An insight into machine-learning algorithms to model human-caused wildfire occurrence

Marcos Rodrigues*, Juan de la Riva
GEOFOREST Group, IUCA, Department of Geography and Land Management, University of Zaragoza, Pedro Cerbuna 12, 50009 Zaragoza, Spain

A R T I C L E   I N F O
Article history:
Received 26 January 2014
Received in revised form 24 February 2014
Accepted 4 March 2014
Available online 2 April 2014

Keywords:
Machine learning
Model
Wildfire
Random Forest
Boosted Regression Tree
Support Vector Machine

A B S T R A C T
This paper provides insight into the use of Machine Learning (ML) models for the assessment of human-caused wildfire occurrence. It proposes the use of ML within the context of fire risk prediction, and more specifically, in the evaluation of human-induced wildfires in Spain. In this context, three ML algorithms—Random Forest (RF), Boosting Regression Trees (BRT), and Support Vector Machines (SVM)—are implemented and compared with traditional methods like Logistic Regression (LR). Results suggest that the use of any of these ML algorithms leads to an improvement in the accuracy—in terms of the AUC (area under the curve)—of the model when compared to LR outputs. According to the AUC values, RF and BRT seem to be the most adequate methods, reaching AUC values of 0.746 and 0.730 respectively. On the other hand, despite the fact that the SVM yields an AUC value higher than that from LR, the authors consider it inadequate for classifying wildfire occurrences because its calibration is extremely time-consuming.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Concern about wildfires and their impacts is an increasing phenomenon. In Mediterranean Europe, increasing trends in the number of fires have been detected in some countries such as Portugal and Spain (San-Miguel Ayanz et al., 2012). This increase in wildfire frequency, with its associated risks to the environment and society (Moreno et al., 2011), calls for better understanding of the processes that control wildfire activity (Bar Massada et al., 2012). Therefore, a better comprehension of the driving forces of fire ignition and of predicting where fires are likely to start are core elements in designing strategies to mitigate wildfire initiation and to identify areas at risk (Finney, 2005). Consequently, efforts to achieve a better understanding of wildfires have been increasing in recent years, and several methods for wildfire risk assessment have been developed using different methodological schemes, variables, and scales (Martínez-Vega et al., 2012). Without being exhaustive, some of the more recent efforts have included those by Amatulli et al. (2006), Bogdos and Manolakos (2013), Chuvieco et al. (2010, 2012), Cooke et al. (2007), Cruz et al. (2013), Duff et al. (2013), Loboda (2009), Martínez et al. (2009), Martínez-Fernández and Koutsias 2011, Padilla and Vega-García (2011), Rodrigues et al. 2014, Romero-Calcerrada et al. (2010), Serra et al. (2013), and Sullivan and Matthews (2013). Within the same context, similar efforts have been invested in modeling fire occurrence (see (Plucinksi, 2011) for an exhaustive review), one of the main input parameters when modeling wildfire risk (Chuvieco et al., 2012).

On the other hand, human beings have a great impact on fire regimes because they alter ignition frequency and fuel fragmentation and suppress fires (Guyette et al., 2002). The dynamics of fire regimes in southern Europe are related mainly to human factors, which are the cause of more than 95% of fires in this region (San-Miguel Ayanz and Camiá, 2009). The analysis of human factors in forest fires is widely recognized as of critical importance for fire danger estimation (Kalabokidis et al., 2002; Martínez et al., 2004a), but the literature on this topic is scarce and mainly site-specific (Krawchuk et al., 2009; Le Page et al., 2010; Martínez et al., 2009). However, in recent years, the role of human factors in fire behavior modeling has been increasing, and several models now include an anthropogenic component in their assessments (Chuvieco et al., 2010, 2012; Loepef et al., 2011). Machine learning (ML) models have shown their predictive accuracy in data mining and other disciplines (Casalegno et al., 2011; Cutler et al., 2007; Díaz-Uriarte and de Andres, 2006; Drake et al., 2006; Li et al., 2011; Marmion et al., 2009; Pino-Mejías et al., 2010; Shan et al., 2006). Previous studies have also proposed ML algorithms to model the spatial distribution of wildfire occurrence or ignition. These algorithms include Regression Trees (RT; Amatulli et al., 2006),
Artificial Neural Networks (Vasconcelos et al., 2001; Vega-Garcia et al., 1996), and more recently, Random Forest (RF; Bar Massada et al., 2012). However, these methods have not been widely used to model human-caused wildfire occurrence at a regional scale or for large occurrence datasets; this is therefore the main goal of this work. This topic will be addressed in greater depth by exploring other stochastic and deterministic ML algorithms and their application to the Spanish territory. Specifically, the performance of Random Forest (RF), Boosted Regression Trees (BRT), and Support Vector Machines (SVM) has been explored, and their outcomes have been compared with those from binary logistic regression (LR), a commonly used technique for probabilistic explanation of human-caused occurrences (Chuvieco et al., 2010; Martínez et al., 2004a, 2009; Vasconcelos et al., 2001; Vega-Garcia et al., 1995).

The main drawback of modeling with only one RT is that this approach is not entirely robust because each division can involve a set of variables with similar discriminatory power. Therefore, small changes in the data can generate very different models. To avoid such problems, researchers have recently shown interest in ensemble learning methods. These methods generate many classifiers and enable grouping of the results in a final classification. Two examples of well-known ensemble methods are boosting and bagging (Breiman et al., 2000; Hastie et al., 2004; Sierra, 2006). Boosting is a method for improving model accuracy, based on the idea that it is easier to find and average many rough empirical rules than to find a single, highly accurate prediction rule (Schapire, 2003). Related techniques—including bagging, stacking and model averaging—also build, then merge results from multiple models, but boosting is unique because it is sequential (Elith et al., 2008). On the other hand, bagging is a technique designed to create training data sets resampled randomly with replacement of original data, i.e., without removing the selected data set before selecting the next subset. Thus, data may be used more than once to train individual classifiers. This property makes bagging methods less sensitive to slight variations in the input data (training changes, outliers, noise ...) and at the same time increases the accuracy of classifications (Breiman, 2001).

RF is an ensemble classifier using decision trees as base classifiers. The RF algorithm was proposed by Breiman (2001) and adds an element of randomness to bagging, increasing the diversity of decision trees by growing them from different subsets. Besides generating each decision tree using a subset of different training elements in each iteration, RF changes the way that the decision tree is generated by the classification. In the creation of decision trees in the CART algorithm, each node is split using the best threshold for all variables introduced, while in RF, the nodes are relatively simple tree models to optimize predictive performance (Elith et al., 2008; Leathwick et al., 2006, 2008). The boosting approach used in BRT places its origins within ML (Schapire, 2002), but subsequent developments in the statistical community have interpreted it as an advanced form of regression (Friedman et al., 2000).

On the other hand, the SVM algorithm is based on making highly reliable predictions, even at the risk of making some mistakes. To this end, SVM tries to find the optimal hyperplane of separation between the classes, i.e., the plane in which the separability between classes is a maximum. The examples located on this hyperplane are called support vectors. These examples are the most difficult to classify since they have lower separability. In the simplest case, two classes in a two-dimensional space in which the data are linearly separable, the optimal hyperplane would be defined by a straight line. For a more detailed description of SVM operation, see Vapnik (1995, 1998).

In this work, several models using these three algorithms were investigated. Their results were compared to LR outcomes, also calculated in this paper, to test their performance. All models were fitted using the same explanatory and dependent variables. The explanatory variables (later introduced and described) were selected on the basis of the authors’ previous experience with models at regional and national scales (Amatulli et al., 2007; Chuvieco et al., 2010, 2012; Martínez et al., 2004a; Vilar et al., 2008), while the dependent binary variable (high/low fire occurrence) was constructed from wildfire observations from 1988 to 2007 in Spain. Spain could be considered as a key area for wildfire modeling since it is, by far, the most fire-affected territory within the European Union (Rodríguez et al., 2013; San-Miguel Ayanz et al., 2012), a fact that justifies its analysis. Results suggest that ML models improve LR both in terms of prediction accuracy and of the spatial pattern of the probability of occurrence. However, SVM requires a more in-depth exploration and optimization to be properly calibrated for wildfire occurrence prediction.

2. Materials and methods

2.1. Study area

The study area covered the whole of peninsular Spain, excluding the Balearic and Canary Islands, as well as the autonomous cities of Ceuta and Melilla, because some parameters needed to develop the methodology were not available in those areas. Therefore, the total area of the study region was approximately 498,000 km². Moreover, the study region was restricted to forested areas. Consequently, urban areas, agricultural areas, and inland water zones were excluded from the assessment, and no data for them are detailed or displayed in the maps. Spain is a territory of wide contrasts which presents a great variety of climatic, topographic, environmental, and other biophysical conditions. These dissimilar conditions also appear when talking about socioeconomic conditions in terms of population systems and productive sectors, geographical structure. Therefore, the complexity of the socioeconomic conditions play a determining role and are especially important when modeling human factors because this complexity is transferred to the relationships between socioeconomic variables and to natural phenomenon like wildfires, making their assessment more difficult.

2.2. Dependent variable

The dependent variable—high/low wildfire occurrence—was built from the Spanish EGIF (General Statistics of Wildfires) database from 1988 to 2007. The EGIF database is one of the oldest “complete” wildfire databases in Europe, beginning in 1968 (Velez, 2001). It has been compiled by the Ministry of Environment, Rural, and Marine Affairs (MARM) using forest fire reports from the various autonomous regions (Moreno et al., 2011). Among other useful information relating fire events, these reports include data regarding the starting location point of each fire. This position is recorded on the basis of a reference 10 × 10 km ICONA grid (used by the firefighting services for approximate location of fire events) and the municipality origin of the ignition. The spatial distribution of fire occurrence (308,893 fires in the period from 1988 to 2007, as shown in Fig. 1) was developed through a combination of the 10 × 10 km grid, a digital map of Spanish municipal boundaries, and the boundaries of the forest area. More specifically the ignition location procedure is based in the method developed by de la Riva et al. (2004). This method is widely recognized and has been used in many wildfire assessment research works in the Spanish territory such as Amatulli et al. (2007), Chuvieco et al. (2010, 2012) and most recently in Rodrigues et al. (2014). The method proposes a multi-step procedure which successively refines and decreases the potential location area of the ignition points by ruling out areas where the fire could not have occurred. Firstly it starts in the 10 × 10 grid with a potential location area of 100 km². Then this area is decreased by intersecting with the boundaries of the municipality origin of the fire. Finally, the location area is restricted to the forest perimeter (MARM, 1997) — since the ignition
location of every wildfire is expected to be in the forest area – to determine the final potential location area. This process leads to a significantly smaller area where the ignition points are then randomly distributed. This allowed us to calculate fire density maps with a spatial resolution of $1 \times 1$ km by overlapping the final ignition points cloud and a $1 \times 1$ km UTM grid (which perfectly fits the $10 \times 10$ grid). Fig. 2 illustrates this procedure. Recent studies have commented that predictions from fire simulations based on random ignitions may produce unrealistic results because the spatial distribution of ignition locations, whether human-caused or natural, is non-random (Bar Massada et al., 2011). However, the lack of explicit location data for wildfire events, especially in the first years of the EGIF dataset, made it impossible to generate a realistic set of locations. On the other hand, in many cases where coordinates have been assigned, the final location seems to be unreliable because it

![Fig. 1. Spatial distribution of ignition points (left) and the dependent variable (right).](image)

![Fig. 2. Procedure for ignition points location. Potential location area is gray-colored. a) $10 \times 10$ Km ICONA grid; b) municipality intersection; c) forest area intersection; d) random point location and intersection with $1 \times 1$ Km grid.](image)
corresponds with unexpected sites such as the corner of the UTM grid or outside the forest area, which are more likely to be false.

The final dependent variable was created on a conceptual framework which assumed that there were no true cases of fire absence. In ignition data, most or all of the fire occurrences are accounted for, which may make it seem as if all other locations in the landscape have no fires. In this context, most previous attempts at fire occurrence modeling had used background subsets of “no occurrence” during the analyzed time span, considering them to be true cases of fire absence (e.g., Chuvieco et al., 2010; Padilla and Vega-García, 2011). However, the fact that these areas did not experience an ignition event during the temporal span of the data set does not mean that they could not feasibly support an ignition event in the future, or that they experience an ignition event during the temporal span of the data set does not mean that they could not feasibly support an ignition event in the future, or that they never ignited in the past (Bar Massada et al., 2012). In line with this reasoning, the dependent variable was developed by classifying the occurrence values into two categories: high occurrence (presence: 27,956 points) in locations with two or more fires, and low occurrence (pseudo-absence or background: 28,188 points) in locations with only one fire (Fig. 1). The authors thought that the consideration of low-occurrence locations as pseudo-absences was more realistic than the creation of random background subsets. The fact that these areas has experienced only one fire event in a long time span (20 years), means that their characteristics are strongly related with low fire frequencies.

2.3. Explanatory variables

The explanatory variables were selected and spatialized on the basis of the authors’ experience with models at regional and national scales (Amatulli et al., 2007; Chuvieco et al., 2010, 2012; Martínez et al., 2004a; Vilar et al., 2008). According to this, the explanatory variables were classified in relation to the typology of the affecting factor (Leone et al., 2003; Martínez et al., 2004b), as follows:

1. Factors related to socio-economic transformations.
   1.1. Abandonment of traditional activities in wildland and rural areas, especially in privately owned forests with no prospect of economic profit. Little or no interest in forest conservation.
   - Forestry and public utilities. Delimitation of the area occupied by forestry areas included in the public utility catalog.
   - Human presence, population increase, and urban growth. More pressure on wildlands.
   - Wildland- Urban Interface (WUI). Area occupied by the 200-m buffer from the line of contact to the forest area. Constructed from the Spanish Forestry Map 1:200000 (MFE200).

2. Factors related to traditional economic activities in rural areas.
   - Agriculture. Fire use to eliminate harvesting wastes and to clean cropland borders.
   - Wildland-Agricultural Interface (WAI). Area occupied by the 200-m buffer from the line of contact to the forest area. Constructed from the Spanish Forestry Map 1:200000 (MFE200).

3. Factors which could cause fire mainly by accident or negligence.
   - Electric lines. Possible cause of ignition by accident.
   - Power lines. Area occupied by the 50-m buffer around the high-, medium-, and low-voltage transport network. Obtained from BCN200.
   - Engines and machines working in or close to forest areas. Possible cause of ignition by accident or negligence.
   - Density of agricultural machinery. Obtained at the municipal level from the Agricultural Census 1999 of the Spanish Statistics Institute (INE).
   - Presence of roads, railways, and tracks and their accessibility. More human pressure on wildland.
   - Railways. Area occupied by the 200-m buffer around the railroad network (excluding the high-speed network). Obtained from a digital cartographic database (BCN200).

4. Factors which could hamper fires.
   - Protected areas. Increasing concern about forest protection.
   - Protected areas. Delimitation of the area occupied by natural protected areas and the Natura 2000 network.

All the predictive variables, as well as the dependent variable, were distributed in space at a resolution of 1 x 1 km. To ensure consistency of results, a collinearity analysis of the explanatory variables was carried out. No collinear variables were found.

2.4. Model calibration and software

The models were fitted using the R statistical software (packages randomForest, gbm, and kernlab). R is an open-source statistical programming language developed as a large collaborative project by statisticians from different countries and disciplines (R Development Core Team, 2008). The total sample obtained from the spatial distribution of the fire reports compiled in the EGIF database (93,573 locations with fire) was separated into a training sample (60% of the population) and a testing sample (40% of the population). Consequently, the calibration sample was made up of 56,144 fire records and the validation sample of 37,429. The explanatory variables were considered or not considered, depending on the model, according to the value of the area under the receiver operating characteristic curve (AUC) of the trained model (see Section 2.5); variables were introduced when they improved the AUC value and dropped when the AUC remained at the same or a lesser value.

2.4.1. RF

RF can be parameterized according to the number of trees averaged in the ensemble forest (ntrees), the number of predictor variables randomly selected at each iteration (mtry), and the minimum number of observations at end nodes (nodesize), which can decrease the length of nodes in tree branches and simplify trees. All combinations of five ntrees levels (1000, 2000, 3000, 4000, and 5000) and three mtry levels (from 1 to 3) were tested. The node size parameter was left at its default value. The values of the parameters in the final model were mtry = 2 and ntrees = 3000. Models with higher values of these parameters did not improve accuracy.

2.4.2. BRT

A BRT model can be tuned using several parameters such as the number of nodes in a tree (tree complexity), the contribution to the model of each tree (learning rate), the proportion of data to be selected at each step (bag fraction), and the average number of trees in the ensemble forest (ntrees). According to Elith et al. (2008), a decreasing (slowing) learning rate increases the value of ntrees required, and in general, a smaller value of learning rate (and a larger value of ntrees) is preferable, conditional on the number of observations and the time available for computation. All combinations of five ntrees levels (1000, 2000, 3000, 4000, and 5000) and five values of learning rate (0.05, 0.01, 0.005, and 0.001) were tested, resulting in optimum values of 3000 for ntrees and a learning rate of 0.005. The values corresponding to tree complexity and bag fraction were set at 5 and 0.5 respectively for each combination of ntrees and learning rate.

2.4.3. SVM

An SVM model requires a large number of parameters to be optimized: kernel functions (linear, polynomial, sigma, or radial basis), cost, the gamma of the kernel function (except the linear kernel), the bias of the kernel function (applicable only to the polynomial sigmold kernel), and finally the polynomial degree (applicable only to the polynomial kernel). For this reason, the optimization of an SVM model is more complicated than optimization of RF or BRT. The SVM model was calibrated using the R package kernlab. The parametrization of the model was done as follows: type = “C – bsvc”, kernel = “rbfdot” with kpar = “sigma(0.1). The cost was set to a range of values from 1 to 10 and was finally left at 1. The authors were aware that further testing of the parameters involved in SVM calibration was needed, but the lack of available computing power and the resulting long run times did not permit proper exploration of these issues.

2.4.4. LR

LR models are statistical models which provide insights into the relationship between a qualitative dependent variable, dichotomous in the present case, and one or more independent explanatory variables, whether qualitative or quantitative. The mathematical expression of LR models is:

\[ f(y_i) = \frac{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)} \]

In this work, the LR model was developed using a forward stepwise procedure in which the explanatory variables were introduced into the model one by one according to the resulting improvement in the model, as measured by the Akaike Information Criterion (AIC).

2.5. Model evaluation and comparison

To calculate and compare the classification accuracy of the four models, the area under the receiver operating characteristic (ROC) curve (AUC; Hanley and McNeil, 1982) was calculated. The ROC curve is a graphical representation of the false-positive error (1 – specificity, where specificity is the proportion of incorrect predictions) versus the true positive rate (also referred as sensitivity or the proportion of correct predictions) for a binary classifier system and for different values of the discrimination threshold (Zhou et al., 2002). The AUC is a threshold-independent metric because it evaluates the performance of a model at all possible threshold values (Franklin, 2010). AUC values ranged from 0.5 to 1, where 0.5 is analogous to a completely random prediction and 1 implies perfect prediction. AUC values between 0.5 and 0.7 denote poor performance, values between 0.7 and 0.9 denote moderately good performance, and values larger than 0.9 denote excellent model performance (McCune and Grace, 2002).
2.6. Evaluating variable importance

The evaluation of variable importance in the models was carried out using two different approaches. The first involves use of model-specific procedures, i.e., the increase in node purity for RF, the relative influence of the variables for BRT, and Z values for LR. The increase in node purity is measured by the gini criterion from all the splits in the forest based on a particular variable (Breiman, 2001). Relative influence measures the number of times that a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Friedman and Meulman, 2003). The relative influence (or contribution) of each variable is scaled so that the sum is 100. The second method for variable importance measurement is based on an AUC procedure as a jackknife estimator of variable importance, as described by Bar Massada et al. (2012). This procedure is based on the fact that a binary classification system can be used to calculate receiver operating characteristic curves (ROC) and to determine the precision of a diagnostic test (Ordóñez et al., 2012). Therefore, the method was based on measuring the change in AUC using the test data, a method which yields directly comparable results across the three ML models and the LR. The approach consists on removing predictor variables from the full model one at a time, training the model, and calculating the AUC using the test data. The difference between the full- and partial-model (without the variable) AUC indicates the contribution of each variable to the model. Therefore, it represents the information provided by a given variable that is not present in other variables. In addition, the AUC of the model was quantified using one variable at a time, the AUC values of single-variable models were compared, and the variables were ranked accordingly. This procedure enabled both a comparison of the accuracy of the models and a significance analysis of the explanatory variables within each model and compared to the other algorithms. Because there is no variable importance method implemented for SVM, this approach will be evaluated only using the AUC.

3. Results

3.1. Model performance

RF and BRT achieved the highest accuracy, reaching AUC values of 0.746 and 0.730 respectively. The SVM model yielded an AUC value of 0.709. The worst model accuracy was associated with the LR model, with a value of 0.686. According to the accuracy threshold proposed in McCune and Grace (2002), the ML models achieved moderate performance, while the LR had rather poor performance. Note also that RF, despite being the model with the fewest predictive variables (as reported in Table 1), reached the highest accuracy of classification (Fig. 3). Table 1 shows a brief summary of model accuracy and the explanatory variables considered in each model.

3.2. Variable importance

According both to model-specific procedures and to AUC jackknife estimation, negligence and accidents due to engines and machines working in or close to forest areas, the use of fire to clean up harvest wastes and crop boundaries, the increase in human presence and pressure near wildlands, and increasing concern for forest protection are the main factors related to human-induced wildfire occurrence in Spain (Fig. 3). The four explanatory variables linked to each typology of causes, i.e.,

![Density of machinery](image1)

![Changes in demographic potential](image2)

![Wildland-Agricultural Interface](image3)

![Protected areas](image4)

![Forestry area in public utility](image5)

![Density of machinery](image6)

![Changes in demographic potential](image7)

![Wildland-Agricultural Interface](image8)

![Protected areas](image9)

![Forestry area in public utility](image10)

**Fig. 3.** Increment in node purity (RF; top left), relative importance (BRT; top right) and absolute Z value (LR; bottom).
density of agricultural machinery (DAM), changes in demographic potential (CDP), wildland-agricultural interface (WAI), and protected areas (PA), make a significant contribution to all models (Table 1) and therefore are included in all of them. DAM and CDP are the main variables in each variable importance method, although WAI reduces its contribution while that of PA increases in the AUC procedure.

When considering each variable individually (univariate models), DAM and CDP are the two predictive variables with the highest explanatory power, with AUC values ranging from 0.669 to 0.687 and 0.630 to 0.659 respectively. This predictive power is supported by examining the models without these variables, where the losses in AUC are also the greatest. In the background, with modest AUC values, appear the rest of the variables, ordered from more to less contribution as follows: PA, WAI, FAPU, Wildland-Urban interface, power lines, and railroads (Table 2).

### 3.3. Spatial distribution of occurrence probability

The mapping of the spatial pattern of predicted wildfire occurrence is significantly different from one model to another (Fig. 5). A visual analysis reveals that RF has the highest spatial variability in predicted values. Despite having the same independent variables (Table 2), BRT and SVM present a very different spatial pattern. In the case of SVM, the pattern seems to be dichotomized, with probability values concentrated in two ranges close to the maximum and minimum values. On the other hand, the spatial distribution is more heterogeneous in BRT and closer to the RF distribution. Finally, the LR map shows a spatial distribution similar to BRT, but the fact that it contains almost no low probability values (in the range from 0 to 0.2) has hindered both its accuracy and its predictive power.

### 4. Discussion

Determining which model type to use for occurrence-distribution modeling is important because the outcomes may have direct management implications. Previous findings from species-distribution modeling (SDM, Franklin, 2010) for wildlife have suggested that ML algorithms may be more suitable than statistical models (Elith et al., 2006). ML models, and more specifically RF, enhance prediction accuracy compared with traditional statistical methods like LR. This improvement is reflected not only in the higher accuracy of the RF model, but also in the fact that fewer predictive variables are required to achieve this performance. In fact, the use of fewer variables is another point in its favor because it is preferable to use models which are as simple as possible, thereby facilitating
interpretation of the results. In addition, RF cartographic outputs (Fig. 5) seem to be more realistic because this method has increased spatial variability and therefore higher discriminatory power in neighboring areas with different occurrence values. Moreover, BRT, SVM, and LR present high area concentrations in some of the classification intervals (low-mid values in BRT, low-high in SVM, and midrange values in LR), while RF seems to be more equally distributed, although most of the probability values are located in the first interval (Fig. 4, Table 3). On the other hand, although the BRT model has similar performance to RF, its calibration and optimization involves more parameters, and therefore it is more difficult and time-consuming to compute. Finally, the authors consider that SVM is a less adequate method for predicting wildfire occurrences compared to the other proposed methods. This is mainly because its calibration is significantly more difficult, its optimization is too time-consuming, and its accuracy does not reach RF or BRT levels, remaining closer to LR. These findings are supported by the results reported in Bar Massada et al. (2012), where the RF algorithm is proposed as the most adequate compared to logit GLM and Maxent models.

As for the explanatory variables, as in most human-dominated landscapes where anthropogenic ignitions surpass natural

**Table 3**

Summary of area (km²) distribution for each interval and model.

<table>
<thead>
<tr>
<th>Interval</th>
<th>RF</th>
<th>BRT</th>
<th>SVM</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very low</td>
<td>0.0–0.2</td>
<td>91,540</td>
<td>31,798</td>
<td>172,777</td>
</tr>
<tr>
<td>Low</td>
<td>0.2–0.4</td>
<td>81,069</td>
<td>139,375</td>
<td>22,401</td>
</tr>
<tr>
<td>Medium</td>
<td>0.4–0.6</td>
<td>40,375</td>
<td>51,327</td>
<td>15,054</td>
</tr>
<tr>
<td>High</td>
<td>0.6–0.8</td>
<td>24,438</td>
<td>29,779</td>
<td>12,088</td>
</tr>
<tr>
<td>Very high</td>
<td>0.8–1.0</td>
<td>19,437</td>
<td>14580</td>
<td>34,539</td>
</tr>
</tbody>
</table>

Fig. 5. Spatial pattern of the probability of wildfire occurrence for each model. Top left is RF, top right BRT, bottom left SVM, and bottom right LR.
ignitions, in peninsular Spain, both human accessibility and population density are likely to be strong predictors of ignition risk (Bar Massada et al., 2012). However, although these factors make an important contribution in the study area (CDP or WUI, Galiana-Martin et al., 2011), other explanatory factors more related to agricultural activities and forest management also influence wildfire occurrence. More specifically, these involve negligence and accidents due to engines and machines working close to the forest areas (DAM), the use of fire in cleanup of harvest wastes and crop boundaries (WAI), and forestry protection policies (PA). Nonetheless, whereas DAM, CDP, WAI, and WUI are factors related to an increased ignition probability, PA intervenes in the opposite way, i.e., by decreasing ignition likelihood, because it is directly linked with the protection and conservation of landscape (Figs. 6 and 7).

In particular, DAM and CDP have proved to be the variables most closely related to fire occurrence in Spain. However, this high predictive power is also linked to the fact that these variables are continuous in nature, i.e., they have values in all the locations throughout the study area. Therefore, DAM and CDP can function as discriminatory variables in all cases, while the other predictors cannot. Note also that the continuous nature of DAM does not arise from there being machinery in all locations, but from the fact that it is obtained as a statistical value reported at the municipal level. Moreover, the influence of DAM should ideally be restricted to mechanized WAI because the focus is on ignitions in forest areas, and therefore the agricultural machinery must be in crop areas close to forest surfaces, i.e., being used for WAI. This fact is supported by the (non-collinear) interaction between DAM and WAI (Figs. 6 and 7), in which high values of probabilities are related to high values both of DAM and of WAI. On the other hand, CDP seem to have an inverse relationship with wildfire probabilities (Figs. 6 and 7), which means that a decrease in the demographic potential involves an increase in the predicted probability. This may appear strange or hard to understand, but might be related to the fact that in certain locations where nowadays the potential is lower than the initial potential, this initial potential has been almost completely overwhelmed. This indicates that the anthropic pressure in these areas has been significantly strong during the CDP time span (1991–2006), leading to an increased occurrence probability.

5. Conclusions

ML models improve the prediction accuracy of traditional regression methods. Either RF or BRT models yield an improvement in accuracy over LR methods for wildfire occurrence assessment, according to AUC values. More specifically, the RF algorithm seems to be the best choice due not only to its higher accuracy, but also to the fact that fewer predictive variables are required to achieve this accuracy. In addition, its calibration is easier because it involves few
parameters. Another advantage of RF is its cartographic outputs, which seem to be more realistic than those from other models due to RF's higher spatial variability and therefore greater spatial discriminatory power. This enables RF to provide a better reflection of variability in wildfire occurrence linked to heterogeneity of landscapes and human activities. SVM appears to be a less adequate method for predicting wildfire occurrences, mainly because its calibration is significantly more difficult, its optimization is too time-consuming, and its accuracy does not reach the levels of RF or BRT, remaining closer to LR. However, despite the similar predictive power of the proposed models, the resulting predictive maps were very different. This was especially noteworthy in the case of SVM, where the spatial patterns seem to be dichotomized, with probability values concentrated in two ranges close to the maximum and minimum values. Nevertheless, no single model or method can be considered as the perfect modeling tool (Elith et al., 2006), and prediction of wildfire occurrences may benefit from using multiple approaches, yielding a range of predictions rather than a single map (Bar Massada et al., 2012).

Regardless of the method considered, DAM and CDP have proved to be the variables most closely related to fire occurrence, although this result is partially due to the continuous nature of these variables and, in the case of DAM, to interactions with other predictive variables like WAI. In any case, fire occurrence in Spain is mainly related to the increase of human pressure on wildlands and to accidents or negligence in the course of agricultural work (Chuvieco et al., 2012).

Acknowledgments

This work was financed by the National I+D Plan of the Spanish Ministry of Science and Innovation: FPI grant BES-2009-023728. The research was conducted within the framework of the following projects: FIREGLOBE: Analysis of fire risk scenarios at the national and global scales (CGL2008-01083/CLI); and PYRORAMA Modeling future scenarios of fire risk and land use changes at national scale (CGL2011-29619-C03-01), subproject 1 of the PYROSKENE project (CGL2011-29619-C03).

References


