteristics of fire.

The fuel moisture codes are [Gro87]:

- Fine Fuel Moisture Code (FFMC): It rates the moisture content of litter and other cured fine fuels. And, it is an indicator of the relative ease of ignition and the flammability of fine fuel. Its value range is $[0, 10]$
- Duff Moisture Code (DMC): It rates the average moisture content of loosely compacted organic layers of moderate depth. And, it gives an indication of fuel consumption in moderate duff layers and medium-size woody material. Its value range is $[0, +\infty)$
- Drought Code (DC): It rates the average moisture content of deep, compact organic layers. And, it is a useful indicator of seasonal drought effects on forest fuels and the amount of smouldering in deep duff layers and large logs. Its value range is $[0, +\infty)$

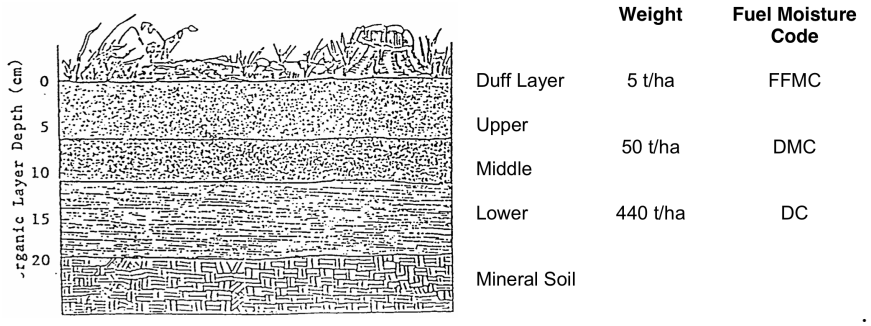


Figure 4.4: Forest floor fuels by fuel moisture codes of the FWI system

The fire behaviour indices are:

- Initial Spread Index (ISI): It rates the expected rate of fire spread. It combines the effects of wind and the FFMC on rate of spread without the influence of variable quantities of fuel. Its value range is $[0, +\infty)$
- Build-Up Index (BUI): It rates the total amount of fuel available for combustion. It combines the DMC and the DC. Its value range is $[0, +\infty)$
- Fire Weather Index (FWI): It rates fire intensity. It combines the Initial Spread Index and the Build-Up Index. It is suitable as a general index of fire danger. Its value range is $[0, +\infty)$

The interdependency between the meteorological variables and the indices is:

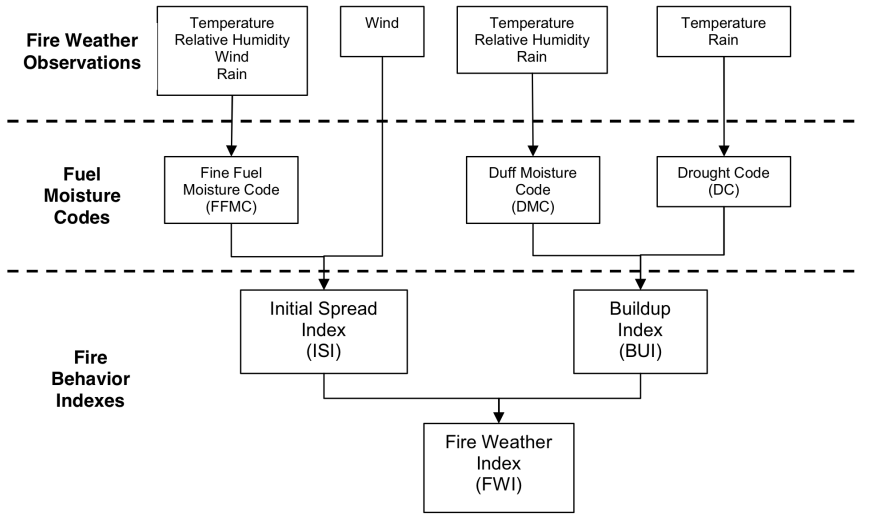


Figure 4.5: Structure of the Canadian Forest Fire Weather Index System

I have chosen as the source for the FWI data the Fire danger indices historical data from the Copernicus Emergency Management Service, part of the Copernicus Programme[6]. The data set is produced by the Copernicus Emergency Management Service for the European Forest Fire Information System (EFFIS).

Copernicus is a European programme for monitoring the Earth. Data is collected by Earth observation satellites and combined with observation data from sensor networks on the Earth's surface. Once collected the data is processed, providing reliable and up-to-date information within six thematic areas. These areas are: land, marine, atmosphere, climate change, emergency management and security.

The FWI indices from this data calculated using weather forecast from historical simulations provided by ECMWF ERA5 reanalysis combining model data and a set of quality-controlled observations.

The data sets span from 1979 to the present with global coverage.

Opening an account is necessary to download the data. Once logged into the system you can select which of the indices to download and the time span.

The downloaded data consists of one compressed file (using the ZIP format) per index containing one file for each day (one data point per day per index) covering the whole surface of the Earth with an spatial resolution of the data is a grid of $0.25° \times 0.25°$ encoded using the NetCDF format.

Figure 4.6: Subset of the Copernicus FWI grid covering continental Portugal

[6] `https://www.copernicus.eu/en`

4.4 Physiography

4.4.1 Elevation

Usually, HCFs occurrence increases as elevation (and slope) decreases. As altitude increases the vegetation loading and temperature decrease (average variation of $-0.65°/100\,\text{m}$), thus rendering difficult the ignition of fire [Seb+08].

Also, HCFS tend to occur in low elevation and gentle slopes where population tend to cluster, making physiography variables a proxy for human activity. However, this depends on the activity, fires related to pastures and forests are in the mountain areas [Cos17].

The source of the elevation data is the EU-DEM v1.0 (there is a newer 1.1 version, but it has not been validated yet).

EU-DEM is a digital surface model (DSM) of 39 countries of the European Economic Area (EEA). It is a hybrid product based on the Shuttle Radar Topography Mission (SRTM)[7] [Far+07] and the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER)[8] global Digital Elevation Model (DEM) data fused by a weighted averaging approach. The SRTM was carried out in 2000. But the data was released in 2014.

The ASTER mission is a cooperative effort between the National Aeronautics and Space Administration (NASA) and Japan's Ministry of Economy Trade and Industry (METI), with the collaboration of scientific and industry organizations in both countries.

The EU-DEM is available online through the Copernicus Land Monitoring Service (CLMS), part of the Copernicus Programme.

Copernicus is a European programme for monitoring the Earth. Data is collected by Earth observation satellites and combined with observation data from sensor networks on the Earth's surface. Once collected the data is processed, providing reliable and up-to-date information within six thematic areas. These areas are: land, marine, atmosphere, climate change, emergency management and security.

The downloadable data are single band raster with values relating to the actual elevation. The dataset is encoded as GeoTIFF with LZW compression ($1000 \times 1000\,km$ tiles) or DEFLATE compression (European mosaics as single files).

For this project, I downloaded two tiles covering the whole continental territory of Portugal. The data download for each tile is a single compressed file (in the ZIP format) containing the tile together with another two files: one containing metadata (in the XML format), and other overview file (in the OVR format) containing the image pyramids. The overview file allows to view the images quickly and efficiently at a variety of scales.

[7] `https://www2.jpl.nasa.gov/srtm/`

[8] `https://asterweb.jpl.nasa.gov/index.asp`

The tiles are single band rasters with values relating to the actual elevation encoded as GeoTIFF with LZW compression ($1000 \times 1000\,km$ tiles).

Figure 4.7: DEM tiles covering continental Portugal

4.4.2 Slope

Usually, HCFs occurrence increases as slope (and elevation) decreases. As slope increases the vegetation loading and temperature decrease, thus rendering difficult the ignition of fire [Seb+08].

Also, HCFS tend to occur in low elevation and gentle slopes where population tend to cluster, making physiography variables a proxy for human activity. However, this depends on the activity, arson and negligence fires occur most often in flat or moderate slopes [Cos17].

The slope data set is a product derived from EU-DEM version 1.0. It is created by projecting the DEM data onto an Inspire compliant grid of 25 meters resolution and computing a slope raster.

This product can be downloaded in both full European coverage as by $1000 \times 1000\,km$ tiles. All products are provided as GeoTIFF in 25 meters resolution.

For this project, I downloaded two tiles covering the whole continental territory of Portugal. The data download for each tile is a single compressed file (in the ZIP format) containing the tile together with a PDF document explaining the conversion between the values in the values in the data and the slope in degrees.

The tiles are single band rasters with values relating to the actual slope encoded as GeoTIFF with LZW compression ($1000 \times 1000\,km$).

Figure 4.8: Slope tiles covering continental Portugal

4.5 Fuel risk factors

4.5.1 Land cover

Human activities have a high importance in assessing the risk of fire occurrence, especially for HCFs where humans are the direct or indirect cause [Vas+08]. To cover the interaction between socioeconomic activities in the natural and anthropogenic environment I am going to use land cover data. In particular, the CORINE Land Cover (CLC) data from 2018.

Furthermore, the landscape composition and its interaction with fire weather directly influences fire occurrence [MCM17].

The CORINE Land Cover (CLC) inventory was initiated in 1985 (reference year 1990). Updates have been produced in 2000, 2006, 2012, and 2018. It consists of an inventory of land cover in 44 classes. CLC uses a Minimum Mapping Unit (MMU) of 25 hectares (ha) for areal phenomena and a minimum width of 100 m for linear phenomena.

The time series is complemented by change layers, which highlight changes in land cover with an MMU of 5 ha. Different MMUs mean that the change layer has higher resolution than the status layer. Due to differences in MMUs the difference between two status layers will not equal to the corresponding CLC-Changes layer.

The Eionet network National Reference Centres Land Cover (NRC/LC) is producing the national CLC databases, which are coordinated and integrated by EEA. CLC is produced by most countries by visual interpretation of high-resolution satellite imagery. In a few countries semi-automatic so-

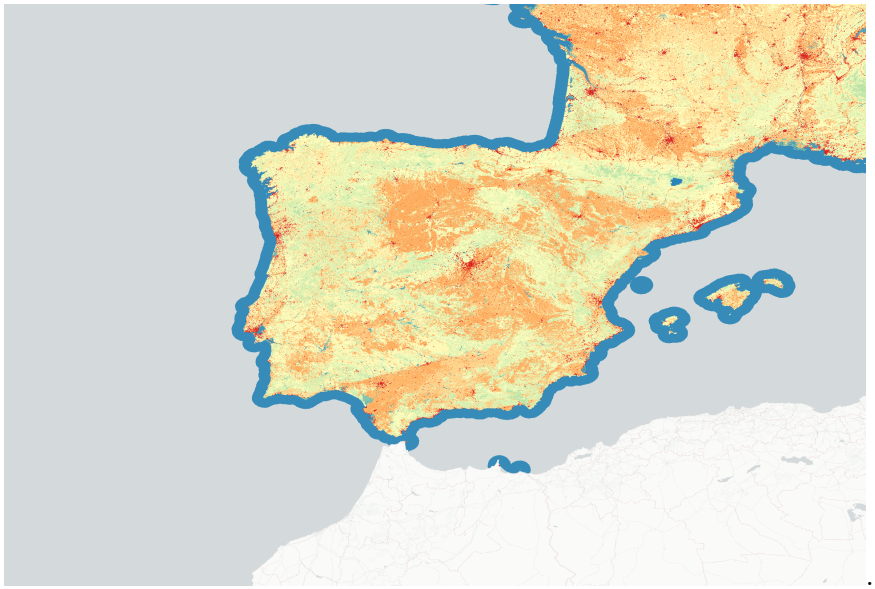


Figure 4.9: CORINE Land Cover centred over the Iberian Peninsula

lutions are applied, using national in-situ data, satellite image processing, GIS integration and generalisation.

The 2012 version of CLC was the first one embedding the CLC time series in the Copernicus programme, thus ensuring sustainable funding for the future. The 2018 version also funded by Copernicus was produced in less than 1 year.

The land classification used by CLC mixes land cover and land use categories [FCW05]. Land cover is the physical material at the surface of the earth. It is determined by direct observation. Whereas land use is the description of how people use the land. It requires socio-economic interpretation of the activities that take place on that surface.

The CORINE level 1 categories are:

id	Description	Category
1	Artificial areas	Cover
2	Agricultural areas	Cover
3	Forest and semi-natural areas	Cover
4	Wetlands	Cover
5	Water bodies	Cover

Table 4.1: CLC level 1

I am going to use the CORINE level 3 categories, both land cover and land use. The categories from class 1 of level 1 are:

id	Description	Category
111	Continuous urban fabric	Cover
112	Discontinuous urban fabric	Cover
121	Industrial or commercial units	Use
122	Road and rail networks and associated land	Use
123	Port areas	Use
124	Airports	Use
131	Mineral extraction sites	Use
132	Dump sites	Use
133	Construction sites	Use
141	Green urban areas	Cover
142	Sport and leisure facilities	Use

Table 4.2: CLC level 3 corresponding to level 1 class "Artificial areas"

The categories from class 2 of level 1 are:

id	Description	Category
211	Non-irrigated arable land	Use
212	Permanently irrigated land	Use
213	Rice fields	Use
221	Vineyards	Use
222	Fruit trees and berry plantations	Use
223	Olive groves	Use
231	Pastures	Use
241	Annual crops associated with permanent crops	Use
242	Complex cultivation patterns	Use
243	Land principally occupied by agriculture, with significant areas of natural vegetation	Use
244	Agro-forestry areas	Use

Table 4.3: CLC level 3 corresponding to level 1 class "Agricultural areas"

The categories from class 3 of level 1 are:

id	Description	Category
311	Broad-leaved forest	Cover
312	Coniferous forest	Cover
313	Mixed forest	Cover
321	Natural grassland	Cover
322	Moors and heathland	Cover
323	Sclerophyllous vegetation	Cover
324	Transitional woodland-scrub	Cover
331	Beaches, dunes, sands	Cover
332	Bare rocks	Cover
333	Sparsely vegetated areas	Cover
334	Burnt areas	Cover
335	Glaciers	Cover

Table 4.4: CLC level 3 corresponding to level 1 class "Forest and semi-natural areas"

The categories from class 4 of level 1 are:

id	Description	Category
411	Inland marshes	Cover
412	Peat bogs	Cover
421	Salt marshes	Cover
422	Salines	Cover
423	Intertidal flats	Cover

Table 4.5: CLC level 3 corresponding to level 1 class "Wetlands"

The categories from class 5 of level 1 are:

id	Description	Category
511	Water courses	Cover
512	Water bodies	Cover
521	Coastal lagoons	Cover
522	Estuaries	Cover
523	Sea and ocean	Cover

Table 4.6: CLC level 3 corresponding to level 1 class "Water bodies"

The CLC data is available through the Copernicus Land Monitoring Service, part of the Copernicus Programme in three formats: as a 100 meter 2018 raster, using the GeoTiff format, as an ESRI Geodatabase, and as a

SQLite database [SQL20], using the GeoPackage format that can be opened with the QGIS software, a free and open-source cross-platform desktop geographic information system (GIS) application.

I downloaded the GeoTiff raster. In addition, I also downloaded a document describing the CORINE Land Cover nomenclature from the technical library of the Copernicus Land Monitoring Service website[9].

4.6 Human factors

4.6.1 Distance to nearest road

Proximity to roads is associated with an increase in HCFs occurrence as more than half of them start along roads since arsonists or careless people use them [ZLS16]. To cover the distance to nearest road I am going to use data from OpenStreetMap.

I am going to download the data through Geofabrik[10], a consulting and software development German company. It offers excerpts of the OpenStreetMap database per country. Geofabrik updates the data sets daily and I am going to use the one extracted on the 12th of April of 2020.

For some countries as is the case for Portugal, it also offers prepared datasets containing subsets of the data (e.g., ways, buildings, etc.) per country as Esri Shapefiles. The data is available as a single file.

4.6.2 Distance to nearest building

HCFs occur most often near settlements [YHS08]. To cover the distance to nearest building I am going to use data from OpenStreetMap.

As with the road infrasctructure data, I am going to download the prepared data set from Geofabrik updated on the 12th of April of 2020.

(a) Road network detail (b) Buildings

Figure 4.10: Human factors data from OpenStreetMap

[9] `https://land.copernicus.eu/user-corner/technical-library`
[10] `https://www.geofabrik.de/`

Chapter 5

Data pre-processing

In this section I am going to describe the pre-processing done to the acquired data. The objective of the different pre-processing tasks is to prepare the data to be easily consumed in later stages.

The data comes from different sources and has to be made compatible before tackling any data analysis task. Furthermore, even multiple datasets from a single provider can present compatibility problems that make crossing the data for analysis a necessary first step.

At a high level, the tasks performed in this chapter are:

- Filter out data not in the spatial (mainland Portugal territory) and temporal (period from 2011 to 2015) frame of interest
- Normalise the Spatial Reference Systems (SRS) into EPSG:4326
- Converting the raw data from sensors into a more usable form

I am going to output all datasets into tabular form (as a series of values each associated with an observation and a variable).

5.1 Human-Caused Fires (HCFs)

The raw fire data is segregated in a file per year so the first step I take is to combine all files into a single data set. Then, I am going to keep only the HCFs identified by the following conditions:

- the value of the field tipo_causa is either “Intencional” for “Negligente”), and
- the value of the field falso_alarme is zero

After these two steps, the number of fires by year are:

From all the variables included in the dataset I am going to select the following:

Year	Fires	HCFs
2011	35,941	12,759
2012	30,740	12,479
2013	27,372	10,357
2014	11,387	3,642
2015	23,175	9,037
Total	128,615	48,274

Table 5.1: Total number of fires and its human caused-fires subset in continental Portugal by year

- x: easting coordinate
- y: northing coordinate
- data_alerta: date when the fire was detected
- area_total: area burnt by the fire (in hectares)

Both coordinates are expressed in meters using the EPSG:20790 Coordinate Reference System (CRS) [Kloa].

Checking the health of the data set only find small number of problematic cases:

Variable	Frequency of zeros	Percentage of zeros
x	8	0.02
y	8	0.02
data_alerta	0	0.00
area_total	333	0.69

Table 5.2: Frequency of zero values in the fire data set

The data set has no missing or infinity values.

The zeros in the location data (variables "x" and "y") are an error since the projected bounds of the CRS used to encode them (EPSG:20790, which covers continental Portugal) are (78230.9913, 5969.3725, 372846.5568, 577613.4152)[1].

Moreover, checking the rest of observations against the bounding box of the CRS reveal 6 more observations out of the extent.

Given that the total of observations is a small percentage ($< 1\%$) I am going to drop them.

I will also drop the fires with burnt area equal to 0 as they are also a small percentage ($< 1\%$) and ascribable to an input error or a value so small that could have been lost in the ingestion of the data.

[1] `https://spatialreference.org/ref/epsg/lisbon-lisbonportuguese-national-grid/`

I am going to convert the CRS of the coordinates from EPSG:20790 to EPSG:4326 as preparation when I combine all data sets.

After pre-processing the source data I have a fire data set with 47,927 observations and 4 variables.

5.2 Major Habitat Types (MHTs)

I am going to use the biomes in the ecoregions data set ("BIOME" attribute) to stratify the other data so I can build a model per biome.

There are only two biomes in continental Portugal

- Temperate broadleaf & mixed forests ("BIOME" attribute equal to 4)
- Mediterranean forests, woodlands & scrub ("BIOME" attribute equal to 12)

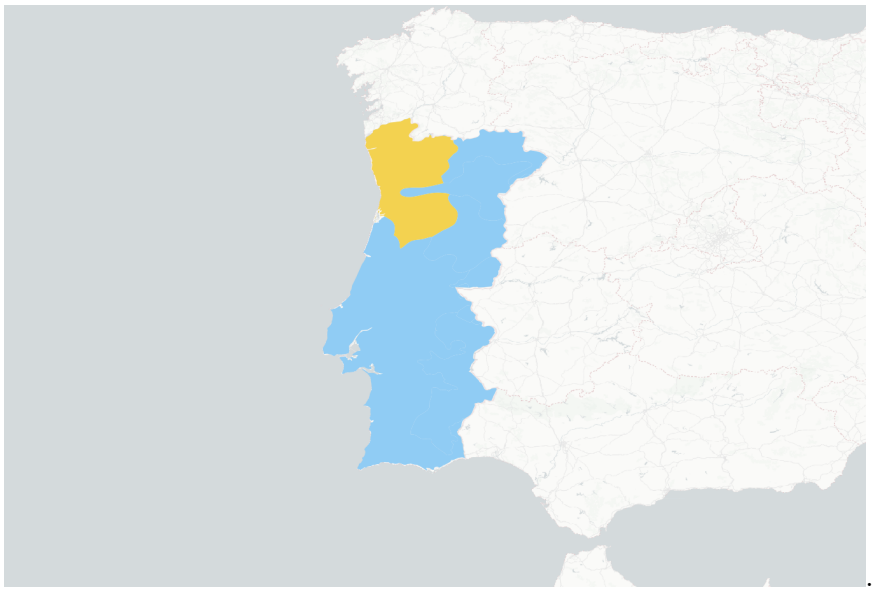


Figure 5.1: Biomes in continental Portugal

The "Temperate broadleaf & mixed forests" biome only covers a small patch in the north-west part corner of the country. The rest corresponds to the "Mediterranean forests, woodlands & scrub" biome.

5.3 Weather factors

5.3.1 Meteorological data

The raw meteorological data is stored in a single file using the CSV format. It contains the following variables:

- LATITUDE: latitude coordinate
- LONGITUDE: longitude coordinate
- DAY: date of the observation
- TEMPERATURE_MAX: Maximum air temperature (°C)
- TEMPERATURE_AVG: Mean air temperature (°C)
- WINDSPEED: Mean daily wind speed at 10m (m/s)
- VAPOURPRESSURE: Vapour pressure (hPa)
- PRECIPITATION: Sum of precipitation (mm/day)
- RADIATION: Total global radiation (KJ/m^2/day)

The latitude and longitude coordinates are expressed in decimal degrees using the EPSG:4326[Klob] CRS.

The raw data contains 345,114 observations from 189 meteorological stations covering the period from 2011 to 2015. This period corresponds to 1,826 days (2012 was a leap year), so:

$$(365 * 4 + 366) * 189 = 345,114$$

Checking the health of the data set only find small number of problematic cases:

Variable	Frequency of zeros	Percentage of zeros
LATITUDE	0	0.00
LONGITUDE	0	0.00
DAY	0	0.00
TEMPERATURE_MAX	2	0.00
TEMPERATURE_AVG	23	0.01
WINDSPEED	51	0.01
VAPOURPRESSURE	0	0.00
PRECIPITATION	239,360	69.36
RADIATION	0	0.00

Table 5.3: Frequency of zero values in the meteorological data set

For temperature data is normal to have variables with observations whose values are equal or less than zero. However, checking the values of the two variables measuring temperatures I find 17 (0.005% of the total) observations where the average temperature is equal to the maximum temperature. This means days where the temperature was constant during the

whole day. Because the percentage of bad observations is so low, I am going to discard them.

The number of zeros for the precipitation variable is high, a 69.36% of all observations. Seems that the variable will be of no use in the training of a model. However, I am not going to discard it yet until I make more tests after a deeper exploration.

After pre-processing the source data, I have a meteorological data set with 345,097 observations and 9 variables.

5.3.2 Canadian Forest Fire Weather Index System

The raw FWI data is stored in a file per index using the NetCDF format. Each file contains data for one of the following indicators:

- Fine Fuel Moisture Code (FFMC)
- Duff Moisture Code (DMC)
- Drought Code (DC)
- Initial Spread Index (ISI)
- Build-Up Index (BUI)
- Fire Weather Index (FWI)

The raw files data sets span from 1979 to the present with global coverage. Therefore, I am going to filter the raw data and only keep the data for the period of study (2011 to 2015) and located inside the bounding box of EPSG:20790. After filtering the data, the data set contains 505,802 observations per indicator.

Checking the health of all the data sets I find that none of them contains missing or infinity values. Some of them contain observations with values equal to 0:

- DMC with 4 observations
- DC with 68 observations
- BUI with 598 observations

These indices can take 0 values as their range is $[0, +\infty)$. Therefore, I am not going to do anything more with the data.

5.4 Physiography

5.4.1 Elevation

I am going to generate an elevation data set by extracting the data from the locations of the fires in the HCFs data set since the rasters are big and compressed making any manipulation computationally costly.

I am going to generate the data set by:

1. Extracting the data separately from the rasters containing the elevation data
2. Combining the elevation data into a single elevation data set

5.4.2 Slope

I am going to generate a slope data set by extracting the data from the locations of the fires in the HCFs data set since the rasters are big and compressed making any manipulation computationally costly.

I am going to generate the data set by:

1. Extracting the data separately from both rasters
2. Combining the elevation data into a single slope data set

As a last step pre-processing the combined slope data I am going to convert the values from the original digital number (DN, a digital number in remote sensing is the value assigned to a pixel in a raster) to the degrees off the horizontal of the surface tangent.

The conversion formula can be found in the product webpage on the Copernicus Land Monitoring Service[2]. It is:

$$slope = \arcsin\left(\frac{DN}{250}\right) \times \frac{180}{\pi}$$

Graphically, the conversion function is:

[2] `https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1-0-and-derived-products/slope`

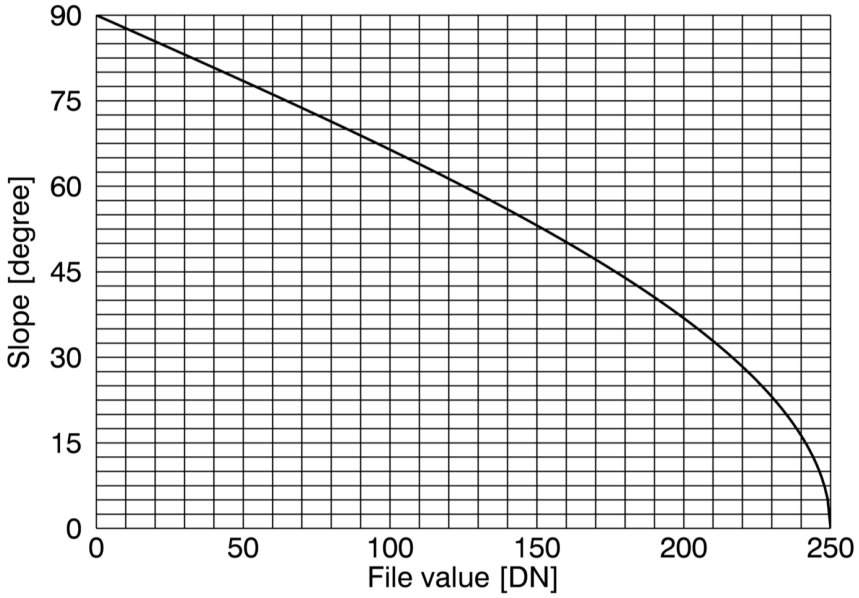


Figure 5.2: Slope conversion curve from DN to decimal degrees over the horizontal

And, the correspondence of some values is:

DN	Slope (degrees)
0	90
25	84.26
50	78.46
75	72.54
100	66.42
125	60
150	53.13
175	45.57
200	36.87
225	25.84
250	0

Table 5.4: Examples of correspondence between the slope as a DN and in decimal degrees over the horizontal

5.5 Fuel risk factors

5.5.1 Land cover

I am going to generate a land cover data set by extracting the level 3 CORINE data from the locations of the fires in the HCFs data set since the rasters are big and compressed making any manipulation computationally costly.

Unlike both physiography data sources, the land cover data is contained in a single raster.

5.6 Human factors

5.6.1 Distance to nearest road

OpenStreetMap classifies as highways [Opea]:

> A highway in OpenStreetMap is any road, route, way, or thoroughfare on land which connects one location to another and has been paved or otherwise improved to allow travel by some conveyance, including motorised vehicles, cyclists, pedestrians, horse riders, and others.

I am going to consider in this work only the main, link and access roads since they are the main ways used by people to move and the OpenStreetMap data contains a large number of non-main highways with diverse degrees of quality and completeness.

The main highways are identified in the OpenStreetMap ontology by one of the following values [Opeb]:

- motorway
- motorway_link
- trunk
- trunk_link
- primary
- primary_link
- secondary
- secondary_link
- tertiary

- tertiary_link
- service

5.6.2 Distance to nearest building

I am not going to pre-process the buildings' data since the quality of the data is not as diverse as it is in the case of the road infrastructure data.

Chapter 6

Data exploration

I am going to describe the data exploration for the whole dataset that I have used trying to understand the data and its characteristics using descriptive statistical and graphical techniques.

The objective of the exploration is to audit the acquired data and gather enough information to process the data so its fit for modelling.

Because, carrying the data exploration process involves the systematic application of multiple techniques to the dataset and the acquired data contains a relatively large number of variables, I am going to describe first which techniques I have used, and second to which variables I have applied them, and a summary of the exploration results.

As I have indicated before, multiple studies that use MHTs to stratify the data [Cos17], and it has been demonstrated that fire risk factors and their importance change between MHTs. Therefore, the biome, as a proxy for the spatial distribution of the data, together with the time dimension, the data belongs to the period from 2011 to 2015, are two possible directions when exploring the data.

I am going to proceed in this way trying to focus the narrative of the data exploration and avoid a long and detailed description of the exploration that would have a detrimental effect on the exposition of this step of the work.

The detailed data exploration is still included in this report but relegated to Appendix C: Detailed data exploration.

6.1 High-level view of a variable by biome

To explore the absolute and relative frequency distribution of a variable grouped by biome I am going to use contingency tables and optionally bar plots. This provides a high-level view of the relation of a variable with the biome.

I have applied this analysis to the following variables:

- Human-caused fires (HCFs)
  - Burnt area
- Fuel risk factors
  - Land cover

From the exploration of these variables is important to highlight that the number of fires and burnt area is almost evenly distributed by biome:

Biome	Count		Burnt area	
Temperate Broadleaf & Mixed Forests	21,590	45.05%	127,331.6	40.55%
Mediterranean Forests, Woodlands & Scrub	26,337	54.95%	186,691.9	59.45%
Total	47,927	100%	314,023.6	100%

Table 6.1: Burnt area observations by biome

And, graphically:

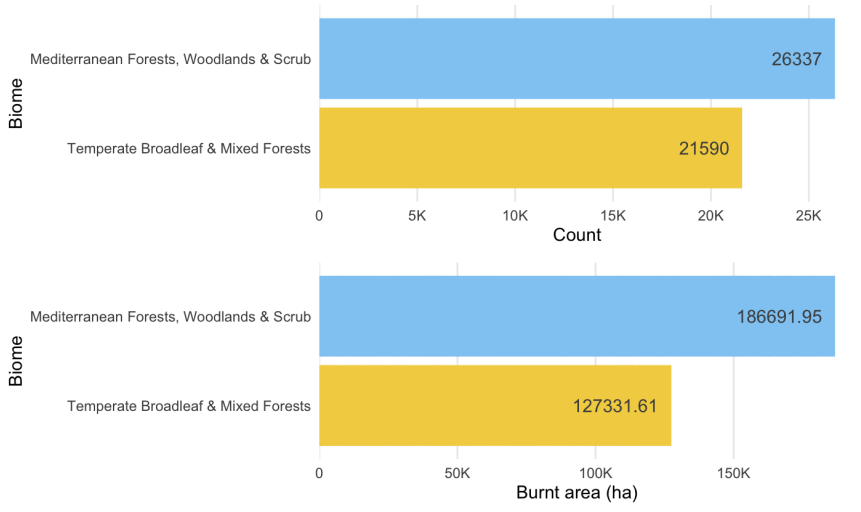


Figure 6.1: HCFs count and burnt area (ha) by biome

Also, the land cover contains 42 different categories for the whole dataset, but only a small number of them contain more than 10% observations.

For the Mediterranean Forests, Woodlands & Scrub biome, that contains the 42 categories, those with more than 10% of the observations are:

id	Description	Frequency	Percentage
242	Complex cultivation patterns	5,400	20.50%
112	Discontinuous urban fabric	3,304	12.55%
243	Land principally occupied by agriculture, with significant areas of natural vegetation	2,914	11.06%

Table 6.2: CLC level 3 categories with more than 10% of observations in Mediterranean Forests, Woodlands & Scrub biome

For the Temperate Broadleaf & Mixed Forests biome, that contains only 37 categories, those with more than 10% of the observations are:

id	Description	Frequency	Percentage
241	Annual crops associated with permanent crops	4,008	18.56%
112	Discontinuous urban fabric	3,272	15.16%
242	Complex cultivation patterns	3,086	14.29%
243	Land principally occupied by agriculture, with significant areas of natural vegetation	3,086	14.29%

Table 6.3: CLC level 3 categories with more than 10% of observations in Temperate Broadleaf & Mixed Forests biome

There are a large number of categories represented but most of them with low cardinality. Thus, the variability and noise might become a problem building a model.

6.2 High-level view of a variable by temporal unit

To explore the absolute and relative frequency distribution of a variable by temporal unit (year or month), I am going to use contingency tables and line plots. This provides a high-level view of the temporal trend of a variable with the biome.

I have applied this analysis only to a single variable:

- Human-caused fires (HCFs)
  - Burnt area

What I have found is that the number of fires in 2014 is low and it has been identified in the literature as one of the "lower lows" years in the complex multi-year temporal pattern of fires in continental Portugal:

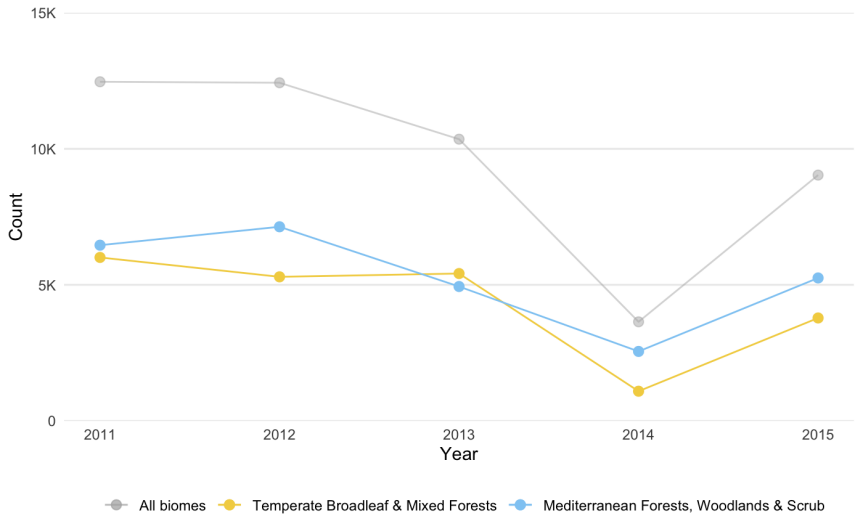


Figure 6.2: Total and grouped by biome yearly count of HCFs in the period 2011–2015

6.3 Shape of variable's distribution by biome

To judge whether there are differences between the distribution of a variable conditioned by the biome I am going to use summary statistics that include: the median and the mean, the first and third quartiles, the minimum and maximum values, and the standard deviation; together with density plots (a sort of smoothed histogram). This provides a view of the shape of a distribution (its spread, the location of its absolute and local maximum and minimum values, its symmetry or skew, and its uniformity).

To complete the numerical data, I am going to use density plots plotting the overlapping distributions of the same variable for each biome, using transparency to highlight were the distributions match and where exist differences. Also, with the same objective, I am going to use a second variant comparing the distribution of a variable by biome against the overall distribution of that variable, again using transparency to highlight the differences.

I decided to use density plots over histograms due to the former not being affected by the number of bins used in their construction. The use of a wrong bin value would make the distribution of the resulting histogram misleading.

I have applied this analysis to the following variables:

- Human-caused fires (HCFs)
  - Burnt area
  - Coordinates (x and y)
- Meteorological data
  - Maximum air temperature
  - Average air temperature
  - Mean daily wind speed at 10m
  - Vapour pressure
  - Sum of precipitation
  - Total global radiation
- Canadian Forest Fire Weather Index (FWI)
  - FFMC
  - DMC
  - DC
  - ISI
  - BUI
  - FWI
- Physiography
  - Elevation
  - Slope
- Human factors
  - Distance to nearest road
  - Distance to nearest building

Exploring the elevation I found that there are values less than zero:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	-1.50	157.73	294.11	350.23	487.07	1,517.43
Mediterranean Forests, Woodlands & Scrub	-2.50	91.76	273.37	340.54	560.50	1,835.37

Table 6.4: Statistical summary of elevation by biome

There are 41 HCFs with negative elevation. They all located along the coast and are plausible but others are located in the sea or estuaries of rivers:

(a) All (b) Detail

Figure 6.3: HCFs with elevation less than zero

Hence, I going to remove the fires with negative elevation when I prepare the data for modelling.

There is also 56 HCFs with zero elevation, but these HCFs in the coast are plausible so I am going to do nothing.

6.4 Shape of variable's distribution by biome and temporal unit

To explore the differences between the distribution of a variable grouped by biome and temporal unit I am going to use overlapping density plots by biome with one overlapping pair per year, and with all the years stacked so comparisons between years can be drawn.

I have applied this analysis to the following variables:

- Meteorological data
  - Vapour pressure
  - Total global radiation
- Canadian Forest Fire Weather Index (FWI)
  - FFMC
- Physiography
  - Elevation

However, I did not found anything noteworthy.

6.5 Distribution of skewed variables

To ascertain whether a skewed variable follows a log-normal or power distribution I am going to use empirical cumulative distribution function (ECDF) plots.

I have applied this analysis to the following variables:

- Human-caused fires (HCFs)
  - Burnt area
- Meteorological data
  - Sum of precipitation

What I have found for both variables is that they follow log-normal distribution, not a power one, as it is made evident in the descending ECDF plot with logarithmic x an y axes. The distribution of a variable following a power law would appear as a straight line, and this is not the case for any of the two.

For the burnt area the plot is:

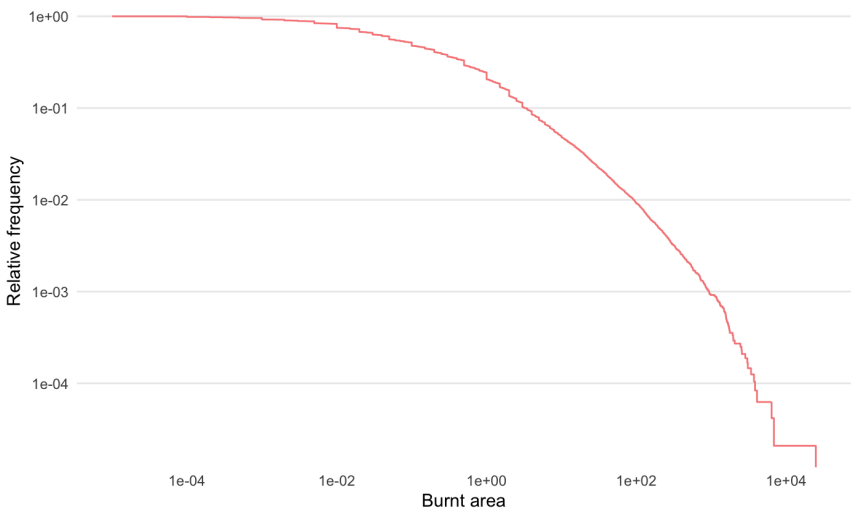


Figure 6.4: Burnt area descending ECDF using logarithmic scale

And, for the precipitation is:

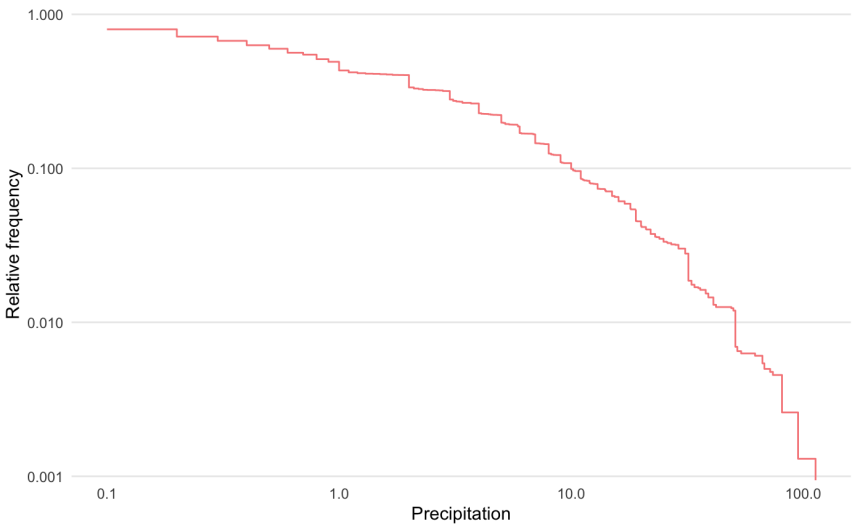


Figure 6.5: Precipitation descending ECDF using logarithmic scale

Portugal is amongst the sunniest areas in Europe. The annual precipitation varies from amounts to 1,450 mm in Braga and 1,100 millimetres in Porto, while it drops to around 700 mm in Lisbon, and to about 500 mm (20 in) in Algarve.

The observations with a value equal to zero by biome are:

Biome	Precipitation = 0		Precipitation > 0	
Temperate Broadleaf & Mixed Forests	19,562	40.82%	2,028	4.23%
Mediterranean Forests, Woodlands & Scrub	23,749	49.55%	2,588	5.40%

Table 6.5: Count of observations by biome with and without precipitation

The percentage of zeros is so high by biome that I am going to drop the variable as too much observations with zero will affect the quality of the model.

6.6 Spread and outliers of a variable

To explore the spread (dispersion), skewness and possible occurrence of outliers of multiple distributions of a quantitative variable by biome or temporal unit at once I am going to use box-and-whisker plots. They provide a way to detect the shift of a variable among distributions. To help understand the spread of a variable values I am going to superimpose to the box-and-whisker plots the scatter plot of the variable observation. To ameliorate

the problem caused by observations occluding other and masking where the values of the distribution are more concentrated I am going to add jitter and transparency to the points.

I have applied this analysis to the following variables:

- Human-caused fires (HCFs)
  - Coordinates (x and y)
- Meteorological data
  - Maximum air temperature
  - Average air temperature
  - Mean daily wind speed at 10m
  - Vapour pressure
  - Total global radiation
- Canadian Forest Fire Weather Index (FWI)
  - FFMC
  - DMC
  - DC
  - ISI
  - BUI
  - FWI
- Physiography
  - Elevation
  - Slope
- Human factors
  - Distance to nearest road
  - Distance to nearest building

From all the variables I am going to highlight the analysis of the slope. The spread of the variable is not uniform due to the how its value is stored in the raw data (as a Digital Number measure by a satellite), and its conversion into degrees over the horizontal:

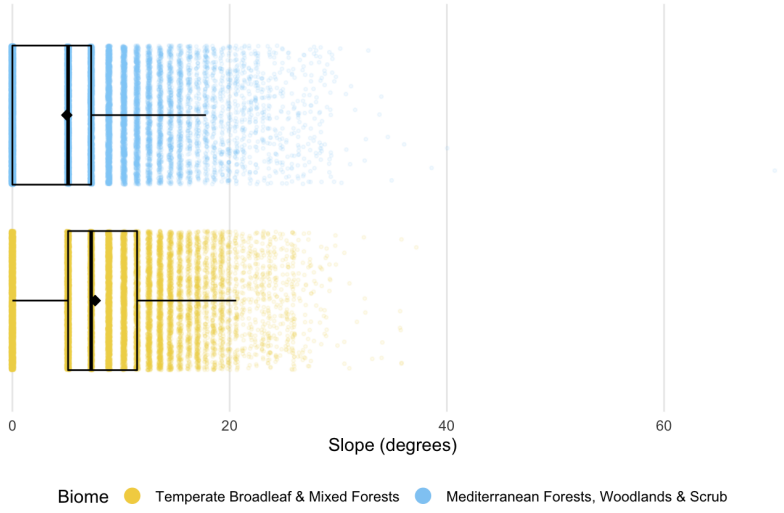


Figure 6.6: Boxplots by biome of slope

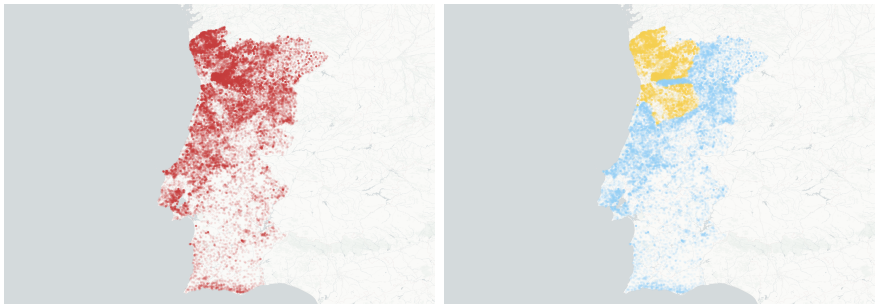
6.7 Geospatial distribution of a variable

To explore the geospatial distribution of a variable I am going to use maps using the EPSG:4326 coordinate reference system and consisting of a layer containing the points indicating the geospatial locations of the variable observations.

I have applied this analysis to only a single variable:

- Major Habitat Types (MHTs)

And, I have found that the biome "Temperate Broadleaf & Mixed Forests" has suffered a larger number of fires in proportion to its extension, thus revealing a stark difference between the northern and southern halves of the territory:



(a) All (b) By biome

Figure 6.7: Location of the HCFs in period 2011–2015

Chapter 7

Feature engineering

I am going to combine in this section the data preparation and feature engineering stages combining all the steps necessary to prepare the data, so it fits the selected model:

- Treating outliers
- Reducing high cardinality in categorical variables
- Assigning the correct data types for each variable (some algorithms only work with certain data types)
- Handling missing data
- Creating new variables

I am going to summarise the feature engineering process and left the detailed account of the steps and results to the Appendix D

7.1 Handling outliers

I am going to handle the outliers found on the data sets considering that in some cases whether a value is abnormal is a matter of perspective.

Also, I am going to try to make the least modifications possible to the data since each change distorts the data and may introduce bias.

For each variable I am going to try 3 methods to detect and treat outliers:

- Bottom/Top x%, based on percentiles. Common values are 0.5%, 1%, 1.5%, 3%, among others.
- Tukey, based on the quartiles. It considers outliers all values outside of the interval $[\text{Q1} - 3\text{IQR}, Q3 + 3\text{IQR}]$ [Tuk81].

- Hampel, based on the median and median absolute deviation (MAD) values. It considers outliers all values outside of the interval [median $-$ $k * \text{MAD}, \text{median} + k * \text{MAD}$] [Ham74].

To visualise graphically the outliers, I am going to use box-and-whisker plots with the outliers depicted as points beyond the whiskers.

To visualise the skewness, before and after removing the outliers, I am going to use different metrics of central tendency and skewness together with density plots marked with lines depicting the location of percentiles (the threshold value indicating a percentage of the observations of the distribution).

Given the data exploration already make I probe the following variables looking for outliers:

- Human-caused fires (HCFs)
  - Burnt area
- Meteorological data
  - Maximum air temperature
  - Average air temperature
  - Mean daily wind speed at 10m
  - Vapour pressure
  - Total global radiation
- Canadian Forest Fire Weather Index (FWI)
  - FFMC
  - DMC
  - DC
  - ISI
  - BUI
  - FWI
- Physiography
  - Elevation
  - Slope
- Human factors
  - Distance to nearest road
  - Distance to nearest building

To treat the outliers, I have choose to clip a feature by assigning all observations flagged as outliers the value of the threshold separating those observations from the rest. Therefore, clipping the feature values instead of removing the observations as the size of the dataset is small and I do not want to reduce its size even more. This is also a especially good strategy when faced with extreme outliers.

7.1.1 Human-caused fires (HCFs): Burnt area

As I saw in the data exploration phase, the burnt area is a highly skewed variable. Thus, to detect outliers I have used the following configuration for each method:

- Bottom/Top method for only the top 5%
- Tukey method
- Hampel method using k = 3

Their application flags the following number and percentage of observations as outliers:

Method	Outliers	Percentage
Bottom / Top	1,091	5.05%
Tukey	2,036	9.43%
Hampel	6,512	30.16%

Table 7.1: Flagged outliers in burnt area

The skewness of this variable was evident by how the standard deviation was large compared to the mean, as reflected in the variation coefficient and kurtosis value. After removing the observations flagged as outliers, all metrics have been reduced:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	5.90	85.62	14.52	45.11	2,710.31
Bottom / Top	0.69	1.41	2.03	3.34	15.79
Tukey	0.44	0.74	1.68	2.23	7.63
Hampel	0.10	0.14	1.37	1.71	4.83

Table 7.2: Skewness metrics before and after imputing outliers in burnt area

And, comparing graphically the observations flagged as outliers in the original variable and all methods:

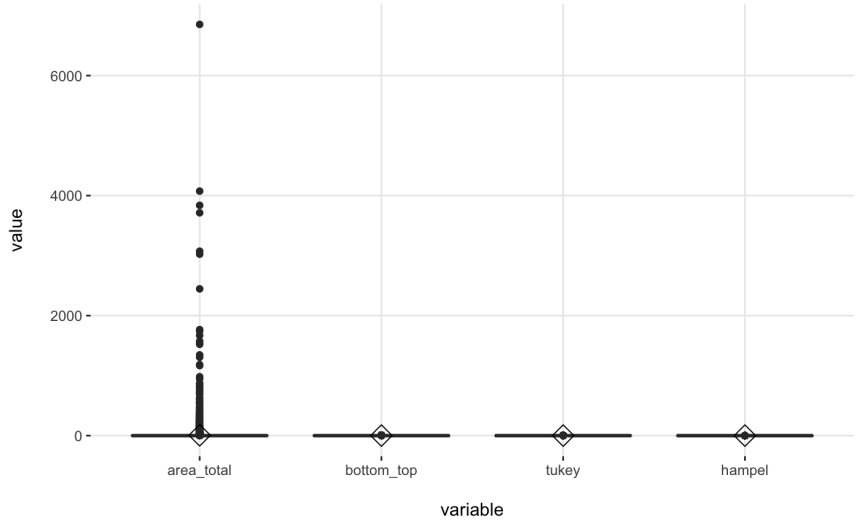


Figure 7.1: Boxplots before and after imputing outliers in burnt area

And, comparing only the result of the three methods:

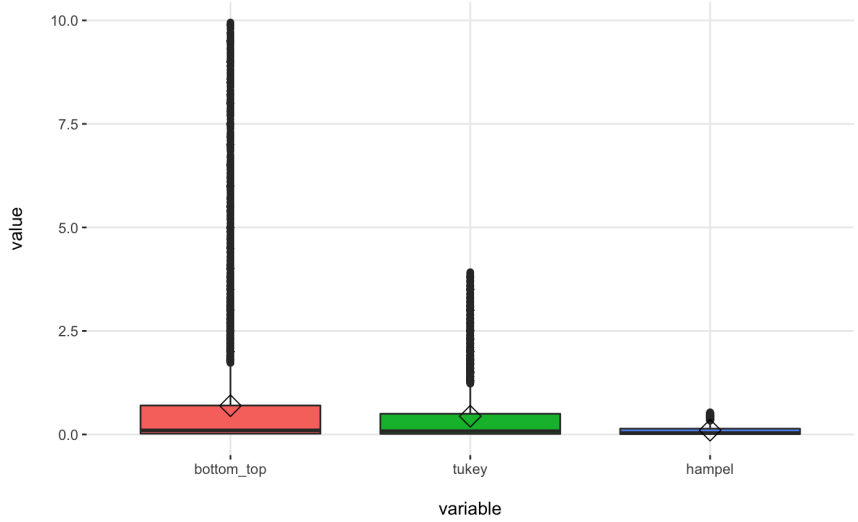


Figure 7.2: Boxplots after imputing outliers in burnt area

The distribution of the new variables is still skewed but with a much smaller tail.

I will use the Bottom/Top method that imputes the least number of observations while at the same time reducing the skewness of the distribution. The density plot for the burnt area treated with the Bottom/Top method is:

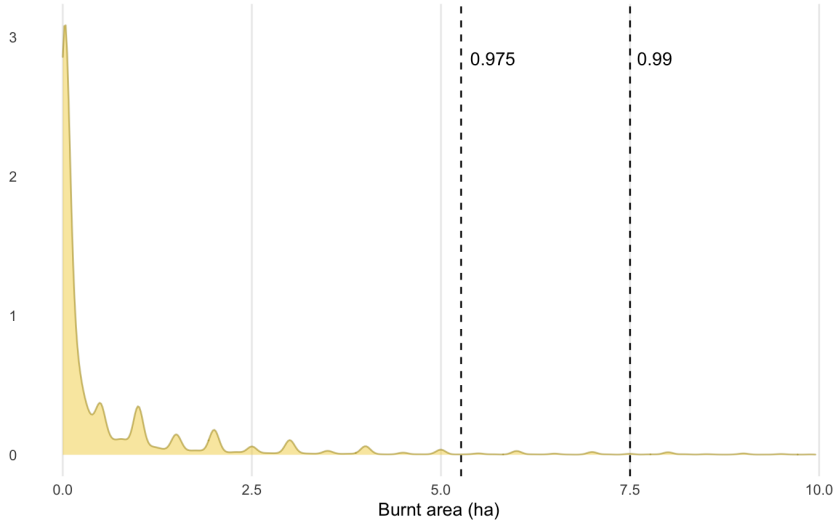


Figure 7.3: Density plot of burnt area after applying the Bottom / Top method

To conclude this section, I want to comment on the nature of the observations flagged as outliers for the burnt area. They are not true outliers but large wildfires that in the last years have become more common and larger. Their study can reveal insights into the conditions that make them more prevalent and so dangerous. However, this goal is beyond the scope of this work and I do not have the domain knowledge necessary to embark in their study.

Although some models, such as gradient-boosting machines (GBM), tend to tolerate outliers better, "noise" may still affect their performance and results. So, when comparing multiple algorithms this steps can be necessary to level the playing field to make a fairer comparison.

For these reasons I have decided to ultimately flag and treat these observations as if they were outliers.

7.1.2 Human factors: distance to closest road

As I saw in the data exploration phase, the distance to closest road is skewed variable with a tall to the right. Thus, to detect outliers I have used the following configuration for each method:

- Bottom/Top method for 5% of the top data
- Tukey method
- Hampel method using k = 3

With the number and percentage of observations flagged as outliers being:

Method	Outliers	Percentage
Bottom / Top	1,080	5.00%
Tukey	599	2.77%
Hampel	2,272	10.52%

Table 7.3: Flagged outliers in distance to nearest road

The difference before with the long right tail and how the skewness of the distribution is reduced after is visible in the box plots:

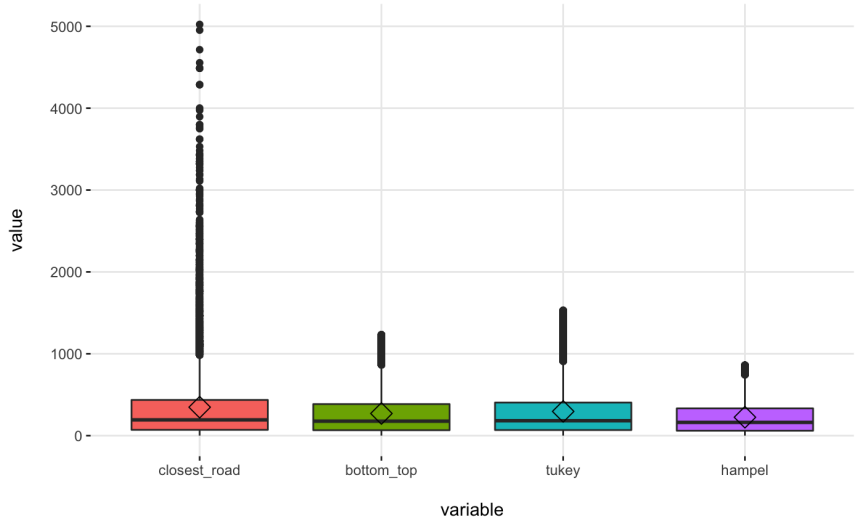


Figure 7.4: Box plots before and after imputing outliers in distance to nearest road

I am going to use the Tukey method to impute the outliers to clip the feature as it has the lowest impact with the result being similar to the Bottom/Top method that flag the double of observations as outliers but do not improve the skewness indicators by a significant margin:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	347.38	447.68	1.29	2.96	16.13
Bottom / Top	270.50	272.33	1.01	1.42	4.50
Tukey	296.02	316.91	1.07	1.65	5.46
Hampel	224.67	204.47	0.91	1.09	3.43

Table 7.4: Skewness metrics before and after imputing outliers in distance to nearest road

This low improvement is also visible in the box plot:

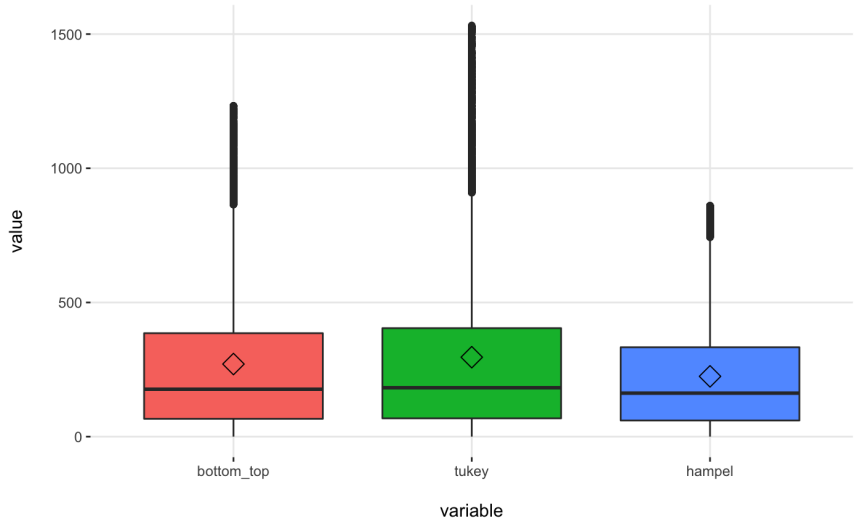


Figure 7.5: Box plots after imputing outliers in distance to nearest road

7.1.3 Human factors: distance to closest building

As I saw in the data exploration phase, the burnt area is a skewed variable with a tail to the right. So, I am going to detect the outliers with using the following configuration for each method:

- Bottom/Top method for 5% of the top data
- Tukey method
- Hampel method using k = 3

And obtaining the following results:

Method	Outliers	Percentage
Bottom / Top	1,084	5.02%
Tukey	230	1.07%
Hampel	1,394	6.46%

Table 7.5: Flagged outliers in distance to nearest building

I am going to impute the outliers using the Tukey method as it has the lowest impact and the improvement of the skewness indicators is enough, with the value of the mean and the standard deviation becoming more balanced:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	748.96	859.41	1.15	2.28	10.77
Bottom / Top	610.08	583.14	0.96	1.04	3.24
Tukey	703.12	735.50	1.05	1.50	5.19
Hampel	583.88	547.56	0.94	0.95	2.95

Table 7.6: Skewness metrics before and after imputing outliers in distance to nearest building

With the reduction of outliers visible in the box plot:

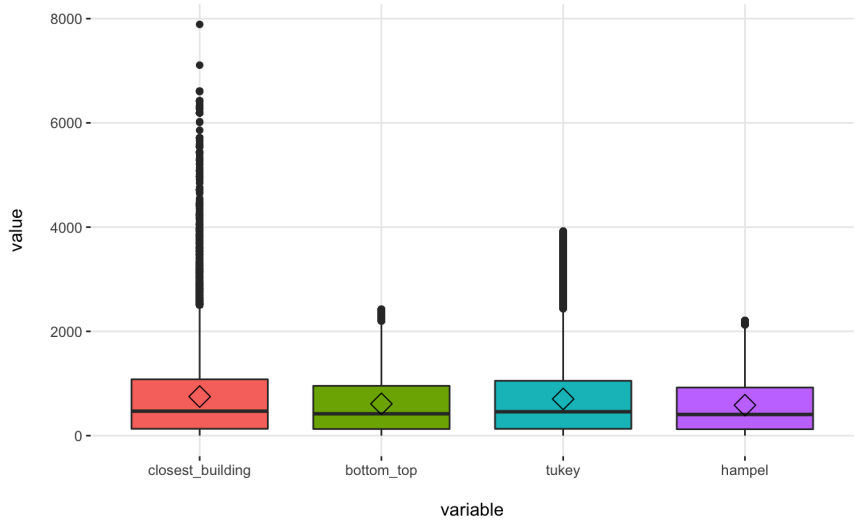


Figure 7.6: Boxplots before and after imputing outliers in distance to nearest building

7.2 Handling data types

The non-numerical variables are data_alerta, and land_cover. In the case of data_alerta I will not apply any transformation since I am going to use data_alerta indirectly through a new variable generated from it.

For the land_cover variable I am going to transform it using one-hot encoding so I can use it with machine-learning algorithms that do not support this type of variables.

But before, due to the variable having only 4 categories with more than 10% of the observations and to avoid the danger of a noisy variable and overfitting I am going to swap the CORINE Land Cover level 3 classification for the corresponding CORINE Land Cover level 2.

After converting the categories, the frequencies and percentages ordered by their frequency are:

Category	Frequency	Percentage
Heterogeneous agricultural areas	10,180	47.15%
Urban fabric	3.434	15.91%
Scrub and/or herbaceous vegetation associations	2,943	13.63%
Forest	2,720	12.60%
Arable land	1,191	5.52%
...	...	...

Table 7.7: Topmost CLC level 2 categories by number of observations

There is now a category that concentrates the 47.15% of all observations, "Heterogeneous agricultural areas." And, only 4 categories have more than 6% of the observations each one and approximately 90% of a total of 14 categories

The reduction from the codification using CLC level 3 has improved but there is still not satisfactory. To finish I am going to aggregate all categories less than a 6% of the observations that represent approximately 10% of all observations. However, the new category will still be the one with less observations.

The final categories are:

Category	Frequency	Percentage
Heterogeneous agricultural areas	10,180	47.15%
Urban fabric	3.434	15.91%
Scrub and/or herbaceous vegetation associations	2,943	13.63%
Forest	2,720	12.60%
Other	2,313	10.71%

Table 7.8: CLC level 2 categories by number of observations after grouping smaller ones

There all only 5 categories, all with al least 10% of the observations. However, now the situation can be dangerous by having less variability and existing the possibility of underfitting.

7.3 Transformations

I am going to drop the variable biome after selecting the data for a single biome to build the models.

Apart from that, the only transformation I am going to perform is to generate a new variable indicating the day of the year, called yday. This new variable may help to capture the fact that early spring and summer are the two periods of the year where the human-caused fires (HCFs) tend to more prevalent.

Chapter 8

Modelling

In this section I am going to build two models using different algorithms, GBM and GLM. The steps I am going to follow are:

1. Select the features to include in the models using a filter method
2. Tuning the models' hyperparameters and training them

To select which features to include I am going to use a metric based on Information Theory trying to discover which features show a greater relationship with the target variable, the area burnt by a human-caused fire.

To train the models I am going to use part of the data, the train dataset, and use it to tune the models' hyperparameters. The rest of the data, the test dataset, will only be used in measuring the performance of the trained models and the application of the eXplainable AI techniques.

Splitting the dataset allow us to measure its performance on data that a model has not seen in training and avoid that it can learn the data so it generalises better with new data and avoid overtraining. For the same reason I am going to use the cross validation technique with 5 folds.

I continue to tune the hyperparameters and traing models until convergence happens. that is, until the new model performance is no better than the previous one.

The steps in the modelling process are:

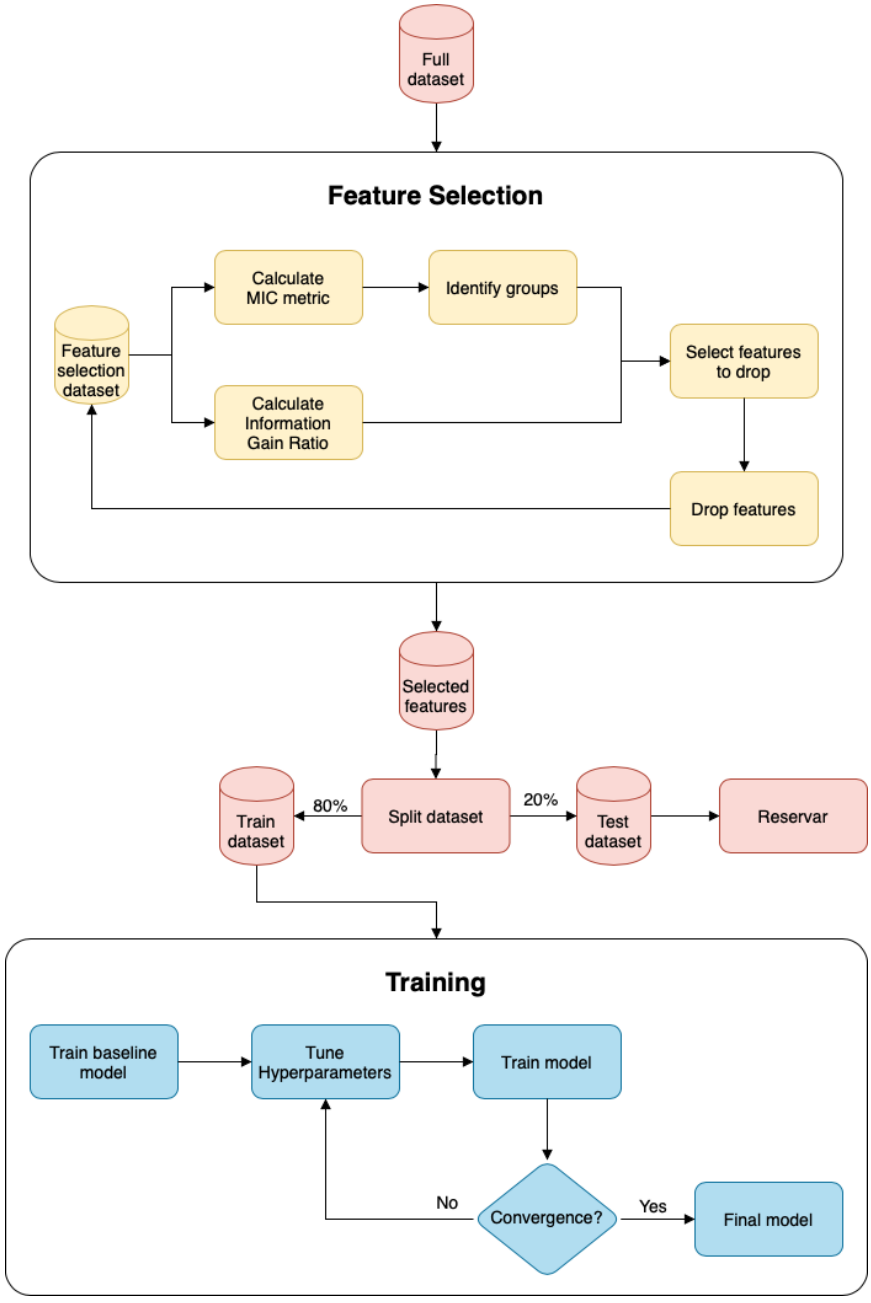


Figure 8.1: Modelling steps

8.1 Algorithms

I am going to train two models using a black-box algorithm, Gradient Boosting Machines (GBM) [Fri00], and a glass-box one, Generalised Linear Models (GLM) [Cra14]. I selected these algorithms becuase I wanted to use and

compare a black-box model against a glass-box model.

For the black-box model I chose GBM becuase it has show a good performance in many areas of application and in many instances. Also, its training is fast, compared to another algorithms.

For the glass-box model I chose GLM)that is used with variables following a log-normal distribution as is the case with the burnt area.

8.2 Feature selection strategies

In feature selection, a learning algorithm must consider the problem and select on which features put its focus, while ignoring the rest. However, this is a problem because the optimal and most relevant features might be not the same that minimise the prediction error [KJ97].

A classification of feature selection methods based on how the selection and learning algorithm steps interact is [Naq11]:

- Filter methods, that analyse the intrinsic properties of the features while ignoring the interaction with the algorithm
- Wrapper methods, that uses a learning algorithm as a "black box" function score different feature subsets
- Embedded methods, that selection is a step in the model training

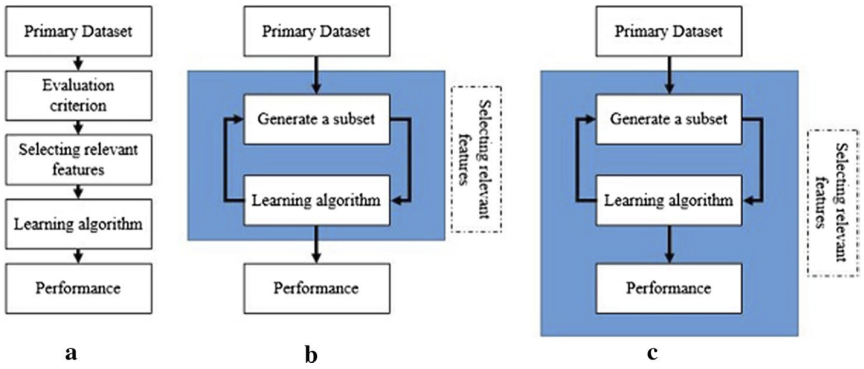


Figure 8.2: Differences between feature selection methods [Tad+19]

Filter methods does not interact with the learning algorithms, as a result, they do not intrinsically include the assumptions of the algorithm. However, this also means that the models trained with the features selected usually have a lower performance that others where the feature selection used a wrapped method.

8.3 Feature selection metrics

I have considered the use of the following methods to select the features:

- Variable importance ranking
- Functional relationship
- Non-linear correlation
- Linear correlation

As the variable importance ranking score, I have considered the information gain ratio metric. Its value is the ratio of the information gain (of each variable with respect to the response variable) to the intrinsic information value (of each variable) [Qui86].

As the functional relationship metric, I have considered the Maximal Information Coefficient (MIC), part of the Maximal Information-based Nonparametric Exploration (MIME) family [Res+11].

The value of the MIC metric indicates whether there is a functional relationship between two variables base on the mutual information. It goes from 0 to 1, with 0 being no correlation and 1 highest correlation.

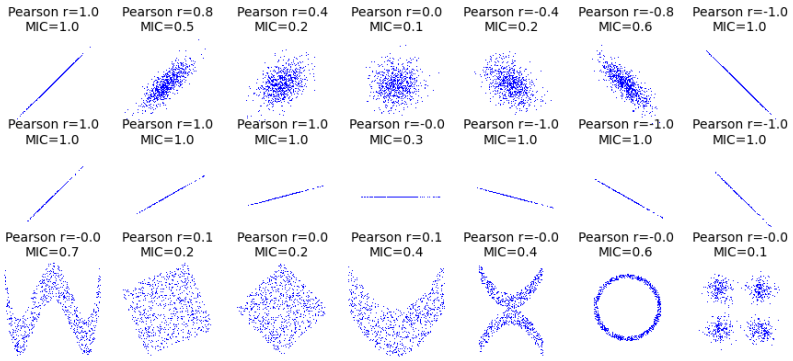


Figure 8.3: Relationships between variables characterised by R^2 and MIC values

As the non-linear relationship metric, I have considered the MICR2 ($MIC - R^2$) metric, also part of the Maximal Information-based Nonparametric Exploration (MIME).

The value of the MICR2 metric indicate whether there is a non-linear relationship between two variables. It is calculated by subtracting R^2 to the MIC metric. It values goes from 0 to 1. A high value indicates a non-linear relationship since a high R^2 indicates a linear relationship.

And, as the linear correlation metric, I have considered the Pearson correlation coefficient [Cra14].

8.4 Strategy followed

The objective is to finish the process with a set of features lowly correlated between them, but highly correlated to the target variable.

I decided on using a filter method using Information Theory metrics to avoid using a machine-learning algorithm that could introduce bias in the selection process since each algorithm sees the features in a different way, that leads to algorithms giving different importance to the same feature.

Moreover, this difference on assigning importance also leads to make more difficult to use surrogate interpretable models to try to understand black-box models.

As I am going to train two models using different algorithms I consider that using a filter method might make the selection process fairer.

For metrics I have selected two metrics based on Information Theory:

- the Information Gain Ratio
- the Maximal Information Coefficient (MIC)

For the Information Gain Ratio, features with a high value exhibit a stronger relationship with the target variable that those with lower ones. And, features with low MIC values between them exhibit a weak functional relationship (linear or not, does not matter).

The steps in the process are:

1. Calculate metrics: the two metrics an be calculated in parallel following the following two groups of the steps:

   (a) Calculate Information Gain Ratio

   And:

   (a) Calculate Maximal Information Coefficient (MIC)

   (b) Identify groups

2. Select feature and drop it

The Information Gain Ratio establishes a ranking of the features in relation to the target variable. Whereas the MIC value identifies groups of variables functionally related.

With this information I select a feature with low Information Gain Ratio with the target variable, but highly functionally related to other features.

I iterate the procedure until the selected features show the desired characteristics or dropping further features leave the data with too few ones.

8.5 Selection of features

I am going to describe the steps in feature selection process and what features were drop in each one.

8.5.1 Starting point

The initial variable importance ranking is:

Variable	Information gain ratio
ISI	0.3727
FWI	0.3727
DC	0.3727
BUI	0.3727
DMC	0.3727
FFMC	0.3725
x	0.3659
y	0.3659
Distance to closest road	0.3634
Distance to closest building	0.3634
Elevation	0.3632
Radiation	0.3395
Vapour pressure	0.2546
Day of year	0.1723
Maximum temperature	0.1665
Average temperature	0.1624
Wind speed	0.1216
Slope	0.1008
Land cover (CLC level 2)	0.0939

Table 8.1: Initial variable importance ranking by information gain ratio

The functional relationships plot using the MIC metric is:

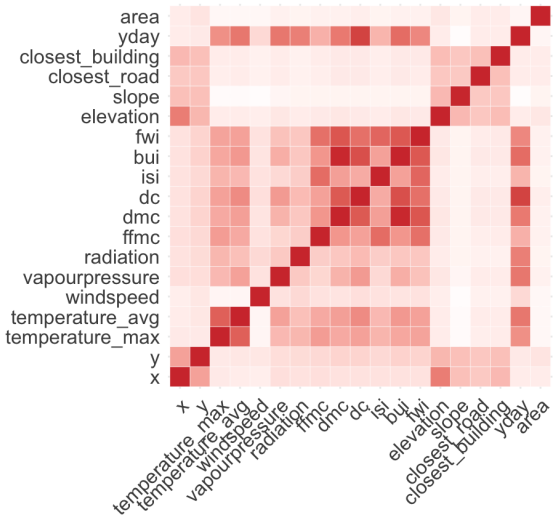


Figure 8.4: MIC matrix for whole modelling data set

Groups of mutually related features can be identified in the functional relationships plot. There is one relating the FWI indicators:

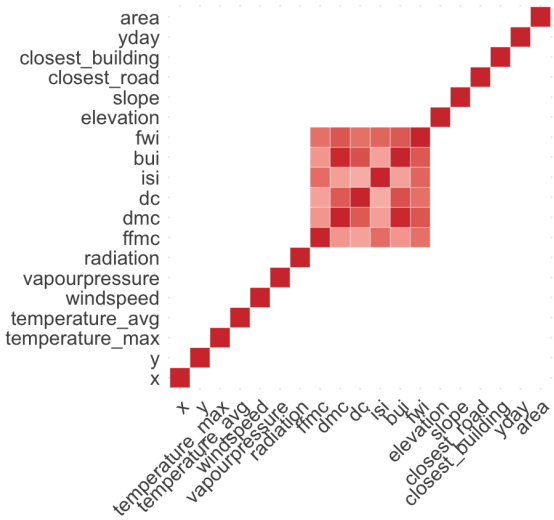


Figure 8.5: MIC matrix detail of FWI indicators

There are also the two temperature features that have a strong relationship between them, and a fair relationship with the FWI indicators:

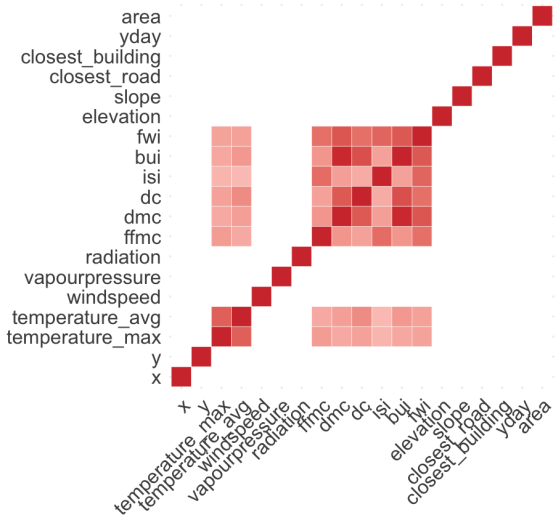


Figure 8.6: MIC matrix detail of FWI indicators and temperature variables

Moreover, the vapour pressure and radiation features can also be included, albeit with a weaker relationship. Thus, this group includes all the weather factors (i.e., meteorological and FWI features) except wind speed:

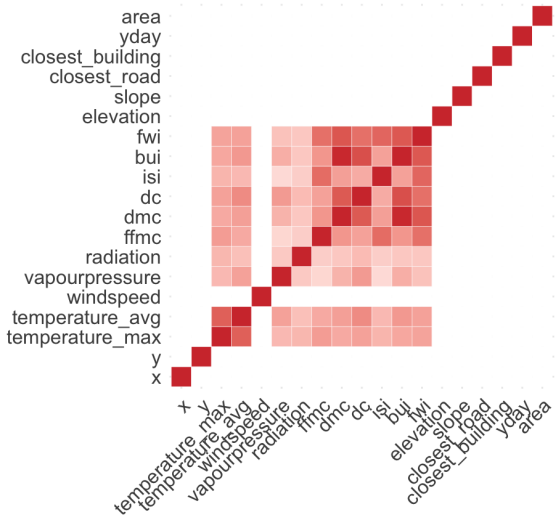


Figure 8.7: MIC matrix detail of weather factors (except for wind speed)

There is another group relating the coordinates, physiography, and human factors features:

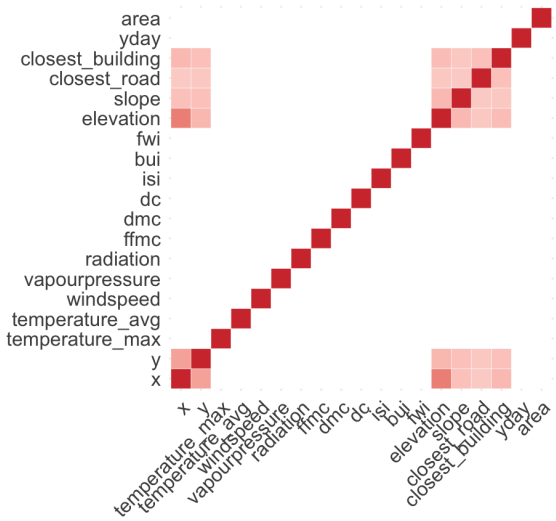


Figure 8.8: MIC matrix detail of coordinates, physiography, and human factors

To finish, the day of year variable that has a strong relationship with many other variables:

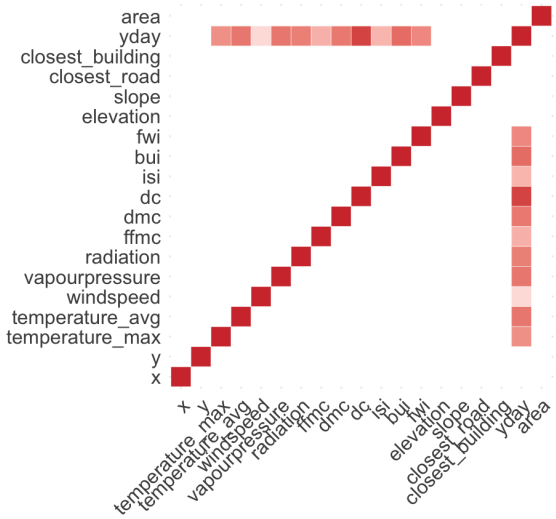


Figure 8.9: MIC matrix detail of day of year

8.5.2 Temperature features

The two temperature variables have a strong relationship, so I am going to remove the average temperature since it has a lower information gain ratio, 0.1624 versus 0.1665. The relative ranking and scores does not change:

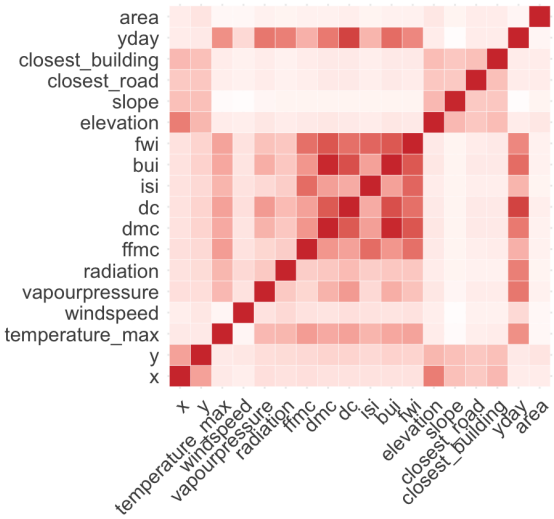


Figure 8.10: MIC matrix after dropping average temperature

8.5.3 Day of year

The day of year variable has a strong or fair relationship with every weather factor because of temporal patterns in the climate. Also, it occupies a low position in the ranking of variables by information gain ratio, immediately before the maximum temperature with a score of 0.1723.

The functional relationships plot using the MIC metric are (the relative ranking and scores does not change):

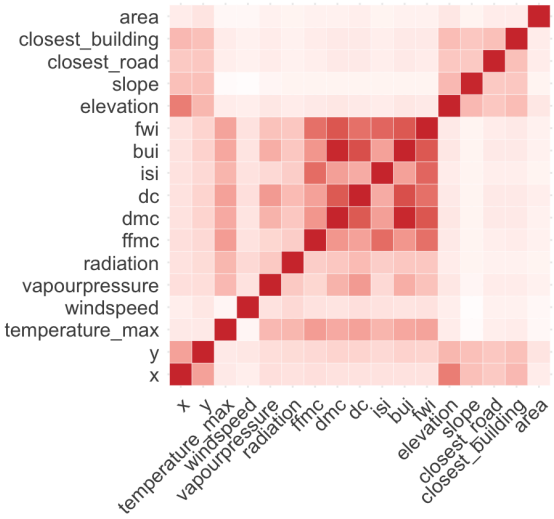


Figure 8.11: MIC matrix after dropping day of year

8.5.4 Coordinates, physiography, and human factors

I am going to drop the slope feature as it has the second-to-last score of all features. Also, I am going to drop the coordinate features since they have a strong relationship between them and all other physiography and human factor features.

The functional relationships plot using the MIC metric are (the relative ranking and scores does not change):

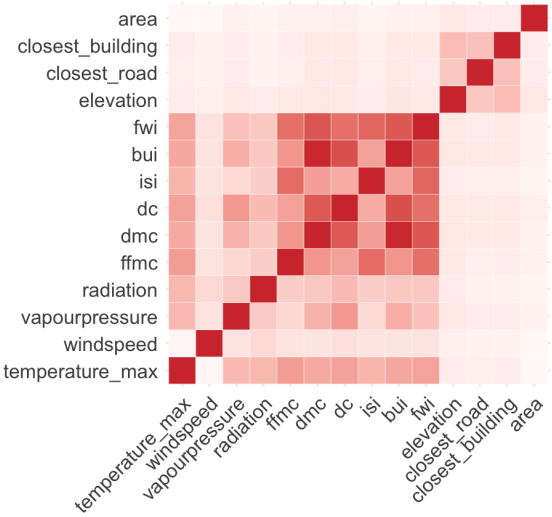


Figure 8.12: MIC matrix after dropping the coordinates and slope features

8.5.5 Canadian Forest Fire Weather Index (FWI) System

The indicators in the Canadian Forest Fire Weather Index (FWI) System have a strong relationship between them because of how they are calculated:

- They are calculated from weather factors: temperature, wind, relative humidity, and rain; and the indicators from previous days.
- BUI is calculated by combining DMC and DC.
- FWI is calculated by combining ISI and BUI.

Therefore, I am going to drop the indicators dependant on others, BUI and FWI:

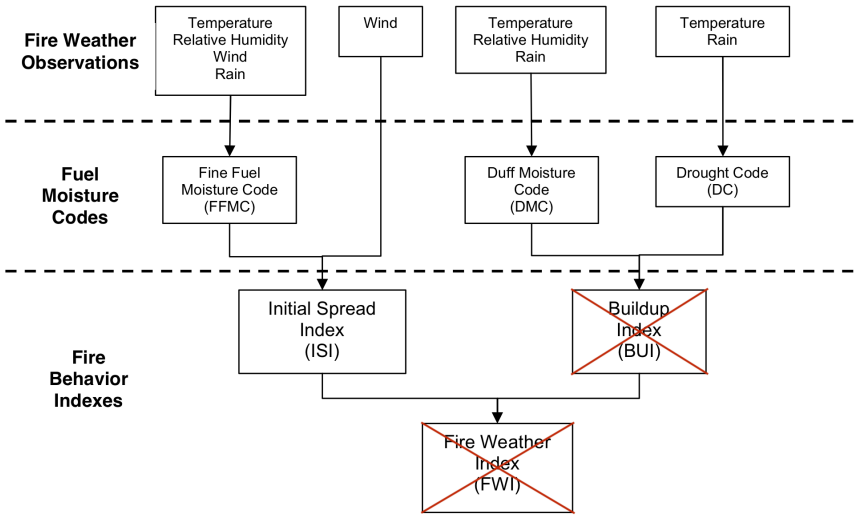


Figure 8.13: Dropped FWI indicators calculated from the other ones

The functional relationships plot using the MIC metric are (the relative ranking and scores does not change):

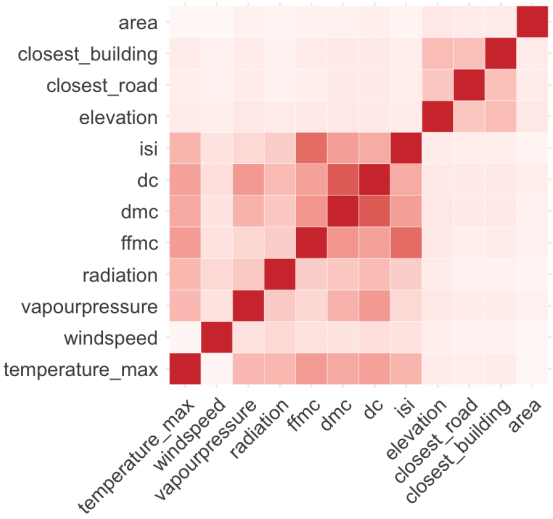


Figure 8.14: MIC matrix after dropping the BUI and FWI indicators

8.5.6 Maximum temperature

The maximum temperature has a fair relationship with the Canadian Forest Fire Weather Index (FWI) System indicators and the meteorological variables except for wind speed, and a low information gain ratio value. Therefore, I am going to drop it.

The functional relationships plot using the MIC metric are (the relative ranking and scores does not change):

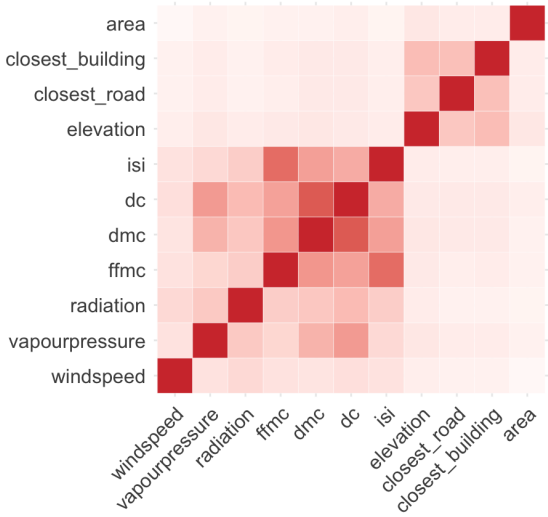


Figure 8.15: MIC matrix after dropping the maximum temperature

8.5.7 Preparing the data for modelling

Although there are still issues in the data set and further feature selection can be done, I am going to use this data set for modelling.

Thus, the final information gain ratio ranking is:

Variable	Information gain ratio
ISI	0.3727
DC	0.3727
DMC	0.3727
FFMC	0.3725
Distance to closest road	0.3634
Distance to closest building	0.3634
Elevation	0.3632
Radiation	0.3395
Vapour pressure	0.2546
Wind speed	0.1216
Land cover (CLC level 2)	0.0939

Table 8.2: Final variable importance ranking by information gain ratio

As a last preparatory step, I am going to split the data into train and test data sets using a 80 / 20 ratio. I am employing this ration because is commonly used and there is nothing that requires taking another course of action such, for example, and imbalanced dataset.

8.6 GBM model training

I am going to tune the hyperparameters performing a cartesian search. I am going to set up the search to stop after 5 rounds in a row where the RMSE metric does not improve by 0.001. The RMSE metrics will be measure using the training dataset, leaving the tetst dataset untouched.

The hyperparameters included in the grid are:

- How many trees to make.
- How deep each tree is allowed to grow.
- Learning rate, lower takes longer and requires a higher number of trees and take more time to train but give a better model.

The number of trees and their depth control the complexity of the model.
Other hyperparameters not included in the grid are:

- Learning rate annealing, the factor to scale the learning rate after each tree is trained. It allows to have a high starting learning rate that then gets gradually lower as more trees are trained.
- Column and row sampling rates, they might improve generalisation (and lower error on the test set).

The rest of the hyperparameters are left with their default values.

Also, I am going to use cross-validation with 5 folds using the training dataset (80% of the original dataset).

8.6.1 GBM baseline

Before tuning the model, I am going to train one with the default hyperparameters to establish a baseline.

The values for the explicitly set hyperparameters, through the grid or directly, are:

- Number of trees = 50
- Maximum depth = 5
- Learning rate = 0.1

- Learning rate annealing = 1.0
- Column sampling rate = 1.0
- Row sampling rate = 1.0

The RMSE value is 2.223.

8.6.2 Tune #1

The values for the hyperparameters in the grid are:

- Number of trees = $\{20, 50, 100, 500, 1000\}$
- Maximum depth = $\{3, 5, 9\}$
- Learning rate = $\{0.001, 0.01, 0.1\}$

The values for the hyperparameters of the best model are:

- Number of trees = 1000
- Maximum depth = 5
- Learning rate = 0.1

The RMSE value is 2.161.

8.6.3 Tune #2

The values for the hyperparameters in the grid are:

- Number of trees = $\{1000, 1125, 1250\}$
- Maximum depth = $\{5, 7, 10\}$
- Learning rate = $\{0.01, 0.1, 1\}$

The values for the hyperparameters of the best model are:

- Number of trees = 1000
- Maximum depth = 5
- Learning rate = 0.1

The RMSE value is 2.161.

8.6.4 Tune #3

The values for the hyperparameters in the grid are:

- Number of trees = $\{900, 1000, 1100\}$
- Maximum depth = $\{3, 5, 7\}$
- Learning rate = $\{0.05, 0.1, 0.5\}$

The values for the hyperparameters of the best model are:

- Number of trees = 900
- Maximum depth = 5
- Learning rate = 0.05

The RMSE value is 2.226.

8.6.5 Tune #4

The values for the hyperparameters in the grid are:

- Number of trees = $\{800, 900, 1000\}$
- Maximum depth = $\{4, 5, 6\}$
- Learning rate = $\{0.01, 0.05, 0.1\}$

The values for the hyperparameters of the best model are:

- Number of trees = 1000
- Maximum depth = 6
- Learning rate = 0.05

The RMSE value is 2.166.

8.6.6 Tune #5

The values for the hyperparameters in the grid are:

- Number of trees = $\{950, 1000, 1050\}$
- Maximum depth = $\{5, 6, 7\}$
- Learning rate = $\{0.025, 0.05, 0.075\}$

The values for the hyperparameters of the best model are:

- Number of trees = 950
- Maximum depth = 6
- Learning rate = 0.05

The RMSE value is 2.166.

8.6.7 Tune #6

The values for the hyperparameters in the grid are:

- Number of trees = $\{925, 950, 975\}$
- Maximum depth = $\{5, 6, 7\}$
- Learning rate = $\{0.025, 0.05, 0.075\}$

The values for the hyperparameters of the best model are:

- Number of trees = 925
- Maximum depth = 6
- Learning rate = 0.05

The RMSE value is 2.163.

I am going to accept the best model of this tune iteration as the final GBM model.

8.7 GLM model training

I am going to train GLM models for the Poisson and Negative Binomial families.

I am going to tune the hyperparameters performing a cartesian search with a grid using the training dataset, leaving the tetst dataset untouched..

The grid will include the following parameters:

- Elastic net regularisation
- Regularization strength

And, hyperparameters not included in the grid are:

- Maximum number of iterations (passes over data) = 1000 (default is 50)

The rest hyperparameters are left with their default values. Note that the GLM model standardise numeric columns by default.

Also, I am going to use cross-validation with 5 folds using the training dataset (80% of the original dataset).

8.7.1 GLM baseline

Before tuning the model, I am going to train one with the default hyperparameters to establish a baseline for each of the families.

The default values are:

- Elastic net regularisation = 0.5
- Lambda = 0.001
- Maximum number of iterations (passes over data) = 50

The default link function for both families is $log()$.
The performance metrics are:

- $\text{RMSE}_{\text{Poisson}} = 2.363$
- $\text{RMSE}_{\text{Negative Binomial}} = 2.363$

8.7.2 Tune Negative Binomial

The values for the hyperparameters in the grid are:

- Elastic net regularisation = $\{0, 0.5, 1\}$

The values for the hyperparameters of the best model are:

- Lambda = 0.266

And, its performance metric is $\text{RMSE}_{\text{Negative Binomial}} = 2.410$ (calculated using only the training dataset).

8.7.3 Tune Poisson

The values for the hyperparameters in the grid are:

- Elastic net regularisation = $\{0, 0.5, 1\}$

The values for the hyperparameters of the best model are:

- Lambda = 0.005

And, its performance metric is $\text{RMSE}_{\text{Poisson}} = 2.363$ (calculated using only the training dataset). I am going to accept this as the final GLM model

8.8 Summary of the training results

To end this chapter I am going to summarise the hyperparameters set and train error per final model, together with the (common) features included in both of them:

Model	Features	Hyperparameters	Train error
GBM	ISI DC DMC FFMC Distance to closes road Distance to closes building	Number of trees = 925 Maximum depth = 6 Learning rate = 0.05 Learning rate annealing = 0.99 Column sampling rate = 0.8 Row sampling rate = 0.8	RMSE = 2.163
GLM	Elevation Radiation Vapour pressure Wind speed Land cover (CLC lavel 2)	Error structure: Poisson Lambda = 0.005	RMSE = 2.363

Part III
Application of Explainable AI techniques to the case of study

Appraisal of some XAI techniques in the context of the case of study with two types of techniques, applicable at model-level and at instance-level.

- At model level: how and why the models make predictions taking into account the whole dataset.
- At instance level: how and why the models make their predictions for specific instances.

Chapter 9

Explainable AI at model level

I am going to describe and apply different techniques for exploration and explanation at model level for the whole data set which are used to try to understand how a model's predictions perform overall.

Techniques at model level can provide insights into the quality of the predictions for a population, assuming the observations used to build the model form a representative sample of the population.

Explainer techniques at model level focus on:

- Model's performance: how good is the model, explored in section 9.1
- Feature's importance: which features and how much contribute to the model's prediction, objective of section 9.2
- Feature's effect: how a feature influences the model's predictions, demonstrated in section 9.3.
- Model's fit: how different are the predictions from the real values, treated in section 9.4.

For the case study I am going to apply different techniques to the models predicting the size of burnt area, their overall performance and how its features influence on the predictions.

9.1 Assessing the quality of a model

Model performance measures can be applied for several purposes:

- Model evaluation: we want to answer the question of how good is the model, that is, how reliable are its predictions and how are the expected errors (in frequency and size).

- Model comparison: we want to choose the more performant model between various candidates.
- Model performance on new data: we want to evaluate a model in production when applied to new data to know whether its performance has worsened.

The metric used to measure the performance of a model depends on the response variable type (continuous, categorical, binary, ...). Most metrics are based on the comparison of expected and predicted values.

The response variable of the case study, the burnt area (in hectares) of a human-caused fire (HCF) is a continuous one. A common metric is the mean-square error (MSE), that is the sum of the squared difference between observed and predicted values:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Since the MSE is on a different scale than the response variable, another popular metric is root-mean-squared-error (RMSE):

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Another popular metric given the sensitivity of MSE to outliers is the median absolute deviation (MAD):

$$\text{MAD} = median\left(|Y_i - median\left(Y_i\right)|\right)$$

All kind of metrics highlight a different aspect of a model or have some shortcoming. Hence, it is customary to use present multiple metrics at the same time. The ones I am going to use are the RMSE and the MAD, and I am going to use the observations from the test set (the 20% of all observations) that I reserved and not used to train the models.

The metrics for the GBM model are:

- $\text{RMSE}_{\text{GBM}} = 2.366$
- $\text{MAD}_{\text{GBM}} = 0.818$

And, for the GLM model are:

- $\text{RMSE}_{\text{GLM}} = 2.402$
- $\text{MAD}_{\text{GLM}} = 0.849$

The global performance difference between the two models is small even after the much longer tuning procedure for the GBM model. Moreover, the values for both models are high to the burnt area values. It might indicate a problem with the selected features.

9.2 Feature's importance

Evaluation of a feature importance can yield insights into:

- Model simplification: features with little importance to a model can lead to discard those features and an increase in the performance of the model.
- Model exploration: comparing features in account of their importance when building a model may help discover correlations between features that may be helpful in improving the final model. They allow to peek into the inner workings of black-box models that otherwise will be beyond our reach and know how the model see the features.
- Validation of domain knowledge: the ranking of features based on their importance may help into appraise a model based on the domain knowledge.
- Generation of new knowledge: by identifying the most important features may lead to new insights into how the features factor to affect the target variable and discovering new mechanisms unbeknown to us.

In this chapter I am going to evaluate global and model-agnostic methods that allow to make contrastive comparison between models, as for example between a glass-box model and a black-box model that can use quite different internal structures and algorithms. Thus, I am going to leave out model-specific methods that could be more precise.

When we find that a feature ranks high when comparing its assigned importance by different types of algorithms, we can mark the feature as candidate to be associated with the target variable. But further studies have to be taken before we can conclude that a correlation between the variables exists.

Moreover, we must take into account that different types of machine learning algorithms can see the features in very different ways assigning a larger importance to different features while at the same time performing both well. And also, we have to be alert to the fact that the presence or not of other features can affect the relationship so adding or removing them might change the interaction between other features and the target variable.

This is also the case we can find when using a surrogate glass-box model to interpret a black-box model trained in the same data but using the predictions of the black-box model as target variable.

Beyond comparing multiple machine-learning algorithms, when we apply this type of techniques to a black-box model we can reveal how it sees the features. This is something can is easily done with glass-box models such as linear models that encode the importance of the variables in the weights assigned by them to the different features. But it is not possible with black-box models, leaving us blind to the inner workings of the model what features, part of the problem domain, it sees as relevant.

The method described in [FRD19] is based on measuring the changes in the fitness of a model when perturbations are introduced in the value of a subset of features. The more important a variable, the larger the reduction in fitness of the model. The perturbations introduced are simply the change of the value of the features of the model (e.g., permuting the values of a feature).

In the case study with a continuous variable, the burnt area in hectares of a human-caused fire, we measure the fitness of both model using the RMSE as loss (error) metric, and how it changes in when the values of the features changes. Thus, the larger the variation in the error as measured by the RMSE, the more important the feature.

The variable importance of the models is:

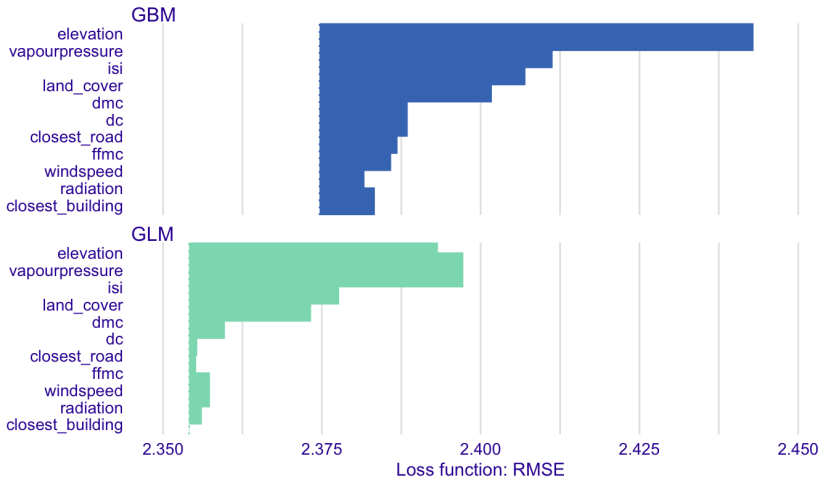


Figure 9.1: Variable importance for the GBM and GLM models

The different models assign different importance to different metrics, as is evident in the different size of the bars for the same feature. Even though, both algorithms seem to assign a high importance to the same group of features:

- Elevation
- Vapour pressure

- ISI
- Land cover
- DMC

Therefore, I can conclude that this group of features are the most important to consider for the prediction of the burnt area by a human-caused fire. Also, this ranking is strikingly different from the one calculated by using the information gain ratio.

However, the perturbations introduced by the method are random in nature so we will get different results depending on the perturbations introduced as it is evident in the following figure:

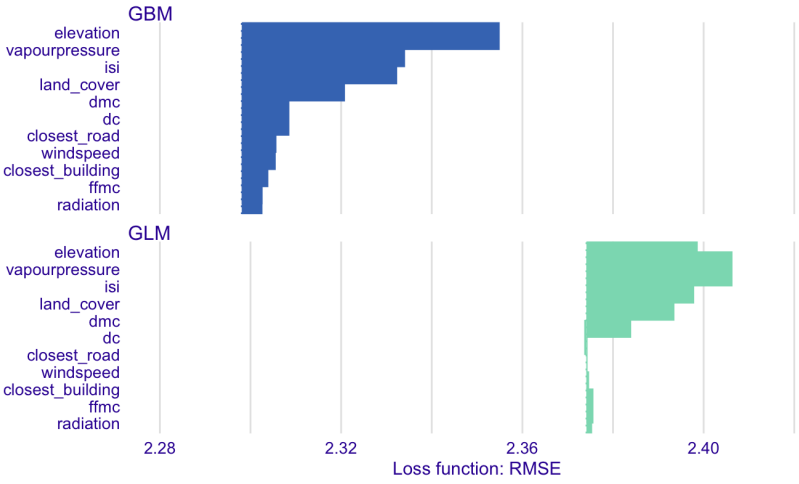


Figure 9.2: Variable importance for the GBM and GLM models

Although the most important variables remain the same, there have been a change of the relative position in the least important ones.

The method can also be applied in the training or test steps that usually use different partition of the data, and different datasets can lead to different variable's importance rankings. Although, it will serve to open a window on what are the variables that an algorithm consider the most important in training, and what are the (possibly different) ones that are more important in making a prediction in the test step, or in production.

To conclude, we can gain valuable insight on how a model sees the features but always considering the sampling variability and randomness present in the method. Furthermore, the different representations consequence of the randomness of the method together with our own expectations can shape our thoughts without ourselves being aware.

9.3 Effects of features in the average prediction

Measuring the effect that a feature has on a prediction helps in understanding the relation between model response and model prediction by summarising the effect of changes on a feature have on the predictions of the model.

In a glass-box model you know the effect of changes on the inputs by examining the importance of the features. For example, in a trained linear regression model with a numerical feature with weight β_j, increasing the value of that feature in one unit changes the estimated output by the value of its weight, when all other features remain the same.

But this is not possible in a black-box model where the connection between the inputs and the outcome of the model is clouded by the own nature of the model.

A technique that tries to reveal this connection between inputs and the outcome is Partial Dependence (PD) profiles (or plots), that measure the effect that a explanatory variable has on a model prediction. They show how the model prediction as a function of a explanatory variable of interest. For example, it can show us whether the relationship between the target and a feature is linear, monotonic or more complex

PD profiles can also be used to explore the local stability of the predictions, that is, how much the predictions change in the face of changes in the feature values. A stable model will change slightly or not at all so in the face of changes in the input, the model will make the same predictions that would have done if the changes were not present. This can be a sign of model will show a good performance on new, unseen data (good generalisation).

9.3.1 Partial Dependence (PD) profiles

PD profiles (or plots) where introduced in [Fri00]. They plot the prediction of a model as function of a selected feature when all others all held constant.

The partial dependence function for regression is defined as [Mol20]:

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C)$$

With x_S being the feature to plot in the PD profile and x_C the rest present in the model $\hat{f}$.

The partial function $\hat{f}_{x_S}$ is estimated by calculating averages from the data or sample:

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

For binary classification models, the PD profile shows the predicted probability for a certain class for different values of the feature of interest. With multi-class classification, a line per class can be plotted.

For example, the PD profile for a cervical cancer classification model and features age and years taking hormonal contraceptives is:

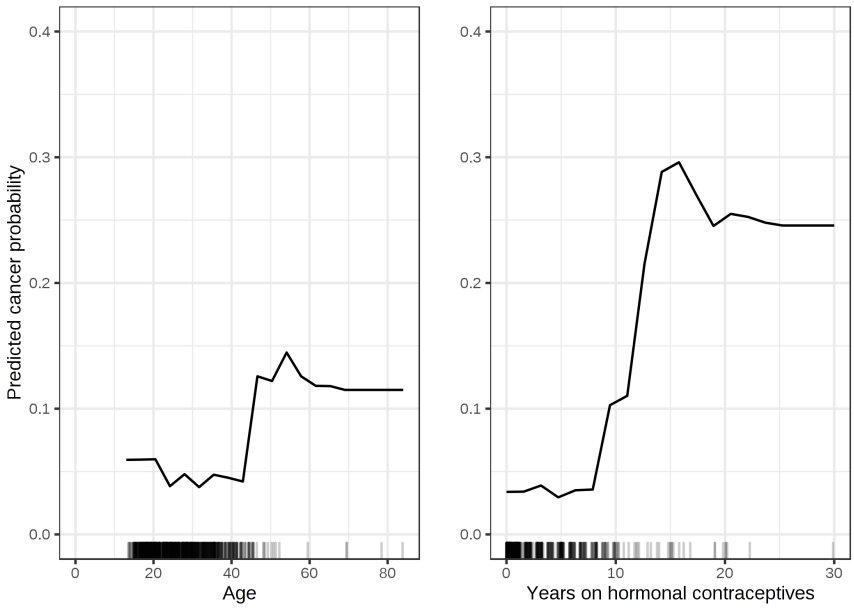


Figure 9.3: PD profile for selected features of cervical cancer classification model

The PD profile shows that for the age the probability of developing cancer is low before reaching 40 years and increases after. While that, for the contraceptives feature it shows that the probability increases with more years taking the contraceptives, especially after 10 years of use [Mol20].

A disadvantage of PD profiles is that holding other features constant can be impossible in the case of existing correlated features in the data and hiding the relationship. This results from the fact that a PD profile show us the average effect of the feature in the predictions of the model using Ceteris Paribus (CP) profiles.

CP profiles is a technique that tries to understand how a single feature affects predictions when all other features all left unchanged (Ceteris paribus is a Latin phrase meaning "other things held constant"), and allowing us to explore the influence of features on the target variable but focusing on a single feature at a time.

For example, in the case of the cervical cancer model example plotting the PD profiles for individual instances for the age feature shows a more complex relationship:

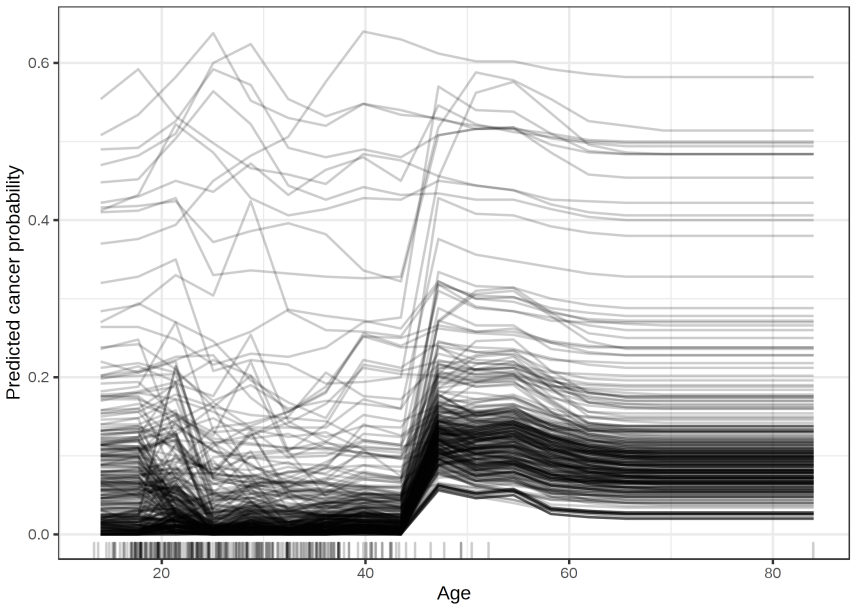


Figure 9.4: PD profiles for the age feature and individual instances

For most women the pattern is consistent with the one show by the PD profile, but for those women with a high probability of developing cancer from an early age the predicted probability doesn't change with age.

For the case study, the comparison of the PD profile against CP profile for the DC feature is):

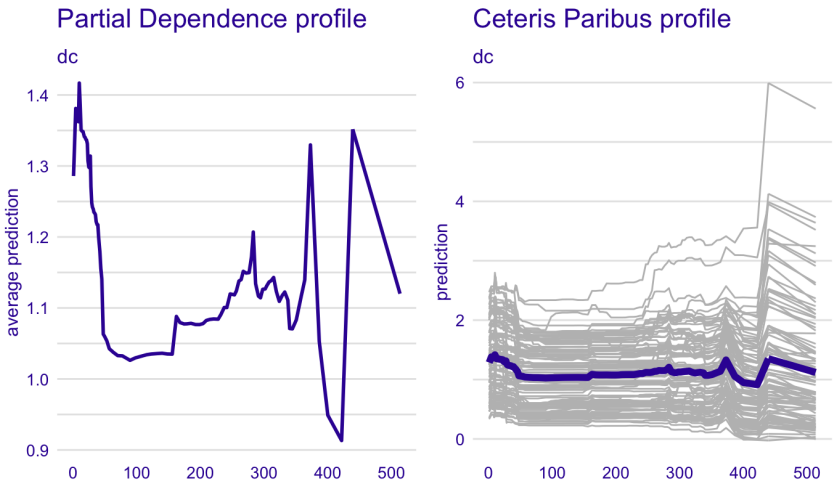


Figure 9.5: Contrasting PD profile and CP profile of DC feature

(Note the difference in scale of the plots).

In the case of the DC indicator, it seems that there are no interactions and the PD profile summarises the average effect the DC feature has on the prediction (i.e., the predicted value as a function of the DC feature value).

However, this not happens with all features in the model as happens with the radiation feature:

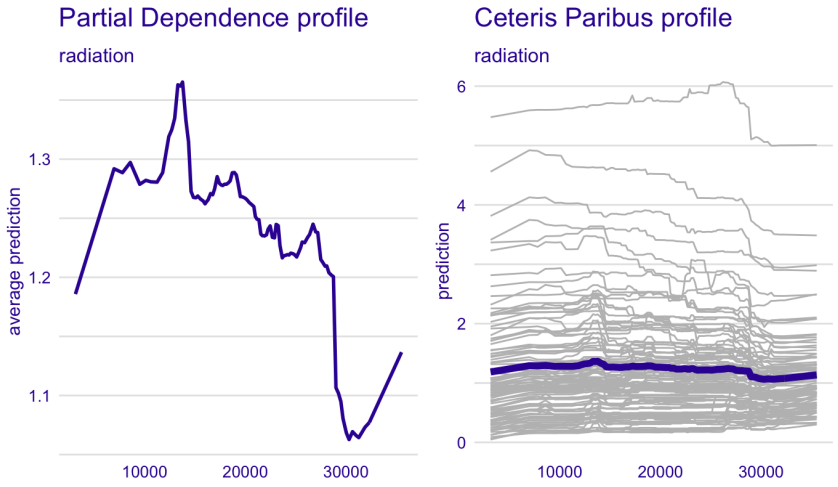


Figure 9.6: Contrasting PD profile and CP profile of radiation feature

Even though some of the observations have a similar shape, there is no clear shared pattern visible. Therefore, we can conclude that there is a more complex relation between the feature and the predictions

9.3.2 Local dependence and accumulated local profiles

To avoid the problem of PD plots with correlated featured [AZ19] introduced Accumulated Local Effects plots that avoid this problem.

The problem of PD plots stems from having to extrapolate in regions where there are no observations, and the uncertainty introduced by the new hypothetical values, especially the further the new data point from the data envelope. The extrapolation occurs because the marginal distribution is less concentrated that the conditional distribution, due to the strong dependence between variables:

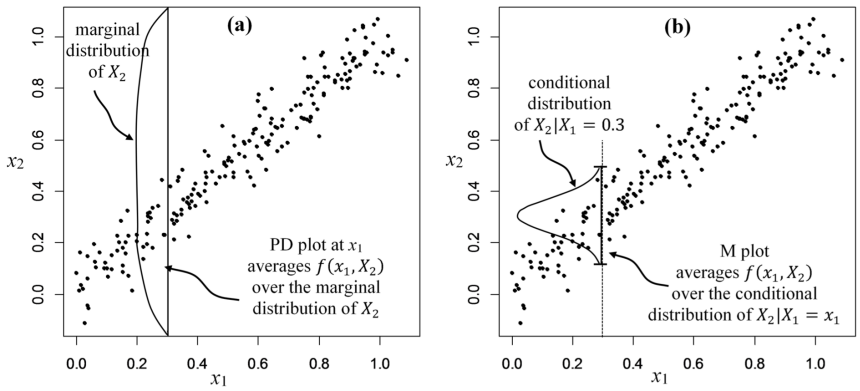


Figure 9.7: Differences between marginal and conditional distributions [AZ19]

For the case study, comparing the PD profile against the local dependence and accumulated local profiles for the distance to closest road feature:

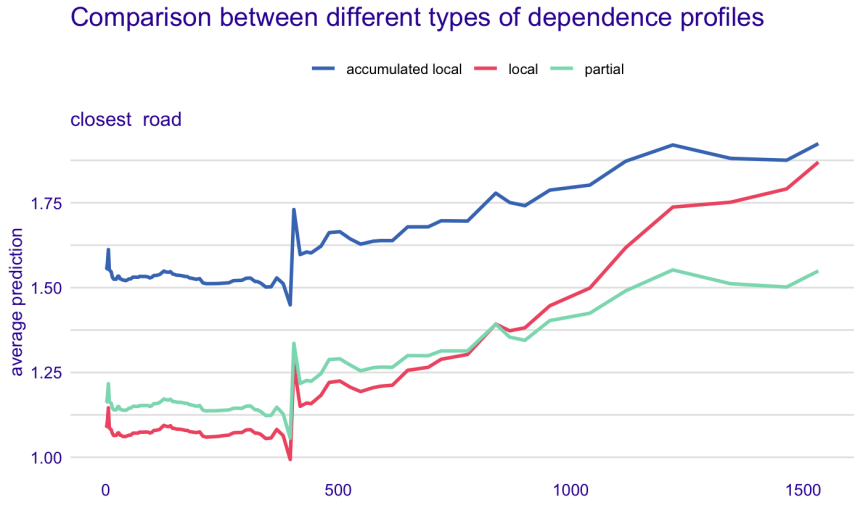


Figure 9.8: Comparison of different dependence profiles to check for correlations between features

The local dependence profile (i.e., the red line) is steeper than the others due to the correlation of the feature with others. However, the accumulated local profile (i.e., the blue line) removes the effects of the correlation and being parallel to the partial dependence profile (i.e., the green line) means that the effects on the model area additive for the distance to closest road feature.

Therefore, we can use this technique to detect features with complex relations to the predictions of the model that some models can have an

easier time capturing.

9.3.3 Clustered PD profiles

Clustered PD profiles are another option when there are correlations between features. When there are no interactions PD profiles are parallel and offer a good summary of a feature. But if not, there are not parallel, making difficult to discern patterns between the effect of a feature and the value of the prediction. Moreover, this also indicates that the PD is hiding a complex relationship and not offering a faithful summary of the relationship of the effect of a feature and the predictions of the model.

Clustered PD tries to find clusters of similar observations whose PD profiles represent that subset of the data. Thus, it can helps us to detect clusters of observations for which a feature has a similar effect on the predictions and indicating patterns in the data that can help us gain new knowledge about the input data or validating domain knowledge.

For the case study, the clustered PD profile for FFMC indicator is:

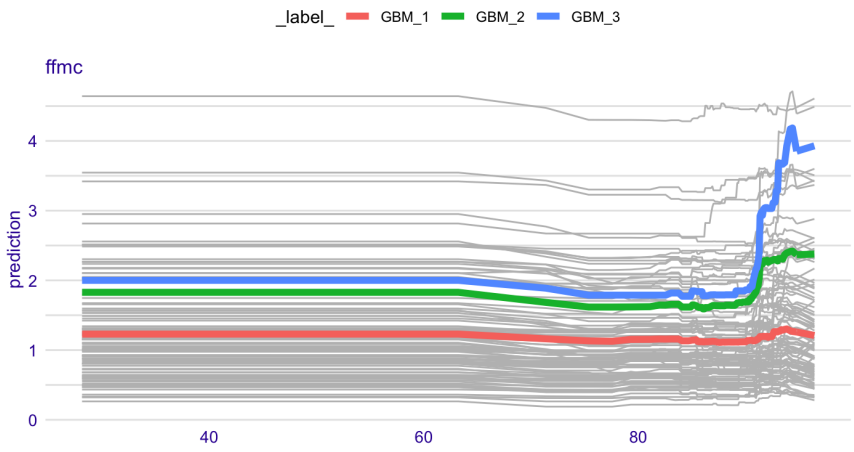


Figure 9.9: Clustered PD profile for FFMC indicator

In the following example, the FFMC feature shows three groups that diverge for high values of the feature in how much the predicted value increases with the value of FFMC. This might indicate groups of observations where the HCFs are ignited by litter and other cured fine fuels since FFMC is an indicator of the ease of ignition and the flammability of fine fuel.

This way by revealing the connection between features and predictions, the clustered PD profile for the FFMC indicator has help us detect the cause of ignition for a group of observations that can lead to detect the conditions

making this kind of human-caused fires more dominant than other fires caused by other sources of ignition.

9.3.4 Grouped PD profiles

Grouped PD profiles are another option when there are correlations between features. When a feature of interest is correlated with a categorical one, a valid approach is to explore the PD profiles of the feature of interest grouped by the values of the categorical feature and investigate whether the grouped observations follow an identifiable pattern.

In the following example, the DC feature PD profiles does not change when we segregate the observations by the land cover categorical feature:

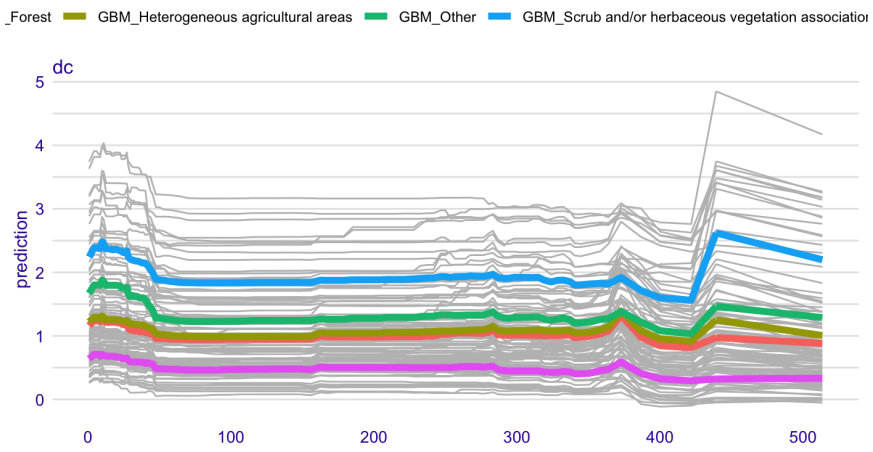


Figure 9.10: Clustered PD profiles for the DC feature

In this case, the technique indicates us that the land cover feature doesn't influence the effect of the DC feature over the predictions, thus indicating that there seems to be no correlation between them.

9.3.5 Contrastive PD profiles

PD profiles can also be used to compare different models trained on the same data set.

For the case study , the comparison between the PD profiles for the vapour pressure feature between the GBM and GLM is:

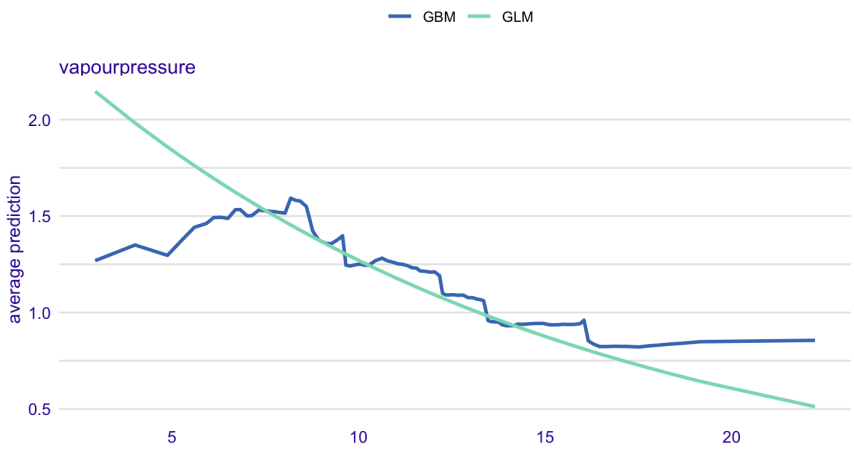


Figure 9.11: PD profiles for the GBM and GLM models for the vapour pressure feature

The biggest differences between the models is largest at the edges revealing the different nature of the models, with GBM being a more flexible model than GLM.

9.4 Residuals

The study of residuals in statistical modelling is generally the first method taught to explore the predictions of a model since they can show us what aspects of a data set have not been captured by the model. Usually graphical techniques are used in the study of residuals.

Chapter 10

Explainable AI at instance level

I am going to describe and apply different techniques for exploration and explanation of model predictions at for a single observation. This level exploration is well suited to how people think using examples and analogies.

Techniques at instance level can provide insights into how a makes a prediction for a single observation so we can:

- Local feature's importance: which features and how much contribute to the outcome of a single prediction, demonstrated in section 10.2.
- Feature's effect on a single prediction: how predictions change in response to changes in the value of the features, treated in section 10.3
- Local model fitness: what is causing erroneous predictions, not explored in this work.

For the case study I am going to apply different techniques to three observations used to predict the size of burnt area, two of them sampled from the test data and another not present neither on the training nor the test data (a synthetic instance). The feature values for the selected observations can be found at section 10.1.

10.1 Instances of interest

For studying techniques at instance level, I am going to select two instances from the test split. The first, together with the observed value and the prediction of the GBM model, is:

Variable	Value
Wind speed (m/s)	3.5
Vapour pressure (hPa)	8.43
Radiation (KJ/m^2/day)	13,586
FFMC	86.35609
DMC	21.02484
DC	19.77554
ISI	5.681571
Elevation (meters)	459.3532
Land cover	Heterogeneous agricultural areas
Distance to closest road (meters)	8.704325
Distance to closest building (meters)	469.2773
Burnt area (ha)	1.7
Prediction (ha)	1.832

Table 10.1: Variable and prediction values of first instance of interest

Also, I am going to include a what-if instance by taking the first instance selected and change the value of one of the features to simulate a what-if scenario to test how the model behaves in the face of the modification. For this example, I am going to change the value of the FFMC indicator from 86.35609 to 96.35609:

Variable	Value
Wind speed (m/s)	3.5
Vapour pressure (hPa)	8.43
Radiation (KJ/m^2/day)	13,586
FFMC	96.35609
DMC	21.02484
DC	19.77554
ISI	5.681571
Elevation (meters)	459.3532
Land cover	Heterogeneous agricultural areas
Distance to closest road (meters)	8.704325
Distance to closest building (meters)	469.2773
Burnt area (ha)	1.7
Prediction (ha)	1.832

Table 10.2: Variable and prediction values of what-if instance

10.2 Local feature importance

Techniques that focus on local feature importance allow us to dissect a prediction into parts, with each one linked to specific features. Hence, applying these techniques to black-box models allows us to discover the inner workings of the model and know which features where judged relevant by it to make a prediction for a specific observation.

Examples of these techniques are:

- Break-down (BD) plots [SB18]
- Shapley Additive Explanations (SHAP) [SK10]
- Local Interpretable Model-agnostic Explanations (LIME) [RSG16]

10.2.1 Explanation of the technique

The underlying idea to BD plots is to calculate the contribution of a feature to the prediction made by the model as the changes in the prediction while maintaining fixed the other features. This idea is summarised in [SB18] with the following figures that apply the algorithm to a logistic regression model trying to predict which employees will leave a company using the HR analytics synthetic dataset from:

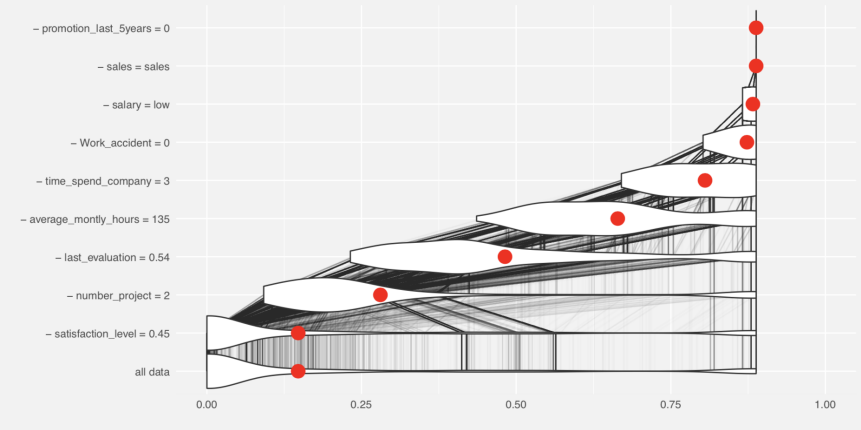


Figure 10.1: Conditional distributions of predictions when a feature is maintained fixed

The violin plots summarize the conditional distributions of the predictions (whose value is indicated by the x-axis), and the red dot the average of the predictions for all input data. The thin black lines in show how individual predictions change after the feature indicated in the y-axis is fixed at the indicated value.

The BD plot summaries the previous figure with the prediction of the model in the first row and the red dot indicating the beginning and end of each rectangle, whose width indicates the size of the contribution):

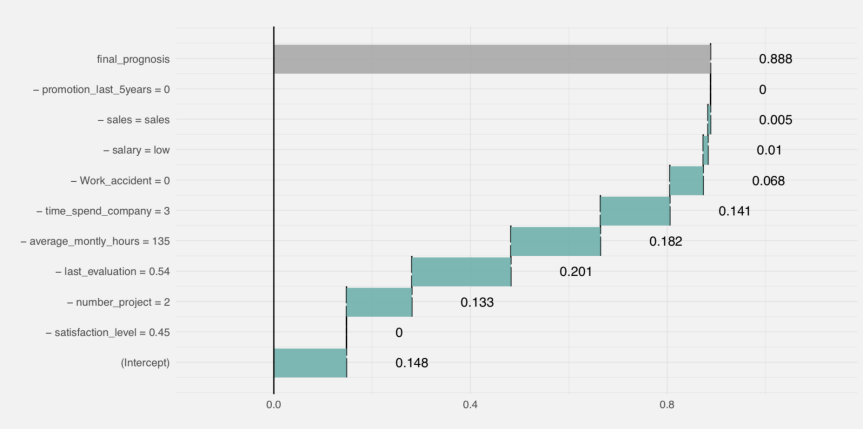


Figure 10.2: BD plot for a single instance of the HR analytics model

In this specific case, all features have a positive or zero contribution to the prediction. But, the bigger contributions to the prediction are the result of the last evaluation and the average number of hours worked by the employee. Other features with a big impact are the extent of time that the employee has been working for the company and the number of projects in which he or she participates.

The disadvantage of BD plots is that in the presence of interactions between features (non-additive models), the algorithm used to generate the BD plots is sensible to the ordering of features chosen so the data produced can lead to erroneous conclusions on the contribution of the different features to the prediction of interest.

This effect is patent in the following example of the application of different random orderings to a synthetic instance on a Random Forest model used to predict the probability of survival in the widely known Titanic dataset. Each subplot corresponds to a random ordering that conditions the value of the features' contribution:

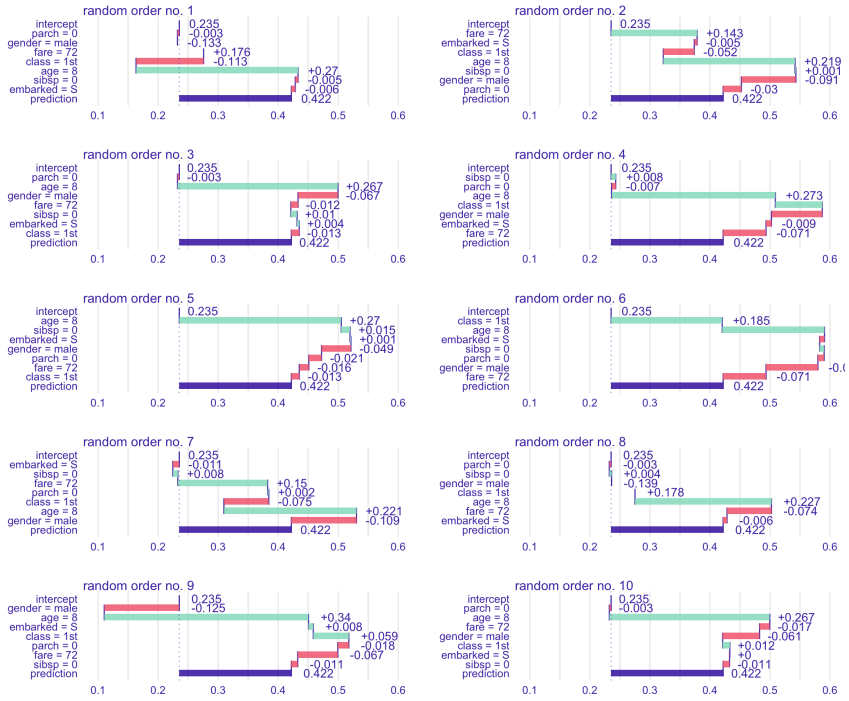


Figure 10.3: BD plots for ten random orderings of features

The dark blue bar in the last row of each BD plot indicate the value of the prediction that is 0.422 and equal in all plots.

The value and sign of the contributions to the probability of survival (red bars indicate a negative contribution, whereas green ones indicate a positive contribution) depend on the ordering with differences between the plots.

The values of the synthetic instance are indicated in the labels of the y-axis of each plot. The red bars indicate a negative contribution to the probability of survival, the green ones indicate a positive contribution.

A technique to assess the local importance of features to a model's prediction that tries to address this problem of the dependency of the explanatory covariates in the presence of interactions is SHapley Additive exPlanations (SHAP) [SK10]. The strategy it used to achieve its objective is averaging the value of a feature's contribution over all (or a large number of) possible orderings.

It is based on Shapely values [Sha53] originally developed for selecting between different lines of action in semi-cooperative games in which either a negotiation among the n players determines their actions or else an arbitrator specifies them.

Continuing with the example of the Titanic dataset, the SHAP values for the same instance are:

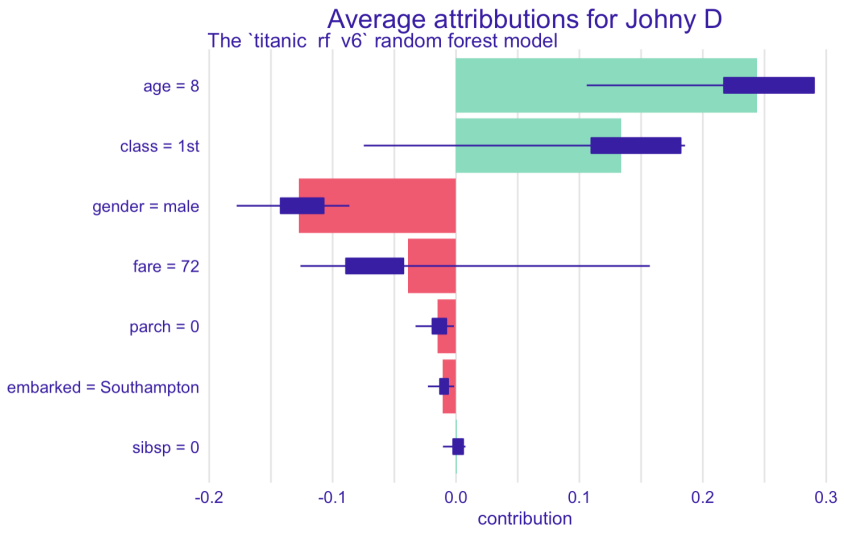


Figure 10.4: SHAP values for synthetic instance and random forest model trained in the Titatnic dataset

The probability assigned by the model to the instance is of 0.422. The plot indicates that the bigger contributions to this score are:

- The age: most of the survivors were children and women
- The class: most of the survivors travelled in first class
- The gender: the person was male so its a factor against his survival since a large part of the survivors were women

The bars indicate the average contributions for random orderings. Red (negative contribution) and green bars (positive contribution) present the averages. Box-plots (in dark blue) summarise the distribution of contributions for each feature across all the orderings.

This way the technique has validates the knowledge over the distribution of people that survived the sinking of the Titanic indicating a model with good performance and allowing us to explain how the model assigns a prediction to specific instances.

However, as with many XAI methods, there might be problems when there is strong correlation between features.

10.2.2 Application to the case study

The result of applying the SHAP values calculated from several randomly selected observations (25 in this case) for the first instance is:

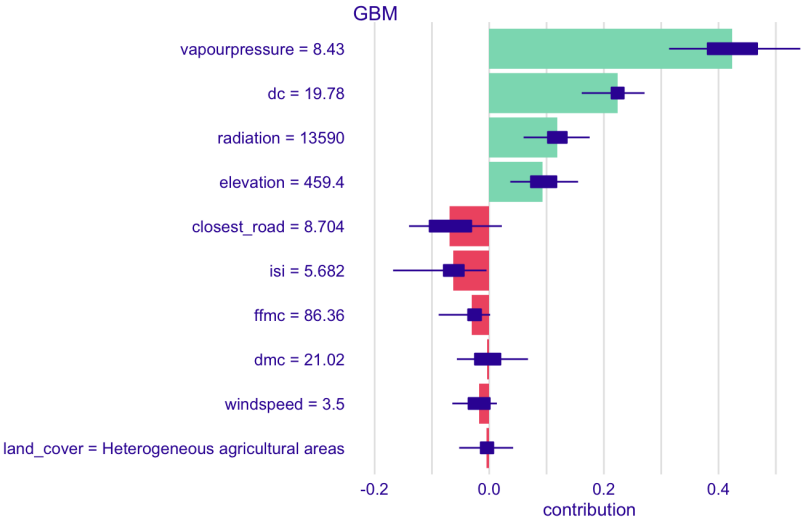


Figure 10.5: Shapely values for the instance #1

The prediction made by the model has a value of 1.832 ha. The model attributes the biggest contribution to the vapour pressure. Such high value indicates that the content of water molecules in the atmosphere acts as a filter that blocks part of the sun radiation dulling its effect.

The second biggest contribution is assigned to the DC index that has a low value for this specific instance. Its value together with the also low value for the DMC index and the value of the vapour pressure indicate a fire typical of those that occur on springtime. This is confirmed by the high value of the FFMC index, which has a small negative contribution to the prediction. However, the fires in Spring are usually wind-driven but the value of the wind speed in the data is the daily average for the station covering the grid square where the fire occurred, so we can not make a conclusion.

The typical relative values of the FFMC, DMC and DC indexes for fires on different periods of the year can be seen in the following figure that depicts them for the zone of the Hogatza river zone in Alaska (U.S.A.) [Nat20]. The values of the indexes and the times of year will change as they need to be adjusted for the local conditions, but the relative relations and patterns will hold:

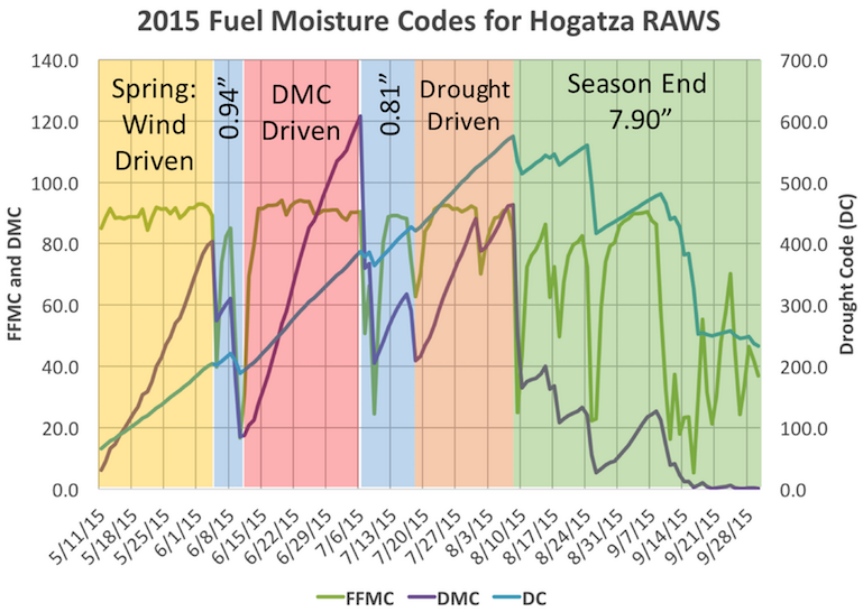


Figure 10.6: FFMC, DMC and DC indexes across the year in the Hogatza river zone in Alaska

The SHAP values of the what-if instance, which increases the value of the FFMC indicator from that of the first instance, is:

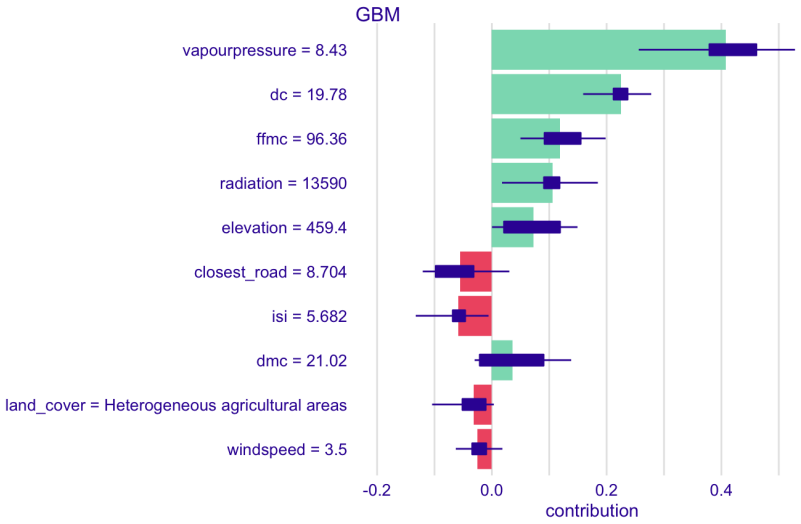


Figure 10.7: Shapely values for the what-if instance with distribution of values

This time the model assigns the biggest contributions again to the vapour pressure and the DC index. However, the FFMC index that for the SHAP

values of the first instance has a small negative contribution now has become the third biggest contributor with a positive contribution. In the light of the changes, the model seems to make a stronger case for a springtime fire.

10.3 Effects of features in a prediction

Techniques focusing on analysing the effect of features in a prediction measure how changes in features cause changes in the prediction. Hence, these techniques allow us to perform sensitivity analysis and measure the change of the predicted value when the features change. For example, with a model the survival probability of cancer patients, the technique would answer the question of how their survival would change if the treatment is changed.

These techniques also allow us to perform local stability tests that would make possible for us to answer what feature changes cause a prediction to change. For example, applying them to a model classifying bank transactions as fraudulent would allow us to answer what changes would flip the prediction of the model and highlighting what it considers to flag a transaction as fraudulent. This can lead to the development of adversarial instances that belong to fraudulent transactions, but the model would not catch.

In addition, these techniques can also be used to perform what-if analysis on individual instances that allows us to understand a model by exploring the influence of different features on the predictions, one feature at a time. A technique that is used to perform this kind of task is Ceteris paribus (CP) profiles. It explores the local curvature of the model response surface using conditional predictions, that is, exploring the effect of changes on a feature has on the prediction while fixing the rest.

The CP profiles for the selected instance using the GBM model are:

The point in each CP profile curve is the actual value of the feature.

The CP profiles of most features remain flat indicating that a change on their values would not affect the value of the prediction. It is only for the ISI and elevation features that the prediction changes, and in both cases, it will increase as the value of the features also increases.

The values of the first instance seem to indicate a typical spring time fire that are usually wind drive so the influence of the ISI is an index integrating the fuel moisture for fine dead fuels and surface wind speed to estimate the fire spread potential seems plausible. This is reinforced by the high value of the FFMC feature that measures the content of fine dead fuels.

However, human-caused fires usually occur in low elevations and their occurrence decrease with altitude since at higher altitudes vegetation and temperature decrease and rendering more difficult the ignition of fire. Hence, the CP profile indicate an instance that requires further investigation.

The CP profiles show us how the prediction will change in response to change in the features, but they fail to show us which ones have the bigger

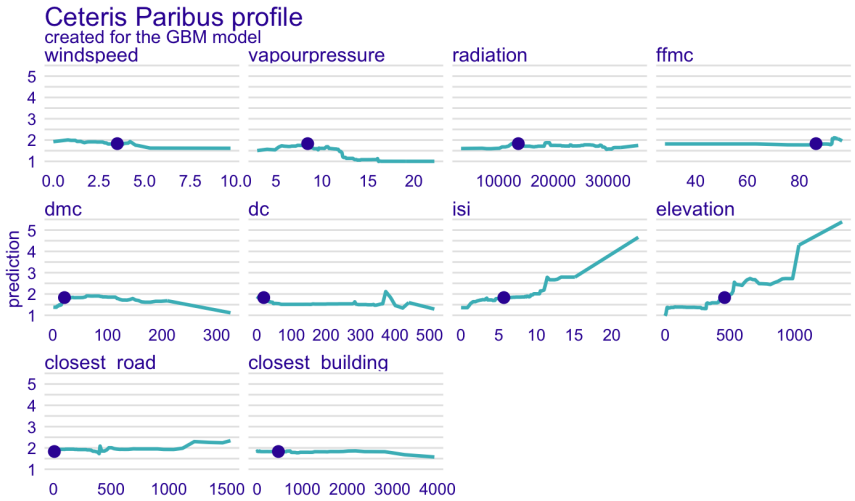


Figure 10.8: CP profiles for all the features and the first selected observation

impact. To do the method used the concept of oscillations of CP profiles.

For features with that has little influence on the prediction the CP profile plot is flat, so the values at any point of the curve are close to the actual value of the feature, indicated by the point in the curve of the plot. But, for features with a high influence on the prediction, the difference between the actual value and the points of the curve will vary wildly.

Therefore, the influence of a feature can be calculated by summing the difference of each value across the curve with the actual value of the feature.

For the case study, the rank of features by influence for the CP profiles of the previous figure are:

Since mostly CP profiles are flat except for those of the elevation and vapour pressure features, these two features rank the highest. However, the rank allows us to know which feature has relatively higher effect, elevation in this case, something that cannot be done simply by visual inspection.

The drawback of this method, as with many others, it is that the presence of correlated features can led to misleading results.

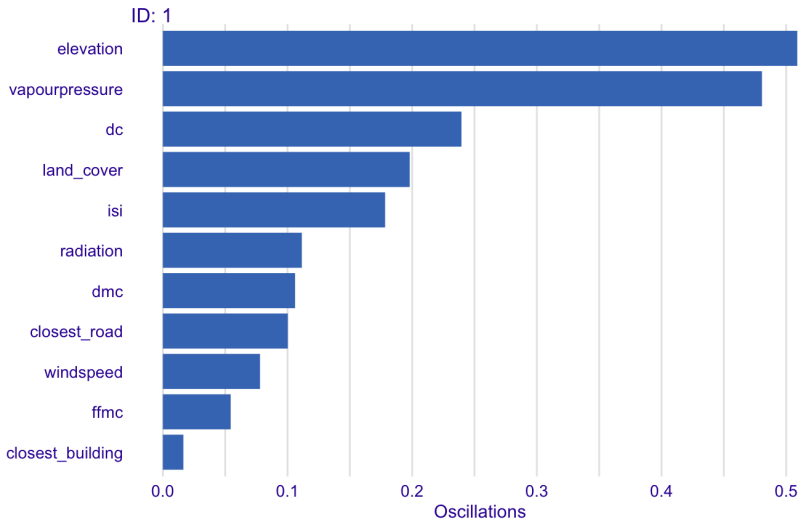


Figure 10.9: CP oscillations for the GBM model and the first observation

Part IV
Conclusions

In this part I describe the conclusions I have reached while working in this project and possible future directions recapping all I have learnt on the course of working on this project.

Chapter 11

Conclusions, and a look to the future

Machine learning is a powerful tool and is our duty that it is used ethically to bring benefits for society. Every other course of action would be a dereliction of duty. We cannot afford to be passive witness or defer responsibility of our contributions as AI establishes itself reaching further and deeper in more applications.

Recently, there has been an increase in the number of articles and initiatives spearheading the use of XAI, first by developing mathematically-sound metrics and a corpus of good practices, and then deploying them in production. However, there is still a long way to go due to the lack of agreement in basic concepts and the fast-growing number of metrics and tools that are a normal sign of new fields. Moreover, the ratification of legislation requiring the explainability of all deployed models might become the catalyst for an explosion of activity and interest in the XAI field.

Nowadays, the use of XAI technique to made models interpretable requires a high degree of expertise in Statistics (and Machine Learning), without forgetting domain knowledge. Without the former, trying to extract meaningful explanations from XAI metrics can lead to making the wrong conclusions, especially in the case of presenting them graphically. And, the importance of domain knowledge is compounded by the nuance of figuring out how to interpret the explanations.

But it has been the lack or domain knowledge that has become a huge handicap in the realisation of the present work. I did not anticipate that the study of wildfires is the confluence of so many diverse disciplines such as ecology, physics, or sociology. The consequence has been that the poor quality of the predictive models, and the difficulty of using the XAI techniques to their full potential. Furthermore, my lack of a deeper statistical and mathematical knowledge has been another impediment to fully understand some of the techniques. Nonetheless, data science projects are multidisci-

plinary affairs, in contrast to the somehow contrived framework that governs projects like this one.

Another obstacle in the realisation of this project has been the large number of new things that I have to learnt, albeit this has been self-inflicted:

- The domain knowledge required by a use case that I have little knowledge of. Before tackling this work I did not know much about wildfires. The multidisciplinary character of this field was a surprise and made more difficult the project (e.g., in judging the more relevant variables to include in the model). Moreover, local conditions and context have a strong influence in the causes and factors influencing wildfires.
- Picking out a programming language (and libraries) in which I have only an experience of no more than three months at the beginning of the project
- A new ML library
- The learning algorithms (GLM and GBM) unbeknown to me

In addition, the realisation of this work while working a full-time job put me under a considerable strain. Regardless of so many difficulties, this has been a fulfilling project.

Out of the project for lack of more time I have left some tasks:

- Use of a larger data set covering a larger span of time given the complexity of the HCFs pattern in mainland Portugal.
- The analysis of the spatio-temporal patterns present in the occurrence of HCFs due the influence of climate and physiography.
- Improve the current model by using a different set of features that can be some of the ones that I have available but discard or new ones.
- Use of different learning algorithms such as XGBoost or deep neural networks.
- Reappraisal of the XAI techniques used.
- • Use of XAI techniques such as prototypes and criticisms that can be used to describe data distributions. The former are observations that are representative of all data whereas the later are observations not well represented by the prototypes.

Appendices

Appendix A

Methodologies

A.1 CRISP-DM

CRISP-DM is a methodology that provides a structured approach to planning a data mining project [Sma20].

The phases of the model are:

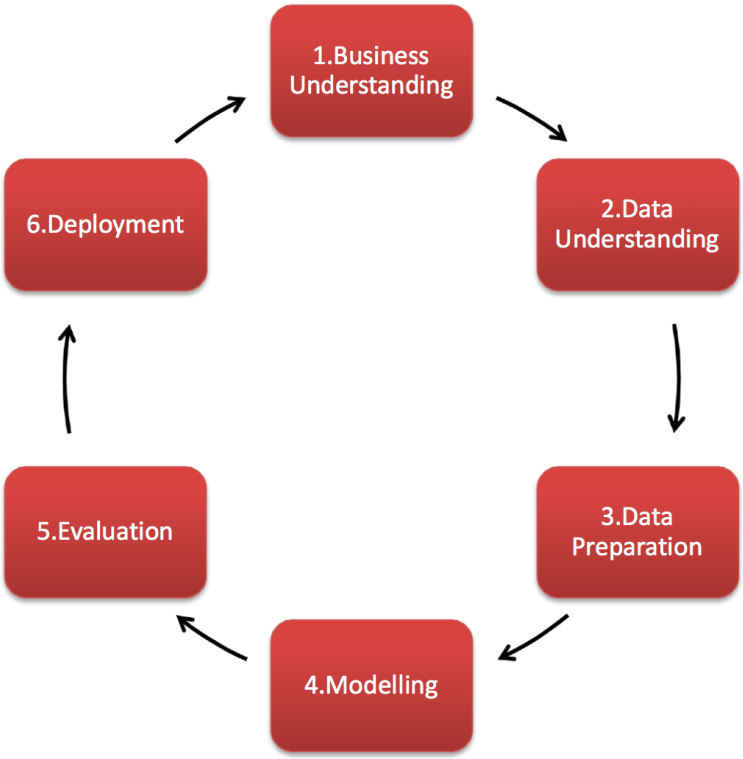


Figure A.1: Phases of CRISP-DM Methodology

This model is an idealised sequence of events and does not try to capture all possible routes through the data mining process. In practice many of the tasks can be performed in a different order and it will often be necessary to backtrack to previous tasks and repeat certain actions:

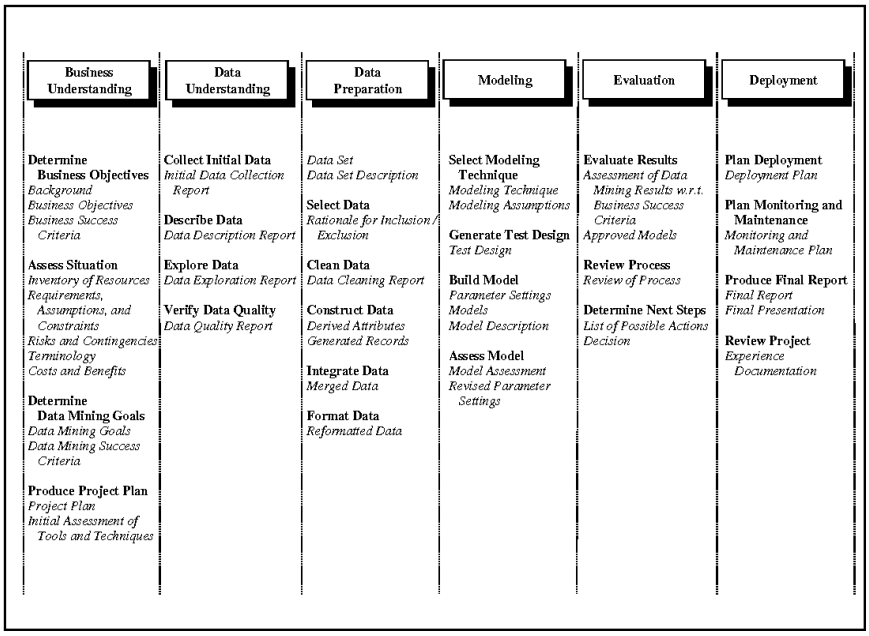


Figure A.2: CRISP-DM Tasks and Outputs

The goals and outputs of the different phases are:

1. 1. Business understanding: The goal is to understand what you want to accomplish from a business perspective and uncover important factors that could influence the outcome of the project. The desired outputs are:

   (a) Description of the objectives

   (b) Project plan

   (c) Business success criteria

2. Data understanding: The goal is to acquire the data listed in the project resources and profile it. The desired outputs are:

   (a) Data exploration report

   (b) Data quality report

3. Data preparation: The goal is to prepare the data for modelling. The single desired outputs is:

(a) Data suitable for modelling

4. Modelling: The goal is to prepare the data for modelling. The desired outputs are:

   (a) Model description
   (b) Trained model

5. Evaluation: The goal is to assess the degree to which the model meets your business objectives. The desired outputs are:

   (a) Model assessment
   (b) Improved model (if necessary)

6. Deployment: The goals is to determine a strategy for their deployment and release the model in production. The desired outputs are:

   (a) Deployment plan
   (b) Monitoring and maintenance plan
   (c) Final report

A.2 Team Data Science Process (TDSP)

TDSP is an agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently [Mic20].

Its key components are:

- A data science lifecycle definition
- A standardised project structure
- Infrastructure and resources for data science projects
- Tools and utilities for project execution

The major stages of its lifecycle that projects typically execute, often iteratively, are:

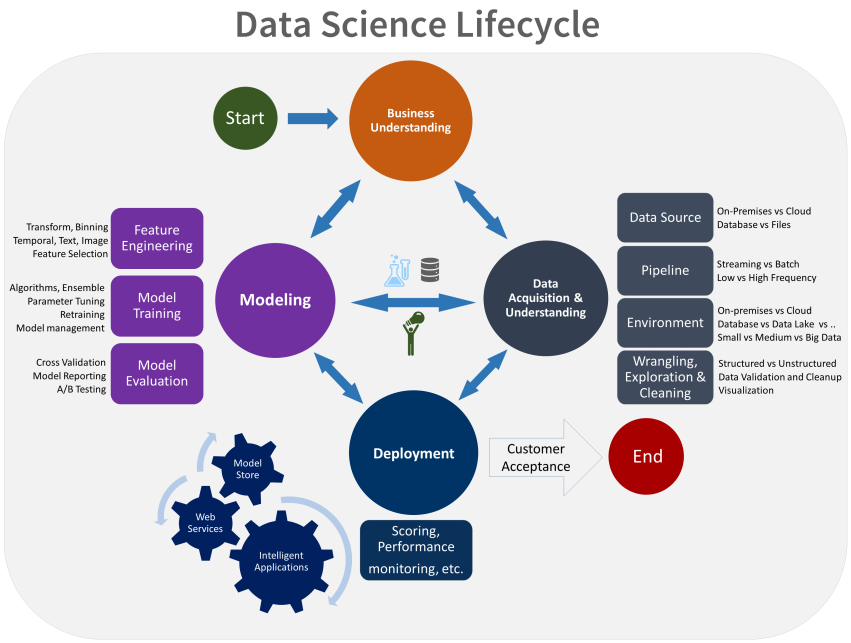


Figure A.3: Team Data Science Process lifecycle

The goals and outputs of each phase are:

1. Business understanding: The goals are to specify the key targets metrics used determine the success of the project and to identify the relevant data sources. The outputs are:

    - Charter document
    - Data sources description
    - Data dictionary

2. Data acquisition and understanding: The goals are to produce a clean, high-quality data set and define the solution architecture. The outputs are:

    - Data quality report
    - Solution architecture

3. Modelling: The goals are to determine the optimal data features and train the model. The outputs are:

    - Feature sets

- Model report

4. Deployment: The goal is to deploy the model with a data pipeline to the production environment for final user acceptance. The outputs are:

    - System status dashboard
    - Final modelling report with deployment details
    - Final solution architecture document

5. Customer acceptance: The goal is to finalise project deliverables.

Appendix B

Metadata

In this appendix I am going to describe the metadata of the data sets used.

The summary of different data sets is:

- Human-Caused Fires (HCFs)
  - Fire occurrence in Portugal from 2011 to 2015
- Major Habitat Types (MHTs)
  - Ecoregions in mainland Portugal
- Weather factors
  - Meteorological data from 2011 to 2015
  - Forest Fire Weather Index (FWI) data from 2011 to 2015
- Physiography
  - Elevation
  - Slope
- Fuel risk factors
  - Land cover in mainland Portugal
- Human factors
  - Distance to nearest road
  - Distance to nearest building

B.1 Human-Caused Fires (HCFs)

The data set is split by year with each year having all or some of the following variables:

- ano: year
- codigo_sgif: code given by the Sistema de Gestão de Informação de Incêndios Florestais (SGIF)
- codigo_anpc: code given by the Autoridade Nacional de Proteção Civil (ANPC)
- tipo: type of fire
- distrito: administrative division
- concelho: administrative division
- freguesia: administrative division
- local: municipality
- ine: code given by the Instituto Nacional de Estatística (INE)
- x: easting coordinate expressed using EPSG:20790 in meters
- y: northing coordinate expressed using EPSG:20790 in meters
- lat: latitude expressed using EPSG:4326 in degree-minute-second units
- lon: longitude expressed using EPSG:4326 in degree-minute-second units
- data_alerta: date of fire detection
- hora_alerta: hour of fire detection
- data_extincao: date of fire extinction
- hora_extincao: hour of fire extinction
- data_primeira_intervencao: date of first firefighting effort
- hora_primeira_intervencao: hour of first firefighting effort
- fonte_alerta: source of the fire warning
- nut: Nomeclatura de Unidades Territoriais (NUT)
- area_povoamento: size of urban area burned (in hectares)

- area_mato: size of shrubland area burned (in hectares)
- area_agricola: size of agricultural area burned (in hectares)
- area_pov_mato: size of urban + shrubland area burned (in hectares)
- area_total: total area burned (in hectares)
- reacendimento: whether the fire was caused by reignition of an extinguished fire
- queimada: whether the fire was caused by pasture renovation or stubble burning
- falso_alarme: whether the fire was a false alarm
- fogacho: whether the fire affected to an area of less than 1 hectare
- incendio: whether the fire affected to an area equal or greater than 1 hectare
- agricola: whether the fire was an agricultural one
- perimetro: perimeter of the area burned (in meters)
- aps: no description available
- causa: source of the fire
- tipo_causa: classification of the source of the fire
- regiao_prof: identifier of the region in the Programas Regionais de Ordenamento Florestal (PROF)
- ugf: no description available

B.2 Major Habitat Types (MHTs)

The geographic extent of the data, given by its bounding coordinates using EPSG:4326, is:

- West Bounding Coordinate: -179.999989
- East Bounding Coordinate: 179.999989
- North Bounding Coordinate: 83.623125
- South Bounding Coordinate: -89.891973

Horizontal Coordinate System Definition:

- Geographic:
  - Latitude Resolution: 0.000001
  - Longitude Resolution: 0.000001
  - Geographic Coordinate Units: Decimal degrees
- Geodetic Model:
  - Horizontal Datum Name: D_WGS_1984
  - Ellipsoid Name: WGS_1984
  - Semi-major Axis: 6378137.000000
  - Denominator of Flattening Ratio: 298.257224

Vertical Coordinate System Definition:

- Altitude Resolution: 0.000010
- Altitude Encoding Method: Explicit elevation coordinate included with horizontal coordinates

Entity Information:

- Entity Type Label: wwf_terr_ecos
- Entity Type Definition: WWF's Terrestrial Ecoregions of the World
- Entity Type Definition Source: `http://www.worldwildlife.org/ecoregions/attributes.htm`

Attribute Information:

- OBJECTID: internal unique feature number.
- Shape: coordinates defining the geometry of features.
- AREA: area of each individual polygon in square kilometers. The attribute area_km2 is a sum of this field for each ecoregion.
- PERIMETER: perimeter.
- ECO_NAME: ecoregion name.
- REALM: biogeographical realm. It can take one of the following values:
  - AA: Australasia.
  - AN: Antarctic
  - AT: Afrotropics

  - IM: IndoMalay
  - NA: Nearctic
  - NT: Neotropics
  - OC: Oceania
  - PA: Palearctic

- BIOME: biome. It can take one of the following values:

  - 1: tropical & subtropical moist broadleaf forests
  - 2: tropical & subtropical dry broadleaf forests
  - 3: tropical & subtropical coniferous forests
  - 4: temperate broadleaf & mixed forests
  - 5: temperate conifer forests
  - 6: boreal forests / taiga
  - 7: tropical & subtropical grasslands, savannas & shrublands
  - 8: temperate grasslands, savannas & shrublands
  - 9: flooded grasslands & savannas
  - 10: montane grasslands & shrublands
  - 11: tundra
  - 12: mediterranean forests, woodlands & scrub
  - 13: deserts & xeric shrublands
  - 14: mangroves

- ECO_NUM: unique number for each ecoregion within each biome nested within each realm.

- ECO_ID: number created by combining REALM, BIOME, and ECO_NUM, thus creating a unique numeric ID for each ecoregion.

- ECO_SYM: ecoregion symbol (used to display the map in Esri ArcInfo)

- eco_code: alphanumeric code that is similar to ECO_ID but a little easier to interpret. The first 2 characters (letters) are the realm the ecoregion is in. The 2nd 2 characters are the biome and the last 2 characters are the ecoregion number.

- GBL_STAT: global status. A 30-year prediction of future conservation status given current conservation status and trajectories. It can take one of the following values:

  - 1: CRITICAL OR ENDANGERED

  - 2: VULNERABLE
  - 3: RELATIVELY STABLE OR INTACT

- G200_NUM: Global 200 number. The Global 200 is the list of ecoregions identified by WWF as priorities for conservation.
- G200_BIOME: Global 200 biome. The biome of the Global 200 region that the ecoregion is a component of. Occasionally a Global 200 region is made up of ecoregions of different biomes. In this case, for an specific ecoregion, the 'G200_BIOME' may be different from the original biome. For a description of each Global 200 biome (1-14), see the attribute BIOME.
- FID': internal unique feature number.
- G200_STAT': Global 200 conservation status. It can take one of the following values: • 1: CRITICAL OR ENDANGERED • 2: VULNERABLE • 3: RELATIVELY STABLE OR INTACT
- Shape_Area: area of feature in internal units squared.
- area_km2: area of the Ecoregion in kilometers squared.
- G200_REGIO: Global 200 name
- Shape_Leng: shape length.

B.3 Weather factors

B.3.1 Meteorological data

The metadata of the data set is:

- Version: 2.0
- Date published: 14/01/2019
- Creator: Directorate D - Sustainable resources / Unit 05 - Food security
- Publisher: FOODSECURITY-MARS4CAST
- Grid spatial projection: Lambert azimuthal equal area
- Grid EPSG code: 3035
- Grid resolution: 25 km
- Period: 01/01/1975–31/12/2018

- Time resolution: 1 day
- Indicators:
  - Maximum air temperature (°C)
  - Minimum air temperature (°C)
  - Mean air temperature (°C)
  - Mean daily wind speed at 10m (m/s)
  - Vapour pressure (hPa)
  - Sum of precipitation (mm/day)
  - Potential evaporation from a free water surface (mm/day)
  - Potential evapotranspiration from a crop canopy (mm/day)
  - Potential evaporation from a moist bare soil surface (mm/day)
  - Total global radiation (KJ/m2/day)
  - Snow depth (cm)

B.3.2 Canadian Forest Fire Weather Index (FWI)

File naming convention:

> ECMWF_FWI_FWI_19790524_1200_spread_v3.1_con.nc PRODUCER_MODEL_[VARIABLE_VARIABLE]_DATE_TIME_FCTYPE_Version_Dataset.nc

1. Producer:
   (a) ECMWF
2. Model:
   (a) FWI
   (b) MARK5
   (c) NFDRS
3. Variable (some variables have _ in name):
   (a) FWI Variables : FWI, BUI, DANGER_RISK,DC,DMC,ISI,FFMC.DSR
   (b) MARK5 Variables: KBDI,DF,ROS,FDI
   (c) NFDRS Variables: SC,ERC,BI,IC
4. Date:
   (a) Date of the Reanalysis

5. Time:
    (a) Time of the Reanalysis
6. Fctype:
    (a) Hr: Reanalysis, ERA5 Reanalysis Short Model at 0.25 Degrees
    (b) En: Ensemble Members, ERA5 Reanalysis ensemble at 0.50 Degrees
    (c) Spread: Ensemble Spread at 0.50 Degrees
    (d) Mean: Ensemble Mean at 0.50 Degrees
7. Version:
    (a) 3.0:
    (b) 3.1: Updated smoothing of data and Drought Coefficients
8. Dataset:
    (a) Con: Consolidated ERA5 Data (2-3 Month Lag behind real time)
    (b) Int: Intermediate ERA5 Data (up to 5 days lag behind real time, product could differ when quality checks and consolidation occurs)

The metadata of the data set is:

- Data type: gridded
- Horizonal coverage: Global Land
- Horizontal resolution Reanalysis: 0.25°x0.25°
- Temporal coverage: 1979 to present
- Temporal resolution: Daily
- File format: NetCDF
- Update frequency: Monthly
- Consolidated Dataset: Based upon the Officially released ERA5 Reanalysis
- Version: 3.1
- Release date: 2019
- Date of Initialisation of Intermediate Dataset: 20190831

- Start Date (Consolidated): 19790103
- End Date (Consolidated): On going
- Fire weather index:
  - Numerical rating of fire intensity. It is suitable as a general index of fire danger
  - Numerical rating
- Initial Spread Index
  - Numerical rating of the expected rate of fire spread
  - Numerical rating
- Build-up Index
  - Numerical rating of the total amount of fuel available for combustion
  - Numerical rating
- Drought Code
  - Numerical rating of the average moisture content of deep, compact organic layers
  - Numerical rating
- Duff Moisture Code
  - Numerical rating of the average moisture content of loosely compacted organic layers of moderate depth
  - Numerical rating
- Fine Fuel Moisture Code
  - Numerical rating of the moisture content of litter and other cured fine fuels.
  - Numerical rating

B.4 Physiography

B.4.1 Elevation

The metadata of the EU-DEM v1.0 is:

- Bounding box:

  - West bounding coordinate: -10.61982
  - East bounding coordinate: 44.82124
  - North bounding coordinate: 71.18545
  - South bounding coordinate: 34.56192
- Coordinate reference system: EPSG:3035 (ETRS89, LAEA)
- Temporal extent: 2000
- Date of publication: Apr 20, 2016
- Lineage: The EU-DEM v1.0 is derived from an automated data fusion process using SRTM and ASTER GDEM digital surface model (DSM) data. Intermap's NEXTMap Europe dataset is utilized to remove any consistent horizontal bias in the GDEM data. The EU-DEM v1.0 is edited to ensure that water features are adequately represented and consistent with the hydrography layer. Residual clouds within the GDEM data are identified and removed same as suspect data extremely differing from the SRTM data. All EU-DEM tiles are edited interactively in a 3D stereo environment. The editing is restricted to the hydrographical features and pits and bumps. In areas above 60 degrees North, the EU-DEM generation process is supported by other DEM data sources provided by the JRC. Water features are flattened (oceans, lakes) and stepped (rivers) based on the hydrography data. The spatial reference system is geographic, lat/lon with horizontal datum ETRS89, ellipsoid GRS80 and vertical datum EVRS2000 with geoid EGG08.
- Spatial resolution: 25 m
- Specification: Commission Regulation (EU) No 1089/2010 of 23 November 2010 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards interoperability of spatial data sets and services, Date of publication: 2010-12-08.

B.4.2 Slope

The metadata of the slope data derived from EU-DEM v1.0 is:

- Bounding box:
  - West bounding coordinate: -10.61982
  - East bounding coordinate: 44.82124
  - North bounding coordinate: 71.18545
  - South bounding coordinate: 34.56192

- Coordinate reference system: EPSG:3035 (ETRS89, LAEA)
- Temporal extent: 2000
- Date of publication: Apr 20, 2016
- Lineage: The EU-DEM v1.0 is derived from an automated data fusion process using SRTM and ASTER GDEM digital surface model (DSM) data. Intermap's NEXTMap Europe dataset is utilized to remove any consistent horizontal bias in the GDEM data. The EU-DEM v1.0 is edited to ensure that water features are adequately represented and consistent with the hydrography layer. Residual clouds within the GDEM data are identified and removed same as suspect data extremely differing from the SRTM data. All EU-DEM tiles are edited interactively in a 3D stereo environment. The editing is restricted to the hydrographical features and pits and bumps. In areas above 60 degress North, the EU-DEM generation process is supported by other DEM data sources provided by the JRC. Water features are flattened (oceans, lakes) and stepped (rivers) based on the hydrography data. The spatial reference system is geographic, lat/lon with horizontal datum ETRS89, ellipsoid GRS80 and vertical datum EVRS2000 with geoid EGG08. The slope dataset has been created from the EU-DEM projected to ETRS89/ETRS-LAEA.
- Spatial resolution: 25 m
- Specification: Commission Regulation (EU) No 1089/2010 of 23 November 2010 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards interoperability of spatial data sets and services, Date of publication: 2010-12-08.

B.5 Fuel risk factors

B.5.1 Land cover

The metadata of CLC 2018 is:

- Bounding Box:
  - Region 1:
    - West = -31.561261
    - East = 44.820775
    - North = 71.409109
    - South = 27.405827
  - Region 2:

    * West = -61.906047
    * East = -60.905616
    * North = 16.607552
    * South = 15.736333
  - Region 3:
    * West = -54.268239
    * East = -51.621253
    * North = 5.851958
    * South = 3.772692
  - Region 4:
    * West = -61.326095
    * East = -60.711516
    * North = 14.970484
    * South = 14.29692
  - Region 5:
    * West = 44.927382
    * East = 45.390135
    * North = -12.546691
    * South = -13.089579
  - Region 6:
    * West = 55.114983
    * East = 55.935919
    * North = -20.77811
    * South = -21.482245

  xs

- Coordinate Reference System: EPSG:3035 (ETRS89, LAEA)
- Temporal extent: 2017-2018
- Date of publication: Jun 14, 2019
- Spatial resolution: Minimum Mapping Unit (MMU): 25 ha
- Conformity: Commission Regulation (EU) No 1089/2010 of 23 November 2010 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards interoperability of spatial data sets and services, Date of publication: 2010-12-08.
- Responsible party: European Environment Agency (EEA) under the framework of the Copernicus programme - copernicus@eea.europa.eu.

The lineage of the CLC Version 20 is: Vector CLC database was provided by National Teams within original CLC1990, CLC2000 update, CLC2006 update, CLC2012 update and CLC2018 update projects. All features in original vector database were classified and digitised based on satellite images with 100 m positional accuracy (according to CLC specifications) and 25 ha minimum mapping unit into the standardized CLC nomenclature (44 CLC classes).

European Corine Land Cover seamless DBs represent the final product of European data integration. The process of data integration started when national deliveries have been accepted and the Database Acceptance Report (DBTA) delivered. Delivered national data were produced in local national systems of all participating countries. Each national Coordinate Reference System (CRS) definition had to be known precisely together with its geometric relationship to a standard system in order to accurately transfer all national data into a standard European coordinate reference - ETRS89/LAEA1052. Mostly, the process itself was carried out by global equation-based transformation to ETRS89 (e.g. seven-parameters Bursa-Wolf methods). The accuracy of a particular transformation ranges from centimetres to meters depending on the method and the quality and number of control points available to define the transformation parameters, but, in any case, the accuracy is far above the actual CLC data resolution (for more details see the DBTA reports for particular country). National data, when transformed into the common European reference, are introduced into tiled pan-European structure and as final step seamless dataset is produced.

In order to achieve production of the real seamless European database, the integration step includes also harmonization of database along country borders. It consists from edge-matching of land cover polygons from the national databases across national borders done by a verification / reinterpretation of the satellite images in the border regions (2 km wide strip along borders). The satellite images from IMAGE2000. CLC1990, CHA9000 and CLC2000 database were harmonized this way, but the order to priority was as following: CLC2000, both geometric and thematic adaptations of all polygons in a 2 km strip along national boundary lines; CHA9000 database to ensure that changes in CLC2000 are consistent with the change database; corrected CLC90 (if provided by the MS); corrections were focused to geometric adaptations in semi-automatical way based on CLC00 and CHA00 databases. Border harmonization step has been skipped for CHA0006, CHA0612, CLC2012 and clc2018 datasets. Simplified border harmonization step for CLC2006 dataset has been created for these countries: CH, NO, XK, TR, IE.

A simplified border matching has been applied:

- $<$ 25 ha polygons are NOT systematically removed (see next bullet).
- Sliver-like polygons (area $<$ cca. 5 ha - soft limit) are generalised to

largest or thematically most similar neighbour.

- CLC-code differences in polygons along two sides of the border are NOT changed

Only polygons with area <= 0,1ha were eliminated in CHA0006, CHA0612, CHA1218, CLC2012 and for CLC2018 datasets and CLC2006 dataset (besides the above-mentioned cases) and in parts newly added in campaigns 2006 and 2012 too.

File name is a combination of campaign year, reference year, inventory year and release version.

General file name format:

[Mapping_Campaign]_[CLC_Reference_Year]_[Created_Inventory_Year]_[Version]

Stable version example with file name CLC2006_CLC2000_V2018_20 means:

- CLC2006_ That the file came from the 2006 mapping campaign (the file content was last modified in this campaign)
- _CLC2000_ That the file captures Land Cover mapping results for 2000 reference year
- _V2018_ That this file comes from a delivery created in 2018 inventory year
- _20 That this is the final stable version

Beta-version example with file name CLC2006_CLC2000_V2018_20b2 means:

- CLC2006_ That the file came from the 2006 mapping campaign (the file content was last modified in this campaign)
- _CLC2000_ That the file captures Land Cover mapping results for 2000 reference year
- _V2018_ That this file comes from a delivery created in 2018 inventory year
- _20b2 That this is the second beta-version

Data type changed from uint8 to uint16 for status rasters in 2018 release.

Some artificial lines (dividing polygons with the same code) can be still present in database due to technical constraints of current ArcGIS technology, but has no impact for dataset contents and can be dissolved for data extracts.

B.6 Human factors

B.6.1 Distance to nearest road

The metadata of Geofabrik Esri Shapefiles from OpenStreetMap containing roads and paths is:

- Encoding: UTF-8
- Geometry: Line (MultiLineString)
- CRS: EPSG:4326 - WGS 84 – Geographic
- Extent:
  - -31.2665834999999994, 32.6332820000000012
  - -6.1748000999999997, 42.2029769000000030
  - Unit: degrees
- Feature count: 950,238
- Attributes:
  - id
    - VARCHAR (4 Bytes)
    - Id of this feature. Unique in this layer.
  - osm_id
    - VARCHAR (10)
    - OSM Id taken from the Id of this feature (node_id, way_id, or relation_id) in the OSM database. In case several features in the OSM database are joined into one feature, this is one of the Ids. This Id is not necessarily unique because one OSM object can result in several geometry objects.
      Also note that when doing shape file exports, this will be exported as a VARCHAR type since shape files don't support long integers.
  - code
    - SMALLINT (2 Bytes)
    - 4 digit code (between 1000 and 9999) defining the feature class. The first one or two digits define the layer, the last two or three digits the class inside a layer.
  - fclass
    - VARCHAR(40)

    - Class name of this feature. This does not add any information that is not already in the “code” field but it is better readable.
  - name
    - VARCHAR(100)
    - Name of this feature, like a street or place name. If the name in OSM contains obviously wrong data such as “fixme” or “none”, it will be empty.
  - ref
    - VARCHAR(20)
    - Reference number of this road (‘A 5’, ‘L 605’,...)
  - oneway
    - VARCHAR(1)
    - Is this a oneway road? “F” means that only driving in direction of the linestring is allowed. “T” means that only the opposite direction is allowed. “B” (default value) means that both directions are ok.
  - maxspeed
    - SMALLINT
    - Max allowed speed in km/h
  - layer
    - SMALLINT
    - Relative layering of roads (-5, ..., 0, ..., 5)
  - bridge
    - VARCHAR(1)
    - Is this road on a bridge? (“T” = true, “F” = false)
  - tunnel
    - VARCHAR(1)
    - Is this road in a tunnel? (“T” = true, “F” = false)

B.6.2 Distance to nearest building

The metadata of Geofabrik Esri Shapefiles from OpenStreetMap containing buildings is:

- Encoding: UTF-8
- Geometry: Line (MultiLineString)
- CRS: EPSG:4326 - WGS 84 – Geographic

- Extent:
  - -31.2661017000000001, 32.5132367000000002
  - -6.1857366000000003, 42.1595287999999968
  - Unit: degrees
- Feature count: 976,014
- Attributes:
  - id
    - VARCHAR (4 Bytes)
    - Id of this feature. Unique in this layer.
  - osm_id
    - VARCHAR (10)
    - OSM Id taken from the Id of this feature (node_id, way_id, or relation_id) in the OSM database. In case several features in the OSM database are joined into one feature, this is one of the Ids. This Id is not necessarily unique because one OSM object can result in several geometry objects.
      Also note that when doing shape file exports, this will be exported as a VARCHAR type since shape files don't support long integers.
  - code
    - SMALLINT (2 Bytes)
    - 4 digit code (between 1000 and 9999) defining the feature class. The first one or two digits define the layer, the last two or three digits the class inside a layer.
  - fclass
    - VARCHAR(40)
    - Class name of this feature. This does not add any information that is not already in the "code" field but it is better readable.
  - name
    - VARCHAR(100)
    - Name of this feature, like a street or place name. If the name in OSM contains obviously wrong data such as "fixme" or "none", it will be empty.
  - type
    - VARCHAR(20)
    - The type of building, if specified in OSM; otherwise empty.

Appendix C

Detailed data exploration

In this appendix I am going to include the complete data exploration.

I am going to se descriptive statistical techniques (numerical and graphical) to explore and audit the data to judge its usefulness for building machine-learning models to predict the burnt area by human-caused fires (HCFs). Also, as the data is georeferenced I also going to use GIS techniques to explore and process the data.

Because fire risk factors and their importance change between MHTs [Cos17], I am going to use them to stratify the data. This course of action is used in multiple studies about wildfires.

It is necessary to mention that all data references the location of the fires and the coordinates are encoded using EPSG:4326 in decimal degrees.

C.1 High-level view of a variable by biome

To explore this aspect of the data I am going to use contingency tables and optionally bar plots to explore a variable grouped by biome.

C.1.1 Human-caused fires (HCFs): burnt area

The total number of observations and the burnt area by biome are:

Biome	Count		Burnt area	
Temperate Broadleaf & Mixed Forests	21,590	45.05%	127,331.6	40.55%
Mediterranean Forests, Woodlands & Scrub	26,337	54.95%	186,691.9	59.45%
Total	47,927	100%	314,023.6	100%

Table C.1: Burnt area observations by biome

Graphically is, the count is:

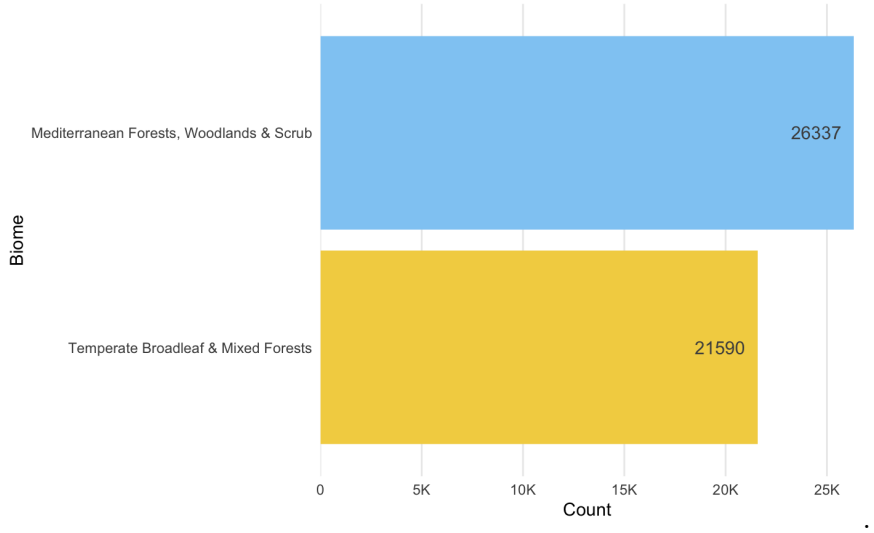


Figure C.1: HCFs count by biome

And, the burnt area is:

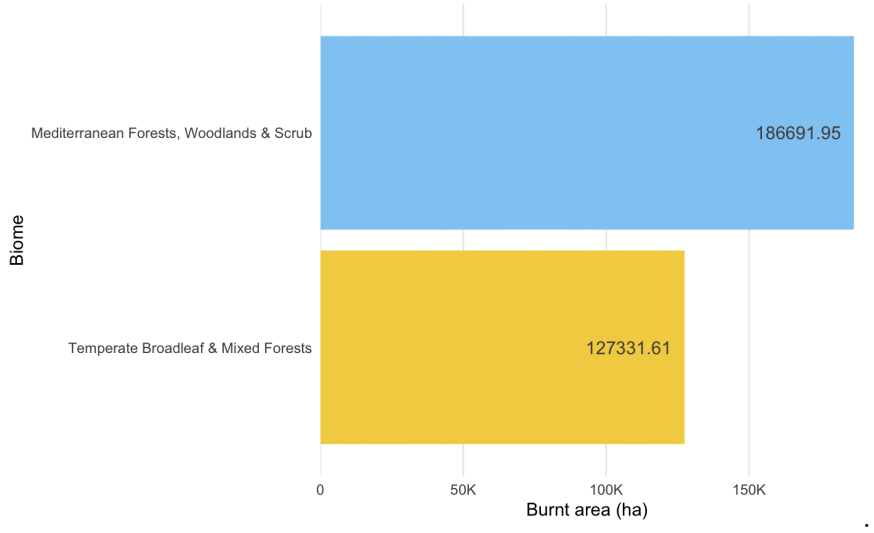


Figure C.2: Burnt area by biome

The count and burnt area is larger on the Mediterranean Forests, Woodlands & Scrub biome than on the Mediterranean Forests, Woodlands & Scrub one, but the difference is small and both metrics are almost evenly distributed between biomes.

C.1.2 Fuel risk factors: land cover

The HCFs are in 42 different land cover categories. However only 4 of them have a number of observations larger than the 10% of the total:

id	Description	Frequency	Percentage
242	Complex cultivation patterns	8,486	17.71%
112	Discontinuous urban fabric	6,576	13.72%
243	Land principally occupied by agriculture, with significant areas of natural vegetation	6,000	12.52%
241	Annual crops associated with permanent crops	5,943	12.40%

Table C.2: CLC level 3 categories with more than 10% of observations

And the topmost categories by biome, beginning with the Temperate Broadleaf & Mixed Forests one are:

id	Description	Frequency	Percentage
241	Annual crops associated with permanent crops	4,008	18.56%
112	Discontinuous urban fabric	3,272	15.16%
242	Complex cultivation patterns	3,086	14.29%
243	Land principally occupied by agriculture, with significant areas of natural vegetation	3,086	14.29%

Table C.3: CLC level 3 categories with more than 10% of observations in Temperate Broadleaf & Mixed Forests biome

While that for the Mediterranean Forests, Woodlands & Scrub biome are:

id	Description	Frequency	Percentage
242	Complex cultivation patterns	5,400	20.50%
112	Discontinuous urban fabric	3,304	12.55%
243	Land principally occupied by agriculture, with significant areas of natural vegetation	2,914	11.06%

Table C.4: CLC level 3 categories with more than 10% of observations in Mediterranean Forests, Woodlands & Scrub biome

There are a large number of categories represented but most of them with

low cardinality. Thus, the variability and noise might become a problem building a model.

C.2 High-level view of a variable by temporal unit

To explore this aspect of the data I am going to use contingency tables and line plots to explore the distribution of a variable by temporal unit (year or month).

C.2.1 Human-caused fires (HCFs): burnt area

The total count of HCFs and burnt area by year is:

Year		Count		Burnt area
2011	12,466	26.01%	56,718.20	18.06%
2012	12,430	25.94%	93,765.88	29.86%
2013	10,355	21.61%	10,7005.39	34.08%
2014	3,641	7.60%	13,795.27	4.39%
2015	9,035	18.85%	42,738.82	13.61%
Total	47,927	100%	314,023.6	100%

Table C.5: HCFs count and burnt area by biome and year

If we count by biome and year, beginning with the Temperate Broadleaf & Mixed Forests biome we have:

Year	Count	Burnt area (ha)
2011	6,008	19,378.28
2012	5,295	22,766.03
2013	5,416	65,078.67
2014	1,089	2,478.625
2015	3,782	17,629.99

Table C.6: Count and burnt area on Temperate Broadleaf & Mixed Forests biome by year

And, continuing with the Mediterranean Forests, Woodlands & Scrub one we have:

Year	Count	Burnt area (ha)
2011	6,458	37,339.92
2012	7,135	70,999.85
2013	4,939	41,926.71
2014	2,552	11,316.643
2015	5,253	25,108.825

Table C.7: Count and burnt area on Mediterranean Forests, Woodlands & Scrub biome by year

Graphically, beginning with the yearly count:

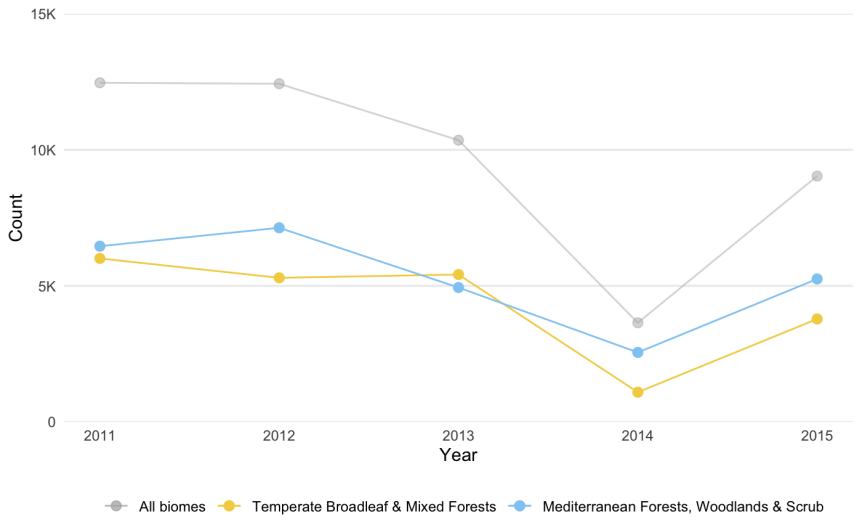


Figure C.3: Yearly count of HCFs in the period 2011–2015: total and by biome

And, the yearly burnt area:

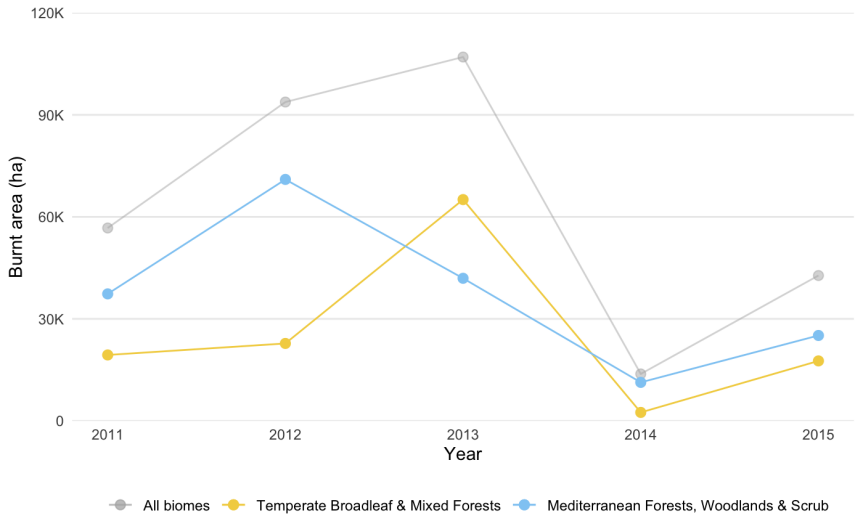


Figure C.4: Yearly burnt area in the period 2011–2015: total and by biome

The same data but with monthly granularity for the count of HCFs:

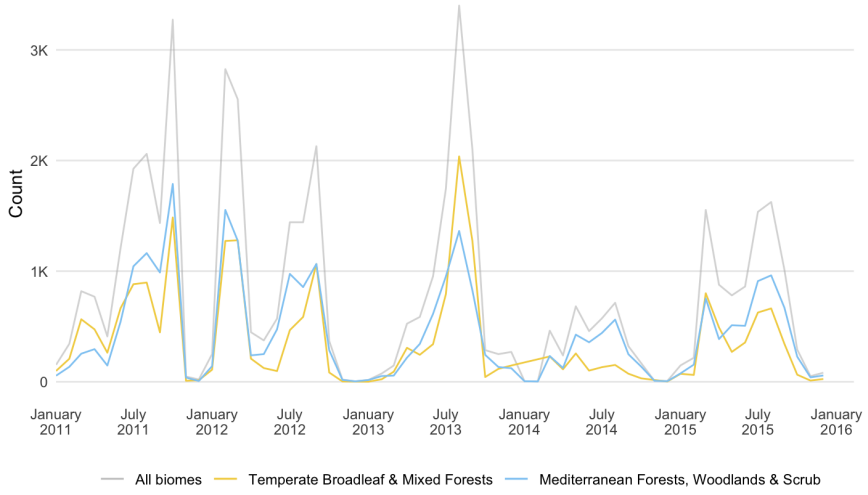


Figure C.5: HCFs count by month in the period 2011–2015: total and by biome

And, for the burnt area:

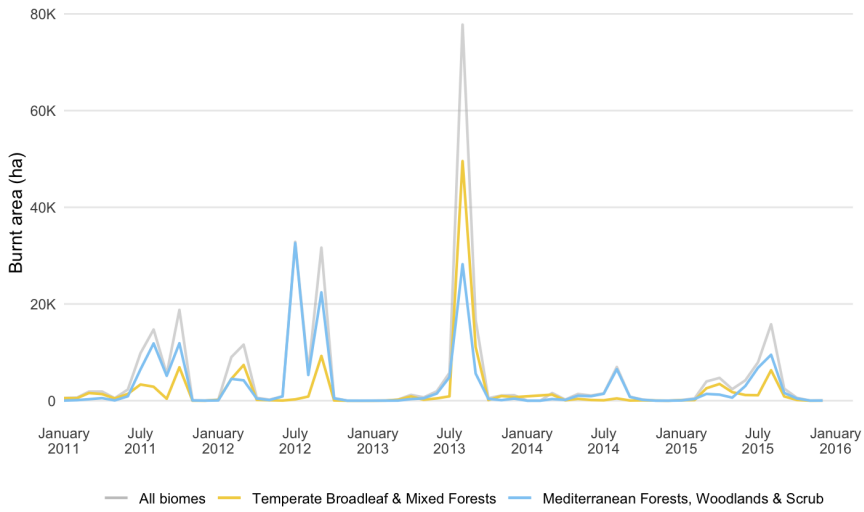


Figure C.6: HCFs burnt area by month in the period 2011–2015: total and by biome

The two periods of the year with higher occurrence of fires, early spring and all all summer (sometimes extending to early autumn), are visible in the plots.

Superimposing the count by year to check for possible seasonal patterns:

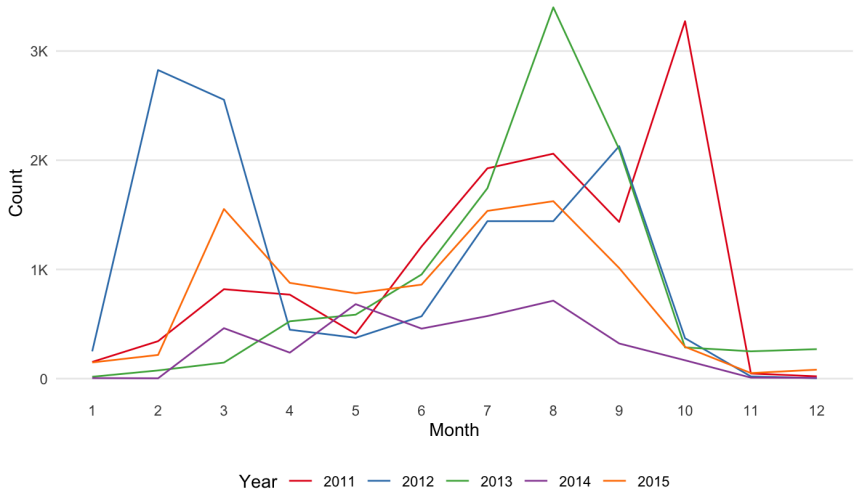


Figure C.7: HCFs occurrence monthly trend by year

And, the burnt area by month:

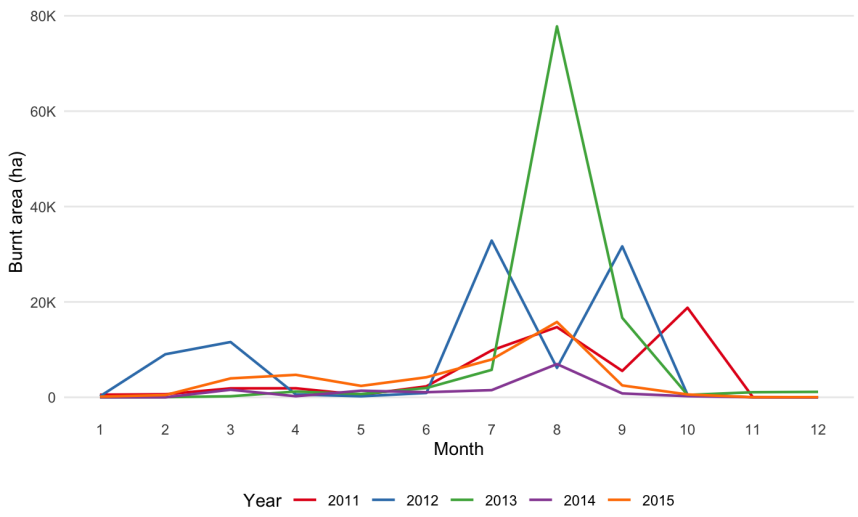


Figure C.8: HCFs burnt area monthly trend by year

It is evident for the data that the period of maximum fire occurrence happens on late summer or early autumn depending on the year. Even though, in some years there is another period of high fire occurrence on early spring.

This is consistent with the literature that describes the seasonal patterns of fires in Portugal shaping a extraordinarily complex multi-year pattern with "high highs" and "low lows" years, with the year 2014 identifiable as one of the "low lows" ones. However, the period of study is too short to appreciate the pattern.

Also, the virulence of HCFs is greater on the summer months.

C.3 Shape of variable's distribution by biome

To explore this aspect of the data I am going to use summary statistics metrics (median and the mean, the first and third quartiles, the minimum and maximum values, and the standard deviation) together with density plots (instead of histograms) of a variable grouped by biome.

C.3.1 Human-caused fires (HCFs): burnt area

The statistical summary of the burnt area is:

Min.	Q_1	Median	Mean	Q_3	Max.
0.00	0.01	0.10	6.55	0.90	24,843.00

Table C.8: Statistical summary of burnt area variable

With its standard deviation being 139.03.

The minimum for the variable is not zero but 0.00001 (zero values where eliminated in the data pre-processing step). Also, the maximum value corresponds to one of the largest wildfires in the history of Portugal[1].

The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	0.00	0.02	0.10	5.90	1.00	6,853.50
Mediterranean Forests, Woodlands & Scrub	0.00	0.01	0.10	7.09	0.78	24,843.00

Table C.9: Statistical summary of burnt area variable by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	85.62
Mediterranean Forests, Woodlands & Scrub	170.77

Table C.10: Standard deviation of burnt area variable by biome

Its evident the influence of the large wildfire in the summary statistics.

C.3.2 Human-caused fires (HCFs): coordinates (x and y)

The coordinates are described by the x (i.e, longitude) and y (i.e., latitude) variables using the EPSG:4326 Coordinate Reference System (CRS) in decimal degrees.

The summary statistics for the x and y variables are:

Variable	Min.	Q_1	Median	Mean	Q_3	Max.
Longitude	-9.48	-8.52	-8.20	-8.11	-7.75	-6.21
Latitude	37.02	40.15	41.03	40.66	41.45	42.15

Table C.11: Statistical summary of x and y variables

The overall density plot and by biome beginning with the x variable:

[1] Incêndio do Algarve foi o segundo maior de sempre em Portugal

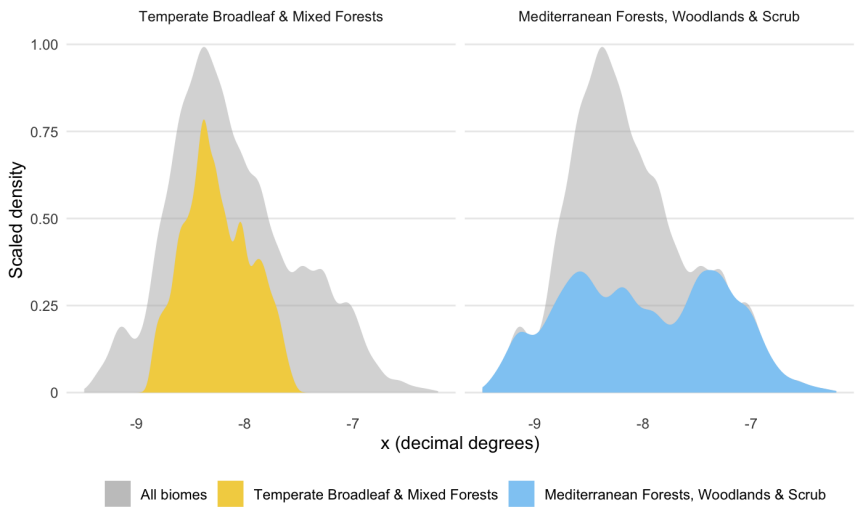


Figure C.9: Density plots by biome of x against overall distribution

And, y:

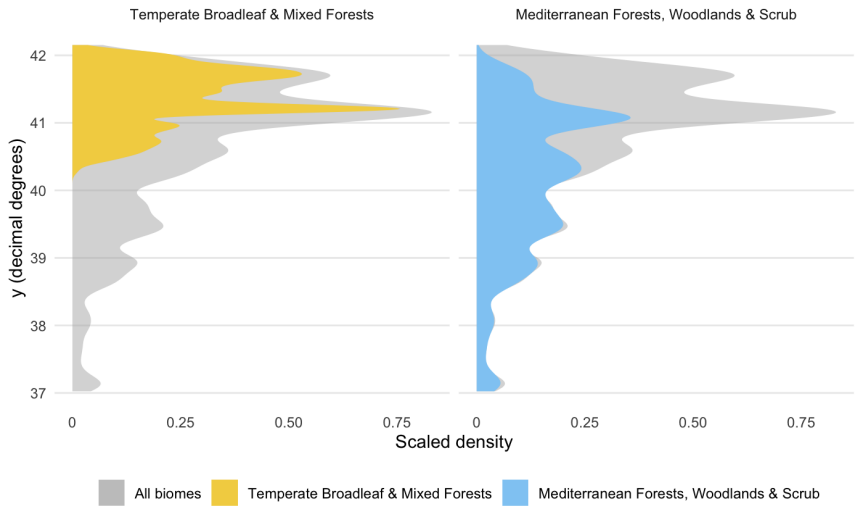


Figure C.10: Density plots by biome of y against overall distribution

The differences on the density plot originate in the following circumstances:

- The number of HCFs is close to being evenly distributed between biomes with 45.05% for the "Temperate Broadleaf & Mixed Forests" biome and 54.95% for the "Mediterranean Forests, Woodlands & Scrub" one.

- The "Temperate Broadleaf & Mixed Forests" occupying a much smaller area in the northern half of continental Portugal as can be seen in the Figure 5.1.

C.3.3 Meteorological data: maximum air temperature

The statistical summary of the maximum air temperature is:

Min.	Q_1	Median	Mean	Q_3	Max.
2.70	21.10	25.30	25.09	29.80	41.80

Table C.12: Statistical summary of maximum temperature

With its standard deviation being 6.34.
The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	2.80	20.60	24.50	24.37	29.00	39.80
Mediterranean Forests, Woodlands & Scrub	2.70	21.60	26.00	25.68	30.40	41.80

Table C.13: Statistical summary of maximum temperature by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	6.12
Mediterranean Forests, Woodlands & Scrub	6.47

Table C.14: Standard deviation of maximum temperature by biome

And, graphically the overall density plot and by biome:

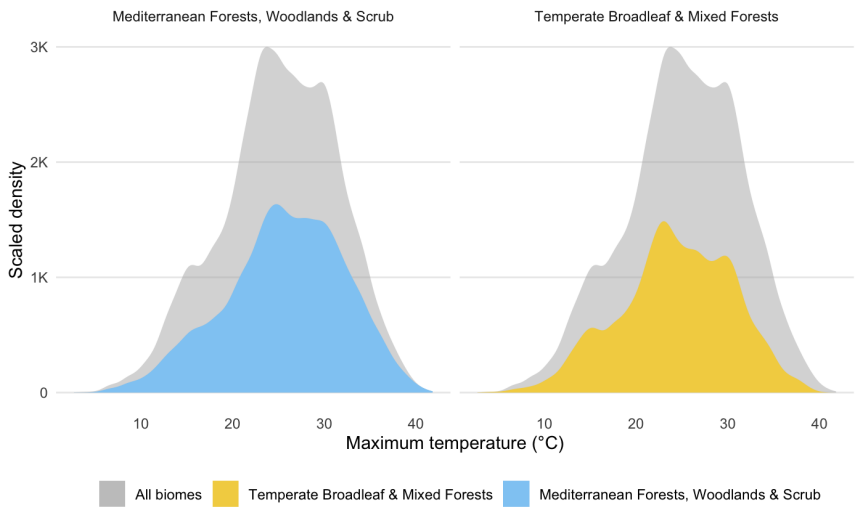


Figure C.11: Density plots by biome against overall distribution of maximum temperature

In contrast to the average air temperature, the distribution of the maximum air temperature by biome is strikingly similar.

C.3.4 Meteorological data: average air temperature

The statistical summary of the average air temperature is:

Min.	Q_1	Median	Mean	Q_3	Max.
-2.00	15.60	19.60	18.78	22.60	34.00

Table C.15: Statistical summary of average temperature variable

With its standard deviation being 5.47.
The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	-0.60	14.90	18.90	18.10	21.90	32.20
Mediterranean Forests, Woodlands & Scrub	-2.00	16.20	20.30	19.34	23.10	34.00

Table C.16: Statistical summary of average temperature variable by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	27.84
Mediterranean Forests, Woodlands & Scrub	30.87

Table C.17: Standard deviation of average temperature variable by biome

And, graphically the overall density plot and by biome:

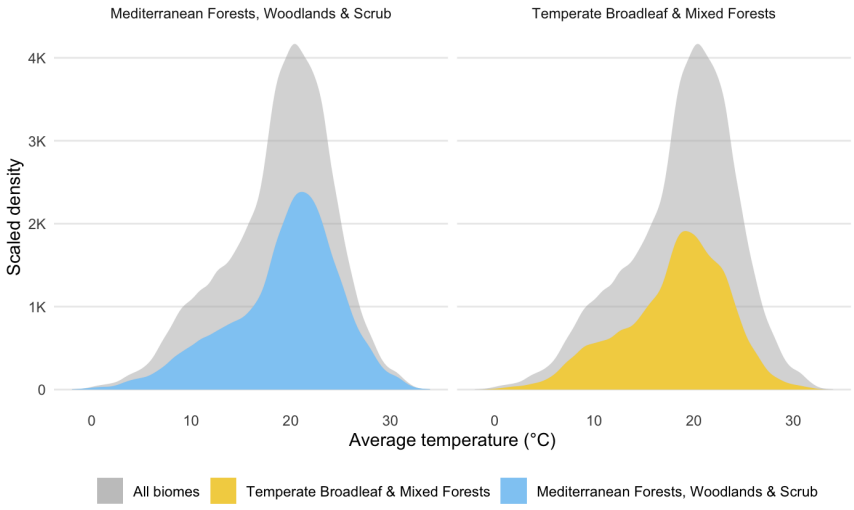


Figure C.12: Density plots by biome against overall distribution of average temperature

The distribution of the average temperature by biome is similar con the exception of smaller values in the case of the "Temperate Broadleaf & Mixed Forests" biome that is situated in the colder half of the territory.

C.3.5 Meteorological data: mean daily wind speed at 10ms

The wind speed variable in this data set represents the mean daily wind speed at 10 meters in meters/second. Its statistical summary is:

Min.	Q_1	Median	Mean	Q_3	Max.
0.00	1.80	2.50	2.74	3.40	10.00

Table C.18: Statistical summary of wind speed

With its standard deviation being 1.26.
The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	0.00	1.80	2.30	2.51	3.00	10.10
Mediterranean Forests, Woodlands & Scrub	0.20	1.90	2.70	2.93	3.70	10.80

Table C.19: Statistical summary of wind speed by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	1.06
Mediterranean Forests, Woodlands & Scrub	1.37

Table C.20: Standard deviation of wind speed by biome

And, graphically the overall density plot and by biome:

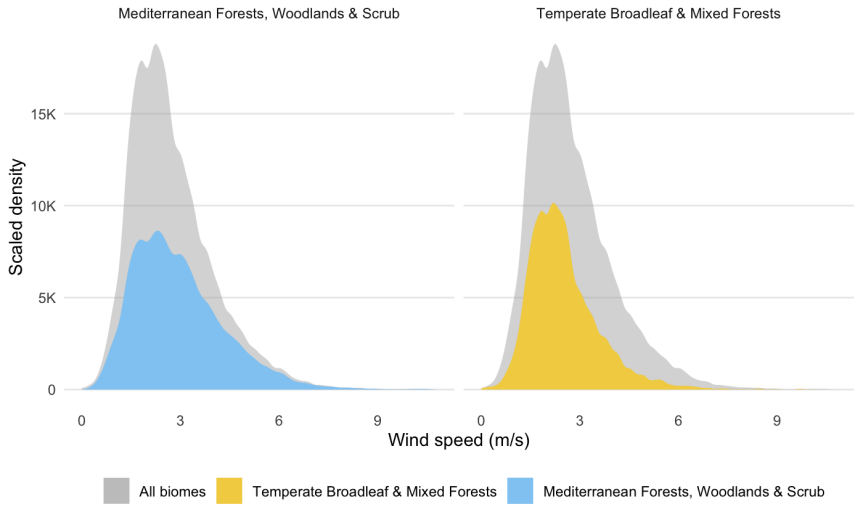


Figure C.13: Density plots by biome against overall distribution of wind speed

The data distribution for each biome is similar with consistently smaller values in the case of the Temperate Broadleaf & Mixed Forests biome.

C.3.6 Meteorological data: vapour pressure

Vapour pressure is one way of measuring the humidity of the air. It is supplied to the atmosphere by evaporation of water from oceans, lakes, wet

land surfaces or from vegetation (transpiration). Its statistical summary is:

Min.	Q_1	Median	Mean	Q_3	Max.
2.51	9.45	12.63	12.32	15.30	23.76

Table C.21: Statistical summary of vapour pressure

With its standard deviation being 3.85.
The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	2.70	9.09	12.34	12.05	15.07	22.34
Mediterranean Forests, Woodlands & Scrub	2.51	9.74	12.87	12.54	15.50	23.76

Table C.22: Statistical summary of vapour pressure by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	3.75
Mediterranean Forests, Woodlands & Scrub	3.91

Table C.23: Standard deviation of vapour pressure by biome

And, graphically the overall density plot and by biome:

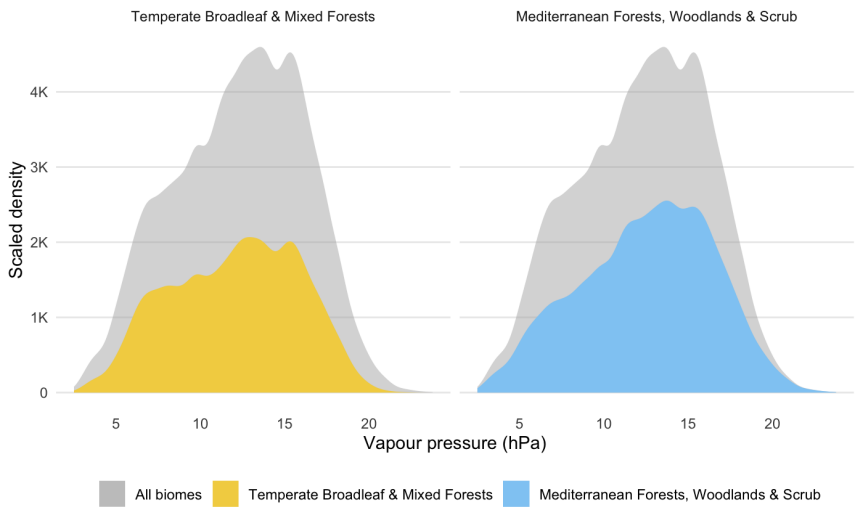


Figure C.14: Density plots by biome against overall distribution of vapour pressure

The distribution of the variable by biome is similar with the Temperate Broadleaf & Mixed Forests experiencing evenly distributed values across its whole range but with a peak in the upper half part of the distribution that ultimately shapes the distribution.

C.3.7 Meteorological data: sum of precipitation

The statistical summary of the precipitation feature, measured in millimetres, is:

Min.	Q_1	Median	Mean	Q_3	Max.
0.00	0.00	0.00	0.41	0.00	113.00

Table C.24: Statistical summary of precipitation

With its standard deviation being 3.28.
The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	0.00	0.00	0.00	0.39	0.00	113.00
Mediterranean Forests, Woodlands & Scrub	0.00	0.00	0.00	0.41	0.00	81.00

Table C.25: Statistical summary of precipitation by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	3.49
Mediterranean Forests, Woodlands & Scrub	3.11

Table C.26: Standard devitation of precipitation by biome

It evident from the data that the distribution is highly skewed.

C.3.8 Meteorological data: total global radiation

The statistical summary of the total global radiation measured in KJ/m^2/day is:

Min.	Q_1	Median	Mean	Q_3	Max.
1,500	16,525	21,800	21,480	26,823	35,761

Table C.27: Statistical summary of radiation

With its standard deviation being 6,461.29.
The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	2,400	16,200	21,163	20,904	25,800	35,511
Mediterranean Forests, Woodlands & Scrub	1,500	16,831	22,500	21,820.55	27,503	35,761

Table C.28: Statistical summary of radiation by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	6,253.34
Mediterranean Forests, Woodlands & Scrub	6,598.40

Table C.29: Standard deviation of radiation by biome

And, graphically the overall density plot and by biome:

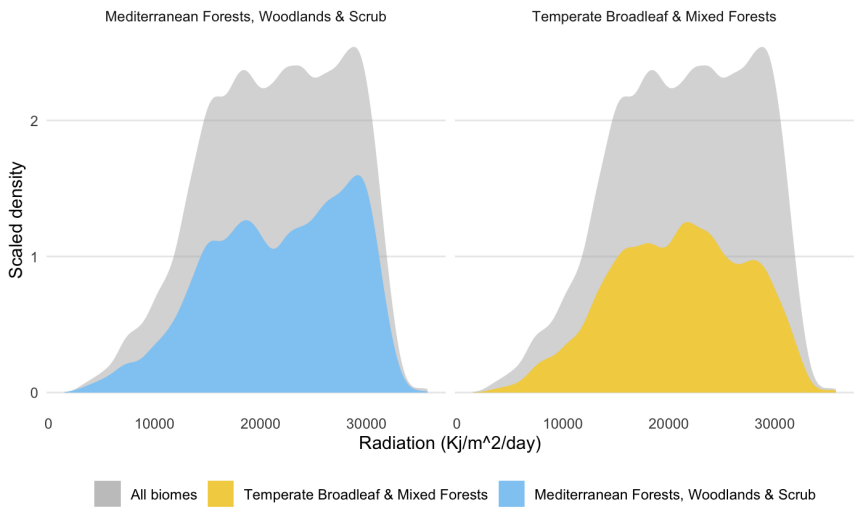


Figure C.15: Density plots by biome against overall distribution of radiation

The distribution is similar, but there are two distinctive peaks in the "Mediterranean Forests, Woodlands & Scrub" biome one whereas the Temperate Broadleaf & Mixed Forests biome distribution shows more uniform, albeit lower, values.

C.3.9 Canadian Forest Fire Weather Index (FWI): FFMC

The statistical summary of FFMC that is unitless is:

Min.	Q_1	Median	Mean	Q_3	Max.
6.09	86.37	89.53	88.43	92.07	98.47

Table C.30: Statistical summary of FFMC

With its standard deviation being 6.11.
The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	18.72	86.22	89.07	88.00	91.43	96.46
Mediterranean Forests, Woodlands & Scrub	6.09	86.54	89.97	88.78	92.55	98.47

Table C.31: Statistical summary of FFMC by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	5.85
Mediterranean Forests, Woodlands & Scrub	6.30

Table C.32: Standard deviation of FFMC by biome

And, graphically the overall density plot and by biome:

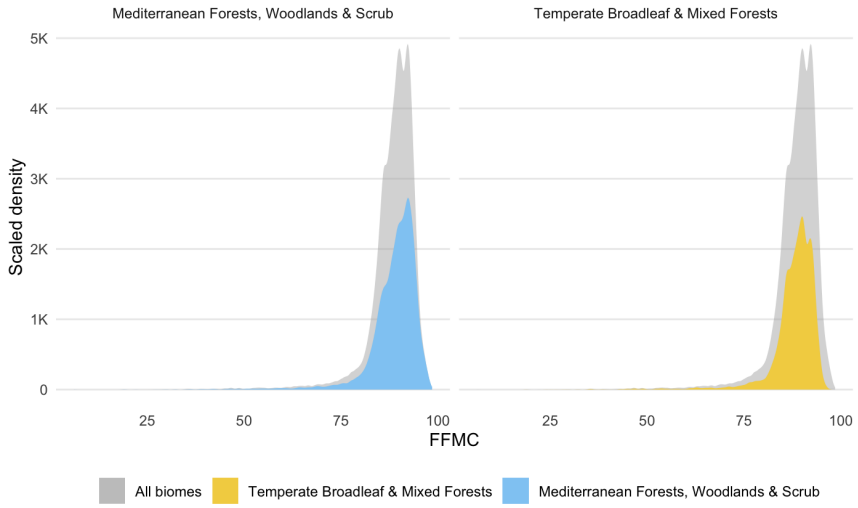


Figure C.16: Density plots by biome against overall distribution of FFMC

The distribution for each biome is similar with high values for the HCFs, but also a high tail towards zero.

C.3.10 Canadian Forest Fire Weather Index (FWI): DMC

The statistical summary of DMC that is unitless is:

Min.	Q_1	Median	Mean	Q_3	Max.
0.07	28.03	76.01	87.93	123.31	702.01

Table C.33: Statistical summary of DMC

With its standard deviation being 74.57.
The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	0.20	24.04	58.71	69.06	103.66	347.43
Mediterranean Forests, Woodlands & Scrub	0.07	34.32	91.17	103.40	142.77	702.01

Table C.34: Statistical summary of DMC by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	52.32
Mediterranean Forests, Woodlands & Scrub	85.69

Table C.35: Standard deviation of DMC by biome

And, graphically the overall density plot and by biome:

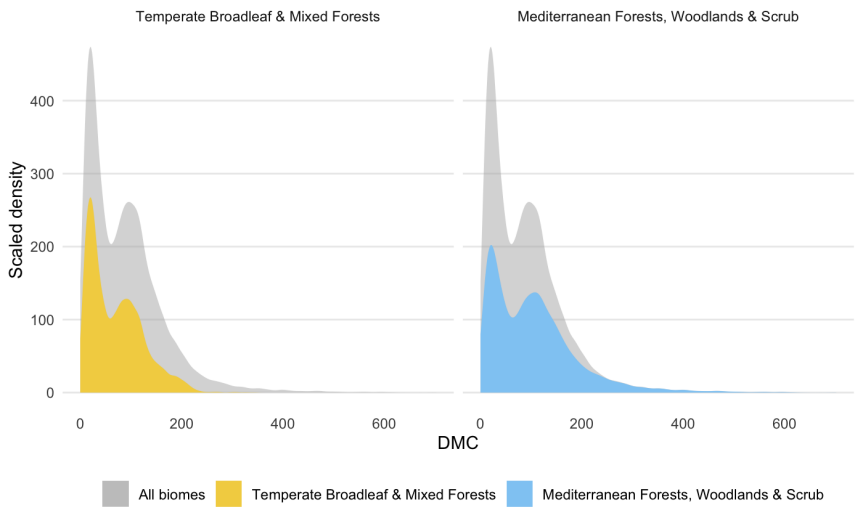


Figure C.17: Density plots by biome against overall distribution of DMC

Even though the distribution in each biome is similar, in the Temperate Broadleaf & Mixed Forests biome the values concentrate in the lower end.

Whereas, in the Mediterranean Forests, Woodlands & Scrub biome there is a significant larger variability with extreme values. Moreover, its maximum value almost doubles the one in the other biome.

C.3.11 Canadian Forest Fire Weather Index (FWI): DC

DC, which is unitless, represents the moisture content of deep (approximately 10–20 cm deep), compact organic layers with a lag of 52 days. Its statistical summary is:

Min.	Q_1	Median	Mean	Q_3	Max.
0.44	48.87	205.04	202.52	319.15	760.15

Table C.36: Statistical summary of DC

With its standard deviation being 147.71.
The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	0.44	33.69	159.00	165.57	281.38	513.48
Mediterranean Forests, Woodlands & Scrub	0.56	83.08	241.53	232.82	353.76	760.15

Table C.37: Statistical summary of DC by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	130.95
Mediterranean Forests, Woodlands & Scrub	153.65

Table C.38: Standard deviation of DC by biome

And, graphically the overall density plot and by biome:

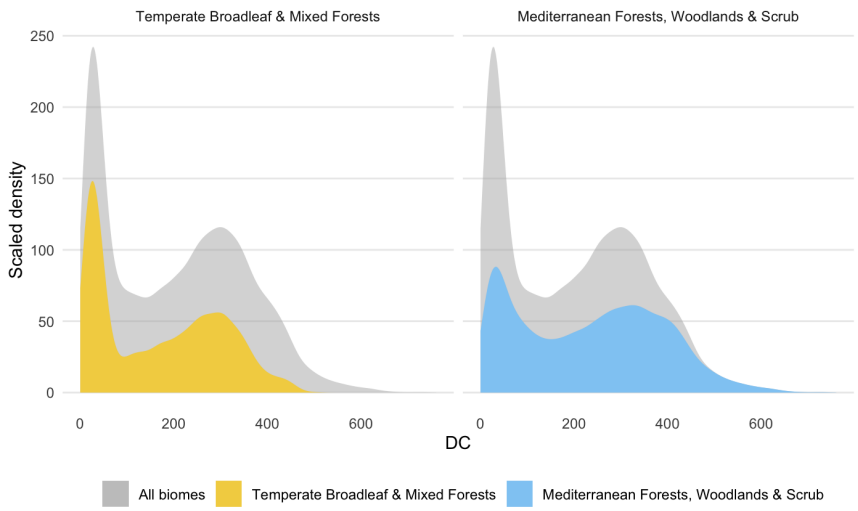


Figure C.18: Density plots by biome against overall distribution of DC

The shape of both distributions is similar. Each one seems like the combination of other two distributions, one closer to zero and another in the other side of their range. This bimodality indicates a divide between HCFs driven by long draughts indicated by high values of DC and other HCFs driven by another causes.

C.3.12 Canadian Forest Fire Weather Index (FWI): ISI

The statistical summary of ISI that is unitless is:

Min.	Q_1	Median	Mean	Q_3	Max.
0.00	4.93	7.28	7.72	9.87	34.46

Table C.39: Statistical summary of ISI

With its standard deviation being 3.92.
The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	0.00	4.61	6.58	6.85	8.78	24.68
Mediterranean Forests, Woodlands & Scrub	0.00	5.30	7.99	8.44	10.88	34.46

Table C.40: Statistical summary of ISI by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	3.14
Mediterranean Forests, Woodlands & Scrub	4.33

Table C.41: Standard deviation of ISI by biome

And, graphically the overall density plot and by biome:

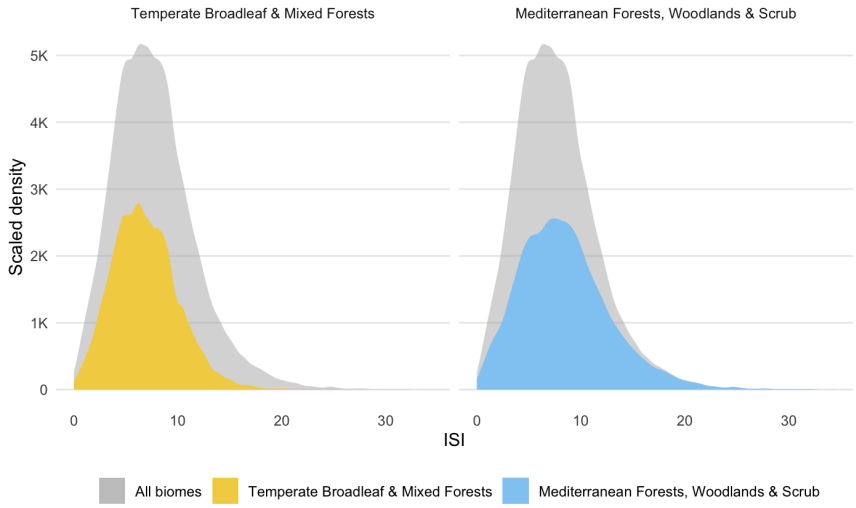


Figure C.19: Density plots by biome against overall distribution of ISI

The distributions are fairly similar with lower values on average in the case of the "Temperate Broadleaf & Mixed Forests" biome.

C.3.13 Canadian Forest Fire Weather Index (FWI): BUI

BUI, which is unitless, indicates the total amount of fuel available for combustion by a moving flame front. Although BUI is a weighted combination of the DMC and DC indicators, with the former having a bigger influence on the final value. For example, a DMC value of zero always results in a BUI value of zero regardless of what the DC value is. But, the influence of DC increases with the value of DMC.

Its statistical summary is:

Min.	Q_1	Median	Mean	Q_3	Max.
0.11	29.93	84.44	92.04	130.20	685.90

Table C.42: Statistical summary of BUI

With its standard deviation being 74.25.

The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	0.24	24.56	65.05	72.34	109.72	344.37
Mediterranean Forests, Woodlands & Scrub	0.11	38.46	100.16	108.18	149.13	685.90

Table C.43: Statistical summary of BUI by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	53.29
Mediterranean Forests, Woodlands & Scrub	84.42

Table C.44: Standard deviation of BUI by biome

And, graphically the overall density plot and by biome:

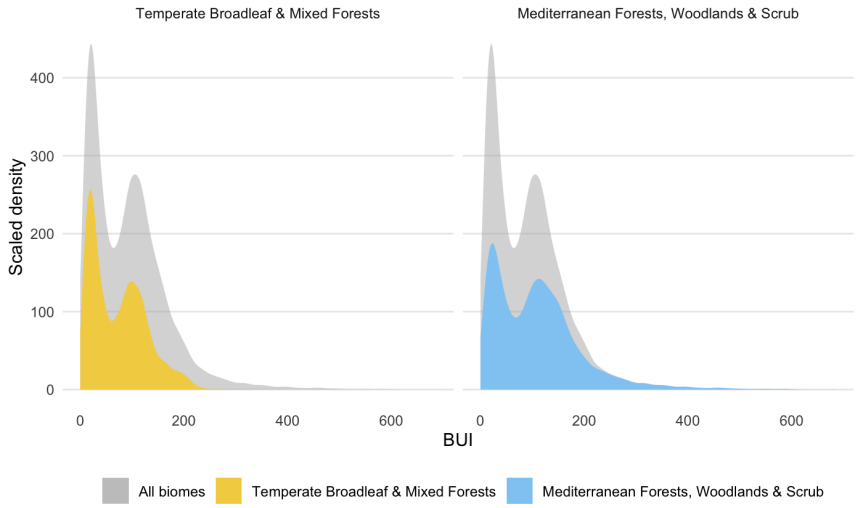


Figure C.20: Density plots by biome against overall distribution of BUI

The shape of the distributions makes evident that BUI is a combination of the DMC and DC indicators.

C.3.14 Canadian Forest Fire Weather Index (FWI): FWI

The statistical summary of FWI that is unitless is:

Min.	Q_1	Median	Mean	Q_3	Max.
0.00	10.13	22.41	22.62	32.95	84.78

Table C.45: Statistical summary of FWI

With its standard deviation being 14.00.
The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	0.00	8.39	18.51	19.13	28.61	64.06
Mediterranean Forests, Woodlands & Scrub	0.00	12.41	26.08	25.48	36.22	84.78

Table C.46: Statistical summary of FWI by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	11.95
Mediterranean Forests, Woodlands & Scrub	14.88

Table C.47: Standard deviation of FWI by biome

And, graphically the overall density plot and by biome:

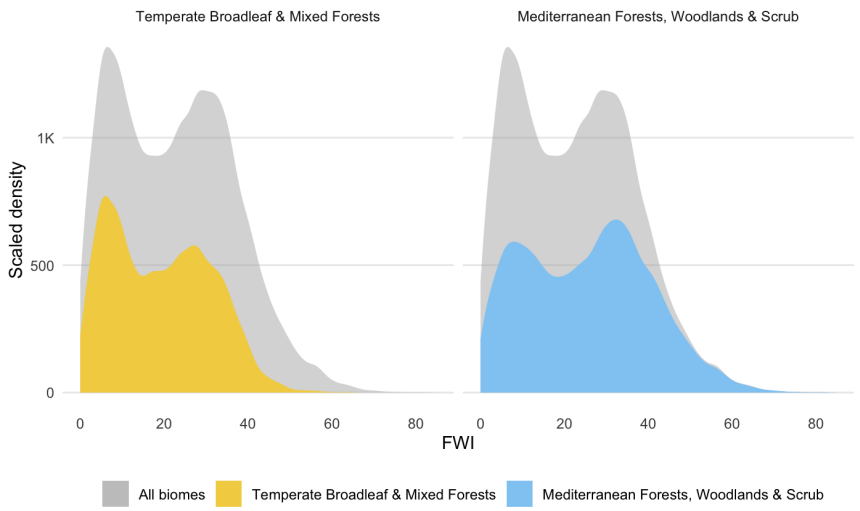


Figure C.21: Density plots by biome against overall distribution of FWI

The FWI is a combination of ISI and BUI representing the fire intensity by combining the rate of fire spread with the amount of fuel being consumed. Therefore, it shows patterns seen on other Forest Fire Weather Indicators:

C.3.15 Physiography: elevation

The elevation of continental Portugal goes from the lowest point 0 at the coast to e point of highest elevation of continental Portugal is Torre located in Serra da Estrela with 1,993 meters above sea level. And, the mean elevation of Portugal is 372 meters.

The statistical summary of the elevation above sea level in meters is:

Min.	Q_1	Median	Mean	Q_3	Max.
-2.50	120.40	286.20	344.90	528.00	1,835.37

Table C.48: Statistical summary of elevation

With its standard deviation being 265.22.

The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	-1.50	157.73	294.11	350.23	487.07	1,517.43
Mediterranean Forests, Woodlands & Scrub	-2.50	91.76	273.37	340.54	560.50	1,835.37

Table C.49: Statistical summary of elevation by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	256.66
Mediterranean Forests, Woodlands & Scrub	271.97

Table C.50: Standard devaition of elevation by biome

There are 41 HCFs with negative elevation. They all located along the coast and are plausible but others are located in the sea or estuaries of rivers:

(a) All (b) Detail

Figure C.22: HCFs with elevation less than zero

Hence, I going to remove the fires with negative elevation when I prepare the data for modelling.

There is also 56 HCFs with zero elevation, but these HCFs in the coast are plausible so I am going to do nothing.

Continuing, the density plot by biome against the overall distribution is:

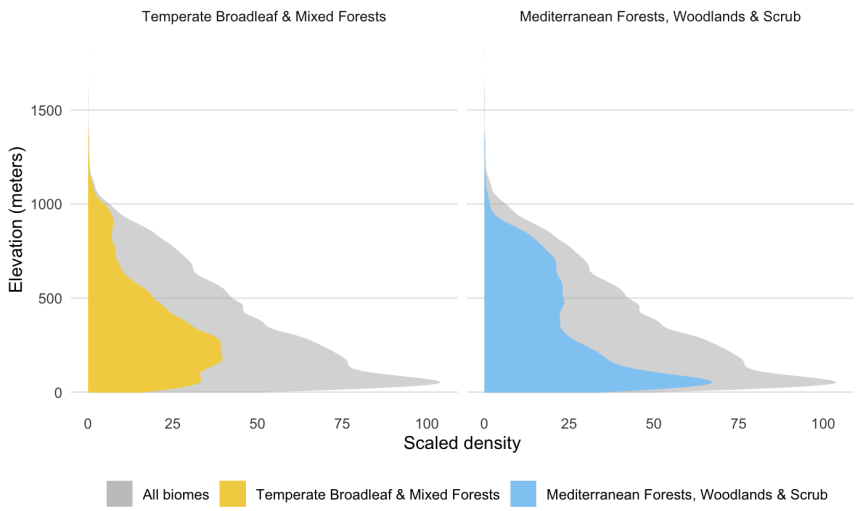


Figure C.23: Density plots by biome against overall distribution of elevation

The "Mediterranean Forests, Woodlands & Scrub" biome cover a more diverse landscape and this is reflected in the density plot with a large number of fires occurring at low elevations and a smaller number at higher elevation.

However, in the "Temperate Broadleaf & Mixed Forests" biome the fire occurrence is concentrated in lower elevations, but not so low than in the other biome.

C.3.16 Physiography: slope

The statistical summary of the slope measured in degrees from the horizontal:

Min.	Q_1	Median	Mean	Q_3	Max.
0.00	0.00	5.13	6.18	8.89	70.12

Table C.51: Statistical summary of slope

With its standard deviation being 5.92.

The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	0.00	5.13	7.25	7.62	11.48	37.25
Mediterranean Forests, Woodlands & Scrub	0.00	0.00	5.13	5.01	7.25	70.12

Table C.52: Statistical summary of slope by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	5.95
Mediterranean Forests, Woodlands & Scrub	5.64

Table C.53: Standard deviation of slope by biome

And, graphically the overall density plot and by biome:

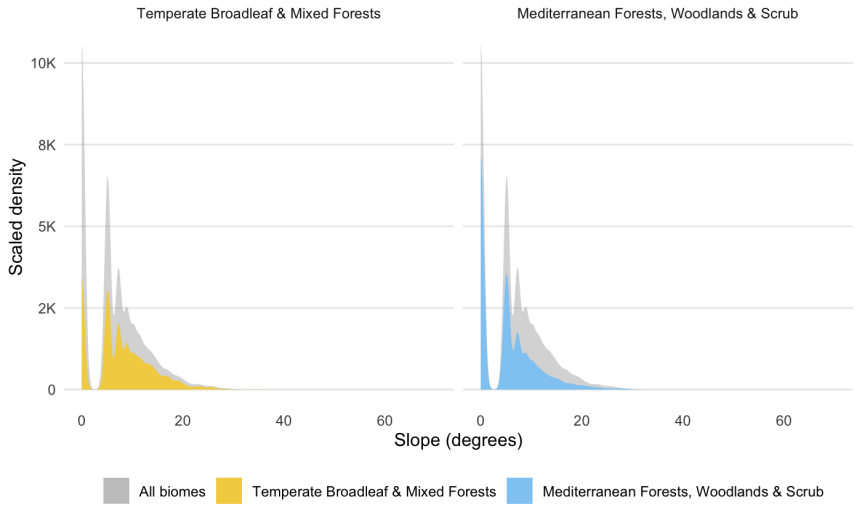


Figure C.24: Density plots by biome against overall distribution of slope

There is a gap in the density plot in the overall distribution and in each one for the biomes. Its origin is on the original data: how the variable is measured in the first place; and augmented by the conversion to decimal degrees. It is easier to appreciate by analysing its spread using box plots.

Also, HCFs in the "Mediterranean Forests, Woodlands & Scrub" occur on gentler slopes.

C.3.17 Human factors: distance to nearest road

The statistical summary of the distance to the nearest main road is:

Min.	Q_1	Median	Mean	Q_3	Max.
0.01	70.06	209.54	410.13	506.77	6,826.63

Table C.54: Statistical summary of distance to nearest road by biome

With its standard deviation being 570.15.

The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	0.02	70.70	192.76	347.38	436.13	5,023.38
Mediterranean Forests, Woodlands & Scrub	0.01	74.18	225.12	461.56	579.51	6,862.63

Table C.55: Statistical summary of distance to nearest road by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	447.68
Mediterranean Forests, Woodlands & Scrub	649.16

Table C.56: Standard deviation of distance to nearest road by biome

And, graphically the overall density plot and by biome:

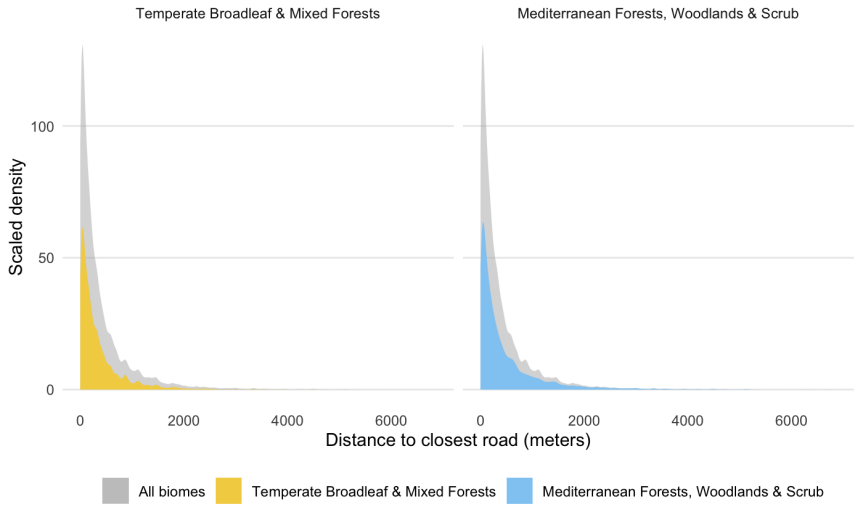


Figure C.25: Density plots by biome against overall distribution of distance to nearest road

Most HCFs happen close to a road, that allow easy access.

C.3.18 Human factors: distance to nearest building

The statistical summary of the distance to the nearest build-up area is:

Min.	Q_1	Median	Mean	Q_3	Max.
0.00	148.40	545.70	896.10	1,288.50	7,923.68

Table C.57: Statistical summary of distance to nearest building

With its standard deviation being 1,016.79.
The same summary but by biome is:

Biome	Min.	Q_1	Median	Mean	Q_3	Max.
Temperate Broadleaf & Mixed Forests	0.00	131.19	469.28	748.96	1,080.42	7,889.37
Mediterranean Forests, Woodlands & Scrub	0.00	163.34	634.18	1,016.66	1,501.96	7,923.68

Table C.58: Statistical summary of distance to nearest building by biome

With the standard deviation by biome is:

Biome	Standard deviation
Temperate Broadleaf & Mixed Forests	859.41
Mediterranean Forests, Woodlands & Scrub	1,115.21

Table C.59: Standard deviation of distance to nearest building by biome

And, graphically the overall density plot and by biome:

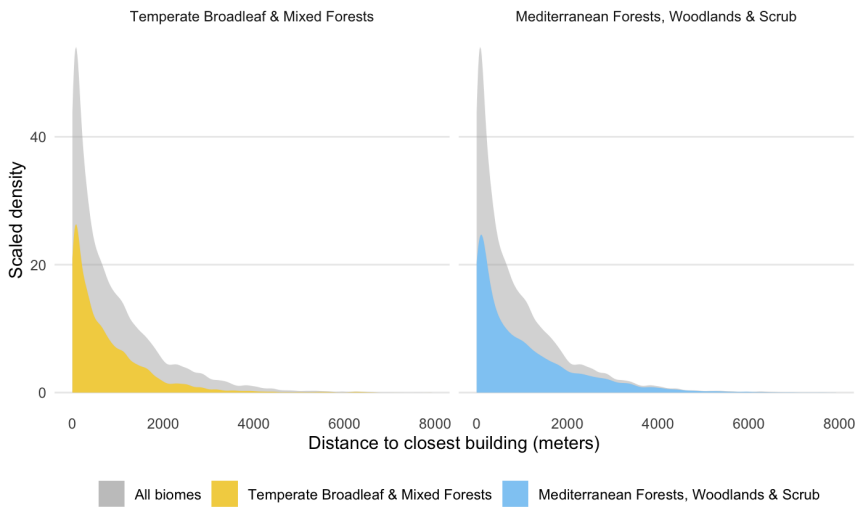


Figure C.26: Density plots by biome against overall distribution of distance to nearest building

Most HCFs happen close to a build-up areas where humans, who are the direct or indirect cause of HCFs, live.

C.4 Shape of variable's distribution by biome and temporal unit

To explore this aspect of the data I am going to use stacked density plots of a variable grouped by year and for each year overlapping density plots by biome

C.4.1 Meteorological data: vapour pressure

Looking at the distribution of the vapour pressure variable by biome and year shows a marked disparity between years and biomes. Low values indicate drier conditions that affect the content of moisture in the soil. Whereas high values are characteristic of the heat waves occurring in the summer:

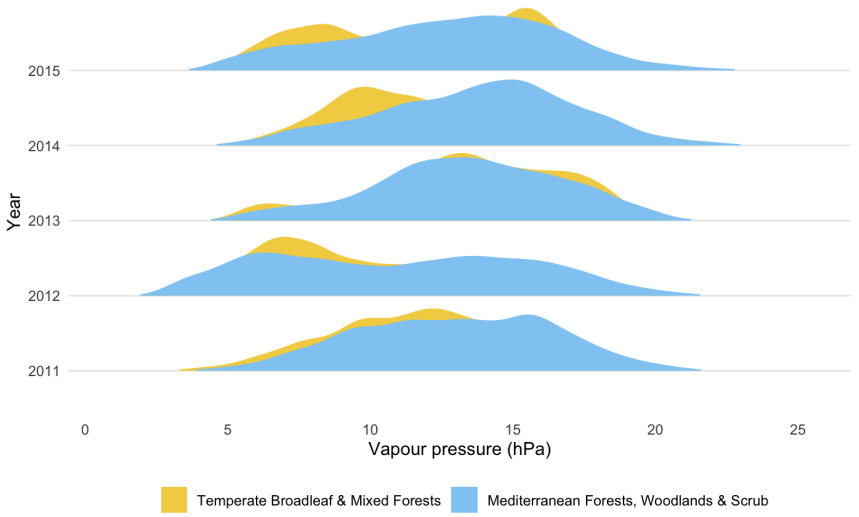


Figure C.27: Density plots by biome and year of vapour pressure

This is different from what we can see when visualising the ungrouped distribution of the variable.

C.4.2 Meteorological data: total global radiation

The distribution of the radiation grouped by year and biome shows a large variability in the range of values and especially between biomes:

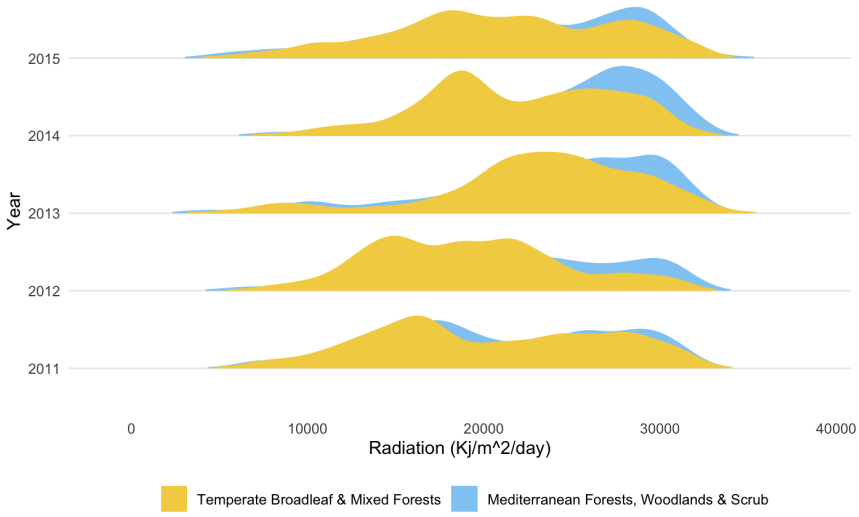


Figure C.28: Density plots by biome and year of radiation

Solar radiation has a strong impact on creating the conditions suitable for the occurrence of fire. However, without considering the slope of the terrain

and the average of radiation for a past period (e.g., monthly average) prior to the fire instead of a daily measurement does not appear to have much significance in the prediction of HCFs.

C.4.3 Physiography: elevation

The distribution of the elevation grouped by year and biome shows some variability in the case of the "Temperate Broadleaf & Mixed Forests" biome, but nothing that merits further investigation:

Figure C.29: Density plots by biome and year of elevation

C.5 Distribution of skewed variables

To explore this aspect of the data I am going to use empirical cumulative distribution function (ECDF) plots to ascertain whether a skewed variable follows a log-normal or power distribution so I know what transformation to apply to a variable.

C.5.1 Human-caused fires (HCFs): burnt area

It is evident from the statistics summary of the variable that the distribution is highly skewed. The empirical cumulative distribution function (ECDF) shows an almost vertical rise at 0:

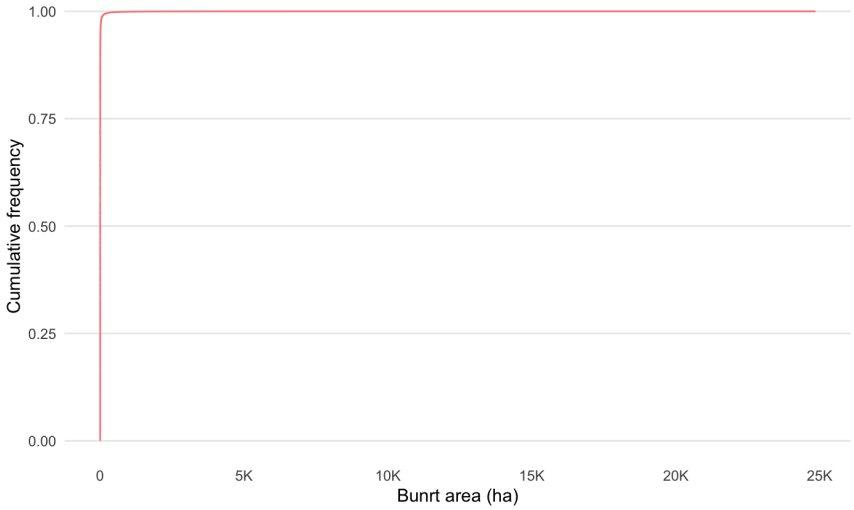


Figure C.30: Burnt area ECDF

Now the question is if it follows a log-normal or power distribution. First, I am going to apply a $log\,(area)$ transformation. The density plot is:

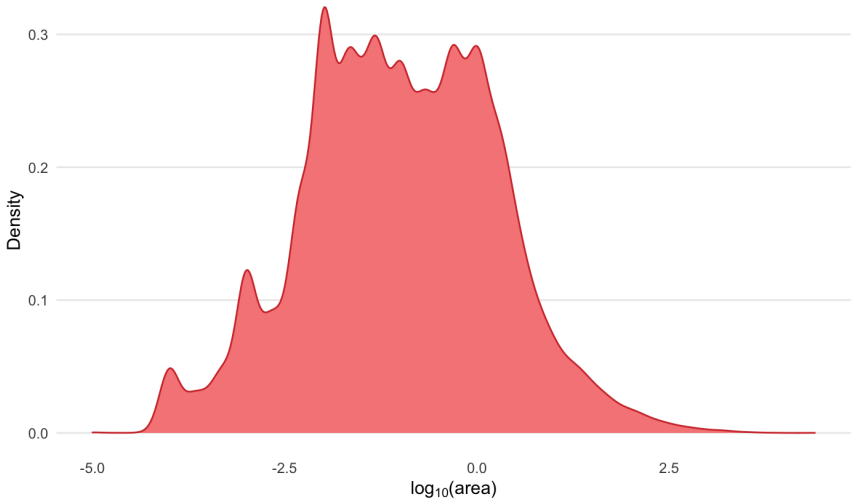


Figure C.31: Density plot of transformed burnt area

And, the empirical cumulative distribution of the transformed variable is:

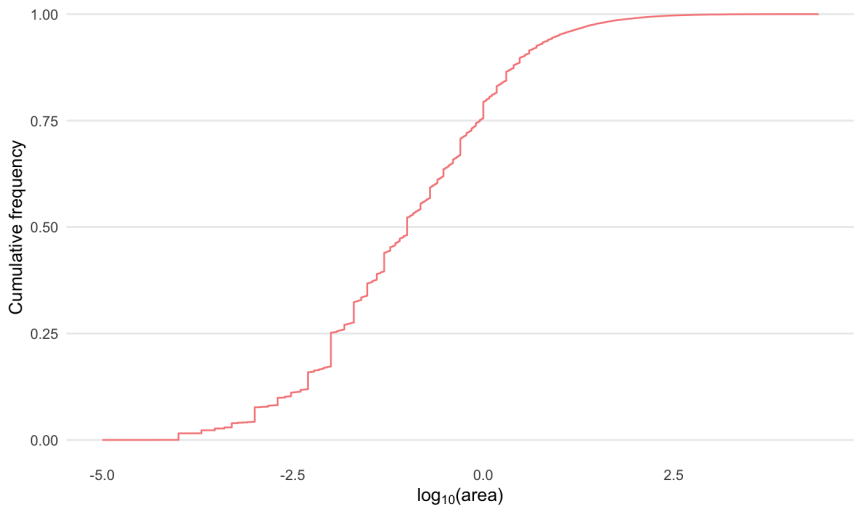


Figure C.32: ECDF of transformed burnt area

To check whether this distribution is not a power law, I plot it as a descending ECDF with logarithmic x and y axes. In this visualization, a power law appears as a perfect straight line. And, this is not the case:

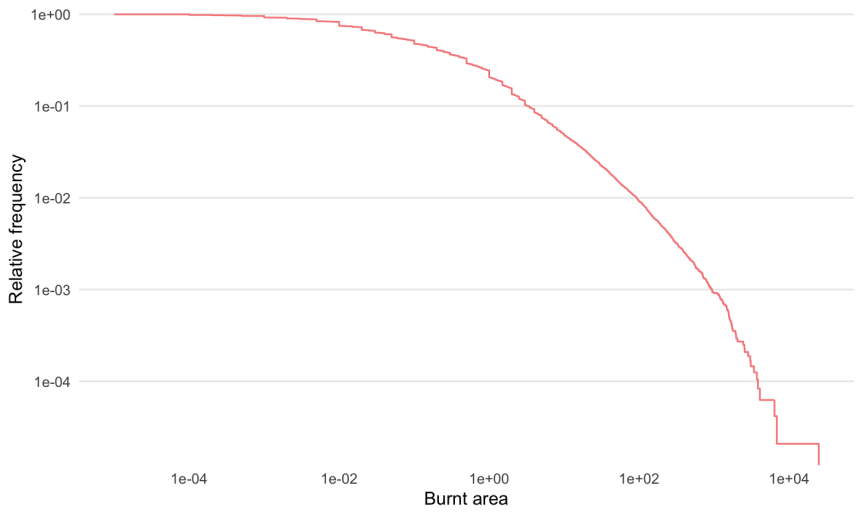


Figure C.33: Burnt area descending ECDF using logarithmic scale

C.5.2 Meteorological data: sum of precipitation

As is evident from the statistical summary of the variable, the distribution is highly skewed containing a large amount of days with human-caused fires ()HCFs) happening but with no (or almost) rain.

Portugal is amongst the sunniest areas in Europe. The annual precipitation varies from amounts to 1,450 mm in Braga and 1,100 millimetres in Porto, while it drops to around 700 mm in Lisbon, and to about 500 mm (20 in) in Algarve.

The empirical cumulative distribution function (ECDF) shows a step rise at 0:

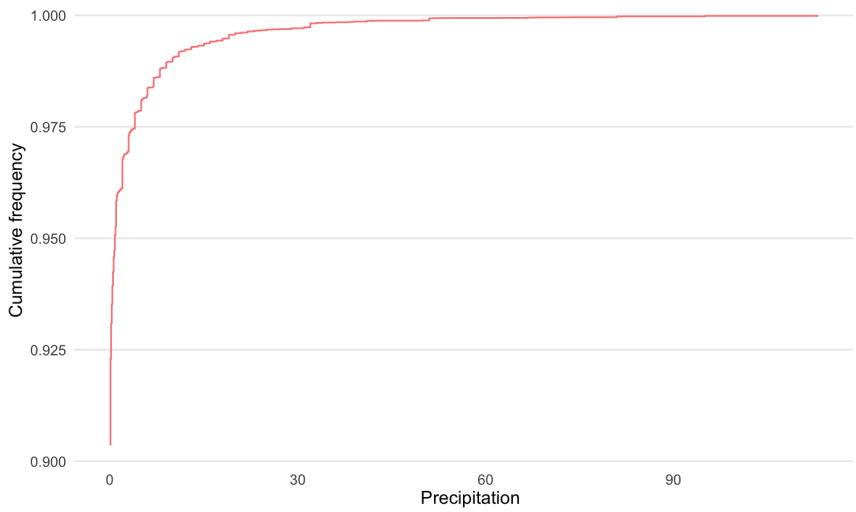


Figure C.34: ECDF of precipitation

Now the question is if it follows a log-normal or power distribution. First, I am going to apply a *log*(precipitation) transformation.

The density plot is fo the transformed variable is:

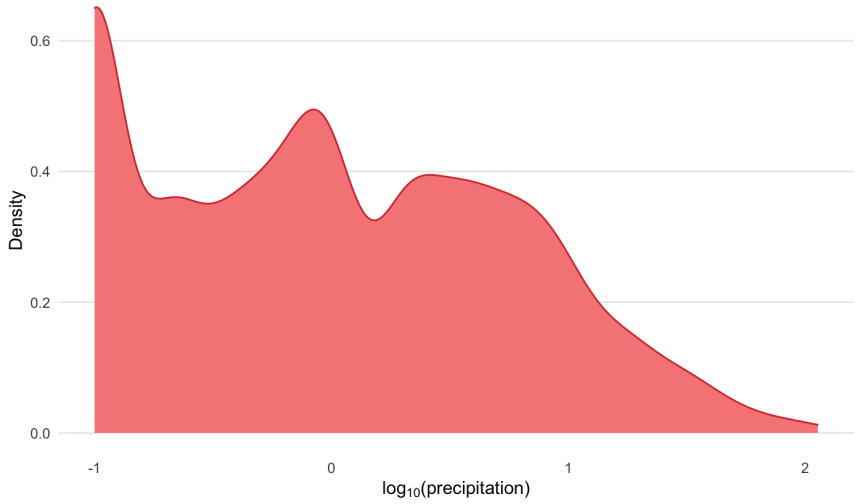


Figure C.35: Density plot of transformed precipitation

And, the empirical cumulative distribution of the transformed variable is:

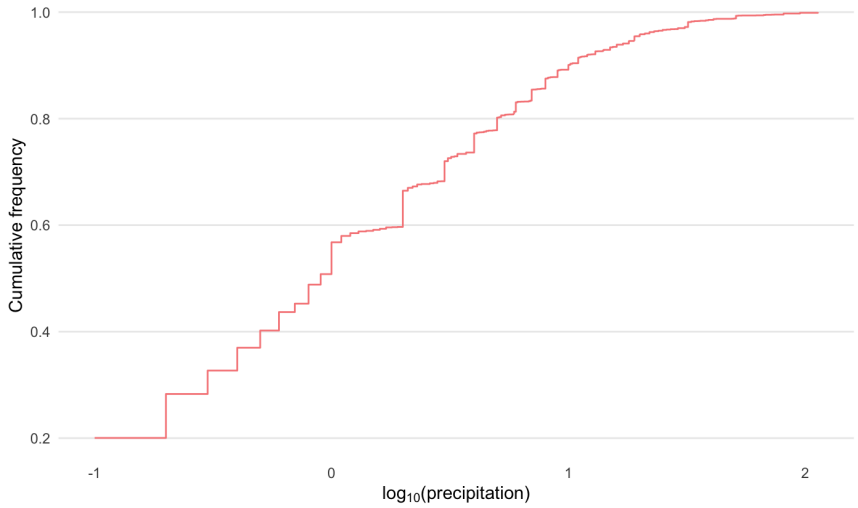


Figure C.36: ECDF of transformed precipitation

To check whether this distribution is not a power law, I plot it as a descending ECDF with logarithmic x and y axes. In this visualization, a power law appears as a perfect straight line. And, this is not the case:

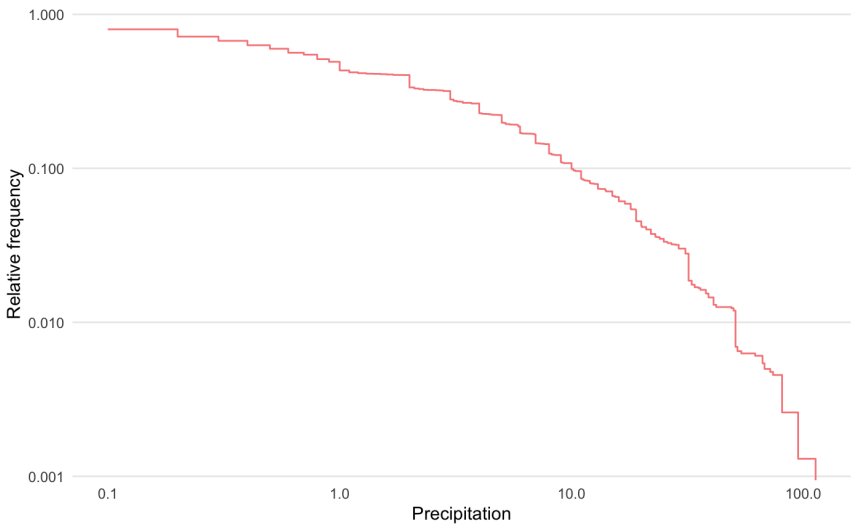


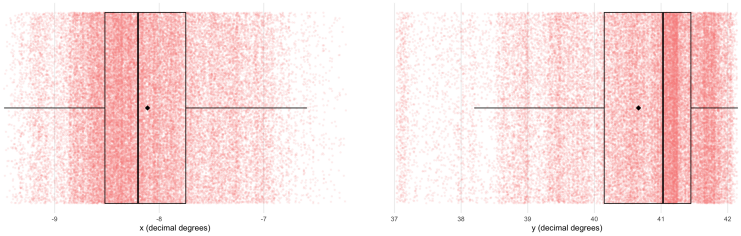
Figure C.37: Precipitation descending ECDF using logarithmic scale

C.6 Spread and outliers of a variable

To explore this aspect of the data I am going to use box-and-whisker plots to check the spread and symmetry of a variable distribution.

C.6.1 Human-caused fires (HCFs): coordinates (x and y)

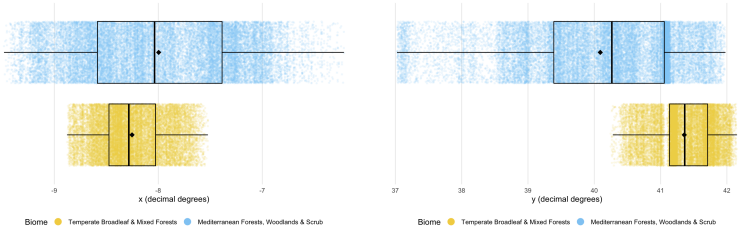
The difference in the geospatial location and area covered by each biome shows as differentes in the spread of each of the coordinate variables, x (i.e., longitude) and y (i.e., latitude):



(a) x (longitude) (b) y (latitude)

Figure C.38: Boxplot of x and y

When the box plot is grouped by biome the, the observations beyond the whiskers (that could be considered outliers) disappear:



(a) x (longitude) (b) y (latitude)

Figure C.39: Boxplot of x and y by biome

C.6.2 Meteorological data: maximum air temperature

The "Temperate Broadleaf & Mixed Forests" biome shows slightly smaller values on average as we can see in the boxplot:

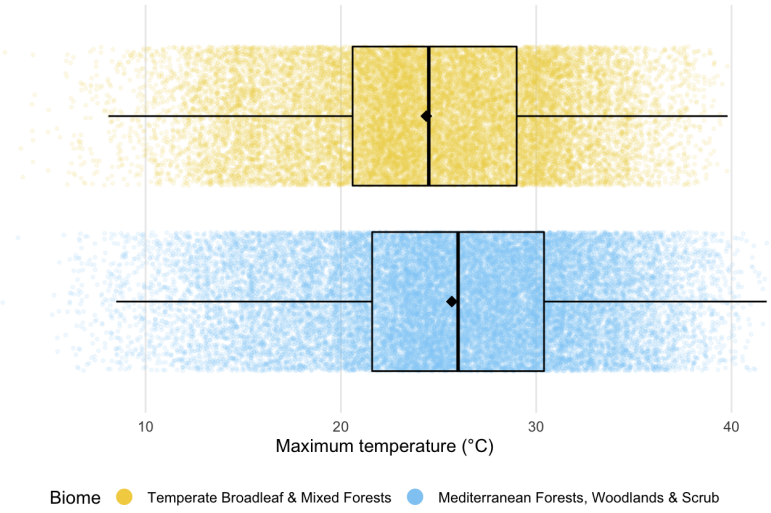


Figure C.40: Boxplot of maximum temperature by biome

Except that there is a lower number of outliers corresponding to low temperatures, the maximum temperatures grouped by year repeats the same patterns as the average temperatures, (low) outliers in each year and a greater range variability and more extreme values in 2012 than the rest of years:

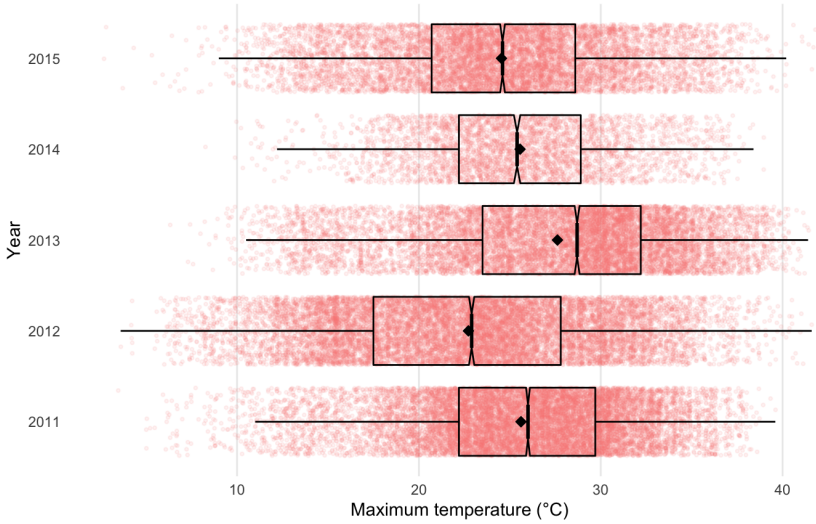


Figure C.41: Boxplot of maximum temperature by year

The similarity is also found when the data is grouped by month:

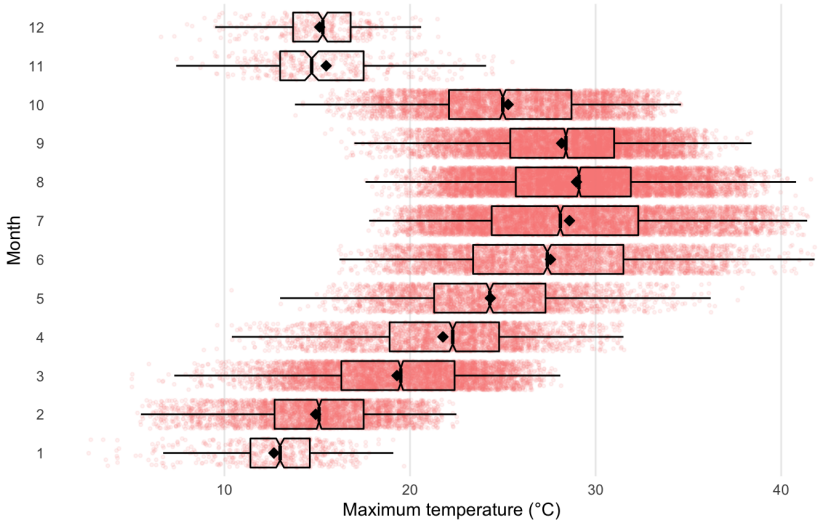


Figure C.42: Boxplot of maximum temperatures by month

The range of maximum temperatures across the year is wide. Moreover, the values in are consistently high temperatures in the summer.

The ocean exerts a great influence in the weather of continental Portugal. It shelters the land from cold winds and night frosts which are exceedingly rare and never intense. Summer lasts from June to mid-September, it is milder on the northern half on the country than the south. Also, the coastal areas are colder than the interior since they are exposed to ocean winds.

Hence, the hotter part of the territory is the interior south. While the north is colder and rainfall more frequent and abundant.

C.6.3 Meteorological data: average air temperature

The spread is similar with smaller values in the case of the "Temperate Broadleaf & Mixed Forests" biome situated in the colder half of the territory:

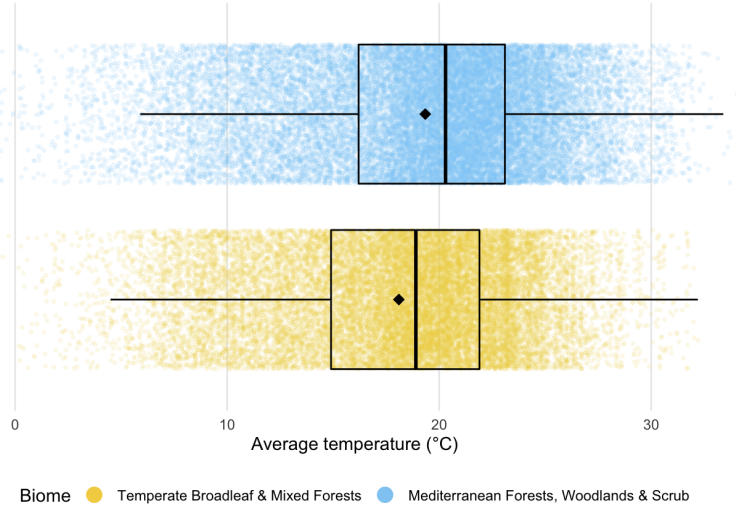


Figure C.43: Boxplot of average temperature by biome

There is a large number of outliers corresponding to low temperatures:

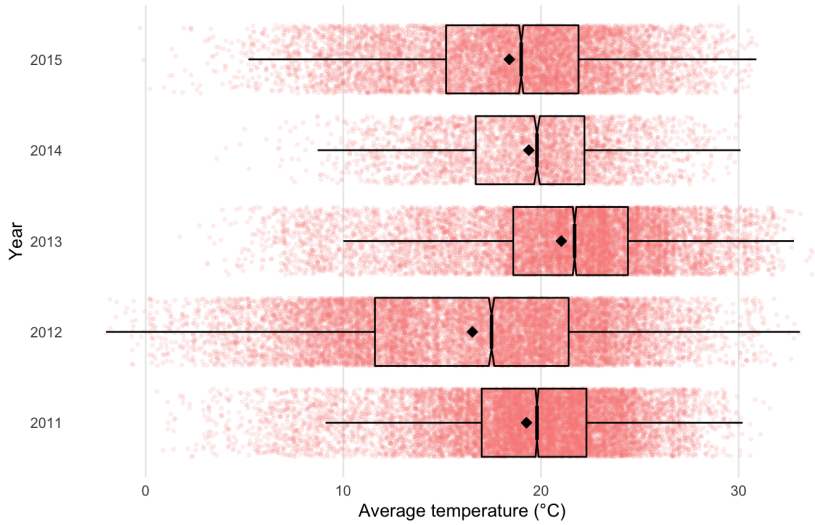


Figure C.44: Boxplot of average temperature by year

This pattern repeats all years except in 2012. However, grouping the data by month results in the range of average temperatures across the year is large with consistent high temperatures in its "extended" Summer that lasts from June to October:

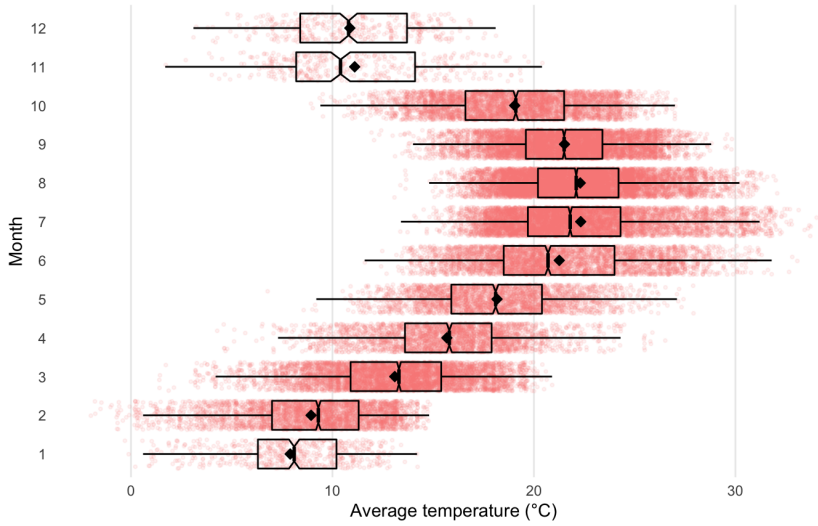


Figure C.45: Boxplot of average temperature by month

The annual average temperature in continental Portugal varies from 12–13 °C in the mountainous interior north to 17–18 °C in the south (in general the south is warmer and drier than the north).

Extreme temperatures occur in the mountains in the interior North and Centre of the country in winter, where they may fall below −10 °C, and in south-eastern parts in the summer, sometimes exceeding 45 °C.

The values are greater than expected but we are calculating the statistics from a small sample, so some deviation can be expected. But there is also the matter of a high variability and the long Summer which can be a consequence of climate change with hotter years.

C.6.4 Meteorological data: mean daily wind speed at 10ms

The data distribution for each biome is similar with consistently small values in the case of the Temperate Broadleaf & Mixed Forests biome:

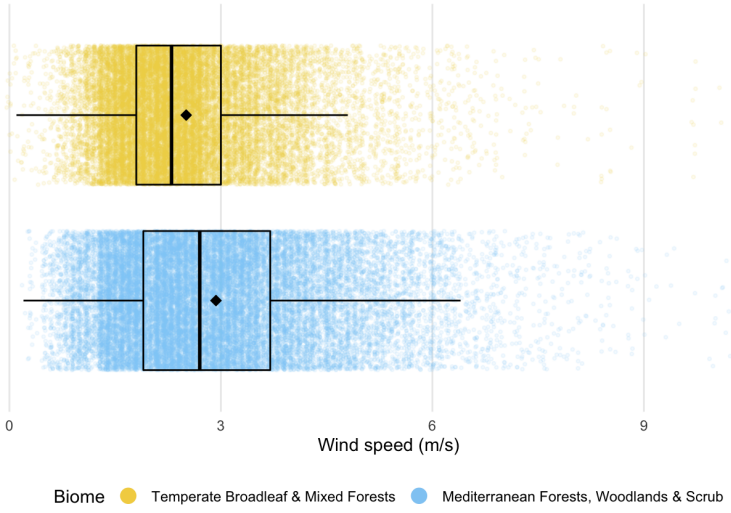


Figure C.46: Boxplot by biome of wind speed

The wind speed data is also consistent in all years of the period of study:

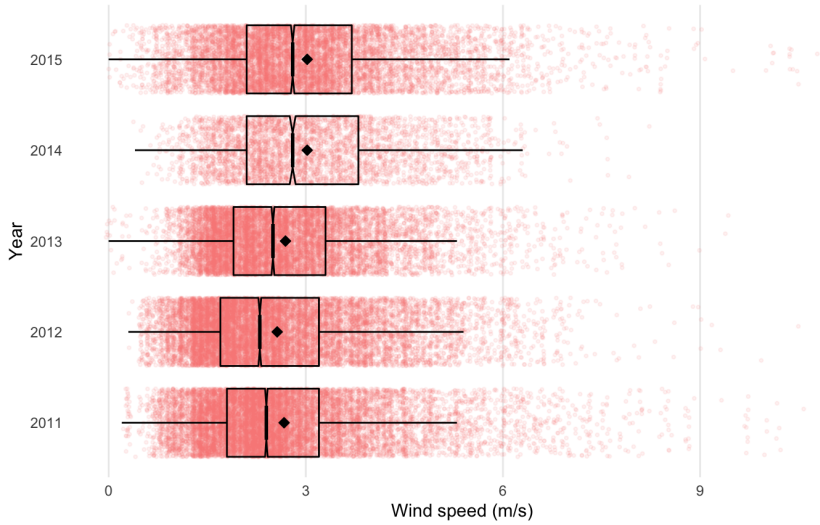


Figure C.47: Boxplot by year of wind speed

C.6.5 Meteorological data: vapour pressure

The distribution by biome is similar but the values of the Temperate Broadleaf & Mixed Forests more concentrated.

Looking at the distribution grouped by year, the values of the year 2014 stands out):

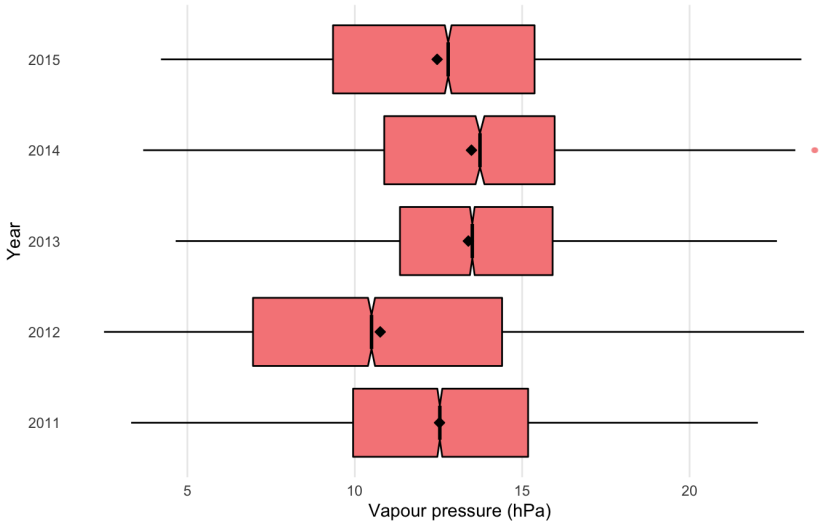


Figure C.48: Boxplot by year of vapour pressure

Although, the data grouped by biome shows that the distribution is similar between them:

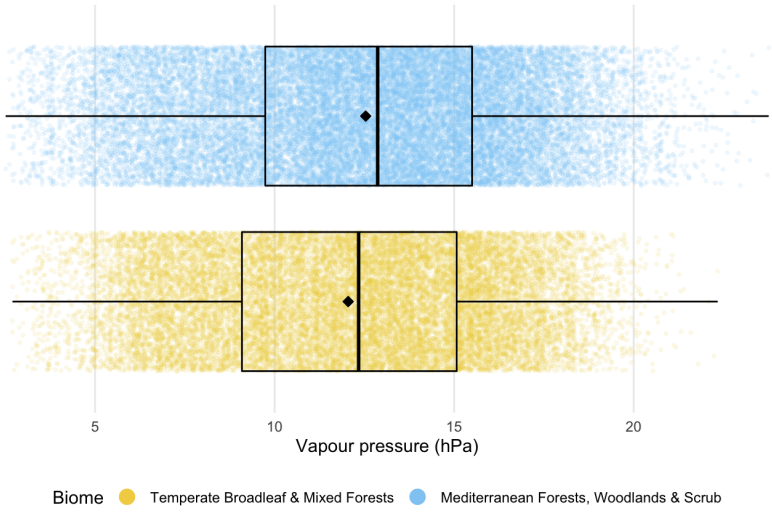


Figure C.49: Boxplot by biome of vapour pressure

And, grouped by month the data shows the variability typical of meteorological data:

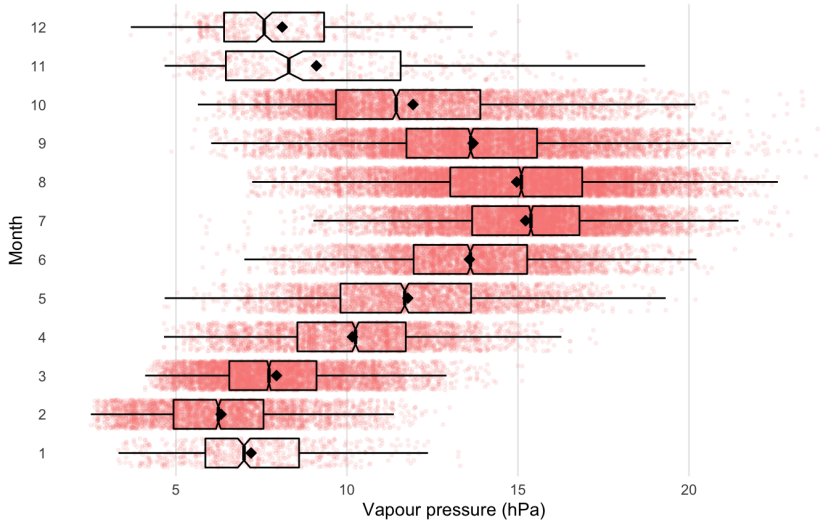


Figure C.50: Boxplot by month of vapour pressure

C.6.6 Meteorological data: total global radiation

The distribution grouped by year shows that the spread for all years is similar with the lower amount of data for the year 2014 patent by the number of superimposed points in the figure:

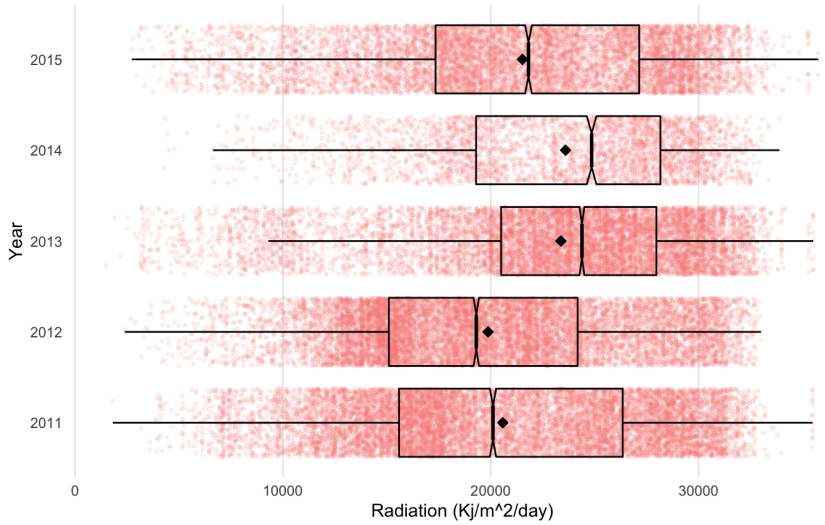


Figure C.51: Boxplot by year of radiation

And, grouped by month the data shows the typical variation of the amount of insolation across the year:

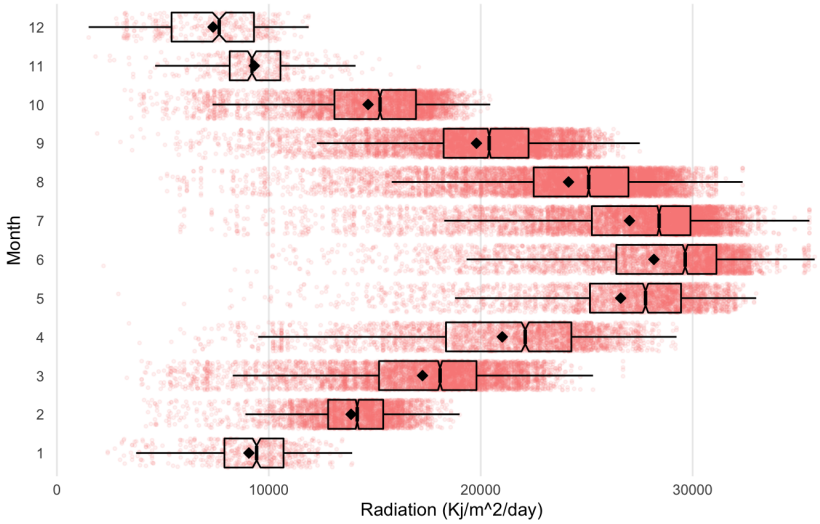


Figure C.52: Boxplot by month of radiation

C.6.7 Canadian Forest Fire Weather Index (FWI): FFMC

The distribution for each biome is similar with high values for the HCFs, but also a high towards zero. Moreover, this pattern persists across ever year:

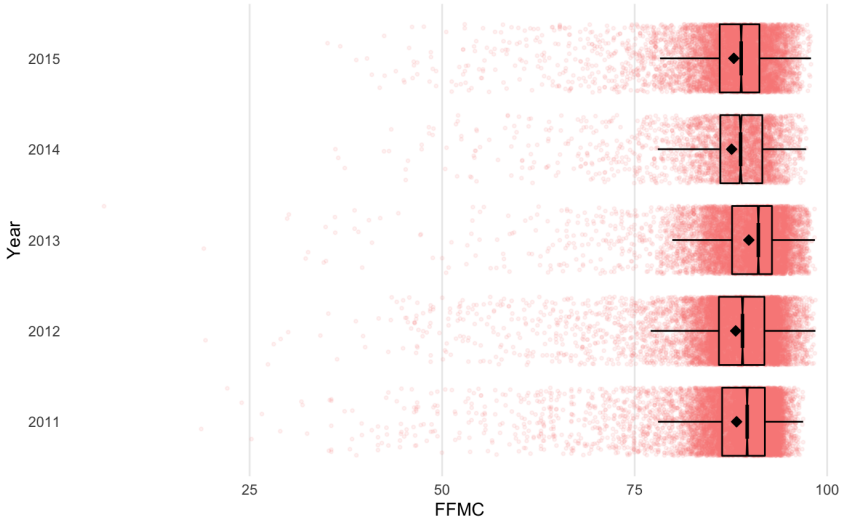


Figure C.53: Boxplot by year of FFMC

And, every month:

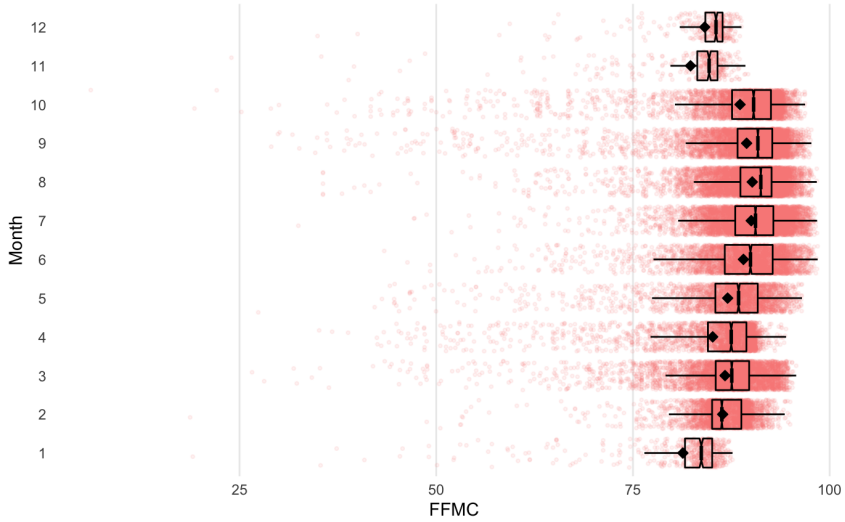


Figure C.54: Boxplot by month of FFMC

There are lower values in average for the winter months than other period since FFMC is calculated from meteorological variables (temperature, relative humidity, wind, and rain).

But, except for the outliers, the values are high indicating an easy ignition and high flammability of fine fuels the days of occurrence of HCFs.

C.6.8 Canadian Forest Fire Weather Index (FWI): DMC

The spread of values in the Mediterranean Forests, Woodlands & Scrub biome shows a larger variability with extreme upper values. Moreover, its maximum value almost doubles the one in the other biome.

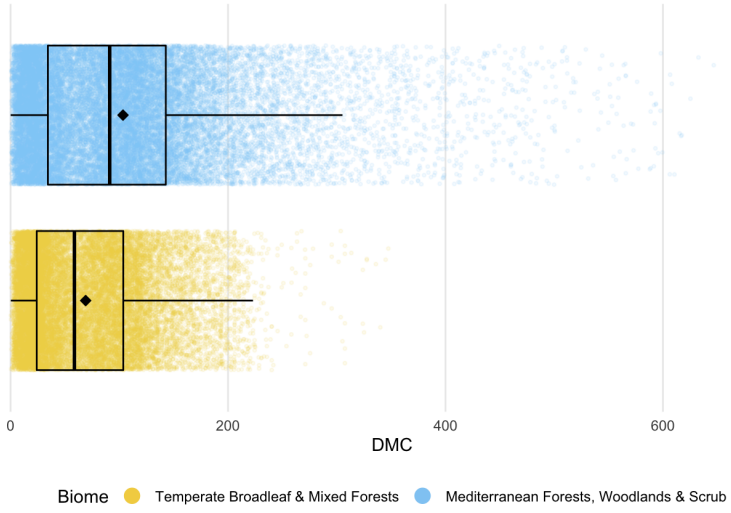


Figure C.55: Boxplots by biome of DMC

There are higher values in average for the summer months since this index represents moisture conditions for the equivalent of 15-day lag indicating fires driven by lower moisture content in loosely compacted organic layers of moderate depth:

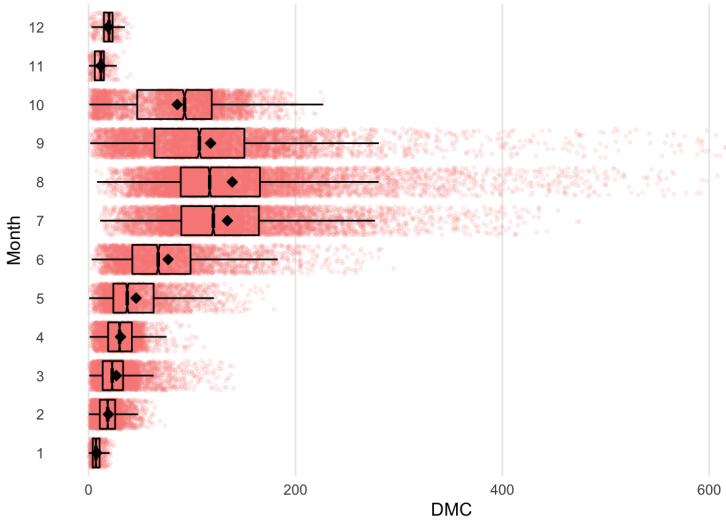


Figure C.56: Boxplots by month of DMC

C.6.9 Canadian Forest Fire Weather Index (FWI): DC

The Mediterranean Forests, Woodlands & Scrub biome is climatological more diverse. The FWI metrics are calculated from meteorological data so

its values are more spread and reaching higher maximum values than the other biome:

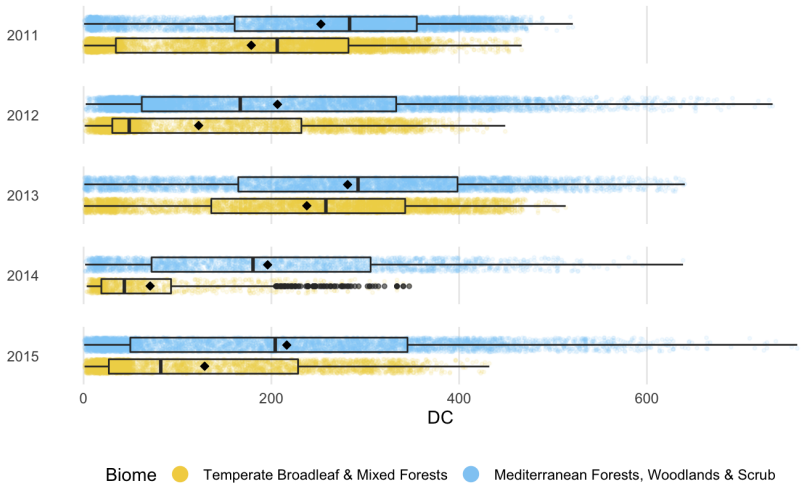


Figure C.57: Boxplots by biome and year of DC

C.6.10 Canadian Forest Fire Weather Index (FWI): ISI

The values for the Temperate Broadleaf & Mixed Forests are on average lower than the other biome but both show outliers in the upper range:

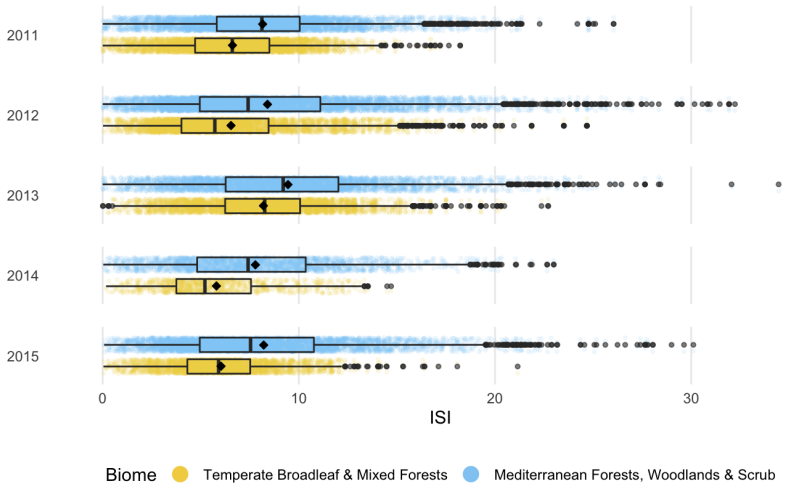


Figure C.58: Boxplots by biome and year of ISI

The median and average values of the indicator can be classified as mod-

erate, but the values above 15–20 can be considered extreme (the threshold values depend on local conditions).

C.6.11 Canadian Forest Fire Weather Index (FWI): BUI

The BUI metric shows higher values on average on the summer months as with other FWI metrics, which are calculated from meteorological data, with the outliers visible for those months standing out:

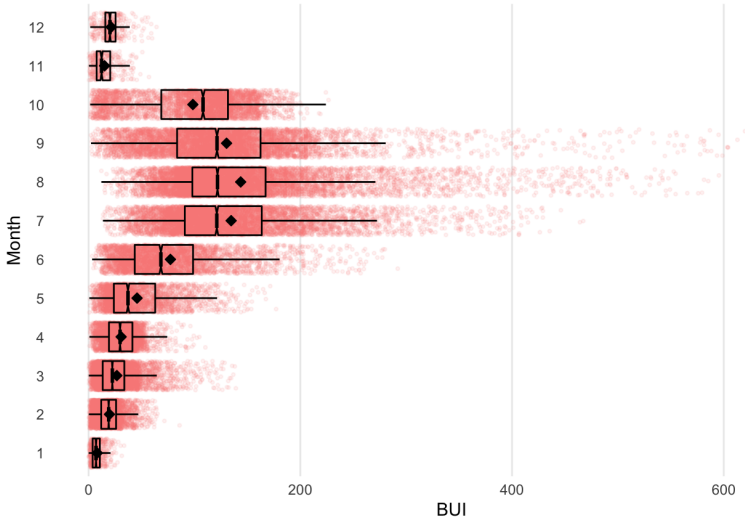


Figure C.59: Boxplots by month of BUI

C.6.12 Canadian Forest Fire Weather Index (FWI): FWI

The patterns already indicated for other variables of the Canadian Forest Fire Weather Index:

- showing higher values on average on the summer months
- values in from the Mediterranean Forests, Woodlands & Scrub biome being more spread and reaching higher maximum values

held also for the FWI metric:

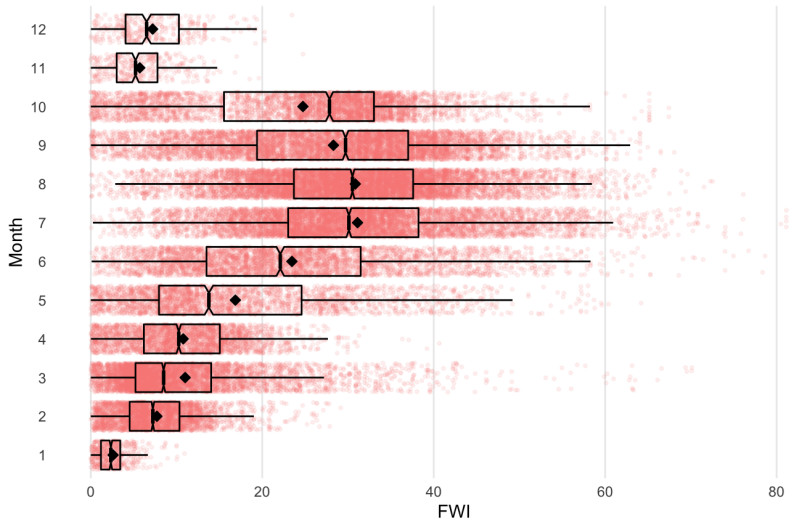


Figure C.60: Boxplots by month of FWI

Values above 30–40 can be considered extreme (the graduation is subject to local conditions), as those reached in the summer and early spring in continental Portugal in both biomes:

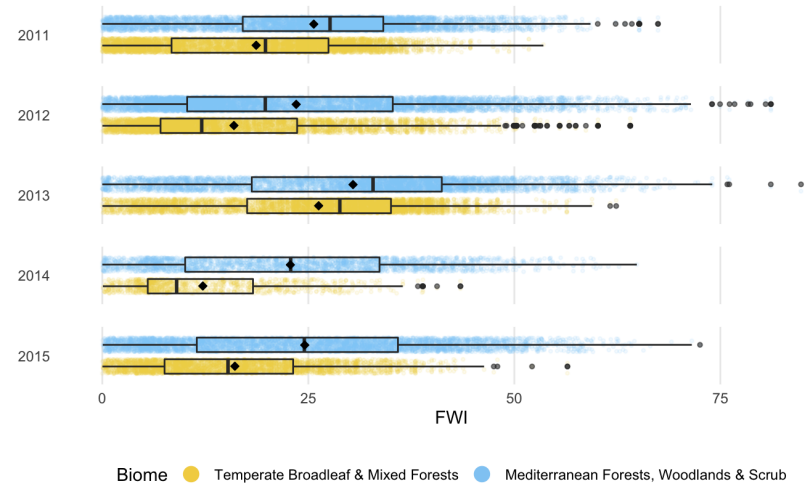


Figure C.61: Boxplots by biome and year of FWI

C.6.13 Physiography: elevation

The "Mediterranean Forests, Woodlands & Scrub" biome cover a more diverse landscape and this is reflected in the density plot with a large number of fires occurring at low elevations and a smaller number at higher elevation.

However, fire occurrence in the Temperate Broadleaf & Mixed Forests biome is concentrated in lower elevations, but not so low than in the other biome.

The distribution of the elevation of HCFs occurrence does not change so much across all the period of study:

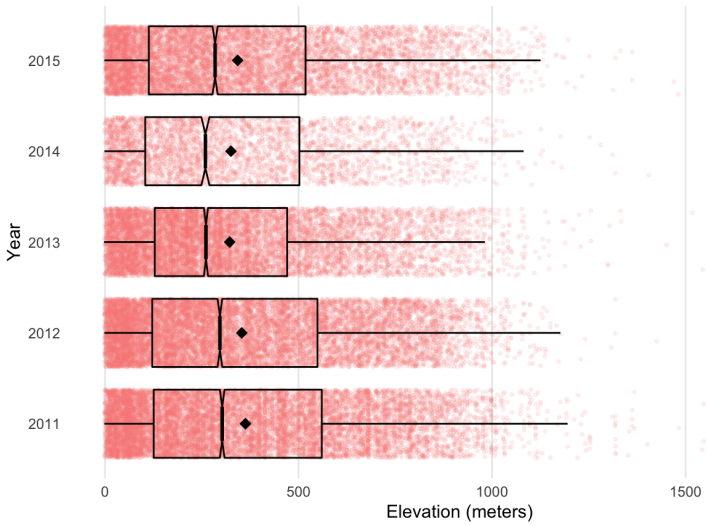


Figure C.62: Boxplots by year of elevation

C.6.14 Physiography: slope

The density plot of the slope distribution for the overall distribution and in the one for each biome there are huge gaps for the same values. This is a consequence of the way the value is encoded in the original data (digital numbers from the satellite sensor) and converted to decimal degrees:

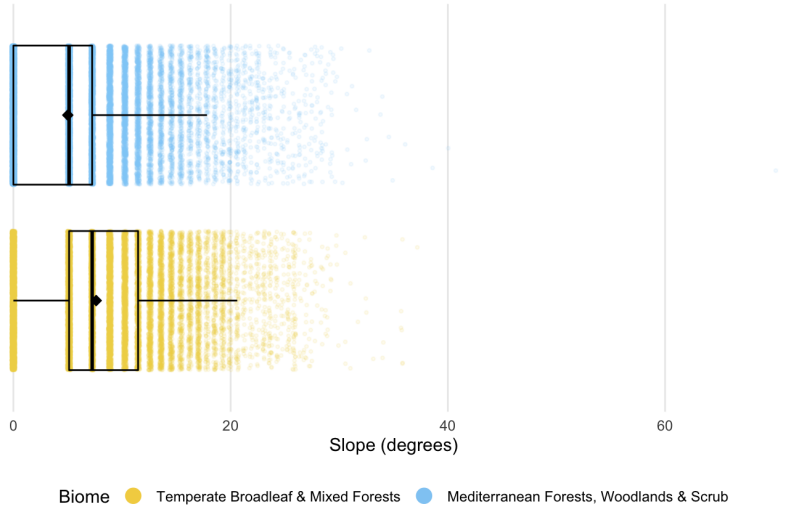
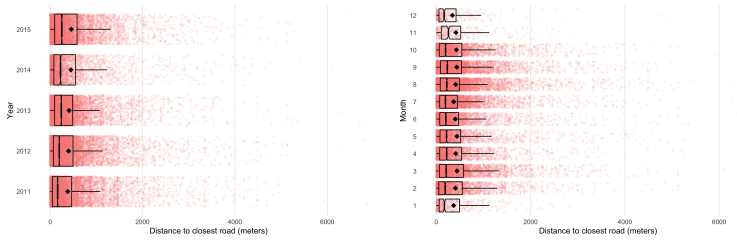


Figure C.63: Boxplots by biome of slope

However, HCFs in the "Mediterranean Forests, Woodlands & Scrub" occur on gentler slopes.

C.6.15 Human factors: distance to nearest road

While most HCFs happen close to a road, which allows easy access, there are a high number of outliers:

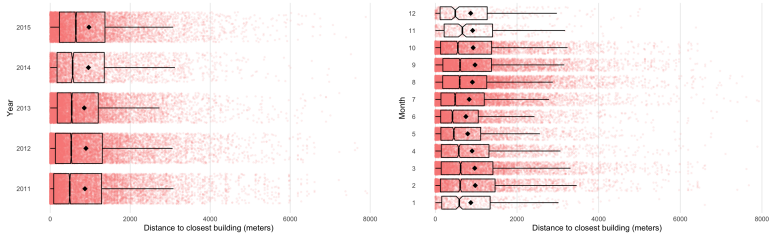


(a) Grouped by year (b) Grouped by month

Figure C.64: Boxplots by year and by month of distance to nearest road

C.6.16 Human factors: distance to nearest building

While most HCFs happen close to a build-up area, where humans that are the direct or indirect cause of HCFs live, there are a high number of outliers:



(a) Grouped by year (b) Grouped by month

Figure C.65: Boxplots by year and by month of distance to nearest building

C.7 Geospatial distribution of a variable

To explore this aspect of the data I am going to use maps (using using the EPSG:4326) to explore the geospatial distribution of a variable.

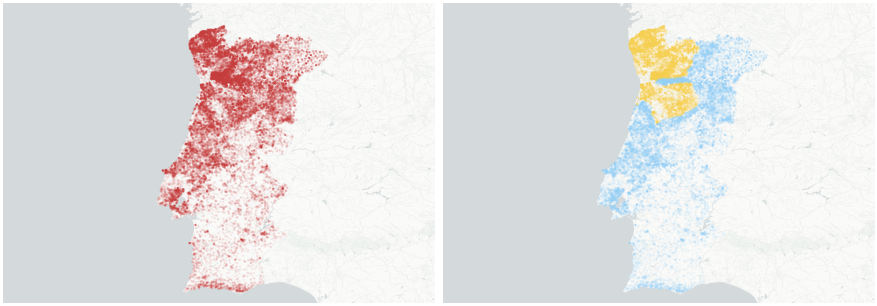
C.7.1 Major Habitat Types (MHTs)

The area occupied by each ecoregion, as can be seen in the Figure 5.1, in squared kilometres is:

Biome	Area (Km^2)	
Temperate Broadleaf & Mixed Forests	79,823.58	15.81%
Mediterranean Forests, Woodlands & Scrub	42,5214.23	84.19%

Table C.60: Biome area in continental Portugal

And, visualising the location of the fires using transparency to make evident the concentration of fires:



(a) All (b) By biome

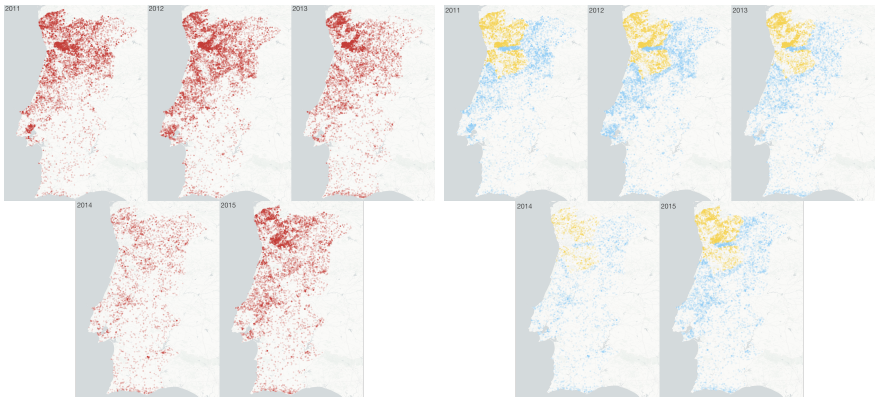
Figure C.66: Location of the HCFs in period 2011–2015

Moreover, the total number of observations and the burnt area by biome, which are:

Biome		Count	Burnt area (ha)	
Temperate Broadleaf & Mixed Forests	21,590	45.05%	127,331.6	40.55%
Mediterranean Forests, Woodlands & Scrub	26,337	854.95%	186,691.9	59.45%
Total	47,927	100%	314,023.6	100 %

Table C.61: HCFs count and burnt area by biome

And again, visualising the location of the fires using transparency to make evident the concentration of fires but by year:



(a) All (b) By biome

Figure C.67: Location of HCFs by year in period 2011–2015

The plots reveal that the biome Temperate Broadleaf & Mixed Forests has suffered a larger number of fires in proportion to its extension revealing a stark difference between the northern and southern halves of the territory.

Appendix D

Feature engineering

This appendix contains the detailed feature engineering step that is summarised in the Chapter 7.

I am going to combine in this section the data preparation and feature engineering stages combining all the steps necessary to prepare the data, so it fits the selected model:

- Treating outliers
- Reducing high cardinality in categorical variables
- Assigning the correct data types for each variable (some algorithms only work with certain data types)
- Handling missing data
- Creating new variables

D.1 Handling outliers

I am going to handle the outliers found on the data sets considering that in some cases whether a value is abnormal is a matter of perspective.

Also, I am going to try to make the least modifications possible to the data since each change distorts the data and may introduce bias.

For each variable I am going to try 3 methods to detect and treat outliers:

- Bottom/Top x%, based on percentiles. Common values are 0.5%, 1%, 1.5%, 3%, among others.
- Tukey, based on the quartiles. It considers outliers all values outside of the interval $[\text{Q1} - 3\text{IQR}, Q3 + 3\text{IQR}]$ [Tuk81].
- Hampel, based on the median and median absolute deviation (MAD) values. It considers outliers all values outside of the interval $[\text{median} - k * \text{MAD}, \text{median} + k * \text{MAD}]$ [Ham74].

To handle the outliers I am going to use feature clipping, that is, to set the values of the observation for the affected feature to the threshold value that the chosen method has decided separates the abnormal values from the rest.

D.1.1 Human-caused fires (HCFs)

Burnt area

I am only going to check the burnt area (i.e., area_total) variable since there makes no sense to check for outliers the location variables.

As I saw in the data exploration phase, the burnt area is a highly skewed variable. This is also evident in how the values of the standard deviation is much higher than the mean:

Mean	5.90
Standard deviation	85.62
Coefficient of variation	14.52
Skewness	45.11
Kurtosis	2,710.31
1^{st} percentile	0.0002
5^{st} percentile	0.002
25^{st} percentile	0.02
50^{st} percentile	0.1
75^{st} percentile	1
95^{st} percentile	10
99^{st} percentile	82.06
IQR	82.06
Range with 80% observations	$[0.005, 3.5]$
Range with 98% observations	$[0.0002, 82.06]$

Table D.1: Central and skewness metrics of burnt area

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for only the top 5%
- Tukey method
- Hampel method using k = 3

The number and percentage of flagged outliers is:

Method	Outliers	Percentage
Bottom / Top	1,091	5.05%
Tukey	2,036	9.43%
Hampel	6,512	30.16%

Table D.2: Flagged outliers in burnt area

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	5.90	85.62	14.52	45.11	2,710.31
Bottom / Top	0.69	1.41	2.03	3.34	15.79
Tukey	0.44	0.74	1.68	2.23	7.63
Hampel	0.10	0.14	1.37	1.71	4.83

Table D.3: Skewness metrics before and after imputing outliers in burnt area

With the range covering the 95% of the values is:

Original	$[0.001, 26]$
Bottom / Top	$[0.001, 5.27]$
Tukey	$[0.001, 3.00]$
Hampel	$[0.0005, 0.5]$

Table D.4: Intervals covering 95% of observation before and after imputing outliers in burnt area

Before the standard deviation was large compared to the mean, as reflected in the variation coefficient and kurtosis value. After, both methods have reduced all the metrics.

Comparing graphically the outliers in the original variable and all methods:

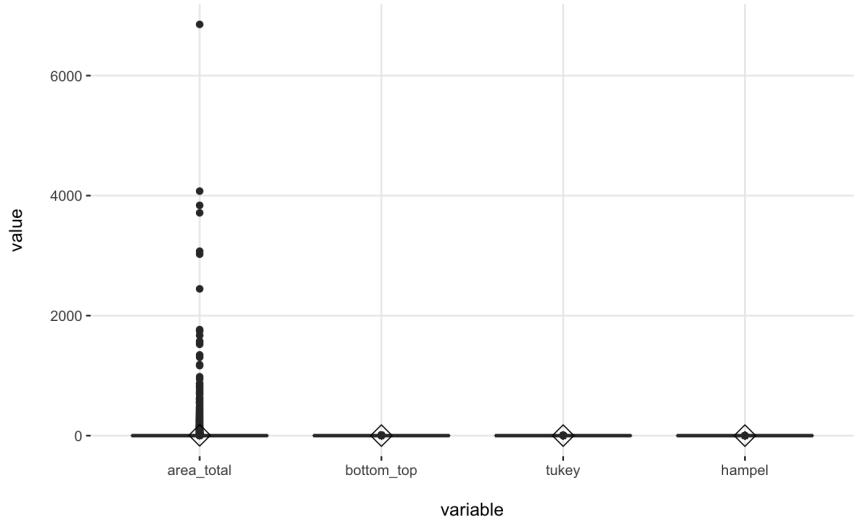


Figure D.1: Box plots before and after imputing outliers in burnt area

And, comparing only the result of the three methods:

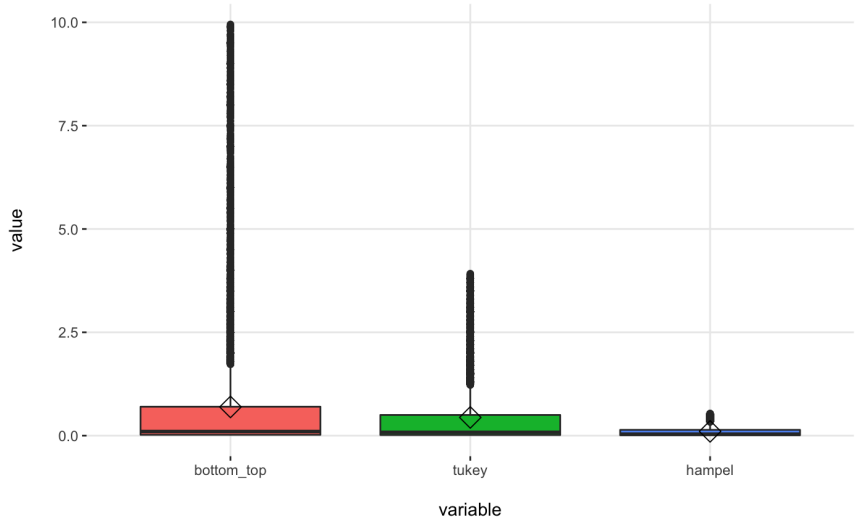


Figure D.2: Box plots after imputing outliers in burnt area

The distribution of the new variables is still skewed but with a much smaller tail.

I will use the Bottom/Top method that imputes the least number of observations while at the same time reducing the skewness of the distribution. The density plot for the burnt area treated with the Bottom/Top method is:

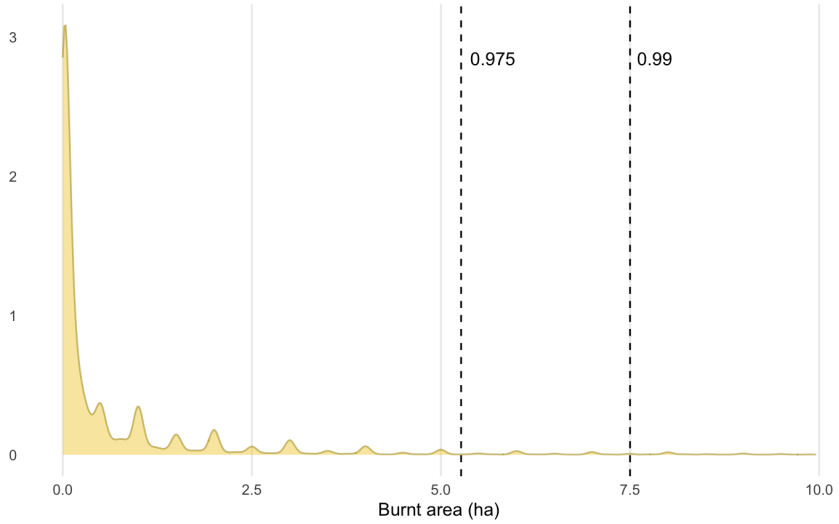


Figure D.3: Density plot of burnt area after applying the Bottom / Top method

D.1.2 Weather factors: meteorological data

Average temperature

As seen in the data exploration phase, average temperature is not a skewed variable. Although there are some observations with extreme low values to the rest. This is evident in the low coefficient of variation and the percentiles:

Mean	18.10
Standard deviation	5.28
Coefficient of variation	0.29
Skewness	-0.51
Kurtosis	2.97
1^{st} percentile	4.7
5^{st} percentile	8.24
25^{st} percentile	14.9
50^{st} percentile	18.9
75^{st} percentile	21.9
95^{st} percentile	25.7
99^{st} percentile	28.5
IQR	7
Range with 80% observations	$[4.7, 28.5]$
Range with 98% observations	$[10.3, 24.2]$

Table D.5: Central and skewness metrics of average temperature

The density plot marking the 0.5th, 5th, 95th, 99.5th percentiles is:

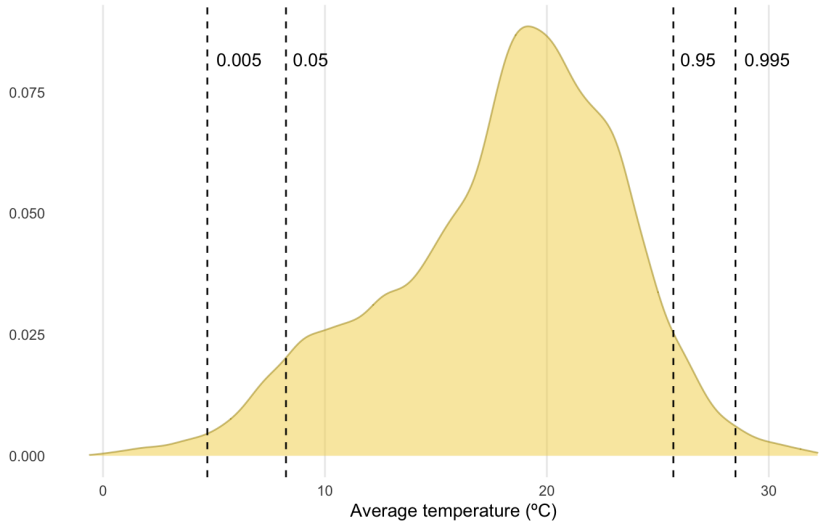


Figure D.4: Density plot of average temperature with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for a 1% of the data split between top and bottom
- Tukey method
- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	232	1.07%
Tukey	0	0%
Hampel	179	0.83%

Table D.6: Flagged outliers in average temperature

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	18.10	5.28	0.29	-0.51	2.97
Bottom / Top	18.12	5.09	0.28	-0.50	2.73
Tukey	19.10	5.28	0.29	-0.51	2.97
Hampel	18.23	5.10	0.28	-0.41	2.76

Table D.7: Skewness metrics before and after imputing outliers in average temperature

And, the range covering the 99% of the values is:

Original	$[3.4, 29.5]$
Bottom / Top	$[4.8, 28.42]$
Tukey	$[3.4, 29.5]$
Hampel	$[5.4, 29.5]$

Table D.8: Intervals covering 99% of observation before and after imputing outliers in average temperature

Because the percentage of flagged outliers is so low and the skewness indicators does not improve significantly, I am not going to treat any values of this variable as outliers and modify it.

Maximum temperature

As seen in the data exploration phase, average temperature is not a skewed variable. Although there are some observations with extreme low values to the rest. This is evident in the low coefficient of variation and the percentiles:

Mean	24.37
Standard deviation	6.12
Coefficient of variation	0.25
Skewness	-0.25
Kurtosis	2.67
1^{st} percentile	10
5^{st} percentile	13.7
25^{st} percentile	20.6
50^{st} percentile	24.5
75^{st} percentile	29
95^{st} percentile	33.8
99^{st} percentile	37
IQR	8.4
Range with 80% observations	$[15.6, 32]$
Range with 98% observations	$[10, 37]$

Table D.9: Central and skewness metrics of maximum temperature

The density plot marking the 0.5th, 5th, 95th, 99.5th percentiles is:

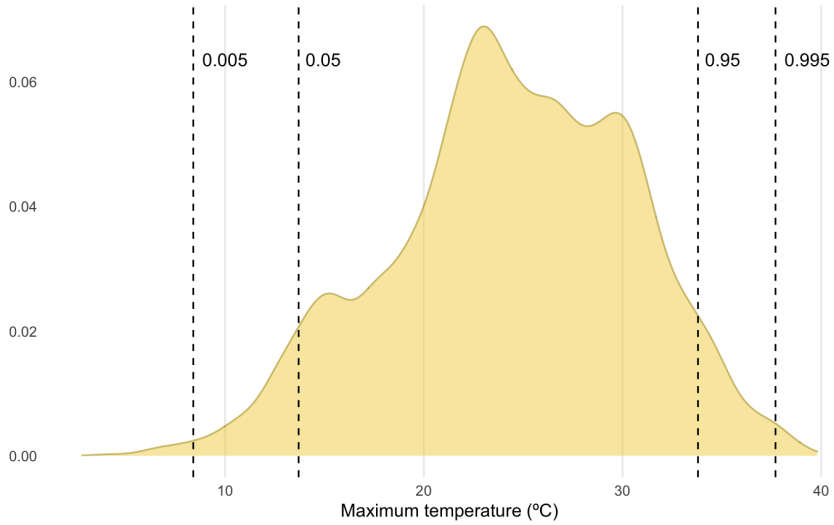


Figure D.5: Density plot of maximum temperature with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for a 1% of the data split between top and bottom
- Tukey method
- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	231	1.07%
Tukey	0	0%
Hampel	12	0.06%

Table D.10: Flagged outliers in maximum temperature

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	24.37	6.12	0.25	-0.25	2.67
Bottom / Top	5.93	0.24	-0.22	-0.50	2.49
Tukey	24.37	6.12	0.25	-0.25	2.67
Hampel	24.38	6.10	0.25	-0.24	2.64

Table D.11: Skewness metrics before and after imputing outliers in maximum temperature

And, the range covering the 99% of the values is:

Original	$[8.4, 37.7]$
Bottom / Top	$[10.0, 36.9]$
Tukey	$[8.4, 37.7]$
Hampel	$[8.69, 37.7]$

Table D.12: Intervals covering 99% of observation before and after imputing outliers in maximum temperature

Because the number of flagger outliers is so low and the skewness indicators does not improve significantly, I am not going to treat any values of this variable as an outliers and modify it.

Wind speed

As I saw in the data exploration phase, the wind speed has a small tail on the right, but it is not a highly skewed variable. Although, there are some observations with extreme low values to the rest. This is evident in the low coefficient of variation and the percentiles:

Mean	2.51
Standard deviation	1.06
Coefficient of variation	0.41
Skewness	1.32
Kurtosis	6.37
1^{st} percentile	0.8
5^{st} percentile	1.2
25^{st} percentile	1.8
50^{st} percentile	2.3
75^{st} percentile	3
95^{st} percentile	4.5
99^{st} percentile	6
IQR	1.2
Range with 80% observations	$[1.4, 3.9]$
Range with 98% observations	$[0.8, 6]$

Table D.13: Central and skewness metrics of wind speed

The density plot marking the 0.5th, 5th, 95th, 99.5th percentiles is:

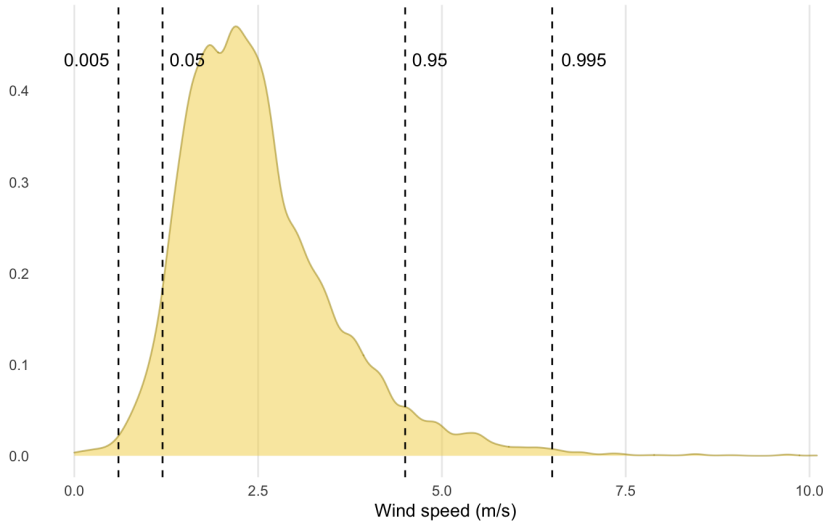


Figure D.6: Density plot of wind speed with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for a 1% of the data split between top and bottom
- Tukey method
- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	249	1.15%
Tukey	103	0.48%
Hampel	665	3.08%

Table D.14: Flagged outliers in wind speed

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	2.51	1.06	0.43	1.32	6.37
Bottom / Top	2.49	0.98	0.39	0.99	4.10
Tukey	2.48	1.00	0.40	0.96	4.18
Hampel	2.40	0.88	0.36	0.54	2.97

Table D.15: Skewness metrics before and after imputing outliers in wind speed

And, the range covering the 99% of the values is:

Original	$[0.6, 6.5]$
Bottom / Top	$[0.8, 5.9]$
Tukey	$[0.6, 6.0]$
Hampel	$[0.6, 4.8]$

Table D.16: Intervals covering 99% of observation before and after imputing outliers in wind speed

Because the number of flagged outliers is so low and the skewness indicators does not improve significantly, I am not going to treat any values of this variable as an outliers and modify it.

Vapour pressure

As I saw in the data exploration phase, the wind speed has a small tail on the right, but it is not a highly skewed variable. Although, there are some observations with extreme low values to the rest. This is evident in the low coefficient of variation and the percentiles:

Mean	12.05
Standard deviation	3.75
Coefficient of variation	0.31
Skewness	-0.15
Kurtosis	2.17
1^{st} percentile	4.20
5^{st} percentile	5.91
25^{st} percentile	9.09
50^{st} percentile	12.34
75^{st} percentile	15.07
95^{st} percentile	17.77
99^{st} percentile	19.18
IQR	5.98
Range with 80% observations	[6.79, 16.84]
Range with 98% observations	[4.20, 19.18]

Table D.17: Central and skewness metrics of vapour pressure

The density plot marking the 0.5th, 5th, 95th, 99.5th percentiles is:

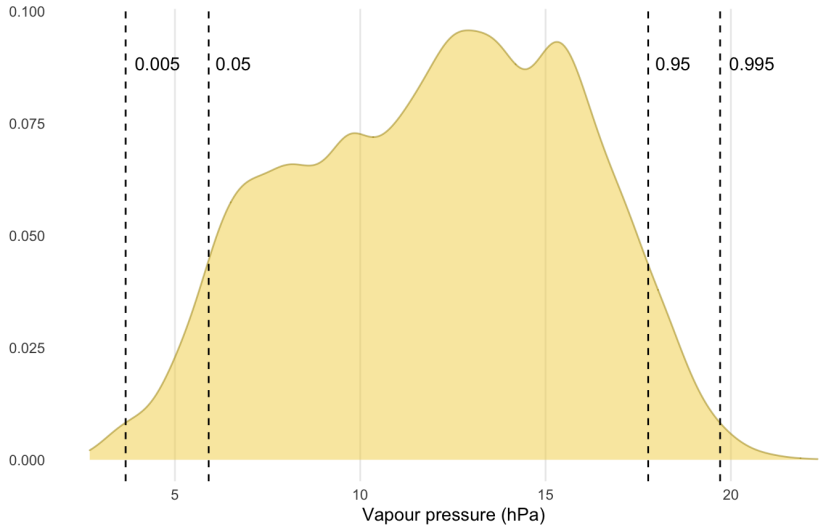


Figure D.7: Density plot of Vapour pressure with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for a 1% of the data split between top and bottom
- Tukey method
- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	220	1.02%
Tukey	0	0%
Hampel	0	0%

Table D.18: Flagged outliers in vapour pressure

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	12.51	3.75	0.31	-0.15	2.17
Bottom / Top	12.51	3.67	0.31	-0.15	2.10
Tukey	12.51	3.75	0.31	-0.15	2.17
Hampel	12.51	3.75	0.31	-0.15	2.17

Table D.19: Skewness metrics before and after imputing outliers in vapour pressure

And, the range covering the 99% of the values is:

Original	[3.67, 19.71]
Bottom / Top	[4.20, 19.18]
Tukey	[3.67, 19.71]
Hampel	[3.67, 19.71]

Table D.20: Intervals covering 99% of observation before and after imputing outliers in vapour pressure

Since only the Bottom/Top method detects outliers and this method will always flags as outliers the prescribed percentage, I am not going to apply any transformation to this variable.

Radiation

As I saw in the data exploration phase, the radiation is not a skewed variable. This is also evident in how the values of the standard deviation is much higher than the mean:

Mean	20,904.13
Standard deviation	6,253.34
Coefficient of variation	0.30
Skewness	-0.19
Kurtosis	2.35
1^{st} percentile	6,925
5^{st} percentile	10,262
25^{st} percentile	16,200
50^{st} percentile	21,163
75^{st} percentile	25,800
95^{st} percentile	30,632.5
99^{st} percentile	32,300
IQR	7
Range with 80% observations	[10.3, 24.2]
Range with 98% observations	[4.7, 28.5]

Table D.21: Central and skewness metrics of radiation

The density plot marking the 0.5th, 5th, 95th, 99.5th percentiles is:

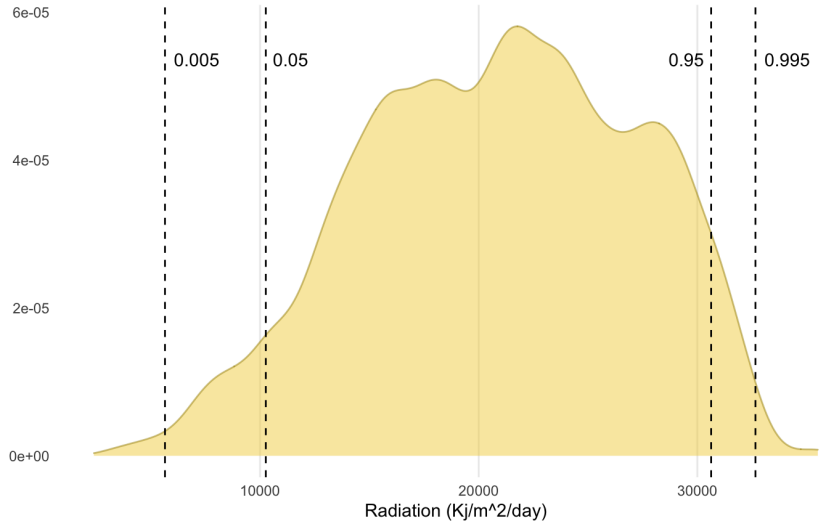


Figure D.8: Density plot of radiation with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for a 1% of the data split between top and bottom
- Tukey method
- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	111	0.51%
Tukey	0	0%
Hampel	0	0%

Table D.22: Flagged outliers in radiation

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	20,904.13	6,253.34	0.30	-0.19	2.35
Bottom / Top	20,904.13	6,253.34	-0.21	-0.15	2.33
Tukey	20,904.13	6,253.34	0.30	-0.19	2.35
Hampel	20,904.13	6,253.34	0.30	-0.19	2.35

Table D.23: Skewness metrics before and after imputing outliers in radiation

And, the range covering the 99% of the values is:

Original	$[5, 650, 32, 660]$
Bottom / Top	$[5, 650, 32, 660]$
Tukey	$[5, 650, 32, 660]$
Hampel	$[5, 650, 32, 660]$

Table D.24: Intervals covering 99% of observation before and after imputing outliers in radiation

Since only the Bottom/Top method detects outliers and this method will always flags as outliers the prescribed percentage, I am not going to apply any transformation to this variable.

D.1.3 Weather factors: Canadian Forest Fire Weather Index (FWI) System

Fine Fuel Moisture Code (FFMC)

As I saw in the data exploration phase, the FFMC is has a tail towards the left. This is evident in the percentiles value:

Mean	88.00
Standard deviation	5.85
Coefficient of variation	0.07
Skewness	-3.65
Kurtosis	24.89
1^{st} percentile	62.35
5^{st} percentile	79.31
25^{st} percentile	86.22
50^{st} percentile	89.07
75^{st} percentile	91.43
95^{st} percentile	93.60
99^{st} percentile	94.79
IQR	7
Range with 80% observations	[83.33, 92.91]
Range with 98% observations	[62.35, 94.79]

Table D.25: Central and skewness metrics of FFMC

The density plot marking the 1th, 5th, 95th, 99th percentiles is:

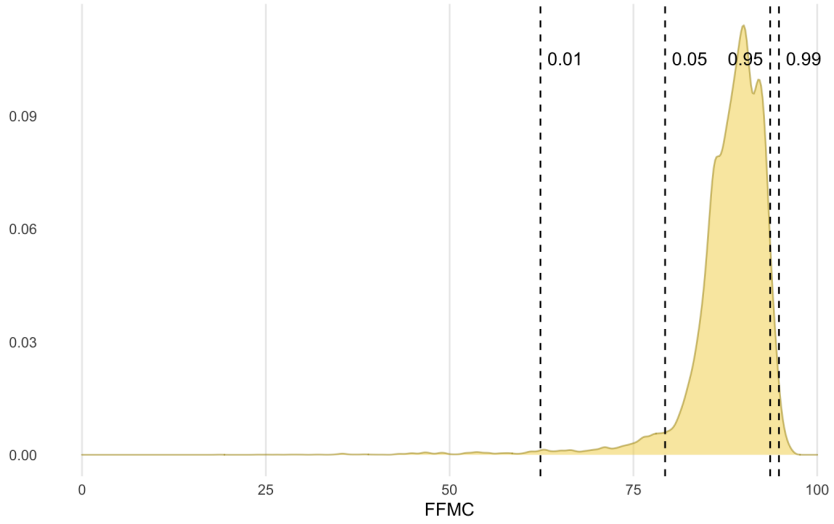


Figure D.9: Density plot of FFMC with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for the 2% of the observations split between the bottom and top
- Tukey method

- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	431	2.00%
Tukey	421	1.95%
Hampel	868	4.02%

Table D.26: Flagged outliers in FFMC

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	88.00	5.85	0.07	-3.65	24.89
Bottom / Top	4.44	0.05	-1.90	-0.15	9.13
Tukey	88.59	3.92	0.04	-1.16	5.20
Hampel	88.88	3.39	0.04	-0.56	3.08

Table D.27: Skewness metrics before and after imputing outliers in FFMC

And, the range covering the 98% of the values is:

Original	[62.35, 94.79]
Bottom / Top	[70.80, 94.36]
Tukey	[75.09, 94.82]
Hampel	[79.23, 94..84]

Table D.28: Intervals covering 98% of observation before and after imputing outliers in FFMC

Because the percentage of flagged outliers is so low and the skewness indicators, even although they improve, are not band to begin with, I am not going to treat any values of this variable as outliers and modify it.

Duff Moisture Code (DMC)

As I saw in the data exploration phase, the DMC has a small tail on the right, but it is not a highly skewed variable. However, the value of the standard deviation is the 75.76% of the mean:

Mean	69.06
Standard deviation	52.32
Coefficient of variation	0.76
Skewness	0.87
Kurtosis	3.38
1^{st} percentile	3.43
5^{st} percentile	8.17
25^{st} percentile	24.04
50^{st} percentile	58.71
75^{st} percentile	103.66
95^{st} percentile	169.30
99^{st} percentile	209.88
IQR	79.62
Range with 80% observations	[12.48, 141.271]
Range with 98% observations	[3.43, 209.88]

Table D.29: Central and skewness metrics of DMC

The density plot marking the 95th, 99th, 99.5th percentiles is:

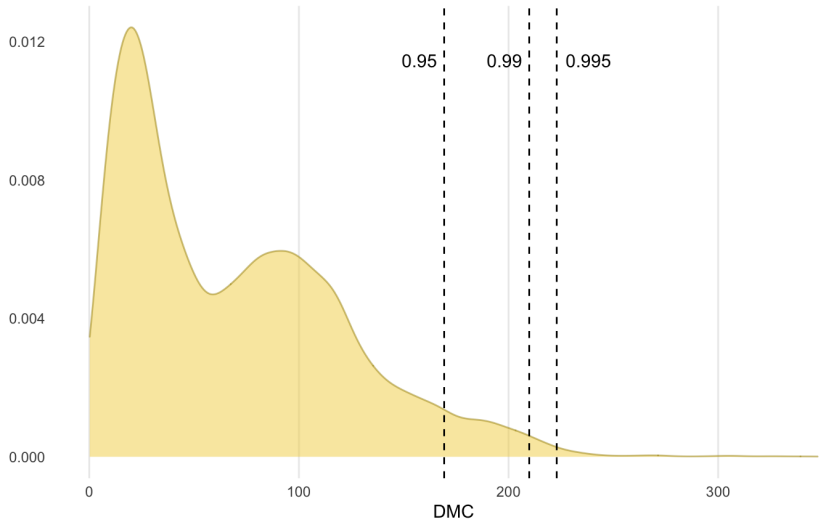


Figure D.10: Density plot of DMC with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for 1% of the top data
- Tukey method

- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	233	1.08%
Tukey	2	0.01%
Hampel	89	0.41%

Table D.30: Flagged outliers in DMC

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	69.06	52.32	0.76	0.87	3.38
Bottom / Top	49.56	0.74	0.69	-0.15	2.62
Tukey	69.03	52.26	0.76	0.86	3.32
Hampel	68.26	50.87	0.75	0.75	2.78

Table D.31: Skewness metrics before and after imputing outliers in DMC

And, the range covering the 99% of the values is:

Original	[2.63, 222.98]
Bottom / Top	[2.63, 202.33]
Tukey	[2.63, 221.60]
Hampel	[2.63, 212.24]

Table D.32: Intervals covering 99% of observation before and after imputing outliers in DMC

Because the percentage of flagged outliers is so low except for the Bottom/Top method, I am not going to treat any values of this variable as outliers and modify it.

Drought Code (DC)

As I saw in the data exploration phase, the DC is not a skewed variable. However, the value of the standard deviation is the 79.09% of the mean:

Mean	165.57
Standard deviation	130.95
Coefficient of variation	0.79
Skewness	0.30
Kurtosis	1.76
1^{st} percentile	4.14
5^{st} percentile	10.03
25^{st} percentile	33.69
50^{st} percentile	159.00
75^{st} percentile	281.38
95^{st} percentile	375.15
99^{st} percentile	440.57
IQR	247.69
Range with 80% observations	[16.62, 342.23]
Range with 98% observations	[4.14, 440.57]

Table D.33: Central and skewness metrics of DC

The density plot marking the 95th, 99th, 99.5th percentiles is:

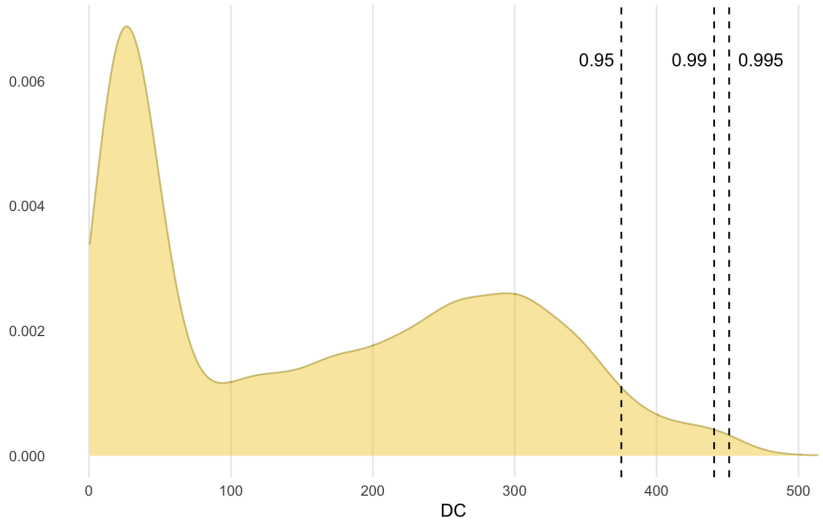


Figure D.11: Density plot of DC with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for 1% of the top data
- Tukey method
- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	217	1.01%
Tukey	0	0%
Hampel	0	0%

Table D.34: Flagged outliers in DC

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	165.57	130.95	0.79	0.30	1.72
Bottom / Top	162.64	128.31	0.27	-0.15	1.64
Tukey	165.57	130.95	0.79	0.30	1.72
Hampel	165.57	130.95	0.79	0.30	1.72

Table D.35: Skewness metrics before and after imputing outliers in DC

And, the range covering the 99% of the values is:

Original	$[3.14, 451.16]$
Bottom / Top	$[3.14, 429.07]$
Tukey	$[3.14, 451.16]$
Hampel	$[3.14, 451.16]$

Table D.36: Intervals covering 99% of observation before and after imputing outliers in DC

Because the percentage of flagged outliers is so low except for the Bottom/Top method, I am not going to treat any values of this variable as outliers and modify it.

Initial Spread Index (ISI)

As I saw in the data exploration phase, the ISI is not a skewed variable:

Mean	6.85
Standard deviation	3.14
Coefficient of variation	0.46
Skewness	0.59
Kurtosis	3.80
1^{st} percentile	0.83
5^{st} percentile	2.20
25^{st} percentile	4.61
50^{st} percentile	6.58
75^{st} percentile	8.78
95^{st} percentile	12.32
99^{st} percentile	15.33
IQR	4.17
Range with 80% observations	[3.02, 10.94]
Range with 98% observations	[0.82, 15.33]

Table D.37: Central and skewness metrics of ISI

The density plot marking the 95th, 99th, 99.5th percentiles is:

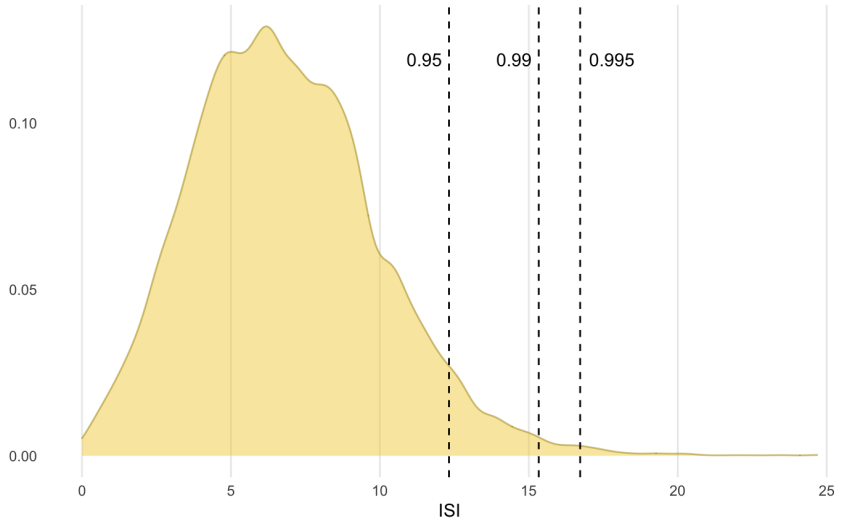


Figure D.12: Density plot of ISI with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for 1% of the top data
- Tukey method
- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	215	1.01%
Tukey	15	0.07%
Hampel	170	0.79%

Table D.38: Flagged outliers in ISI

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	6.85	3.14	0.46	0.59	3.80
Bottom / Top	6.74	2.95	0.44	0.28	2.70
Tukey	6.84	3.11	0.45	0.51	3.39
Hampel	6.76	2.98	0.44	0.31	2.76

Table D.39: Skewness metrics before and after imputing outliers in ISI

And, the range covering the 99% of the values is:

Original	$[0.53, 16.72]$
Bottom / Top	$[0.53, 14.62]$
Tukey	$[0.53, 16.65]$
Hampel	$[0.53, 14.94]$

Table D.40: Intervals covering 99% of observation before and after imputing outliers in ISI

Because the percentage of flagged outliers are so low, I am not going to treat any values of this variable as outliers and modify it.

Buildup Index (BUI)

As I saw in the data exploration phase, the BUI has a small tail on the right, but it is not a highly skewed variable. However, the value of the standard deviation is the 73.67% of the mean:

Mean	72.34
Standard deviation	53.29
Coefficient of variation	0.74
Skewness	0.71
Kurtosis	3.00
1^{st} percentile	3.17
5^{st} percentile	8.07
25^{st} percentile	24.56
50^{st} percentile	65.05
75^{st} percentile	109.72
95^{st} percentile	171.04
99^{st} percentile	209.06
IQR	85.16
Range with 80% observations	[12.55, 143.43]
Range with 98% observations	[3.17, 209.06]

Table D.41: Central and skewness metrics of BUI

The density plot marking the 95th, 99th, 99.5th percentiles is:

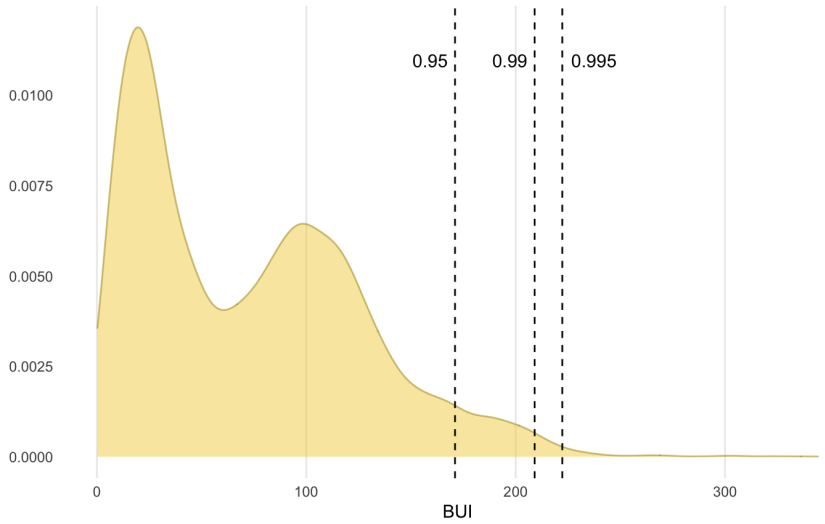


Figure D.13: Density plot of BUI with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for 1% of the top data
- Tukey method

- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	234	1.08%
Tukey	0	0%
Hampel	44	0.20%

Table D.42: Flagged outliers in BUI

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	72.34	53.29	0.74	0.71	3.00
Bottom / Top	70.58	50.75	0.72	0.55	2.35
Tukey	72.34	53.29	0.74	0.71	3.00
Hampel	71.89	52.42	0.73	0.62	2.57

Table D.43: Skewness metrics before and after imputing outliers in BUI

And, the range covering the 99% of the values is:

Original	[2.33, 222.26]
Bottom / Top	[2.33, 202.11]
Tukey	[2.33, 222.26]
Hampel	[2.33, 216.10]

Table D.44: Intervals covering 99% of observation before and after imputing outliers in BUI

Because the percentage of flagged outliers is so low, I am not going to treat any values of this variable as outliers and modify it.

Fire Weather Index (FWI)

As I saw in the data exploration phase, the FWI is not a skewed variable:

Mean	19.13
Standard deviation	11.95
Coefficient of variation	0.62
Skewness	0.30
Kurtosis	2.15
1^{st} percentile	0.55
5^{st} percentile	2.38
25^{st} percentile	8.39
50^{st} percentile	18.51
75^{st} percentile	28.61
95^{st} percentile	38.66
99^{st} percentile	45.58
IQR	20.22
Range with 80% observations	[4.24, 35.16]
Range with 98% observations	[0.55, 45.58]

Table D.45: Central and skewness metrics of FWI

The density plot marking the 95th, 99th, 99.5th percentiles is:

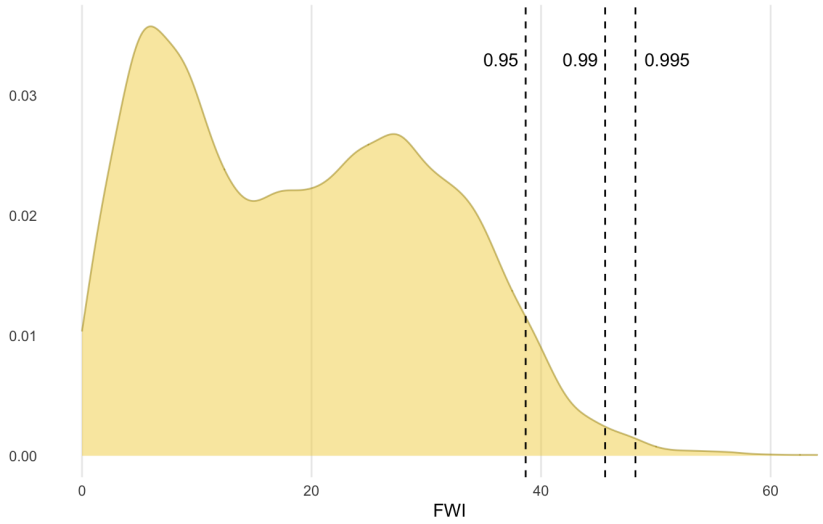


Figure D.14: Density plot of IFWI with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for 1% of the top data
- Tukey method
- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	220	1.02%
Tukey	0	0%
Hampel	5	0.02%

Table D.46: Flagged outliers in FWI

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	19.13	11.95	0.62	0.30	2.15
Bottom / Top	18.81	11.57	0.62	0.20	1.88
Tukey	19.13	11.95	0.62	0.30	2.15
Hampel	19.12	11.93	0.62	0.29	2.12

Table D.47: Skewness metrics before and after imputing outliers in FWI

And, the range covering the 99% of the values is:

Original	$[0.32, 48.22]$
Bottom / Top	$[0.32, 43.96]$
Tukey	$[0.32, 48.22]$
Hampel	$[0.32, 47.96]$

Table D.48: Intervals covering 99% of observation before and after imputing outliers in FWI

Because the percentage of flagged outliers are so low, I am not going to treat any values of this variable as outliers and modify it.

D.1.4 Physiography

Elevation

As I saw in the data exploration phase, the elevation is not a skewed variable. However, the value of the standard deviation is the 73.28% of the mean:

Mean	350.23
Standard deviation	256.66
Coefficient of variation	0.73
Skewness	0.94
Kurtosis	3.36
1^{st} percentile	5.52
5^{st} percentile	31.43
25^{st} percentile	157.73
50^{st} percentile	294.11
75^{st} percentile	487.07
95^{st} percentile	897.36
99^{st} percentile	1,032.59
IQR	329.34
Range with 80% observations	[61.13, 746.58]
Range with 98% observations	[5.52, 1, 032.59]

Table D.49: Central and skewness metrics of elevation

The density plot marking the 95th, 99th, 99.5th percentiles is:

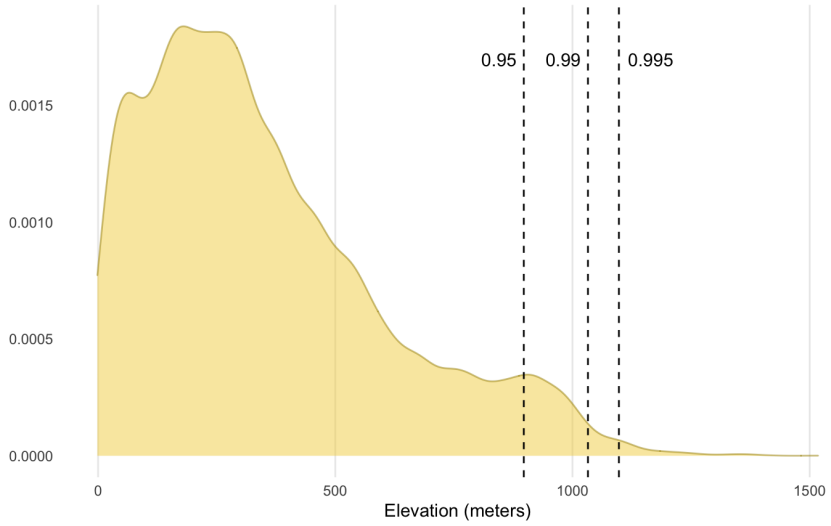


Figure D.15: Density plot of elevation with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for 1% of the top data
- Tukey method
- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	216	1.00%
Tukey	1	0%
Hampel	370	1.71%

Table D.50: Flagged outliers in elevation

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	350.23	256.66	0.73	0.94	3.36
Bottom / Top	342.36	245.46	0.72	0.84	3.03
Tukey	350.18	256.54	0.73	0.94	3.34
Hampel	337.49	239.59	0.71	0.82	2.99

Table D.51: Skewness metrics before and after imputing outliers in elevation

And, the range covering the 99% of the values is:

Original	$[2.60, 1,097.86]$
Bottom / Top	$[2.60, 1,001.92]$
Tukey	$[2.60, 1,097.86]$
Hampel	$[2.60, 979.68]$

Table D.52: Intervals covering 99% of observation before and after imputing outliers in elevation

Because the percentage of flagged outliers are so low, I am not going to treat any values of this variable as outliers and modify it.

Slope

As I saw in the data exploration phase, the BUI has a small tail on the right, but it is not a highly skewed variable. However, the value of the standard deviation is the 78.03% of the mean:

Mean	7.62
Standard deviation	5.94
Coefficient of variation	0.78
Skewness	0.64
Kurtosis	3.36
1^{st} percentile	0
5^{st} percentile	0
25^{st} percentile	5.13
50^{st} percentile	72.5
75^{st} percentile	11.48
95^{st} percentile	18.56
99^{st} percentile	24.77
IQR	6.35
Range with 80% observations	$[0, 15.42]$
Range with 98% observations	$[0, 24.77]$

Table D.53: Central and skewness metrics of slope

The density plot marking the 95th, 99th, 99.5th percentiles is:

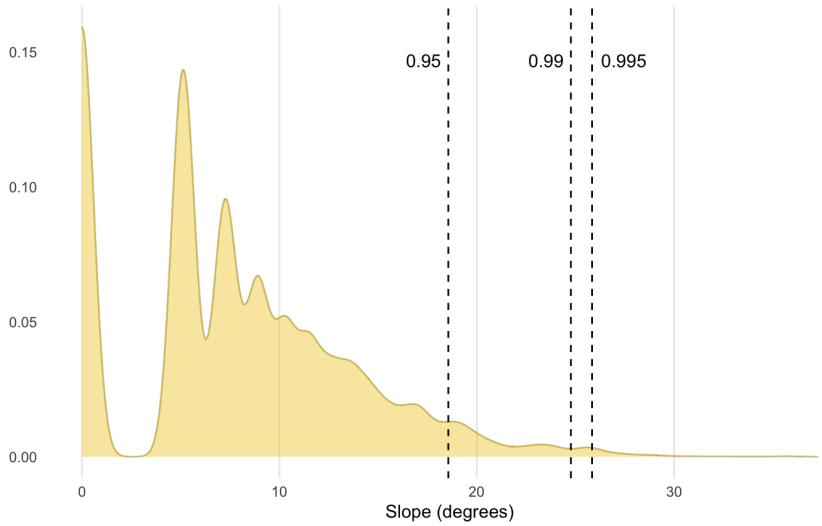


Figure D.16: Sensity plot of slope with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for 1% of the top data
- Tukey method

- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	217	1.01%
Tukey	19	0.09%
Hampel	525	2.43%

Table D.54: Flagged outliers in slope

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	7.62	5.95	0.78	0.64	3.36
Bottom / Top	7.43	5.64	0.76	0.43	2.69
Tukey	7.60	5.90	0.78	0.59	3.15
Hampel	7.20	5.37	0.75	0.29	2.37

Table D.55: Skewness metrics before and after imputing outliers in slope

And, the range covering the 99% of the values is:

Original	[0.00, 25.84]
Bottom / Top	[0.00, 23.07]
Tukey	[0.00, 25.84]
Hampel	[0.00, 19.95]

Table D.56: Intervals covering 99% of observation before and after imputing outliers in slope

Because the percentage of flagged outliers are so low, I am not going to treat any values of this variable as outliers and modify it.

D.1.5 Human factors

Distance to closest road

As I saw in the data exploration phase, the distance to closest road is skewed variable with a tail to the right. This is also evident in how the value of the standard deviation is higher than the mean:

Mean	347.38
Standard deviation	447.67
Coefficient of variation	1.29
Skewness	2.96
Kurtosis	16.13
1^{st} percentile	1.55
5^{st} percentile	8.90
25^{st} percentile	70.70
50^{st} percentile	192.76
75^{st} percentile	436.13
95^{st} percentile	1,233.49
99^{st} percentile	2,143.69
IQR	365.43
Range with 80% observations	$[21.31, 868.34]$
Range with 98% observations	$[1.55, 2, 143.69]$

Table D.57: Central and skewness metrics of distance to closest road

The density plot marking the 95th, 99th, 99.5th percentiles is:

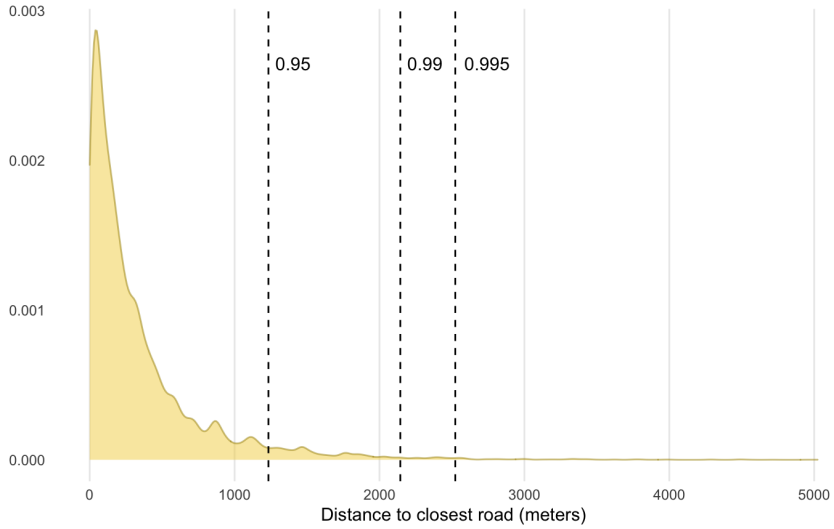


Figure D.17: Density plot of distance to closest road with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for 5% of the top data
- Tukey method
- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	1,080	5.00%
Tukey	599	2.77%
Hampel	2,272	10.52%

Table D.58: Flagged outliers in distance to nearest road

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	347.38	447.68	1.29	2.96	16.13
Bottom / Top	270.50	272.33	1.01	1.42	4.50
Tukey	296.02	316.91	1.07	1.65	5.46
Hampel	224.67	204.47	0.91	1.09	3.43

Table D.59: Skewness metrics before and after imputing outliers in distance to nearest road

And, the range covering the 95% of the values is:

Original	$[4.71, 1,602.11]$
Bottom / Top	$[4.35, 1,040.75]$
Tukey	$[4.56, 1,208.85]$
Hampel	$[4,01,739.73]$

Table D.60: Intervals covering 99% of observation before and after imputing outliers in distance to nearest road

Before the standard deviation was large compared to the mean. After, both methods have reduced all the metrics.

Comparing graphically the outliers in original variable and all methods:

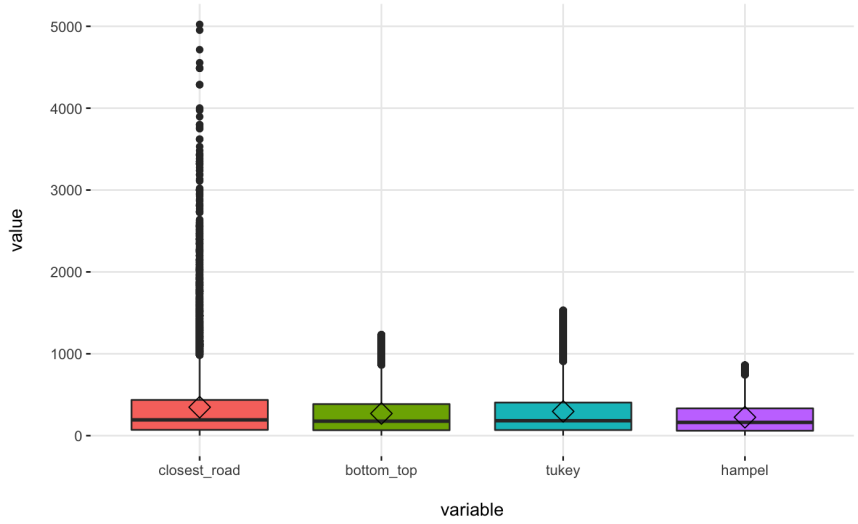


Figure D.18: Box plots before and after imputing outliers in distance to nearest road

And, comparing only the result of the three methods:

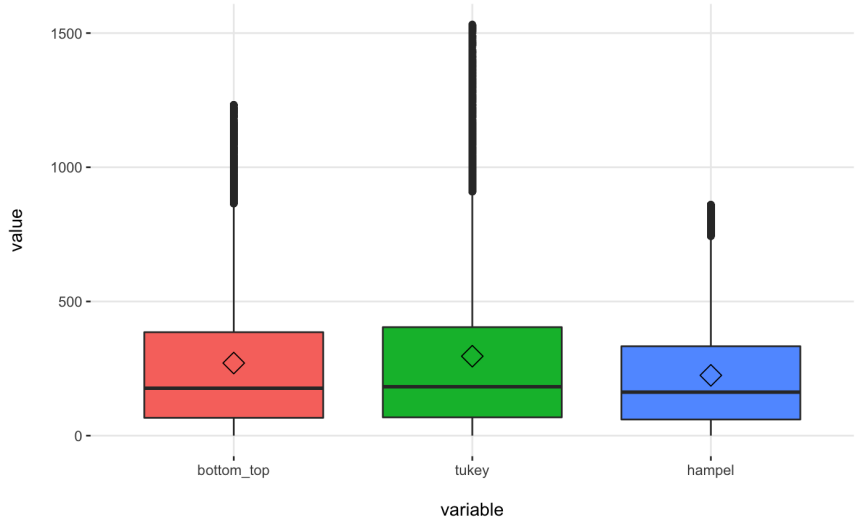


Figure D.19: Box plots after imputing outliers in distance to nearest road

The distribution of the new variables is still skewed but with a much smaller tail. I will use the Tukey method that imputes the least number of observations of all methods but still offers a good improvement of the skewness indicators.

The density plot for the distance to the closest road treated with the Tukey method is:

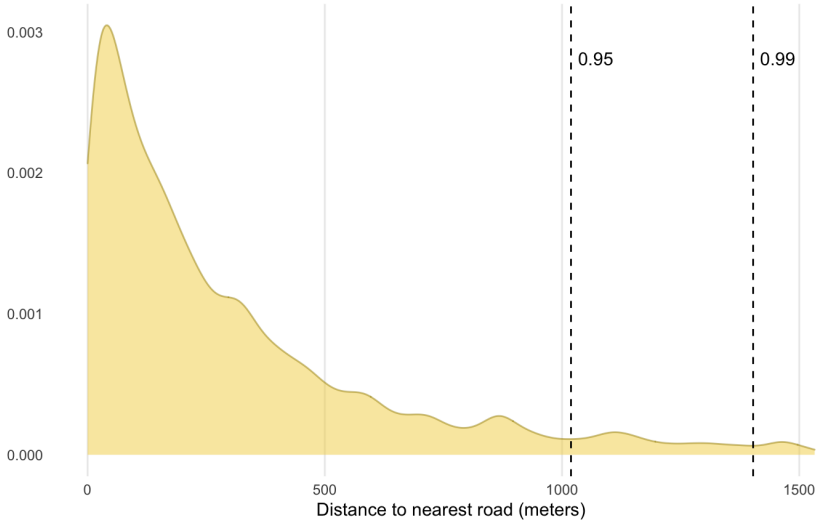


Figure D.20: Density plot of distance to closest road after applying the Tukey method

Distance to closest building

As I saw in the data exploration phase, the burnt area is a highly skewed variable. This is also evident in how the values of the standard deviation is much higher than the mean:

Mean	748.96
Standard deviation	859.41
Coefficient of variation	1.15
Skewness	2.28
Kurtosis	10.77
1^{st} percentile	0
5^{st} percentile	9.54
25^{st} percentile	131.19
50^{st} percentile	469.28
75^{st} percentile	1,080.42
95^{st} percentile	2,427/14
99^{st} percentile	3.964.60
IQR	949.23
Range with 80% observations	$[29.52, 1, 775.96]$
Range with 98% observations	$[0, 3, 964.60]$

Table D.61: Central and skewness metrics of distance to closest building

The density plot marking the 95th, 99th, 99.5th percentiles is:

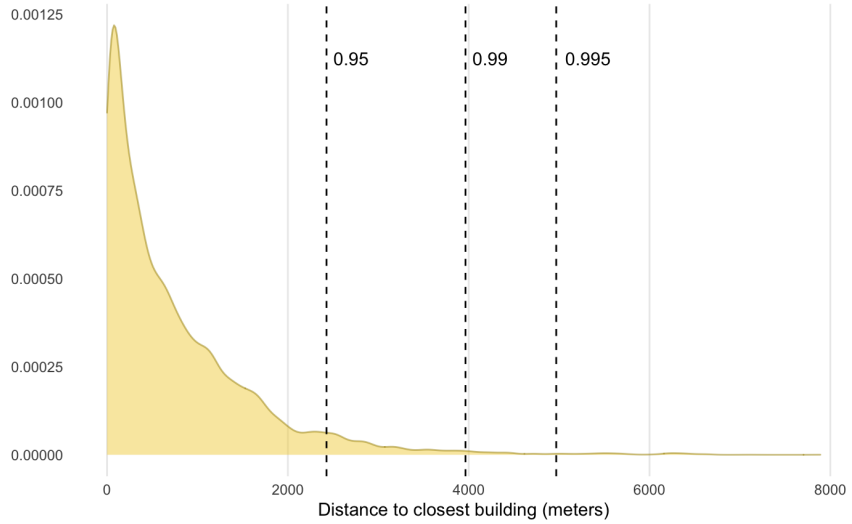


Figure D.21: Distance to nearest building density plot with marked percentiles

I am going to detect the outliers with the using the following configuration for each method:

- Bottom/Top method for 5% of the top data
- Tukey method
- Hampel method using k = 3

The number and percentage of flagged outliers are:

Method	Outliers	Percentage
Bottom / Top	1,084	5.02%
Tukey	230	1.07%
Hampel	1,394	6.46%

Table D.62: Flagged outliers in distance to nearest building

And, the skewness indicators before and after are:

Method	Mean	Standard deviation	Coefficient & of variation	Skewness	Kurtosis
Original	748.96	859.41	1.15	2.28	10.77
Bottom / Top	610.08	583.14	0.96	1.04	3.24
Tukey	703.12	735.50	1.05	1.50	5.19
Hampel	583.88	547.56	0.94	0.95	2.95

Table D.63: Skewness metrics before and after imputing outliers in distance to nearest building

And, the range covering the 95% of the values is:

Original	$[2.40, 3,014.66]$
Bottom / Top	$[2.20, 2,046.08]$
Tukey	$[2.40, 2,712.13]$
Hampel	$[2.17, 1,893.56]$

Table D.64: Intervals covering 99% of observation before and after imputing outliers in distance to nearest building

Before the standard deviation was large compared to the mean, as reflected in the variation coefficient and kurtosis value. After, both methods have reduced all the metrics.

Comparing graphically the outliers in the original variable and all methods:

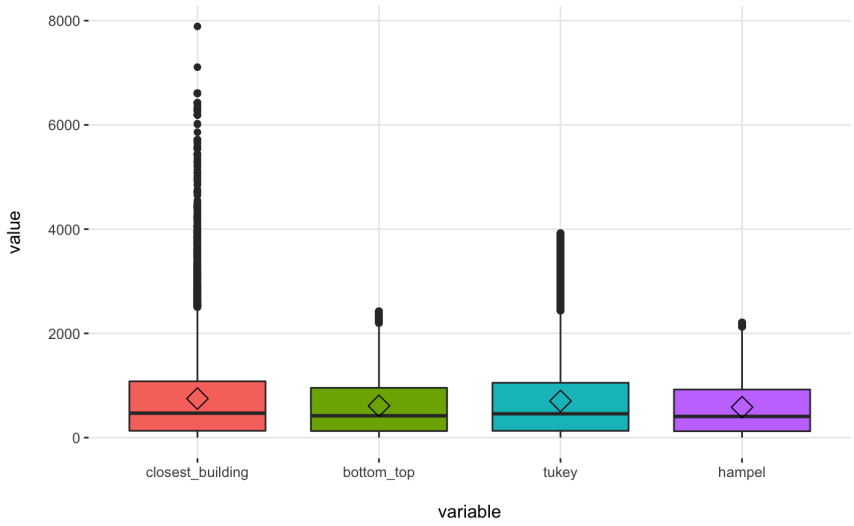


Figure D.22: Box plots before and after imputing outliers in distance to nearest building

And, comparing only the result of the three methods:

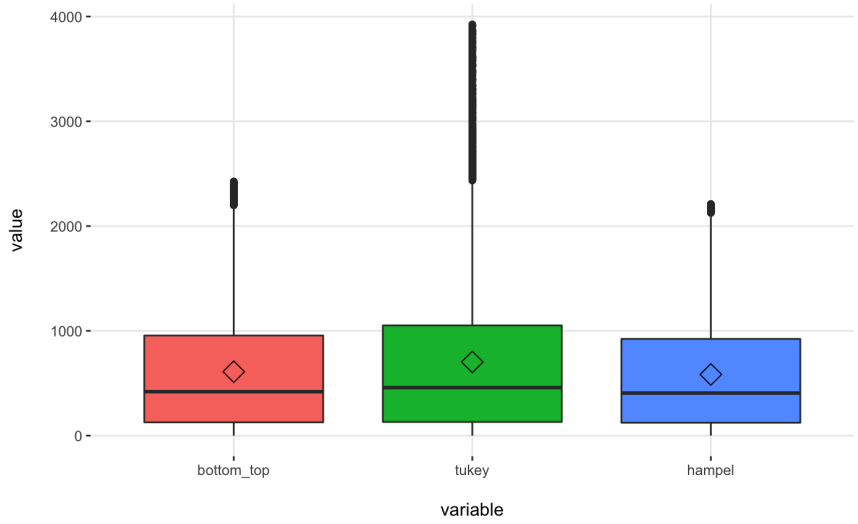


Figure D.23: Box plots after imputing outliers in distance to nearest building

The distribution of the new variables is still skewed but with a much smaller tail.

I will use the Tukey method that imputes the least number of observations of all methods but still offers a good improvement of the skewness indicators.

The density plot for the distance to closest building treated with the Tukey method:

The non-numerical variables are data_alerta, and land_cover. In the case of data_alerta I will not apply any transformation since I am going to use data_alerta indirectly through a new variable generated from it.

For the land_cover variable I am going to transform it using one-hot encoding so I can use it with machine-learning algorithms that do not support this type of variables.

But before, due to the variable having only 4 categories with more than 10% of the observations and to avoid the danger of a noisy variable and overfitting I am going to swap the CORINE Land Cover level 3 classification for the corresponding CORINE Land Cover level 2.

After converting the categories, the frequencies and percentages ordered by their frequency are:

There is now a category that concentrates the 47.15% of all observations, "Heterogeneous agricultural areas." And, only 4 categories have more than 6% of the observations each one and approximately 90% of a total of 14 categories

The reduction from the codification using CLC level 3 has improved but there is still not satisfactory. To finish I am going to aggregate all categories

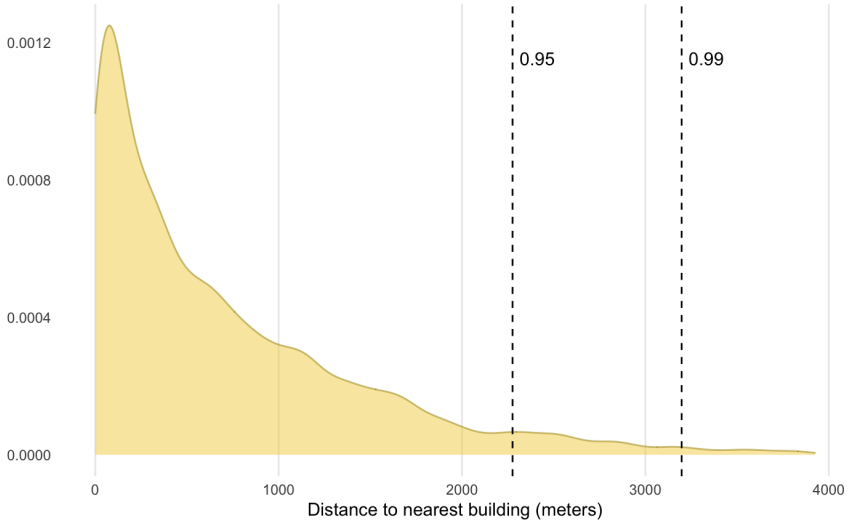


Figure D.24: Density plot of distance to closest building after applying the Tukey method

Category	Frequency	Percentage
Heterogeneous agricultural areas	10,180	47.15%
Urban fabric	3.434	15.91%
Scrub and/or herbaceous vegetation associations	2,943	13.63%
Forest	2,720	12.60%
Arable land	1,191	5.52%
...	...	...

Table D.65: Topmost CLC level 2 categories by number of observations

less than a 6% of the observations that represent approximately 10% of all observations. However, the new category will still be the one with less observations.

The final categories are:

Category	Frequency	Percentage
Heterogeneous agricultural areas	10,180	47.15%
Urban fabric	3.434	15.91%
Scrub and/or herbaceous vegetation associations	2,943	13.63%
Forest	2,720	12.60%
Other	2,313	10.71%

Table D.66: CLC level 2 categories by number of observations after grouping smaller ones

There all only 5 categories, all with al least 10% of the observations. However, now the situation can be dangerous by having less variability and existing the possibility of underfitting.

D.2 Transformations

I am going to drop the variable biome after selecting the data for a single biome to build the models.

Apart from that, the only transformation I am going to perform is to generate a new variable indicating the day of the year, called yday. This new variable may help to capture the fact that early spring and summer are the two periods of the year where the human-caused fires (HCFs) tend to more prevalent.

Bibliography

Modelling

Crawley, Michael J. *Statistics: An Introduction Using R*. 2nd. Wiley, 2014. 354 pp. ISBN: 978-1-118-94109-6.

European Commission. *Monitoring Agricultural ResourceS (MARS)*. EU Science Hub. June 26, 2015. URL: https://ec.europa.eu/jrc/en/mars (visited on 04/20/2020).

Farr, Tom G. et al. "The Shuttle Radar Topography Mission". In: *Reviews of Geophysics* 45.2 (2007). DOI: 10.1029/2005RG000183. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2005RG000183. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005RG000183.

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously". In: (Dec. 23, 2019). arXiv: 1801.01489. URL: http://arxiv.org/abs/1801.01489 (visited on 05/10/2020).

Fisher, P., A.J. Coomber, and R.A. Wadsworth. *Land Use and Land Cover: Contradiction Or Complement*. 2005.

Friedman, Jerome. "Greedy Function Approximation: A Gradient Boosting Machine". In: *The Annals of Statistics* 29 (Nov. 28, 2000). DOI: 10.1214/aos/1013203451.

Groot, William J. de. "Interpreting the Canadian Forest Fire Weather Index (FWI) System". In: *Fourth Central Regional Fire Weather Committee Scientific and Technical Seminar*. Apr. 1987, pp. 3–14.

Hampel, Frank R. "The Influence Curve and Its Role in Robust Estimation". In: *Journal of the American Statistical Association* 69.346 (1974), pp. 383–393. ISSN: 0162-1459. DOI: 10.2307/2285666. URL: https://www.jstor.org/stable/2285666 (visited on 05/03/2020).

Klokan Technologies GmbH. *Lisbon (Lisbon) / Portuguese National Grid - EPSG:20790*. URL: http://epsg.io/20790 (visited on 05/13/2020).

— *WGS84 - World Geodetic System 1984, used in GPS - EPSG:4326*. URL: https://epsg.io/4326 (visited on 05/13/2020).

Kohavi, Ron and George H. John. "Wrappers for feature subset selection". In: *Artificial Intelligence*. Relevance 97.1 (Dec. 1, 1997), pp. 273–324. ISSN: 0004-3702. DOI: 10.1016/S0004-3702(97)00043-X. URL: http://www.sciencedirect.com/science/article/pii/S000437029700043X (visited on 05/06/2020).

Marchal, Jean, Steve G. Cumming, and Eliot J. B. McIntire. "Land cover, more than monthly fire weather, drives fire-size distribution in Southern Québec forests: Implications for fire risk management". In: *PLOS ONE* 12.6 (June 13, 2017). Publisher: Public Library of Science. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0179294. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0179294 (visited on 04/23/2020).

Microsoft Corporation. *Team Data Science Process Documentation*. Library Catalog: docs.microsoft.com. 2020. URL: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/ (visited on 05/16/2020).

Naqvi, Ghayur. "A Hybrid Filter-Wrapper Approach for FeatureSelection". PhD thesis. Örebro University, 2011. 104 pp. URL: http://oru.diva-portal.org/smash/record.jsf?pid=diva2%3A567115&dswid=8063 (visited on 05/06/2020).

OpenStreetMap Wiki contributors. *Highways*. URL: https://wiki.openstreetmap.org/w/index.php?title=Highways&oldid=1905430 (visited on 05/14/2020).

— *Key: highway*. URL: https://wiki.openstreetmap.org/w/index.php?title=Key:highway&oldid=1964792 (visited on 05/14/2020).

Quinlan, J. R. "Induction of decision trees". In: *Machine Learning* 1.1 (Mar. 1, 1986), pp. 81–106. ISSN: 1573-0565. DOI: 10.1007/BF00116251. URL: https://doi.org/10.1007/BF00116251 (visited on 05/05/2020).

República Portuguesa. *Instituto da Conservação da Natureza e das Florestas (ICNF)*. Library Catalog: www2.icnf.pt. 2020. URL: http://www2.icnf.pt/portal (visited on 05/13/2020).

Reshef, David N. et al. "Detecting Novel Associations in Large Data Sets". In: *Science* 334.6062 (Dec. 16, 2011). Publisher: American Association for the Advancement of Science Section: Research Article, pp. 1518–1524. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1205438. URL: https://science.sciencemag.org/content/334/6062/1518 (visited on 05/06/2020).

Sebastián-López, Ana et al. "Integration of socio-economic and environmental variables for modelling long-term fire danger in Southern Europe". In: *European Journal of Forest Research* 127.2 (Mar. 2008), pp. 149–163. ISSN: 1612-4669, 1612-4677. DOI: 10.1007/s10342-007-0191-5. URL: http://link.springer.com/10.1007/s10342-007-0191-5 (visited on 04/23/2020).

Smart Vision Europe. *What is the CRISP-DM methodology?* Library Catalog: www.sv-europe.com. 2020. URL: https://www.sv-europe.com/crisp-dm-methodology/ (visited on 05/16/2020).

Tadist, Khawla et al. "Feature selection methods and genomic big data: a systematic review". In: *Journal of Big Data* 6.1 (Aug. 27, 2019), p. 79. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0241-0. URL: https://doi.org/10.1186/s40537-019-0241-0 (visited on 05/16/2020).

Vasilakos, Christos et al. "Identifying wildland fire ignition factors through sensitivity analysis of a neural network". In: *Natural Hazards* 50 (July 1, 2008), pp. 125–143. DOI: 10.1007/s11069-008-9326-3.

Yang, Jian, Hong S. He, and Stephen R. Shifley. "Spatial controls of occurrence and spread of wildfires in the Missouri Ozark highlands". In: *Ecological Applications* 18.5 (2008), pp. 1212–1225. DOI: 10.1890/07-0825.1. eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/07-0825.1. URL: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/07-0825.1.

Zhang, Yang, Samsung Lim, and Jason Sharples. "Modelling spatial patterns of wildfire occurrence in South-Eastern Australia". In: *Geomatics, Natural Hazards and Risk* 7 (Mar. 3, 2016), pp. 1–16. DOI: 10.1080/19475705.2016.1155501.

eXplainable AI

Apley, Daniel W. and Jingyu Zhu. "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models". In: (Aug. 19, 2019). arXiv: 1612.08468. URL: http://arxiv.org/abs/1612.08468 (visited on 05/10/2020).

Barredo Arrieta, Alejandro et al. *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.* Oct. 22, 2019.

Burrell, Jenna. "How the machine 'thinks': Understanding opacity in machine learning algorithms". In: *Big Data & Society* 3.1 (2016). DOI: 10.1177/2053951715622512. eprint: https://doi.org/10.1177/2053951715622512. URL: https://doi.org/10.1177/2053951715622512.

Doran, Derek, Sarah Schulz, and Tarek R. Besold. "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives". In: *arXiv* (Oct. 2, 2017). arXiv: 1710.00794. URL: http://arxiv.org/abs/1710.00794 (visited on 04/04/2020).

European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and content ealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).* Legislative Body: EP, CONSIL

Library Catalog: EUR-Lex. May 4, 2016. URL: http://data.europa.eu/eli/reg/2016/679/oj/eng (visited on 04/11/2020).

Gilpin, Leilani H. et al. "Explaining Explanations: An Overview of Interpretability of Machine Learning". In: *arXiv* (Feb. 3, 2019). arXiv: 1806.00069. URL: http://arxiv.org/abs/1806.00069 (visited on 04/07/2020).

Gunning, David. *DARPA's explainable artificial intelligence (XAI) program*. Association for Computing Machinery, 2019. DOI: https://doi.org/10.1145/3301275.3308446.

Hamon, Ronan et al. *Robustness and explainability of Artificial Intelligence: from technical to policy solutions*. OCLC: 1140718927. 2020. ISBN: 978-92-76-14660-5. URL: https://op.europa.eu/publication/manifestation_identifier/PUB_KJNA30040ENN (visited on 04/10/2020).

Holzinger, Andreas et al. "What do we need to build explainable AI systems for the medical domain?" In: *arXiv* (Dec. 28, 2017). arXiv: 1712.09923. URL: http://arxiv.org/abs/1712.09923 (visited on 04/04/2020).

Lipton, Zachary C. "The Mythos of Model Interpretability". In: *arXiv:1606.03490 [cs, stat]* (Mar. 6, 2017). arXiv: 1606.03490. URL: http://arxiv.org/abs/1606.03490 (visited on 04/04/2020).

Molnar, Christoph. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/. June 29, 2020.

National Wildfire Coordination Group. *Fire Weather Index (FWI) System*. 2020. URL: https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system (visited on 07/05/2020).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.

Roff, Heather M. and Richard Moyes. "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons". In: Briefing paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons. Arizona State University. Article 36, Apr. 2016. URL: http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf (visited on 03/30/2020).

Rudin, Cynthia. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: *arXiv* (Sept. 21, 2019). arXiv: 1811.10154. URL: http://arxiv.org/abs/1811.10154 (visited on 04/10/2020).

Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller. "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models". In: *arXiv* (Aug. 28, 2017). arXiv: 1708.08296. URL: http://arxiv.org/abs/1708.08296 (visited on 04/05/2020).

Schwarz, Elke. *The (im)possibility of meaningful human control for lethal autonomous weapon systems*. International Committee of the Red Cross. Aug. 29, 2018. URL: https://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems/ (visited on 05/13/2020).

Shapley, L. S. "A Value for n-Person Games". In: *Contributions to the Theory of Games (AM-28)*. Vol. 2. Princeton University Press, 1953, pp. 307–318.

Sharkey, Noel. "Guidelines for the Human Control of Weapons Systems". In: Working paper for Convention on Certain Conventional Weapons. International Committee for Robot Arms Control. Apr. 2018. URL: https://www.icrac.net/wp-content/uploads/2018/04/Sharkey_Guideline-for-the-human-control-of-weapons-systems_ICRAC-WP3_GGE-April-2018.pdf (visited on 03/30/2020).

Sheridan, T.B. "Function allocation: algorithm, alchemy or apostasy?" In: *International Journal of Human-Computer Studies* 52.2 (2000), pp. 203–216. ISSN: 1071-5819. DOI: https://doi.org/10.1006/ijhc.1999.0285. URL: http://www.sciencedirect.com/science/article/pii/S1071581999902859.

Staniak, Mateusz and Przemyslaw Biecek. "Explanations of model predictions with live and breakDown packages". In: *R J.* 10 (2018), p. 395.

Strumbelj, Erik and Igor Kononenko. "An Efficient Explanation of Individual Classifications Using Game Theory". In: *J. Mach. Learn. Res.* 11 (Mar. 2010), pp. 1–18. ISSN: 1532-4435.

Wagner, Ben. "Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems". In: *Policy & Internet* 11.1 (2019), pp. 104–122. DOI: 10.1002/poi3.198. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.198. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.198.

Human-Caused Fires (HCFs)

Beighley, Mark and A.C. Hyde. "Portugal Wildfire Management in a New Era: Assessing Fire Risks, Resources and Reforms". In: (2018).

Costafreda Aumedes, Sergi. "Spatio-temporal analysis of human-caused fire occurrence patterns in Spain". Accepted: 2017-02-28T09:16:48Z Publication Title: TDX (Tesis Doctorals en Xarxa). PhD thesis. Universitat de Lleida, Feb. 17, 2017. URL: http://www.tdx.cat/handle/10803/400822 (visited on 04/11/2020).

Food and Agriculture Organization of the United Nations. *Global Forest Resources Assessment 2005: progress towards sustainable forest management*. FAO forestry paper 147. OCLC: 255422261. Rome, 2006. 320 pp. ISBN: 978-92-5-105481-9.

— *Global Forest Resources Assessment 2006: A thematic study prepared in the framework of the Global Forest Resources Assessment 2005*. FAO forestry paper 151. Rome, 2007. 135 pp. ISBN: 978-92-5-105666-0.

Lierop, Pieter van et al. "Global forest area disturbance from fire, insect pests, diseases and severe weather events". In: *Forest Ecology and Management* 352 (Sept. 2015), pp. 78–88. ISSN: 03781127. DOI: 10.1016/j.foreco.2015.06.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378112715003369 (visited on 04/18/2020).

Moritz, Max A. et al. "Climate change and disruptions to global fire activity". In: *Ecosphere* 3.6 (2012). _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/ES11-00345.1. ISSN: 2150-8925. DOI: 10.1890/ES11-00345.1. URL: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/ES11-00345.1 (visited on 04/13/2020).

Natural Resources Canada. *Canadian Forest Fire Weather Index (FWI) System*. URL: https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi (visited on 04/18/2020).

Olson, David et al. "Terrestrial Ecoregions of the World: A New Map of Life on Earth". In: *BioScience* 51 (Nov. 1, 2001), pp. 933–938. DOI: 10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2.

Pereira, Mário et al. "Synoptic patterns associated with large summer forest fires in Portugal". In: *Agricultural and Forest Meteorology* 129 (Mar. 1, 2005), pp. 11–25. DOI: 10.1016/j.agrformet.2004.12.007.

Turco, Marco et al. "Climate drivers of the 2017 devastating fires in Portugal". In: *Scientific Reports* 9.1 (Oct. 10, 2019), pp. 1–8. ISSN: 2045-2322. DOI: 10.1038/s41598-019-50281-2. URL: https://www.nature.com/articles/s41598-019-50281-2 (visited on 04/18/2020).

Software

Biecek, Przemyslaw. *DALEX: moDel Agnostic Language for Exploration and eXplanation*. Version 1.2.0. Apr. 22, 2020. URL: https://modeloriented.github.io/DALEX/index.html.

— *ingredients: Effects and Importances of Model Ingredients*. Version 1.2.0. Apr. 22, 2020. URL: https://modeloriented.github.io/ingredients/.

Casas, Pablo. *funModeling: Exploratory Data Analysis and Data Preparation Tool-Box*. Version 1.9.3. Oct. 9, 2019. URL: https://livebook.datascienceheroes.com.

Dowle, Matt. *data.table: Extension of data.frame*. Version 1.12.8. Dec. 8, 2019. URL: http://r-datatable.com.

Filosi, Michele. *minerva: Maximal Information-Based Nonparametric Exploration for Variable Analysis*. Version 1.5.8. May 24, 2019. URL: http://www.exploredata.net.

Git Maintainers. *Git: Fast, Scalable, Distributed Revision Control System*. Version 2.26.2. The Git Project, Apr. 20, 2020. URL: https://rmarkdown.rstudio.com.

H2O Dev Team. *H2O: Fast Scalable Machine Learning For Smarter Applications*. Version 3.30.0.1. H2O.ai, Apr. 3, 2020. URL: https://www.h2o.ai.

Hijmans, Robert J. *raster: Geographic Data Analysis and Modeling*. Version 3.1-5. Apr. 18, 2020. URL: https://rspatial.org/raster.

Kassambara, Alboukadel. *ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'*. Version 0.1.3. May 19, 2019. URL: http://www.sthda.com/english/wiki/ggcorrplot.

Lin, Thomas. *patchwork: The Composer of Plots*. Version 1.0.0. Dec. 1, 2019. URL: https://patchwork.data-imaginist.com.

MacTeX TeXnical working group. *MacTeX*. Version 2020. TeX Users Group (TUG), Apr. 8, 2020. URL: https://rmarkdown.rstudio.com.

Milton Bache, Stefan. *magrittr: A Forward-Pipe Operator for R*. Version 1.5. Nov. 22, 2014. URL: https://magrittr.tidyverse.org.

Pebesma, Edzer. *sf: Simple Features for R*. Version 0.9-2. Apr. 14, 2020. URL: https://r-spatial.github.io/sf/.

PostGIS Project Steering Committee (PSC). *PostGIS: Spatial and Geographic objects for PostgreSQL*. Version 3.0.0. Open Source Geospatial Foundation Project, Oct. 20, 2019. URL: https://postgis.net.

PostgreSQL Development Team. *PostgreSQL Open Source Object-relational Database System*. Version 12.1. The PostgreSQL Global Development Group, Nov. 14, 2019. URL: https://www.postgresql.org.

QGIS Development Team. *QGIS Geographic Information System*. Version 3.10.2. Open Source Geospatial Foundation Project, Jan. 23, 2020. URL: http://qgis.osgeo.org/.

R Core Team. *R: A Language and Environment for Statistical Computing*. Version 3.6.1. R Foundation for Statistical Computing, July 5, 2019. URL: https://www.r-project.org.

Ripley, Brian. *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. Version 7.3-51.5. Dec. 20, 2019. URL: http://www.stats.ox.ac.uk/pub/MASS4/.

RStudio Team. *RStudio: Integrated Development Environment for R*. Version 1.2.5033. RStudio PBC, Dec. 4, 2019. URL: https://rstudio.com.

Spinu, Vitalie. *lubridate: Make Dealing with Dates a Little Easier*. Version 1.7.8. Apr. 3, 2020. URL: http://lubridate.tidyverse.org.

SQLite Development Team. *SQLite*. Version 3.31.1. Jan. 27, 2020. URL: https://www.sqlite.org/.

Sumner, Michael. *ncmeta: Straightforward 'NetCDF' Metadata*. Version 0.2.0. Oct. 22, 2019. URL: https://github.com/hypertidy/ncmeta.

Sumner, Michael. *tidync: A Tidy Approach to 'NetCDF' Data Exploration and Extraction*. Version 0.2.3. rOpenSci, Nov. 6, 2019. URL: https://docs.ropensci.org/tidync/.

Wickham, Hadley. *dplyr: A Grammar of Data Manipulation*. Version 0.8.5. Mar. 5, 2020. URL: http://dplyr.tidyverse.org.

— *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. Version 3.3.0. Mar. 5, 2020. URL: http://ggplot2.tidyverse.org.

— *reshape2: Flexibly Reshape Data*. Version 1.4.4. Apr. 9, 2020. URL: https://github.com/hadley/reshape.

— *scales: Scale Functions for Visualization*. Version 1.1.0. Dec. 4, 2019. URL: https://scales.r-lib.org.

— *tidyr: Tidy Messy Data*. Version 1.0.2. Jan. 23, 2020. URL: https://tidyr.tidyverse.org.

Xie, Yihui. *bookdown: Authoring Books and Technical Documents with R Markdown*. Version 0.17. Mar. 5, 2020. URL: https://github.com/rstudio/bookdown.

— *knitr: A General-Purpose Package for Dynamic Report Generation in R*. Version 1.27. Feb. 6, 2020. URL: https://yihui.org/knitr/.

— *rmarkdown: Dynamic Documents for R*. Version 2.1. Jan. 20, 2020. URL: https://rmarkdown.rstudio.com.

Zeileis, Achim. *colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes*. Version 1.4-1. Mar. 18, 2019. URL: http://colorspace.R-Forge.R-project.org.