Type of the Paper (Article)

A machine learning based approach for wildfire susceptibility mapping. The case study of Liguria region in Italy.

Marj Tonini[1,*], Mirko D'Andrea[2], Guido Biondi[2], Silvia Degli Esposti[2], Andrea Trucchia[2], Paolo Fiorucci[2]

[1] Institute of Earth Surface Dynamics, Faculty of Geosciences and Environment, University of Lausanne, marj.tonini@unil.ch

[2] CIMA Research Foundation; paolo.fiorucci@cimafoundation.org

* Correspondence: marj.tonini@unil.ch; +41 21 692 35 37

Abstract: Wildfire susceptibility maps display the wildfires occurrence probability, ranked from low to high, under a given environmental context. Current studies in this field often rely on expert knowledge, including or not statistical models allowing to assess the cause-effect correlation. Machine learning (ML) algorithms can perform very well and be more generalizable thanks to their capability of learning from and make predictions on data. Italy is highly affected by wildfires due to the high heterogeneity of the territory and to the predisposing meteorological conditions. The main objective of the present study is to elaborate a wildfire susceptibility map for Liguria region (Italy) by applying Random Forest, an ensemble ML algorithm based on decision trees. Susceptibility was assessed by evaluating the probability for an area to burn in the future considering where wildfires occurred in the past and which are the geo-environmental factors that favor their spread. Different models were compared, including or not the neighboring vegetation and using an increasing number of folds for the spatial-cross validation. Susceptibility maps for the two fire seasons were finally elaborated and validated and results critically discussed highlighting the capacity of the proposed approach to identify the efficiency of fire fighting activities.

Keywords: wildfires; susceptibility mapping; machine learning; random forest; model validation; Liguria region

1. Introduction

Mapping current and past hazardous events, such as landslides, flooding, or wildfires, represents a precious information for addressing prevention-planning programs aiming to reduce human and material losses. Raw information is generally stored on national and regional multi-temporal inventories reporting the occurrences from multiple source, aa paper datasheets, field surveys, interpretation of aerial photograph and, more recently, remotely sensed imagery. These inventories represent key data for the elaboration of hazard and risk maps. Namely, hazard maps portray the zonation of the spatio-temporal probability of events, while the expected damages or losses are assessed and represented on risk maps. Furthermore, according to the assumption that future events are expected to occur under similar conditions as the observed past ones, susceptibility map indicates zones with a potential to experience a particular hazard in the future based solely on the intrinsic local proprieties of a site and expressed in term of relative spatial likelihood. Although these

concepts are well-consolidated in the research area related with the risk assessment, especially for landslides [1–4] the need exist for elaborating susceptibility and risk maps for other natural hazards and to develop new quantitative and robust methods supporting their production.

In the field of wildfires risk assessment, fire risk has been defined as a quantitative or qualitative indicator of the probability of an area to be source of ignition by natural or artificial means in a certain period of time [5–8]. In this context, modeling fire risk represents a modern tool to support forest protection plans and to address fuel management strategies in order to reduce fires' consequences [9–11]. More in general, risk and susceptibility analyses are of great importance for land use planning, civil protection and risk reduction programs. A number of techniques were recently developed to monitoring and mapping the spatial distribution of burned area and to predict area at risk for wildfires. These, often, involve the implementation of physically based models integrated into a Geographic Information System (GIS), relying on expert knowledge to estimate the predisposing factors or including statistical analyses and modeling to assess the variables' importance [12–22]. Lately, the comparison of deterministic physically/statistically based and stochastics approaches, highlighted the benefit of using data driven methods [23–31] able to extract knowledge directly from data. In comparison with deterministic methods, which, given a set of initial conditions (i.e. predisposing factors) always perform the same results, stochastics models assume that results obtained by the combination of independent predisposing factors can be slightly different as a consequence of the randomness of the process.

Although the terms "risk" and "susceptibility" are often used as synonymous, hereinafter we refer to susceptibility mapping meaning that the propensity of an area to wildfire occurrence is assessed with no consideration for the magnitude of a single event or its temporal dimension. Indeed, only the spatial probability for an area to burn in the future, assessed by defining a rank from low to high, is evaluated. This quantitative evaluation is performed considering two aspects: where wildfires occurred in the past, in terms of burned area, and which are the geo-environmental and anthropogenic predisposing factors that favor their spread. In this regard, it is worth noting that meteorological factors, like wind speed and wind direction, temperature, humidity, and rainfall are considered as triggering and not predisposing factors. The trigger is a local condition that cause a risk to occur if and only if the area is susceptible to that risk, while the susceptibility is assessed based only on predisposing factors, which are stable over time. The proximity to road and pathway networks and to urban and recreation areas are the most frequently mentioned as predisposing human factors for wildfire [32–39]. As regards geo-environmental variables, those related vegetation type and topography result to be the most significant drivers, especially in Mediterranean-type regions [40–43].

Italy is particularly affected by wildfires, because of the high topographic and vegetation heterogeneity of the territory, as well as the favorable climatic conditions that characterize the entire Mediterranean basin. Wildfires are more frequent and larger in summer season (May - October) than in winter season (November - April) in almost all the Mediterranean countries, since the first is hotter and dryer. Liguria region, in Italy, represents an exception because it is highly affected by wildfires during the entire year and the number of wildfires and burned area can be higher in winter than in summer. Nevertheless, the spatial distribution of burned areas differs in the two seasons, probably due to the vegetation phenology at different altitudes in terms of plant senescence.

Therefore, it is important to assess the wildfires susceptibility of this region separately for the summer and the winter season.

In this study, we adopted a stochastic approach based on machine learning algorithm (ML) to elaborate wildfire susceptibility mapping for Liguria region. ML includes a class of algorithms for the analysis, modelling and visualization of environmental data, and performs particularly well to model environmental hazard, which naturally present a complex and non-linear behavior [44,45]. Our model includes the inspection and selection of predisposing factors acting as independent variables. Specifically, two models were compared, including or not the neighboring vegetation and using an increasing number of folds for the spatial-cross validation.

2. Study Area

The study area is the administrative Region of Liguria (Italy). It covers a total area of 5400 km^2 lying between the Cote d'Azur (France) and Tuscany (Italy) on the northwest coast of the Tyrrhenian Sea. This Mediterranean region is characterized by complex topography, with a slope higher than 40% for the 50% of the total area, and dense vegetation, with more than 70% of the total area covered by forests (Figure 1).

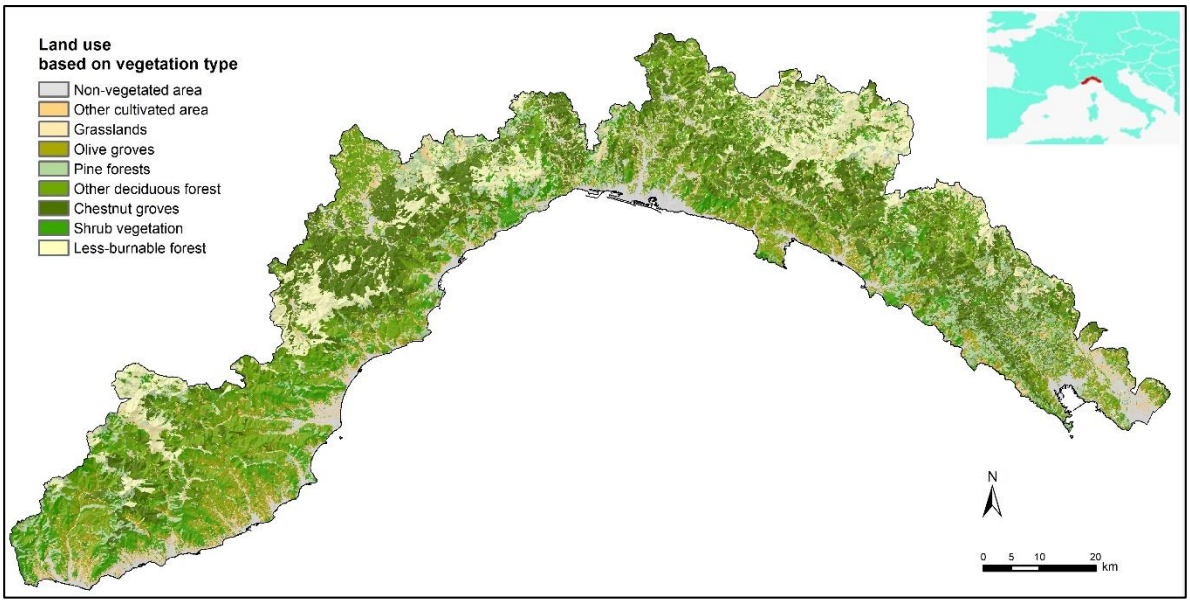

Figure 1. Study areas with the location map. Land use results from the aggregation of the original classes represented on the regional map of forest types, provided by Liguria Region.

In the investigated period, 1997-2017, an average of 365 wildfires burn an area of 55 km^2 per year and are a recurrent phenomenon both in summer and winter season. Winter fire regime is mainly due to frequent extremely dry winds from the north in condition of curing for most of the herbaceous species and the number of wildfires and burned area can be higher than in summer (Figure 2).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2020 doi:10.20944/preprints202001.0385.v1

Peer-reviewed version available at Geosciences 2020, 10, 105; doi:10.3390/geosciences10030105

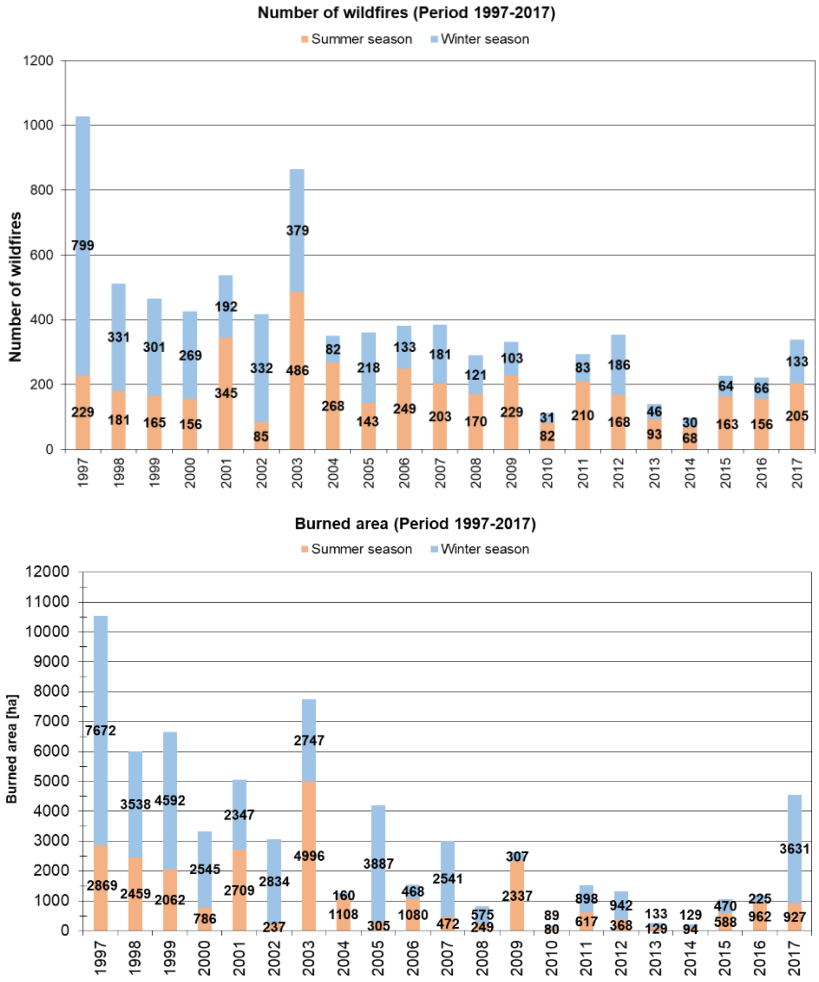

Figure 2. Yearly frequency of wildfires (on the top) and burned area (on the bottom) in Liguria region (Italy) in the last two decades (1997 – 2017) during the summer and the winter season.

Since the 1950 forest covers were limited to 30% of the total area, because most of the areas were subject to grazing activities. After the Second World War, rural communities were engaged in many reforestation programs using different Mediterranean pine species. The widespread abandon of agricultural and grazing activities lead to a large spread of pines and shrub species, frequently affected by fires. The urban development and rural abandonment lead the region to face a very large extension of the Wildland Urban Interface.

3. Materials and Methods

3.1. Dataset: burned area and predisposing factors

The dependent variable of our model, allowing to assess the susceptibility of the region to wildfires, is the burned area available as mapped fire perimeters and spanning a 20-years period (1997-2017). This dataset have been acquired and elaborated as shapefile format by the regional forestry service, on the base of GPS-survey and subsequent digitalization over the cadastral map (scale 1:10 000). It is worth observing that fire perimeters are also affected by the capacity of intervention which has not been considered in the analysis because the lack of homogenized data.

The seasonality of the fire regime was also considered, partitioning the dataset of burned areas in two macro seasons: winter (November-April) and summer (May-October) (Figure 3)

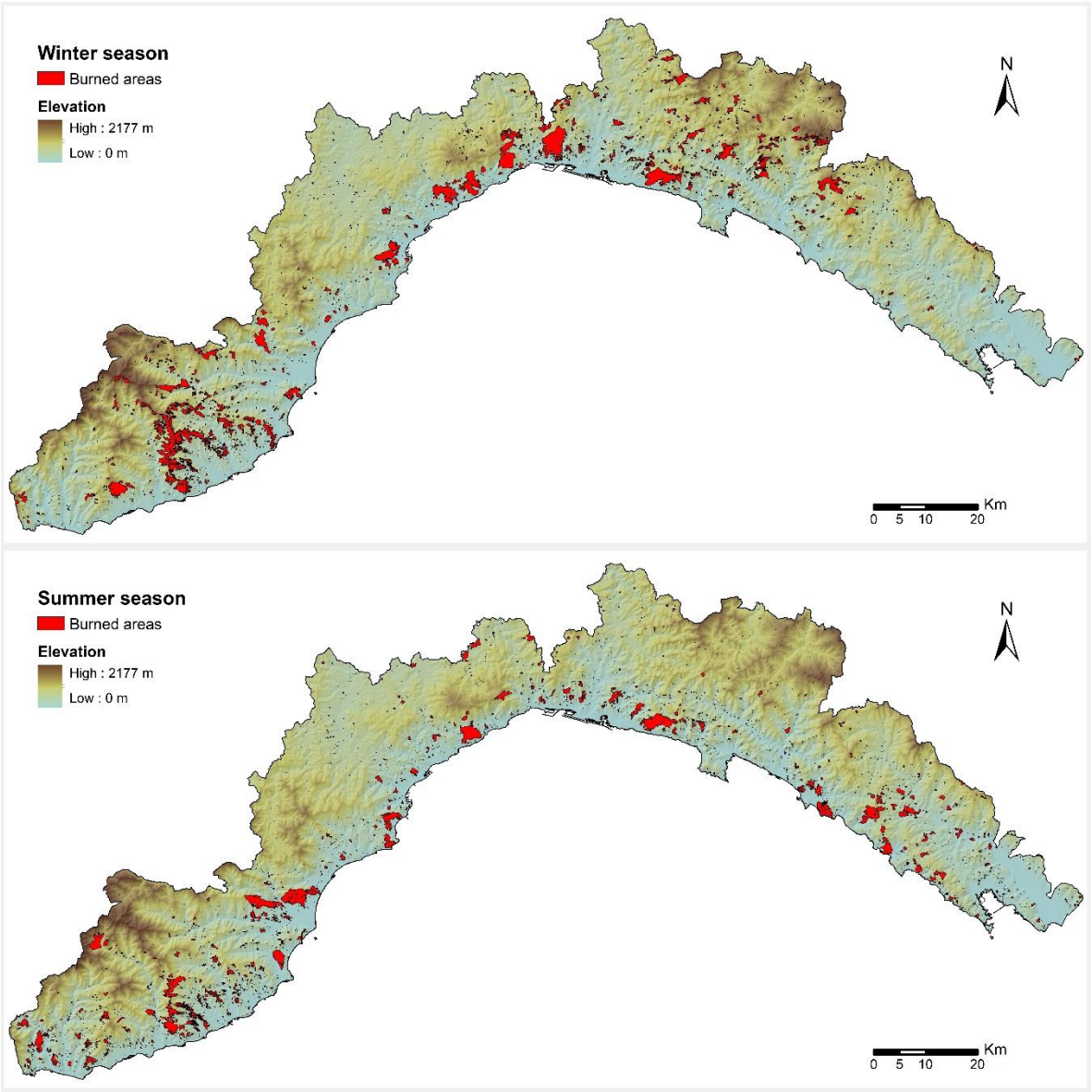

Figure 3. Spatial distribution of burned areas during the winter and summer season in Liguria region (Italy)

The following independent variables provide a detailed knowledge of topography and land cover, allowing to understand the main features involved in wildfire occurrences and their behavior: DEM (altitude) and derivatives (slope, northness and eastness), distance to an anthropogenic features (urban area, road, pathways, crops), protected area, vegetation type and neighboring vegetation. In more details, northness and eastness, corresponding respectively to the cosine and to the minus sine of aspect angle, were considered instead of the pure aspect angle (i.e. the terrain orientation) to avoid the use of a circular variable. The distances-values were evaluated computing, for each pixel, the Euclidean distance to the closest considered element. Protected areas were introduced as a binary variable, computing for each pixel if it falls inside or outside to these delimited areas. The vegetation type was obtained joining the land use map (scale 1:10 000) - allowing to select the non-vegetated

areas - and all the vegetation types out of the forest species (i.e. agricultural areas and grasslands), with the regional map of forest types (scale 1:25 000). This map includes more than hundred classes, which were aggregated for the purposes of the present study into 37 classes, based on the flammability of each vegetation type. The non-vegetated areas (e.g. urban area, industrial and commercial units, road and rail network, port and airport area, bare rocks, water bodies), were unified under the label "non-flammable area" and kept out in the final step, consisting in the elaboration of the wildfires susceptibility area. All these digital layers (Table 1) were provided by the authority of Liguria Region and available on the official geo-portal (https://geoportal.regione.liguria.it/). The ensemble of the spatial layers was pre-processed and resampled to match with the same spatial resolution of 100 meters.

A set of new variables was computed for each pixel by considering the percentage of each land cover type, including both the vegetated (37 classes) and the non-flammable areas (1 class), within a 300 by 300 meters neighborhood distance. This resulted into 38 additional variables to the basic model, which include only 10 variables (Table 1). To prove if the neighboring vegetation type allows to rise the model accuracy, and consequently if this factor need to be included in this kind of study, two models were tested: the first which does not consider the neighboring vegetation (hereafter defined "standard model") and the second accounting for this factor (hereafter defined "neighboring vegetation model").

Table 1 List of predisposing factors and their characteristics

Independent variables	Acquisition scale	Variable type	Range	# of variables
DEM	1 : 5 000	Numerical (meters)	0-2132	1
Slope	-	Numerical (degree)	0-60	1
Northness & Eastness	-	Numerical	[-1,+1]	2
Distance to anthropogenic features	1 : 10 000	Numerical (meters)	0-9000	4
Protected area	1 : 25 000	Binary	0 or 1	1
Vegetation type	1 : 25 000	Categorical	37 classes	1
Neighboring vegetation	-	Numerical (percentage)	[0,100]	38

3.2 Methods: machine learning approach

ML is based on algorithms capable of learning from and make predictions on data, through the modeling of the hidden relationships between a set of input and output variables, representing respectively the predisposing factors (independent variables) and the occurrences of the phenomenon (dependent variable). A training procedure is carried out to calibrate the parameters of the model: the optimal parameters are the ones that minimize the error on the validation dataset, as this is a sign of overfitting to the training dataset. Lastly, to provide an unbiased evaluation of the final model and to assess its performance (i.e. generalization) this one is predicted over unused observations, defined as the testing dataset. Model generalization refers to the ability of the model to perform good prediction on new/previously unseen data, drawn from the same distribution as the ones used to create the model. This concept is closely related with the overfitting: this happen when the model performs well on the training data but is unable to predict new data. Thus, a model that perfectly fit the training data normally displays a large generalization error.

A binary classification problem, such as the prediction of burned and unburned areas in the present case study, can be solved counting how many times each observation is

classified as positive or negative and normalizing the result over the total number of predictions. This provides probabilistic outputs, which can be used to elaborate susceptibility maps identifying areas affected by fires over a rank from low to high. The proposed approach involves the generation of pseudo-absences, to ascertain unburned area. Indeed, to assure a good generalization of the model, avoiding the overestimation of the low classes, pseudo-absences need to be generated in all cases where they are not explicitly expressed (e.g. wildfires location is known, normally as mapped burned areas, but the not-burned areas have to be defined). We solved this problem generating randomly a number of absences equal to the number of presences (i.e. equals to the number of pixels burning).

3.2.1 Random Forest

Analyses were performed using Random Forest [46], an ensemble ML algorithm based on decision trees. The most important hyperparameters that need to be specified are the number of decision trees (*ntree*) and the number of variables randomly sampled as candidates at each split (*mtry*). As general rule, *ntree* should be large enough to ensure that every observation (i.e. input rows) gets predicted at least a few times; the standard value for *mtry* in classification problems is equal to the square root of the number of predictors (i.e. the independent variables), but this value can be optimized. Operationally, the algorithm generates *ntree*-subsets of the training dataset by bootstrapping (i.e. random sampling with replacement), each subset counting about two-third of the observations. Then a decision tree is generated for each subset considering a reduced number of variables (*mtry*) randomly selected: at each node, the Gini impurity is computed and the variable that minimize this value is selected as the best one for the split. Gini impurity measures how often a randomly chosen observation from the training dataset is incorrectly labeled if it is labeled randomly, according to the distribution of the labels in the subset. This process is iterated up to the maximum level, or it can stop when each node contains less than a fixed number of data points. For a classification problem, the prediction of new data is finally computed taking the maximum voting, which can be converted into a probabilistic output. The model's hyperparameters where optimized by evaluating the prediction-error on those observations that were not used in the training subsets (called "out-of-bag"). In this study, hyperparameters were set to 750 for *mtree* and to the round up square root of the number of predictor factors for *mtry*, both optimized by applying a trial and error process.

3.2.2. Model validation

In machine learning, the dataset is usually split into training, validation, and testing, typically including respectively the 60% to 80%, 10% to 20% and 5% to 20% observations of the original. The training dataset is needed to train the model used to get predictions on new data on the validation dataset. The ultimate purpose of the validation dataset is to provide an unbiased evaluation of the model's fitness. Indeed, a good model is the one which gives accurate predictions on new data and avoids overfitting and underfitting. Finally, the test dataset contains data that has never been used in the training step and helps with the final model evaluation or to compare different models.

When dealing with a spatial environmental phenomenon, if the validation dataset is selected randomly, observations can be located close to the training ones, leading to an over-estimation of the predictive performance of the model. This circumstance is known as "spatial autocorrelation", meaning that observations close to each other hold similar characteristics. To overcome this issue, training and validation data have to be selected far

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2020 doi:10.20944/preprints202001.0385.v1

Peer-reviewed version available at Geosciences 2020, 10, 105; doi:10.3390/geosciences10030105

enough apart in the geographic space. In the present study we adopted the spatial k-fold cross validation and then we evaluated the performance of different models over an independent testing dataset. Methodologically, the k-cross validation consists in splitting the dataset into k groups, holding out a set at a time, training the model on the remaining k-1 sets, and finally testing the model on the hold out set. The process is repeated for each subset and the evaluation scores resulting from each model are finally averaged. In the same way, the final prediction value is computed as the arithmetical mean among all the predictions estimated from each folding. We evaluated three models, considering one, five, and nine-folds. One-fold corresponds to the random selection of the validation dataset including, in our case, 20% of the observations. In the case on five and nine folds, data were selected dividing the space into spatial block of 15 by 15 kilometers: this resulted in nearby 50 boxes covering the entire study area, which gave rise for each fold to include about 9 and 6 blocks respectively, distributed randomly (Figure 4).

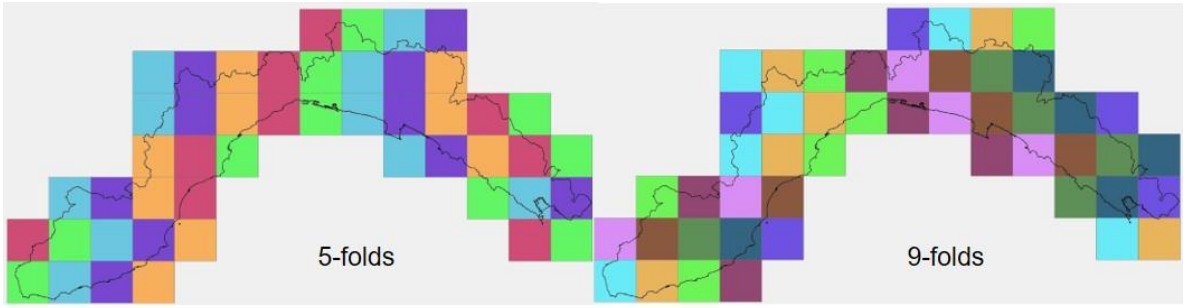

Figure 4: spatial arrangement of the blocks for the 5- and 9-folds adopted in the work (each color corresponds to a single fold).

The testing dataset was defined by splitting the original dataset by years: mapped burned areas observed in the period 1997-2011 were used in the training and validation step, while the last six years (2012-2017) were hold out for testing the predictive performance of the model. This was assessed by computing the fraction of the area with a probabilistic predicted value of burning within certain threshold, falling inside the testing burned areas. We expect that low classes hold higher values for the total area, but lower values for the fraction of this area covering a burning area on the testing dataset. Vice-versa, only a small area will be allocated into the high classes, but this will mainly belong to the burning testing area. To define the classes, we chose the following percentile rank range: 25%, 50%, 75%, 90%, and 95%. These limits correspond to different values of the probabilistic output (*Prob_value*) resulting from each model (Table 2), allowing to a flexible interpretation of the results. Moreover, the root mean square error (RMS) was computed (Table 3) based on the difference between the predicted and the observed values, where observations can only assume the value "one" (if the area burned) or "zero" (if the area did not burned), while the prediction results in a probabilistic value expressed as floating number within this range.

3. Results

The main results obtained by our approach are the following: i) comparison between the "standard model" and the "neighboring vegetation model"; ii) comparison between the random selection of testing dataset (in space) *versus* five- and nine-folds spatial cross validation; iv) prediction values as main output of random forest, allowing to elaborate susceptibility maps of wildfires for the winter and the summer season.

3.1 Models comparison

Values indicated in Table 2 allow to compare and evaluate the prediction performance of different models: the standard *vs* the neighboring vegetation model and the use of one- *vs* five- *vs* nine-folds spatial cross validation for the winter (values above) and the summer season (values below). In addition to the percentile ranking ranges (*Classes*), the corresponding probabilistic output value (*Prob_value*) is specified. The field "*Total area*" indicates the surface predicted by the model to fall within each interval, normalized over the entire area. The field "*Predicted BA*" represents the fraction of the "*Total area*" falling inside the burned area detected in the testing dataset. For example, looking at one-fold standard model, the first class (25th percentile) indicates that only 5.07% of the 25% of the area with the lowest probabilistic values, corresponding to < 0.13, falls inside the burned area in the testing dataset. On the opposite extreme, the 95th percentile indicates that about 47% of the 5% of the area with the highest probabilistic values of burning, corresponding in this case to > 0.91, falls inside the burned area in the testing dataset. Thus, a way to compare the different models is to evaluate which one predicts the largest area with the highest probability of burning falling inside the burned area in the testing dataset (hereafter defined for brevity "predicted burned area"), and vice-versa for the lower classes. To facilitate the comparison, the last 75th percentile was considered and discussed in the following. Results show that the neighboring vegetation model performs better than the standard model in both the seasons. Indeed, the predicted burned area is higher both for the one- and the five-folds cross validation when the model includes the information concerning the type of vegetation and non-flammable area in the neighboring of each pixel. The increment in this case is of about three-percentage points or more, while, when comparing with the nine-folds cross validation, the predicted burned area increases of only 0.87 % in the winter season and 0.02% in the summer season. So, using nine- instead of five-folds allows to slightly increase the prediction performance of the model, face to the fact that the training algorithm is much more computationally intensive as it has to be rerun from scratch five more times. On the other hand, the prediction performance of using five-folds cross validation compared with the one-fold increases of about four-percentage points in the winter season, both when considering the standard and the neighboring vegetation model, and 2.24% and only 0.07% for the equivalent models in the summer season. Despite this last low value, which will be discussed later, overall it results that using five-folds cross validation give better performances of the model than selecting randomly 20% of the dataset for validation.

The root-mean-square errors computed for all the models (Table 3) confirms these results. As expected, all RMS-errors are lower than 0.5, as on average each model performs good prediction of burned and unburned areas. The neighboring vegetation model performs better than the standard one in both the seasons. Increasing the number of folds in the cross validation allows to increase the model's performance in all the cases, except that in the case of the neighboring vegetation model in summer season which performs the same regardless the number of folds.

All the models perform better in winter than in summer season. Indeed, in winter, the testing burned area values for the last 75th percentile ranges from about 80% (1-fold "standard model") to 87% (5- and 9-folds "neighboring vegetation model"). In summer, these values are quite lower, ranging from about 70% (1-fold "standard model") to 75% (5- and 9- folds "neighboring vegetation model"). These performances are confirmed by the RMS-error, showing lower values in summer than in winter for the same model.

To resume, the neighboring vegetation model using five-folds cross validation results to be the one performing better. This last model was than evaluated by computing the fraction of predicted burning area above the 75th percentile falling inside the testing burned areas for each single year and by season. Values are expressed as percentage and as number of pixels (Table 4). This allowed to investigate the influence of the testing dataset on the predictive performances of the model. In the winter season, the model performed well every years, with value ranging from 83.4% in 2013 to 91.7% in 2016. In the summer season, performances are still quite good in the first five testing periods (years 2012 to 2016) while in the last period (year 2017) the assessed value drops to 45.7%. At the same time, we notice that the size of the burned area, which can highly differs in each testing period, does not compromise the predictive performance of the model. More specifically, this result seems to imply that something during the 2017 wildfire summer season in Liguria region does not worked as predicted by the model.

geosciences

MDPI

Table 2. Validation of the models (see Section 3.2.2)

Winter season		1-fold cross validation				5-folds cross validation				9-folds cross validation	
		standard model		neighboring vegetation		standard		neighboring vegetation		neighboring vegetation	
Classes	Total Area (%)	Testing BA (%)	Prob_value	BurnArea (%)	Prob_value	BurnArea (%)	Prob_value	BurnArea (%)	Prob_value	BurnArea (%)	Prob_value
25%	25	5.07	0.13	3.85	0.09	4.42	0.11	3.50	0.07	3.36	0.07
50%	25	4.95	0.25	3.40	0.22	3.44	0.23	3.27	0.18	3.17	0.17
75%	25	10.52	0.48	8.80	0.47	8.90	0.43	6.22	0.39	6.44	0.39
90%	15	17.64	0.78	14.98	0.74	15.77	0.70	13.05	0.68	11.91	0.69
95%	5	14.38	0.91	15.26	0.87	15.67	0.85	17.70	0.83	16.43	0.85
100%	5	47.26	1.00	52.86	1.00	51.78	1.00	56.26	1.00	58.69	1.00
>75%		79.28		83.10		83.22		87.01		87.03	

Summer season		1-fold cross validation				5-folds cross validation				9-folds cross validation	
		standard model		neighboring vegetation		standard		neighboring vegetation		neighboring vegetation	
Classes	Total Area (%)	Testing BA (%)	Prob_value	Testing BA (%)	Prob_value	Testing BA (%)	Prob_value	Testing BA (%)	Prob_value	Testing BA (%)	Prob_value
25%	25	4.71	0.08	1.04	0.04	4.04	0.06	0.80	0.04	0.80	0.04
50%	25	7.52	0.23	4.64	0.17	9.39	0.19	5.08	0.14	5.54	0.14
75%	25	17.94	0.51	18.27	0.41	15.16	0.44	19.77	0.35	18.44	0.35
90%	15	24.45	0.78	26.19	0.70	23.31	0.69	21.51	0.65	22.11	0.66
95%	5	14.06	0.91	14.60	0.87	14.60	0.83	15.06	0.83	14.73	0.85
100%	5	30.66	1.00	33.43	1.00	33.50	1.00	37.71	1.00	38.31	1.00
>75%		69.17		74.22		71.41		74.28		75.15	

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2020 doi:10.20944/preprints202001.0385.v1

Peer-reviewed version available at Geosciences 2020, 10, 105; doi:10.3390/geosciences10030105

Table 3 Root-mean-square error based on the difference between the predicted and the observed value

Winter season	1-fold	5-folds	9-folds
Standard model	0.407	0.380	-
Neighboring vegetation	0.377	0.354	0.351
Summer season	1-fold	5-folds	9-folds
Standard model	0.437	0.428	-
Neighboring vegetation	0.411	0.411	0.411

Table 4 Model validation evaluated by computing the percentage of the predicted burning area above the 75th percentile falling inside the testing burned areas for each single year (BA>75%). The number of pixel above and below this value for each season are also shown.

	Winter season				Summer season				
	BA > 75 %	BA >75%	BA < 75%	Tot_Winter	BA > 75 %	BA >75%	BA < 75%	Tot_Summer	TOT_Year
Year	*(%)*	*(# pixels)*	*(# pixels)*	*(# pixels)*	*(%)*	*(# pixels)*	*(# pixels)*	*(# pixels)*	*(# pixels)*
2012	84.1	844	159	1003	77.8	337	96	433	1436
2013	83.4	121	24	145	86.9	140	21	161	306
2014	86.0	117	19	136	92.8	103	8	111	247
2015	91.2	465	45	510	84.9	535	95	630	1140
2016	91.7	220	20	240	92.7	936	74	1010	1250
2017	86.4	3144	496	3640	**45.7**	449	534	983	4623
TOT				5674				3328	9002

3.1 Susceptibility mapping

Random forest gives as output a probabilistic value, expressing the probability for each pixel of burning under the assumption of a set of predisposing variables. These values were used to elaborate wildfire susceptibility maps in the summer (Figure 5.a) and winter (Figure 5.b) season. Only the results of the neighboring vegetation model using five-folds cross validation, which is the one performing better, were retained for this purpose. The two seasons display a different behavior in relation with the predicted susceptibility to wildfires.

Higher classes (above the 90th percentile) are closer to the coast in summer and develop along the interior, at higher altitude, in winter. This can be due to the state of the vegetation: in winter, vegetation is more stressed and senescent at higher altitude, due to the lower temperature in altitude. On the contrary, in summer vegetation is dryer and more burnable at lower elevations because of the high temperature and the dry weather. Finally, even though the two models implemented separately for the winter and the summer season used the same independent variables as input, Random Forest succeeded in discriminating among the two pattern in the distribution of the wildfire susceptible areas thanks to the training procedure.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2020 doi:10.20944/preprints202001.0385.v1

Peer-reviewed version available at Geosciences 2020, 10, 105; doi:10.3390/geosciences10030105

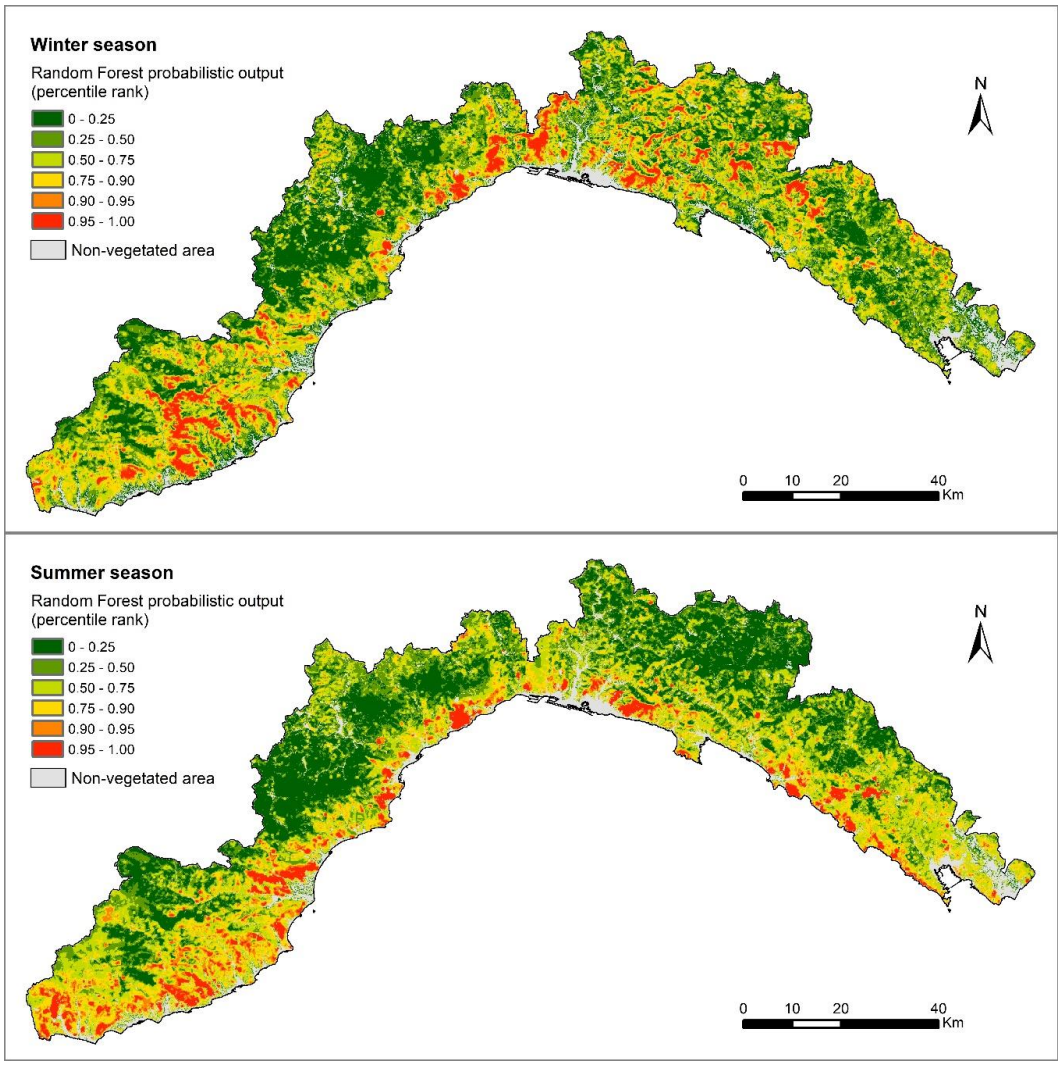

Figure 5. Wildfires susceptibility map in the winter and summer season for Liguria region (Italy)

4. Discussion and conclusions

In the present paper, we introduced an innovative Machine Learning approach, based on Random Forest, allowing to elaborate the wildfire susceptibility map for Liguria region (Italy). Susceptibility was assessed by evaluating the probability for an area to burn in the future taking into account the spatial extension of past-burned areas and the geo-environmental factors that favor their occurrence (i.e. altitude and its derivatives, distance to an anthropogenic feature, protected area, vegetation type). An alternative model, including the neighboring vegetation type at each location, was developed and compared with the standard one. For validation purposes, we adopted the spatial k-fold cross validation (with k= 1, 5 and 9) and then we evaluated the predictive performances of the different models over an independent testing dataset. This last was defined by splitting the original dataset, spanning from 1997 to 2017, by years and holding out the last six years for testing purposes. Finally, we compared the standard *vs* the neighboring vegetation model and the use of one- *vs* five- *vs* nine-folds spatial cross validation, both for the winter and the summer season. It results that: 1) neighboring vegetation model performs better than the standard one in both the seasons; 2) five results to be the optimal number of folds based on model performances; 3) all the models perform better in winter than in summer season.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2020 doi:10.20944/preprints202001.0385.v1

Peer-reviewed version available at Geosciences 2020, 10, 105; doi:10.3390/geosciences10030105

The implemented models globally shows a high capacity to discriminate most of the burned area within the 75th percentile for most of the testing periods (2012 to 2017) and in both the wildfire seasons. The neighboring vegetation model using five-folds cross validation gives the best performances and was retained to elaborate winter and summer wildfires susceptibility maps. The performance of this model in the summer 2017 testing-period manifests the only notable exception to the overall good prediction capability, and it will be object of discussion in the following. Instead of considering this fact a downside of the proposed modeling framework, it indeed shifted the attention towards a specific situation, the fire management in summer 2017. As a matter of fact, it is worth considering that the Italian Forestry Corp, in charge of fire management since 1984, was dismissed at the end of 2016 (DLGS 19/08/2016 n. 177). To counter this, Liguria Region named Italian Fire Fighters (CNVVF) in charge of forest fire management. Before 2017, the role of CNVVF was limited to the management of fire in the Wildland Urban Interface and mainly restricted to the safeguard of civilians (protection of houses and infrastructures). In January 2017 a week of severe fire danger caused the burning of about 3 000 ha, most of them characterized by high or extreme fire risk. However, the summer season of 2017 has been a year peculiar not only in terms of fire management procedures, but also with respect to the meteorological conditions. This season was in fact characterized by a long drought, but with a remarkable average relative humidity higher than 67%. In Figure 6 the distribution of the relative humidity observed by 40 meteorological stations in the period ranging from 01/05/2017 to 31/10/2017 is reported. Only a couple of days were characterized by relative humidity lower than 40%.

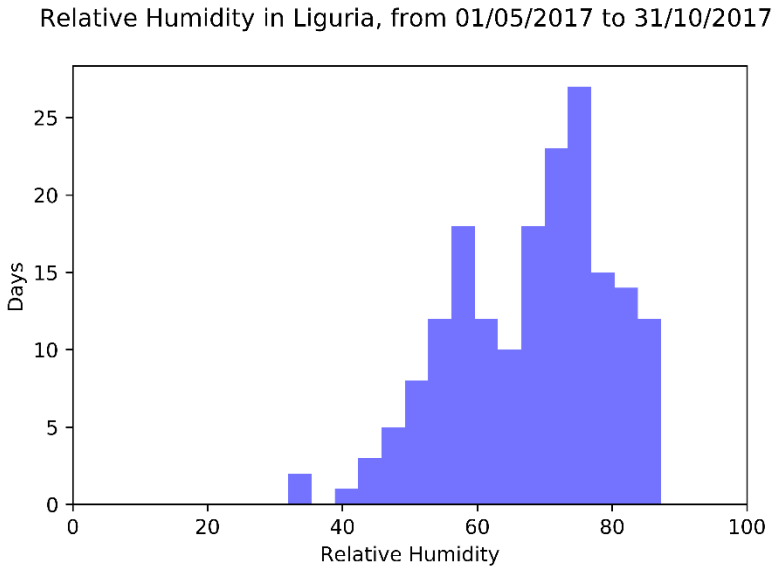

Figure 6. Histogram of daily average relative humidity from May to October observed from 40 meteorological stations scattered all over Liguria

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2020 doi:10.20944/preprints202001.0385.v1

Peer-reviewed version available at Geosciences 2020, 10, 105; doi:10.3390/geosciences10030105

In this frame period, 205 wildfires burned a total area of 927 ha. Only six events burned an area greater than 50 ha each, resulting in 554 ha (corresponding to 60% of the total burned area). These six events were analyzed in detail for a better understanding of the implication that the new management can have on the model performance and prediction capabilities.

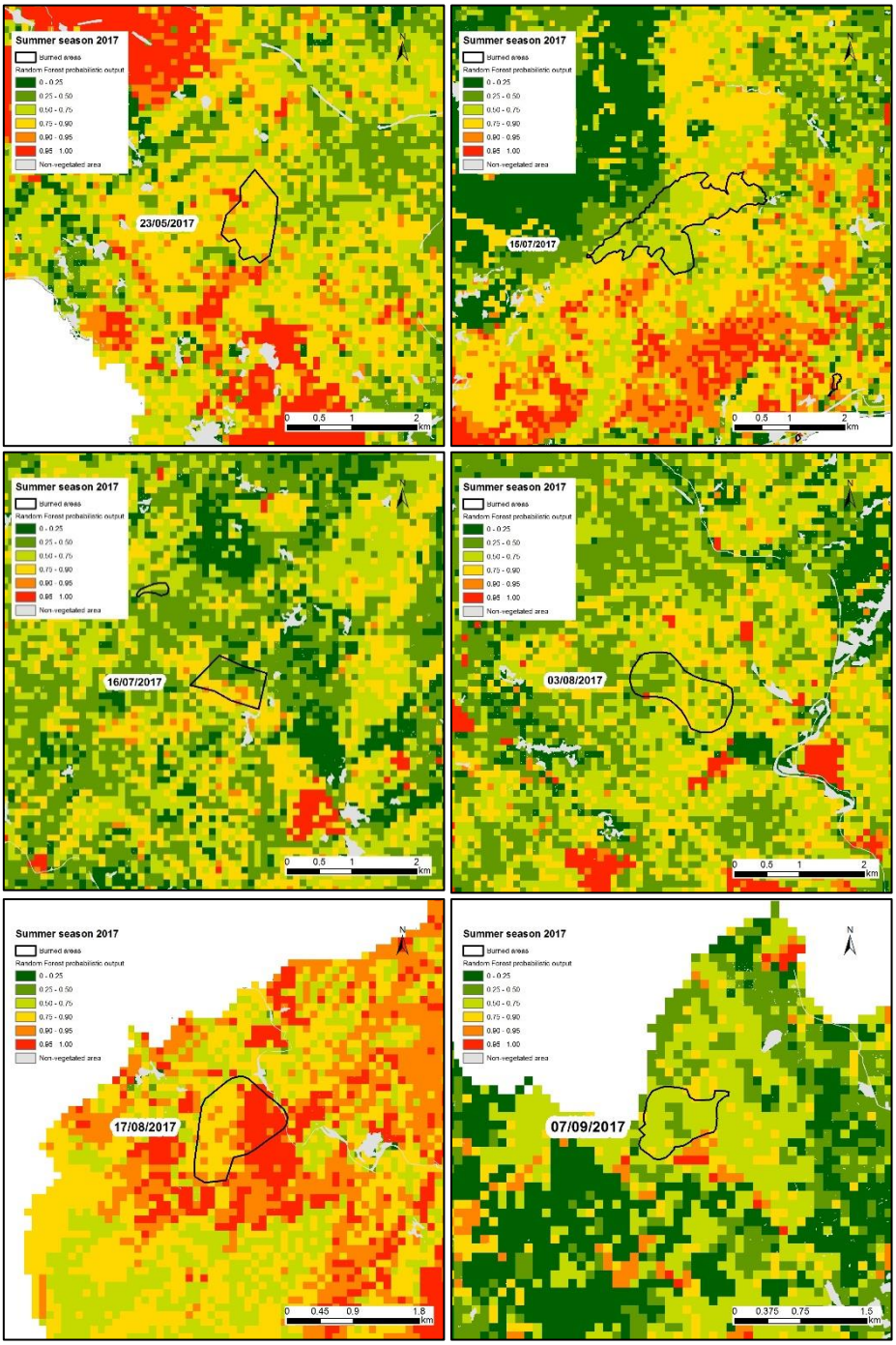

Figure 7 The six large wildfire occurrences of Summer 2017, along with the fire susceptibility map.

As it is evident in Figure 7, most of the burned area within these six large wildfire occurrences are characterized by a middle-to-low level susceptibility, reflecting that Random Forest cannot take into account the management issues of 2017. Despite the rapid capacity of intervention of CNVVF, the fire brigade did not considered the mop up phase, which caused each single fires to be reignited several times, extending the fire propagation

for many days. These considerations put in evidence the capacity of the proposed approach to identify also the efficiency of fire fighting activities, specifically if fire extinction procedures are handled with modalities which differ compared with the ones used in the past, and implicitly valued into the model. Before 2017, the different tactic of fire management, which included the mop up phase, could have resulted in lower burned areas which, by the way, were identified by the models as result from the susceptibility map.

In conclusion, the proposed approach proved to be globally effective to deal with large, high dimensional spatial data, which reflects the great flexibility of machine learning in general. In particular, one advantage of Random Forest is its capability of handling directly categorical variable, such as the land use classes or the vegetation type. Ultimately, the results of the present study highlights the importance of accurately select the predisposing factors and the parameters in the model, and of taking into account possible changes of surrounding conditions which can affect the validity of the model in space and in time.

References

1. Mihalić Arbanas, S. Landslide Hazard, Risk Assessment and Prediction: Landslide Inventories and Susceptibility, Hazard Mapping Methods, Damage Potential—Part 2. In *Advancing Culture of Living with Landslides*; Mikos, M., Tiwari, B., Yin, Y., Sassa, K., Eds.; Springer International Publishing: Cham, 2017; pp. 695–698 ISBN 978-3-319-53497-8.
2. Hervás, J.; Bobrowsky, P. Mapping: Inventories, Susceptibility, Hazard and Risk. In *Landslides – Disaster Risk Reduction*; Sassa, K., Canuti, P., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2009; pp. 321–349 ISBN 978-3-540-69966-8.
3. Van Den Eeckhaut, M.; Hervás, J. State of the art of national landslide databases in Europe and their potential for assessing landslide susceptibility, hazard and risk. *Geomorphology* **2012**, *139–140*, 545–558.
4. Fell, R.; Corominas, J.; Bonnard, C.; Cascini, L.; Leroi, E.; Savage, W.Z. Guidelines for landslide susceptibility, hazard and risk zoning for land-use planning. *Eng. Geol.* **2008**, *102*, 99–111.
5. Finney, M.A. The challenge of quantitative risk analysis for wildland fire. *For. Ecol. Manag.* **2005**, *211*, 97–108.
6. Hardy, C.C. Wildland fire hazard and risk: Problems, definitions, and context. *For. Ecol. Manag.* **2005**, *211*, 73–82.
7. Watts, J.M.; Hall, J.R. Introduction to Fire Risk Analysis. In *SFPE Handbook of Fire Protection Engineering*; Hurley, M.J., Gottuk, D., Hall, J.R., Harada, K., Kuligowski, E., Puchovsky, M., Torero, J., Watts, J.M., Wieczorek, C., Eds.; Springer New York: New York, NY, 2016; pp. 2817–2826 ISBN 978-1-4939-2564-3.
8. Meacham, B.J.; Charters, D.; Johnson, P.; Salisbury, M. Building Fire Risk Analysis. In *SFPE Handbook of Fire Protection Engineering*; Hurley, M.J., Gottuk, D., Hall, J.R., Harada, K., Kuligowski, E., Puchovsky, M., Torero, J., Watts, J.M., Wieczorek, C., Eds.; Springer New York: New York, NY, 2016; pp. 2941–2991 ISBN 978-1-4939-2564-3.
9. Brillinger, D.R.; Preisler, H.K.; John W. Benoit Risk Assessment: A Forest Fire Example. *Lect. Notes-Monogr. Ser.* **2003**, *40*, 177–196.

10. Catry, F.X.; Rego, F.C.; Bação, F.L.; Moreira, F. Modeling and mapping wildfire ignition risk in Portugal. *Int. J. Wildland Fire* **2010**, *18*, 921–931.
11. Ager, A.A.; Finney, M.A.; Kerns, B.K.; Maffei, H. Modeling wildfire risk to northern spotted owl (Strix occidentalis caurina) habitat in Central Oregon, USA. *For. Ecol. Manag.* **2007**, *246*, 45–56.
12. Eugenio, F.C.; dos Santos, A.R.; Fiedler, N.C.; Ribeiro, G.A.; da Silva, A.G.; dos Santos, Á.B.; Paneto, G.G.; Schettino, V.R. Applying GIS to develop a model for forest fire risk: A case study in Espírito Santo, Brazil. *J. Environ. Manage.* **2016**, *173*, 65–71.
13. Teodoro, A.C.; Duarte, L. Forest fire risk maps: a GIS open source application – a case study in Norwest of Portugal. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 699–720.
14. Gai, C.; Weng, W.; Yuan, H. GIS-Based Forest Fire Risk Assessment and Mapping. In Proceedings of the 2011 Fourth International Joint Conference on Computational Sciences and Optimization; IEEE: Kunming and Lijiang City, China, 2011; pp. 1240–1244.
15. Pourghasemi, H.R. GIS-based forest fire susceptibility mapping in Iran: a comparison between evidential belief function and binary logistic regression models. *Scand. J. For. Res.* **2016**, *31*, 80–98.
16. Mohammadi, F.; Bavaghar, M.P.; Shabanian, N. Forest Fire Risk Zone Modeling Using Logistic Regression and GIS: An Iranian Case Study. *Small-Scale For.* **2014**, *13*, 117–125.
17. Vadrevu, K.P.; Eaturu, A.; Badarinath, K.V.S. Fire risk evaluation using multicriteria analysis—a case study. *Environ. Monit. Assess.* **2010**, *166*, 223–239.
18. Kant Sharma, L.; Kanga, S.; Singh Nathawat, M.; Sinha, S.; Chandra Pandey, P. Fuzzy AHP for forest fire risk modeling. *Disaster Prev. Manag. Int. J.* **2012**, *21*, 160–171.
19. Carmel, Y.; Paz, S.; Jahashan, F.; Shoshany, M. Assessing fire risk using Monte Carlo simulations of fire spread. *For. Ecol. Manag.* **2009**, *257*, 370–377.
20. Catry, F.X.; Rego, F.C.; Bação, F.L.; Moreira, F. Modeling and mapping wildfire ignition risk in Portugal. *Int. J. Wildland Fire* **2009**, *18*, 921.
21. Chuvieco, E.; Aguado, I.; Yebra, M.; Nieto, H.; Salas, J.; Martín, M.P.; Vilar, L.; Martínez, J.; Martín, S.; Ibarra, P.; et al. Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. *Ecol. Model.* **2010**, *221*, 46–58.
22. Chuvieco, E.; Salas, J. Mapping the spatial distribution of forest fire danger using GIS. *Int. J. Geogr. Inf. Syst.* **1996**, *10*, 333–345.
23. Pourtaghi, Z.S.; Pourghasemi, H.R.; Aretano, R.; Semeraro, T. Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques. *Ecol. Indic.* **2016**, *64*, 72–84.
24. Leuenberger, M.; Parente, J.; Tonini, M.; Pereira, M.G.; Kanevski, M. Wildfire susceptibility mapping: Deterministic vs. stochastic approaches. *Environ. Model. Softw.* **2018**, *101*, 194–203.
25. Arpaci, A.; Malowerschnig, B.; Sass, O.; Vacik, H. Using multi variate data mining techniques for estimating fire susceptibility of Tyrolean forests. *Appl. Geogr.* **2014**, *53*, 258–270.
26. Oliveira, S.; Oehler, F.; San-Miguel-Ayanz, J.; Camia, A.; Pereira, J.M.C. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *For. Ecol. Manag.* **2012**, *275*, 117–129.
27. Ngoc Thach, N.; Bao-Toan Ngo, D.; Xuan-Canh, P.; Hong-Thi, N.; Hang Thi, B.; Nhat-Duc, H.; Dieu, T.B. Spatial pattern assessment of tropical forest fire danger at Thuan Chau area

(Vietnam) using GIS-based advanced machine learning algorithms: A comparative study. *Ecol. Inform.* **2018**, *46*, 74–85.

28. Satir, O.; Berberoglu, S.; Donmez, C. Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem. *Geomat. Nat. Hazards Risk* **2016**, *7*, 1645–1658.
29. Tehrany, M.S.; Jones, S.; Shabani, F.; Martínez-Álvarez, F.; Tien Bui, D. A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using LogitBoost machine learning classifier and multi-source geospatial data. *Theor. Appl. Climatol.* **2019**, *137*, 637–653.
30. Rodrigues, M.; de la Riva, J. An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environ. Model. Softw.* **2014**, *57*, 192–201.
31. Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Aryal, J. Forest Fire Susceptibility and Risk Mapping Using Social/Infrastructural Vulnerability and Environmental Variables. *Fire* **2019**, *2*, 50.
32. Ganteaume, A.; Camia, A.; Jappiot, M.; San-Miguel-Ayanz, J.; Long-Fournel, M.; Lampin, C. A Review of the Main Driving Factors of Forest Fire Ignition Over Europe. *Environ. Manage.* **2013**, *51*, 651–662.
33. Tonini, M.; Parente, J.; Pereira, M.G. Global assessment of rural–urban interface in Portugal related to land cover changes. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 1647–1664.
34. Conedera, M.; Tonini, M.; Oleggini, L.; Vega Orozco, C.; Leuenberger, M.; Pezzatti, G.B. Geospatial approach for defining the Wildland-Urban Interface in the Alpine environment. *Comput. Environ. Urban Syst.* **2015**, *52*, 10–20.
35. Zumbrunnen, T.; Menéndez, P.; Bugmann, H.; Conedera, M.; Gimmi, U.; Bürgi, M. Human impacts on fire occurrence: a case study of hundred years of forest fires in a dry alpine valley in Switzerland. *Reg. Environ. Change* **2012**, *12*, 935–949.
36. Badia, A.; Serra, P.; Modugno, S. Identifying dynamics of fire ignition probabilities in two representative Mediterranean wildland-urban interface areas. *Appl. Geogr.* **2011**, *31*, 930–940.
37. Romero-Calcerrada, R.; Barrio-Parra, F.; Millington, J.D.A.; Novillo, C.J. Spatial modelling of socioeconomic data to understand patterns of human-caused wildfire ignition risk in the SW of Madrid (central Spain). *Ecol. Model.* **2010**, *221*, 34–45.
38. Castillo Soto, M.E. The identification and assessment of areas at risk of forest fire using fuzzy methodology. *Appl. Geogr.* **2012**, *35*, 199–207.
39. Vilar, L.; Woolford, Douglas.G.; Martell, D.L.; Martín, M.P. A model for predicting human-caused wildfire occurrence in the region of Madrid, Spain. *Int. J. Wildland Fire* **2010**, *19*, 325.
40. Mermoz, M.; Kitzberger, T.; Veblen, T.T. LANDSCAPE INFLUENCES ON OCCURRENCE AND SPREAD OF WILDFIRES IN PATAGONIAN FORESTS AND SHRUBLANDS. *Ecology* **2005**, *86*, 2705–2715.
41. Carmo, M.; Moreira, F.; Casimiro, P.; Vaz, P. Land use and topography influences on wildfire occurrence in northern Portugal. *Landsc. Urban Plan.* **2011**, *100*, 169–176.
42. Harris, L.; Taylor, A.H. Previous burns and topography limit and reinforce fire severity in a large wildfire. *Ecosphere* **2017**, *8*, e02019.
43. Moreira, F.; Vaz, P.; Catry, F.; Silva, J.S. Regional variations in wildfire susceptibility of land-cover types in Portugal: implications for landscape management to minimize fire hazard. *Int. J. Wildland Fire* **2009**, *18*, 563.

44. Kanevski, M.; Pozdnoukhov, A.; Timonin, V. *Machine learning for spatial environmental data: theory, applications and software*; Environmental sciences, environmental engineering; Epfel: Lausanne, 2009; ISBN 978-0-8493-8237-6.
45. *Machine Learning Techniques Applied to Geoscience Information System and Remote Sensing*; MDPI, 2019; ISBN 978-3-03921-216-3.
46. Breiman, L. Random Forest. *Mach. Learn.* **2001**, *45*, 5–32.