A Machine Learning-based approach for mapping Wildfire Risk in Italy

Iris Haake (Matr.-Nr. 3647626) [1] & Stefanie Kern (Matr.-Nr. 3677753) [2]

[1] Faculty of Geography (FB19), Philipps-Universität Marburg, 35037 Marburg, Germany - Haakei@students.uni-marburg.de
[2] Faculty of Geography (FB19), Philipps-Universität Marburg, 35037 Marburg, Germany - Kernst@students.uni-marburg.de

ABSTRACT

As a result of global warming, drought and heat waves dominate summer events in Central/Southern Europe and forest fires are increasing in intensity and frequency. In this study, the fire occurrence and probability in Italy in July 2022 is investigated. Italy can be divided into three areas with different climatic, topographic, and ecological characteristics, which have a unique susceptibility to wildfires. To predict the probability of a fire event in Italy, variables on climatic, topographical, and anthropogenic conditions are used. This is based on the Random Forest machine learning method. With the help of the prediction variables and historical data on forest fire events, this model is trained, validated, and applied for July 2022. Especially in southern Italy, Liguria, Sicily and Sardinia, there is an increased risk of forest fires. Geographical location, proximity to people and climatic factors have a major influence. The model has a good prediction accuracy, although the fire pixels are overestimated. The selection and importance of predictor variables needs to be further investigated and compared to other model algorithms. Mapping wildfire probability for a single month can help adapting strategies to manage wildfires during periods especially at risk.

KEYWORDS

Fire Risk, Wildfire, Italy, Mediterranean Region, Random Forest, Machine Learning, Feature Importance

1. INTRODUCTION

Southern Europe has been experiencing a stable high pressure area for several weeks, causing persistent drought, heat waves reaching up to 46 °C and an elevated risk of wildfires (ECMWF 2023). The areas most affected include Spain, southern Italy, Sardinia and Sicily (WMO 2023). According to ECMWF 2023, both June and the first half of July 2023 show the highest temperatures in Europe recorded to date. The vulnerability to wildfires is also reflected in the number of wildfires in Italy in 2022 (Fig. 1), which is significantly higher in June and July compared to the long-term average from 2006-2022. In 2022, a total of 58,751 ha of forest burned in Italy (EFFIS 2023). The increasing risk of wildfires due to extreme weather conditions is an enormous challenge for the population. With the help of models, areas at risk of wildfires can be identified and prepared for such events in a preventive manner.

Machine learning has been used in previous studies to identify environmental hazards and risks, as these are usually caused by a non-linear interaction of several variables and machine learning models such as Random Forest are well suited to reflect these relationships (Tonini et al. 2020; Trucchia et al. 2022; Iban & Sekertekin 2022). Previous studies investigated the susceptibility of the Italian region of Liguria (Tonini et al. 2020) and all of Italy (Trucchia et al. 2022) to wildfire. Susceptibility to wildfire has also been analyzed in other Mediterranean countries such as Portugal (Carmo et al. 2011) and Turkey (Iban & Sekertekin 2022) as well as for the whole eastern Mediterranean landscape including 13 countries (Trucchia et al. 2023). Several variables enhancing the risk of wildfires in the Mediterranean region have been identified by the named studies. An important factor is the proximity to anthropogenic features such as urban areas, agriculture, roads, recreation areas and pathways, as most wildfires in the Mediterranean region are ignited by humans (Ganteaume et al. 2013), in Italy even 98 % of all wildfires (Trucchia et al. 2022). Geo-environmental variables that have an impact on fire risk include land cover, location, and topography, more specifically elevation, slope, and aspect. Meteorological factors such as mean temperature, total precipitation, wind speed, wind direction and humidity were also stated.

The studies by Tonini et al. 2020 and Trucchia et al. 2022, which analyzed the susceptibility to wildfire for Liguria and the whole of Italy respectively, considered the situation in winter and summer on a spatial scale. However, no study has yet examined the wildfire risk in a clearly defined month. In Italy, most of the wildfires of the year have historically occurred in July (Fig. 1) and

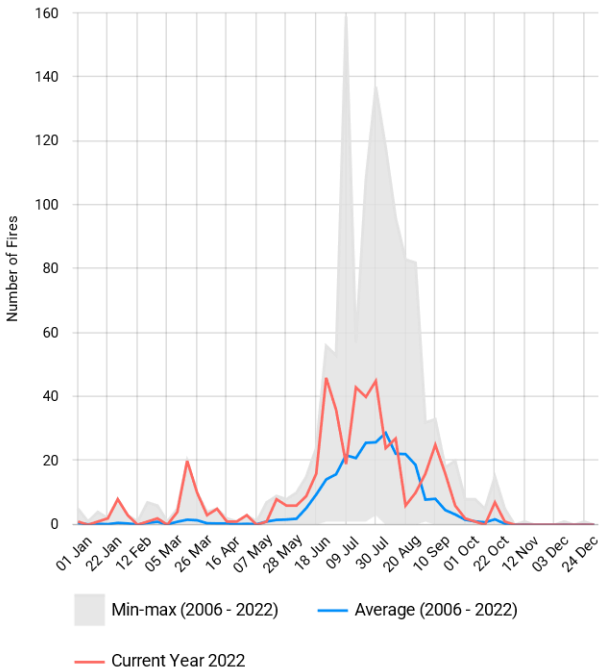

Figure 1: Weekly number of fires in Italy 2022 compared to the long-term average from 2006-2022 (EFFIS 2023).

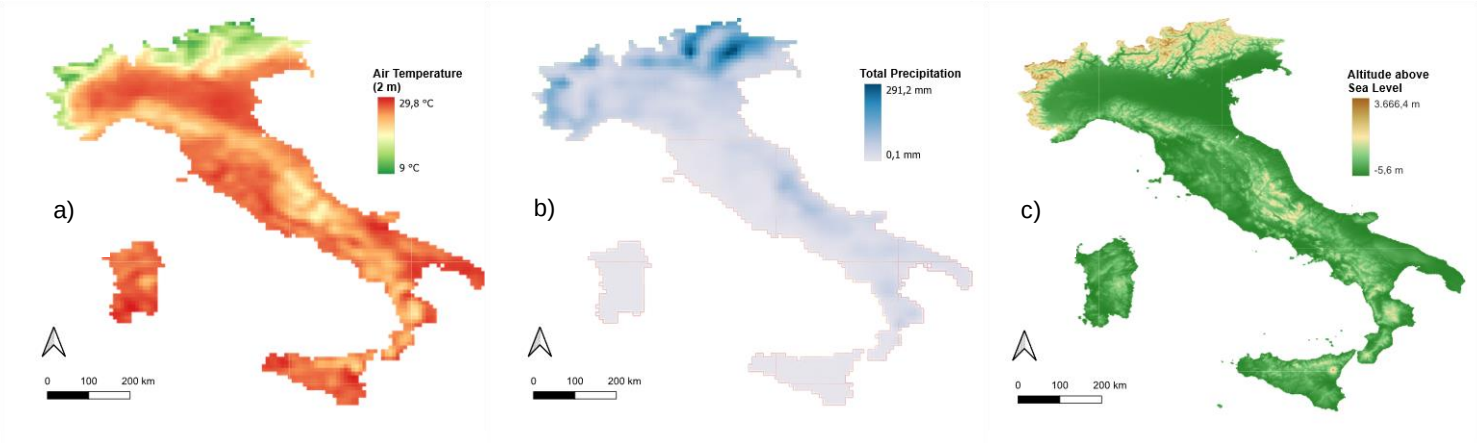

Figure 2: a) Elevation (derived from Copernicus DEM), b) Mean air temperature in July 2022, measured 2 m above ground (derived from Copernicus Climate Data Store), c) Total precipitation in July 2022 (derived from Copernicus Climate Data Store)

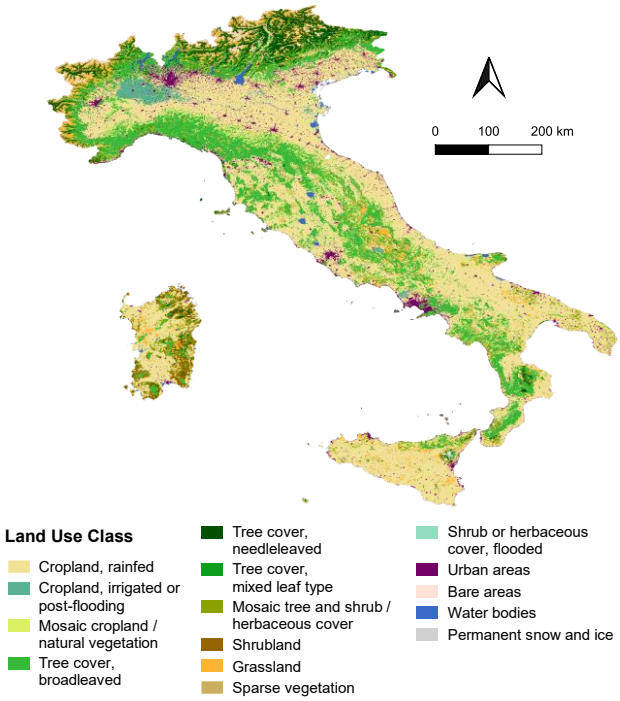

Figure 3: Land Use Cover 2020 (derived from Copernicus Land Cover Data)

a separate analysis of this month is necessary to better prepare for wildfire events. Therefore, the present study analyses the risk of wildfires in July 2022 for the whole of Italy. The definition of fire risk used is the one provided by Tonini et al. 2020, which defines fire risk as a "quantitative [...] indicator of the likelihood that an area would burn in a certain period of time" (p. 2).

The present study also includes more climatic and topographic variables, as well as the NDVI, to better represent the complex factors that favor wildfires. In summary, this study investigated the research question: What is the wildfire risk in Italy in July 2022 and which variables influence it to what extent?

2. Study Area

The study area is the country of Italy, located in southern Europe, with the islands of Sardinia and Sicily. Italy is surrounded by the Mediterranean Sea on three sides and is essentially divided into three areas with different climatic, ecological, and topographical characteristics (Fig. 2 and 3). In the Italian Alps, the highest region (around 1000 to 3666 m), the predominant land cover types are coniferous and mixed forests, bare areas, and grasslands. In July 2022, mean temperatures ranged from 9 to 16 °C and total precipitation from 130 to 290 mm. The second division is the Apennine Mountain range (around 400 to 2700 m), which stretches across Italy from north to south. It is characterized by broad-leaved trees and grassland, with an average temperature of 19-25 °C and precipitation of 20-100 mm in July 2022. The third division (around 0 to 400 m) includes the Po plain in the north, the coastal areas, and the islands. These regions consist mainly of cropland, with occasional scrubland, and the largest urban areas are located here. The Po Valley had slightly higher rainfall in July 2022, with 20-60 mm, while the remaining areas had 20 mm or less, with many regions receiving no rain at all. This division had the highest average temperatures of 26-30 °C.

3. Data

To predict the risk of wildfires in Italy, variables related to climate, land use and terrain characteristics are selected and further processed at different data levels. In addition to the raster data, a CSV-file with detailed information on historical fire events is used. All data used are summarised in Table 1 with their characteristics and source.

3.1 Fire Events

Historical fire events in Italy are used to derive the target variable “fire status” for the prediction. The raw data originally has a temporal resolution of one day, so an aggregation to 4 days is done. A reclassification of the scan (< 1.5) and confidence (>= 30) parameters is necessary, as these otherwise lead to bias and a lower quality of the prediction of fire in the model (NASA 2021). In contrast, if too many pixels are removed, information can be lost and thus negatively affect the validity and reliability of the model. In a final step, the target variable fire status (0: no fire, 1: fire) is derived for each pixel from the processed fire data set.

3.2 Raster

In addition to fire status as the target variable, the following independent variables are added to predict fire areas: Land Cover (LUC), NDVI, Land Surface Temperature (LST), climate variables (ERA5) and Digital Elevation Model (DEM). Further

variables were derived from these original data. Thus, proximity of agricultural areas and proximity of urban areas were derived from the LUC data; relative humidity, vapor pressure, saturation vapor pressure and vapor pressure deficit (VPD) were calculated from the ERA5 data; aspect, slope, curvature and tpi were derived from the DEM.

To create a dataset with all variables, a reprojection to the reference grid of the NDVI (WGS84 and Study Area), an upscaling/downscaling to obtain the same spatial (1 km) as well as temporal resolution (monthly) is performed. With the reclassification of LUC (classes between 30 and 150), NDVI (>= -0.2) and LST (>= 200 K), only value ranges that influence forest fire occurrence will be included in the analysis. To avoid autocorrelations, some variables (i.e., evaporation, vapor pressure) with high correlation values (> 0.90) are removed from the dataset.

3.3 Final Dataset

Based on the extraction of all pixel values of the independent variables from the respective fire/non-fire coordinates of the fire events, a final data set (CSV file) can be generated. Then the target variable fire status is appended to data set. To apply the trained model to predict wildfire risk in July 2022, a separate dataset is created with all pixel values for that time period.

Table 1: Datasets used to prepare the prediction dataset, including characteristics and source.

Variables	Format	Coverage	Resolution		Time Span	Source
			Spatial	Temporal		
Fire Data	CSV	Italy	1 km	Daily	2001-01 to 2022-12	NASA
Land Use Classes	NetCDF	Global	300 m	Yearly	2001-2020	Copernicus
NDVI		Italy	1 km	Monthly	2001-01 to 2022-12	NASA USGS download via AppEEARS
Land Surface Temperature			1 km	8-Day		NASA USGS download via AppEEARS
ERA5 Climate Data			0.1° (~ 9 km)	Monthly		Copernicus Climate Data Store
Digital Elevation Model		Europe	25 m	-	2011	Copernicus

4. METHODS

4.1 Random Forest

To make predictions of fire/non-fire occurrence and fire probability in July 2022, first, a model based on training data (2001-2017) needs to be built. Subsequently, this data set is again divided into training data (80%) and validation data (20%). This serves to train the model using the training data and to check its accuracy with the validation data. For this purpose, the Random Forest algorithm is used. This method works using randomised decision trees. Each of them decides based on the input data whether there is a fire (1) or no fire (0). Due to the averaging of all classification results, the variance can be minimised, and overfitting reduced (Scikit-Learn 2023a). To validate the RF-model (Chapter 4.3), the independent test dataset (2018-2022) is taken to account.

4.2 Selection of important features

To select the features that are relevant for prediction, the feature importance was first calculated. The importance is determined by the mean decrease in impurity, which includes the fraction of the input sample to which the feature contributes, as well as the decrease in impurity that occurs in all splits of the trees made based on that feature (Scikit-Learn 2023b). Forward feature selection was then performed with a predefined number of 19 features. The feature that achieved the best cross-validation score was iteratively added to the set of predictor variables until the set contained 19 variables (Scikit-Learn 2023c). With these 19 variables, the final model was built and applied to the July 2022 dataset to determine fire occurrence and probability.

4.3 Confusion Matrix

A confusion matrix is used to validate the model accuracy. The focus is on the target variable fire status. The frequency of correct/incorrect prediction of fire/non-fire events is determined by comparing the results with the actual values from the independent test data set (2018-2022). In this case of two classes, there are four options to predict. If an actual fire pixel is predicted to be a fire pixel, it is true positive (TP). In contrast, if an actual fire pixel is predicted to be a non-fire pixel, it is a false negative (FN). If an actual non-fire pixel is predicted to be a non-fire pixel, it is called true negative (TN). Lastly, when a non-fire pixel is predicted to be a fire pixel, this is referred to as a false positive (FP). For good model prediction performance, TP and TN should have the highest number of predicted pixel values and thus FP and FN the lowest. Further metrics can be derived from the confusion matrix, which contribute to the assessment of model quality. These differ in the accounting of the four categories (TP, TN, FP, FN) and are defined as follows (Scikit-Learn 2023d; Arias-Duart et al. 2023, p. 3710):

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

$$F1 = (2 * TP) / (2 * TP + FP + FN) \quad (4)$$

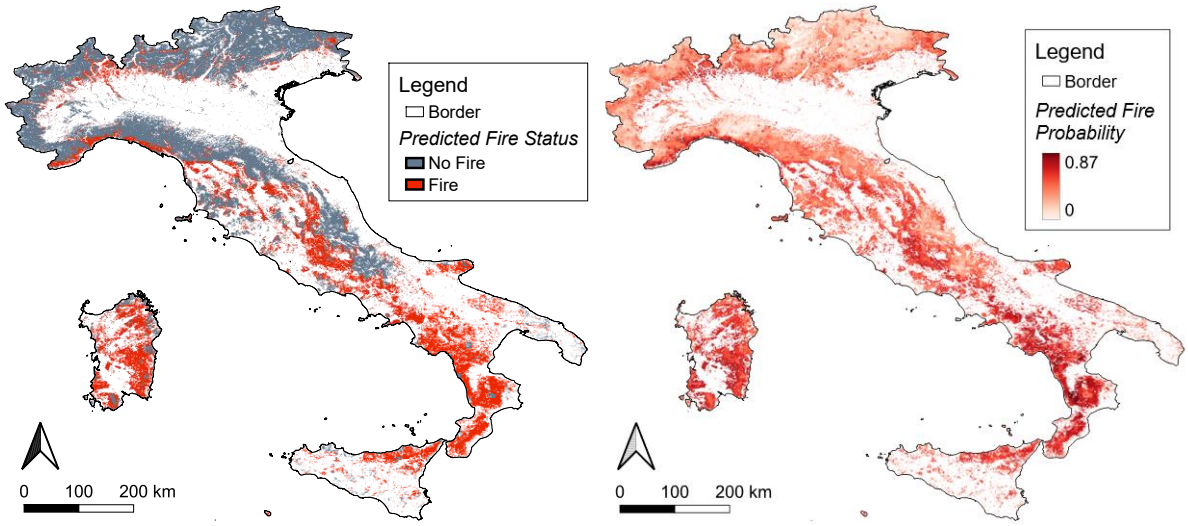

Figure 4: Predicted Fire Status and Fire Probability in July 2022.

5. Results

5.1 Fire Risk

Both the map of predicted fire status and the map of forest fire probability (Fig. 4) serve to classify the risk of forest fire in Italy. When looking at the predicted classes, it becomes clear that there is a north-south contrast. In the regions around the Italian Alps predominantly no fire is predicted, whereas in Sardinia, Sicily, and southern Italy in particular, fire is the dominant class. The Ligurian coast, which falls into the fire class, is particularly notable. If the probability of wildfires is also considered, Liguria has one of the highest risks. Southern Sardinia, eastern Sardinia and the regions of Calabria and Basilicata also have a high probability of wildfires.

5.2 Feature Importance and Forward Feature Selection

According to the feature importance (Fig. 5), variables related to geographical location, such as x and y coordinates, DEM, slope and proximity of urban areas, as well as climatic conditions, including surface pressure (*sp*), surface latent heat flux (*slhf*), wind speed (*v10* and *u10*), surface sensible heat flux (*sshf*), skin reservoir content (*src*), total precipitation (*tp*), air temperature (*t2m*) and relative humidity (*rel_hum*), are particularly important for predicting wildfires. On the other hand, NDVI, solar radiation (*str*, *ssr*) and land cover, among others, have less influence.

The following variables were selected by the forward feature selection and included in the model: x and y coordinates, land use class (*lccs_class*), air temperature (*t2m*), skin reservoir content (*src*), surface latent heat flux (*slhf*), solar radiation (*str*), surface pressure (*sp*), surface sensible heat flux (sshf), total precipitation (*tp*), volumetric soil water (*swvl1*), relative humidity (*rel_hum*), vapor pressure deficit (*VPD*), land surface temperature (*LST*), proximity to agriculture, proximity to urban areas, aspect, slope, tpi.

5.3 Quality Assessment

The reliability of the model can be determined via a Confusion Matrix (Fig. 6) and the four metrics accuracy, precision, recall and F1 (Tab. 2). From the Confusion Matrix, the correctly predicted fire as well as non-fire pixels, holding a total of 1283 pixels, represent a large majority. According to the accuracy, these comprise about 83 % of the total sample number of 1542 pixels. Considering the incorrectly predicted pixels (FP, FN), almost twice as many actual non-fire pixels were classified as fire pixels as vice versa. Since these misclassifications are included in the other metrics, this results in lower average model quality values of 0.79 for precision, 0.76 for recall and 0.77 for F1. In general, the accuracy of all indices is significantly worse for non-fire pixels (0) compared to fire pixels (1). For the weighted mean, all measures of model reliability calculate a value of 0.83.

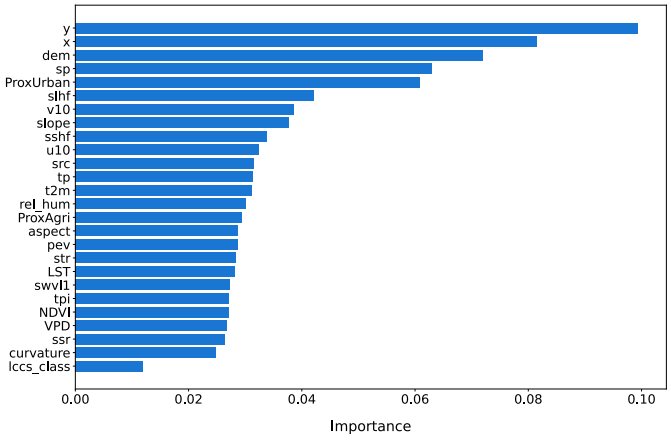

Figure 5: Feature Importances of predictor variables.

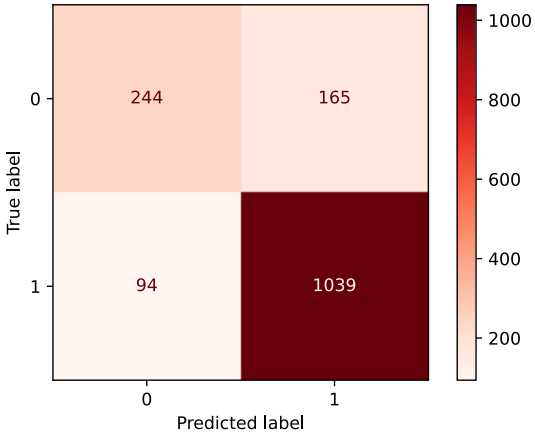

Figure 6: Confusion Matrix of the Random Forest Model.

Table 2: Metrics of the Random Forest model.

		Precision	Recall	F1	Support
Fire Status	0	0.72	0.60	0.65	409
	1	0.86	0.92	0.89	1133
Accuracy				0.83	1542
Macro average		0.79	0.76	0.77	1542
Weighted average		0.83	0.83	0.83	1542

6. Discussion

Comparing the results with the land use class map (Fig. 3 and 4), it is noticeable that most fires were predicted for areas of deciduous forest, grassland and shrubland. It seems strange that the model did not classify the land use class as important (Fig. 5). Trucchia et al. (2023) also classified this variable as unimportant, but it was of significance in Trucchia et al. 2022 and Carmo et al. 2011. In addition, Tonini et al. 2020 and Trucchia et al. 2022 introduced neighboring vegetation as a variable associated with land use and found that the highest fire risk was associated with shrubland and rural areas as neighboring vegetation.

The areas with the highest fire occurrence and probability in the present study were those with little to no precipitation (Fig. 2b) and intermediate temperatures (Fig. 2a). Even these variables were considered less important by the model than other climate variables such as surface pressure, surface latent heat flux and wind speed, although some of these variables are of course related to air temperature. However, the north-south difference in temperature and precipitation is also reflected in fire occurrence and risk. In previous studies (Trucchia et al. 2022; Trucchia et al. 2023), temperature and precipitation were considered to be much more important. The pattern of predictions (Fig. 4) is strongly oriented towards elevation (Fig. 2c), which was classified as the most important variable after coordinates. There is a high risk of fire at middle altitudes, whereas the risk is lower in the Alps and in the lowlands. Other studies have also identified elevation as an important predictor variable (Carmo et al. 2011; Trucchia et al. 2022; Trucchia et al. 2023). It should be noted that when the model was applied to the July dataset, many pixels were generated without values. These NA values are colored white in the result maps (Fig. 4) and have an impact on the interpretation, as the influence of the predictor variables on these areas is unknown and, for example, the areas with the highest temperature values are often NA. Other variables considered important by the model are slope and proximity to urban areas. This was also supported by Carmo et al. (2011), Trucchia et al. (2022) and Trucchia et al. (2023), the latter also identifying population density and proximity to roads as important human-related variables.

Prediction quality for fire occurrence and risk heavily relies on variables and the chosen model algorithm. Although Random Forest is well suited to answering the research question, including several model algorithms and feature selection methods could fine-tune the choice of predictor variables. There were some discrepancies between the feature importance and the forward feature selection. Variables like elevation and windspeed were deemed important, but not included in the model. On the other hand, land cover was the least important variable, but was included in the model. In addition, the temporal dimension of the datasets used needs to be considered. Training data spanned from 2001 to 2017, with 2018 to 2022 as the test set. However, especially since 2018, there have been drought years in Italy and the fire situation has become more severe (EFFIS 2023). This was not considered when building the model. Furthermore, the training dataset's uneven pixel numbers (6415 'fire' pixels vs. 2509 'no fire' pixels) led to overfitting, favoring fire predictions. This contributed to overclassified fire pixels in maps (Fig. 4) and imbalanced F1 scores (0.89 for fire and 0.65 for non-fire, see Tab. 2). With an overall accuracy of 0.83, the model used is useful for getting a first impression of the distribution of fire risk in July 2022, and with some adjustments to the prediction dataset, the predictive power could be increased.

7. Conclusion

This study analyzed Italy's wildfire risk in July 2022 and the factors influencing it. Using the Random Forest algorithm, the research aimed to predict fire occurrence and probability using climate, topography, and land-use variables.

The findings showcased distinct wildfire patterns across Italy. Northern areas, like the Italian Alps, had lower fire risk than southern regions such as Sardinia, Sicily, and southern Italy, with Liguria also at risk. Variables like coordinates, slope, and urban proximity played key roles in predicting wildfires. Climatic conditions, including surface pressure, heat flux, and humidity, also influenced fire likelihood.

However, the model's feature analysis showed differences from prior studies. Land cover and elevation, although significant in other studies, seemed to have a lesser influence on the current model's predictions. Similarly, temperature and precipitation, while important, were not as dominant as in certain prior studies.

While the methodology used in this study revealed some overfitting concerns, leading to potential overestimation of fire occurrences, the model still offered valuable insights into July 2022's fire risk distribution.

In conclusion, this study illuminates the complex factors driving wildfire risk in Italy. By focusing on a specific month, it deepens our understanding of interactions between climate, topography, and human actions. Results can guide strategies for managing wildfire threats in the Mediterranean, especially during heightened risk periods.

Supplementary Material

All programming source scripts as well as raw and processed data have been made available in the corresponding GitHub repository.

References

Arias-Duart, A. / Mariotti, E. / Garcia-Gasulla, D. / Alonso-Moral, J. M. (2023): A Confusion Matrix for Evaluating Feature Attribution Methods. Provided by: Computer Vision Foundation, p. 3708-3713.

Carmo, M. / Moreira, F. / Casimiro, P. / Vaz, P. (2011): Land use and topography influences on wildfire occurrence in northern Portugal. In: Landscape and Urban Planning 100(1-2), p. 169-176 (doi).

Copernicus (2016): EU-DEM v1.1. Available online: https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1?tab=download (retrieved on 15.08.2023).

Copernicus (2020): Land cover classification gridded maps from 1992 to present derived from satellite observations. Available online: https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-land-cover?tab=overview (retrieved on 15.08.2023).

Copernicus (2023): ERA5-Land monthly averaged data from 1950 to present. Available online: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land-monthly-means?tab=overview (retrieved on 15.08.2023).

ECMWF (2023): European heatwave July 2023. Available online: https://www.ecmwf.int/en/about/media-centre/science-blog/2023/european-heatwave-july-2023 (retrieved on 25.07.2023).

EFFIS (2023): Seasonal Trend for Italy. Available online: https://effis.jrc.ec.europa.eu/apps/effis.statistics/seasonaltrend (retrieved on 25.07.2023).

Ganteaume, A. / Camia, A. / Jappiot, M. / San-Miguel-Aynaz, J. / Long-Fournel, M. / Lampin, C. (2013): A Review of the Main Driving Factors of Forest Fire Ignition Over Europe. In. Environmental Management 51, p. 651-662 (doi).

Iban, M.C. / Sekertekin, A. (2022): Machine learning based wildfire susceptibility mapping using remotely sensed fire data and GIS: A case study of Adana and Mersin provinces, Turkey. In: Ecological Informatics 69, p. 101647 (doi).

NASA (2021): MCD14DL-NRT. Available online: https://www.earthdata.nasa.gov/learn/find-data/near-real-time/firms/mcd14dl-nrt#ed-firms-attributes (retrieved on 22.07.2023).

NASA (2023a): Fire Information for Resource Management System. Available online: https://firms.modaps.eosdis.nasa.gov/download/ (retrieved on 15.08.2023).

NASA (2023b): MOD13A3v061. MODIS/Terra Vegetation Indices Monthly L3 Global 1 km SIN Grid. Available online: https://lpdaac.usgs.gov/products/mod13a3v061/ (retrieved on 15.08.2023).

NASA (2023c): MOD11A2v061. MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1 km SIN Grid. Available online: https://lpdaac.usgs.gov/products/mod11a2v061/ (retrieved on 15.08.2023).

Scikit-Learn (2023a): Forests of randomized trees. Available online: https://scikit-learn.org/stable/modules/ensemble.html#random-forests (retrieved on 23.07.2023).

Scikit-Learn (2023b): Feature Importance Evaluation. Available online: https://scikit-learn.org/stable/modules/ensemble.html#feature-importance-evaluation (retrieved on 13.08.2023).

Scikit-Learn (2023c): Sequential Feature Selection. Available online: https://scikit-learn.org/stable/modules/feature_selection.html#sequential-feature-selection (retrieved on 13.08.2023).

Scikit-Learn (2023d): Confusion Matrix. Available online: https://scikit-learn.org/stable/modules/model_evaluation.html#confusion-matrix (retrieved on 23.07.2023).

Tonini, M. / D'Andrea, M. / Biondi, G. / Degli Esposti, S. / Trucchia, A. / Fiorucci, P. (2020): A Machine Learning-Based Approach for Wildfire Susceptibility Mapping. The Case Study of the Liguria Region in Italy. In: Geosciences 10(3), p. 105 (doi).

Trucchia, A. / Meschi, G. / Fiorucci, P. / Gollini, A. / Negro, D. (2022): Defining Wildfire Susceptibility Maps in Italy for Understanding Seasonal Wildfire Regimes at the National Level. In: Fire 5(1), p. 30 (doi).

Trucchia, A. / Meschi, G. / Fiorucci, P. / Provenzale, A. / Tonini, M. / Pernice, U. (2023): Wildfire hazard mapping in the eastern Mediterranean landscape. In: International Journal of Wildland Fire 32(3), p. 417-434 (doi).

WMO (2023): Simultaneous heatwaves hit northern hemisphere in summer of extremes. Available online: https://public.wmo.int/en/media/news/simultaneous-heatwaves-hit-northern-hemisphere-summer-of-extremes (retrieved on 25.07.2023).