UNIVERSITY OF TORINO

DIPARTIMENTO DI MATEMATICA GIUSEPPE PEANO

SCUOLA DI SCIENZE DELLA NATURA

M.Sc. in Stochastics and Data Science

Final dissertation

Human-Caused Wildfire Ignition Probability Estimation

Supervisor: prof. Marco Grangetto
Co-supervisor: Andrea Bragagnolo, Ph.D

Candidate: Simone Genovese

ACADEMIC YEAR 2024/2025

Contents

Chapter 0

Introduction

Although they can be triggered by natural events such as lightnings, human-caused wildfires account for the 85-90% of ignitions worldwide[1]. These include either intentional and unintentional fires. Uncontrolled wildfires threaten human assets as well as the local biosphere. For this reason, being able to evaluate the fire risk of vulnerable areas is a key point for efficiently preventing the damage.

This thesis proposes a statistical approach to the problem of the prediction of the human-caused ignitions probability through the tools of Machine Learning, based on historical fire detections in the region of Piedmont, Italy. The goal is not only to build a model that can be useful for future predictions, but also to look for patterns and filter the data actually useful for prediction, hence a sufficient level of interpretability of the model is desirable. The National Agency of Environmental Protection (ARPA) reported as the main causes of forest fires agricultural activities, cultural abandonment and negligence and hunting activities among the unintentional ones, arson, vandalism, building speculation and waste elimination among the intentional ones[2]. These motivations are effected by socioeconomic conditions of the population.

The first chapter is dedicated to a review of the scientific literature on the topic of fire risk, and specifically about ignition. A theoretical section to define the problem is required since the targets can be diverse and each term in environmental sciences can cover different aspects of the phenomenons due to the strong interdependence of the human and natural systems. The results of the papers have been mainly searched among the publications of the European FirEUrisk project which gives the baselines for the modelization and suggestions on the type of data and procedure to use. In the second chapter the study method of this thesis is explained in detail, from the data gathering to the models' training and evaluation. It contains also a small theoretical excursus on a spatial interpolation technique known as Kriging that determines a core step of the workflow, since it gives the target in its complete form. Three different datasets have been used since the observations had to be sampled and the techniques involved showed some limitations. The model chosen is the Random Forest algorithm whose strengths satisfy the model requirements previously described. Chapter 3 concludes with the analysis of the models' results and data patterns correlated to the target probability, and a brief comparison between similar previous work and this thesis is carried out. The topic addressed remains open and this work is intended as a step for the improvement of the wildfires prevention techniques in Piedmont.

This thesis was made in collaboration with LINKS Foundation - AI, Data & Space division contextually to a curricular internship at the end of the Master's degree course. Special thanks to Ph.D. Andrea Bragagnolo as the internship's tutor and to Nicola Bavaro, student of the Politecnico di Torino for the joint work in data collection.

Chapter 1

State of the art

In 2021 the European Commission's annual report of forest fires recorded a total of over 5500 km^2 burnt land (of which 1000 km^2 are in protected areas) in 22 out of 27 countries[3]. The increase of the strength and frequency of fire events in the new millennium, also caused by climate changes, awakened the interest of monitoring and studying the phenomenon. As a threat to biodiversity, people and economy, the policy to face the problem in the New EU Forest Strategy for 2030 (Brussels, 16/07/2021) highlights the necessity to guarantee resilience of forests against wildfires, pests and diseases together with the creation of positive spill over effects due to risk management practices. Speaking of wildfires specifically, EU consistently contributed to the project named FirEUrisk whose objective was «to develop, evaluate and disseminate a science-based integrated strategy to: 1) expand current wildland fire risk assessment systems, including critical factors of risk previously not covered; 2) produce effective measures to reduce current fire risk conditions, and 3) adapt management strategies to expected future climate and socio-economic changes»[4]. The project counts over 60 scientific publications that cover several aspects of the phenomenon, but most of all define a baseline to conduct researches from. A relevant problem to face is to make up a predictive tool that can be employed for the definition of intervention strategies by local authorities dealing with wildfires.

1.1 Definition of Risk

In the late 1960's the Canadian Forest Service developed a well known estimation of fire risk currently used in several countries: the Fire Weather Index (FWI[5]), whose method and definition is currently kept up to date. It is composed of a combination of fuel moisture, wind, temperature, precipitation and humidity indexes and maps the risk of fire on a given region. Though, the concept of *risk* can be vague if not properly defined, hence a good displacement of its parts is needed. The FirEUrisk project gives an in-depth analysis of the various components and sub-components which allows to focus on one task at a time. Risk is assessed by integrating the information determined by 3 macro components: danger, exposure and vulnerability. The European Forest Fire Information System (EFFIS[6]) includes the three components in its estimations together with the final risk mapping.

Danger Also known as Hazard (Danger is more commonly used in the wildfire management community), its assessment considers human and natural causes leading to a fire ignition, as well as those factors that must be present to start a fire or affect fire behavior, including fuel availability and moisture status, slope and weather conditions. Fire Ignition refers to the fire occurrence itself. It can be divided into natural and human ignition (either accidental and intentional), defined according to the main fire causes. The cause is identified as the heat source, which can start a fire together with the favorable presence of fuels and oxygen. The propagation phase instead counts as a different component of the danger since it depends on humidity and wind conditions as well as the morphology of the terrain.

This work puts its focus specifically on the ignition part.

Exposure Exposure indicates the extent to which people, infrastructures and other tangible human assets, as well as ecosystems, could be affected by wildfires. Exposition does not only refer to direct contact with the fire front but also indirectly through the dispersion of smoke, or by fire-caused changes in hydrological cycles or soil erosion. The exposure component forms a link between danger and vulnerability. Studies on wildfire exposure are often based on the intersection of predictions of fire behavior models with the endangered elements (assets, people, woodlands...).

Vulnerability Intuitively speaking, a vulnerable territory would be an area with high values that in case will be lost by the effects of fire, together with the weak ability to resume those values. Vulnerability refers to the potential damages caused by wildfires on a particular territory, including the losses directly caused by fires but also the degree of ability to recover afterwards. The approach perspectives are several since the objects of potential fire damages varies among human and ecological assets. From an ecological point of view, vulnerability has been defined as the interaction among exposure, resilience (also named sensitivity) and recovery potential (also called adaptive capacity). Societal vulnerability to wildfires may be identified as a combination of the magnitude of the socioeconomic impacts deriving from wildfires, and the inability of local societies to adapt to the consequences of the exposure to wildfires.

1.2 The problem of the target and the estimation techniques

Scientific literature spreads among different approaches to the wildfire prediction problem. The main target itself varies as the data availability and application purposes change. Speaking of the ignition part, estimates are based on historical data. This includes either point-wise fires (ignition point on given coordinates) or aggregated fires within regular grids (pixels, quadrants...) or within irregular administrative divisions (area units such as districts, provinces, townships). Furthermore one can base the estimation on the whole burnt area, which is reported either by local institutions and by remote sensing tools. There are available datasets of past fires and total burnt area at European scale such as in the EFFIS database and at world scale such as the ONFIRE[7] dataset. The limitations of these methods lie on the spatial resolution and time span[8], but most of all on the problem of the *negative sampling* occurring when target data has the form of ignition points: the absence of fire events must be someway recorded to train the discrimination ability of the algorithms. In general, regression methods are widely applied since they are easy to understand and simple, but there's evidence of suffering from multicollinearity[9]. Complex techniques such as Classification and Regression Trees, Artificial Neural Networks, Support Vector Machines or Generalized Additive Models have been introduced as alternatives to traditional statistical methods, especially when dealing with large databases, non-linear patterns and variables that are highly correlated or not normally distributed. Ensembles and mixed models are able to improve the result scores, and reliability of the models as long as a model validation is carried out.

In the works by Mario Elia et al.[10] and by Mariana Dondo Bühler et al.[11] the target fires have been aggregated into administrative areas, the number of ignitions was the one to predict through regression models or ANNs. The studies by Pedro Almeida et al.[12] and by Annalie Dorph et al.[13] worked instead on ignition points and probability as a number between 0.0 and 1.0. The strategy to have a defined target was to gather the positive historical events (probability 1) and to simulate negative events (probability 0) along the spatial domain, then a technique to build a continuous gradient map has been applied such that all the map points could have a probability assigned. Tree-based algorithms such as Random Forests have been employed here, as well as regression (or classification) models as GLM. A hybrid technique between point fires and aggregation is the one proposed by Jiangxia Ye et al.[14] which divides the map through a regular grid of binary cells: 1 if an ignition occurred, 0 otherwise. Then, a Bayes model has been applied where the target areas were larger and comprehensive of multiple grid cells such that a discrete distribution could have been computed by counting the proportion of positive cells.

1.3 Wildfire ignition predictive variables

Over the definition of the risk earlier studies have figured out what components to consider as causes or simply correlated to fire events, first observed and then tested through statistical techniques. The work of FirEUrisk investigates on the risk assessment systems currently used by several government institutions and their methods. They explored several sets of features that may show relevant relationships with wildfires. Independently on the form of the target, which depends either on the objective of the work and on data availability, one can observe common results about which features correlate most among the works.

Weather variables used commonly include atmospheric and precipitation data daily reported or aggregated in time intervals depending on model needs. Combinations of high temperature, calm wind and low humidity give favor to fire ignition[15]. Also topographic characteristics of the place such as altitude and slope determine whether the ignition will grow up to a relevant size.

Essential component is moisture. Beside local mappings of the terrain and vegetation, a powerful tool to model wildland fuels is remote sensing. A widely used model of this type in Europe is the CORINE Land Cover model[16]. Part of this coverage is devolved to cultivations, grazing or protected areas: here is where human activities make their impact on the wildfires' dynamics. The usage intention of the land and the forest coverage gain relevance due to the importance that the human gives to the area, and so their interest to keep it untouched[17,18].

A main factor is the accessibility of human activity in the area of ignition. The correlation with proximity to roads and road density confirms this feature as relevant by two perspectives: ones considers roads as signs of urbanization which generally decreases the likelihood of fires, while others focus more on this indicator as presence of human activity inside flammable areas[17,19]. In both cases, roads must be combined with other components in order to extract insights. Types of cultivations, usage intention of the terrains and basic characteristics of agricultural systems play a crucial role on the phenomenon's behavior. The Global Roads Inventory Project (GRIP[20]) has developed a vector dataset for the exact use of global environmental and biodiversity assessment. Other studies put the stress on the human component by looking for those features that can be indirectly correlated to fire phenomena such as demographic data, degree of education and employment, population's wealth. It has been proved that denser population areas together with a less developed economy and other data are positively correlated to fire occurrence. The review by Costafreda et al.[9] gathered information about the statistical correlations among the predictive variables and the targets regardless of its form. Among the positive correlated ones we can find the presence of cultivations and/or forest and grasslands, the roads and railroads density, temperature and days without precipitation. Among the negative correlated ones we can find the distance to roads and settlements, humidity, urbanization of the area.

The thesis work that follows takes into account these results in order to compare with the estimation insights. As a starting point, the most wide and complete data features have been included, looking for them in the fields previously mentioned: socioeconomic, demographic, topographic, environmental and atmospheric.

Chapter 2

Method

The target data availability drove the choice of the land to work on. The region of Piedmont has been chosen due to the contextual familiarity and the presence of a large history of wildfires stored in the institutional websites. Once identified the type of data to search for, a massive collection of open-source resources made the raw dataset of features that were first filtered and then processed as tabular data, in case encoded. The target definition was the first problem to face. The main objective became to test the performances of Random Forest algorithms on three different sampling methods of the target, which has been chosen to be a discrete probability of ignition (from 0.0 to 1.0 with 0.1 step). Then the interpretation of the models were able to profile the areas depending on the outcomes. The region of Piedmont has currently an available open web board on the Regional Agency for Environmental Protection (ARPA[21]) to report the risk of several environmental hazards including wildfires. This work is intended to be independent on the current system and it wants to explore the topic on a methodological level such that one can apply the same procedure in different geographical areas, as well as in finer resolution (e.g. Italy, Europe...).

2.1 Data collection

All data sources are listed in Table A.1. References to geographical coordinates have been all standardized to WGS 84 / World Pseudo Mercator (EPSG code 3857). Each dataset contains a *geometry* feature or reference variables to the respective townships through name or code identifier, hence the observations are gathered by (available) years and municipality.

Municipality boundaries. The dataset from RNDT contains a list of *Polygon* shape objects bounding the municipality administrative areas. Each polygon is provided of its national unique 6-digits identifier code from the National Istitute of Statistics (ISTAT), the name of the municipality and other attributes like the belonging region and province. Only the municipalities of the region of Piedmont have been kept by filtering on region code 1. Note that administrative areas change throughout the years, leading to misalignment or missing areas. From now on, as one municipality is not found in the boundaries dataset it is manually added if possible, otherwise it is ignored and will result in NA data in the final training dataset in case a searched point falls in an empty area. Moreover, each data referring to a certain municipality is merged with the geometry in this boundaries dataset.

Agriculture. Data from finer granularity (province, region...) has been removed. Since the variable names vary between 2010 and 2020, they have been standardized to the 2020 notation. Regarding grazing, buffalos have been included into cattles and goats into sheeps. Hives data are missing in 2010 hence they have been filled with 2020 data. Exceeding columns in 2010 have been dropped.

Altimetry. Data contains the main statistics of height of subareas of each municipality. Only the integer median variable *MEDIANA* will be used for the weather regression Kriging (see 2.2.2).

Weather. Data have been reported by more than 300 weather stations spread throughout the region. The most of the parameters are detect from all the stations, with little exceptions like the solar radiation, which is replaced with NA values when missing. The coordinates and altitude of each station are joined with the dataset.

Education and Employment. They contain the number of people for each education level and employment condition categorized by age. Data from finer granularity (province, region...) has been removed.

Incomes. They contain the total amount of money and people referred for each income category. Some columns appear only in more recent years. They have been deleted to standardize the variables.

IFC. Geographical references show up in the form of centroids of the municipality, hence they have been merged with the boundaries by looking for the polygon containing the centroid. No ambiguity has been found.

Demography. The datasets store population numbers per age and per sex. The ones containing columns per sex also contain a column referring to age, hence the columns have been standardized by age intervals. Furthermore, this partition coincides with the partition gave in the employment and education data.

Target wildfires. The regional database contains a list of more than 6000 past wildfires since 1997 provided of a unique code identifier, space and time coordinates of the event and other detail about the place. The pairs of coordinates are converted to *Point* shape objects. About the 25% of ignition points has the burnt area attached which states the final size of the fire. These areas in particular were compared to the ones reported by the EFFIS dataset and it resulted that the Copernicus' ones are fewer and less precise than the region's website's ones. Hence, the more complete dataset has been used.

Vegetation and Roads. Coverage maps have been found in files in *LineString* shape. They have been buffered of a 0.5 meters radius for the vegetation and average width for the roads (depending on road lanes): highways 14m, primary roads 7m, secondary roads 5m, tertiary roads 2.5m, local roads 2.5m. This has been done in order to make them *Polygon* shapes whose area can be actually measured. Both datasets have observation detecting a certain type of road/vegetation and its extension through the *LineString* object. Hence no merge with municipality boundaries has been done. In 2.3 this data will be included in the observations by computing the percentage coverage of a certain road/vegetation type in a buffer of 1 kilometer around the observed point.

Sentinel-2 remote sensing. Data consists in available images with 10 meters resolution within 7 days before the fire. The images are cropped around the fire area boundaries and their size set to 256x256 pixels. The available bands are 12 and they have been normalized. Moreover, a further channel is added composed of red pixels on the burnt area and black otherwise. In order to reduce satellite to a table, for each input channel (size 256x256), a slider halves the size of the image with a kernel of size 3, stride 2, padding 1, until only 1 pixel is extracted. The value of that pixel is stored in a variable that represents that channel.

2.2 Kriging

Kriging is a method of spatial interpolation that originated in the field of mining geology. It is named after South African mining engineer Danie Krige. The necessity under spatial interpolation is to transform punctual measurements into a continuous gradient map of values. The difference between Kriging and other methods like Inverse Distance Weights and Linear Regression is that it uses the spatial correlation between sampled points to interpolate the values in the spatial field rather than building a distribution model. The Kriging predictor is an exact interpolator, meaning that each interpolated value is calculated to minimize the prediction error for that point.

To set up the Kriging environment, it must be given a quantity $z(\mathbf{x})$ function of the spatial vector $\mathbf{x} = (x, y)$. It is given a distance (usually the Euclidean distance) on the coordinate space where the spatial vectors live (usually $\mathbb{R}^2$ in 2 dimensions).

The variogram The variogram (or semivariogram) is a function representing the variance of the values z against the spatial distance between the domain points $\mathbf{x}$. This scatter plot is named *raw* variogram. Instead, the *experimental* variogram is a smooth line through the raw one. It is thus created: the axis of separation distance is divided into consecutive intervals whose maximums are called *lags*. Intervals between one lag and another are called *lag intervals* and their representative lag is their maximum. Depending on spatial in-between distances, pairs of data points $\mathbf{x}$ are grouped into subsets (not necessarily disjoint) relative to the lags, then the variance of values of each subset is computed. The experimental variogram is a scatter plot of points $(lag, variance)$ where the variance value is the sum of square differences of all the pairs of points whose spatial distance lies inside the lag interval.

As the experimental variogram is built, an approximation model is required. The model function must follow some constraints:

- non negativity: this follows from the non negativity of the variance
- continuity and eventual constancy: above a certain distance value called *range* the function should not vary anymore since the overall sample variance has been reached. This variance value is called *sill*
- isotropy (the spatial correlation does not depend on the direction but only on the absolute distance), stationarity (the variogram model is the same across the whole space)

Thus it follows that the variogram model has the form of a piecewise function of the form

$$\gamma(d) = \begin{cases} f(d) + n & d \leq r \\ s & d > r \end{cases}$$

where d is the distance value at which to calculate the variogram, $r \geq 0$ is the range, $s \geq 0$ is the sill, n is a correction factor called *nugget*, and f is a positive continuous function such that $\lim_{d \to r} f(d) = s$. Usually, instead of the sill, one speaks of the partial sill $psill = sill - nugget$.

Common models for the f function include: Gaussian, exponential, spherical, hole-effect, nugget-effect, linear, power, logarithmic.

Interpolation using Kriging Once modeled the variogram, we define the random function

$$Z(\mathbf{x}) := \sum_{i=1}^{n} \lambda_i z(\mathbf{x}_i)$$

that is, the value of z to estimate in a given point $\mathbf{x}$ of the domain is a linear combination of the data values. We will consider the generic problem: given n measurements of z at locations with spatial coordinates $x_1, x_2, \ldots, x_n$, estimate the value of Z at point x_0. The λ_i are defined based on the variogram model with the following method:

- we want the random function to be an unbiased estimator. This means that $\sum_{i=1}^{n} \lambda_i = 1$
- we want the estimator to be the optimal one, i.e. to have the minimum variance possible. Hence, one has to minimize the function

  $$\mathbb{E}[(\hat{Z}_0 - z(\mathbf{x}_0))^2] = -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j\gamma(\|\mathbf{x}_i - \mathbf{x}_j\|) + 2\sum_{i=1}^{n} \lambda_i\gamma(\|\mathbf{x}_i - \mathbf{x}_0\|)$$

  under the previous constraint on the λ_i coefficients. The value γ is the variogram computed in the distance $\|\mathbf{x}_i - \mathbf{x}_0\|$.

This method is known as **Ordinary Kriging**, the most popular method, which needs only the variogram to be computed.
Other methods include:

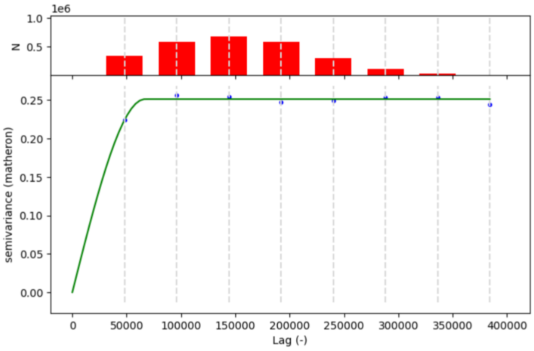
- *Simple Kriging.* The most straightforward, but less general. It assumes the expectation of the random field is known, hence no need of variogram modeling is required, but it does not replicate sampling variations as effectively as Ordinary Kriging
- *Universal Kriging.* It generalizes the Ordinary Kriging by assuming a spatial trend of values varying in a deterministic way. Therefore the trend is modeled usually as a polynomial function. The variance is assumed stationary anyway
- *Regression Kriging.* It estimates the spatial trend of the values through a regression model (either a Machine Learning model) and adds the residuals modeled through an Ordinary Kriging. This method is useful when the trend is assumed to be predicted by another spatial variable and if the response variable distribution follows a normal distribution

2.2.1 Target interpolation

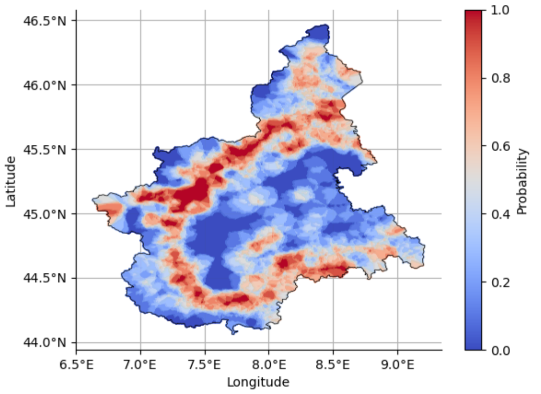
The usage of Kriging interpolation for the target variable has been inspired by the work by Almeda et al.[12] The point values were thus assigned: 1.0 to the historical ignition points and 0.0 to an equal number of points randomly extracted where no ignitions happened, at least 1 kilometer far from the past fires. A Moran I test on these points shows a theoretical index close to 0 (which is desirable) and an actual index of 0.35 with a p-value of 0.001. This means that there's a moderate positive spatial autocorrelation, similar values are not strongly clustered together. Since only the 1 values are "real" and all the zero values are uniformly sampled along the map, no spatial trend is assumed. Then a *spherical* Ordinary Kriging model has been trained[24]. The variogram model has the following form:

$$\gamma(d) = \begin{cases} p * (\frac{3d}{2r} - \frac{d^3}{2r^3}) + n & d \leq r \\ p + n & d > r \end{cases}$$

The choice of the spherical model has been done by GridSearch by minimization of the RMSE score. Figure 2.1a shows the variogram fit on the data points and Figure 2.1b shows the probability interpolation over the region map.



(a) Variogram model fit on fire target data. The red histogram displays the amount of data points relative to each lag interval have been used to compute the (semi)variance



(b) Probability map interpolated by the Kriging model

2.2.2 Weather interpolation

As said in 2.1, weather data are gathered in the form of reports from weather stations whose coordinates and altitude are known. Hence one can extend

the values to all coordinates through a Kriging algorithm that can take advantage of the altitude data as a predictive variable.
Here's how the weather interpolation is made:

- if there is no value available for that parameter in the specified time period, return NA values
- else, if the parameter is constant valued throughout all the stations, or if the algorithm fails due to small variability of values, return the most frequent value reported
- else, train a Regression Kriging algorithm with a Gradient Boosting Regressor base model that takes as explanatory variable only the altitude of the points

Given a random point in the map, its altitude is assigned as the median of its municipality's altitudes and the respective weather parameters are computed through the trained interpolation model.

2.3 Dataset creation

Sampling the target Other points have been uniformly picked up from the map and the probability values have been assigned by the interpolation model, as well as the previous points available, hence also the positives and negatives used for the interpolation have been recalculated. This operation has the advantage to reduce the impact of spatial outliers and to have the target less imbalanced. This set of *point* → *probability* items is the target. Regarding the date of ignition, it has been uniformly sampled for each point. From now on, we refer to this points as target points, and this sample as **Sample A**. It is composed of 4226 observations.

Further processing All features have been converted to numeric formats (int or float type) and duplicate rows dropped. All variables municipality-referred are merged by the corresponding *Polygon* shape. When merging by year, since not all the values are available for each year, the past closest available ones have been chosen; if not available, the choice fell on the closest ones in the future. For instance, if education data are missing in 2017 but present in 2018, the values of 2018 are used also for 2017 data. However, due to a large amount of missing values, only the agriculture data of year 2020 has been added to all yearly datasets. Regarding satellite data, only 3 images among the potential 7 available for each fire have been kept, selected as the three most recent ones. Another operation done was to encode ordinarily three class variables: *IFC (valore)* (fragility index description), *Fascia demografica* (population size interval), *Grado di urbanizzazione* (urbanization degree), *GP_RTP* (road type).

Remark 2.1. The choice of the ordinal encoding goes along with the hierarchy of values that the three variables contained.

The data loader Feature values must be provided to each target point. First, the municipality containing the target point is identified and all the corresponding values are assigned. Then, as anticipated, a buffer of 1 kilometer has been drawn around the point and the coverage percentages have been computed for each category of road and each class value of all vegetation variables. Lastly, the weather values are computed given the altitude of the point. If the altitude is not available, all weather parameters' values have been set as NA. The final features are listed in Table A.2.

Remark 2.2. Due to the sampling method it is rare to have duplicated rows since at least the roads and vegetation vary between kilometers.

Correlation analysis Pearson's and Spearman's pairwise correlations have been analyzed. There is no relevant difference among the samples since most of the variable's values depend on the municipality where they come from. It is observed a general low correlation (in absolute value) especially among the variables belonging to different fields (for example, weather against population). Conversely, they have been identified some correlated blocks among variables from similar fields. In particular, the variables from incomes, employment, population and education data are extremely correlated among themselves (score > 0.95) and highly correlated between each other (score 0.80 - 0.90). This is an information about the distribution of the various categories of the population: proportions among income brackets, ages and social status are generally kept along the region. No relevant correlation can be appreciated between predictive features and the target: any linear (or monotone) prediction model would not fit well.

Missing values handling Due to incompleteness of the datasets imported, it happens that some municipalities have no value in certain variables or years. In order not to let those result in many missing values, a further imputation has been done. It consists in averaging the values of the neighboring municipalities for each missing variable. For instance, let's say that the city 001001 has no *total_cattles* data. All the neighboring municipalities are detected and their *total_cattles* values get averaged to impute the missing one. Over this first imputing a small percentage (~30 observations) of records show missing values. They will be imputed through a KNN algorithm *after* the split between train, validation and test dataset to avoid data leakage. The algorithm takes as hyperparameters: 10 nearest neighbors to consider, and value averaging weighted on distance of the points.

Train test split Here the dataset has been split into train sample (85%) and test sample (15%). The test sample will be used only for evaluating final performance, while the train/validation dataset will be used for feature and model selection, as well as the evaluation of final train scores to be compared with the test scores.

2.3.1 Resampling the target

So far the sampling procedure shows the following limitations:

- target values are hilghly imbalanced in favor of the extreme probabilities (zeros and ones)

- randomization of the time coordinates gives less reliability to weather data

A different approach is needed. As the target Kriging has been done only with positives and negatives, one can build a regular grid of points over the map and predict the target through the model, such that the Kriged map itself becomes a distribution of target probabilities, and one can sample from it to balance the values.

The first resampling has been done this way: given the Kriging model trained with positives and negatives previously generated, 1000 observations per each class have been extracted, plus the fire year has been uniformly sampled between 2016 and 2024 (included). From now on we will refer to this sample as **Sample B**. It is composed of 11000 observations.

The second resampling involved drastic changes to the interpolation process. First, all ignition points have been used, not only the ones with burnt area attached. Fires have been divided per year and *season* with the following rule: january-march (winter), april-june (spring), july-september (summer), october-december (autumn). Now, for each season and year, an equal number of negatives have been generated distanced at least 1 kilometer from the positives. Then for each year and season an ordinal spherical model with the same previous parameters have been trained. Finally, for each model and for each target class (0.0 to 1.0) at most 50 extractions completed the final sample. In case of a lower cardinality of the minor class, that number of extractions are done instead. Thus each target point is now balanced between classes and within the same season. From now on we will refer to this sample as **Sample C**. It is composed of 13000 observations. The new weather features are listed in Table A.2.

2.4 Feature selection and model training

RFE selection After the drop of the variables with no variance, the method to select which variables would have trained the final model was the Recursive Feature Elimination through Random Forest algorithms (either regression and classification), ranked by feature importance (sum of Gini indexes). Since the model uses Decision Tree algorithms, an orthogonal rotation to principal component may increase the performances. Therefore also the rotated database has been tested. Different samples produced different set of features based on the peak of the validation score until the train score did not remarkably decrease, reducing from 358-378 features to its 5-10%:

- Sample A: 25 features (in the form of principal components)
- Sample B: 39 features
- Sample C: 20 features

Regarding Sample A, all the Principal Components mainly encode for vegetation types and categories. Also, satellite variables included in those 25 ones have been added later to test them in 2.4.

PCA reduction As in 2.2 the features show several linear correlations, a different feature selection approach tested was to keep only the first n principal components in equal number of the final features selected in RFE. So 25 PCs for Sample A, 39 for Sample B and 20 for Sample C. This selection kept respectively 61.8% , 61.3% and 60.1% of the variance of the original features.

Human-related features filter All the original features have been kept so far before the algorithmic selection. The variables can be split between environmental data (weather, vegetation...) and human-related data (population, urbanization, incomes...). Since the focus of the work is on human-caused fires, it would be nice to study the behavior of the predictive models under human features only, but most of all the importance rankings of these features in order to find correlations (either spurious) with the target. As observed in 2.2 there is a consistent block of extremely linearly correlated variables that here compose a large proportion of the dataset, so they were orthogonally rotated and its number reduced through a PCA to 9 components for all the three samples. The choice has been made by elbow method. The 9 principal components account for ~87% of the total variance.

Image ablation study As the most complex data to load, handle and transform, it would be nice to test the predictive power of satellite images in the form used for this work. Hence, a simple study of final scores in case of the removal of all the image features leads to this information. The scores were all computed by training new models with the selected PCs of Sample A and without Sentinel's variables.

2.4.1 Random Forest model training

As reviewed in 1.2, regression models such as Logistic and Poisson suffer from multicollinearity, which has been already found in this work's dataset. Therefore, based on literature review, a Random Forest model[22] has been chosen as the final one. Random Forests are ensembles of Decision Tree models where each tree is trained on a limited number of observations and features, which has been proved to improve the performance of the elementary models and to reduce overfitting, but they increase the complexity of the model and make interpretability harder. Both Regression and Classification Forests have been used since they both have pro and cons for the specific task: classification models may be more accurate, whereas regression models may catch uncertain probabilities by averaging the neighboring observations. Though, when evaluating final performance, the output of the regression models have been discretized to decimal probabilities (0.0, 0.1, 0.2,. . . , 0.9, 1.0). Regression Forests have been tuned by R2 score, while Classification Forests by Accuracy score (see below for the mathematical definition), both on 10-fold cross validation. In order to have more general results, the validation scores of the grid search has been evaluated starting from 3 different random seeds and then averaged per each combination of hyperparameters. Note that the scores of $max\_depth = 50$ and $max\_depth = 100$ coincide meaning that leaves' purity is reached with at most 50 consecutive splits. The final chosen parameters for regression and classification forests are are listed in Table A.3.

Evaluation metrics definition Given y_i the true value of the $i\,th$ observed probability and f_i its predicted value, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ the mean of the observed data, n the size of the sample, p the number of classes, $\mathbb{1}(\cdot)$ the indicator function, the following evaluation metrics are defined:

$$R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i = f_i)$$

$$Adjusted\ MAE = \frac{1}{p \cdot n}\sum_{i=1}^{n} |y_i - f_i|$$

All the models have been evaluated on these three metrics. The R2 score determines the proportion of variance of the observed values that had been kept by the predicted ones. The accuracy score states the proportion of observed values correctly predicted. The adjusted Mean Absolute Value score is a measure of the errors of the predictions i.e. a dispersion measure around the main diagonal of the confusion matrix between true and predicted values. The choice of the $1/p$ adjustment (corresponding to a $1/11$ multiplicative factor for all the trained models) over the regular MAE has been done only to resize the score numbers.

Chapter 3

Results

3.1 Scores and feature importance

Figure 3.1 shows the accuracy scores of train and test datasets predictions over the sampling method and feature selection method. As expected, the models all overfit. Results show a general better performance on Classification Forests. Among the three samples, Sample B fits the best models and achieves the highest test accuracy and smaller Mean Absolute Error, which means that the confusion matrix displays almost no out-of-sense predictions, meaning that all the non-correctly predicted probabilities are closer to the real value rather than for the other samples. This can be associated to its class-balanced target set and the presence of all the real past fires in it. The confusion matrix with almost no out-of-sense predictions, meaning that all the non-correctly predicted probabilities are closer to the real value rather than for the other samples. The ablation study on the satellite data revealed unchanged scores between the models with and without images. Furthermore, the complete models rely on the principal components only as the first five variables, hence we conclude that satellite images *converted to single tabular values form* are useless.

It has been chosen to proceed with the analysis of the best performative model among the proposed ones (sample B, Classification Forest, RFE selection). To better understand the distribution of the misclassifications, if one considered as valid predictions the ones one cell close to the main diagonal of the confusion matrix (prediction error = 0.1), the test accuracy would increase to 0.767. With a prediction error of at most 0.2, it would reach 0.912. With all the predictions (either train and test), a further Kriging interpolation with the same variogram model of the target in 2.2.1 has been trained and its result showed in Figure 3.2a. It is observed that the map distribution looks smoother than the true one, with evident high-probability zones on the north-west chain of the Alps as for the historical map, and low-probability zones on the central, flatter areas and on the highest mountains along the boundaries. The interpolated map reports 91.72% of misclassified points, either positive and negative, with a few exceptions around clear areas of zero values corresponding to the extreme probabilities. Yet, Figure 3.2b displays at most the 20% of error since the distribution of the residuals is highly concentrated around the origin. The overestimations are heavier and more scattered than the underestimations, which are limited to the region's edges. Here if one considered as valid predictions the ones not exceeding the 10% of error the misclassification percentage would reduce to the 11.52%.

Figure 3.3 shows the first 20 most important variables for the model to predict the ignition probabilities, i.e. the variables that have been used more by the several trees composing the Forest on splits close to the root nodes. As all the other models that used vegetation data report, the total coverage of vegetation within 1 kilometer radius figures as a must-have indicator. Intuitively, locations with sparse or absent fuel such as urbanized areas and the highest mountains have the lowest probabilities since there is lack of one of the three pillars of the ignition phase. Conversely, countrysides, hills and lower mountains are much more covered and thus more prone to fires.

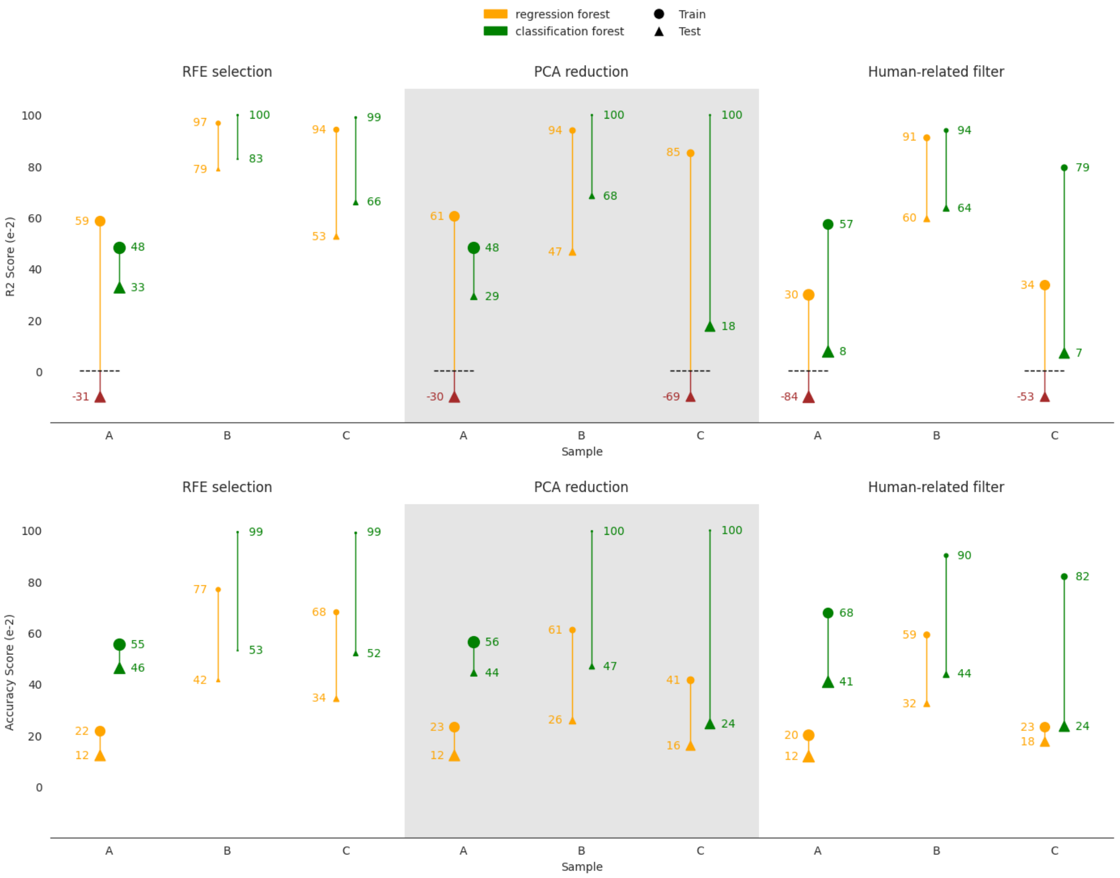


Figure 3.1: R2 scores (top) and accuracy scores (bottom) for each sample and each feature selection method. The size of the value points is proportional to the Mean Absolute Error

Note that the 2nd, 3rd, 4th and 10th most important variables are types of vegetation that spatially overlap the trend of past ignitions (probability 1). The 5th, 6th and 7th features refer to the incomes field. The 5th refers to the median income range of the region while the 7th refers to a richer part of the population. The 6th feature is a measure of the income of enterprises that do not exceed annual incomes of 300k - 700k euros[28] which can be identified as the small and medium-size companies. The four SOC variables refer ordinarily to: the ratio between population sizes of inactive ages (until 14 y.o. and over 65 y.o.) and active ages (15-64 y.o.), the percentage of the adult population with lower education level, The employment rate and the population growth rate. AMB_01 and AMB_02 are ecological indicators: one measures the high-emission motorization rate, the other one the unsorted urban waste collected. The three remaining variables describe the usage of the terrain: land consumption amount (TER_02), the surface left to wild woodlands (Superficie a boschi), and the accessibility time index to essential services which is an indicator of urbanization (TER_03).

From Figure A.1 it is possible break through the relationships among the features and the targets' discrimination. The first 4 graphs (detail in Figure 3.4) tell a clear information: the more vegetation around, the higher is their ignition probability. The same information is given by graphs number 8, 10 and 20 with a remarkable difference: the medians for the smaller probabilities are null. This implies that those types of vegetation are limited to specific geographic areas and absent in general throughout the map. This means that for those observations the model overfitted on a feature with small variance. Different case is the 19th graph, since the purpose of the vegetation coverage (wildland or controlled) seems to influence the ignition rate. Extreme probabilities are characterized by the smallest wildland zones. Looking back to the first 4 graphs, it looks like the low probabilities are characterized by no vegetation at all, while the highest ones with large woods extensions but not left to the wild, somehow controlled and monitored by humans. Average probabilities are related to more wildland coverage, probably due to the inaccessibility by human activities as the graph of TER_03 barely reports, for sure correlated with the lowest land consumptions scores in TER_02. AMB_02 puts a neat distinction between low and high probabilities according to the mean values as well as AMB_01. Hence, the inhabitants that produce little rubbish but has high motorization emissions are more prone to ignite. This type of townships can be identified as the ones with those cities that have low consumptions but need to move a lot by personal vehicles, as the cities far from the metropolitan areas, presumably in the countrysides, hills and low mountains. The graphs number 5, 6, 7 and 9 disagree between mean and median values, which gives an insight about the population density. Greater probabilities show up many times in places with many of the median wealth range and richer range, as well as in zones with no enough people to reach the average wealth of Piedmont's townships. Note that for the 0 probability instead the distribution is more skewed on values higher than the mean, but moderate with respect to the other classes. The highest and lowest income ranges were also the variables describing the first important principal component in the study with only human-related data. Due to the high correlation between income variables and employment and education variables, one cannot deduce any clear cause-effect relationship. Though one can profile the socioeconomic conditions of the cities in terms of the predicted probability. Lastly, it is evident from the 13th graph that the past fires ignited most in administrative areas that register the lowest percentages of people with low education, and surrounded by people with the highest ratios of this index, since the closest high probabilities are as well close in space due to the interpolation technique. It is reasonable to assume that people and human activities have easy access to neighboring municipalities, hence this data suggests that ignition propensity takes benefit from areas with heterogeneous education level.

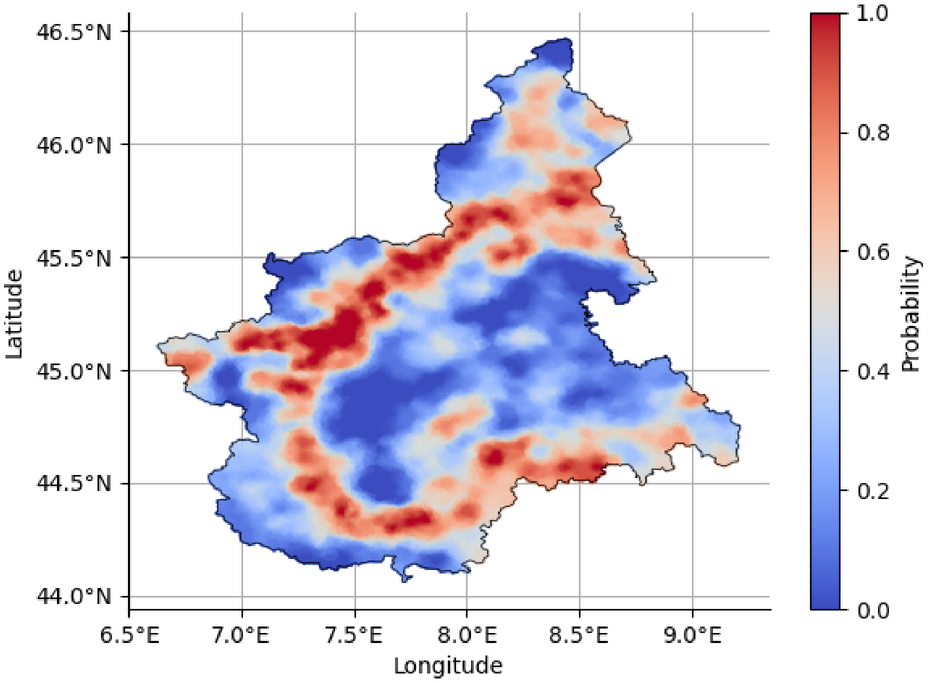
Tables with all the final scores are listed in A.3. Feature importance for all the models are listed in A.4.

3.2 Comparison with previous works

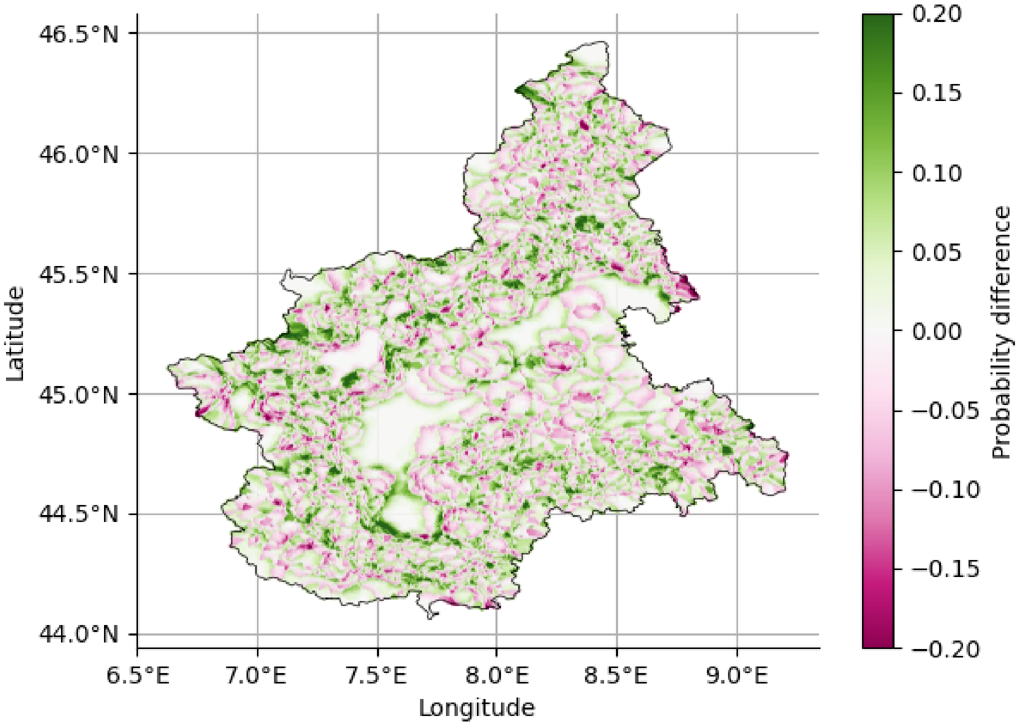
Here follows a brief comparison between the results achieved in this work and two similar ones with the same purpose. It is important to remark that the differences among the three works lie on the *geographic area*, the predictive variables used, the interpolation method and the models used, hence they are not properly comparable from a performative point of view, but qualitatively one can appreciate some interesting convergences and divergences.

Pedro Almeida et al.[12] worked on the entire surface of Portugal and applied a Kriging model by an exponential variogram as it fits better on their data. It is important to notice that part of Portugal's forests are covered by tree species resistant to fire, which is not the case in Piedmont. Though the Italian region's main species belong to the type of chestnut trees and pines which have a strong regeneration ability[2], hence the alpine forests would result more virtuous in terms of resilience on a hypothetical study on the vulnerability component of risk.

Annalie Dorph et al.[13] worked on the Vicotria's district in Australia. Accuracy of classification forests reached the 80-90% with an average classification error between 0.1 and 0.2. They registered a stronger predictive impact of house density and distance to roads. This has some similarities with the results achieved for this work's study on human-related features only, since density of roads appeared as top features explaining the most important principal components. The great difference is that the work on the Victoria's region included a compound ignition index obtained by weather data (Forest Fire Danger Index) that has been always ranked among the top 3 explanatory variables. Moreover, topographic characteristics of the terrain such as slope and altitude, not included in this work, scored a moderate predictive power.



(a) The Kriged map trained on the predicted values by sample B, Classification Forest, RFE selection



(b) The prediction error map (within the 20%) between the predictions and the true values

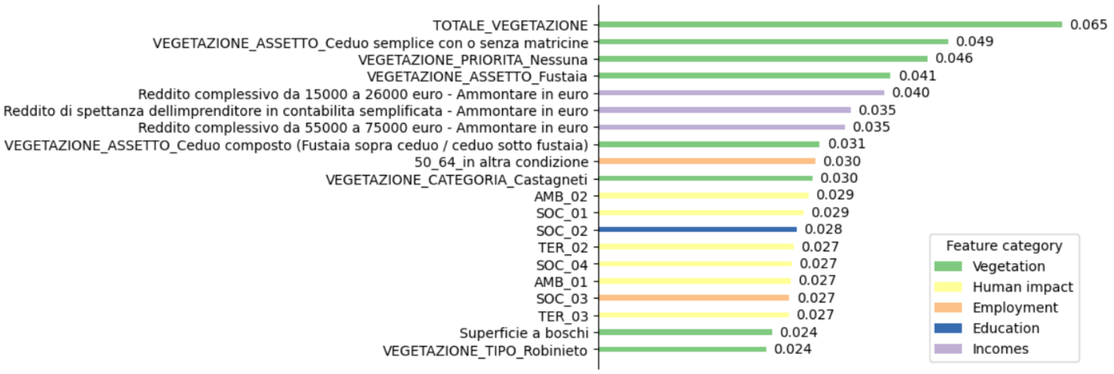


Figure 3.3: The 20 most important features and the respective scores and data categories (sample B, Classification Forest, RFE selection)

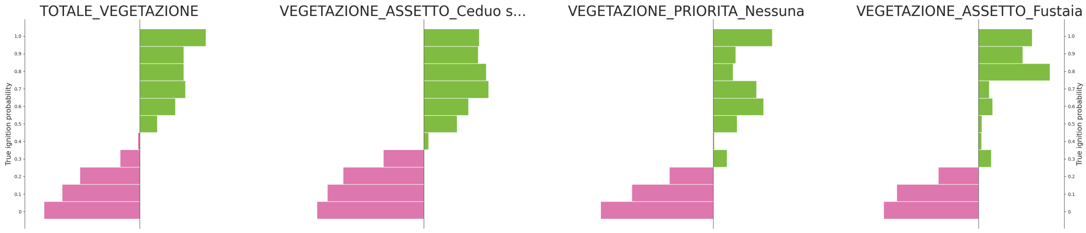


Figure 3.4: The graphs represent the probability class means (bars) of the 4 most important variables with respect to the overall variable's mean (the central vertical line). The pink bars are class mean values smaller than the overall mean, the green bars are class mean values larger than the overall mean

Chapter 4

Conclusions

This thesis addressed the problem of the prediction of the point-wise wildfire ignition probability in the Italian region of Piedmont. Based on historical fires and related environmental and socioeconomic data, the proposed solution was a Random Forest model trained on a spatially interpolated target probability map through a method named Kriging, whose strength is the exploitation of the isotropic distribution of the target's values variance with respect to the spatial distance between the data points.

After the detailed definition of the problem, the first part of chapter 2 has been dedicated to the creation of the dataset by merging all the data based on the municipalities they belong to, encoding categorical features like roads and vegetation, and most of all it took place the Kriging interpolation of the target and the weather variables. Then, from the interpolated map there have been sampled other intermediate target values, and two more samples have been generated in order to get a stronger class balance and larger data. Three methods of feature selection were tried in order to reduce the dataset dimension. The most effective one was the Recursive Feature Elimination through Random Forest models, the same algorithm chosen for the final modelization. All the models overfit as expected and perform better than random guessing. For each model including environmental data, these last one types of features overcome the human-related variables in importance score. The best performing model was a Classification Forest trained on the balanced sample of the Kriged map with all the past fires, on the features selected through RFE on the same type of algorithm. Here, the best discriminating features are four vegetation types spread along the Alps overlapping the past fires. Then a list of incomes data and townships' indicators show relevant links with the target, despite less powerful importance scores. Since there is too high correlations among incomes, education levels and employment features, one cannot extrapolate insights about cause-effect relationships with the ignition probability.

The method resulted as a powerful tool to identify the fields of the mostly useful variables to discriminate the interpolated probabilities. From this point, a possible improvement is to apply other types of algorithms like in the field of Contrastive Learning in order to assign similar ignition probabilities to points with similar profile of variables. The limitations of this method rely on the dependence on the interpolation method of the target and the interdependence of human-related features. To reduce the impact of the interpolation one can predict no more on the single ignition points but on the number of ignition occurrences of each municipality, which is also coherent with the most of the data townships-based. In terms of interpretability, the study of spurious correlations among variables and a finer encoding of categorical variables may help the models to focus on specific features that link the places conditions to fire tendency.

Bibliography

[1] Emilio Chiuveco et al. (2023): Towards an Integrated Approach to Wildfire Risk Assessment: When, Where, What and How May the Landscapes Burn. Fire, 10.3390

[2] Simona Barbarino et al. (2012): Gli incendi boschivi nelle Alpi: Conoscenza, previsione e cooperazione per difendere il nostro patrimonio forestale, Arpa Piemonte

[3] https://environment.ec.europa.eu/topics/forests/forest-fires_en

[4] FIREURISK - DEVELOPING A HOLISTIC, RISK-WISE STRATEGY FOR EUROPEAN WILDFIRE MANAGEMENT, DOI 10.3030/101003890

[5] C.E. Van Wagner (1974): Structure of the Canadian forest Fire Weather Index. Department of the environment, Canadian Forestry Service, Publication No.

[6] https://forest-fire.emergency.copernicus.eu/

[7] Andrina Gincheva et al. (2024): A monthly gridded burned area database of national wildland fire data. Scientific data, 10.1038

[8] Giovanni Laneve, Marco Di Fonzo, Valerio Pampanoni, Ramon Bueno Morles (2024): Progress and Limitations in the Satellite-Based Estimate of Burnt Areas. Remonte sensing, 10.3390

[9] Sergi Costafreda-Aumedes, Carles Comas, Cristina Vega-Garcia (2017): Human-caused fire occurrence modelling in perspective: a review. International Journal of Wildland Fire, 10.1071

[10] Mario Elia et al. (2020): Estimating the probability of wildfire occurrence in Mediterranean landscapes using Artificial Neural Networks. Environmental Impact Assessment Review, 10.1016

[11] Mariana Dondo Bühler, Mónica De Torres Curth, Lucas Alejandro Garibaldi (2013): Demography and socioeconomic vulnerability influence fire occurrence in Bariloche (Argentina). Landscape and Urban Planning, 10.1061

[12] Pedro Almeida, Isilda Cunha Menezes, Ana Isabel Miranda (2024): A Human Behavior Wildfire Ignition Probability Index for Application to Mainland Portugal. Fire, 10.3390

[13] Annalie Dorph, Erica Marshall, Kate A. Parkins, Trent D. Penman (2022): Modelling ignition probability for human- and lightning-caused wildfires in Victoria, Australia. Natural Hazards and Earth System Sciences, 10.5194

[14] Jiangxia Ye et al. (2017): Modeling the spatial patterns of human wildfire ignition in Yunnan province, China. Applied Geography, 10.1016

[15] Rita Durão, Catarina Alonso, Célia Gouveia (2022): The Performance of ECMWF Ensemble Prediction System for European Extreme Fires: Portugal/Monchique in 2018. Atmosphere, 10.3390

[16] https://land.copernicus.eu/en/products/corine-land-cover

[17] Leone D. Mancini Piermaria Corona, Luca Salvati (2018): Ranking the importance of Wildfires' human drivers through a multi-model regression approach. Environmental Impact Assessment Review, 10.1016

[18] Luiz Felipe Galizia, Thomas Curt, Renaud Barbero and Marcos Rodrigues (2021): Understanding fire regimes in Europe. International Journal of Wildland Fire, 10.1071

[19] Pere Joan Gelabert et al. (2025): Assessing human-caused wildfire ignition likelihood across Europe. The EGU interactive community platform, 10.5194

[20] Meijer, J.R., Huijbregts, M.A.J., Schotten, C.G.J. and Schipper, A.M. (2018): Global patterns of current and future road infrastructure. Environmental Research Letters, 13-064006

[21] https://www.arpa.piemonte.it/bollettino/bollettino-pericolo-incendi-boschivi

[22] Gilles Louppe (2014): Understanding Random Forest, University of Liège

[23] P. K. Kitanidis (1997): Introduction to Geostatistics: Applications to Hydrogeology. Cambridge University Press

[24] https://geostat-framework.readthedocs.io/projects/pykrige/en/stable/variogram_models.html

[25] Ing. Antonella Vecchio, Dr. Marco Falconi: Approfondimenti di statistica e geostatistica. Agenzia per la Protezione dell'Ambiente e per i Servizi Tecnici (APAT)

[26] https://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#//009z00000076000000.htm

[27] https://www.publichealth.columbia.edu/research/population-health-methods/kriging-interpolation

[28] https://www.agenziaentrate.gov.it/portale/archivio/modelli-e-istruzioni/modelli-2008-2016/

modelli-di-dichiarazione/2007/unico-pf-2007/fascicolo_3/1_istruzioni_per_la_compilazione_dei_quadri_aggiuntivi_al_modello_base/6_istruzioni_per_la_compilazione_del_quadro_rg.html

[29] Caitlyn Reilley et al. (2023): The Influence of Socioeconomic Factors on Human Wildfire Ignitions in the Pacific Northwest, USA. Fire, 10.3390

Appendix A

Annex

A.1 Data and features

Table A.1: Raw data sources

Data	Source	Years
Wildfires database	Geoportale Piemonte	1997-2024
Municipality boundaries	RNDT	2025
Altimetry	ISTAT	2016
Satellite	Copernicus Sentinel-2	2016-2024
Weather reports	ARPA Piemonte	1988-2025
Incomes	Ministero dell'Economia e delle Finanze	2000-2023
Demographic data	ISTAT	1997-2024
Employment	ISTAT	2018-2022
Education	ISTAT	2018-2022
Agriculture	ISTAT	2010, 2020

Data	**Source**	**Years**
Vegetation map	Servizi Regione Piemonte	2016
Road map	Grip global roads database[20]	2018
IFC	ISTAT	2018,2019,2021

Table A.2: Data variables after feature engineering

Variable name	**Unit**	**Description**
Weather Reports *Variables in italics refer to Sample C (see 2.3.1)*		
tmedia / *Temperatura media_season*	°C	Average temperature
- / *Temperatura media dei massimi_season*	°C	Average maximum temperature
tmax / *Temperatura massima_season*	°C	Maximum temperature
- / *Temperatura media dei minimi_season*	°C	Average minimum temperature
tmin / *Temperatura minima_season*	°C	Minimum temperature
- / *Giorni piovosi pioggia dalle 0 alle 0_season*	int	Number of rainy days counted from 0am to 0am
ptot / *Precipitazione dalle 0 alle 0_season*	mm	Total height of precipitations counted from 0am to 0am
vmedia / *Velocità media del vento_season*	m/s	Average wind speed
- / *Direzione massima raffica_season*	°	Direction of the maximum gust of wind

Variable name	**Unit**	**Description**
vraffica / *Velocità massima raffica di vento_season*	m/s	Speed of the maximum gust of wind
settore_prevalente / *Settore Prevalente_season*	int	Main wind sector (0 - 16)
tempo_premanenza / *Tempo di permanenza nel settore_season*	min	Maximum daily time spent by the wind in its main sector
durata_calma / *Calma di vento_season*	min	Daily time of calm wind (avg wind speed < 0,3 m/s)
umedia / *Umidità media_season*	%	Average air humidity percentage
umin / *Umidità minima_season*	%	Minimum air humidity percentage
umax / *Umidità massima_season*	%	Maximum air humidity percentage
rtot / *Radiazione totale_season*	MJ/m2	Total daily solar radiation over surface unit
hdd_base18 / -	°C	sum of the difference among fixed temperature 18°C and average daily temperature
hdd_base20 / -	°C	sum of the difference among fixed temperature 20°C and average daily temperature

Variable name	**Unit**	**Description**
cdd_base18 / -	°C	sum of the difference among average daily temperature and climate comfort temperature 21°C. Only positive difference factors greater than 3 are counted
day_count / -	int	day number from the start date to the end date of the dataset
- / *season*	class	1 = winter, 2 = spring, 3 = summer, 4 = autumn
YYYY / *YYYY*	int	year of the event
	Income data	
Numero contribuenti	int	Number of taxpayers
Reddito da fabbricati - Frequenza	int	Number of incomes from buildings
Reddito da fabbricati - Ammontare in euro	€	Total income from buildings
Reddito da lavoro dipendente e assimilati - Frequenza	int	Number of incomes from employed work
Reddito da lavoro dipendente e assimilati - Ammontare in euro	€	Total income from employed work
Reddito da pensione - Frequenza	int	Number of incomes from boards
Reddito da pensione - Ammontare in euro	€	Total income from boards

Variable name	Unit	Description
Reddito di spettanza dellimprenditore in contabilita ordinaria - Frequenza	int	Number of incomes attributable to the entrepreneur in ordinary accounting
Reddito di spettanza dellimprenditore in contabilita ordinaria - Ammontare in euro	€	Total income attributable to the entrepreneur in ordinary accounting
Reddito di spettanza dellimprenditore in contabilita semplificata - Frequenza	int	Number of incomes attributable to the entrepreneur in simplified accounting
Reddito di spettanza dellimprenditore in contabilita semplificata - Ammontare in euro	€	Total income attributable to the entrepreneur in simplified accounting
Reddito da partecipazione - Frequenza	int	Number of partecipation incomes
Reddito da partecipazione - Ammontare in euro	€	Total partecipation income
Reddito imponibile - Frequenza	int	Number of taxable incomes
Reddito imponibile - Ammontare in euro	€	Total taxable income
Imposta netta - Frequenza	int	NaN

Variable name	**Unit**	**Description**
Imposta netta - Ammontare in euro	€	Total net tax
Reddito imponibile addizionale - Frequenza	int	Number of net taxes
Reddito imponibile addizionale - Ammontare in euro	€	Total additional taxable income
Addizionale regionale dovuta - Frequenza	int	Number of additional regional tax due
Addizionale regionale dovuta - Ammontare in euro	€	Total additional regional tax due
Addizionale comunale dovuta - Frequenza	int	Number of additional municipal tax due
Addizionale comunale dovuta - Ammontare in euro	€	Total additional municipal tax due
Reddito complessivo minore o uguale a zero euro - Frequenza	int	Number of incomes less or equal than zero euros
Reddito complessivo minore o uguale a zero euro - Ammontare in euro	€	Total income less or equal than zero euros
Reddito complessivo da 0 a 10000 euro - Frequenza	int	Number of incomes from 0 to 10000 euros
Reddito complessivo da 0 a 10000 euro - Ammontare in euro	€	Total income from 0 to 10000 euros

Variable name	**Unit**	**Description**
Reddito complessivo da 10000 a 15000 euro - Frequenza	int	Number of incomes from 10000 to 15000 euros
Reddito complessivo da 10000 a 15000 euro - Ammontare in euro	€	Total income from 10000 to 15000 euros
Reddito complessivo da 15000 a 26000 euro - Frequenza	int	Number of incomes from 15000 to 26000 euros
Reddito complessivo da 15000 a 26000 euro - Ammontare in euro	€	Total income from 15000 to 26000 euros
Reddito complessivo da 26000 a 55000 euro - Frequenza	int	Number of incomes from 26000 to 55000 euros
Reddito complessivo da 26000 a 55000 euro - Ammontare in euro	€	Total income from 26000 to 55000 euros
Reddito complessivo da 55000 a 75000 euro - Frequenza	int	Number of incomes from 55000 to 75000 euros
Reddito complessivo da 55000 a 75000 euro - Ammontare in euro	€	Total income from 55000 to 75000 euros
Reddito complessivo da 75000 a 120000 euro - Frequenza	int	Number of incomes from 75000 to 120000 euros

Variable name	**Unit**	**Description**
Reddito complessivo da 75000 a 120000 euro - Ammontare in euro	€	Total income from 75000 to 120000 euros
Reddito complessivo oltre 120000 euro - Frequenza	int	Number of incomes over 120000 euros
Reddito complessivo oltre 120000 euro - Ammontare in euro	€	Total income over 120000 euros
Reddito da lavoro autonomo - Frequenza	int	Number of incomes from self-employment
Reddito da lavoro autonomo - Ammontare in euro	€	Total income from self-employment
Bonus spettante - Frequenza	int	Number of bonus dues
Bonus spettante - Ammontare in euro	€	Total bonus due
	Demographic data	
Totale	int	Total population
meno di 9 anni	int	Population under 9 years old
9-14 anni	int	Population between 9 and 14 years old
15-24 anni	int	Population between 15 and 24 years old
25-49 anni	int	Population between 25 and 49 years old

Variable name	Unit	Description
50-64 anni	int	Population between 50 and 64 years old
piu di 65 anni	int	Population over 65 years old
Employment data		
occupato	int	Number of employed people
in cerca di occupazione	int	Number of people looking for an employment
non forze di lavoro	int	Unemployable people
percettore/rice di una o più pensioni per effetto di attività lavorativa precedente o di redditi da capitale	int	Number of retired people
studente/ssa	int	Number of students
casalinga/o	int	Number of homemakers
in altra condizione	int	Number of people in employment condition not previously specified
totale	int	Total number of people
15_24_forze di lavoro	int	Number of employable people from 15 to 24 years old
15_24_occupato	int	Number of employed people from 15 to 24 years old
15_24_in cerca di occupazione	int	Number of people looking for an employment from 15 to 24 years old
15_24_non forze di lavoro	int	Number of unemployable people from 15 to 24 years old

Variable name	**Unit**	**Description**
15_24_percettore/rice di una o più pensioni per effetto di attività lavorativa precedente o di redditi da capitale	int	Number of retired people from 15 to 24 years old
15_24_studente/ssa	int	Number of students from 15 to 24 years old
15_24_casalinga/o	int	Number of homemakers from 15 to 24 years old
15_24_in altra condizione	int	Number of people from 15 to 24 years old in employment condition not previously specified
15_24_totale	int	Total number of people from 15 to 24 years old
25_49_forze di lavoro	int	Number of employable people from 25 to 49 years old
25_49_occupato	int	Number of employed people from 25 to 49 years old
25_49_in cerca di occupazione	int	Number of people looking for an employment from 25 to 49 years old
25_49_non forze di lavoro	int	Number of unemployable people from 25 to 49 years old
25_49_percettore/rice di una o più pensioni per effetto di attività lavorativa precedente o di redditi da capitale	int	Number of retired people from 25 to 49 years old

Variable name	Unit	Description
25_49_studente/ssa	int	Number of students from 25 to 49 years old
25_49_casalinga/o	int	Number of homemakers from 25 to 49 years old
25_49_in altra condizione	int	Number of people from 25 to 49 years old in employment condition not previously specified
25_49_totale	int	Total number of people from 25 to 49 years old
50_64_forze di lavoro	int	Number of employable people from 50 to 64 years old
50_64_occupato	int	Number of employed people from 50 to 64 years old
50_64_in cerca di occupazione	int	Number of people looking for an employment from 50 to 64 years old
50_64_non forze di lavoro	int	Number of unemployable people from 50 to 64 years old
50_64_percettore/rice di una o più pensioni per effetto di attività lavorativa precedente o di redditi da capitale	int	Number of retired people from 50 to 64 years old
50_64_studente/ssa	int	Number of students from 50 to 64 years old
50_64_casalinga/o	int	Number of homemakers from 50 to 64 years old

Variable name	Unit	Description
50_64_in altra condizione	int	Number of people from 50 to 64 years old in employment condition not previously specified
50_64_totale	int	Total number of people from 50 to 64 years old
65_forze di lavoro	int	Number of employable people over 65 years old
65_occupato	int	Number of employed people over 65 years old
65_in cerca di occupazione	int	Number of people looking for an employment over 65 years old
65_non forze di lavoro	int	Number of unemployable people over years old
65_percettore/rice di una o più pensioni per effetto di attività lavorativa precedente o di redditi da capitale	int	Number of retired people over 65 years old
65_studente/ssa	int	Number of students over 65 years old
65_casalinga/o	int	Number of homemakers over 65 years old
65_in altra condizione	int	Number of people over 65 years old in employment condition not previously specified
65_totale	int	Total number of people over 65 years old
Education data		

Variable name	**Unit**	**Description**
analfabeti	int	Number of illiterate people
alfabeti privi di titolo di studio	int	Number of illiterate people with no qualification
licenza di scuola elementare	int	Number of people with elementary school diploma
licenza di scuola media inferiore o di avviamento professionale	int	Number of people with middle school diploma or vocational training diploma
diploma di istruzione secondaria di II grado o di qualifica professionale (corso di 3-4 anni) compresi IFTS	int	Number of people with high school degree or professional qualification
diploma di tecnico superiore ITS o titolo di studio terziario di primo livello	int	Number of people with bachelor degree or higher technical diploma
titolo di studio terziario di secondo livello e dottorato di ricerca	int	Number of people with master's degree and Ph.D.
titolo di studio terziario di secondo livello	int	Number of people with master's degree
dottorato di ricerca/diploma accademico di formazione alla ricerca	int	Number of Ph.D.
totale	int	Total number of people

Variable name	Unit	Description
9_24_nessun titolo di studio	int	Number people with no qualification between 9 and 24 years old
9_24_licenza di scuola elementare	int	Number people with elementary school diploma between 9 and 24 years old
9_24_licenza di scuola media inferiore o di avviamento professionale	int	Number of people with middle school diploma or vocational training diploma between 9 and 24 years old
9_24_diploma di istruzione secondaria di II grado o di qualifica professionale (corso di 3-4 anni) compresi IFTS	int	Number of people with high school degree or professional qualification between 9 and 24 years old
9_24_diploma di tecnico superiore ITS o titolo di studio terziario di primo livello	int	Number of people with bachelor degree or higher technical diploma between 9 and 24 years old
9_24_titolo di studio terziario di secondo livello e dottorato di ricerca	int	Number of people with master's degree and Ph.D. between 9 and 24 years old
9_24_totale	int	Total number of people between 9 and 24 years old
25_49_nessun titolo di studio	int	Number people with no qualification between 25 and 49 years old

Variable name	Unit	Description
25_49_licenza di scuola elementare	int	Number people with elementary school diploma between 25 and 49 years old
25_49_licenza di scuola media inferiore o di avviamento professionale	int	Number of people with middle school diploma or vocational training diploma between 25 and 49 years old
25_49_diploma di istruzione secondaria di II grado o di qualifica professionale (corso di 3-4 anni) compresi IFTS	int	Number of people with high school degree or professional qualification between 25 and 49 years old
25_49_diploma di tecnico superiore ITS o titolo di studio terziario di primo livello	int	Number of people with bachelor degree or higher technical diploma between 25 and 49 years old
25_49_titolo di studio terziario di secondo livello e dottorato di ricerca	int	Number of people with master's degree and Ph.D. between 25 and 49 years old
25_49_totale	int	Total number of people between 25 and 49 years old
50_64_nessun titolo di studio	int	Number people with no qualification between 50 and 64 years old
50_64_licenza di scuola elementare	int	Number people with elementary school diploma between 50 and 64 years old

Variable name	Unit	Description
50_64_licenza di scuola media inferiore o di avviamento professionale	int	Number of people with middle school diploma or vocational training diploma between 50 and 64 years old
50_64_diploma di istruzione secondaria di II grado o di qualifica professionale (corso di 3-4 anni) compresi IFTS	int	Number of people with high school degree or professional qualification between 50 and 64 years old
50_64_diploma di tecnico superiore ITS o titolo di studio terziario di primo livello	int	Number of people with bachelor degree or higher technical diploma between 50 and 64 years old
50_64_titolo di studio terziario di secondo livello e dottorato di ricerca	int	Number of people with master's degree and Ph.D. between 50 and 64 years old
50_64_totale	int	Total number of people between 50 and 64 years old
65_nessun titolo di studio	int	Number people with no qualification over 65 years old
65_licenza di scuola elementare	int	Number people with elementary school diploma over 65 years old
65_licenza di scuola media inferiore o di avviamento professionale	int	Number of people with middle school diploma or vocational training diploma over 65 years old

Variable name	**Unit**	**Description**
65_diploma di istruzione secondaria di II grado o di qualifica professionale (corso di 3-4 anni) compresi IFTS	int	Number of people with high school degree or professional qualification over 65 years old
65_diploma di tecnico superiore ITS o titolo di studio terziario di primo livello	int	Number of people with bachelor degree or higher technical diploma over 65 years old
65_titolo di studio terziario di secondo livello e dottorato di ricerca	int	Number of people with master's degree and Ph.D. over 65 years old
65_totale	int	Total number of people over 65 years old
Agriculture data		
Superficie totale - ettari	hm2	Total agricultural area
Superficie agricola utilizzata - ettari	hm2	Total agricultural area actually used
seminativi	hm2	Area dedicated to to arable land
sementi e piantine	hm2	Area dedicated to seeds and seedlings
cereali per la produzione di granella	hm2	Area dedicated to wheats for the production of grain
altri legumi secchi e colture proteiche	hm2	Area dedicated to other dried legumes and protein crops
patata	hm2	Area dedicated to potatoes
barbabietola da zucchero	hm2	Area dedicated to sugar beet

Variable name	**Unit**	**Description**
piante sarchiate da foraggio	hm2	Area dedicated to forage root crops
piante industriali	hm2	Area dedicated to industrial plants
ortive protette in serra e tunnel accessibili all'uomo	hm2	Area dedicated to protected vegetables in greenhouses and tunnels accessible to humans
fiori e piante ornamentali in piena aria	hm2	Area dedicated to flowers and ornamental plants in the open air
foraggere avvicendate	hm2	Area dedicated to alternate forage crops
terreni a riposo	hm2	Area dedicated to fallow land
coltivazioni legnose agrarie	hm2	Area dedicated to agricultural woody crops
vite	hm2	Area dedicated to vine
olivo per la produzione di olive da olio	hm2	Area dedicated to olive tree for the production of oil olives
agrumi	hm2	Area dedicated to citrus fruits
piante ornamentali da vivaio	hm2	Area dedicated to ornamental nursery plants
prati permanenti e pascoli	hm2	Area dedicated to permanent meadows and pastures
Altra superficie rispetto a quella agricola utilizzata, a legna, a boschi e non utilizzata	hm2	Other surface area than the agricultural one used for wood, woods and not used
totale bovini e bufalini	hm2	Area dedicated to cattle and buffalo
totale suini	hm2	Area dedicated to pigs
totale ovini e caprini	hm2	Area dedicated to sheep and goats

Variable name	Unit	Description
totale avicoli	hm2	Area dedicated to poultry
totale alveari	hm2	Area dedicated to hives
Vegetation data		
CATEGORIA	class	Vegetation category
TIPO	class	Vegetation type
INTERVENTO	class	Planned intervention for vegetation
PRIORITA	class	Intervention priority
DESTINAZ	class	Destination use of the vegetation
ASSETTO	class	Vegetation management status
Road data		
GP_RTP	class	Road type 1 = Highways, 2 = Primary roads, 3 = Secondary roads, 4 = Tertiary roads, 5 = Local roads
Fragility indexes		
IFC (decili)	decile classes	Compound index of municipal fragility
IFC (valori)	ordinal class	Compound index of municipal fragility
AMB_01	int	High emission motorization rate per 100 inhabitants
AMB_02	Kg per inhabitant	Unsorted urban waste collection per inhabitant
AMB_03	%	Protected areas
ECO_01	ventile classes	Density of local industrial and service units
ECO_02	ventile classes	Employees in low-productivity local units in the industry and services sector

Variable name	Unit	Description
SOC_01	%	Adjusted population dependency ratio
SOC_02	%	Population aged between 25 and 64 with an educational qualification no higher than lower secondary school or vocational school leaving qualification
SOC_03	%	Employment rate (20-64 years)
SOC_04	int	Population growth rate per 1000 inhabitants
TER_01	%	Landslide risk area
TER_02	%	Land consumption
TER_03	min	Index of accessibility to essential services
Fascia demografica	ordinal class	Demographic range
Grado di urbanizzazione	ordinal class	Urbanization degree

A.2 Hyperparameters tuning

Table A.3: Hyperparameters tuned for Random Forest models

RFE selection			
	Sample A	**Sample B**	**Sample C**
reg. RF			
max_depth	50	50	50
min_sample_leaf	5	1	1

class. RF			
max_depth	50	50	50
min_sample_leaf	10	1	1
PCA reduction			
	Sample A	**Sample B**	**Sample C**
reg. RF			
max_depth	50	50	50
min_sample_leaf	5	1	1
class. RF			
max_depth	50	50	50
min_sample_leaf	10	1	1
Human-related filter			
	Sample A	**Sample B**	**Sample C**
reg. RF			
max_depth	50	50	50
min_sample_leaf	5	1	5
class. RF			
max_depth	100	100	100
min_sample_leaf	5	1	1

A.3 Final scores

Table A.4: Result scores for Sample A

RFE selection			
	R2 score	**accuracy score**	**adj. MAE**
regression	train 0.587 test -0.310	train 0.217 test 0.121	train 15.324 test 24.863

classification	train 0.483 test 0.328	train 0.553 test 0.464	train 21.176 test 26.199
PCA reduction			
	R2 score	**accuracy score**	**adj. MAE**
regression	train 0.606 test -0.302	train 0.232 test 0.121	train 14.957 test 25.088
classification	train 0.481 test 0.291	train 0.565 test 0.445	train 20.683 test 13.905
Human-related filter			
	R2 score	**accuracy score**	**adj. MAE**
regression	train 0.299 test -0.835	train 0.201 test 0.118	train 18.494 test 27.795
classification	train 0.572 test 0.077	train 0.678 test 0.409	train 15.361 test 27.666

Table A.5: Result scores for Sample B

RFE selection			
	R2 score	**accuracy score**	**adj. MAE**
regression	train 0.970 test 0.789	train 0.769 test 0.416	train 2.666 test 9.460
classification	train 0.999 test 0.829	train 0.994 test 0.532	train 0.087 test 8.130
PCA reduction			
	R2 score	**accuracy score**	**adj. MAE**

regression	train 0.939 test 0.467	train 0.611 test 0.259	train 4.607 test 14.430
classification	train 0.999 test 0.683	train 0.996 test 0.468	train 0.063 test 11.972
Human-related filter			
	R2 score	**accuracy score**	**adj. MAE**
regression	train 0.913 test 0.597	train 0.593 test 0.323	train 5.409 test 13.100
classification	train 0.939 test 0.637	train 0.902 test 0.439	train 2.182 test 13.100

Table A.6: Result scores for Sample C

RFE selection			
	R2 score	**accuracy score**	**adj. MAE**
regression	train 0.942 test 0.526	train 0.682 test 0.344	train 3.870 test 12.241
classification	train 0.991 test 0.660	train 0.992 test 0.519	train 0.238 test 11.079
PCA reduction			
	R2 score	**accuracy score**	**adj. MAE**
regression	train 0.853 test -0.692	train 0.414 test 0.158	train 7.333 test 20.375
classification	train 1.0 test 0.175	train 1.0 test 0.245	train 0.0 test 22.722
Human-related filter			

	R2 score	**accuracy score**	**adj. MAE**
regression	train 0.336 test -0.531	train 0.231 test 0.175	train 14.456 test 20.562
classification	train 0.794 test 0.071	train 0.820 test 0.235	train 5.267 test 23.709

A.4 Feature importance

Table A.7: Feature importance for Sample A

RFE selection	
regression	**classification**
PC4	PC4
PC2	PC2
PC12	PC3
PC10	PC12
PC5	PC5
PC3	sentinel_34
PC15	PC10
sentinel_34	PC15
PC6	PC6
sentinel_4	sentinel_18
sentinel_13	sentinel_13
sentinel_12	sentinel_22
sentinel_28	sentinel_12
sentinel_18	sentinel_3
sentinel_10	sentinel_10
sentinel_25	sentinel_4

sentinel_22	sentinel_25
sentinel_3	sentinel_28
sentinel_23	sentinel_32
sentinel_32	sentinel_23
sentinel_21	sentinel_35
sentinel_35	sentinel_36
sentinel_8	sentinel_21
sentinel_36	sentinel_30
sentinel_30	sentinel_8
PCA reduction	
PC4	PC2
PC2	PC4
PC12	PC3
PC15	PC12
PC3	PC5
PC5	PC1
PC9	PC15
PC13	PC6
PC11	PC14
PC6	PC13
PC1	PC11
PC17	PC18
PC14	PC25
PC7	PC8
PC16	PC17
PC21	PC9
PC25	PC16
PC23	PC10
PC24	PC21

PC8	PC19
PC19	PC24
PC20	PC20
PC22	PC7
PC18	PC23
PC10	PC22
Human-related filter	
regression	**classification**
PC3	PC2
PC6	PC6
PC2	PC3
PC4	PC1
PC1	PC4
PC8	PC8
PC5	PC5
PC9	PC9
PC7	PC7

Table A.8: Feature importance for Sample B

RFE selection	
regression	**classification**
VEGETAZIONE_CATEGORIA_Castagneti	TOTALE_VEGETAZIONE
TOTALE_VEGETAZIONE	VEGETAZIONE_ASSETTO_Ceduo semplice con o senza matricine
TER_02	VEGETAZIONE_PRIORITA_Nessuna

VEGETAZIONE_TIPO_Pioppeti	VEGETAZIONE_ASSETTO_Fustaia
Superficie a boschi	Reddito complessivo da 15000 a 26000 euro - Ammontare in euro
Reddito complessivo da 15000 a 26000 euro - Ammontare in euro	Reddito di spettanza dellimprenditore in contabilita semplificata - Ammontare in euro
TER_03	Reddito complessivo da 55000 a 75000 euro - Ammontare in euro
terreni a riposo	VEGETAZIONE_ASSETTO_Ceduo composto (Fustaia sopra ceduo / ceduo sotto fustaia)
Superficie agricola utilizzata - ettari	50_64_in altra condizione
piante industriali	VEGETAZIONE_CATEGORIA_Castagneti
cereali per la produzione di granella	AMB_02
AMB_03	SOC_01
AMB_02	SOC_02
VEGETAZIONE_TIPO_Robinieto	TER_02
totale ovini	SOC_04
VEGETAZIONE_TIPO_Querceto di rovere a Potentilla alba	AMB_01
Altra superficie rispetto a quella agricola utilizzata, a legna, a boschi e non utilizzata	SOC_03
VEGETAZIONE_ASSETTO_Cedu composto (Fustaia sopra ceduo / ceduo sotto fustaia)	TER_03
SOC_01	Superficie a boschi

Superficie per coltivazioni arboricole da legna	VEGETAZIONE_TIPO_Robinieto
patata	Superficie agricola utilizzata - ettari
Reddito complessivo da 55000 a 75000 euro - Ammontare in euro	cereali per la produzione di granella
SOC_02	Superficie totale - ettari
AMB_01	ECO_01
Reddito di spettanza dellimprenditore in contabilita semplificata - Ammontare in euro	terreni a riposo
VEGETAZIONE_ASSETTO_Fustaia	totale ovini
SOC_04	foraggere avvicendate
SOC_03	totale alveari
VEGETAZIONE_PRIORITA_Nessuna	AMB_03
VEGETAZIONE_ASSETTO_Ceduo semplice con o senza matricine	patata
ECO_01	Altra superficie rispetto a quella agricola utilizzata, a legna, a boschi e non utilizzata
totale alveari	coltivazioni legnose agrarie
coltivazioni fruttifere	coltivazioni fruttifere
Superficie totale - ettari	piante industriali
vite	VEGETAZIONE_TIPO_Pioppeti
50_64_in altra condizione	vite
foraggere avvicendate	Superficie per coltivazioni arboricole da legna

coltivazioni legnose agrarie	ortive protette in serra e tunnel accessibili all'uomo
ortive protette in serra e tunnel accessibili all'uomo	VEGETAZIONE_TIPO_Querceto di rovere a Potentilla alba
PCA reduction	
PC2	PC4
PC4	PC2
PC6	PC1
PC15	PC6
PC1	PC3
PC11	PC7
PC3	PC11
PC7	PC8
PC9	PC9
PC18	PC15
PC39	PC36
PC36	PC13
PC29	PC33
PC33	PC5
PC26	PC19
PC13	PC10
PC19	PC24
PC16	PC12
PC5	PC39
PC34	PC34
PC10	PC29
PC38	PC28
PC24	PC18
PC20	PC32

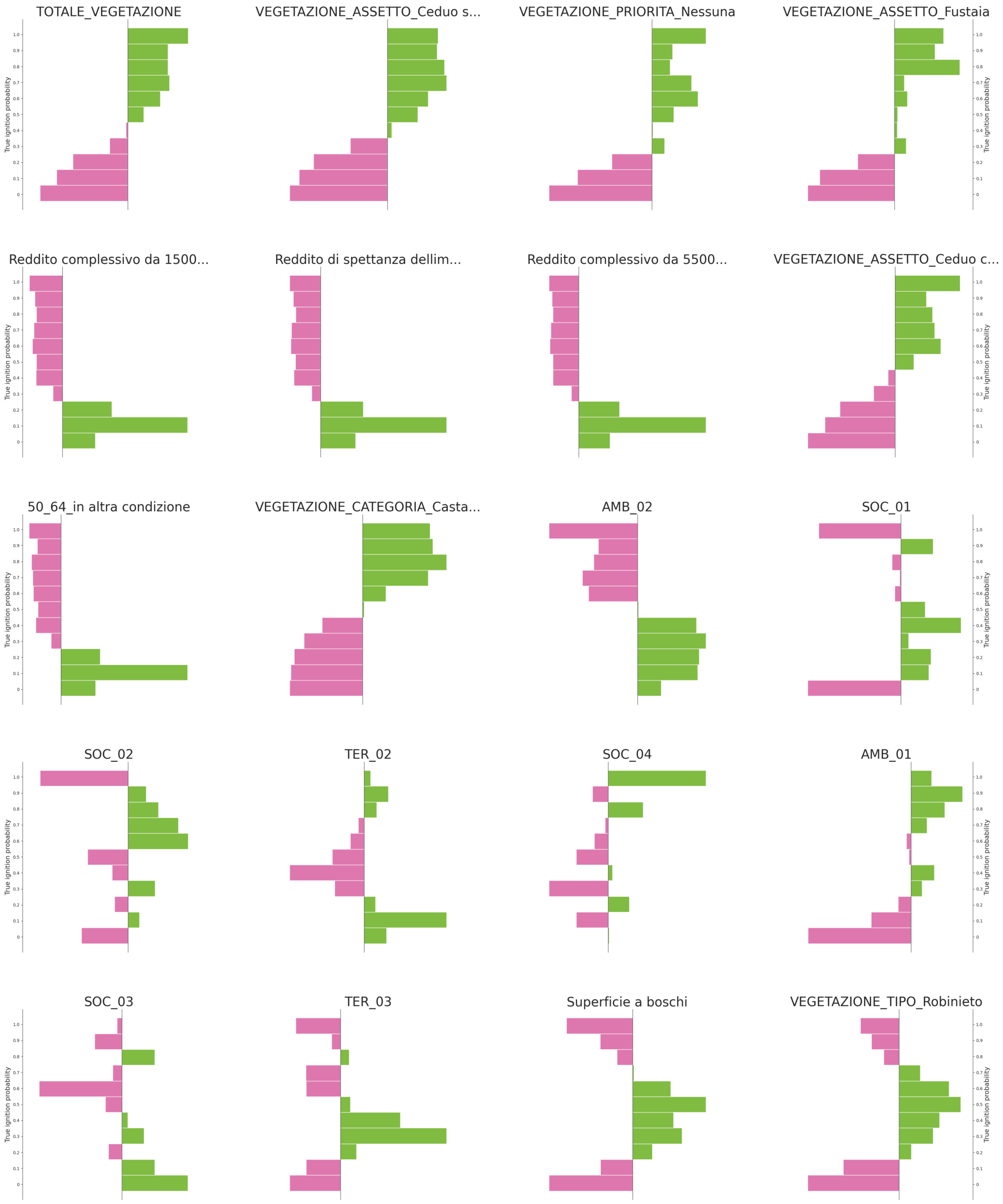
PC35	PC27
PC22	PC21
PC8	PC22
PC28	PC38
PC21	PC16
PC27	PC23
PC23	PC20
PC32	PC37
PC37	PC26
PC14	PC35
PC12	PC25
PC30	PC30
PC31	PC17
PC25	PC31
PC17	PC14
Human-related filter	
regression	**classification**
PC2	PC4
PC3	PC2
PC4	PC3
PC6	PC1
PC1	PC6
PC5	PC5
PC9	PC9
PC8	PC8
PC7	PC7

Table A.9: Feature importance for Sample C

RFE selection	
regression	**classification**
piante industriali	Precipitazione dalle 0 alle 0 (mm)_season
Addizionale comunale dovuta - Frequenza	Temperatura massima (°C)_season
Umidita' media (%)_season	Giorni piovosi pioggia dalle 0 alle 0_season
Precipitazione dalle 0 alle 0 (mm)_season	Umidita' minima (%)_season
Superficie agricola utilizzata - ettari	Umidita' media (%)_season
Velocita' media del vento (m/s)_season	Umidita' massima (%)_season
Umidita' massima (%)_season	Tempo di permanenza nel settore (min)_season
coltivazioni fruttifere	Temperatura minima (°C)_season
AMB_01	Velocita' media del vento (m/s)_season
Umidita' minima (%)_season	Settore Prevalente_season
TER_03	Direzione massima raffica (°)_season
Giorni piovosi pioggia dalle 0 alle 0_season	Radiazione totale (MJ/mq)_season
SOC_03	Addizionale comunale dovuta - Frequenza
Temperatura massima (°C)_season	AMB_01
Settore Prevalente_season	SOC_03

Radiazione totale (MJ/mq)_season	TER_03
Direzione massima raffica (°)_season	Superficie agricola utilizzata - ettari
Temperatura minima (°C)_season	Superficie a boschi
Tempo di permanenza nel settore (min)_season	coltivazioni fruttifere
Superficie a boschi	piante industriali
PCA reduction	
PC2	PC2
PC7	PC14
PC11	PC7
PC6	PC6
PC14	PC1
PC13	PC11
PC3	PC9
PC12	PC12
PC16	PC3
PC1	PC16
PC9	PC13
PC20	PC19
PC8	PC8
PC19	PC15
PC18	PC4
PC4	PC5
PC15	PC17
PC5	PC20
PC17	PC10
PC10	PC18

Human-related filter	
regression	**classification**
PC4	PC4
PC2	PC9
PC9	PC2
PC1	PC6
PC3	PC1
PC8	PC8
PC5	PC3
PC6	PC5
PC7	PC7


TOTALE_VEGETAZIONE
VEGETAZIONE_ASSETTO_Ceduo s...
VEGETAZIONE_PRIORITA_Nessuna
VEGETAZIONE_ASSETTO_Fustaia
Reddito complessivo da 1500...
Reddito di spettanza dellim...
Reddito complessivo da 5500...
VEGETAZIONE_ASSETTO_Ceduo c...
50_64_in altra condizione
VEGETAZIONE_CATEGORIA_Casta...
AMB_02
SOC_01
SOC_02
TER_02
SOC_04
AMB_01
SOC_03
TER_03
Superficie a boschi
VEGETAZIONE_TIPO_Robinieto
True ignition probability
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

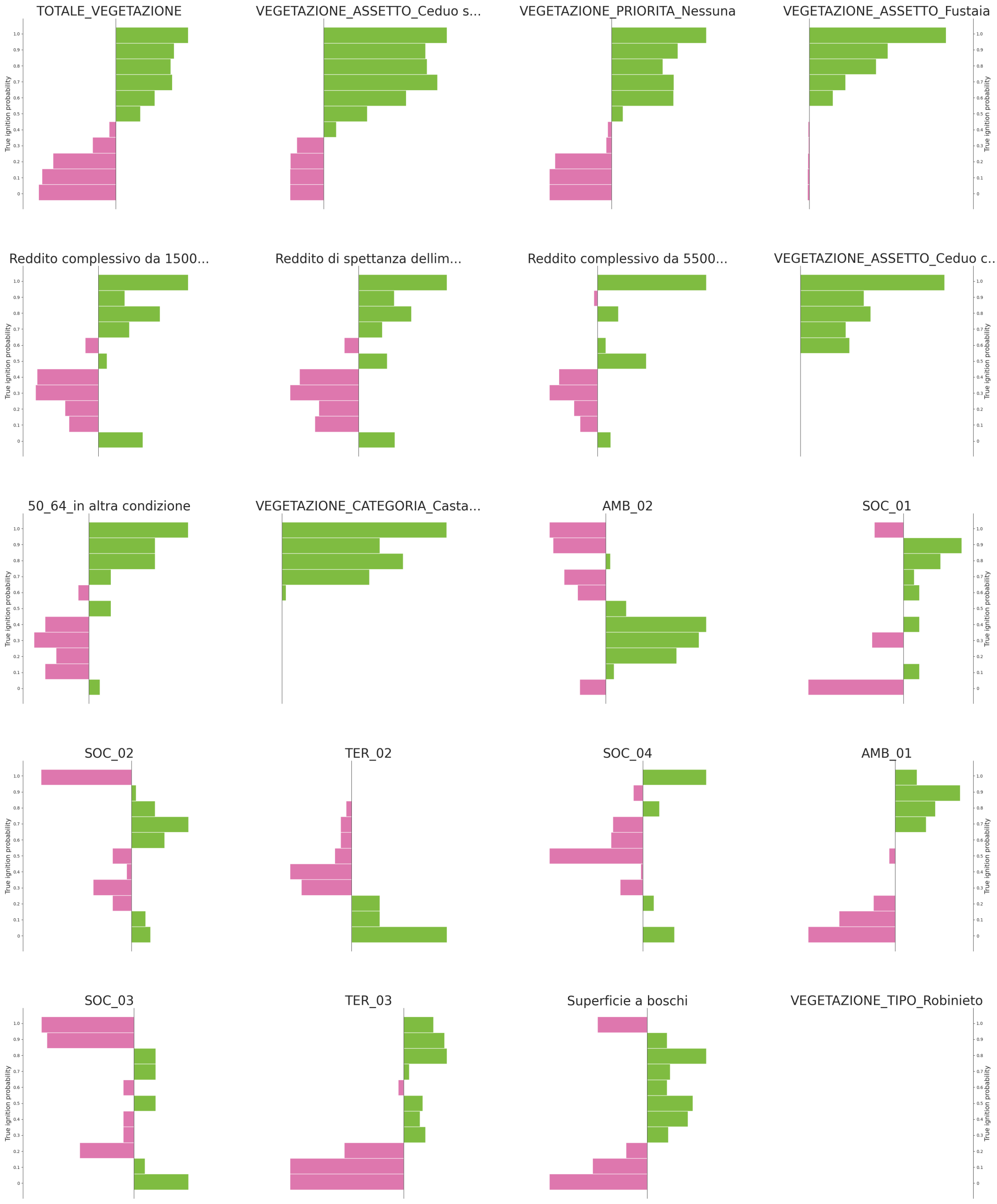


Figure A.1: The bars represent the probability class means (top) and medians (bottom) of the variable displayed in the graph with respect to the overall variable's mean/median (the central vertical line). The pink bars identify the classes whose variable values are averagely lower than the overall average variable value, in green the classes whose variable values are averagely higher than the overall average variable value