Human and Ecological Risk Assessment, 19: 254–267, 2013

ISSN: 1080-7039 print / 1549-7860 online
DOI: 10.1080/10807039.2012.683736

STATISTICAL MODELS AND METHODS ARTICLE

A Bootstrap-Based Covariate Selection Method for Modeling the Risk of Lightning-Induced Fires at a Local Scale: A Case Study in Northwest Spain

Celestino Ordóñez,[1] Javier Roca-Pardiñas,[2] Fernando Castedo-Dorado,[3] and José R. Rodríguez-Pérez[3]
[1]Department of Mining Exploitation and Prospecting, University of Oviedo, Oviedo, Spain; [2]Department of Statistics and Operations Research, University of Vigo, C/Torrecedeira, Vigo, Pontevedra, Spain; [3]Geomatics Engineering Research Group, University of León, Ponferrada, León, Spain

ABSTRACT

In lightning-induced fire risk prediction models, the number of potential predictors is usually high, with some redundancy among them. It is therefore important to select the best subset of predictors that obtain models with the greatest discrimination capacity. With this aim in mind, the logistic generalized linear model was used to estimate lightning-induced fire occurrence using a case study of the province of León (northwest Spain). A bootstrap-based test was used to obtain the optimal number of predictors and to model this optimal number of predictors displaying the largest area under the receiver operating characteristics curve. The results show that of the 16 variables initially considered, only three were necessary to obtain the model with the best discriminatory capacity for estimating lightning-induced fire occurrence. Moreover, this model can be considered equivalent to another nine alternative models with three covariates. Both the optimal and the equivalent models are useful in the spatially explicit assessment of fire risk, the planning and coordination of regional efforts to identify areas at greatest risk, and the design of long-term wildfire management strategies. The methodology used for this case study can be applied to other wildfire risk assessment situations where multiple and interconnected covariates are available.

Received 4 July 2011; revised manuscript accepted 23 October 2011.
Address correspondence to Celestino Ordóñez, Department of Environmental Engineering, University of Vigo, C/Maxwell s/n, Campus Lagoas-Marcosende, 36310-Vigo, Pontevedra, Spain. E-mail: cgalan@uvigo.es

Key Words: bootstrap-based test, variable selection, logistic generalized linear model, area under the ROC curve, lightning-induced fire risk.

INTRODUCTION

In the countries around the Mediterranean Basin and in other Mediterranean-type areas of the world, forest fires are predominantly caused by humans (Vázquez and Moreno 1998). Lightning is also an important ignition cause (Pyne *et al.* 1996), however, and lightning-ignited fires can burn larger areas of forest than those caused by humans because of remoteness and aggregation in time and space (Podur *et al.* 2003).

Several investigators have reported that lightning-induced fires do not occur at random but tend to start in specific places (Vankat 1985). The effectiveness of individual lightning strikes in igniting a forest fire is potentially affected by the following factors: variations in lightning properties such as quantity, polarity and intensity; forest fuel moisture properties resulting from recent weather conditions including rainfall, temperature, and humidity; topographic variables that may affect the aforementioned variables (Diaz-Avalos *et al.* 2001); and rates of combustion, which vary according to the type of fuel. The relative importance of all these variables varies according to the scale considered; however, it is unrealistic to present a general model for large scales, thus making it advisable to develop models at local or regional scales (Pacheco *et al.* 2009).

In order to prevent, minimize, and mitigate the effects of lightning-induced forest fires, a priori risk analysis and maps indicating vulnerable areas are very useful (Bonazountas *et al.* 2005). In the literature, different statistical methods have been used to document spatial patterns of lightning-induced fires, namely, linear regression, logistic regression, multivariate discriminant analysis, classification and regression trees, neural networks, and so on. Logistic regression analysis has been particularly successful in predicting fire occurrence and in examining the most critical factors involved in fire incidence (García *et al.* 1995; Vasconcelos *et al.* 2001; Andrews *et al.* 2003; Wotton and Martell 2005; Martinez *et al.* 2009; Vilar *et al.* 2010).

One of the main problems associated with the development of a logistic fire risk model is identifying the best subsets of predictors that establish the model or models with the best discriminatory capacity. This problem is important both in human-caused and lightning-induced risk assessment because the number of potential covariates is often very high and many predictors are mutually redundant (Martínez *et al.* 2009). In general, the more variables added to a model, the better the apparent fit of the observed data; however, the inclusion of irrelevant variables can increase the variance of the ensuing estimates (leading to a loss in the predictive capacity of the model) and can make the fitted model difficult to interpret.

Automatic selection procedures such as stepwise selection or backward elimination are usually not very appropriate because the selected models can contain irrelevant variables (Hosmer and Lemeshow 2000). In addition, for practical applications of the model, some of the potential explanatory variables are not easily available or cannot be obtained accurately. The bootstrap-based test used in this study overcomes these drawbacks since it detects optimal combination of the variables that best

discriminate between groups (two, in our case—presence and absence), producing parsimonious prediction models.

The aim of this research was to test the bootstrap method for developing a lightning-induced fire prediction model, using data for a case study referring to the province of León (northwest Spain). The applicability of this methodology for fire risk assessment is also discussed.

METHODS AND MATERIALS

Data

The data used were obtained from five different sources over a 6-year period (2002–2007), although only the months of May to September (inclusive) were considered because lightning strikes and lightning-caused fires occurred almost exclusively in this period (Castedo-Dorado *et al.* 2011). The period 2002–2007 was selected because: (1) the year 2002 was the first year in which wildfire ignition points were recorded using X,Y coordinates; (2) it admits the assumption that land cover does not differ from that of 2003, the year for which land cover data was available; and (3) since 2001 flash detection efficiency has exceeded 90% and average location accuracy has improved to within 0.5 km.

The lightning location database (supplied by AEMET, the Spanish Meteorological Agency) provides information such as lightning intensity, polarity, date and time, estimated coordinates of lightning strikes, and quality of location estimates. In order to reduce the uncertainty of the number of strikes in each type of land cover, only flashes with $\chi 2$ equal to or less than 2 and a long semi-axis radius of the error ellipse equal to or less than 1.5 km were used (Nieto *et al.* 2006); this is important in the studied area (the province of León) because it is characterized by fragmented land cover. Therefore the database used for this study included 78,256 lightning flashes.

The Spanish Ministry of the Environment and Rural and Marine Affairs provided data on ignition locations of lightning-induced wildfires. This database includes information about detection time, ignition location, and estimated cause of ignition. Topographic variables (altitude, aspect and slope) were calculated using the digital elevation model (DEM) provided by the Spanish National Cartographic Database.

The composition and structure of land cover was obtained from the digital Spanish Forest Map for the province of León (Spanish Ministry of the Environment 2003). To deal with more homogeneous types of land cover, the categories were grouped in the following classes: coniferous woodland, broadleaf woodland, mixed woodland, non-combustible areas, gallery woodland, open woodland, mosaics of woodland and others, shrubland, grassland, recently deforested areas, and recent forest plantations and reforestations.

Provided by AEMET was daily meteorological data (hourly or daily observations of temperature, relative humidity, and rainfall) that were derived from raw data recorded at 344 weather stations located in León and nearby provinces. Geostatistical methods were used to interpolate weather station data. Universal kriging was used for rainfall and relative humidity data calculations, whereas co-kriging interpolation was used to model temperature.

The studied area (province of León, Figure 1A) was partitioned in pixels of 3 × 3 km, resulting in a total of 1882 pixels. All digital information available for the

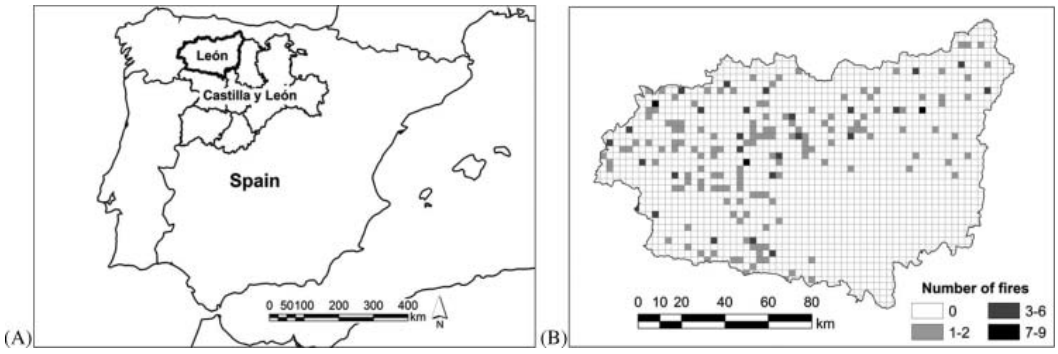

Figure 1. Studied area. (A) The province of León, Spain and (B) the location and number of observed lightning-induced fires for 2002–2007.

analysis was converted to this 3 × 3 km spatial resolution. This particular grid size was selected because it represents a compromise between resolution requirements, interpolation accuracy, and computation cost. Lightning-induced fires occurred in 179 pixels in the period 2002–2007 (Figure 1B).

According to previous studies (Castedo-Dorado *et al*. 2011), 16 variables related to lightning activity, topography, vegetation cover, lightning–vegetation cover interaction, and meteorology were selected for modeling and predicting lightning-induced fires (Table 1).

Methods

The spatial patterns of occurrence were analysed by seeking relationships between the presence/absence of lightning-caused forest fires in a 3 × 3 km grid and the potential explanatory variables converted to this scale by a geographic information system.

Table 1. Potential explanatory variables used in the bootstrap analysis.

Variable	Meaning
X_1	Mean altitude (m)
X_2	Mean slope (%)
X_3	Number of strikes (flashes)
X_4	Number of strikes in shrubland
X_5	Mean intensity (kA) of negative flashes
X_6	Mean intensity (kA) of positive flashes
X_7	Number of positive flashes
X_8	Percentage of coniferous woodland
X_9	Percentage of agricultural areas
X_{10}	Percentage of forested areas
X_{11}	Number of strikes in coniferous woodland
X_{12}	Number of strikes in broadleaf woodland
X_{13}	Number of strikes in forested areas
X_{14}	Number of strikes in agricultural areas
X_{15}	Number of thunderstorms days without rainfall
X_{16}	Number of days with moisture lower than average moisture

The 1882 pixels representing the province of León were coded with 0 or 1, representing, respectively, the absence or occurrence of one or more lightning-induced ignitions in the period 2002–2007. In a logistic regression framework, discrimination between the two possible states of the binary response depends on a set of p continuous covariates (X_j). The expression of a logistic generalized linear model (GLM) (McCullagh and Nelder 1989) is given by:

$$p\,(\mathbf{X}) = p\,(\mathbf{Y} = 1|\mathbf{X}) = \frac{\exp\left(\alpha + \alpha_1 \cdot X_1 + \ldots + \alpha_p \cdot X_p\right)}{1 + \exp\left(\alpha + \alpha_1 \cdot X_1 + \ldots + \alpha_p \cdot X_p\right)} \tag{1}$$

where X_j ($j = 1, \ldots, p$) are the explanatory covariates, α is a constant, and α_j ($j = 1, \ldots, p$) are the unknown parameters.

The proposed bootstrap-based test for covariate selection selects a GLM containing a subset of q variables ($q \leq p$), and eliminates the remainder from the model, according to an optimal criterion based on a receiver operation characteristic (ROC) curve (Swets and Pickett 1982). In practice, the area under the ROC curve (AUC) is one of the most widely used criteria for comparing the discriminatory capacity of a series of binary response regression models and the AUC is more suitable when the variable is Boolean.

The ROC curve relies on false/true–positive/negative tests, where sensitivity is the proportion of event responses that were predicted to be events and specificity is the proportion of non-event responses that were predicted to be non-events. The plot of sensitivity (*i.e.*, hit rate) versus 1-specificity (*i.e.*, false alarm rate) is the ROC curve; the AUC measures the accuracy of the detection system and does not require any assumptions concerning the shape, form, or underlying signal and noise distributions (Saveland and Neuenschwander 1990). This statistic is measure of model discrimination: 0.5 suggests no discrimination, 0.7–0.8 suggests acceptable discrimination and 0.8–0.9 suggests excellent discrimination (Hosmer and Lemeshow 2000).

Bootstrap-Based Test for Covariate Selection

In order to determine the variables to be introduced into the GLM in (1), a bootstrap-based test was used. For a given size k, let AUC (k) be the AUC obtained with the best subset of k variables:

$$\mathrm{AUC}\,(k) = \max_{1 \leq j_1 < j_2 < \ldots < j_k \leq p} \mathrm{AUC}_{j_1, \ldots, j_k} \tag{2}$$

where $\mathrm{AUC}_{j_1, \ldots, j_k}$ is the AUC obtained from the ROC constructed with the theoretical probabilities given by the GLM:

$$p_{j_1, \ldots, j_k}(\mathbf{X}_i) = \frac{\exp\left(\hat{\alpha} + \hat{\alpha}_{j1} \cdot X_{j1} + \ldots + \hat{\alpha}_{jk} \cdot X_{jk}\right)}{1 + \exp\left(\hat{\alpha} + \hat{\alpha}_{j1} \cdot X_{j1} + \ldots + \hat{\alpha}_{jk} \cdot X_{jk}\right)} \tag{3}$$

Given a subset of size q, consideration is given to a test for the null hypothesis:

$$\mathbf{H}_0\,(q) : \mathrm{AUC}\,(q) \geq \max_{r > q} \mathrm{AUC}\,(r) \tag{4}$$

versus the general hypothesis:

$$\mathbf{H}_1 : \mathrm{AUC}\,(q) < \max_{r > q} \mathrm{AUC}\,(r) \tag{5}$$

That is to say, under the null hypothesis, the maximum AUC is obtained with some subset of q covariates whereas the maximum AUC is obtained with some combination of $r > q$ covariates.

To test $\mathbf{H}_0$, given the sample $\{X_i, Y_i\}_{i=1}^n$ of (X,Y), we used the following test statistics:

$$\begin{aligned} &\hat{T} = \max_{r \geq q} \widehat{\mathrm{AUC}}\,(r) - \widehat{\mathrm{AUC}}\,(q) \\ &\text{with} \quad \widehat{\mathrm{AUC}}\,(k) = \max_{1 \leq j_1 < j_2 < \ldots < j_k \leq p} \widehat{\mathrm{AUC}}_{j_1, \ldots, j_k} \end{aligned} \tag{6}$$

where $\widehat{\mathrm{AUC}}_{j_1, \ldots, j_k}$ is the AUC obtained from the estimated probabilities $\hat{p}_{j_1, \ldots, j_k}(X_i)$ for $i = 1, \ldots, n$, obtained by fitting the model in Eq. (3) and leaving out the i^{th} data point.

This estimation can be carried out using the local scoring algorithm (Opsomer 2000). Briefly, the local scoring algorithm is analogous to the use of iteratively reweighted least squares (McCullagh and Nelder 1989) for solving likelihood and nonlinear regression equations.

Note that if the null hypothesis is verified, then $\hat{T}$ should be close to zero, but will generally be positive. Therefore, the test rule for checking $\mathbf{H}_0$ is that the null hypothesis is rejected if $\hat{T}$ is large enough. In order to detect if the true T is significantly positive, it is necessary to build an interval $[a, \infty)$where the hypothetical T value is placed with a determined probability. That is to say, the lower a are calculated such that the following probability is complied with: $p(\hat{T} > a) = 1 - \alpha$. Therefore, the test rule for checking $\mathbf{H}_0$, with an asymptotic significance level α, is that the null hypothesis is rejected if $a > 0$. However, in order to obtain a it is necessary to know the distribution of the estimate for $\hat{T}$. It is well known, nevertheless, that the asymptotic theory for determining such percentiles is not closed; therefore, resampling methods such as bootstrap introduced by Efron (1979) are widely used for this purpose.

The steps to obtain the value for a are as follows:

- Obtain $\hat{T}$ from the sample data as explained above.
- For $b = 1$ to B (*e.g.*, B = 1000), simulate the bootstrap sample $\left\{\left(\mathrm{X}_\mathrm{i}^{*\mathrm{b}}, \mathrm{Y}_\mathrm{i}^{*\mathrm{b}}\right)\right\}_{\mathrm{i}=1}^{\mathrm{n}}$ by randomly sampling the n items from the original dataset $\{(\mathrm{X_i}, \mathrm{Y_i})\}_{\mathrm{i}=1}^{\mathrm{n}}$ with replacement (*i.e.*, each individual value (X_i, Y_i) has a probability n^{-1} of re-occurring) and obtain the bootstrap estimates $\hat{\mathrm{T}}^{*\mathrm{b}}$.

Finally, the value of a is given by $a = \hat{T}^{\alpha/2}$ where $\hat{T}^p$ represents the percentile p of the bootstrapped estimates $\hat{T}^{*1}, \ldots, \hat{T}^{*B}$. The test rule based on T consists of rejecting the null hypothesis if $a > 0$. Applying this test to $q = 1, \ldots, p-1$ is an important step in a covariate selection procedure. If H_0 is not rejected, only the subset of the q covariates $X_{j_1}, \ldots, X_{j_q}$. that maximize $\widehat{\mathrm{AUC}}(q)$ will be retained, with the remaining covariates eliminated from the model. In all other cases, the test will be repeated with $q+1$ variables until the null hypothesis is not rejected. More details

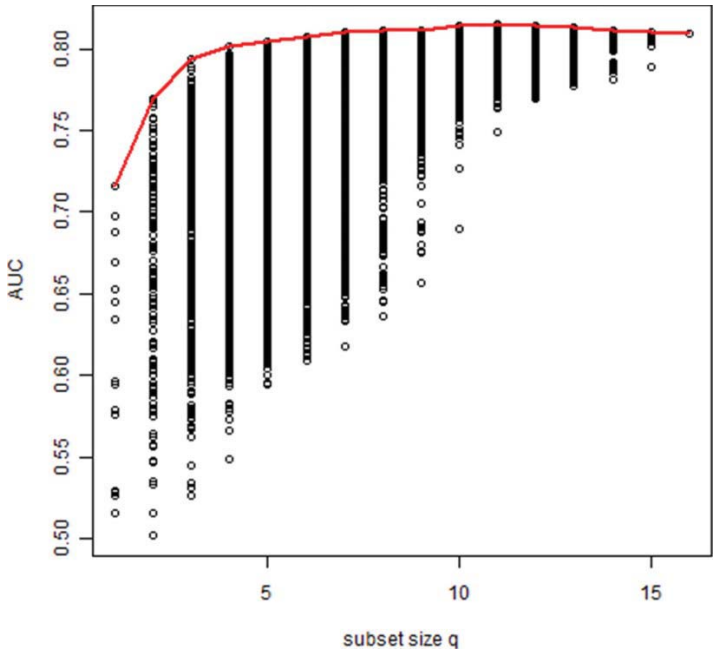

Figure 2. All possible subset models for the example. For each subset size, the area under the ROC curve (AUC) is shown for each model. (Color figure available online.)

of the bootstrap methodology for variable selection are given in Roca-Pardiñas *et al.* (2009).

RESULTS AND DISCUSSION

A plot of the AUC corresponding to all the possible model combinations (from 1 to 16 explanatory variables) is shown in Figure 2. In each subset, q represents the number of variables included in the model. The evolving curve corresponding to the models with the highest AUC values is also presented (Table 2). Also included, along with the increases in the AUC obtained in response to rises in q, is the number of new variables included in the model in order to obtain this maximum value each time. The increases in the AUC values are plotted in Figure 3. Based on Table 2, it can be observed that if only one variable ($q = 1$) is selected, the best AUC is obtained for the variable X_{10} (forested area). In this case, the area equals 0.716, which suggests acceptable discriminatory ability.

The AUC increases as the number of variables included in the model rises to a given q. Furthermore, it should be noted that, as new variables enter, the AUC continues to increase, although more gradually, as is evident from the Δ_{AUC} column in Table 2 and from Figure 3. This is indicative of the fact that, as the number of variables increases, the new variables included in the model represent a refined behaviour with respect to those that are already present. The result is that the improvements in the AUC value are increasingly subtle. Moreover, when $q > 11$, the value of the AUC is reduced.

Table 2. AUC and increment in AUC ($\hat{T}$) obtained with each selected model of size q.

q	AUC	$\hat{T}$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
1	0.716	—										x						
2	0.769	0.053	x									x						
3	**0.794**	**0.025**	**x**									**x**	**x**					
4	0.802	0.008	x									x	x	x				
5	0.804	0.002	x								x		x	x			x	
6	0.807	0.003	x				x					x	x	x			x	
7	0.810	0.003	x				x	x	x			x	x		x			
8	0.811	0.001	x				x	x	x	x		x	x		x			
9	0.812	0.001	x	x	x		x	x	x		x	x	x					
10	0.814	0.002	x	x	x		x	x	x		x	x	x			x		
11	**0.815**	**0.001**	**x**	**x**			**x**	**x**	**x**	**x**	**x**	**x**	**x**		**x**	**x**		
12	0.814	–0.001	x	x		x	x	x	x	x	x	x	x		x	x		
13	0.813	–0.001	x	x	x	x	x	x	x		x	x	x	x	x	x		
14	0.811	–0.002	x	x	x	x	x	x	x		x	x	x		x	x	x	x
15	0.810	–0.001	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
16	0.809	–0.001	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

For $q > 11$ (marked in bold), the AUC decreases.

From the analysis of the way in which the AUC evolves in response to the inclusion of new variables, it is possible to deduce that there is an optimal intermediate point between the number of variables that are considered in the model (preferably low) and the AUC value (preferably high). To delimit this point, the test for the

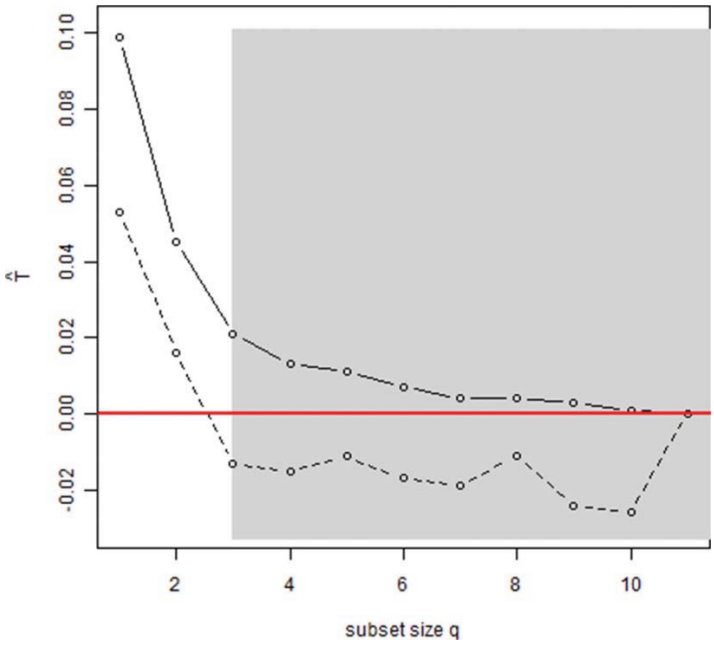

Figure 3. Value of $\hat{T}$ (continuous line) and the corresponding limit a (broken line) for the 95% confidence interval of T. (Color figure available online.)

null hypothesis $\mathbf{H}_0(q)$ described in the Methods section was applied for each q. For a 5% significance level, the null hypothesis was rejected until $q = 3$ and accepted thereafter. Therefore, it can be concluded that the best subset of variables are X_1 (altitude), X_{10} (forested area), and X_{11} (number of strikes in coniferous woodland). The AUC for this model indicated acceptable discriminatory ability (area, 0.7942). For this case, the parameter estimates for the logistic GLM were $\alpha = -4.28$, $\alpha_1 = -0.00156$, $\alpha_{10} = 0.00404$, and $\alpha_{11} = 0.0528$. All the parameter estimates were significant at the 1% level.

The procedure outlined above serves to determine the number of variables to be included in the model. In addition, it ensures that the best choice of variables of size q. In practice, however, various statistically equivalent optimal models of size q can be obtained. Taking into account the corresponding test for the null hypothesis $\mathbf{H}_0(q)$, included in Table 3 are some models with $q = 3$ that are equivalent to the optimal model. Nine equivalent models can be obtained, all of them with very similar discriminatory ability (area, 0.7784–0.7888). Variables X_1, X_{10}, or X_{11} are always included in the equivalent models and this combination has been chosen because it coincides with that obtained previously (Castedo-Dorado *et al.* 2011). Other combinations are possible, but the advantage of the proposed method is, in fact, that it is possible to choose the combination that is easiest to interpret.

According to the α_1 parameter estimate, altitude (X_1 variable) had a negative effect on the probability of occurrence of lightning-induced fires; that is, lower and intermediate elevations were found to be the most prone to fire. Although several studies have shown a positive relationship between altitude and lightning occurrence (Dissing and Verbyla 2003)—even in the studied region (Rivas Soriano *et al.* 2001, 2005)—higher rainfall and lower temperatures at higher altitudes may cause a negative link (Martín and Means 1982; Díaz-Avalos *et al.* 2001). Moreover, the altitudinal ecological limit of woodlands may also be important (Dissing and Verbyla 2003). These outcomes suggest that lightning-induced fires occur at altitudes where fuel continuity and moisture are not limiting factors (Martín and Means 1982). Other authors have reported similar results for lightning-induced fires in other regions of Spain (Nieto *et al.* 2006).

Table 3. Models that are equivalent to the optimal with $q = 3$.

AUC	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
79.42	x									x	x					
78.88										x	x	x				
78.64								x		x	x					
78.21								x	x		x					
77.98									x		x	x				
77.97	x								x						x	
77.96	x									x					x	
77.95	x								x		x					
77.90				x				x			x					
77.84					x					x	x					

According to Table 3, the other topographic variable tested (X_2, slope) was not found to be significant for any of the models equivalent to the optimal model. Similar results were found in other studies (McRae 1992 for Australia). However, Díaz-Ávalos *et al.* (2001), Wierzchowski *et al.* (2002), and Conedera *et al.* (2006) found that lightning-induced fires mainly occurred on steeper slopes.

The presence of forested areas (variable X_{10}) was found to be significant both in the optimal model and in four of the nine equivalent models. It was positively associated with lightning-caused fire ignitions, confirming the results of Vázquez and Moreno (1998), who found that lightning-induced fires in Spain affected a greater proportion of woodlands than human-induced fires. This may be the result of the canopy sheltering the forest floor from rainfall associated with lightning (Kourtz and Todd 1992).

More surprising, perhaps, is the significance of the number of strikes in coniferous woodland (variable X_{11}) in most of the equivalent models, bearing in mind that this type of vegetation represents less than 25% of the woodland area in the province of León (Junta de Castilla y León 2005). This seems to confirm that some types of vegetation cover are more prone to lightning-induced fire than others (Manry and Knight 1986; Granstrom 1993; Dissing and Verbyla 2003; Krawchuk *et al.* 2006; Evett *et al.* 2008). The fact that variable X_{11} is contained in almost all the optimal models is some indicator of its importance, although this has not been proved mathematically.

Lightning ignition may occur when the electrical current ignites fine fuels on the forest floor (usually duff) at the base of a tree (Latham and Williams 2001) or when living trees act as lightning conductors (Ogilcie 1989). Differences in duff layer, produced by differences in vegetation type, would also result in different rates of heating and therefore differences in flammability (Latham and Williams 2001). Duff layer of needles under conifers is a more suitable fire ignition source than the duff layer in a hardwood stand (Flannigan and Wotton 1991). Deciduous species decrease the duff depth and, consequently, the probability of ignition (Latham and Schlieter 1989).

In addition, the high flammability of coniferous species (due to their high resin and essential oil content) and the abundance of highly flammable species in the understory of the *Pinus* stands (*Erica* sp., *Genistella tridentata*, *Calluna vulgaris*, *etc.*) in the province of León may help fire propagation after ignition (Vélez 1990; Bond and Van Wilgen 1996).

The bootstrap methodology revealed no relationship between lightning density (variables X_3 and X_7) and the characteristics of lightning discharges (charge and intensity, variables X_5 and X_6), and fuel ignition. Regarding lightning density variables, the explanation is that higher densities of flashes are generally followed by greater rainfall (Rorig and Ferguson 1999; Rorig and Ferguson 2002; Álvarez-Lamata 2005). A similar result was found for Spain by Vázquez and Moreno (1993), suggesting that a very large number of lightning strikes is needed for several ignitions to take place.

The resulting map of the probability of occurrence of lightning-induced fires for the period 2002–2007 is shown in Figure 4. Probability estimates were obtained from the optimal model since it had the best discriminatory capacity. The highest probability of fire is at intermediate altitude in the province where coniferous

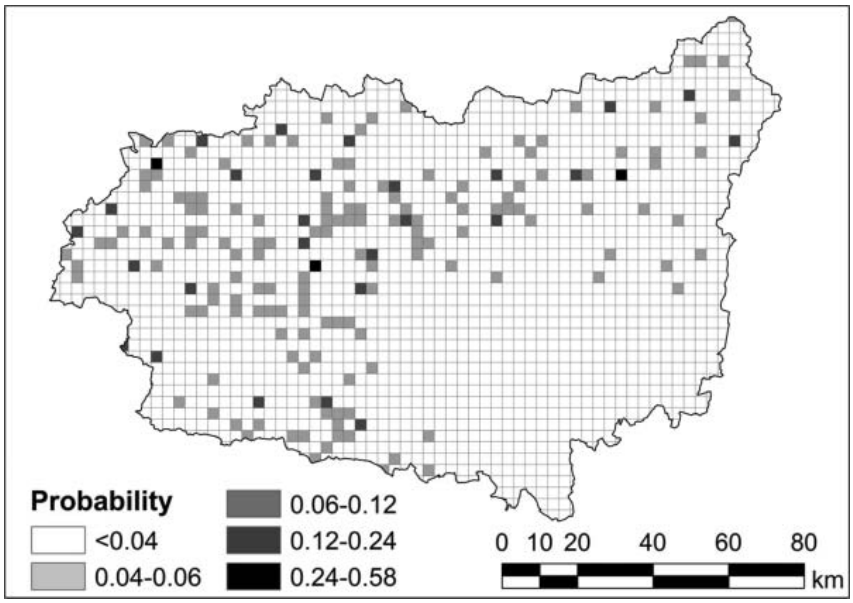

Figure 4. Spatial distribution of the probability of occurrence of lightning-induced fires in the province of León in 2002–2007 (pixels, 3 × 3 km). Probabilities were based on the optimal model (variables X_1, X_{10}, and X_{11}).

woodlands are the dominant type of land cover. Comparing Figures 1B and 4, it can be observed that there is general agreement regarding the most lightning-induced prone areas. The overestimation is probably due to the fact that the number of pixels where lightning fires do not occur is much greater than the percentage of pixels where lighting fires do occur. In fact, in a previous analysis of the data using logistic regression and automatic selection procedures for the independent variables (Castedo-Dorado *et al.* 2011), the percentage of pixels where a lightning-caused fire can be expected is also over-predicted. However, it must be borne in mind that errors resulting in overestimation of the number of pixels in relation to fire occurrence are of much less important than errors resulting in underestimation.

CONCLUSIONS

We used ROC analysis based on the logistic GLM for binary responses to determine the number of covariates q and select the best subsets of size q that would establish the model with best discriminatory capacity for estimating lightning-induced fire occurrence for a case study of León province in northwest Spain. Of the 16 variables initially considered, only three were necessary to obtain an optimal model, based on altitude, forested area and the number of strikes in coniferous woodland as independent variables. The AUC for this model indicated acceptable discriminatory ability (area, 0.7942). This optimal model can be considered equivalent to another nine models with three covariates. This is very interesting from a practical point of view, because it models enables models to be generated that include variables of interest (*i.e.*, easy to measure, easy to obtain accurately, *etc*).

This statistical methodology can be usefully applied to the spatially explicit assessment of fire risk, for combination with models such as FARSITE (Finney 1998), to plan and coordinate efforts to identify areas at greatest risk and to design long-term wildfire management strategies. In addition, both the optimal and equivalent models can easily be integrated with other sources of risk (*e.g.*, socioeconomic causes) in operational wildfire risk systems, within geographic information systems.

The proposed method is time consuming to apply but is easily automated, as the same model is used for different combinations of independent variables. The methodology used for this case study can be applied to other wildfire risk assessment situations where multiple and interconnected covariates are available. Studies can also be carried out at different spatial scales.

Future research will cover the application of a spatial regression logistic model to solving the same problem so as to take into account possible spatially correlated variables.

ACKNOWLEDGMENTS

The Spanish meteorology agency, AEMET, provided the lightning strike and meteorological data. This research was supported by funding provided by the Diputación of León for the project titled "Modelización de la probabilidad espacial y temporal de ocurrencia de incendios forestales por rayo en la provincia de León."

REFERENCES

Álvarez-Lamata E. 2005. Los incendios forestales y las condiciones meteorológicas en Aragón. In: 4th Congreso Forestal español, Zaragoza, September, 26–30. Sociedad Española de Ciencias Forestales y Gobierno de Aragón [In Spanish]

Andrews PL, Loftsgaarden DO, and Bradshaw LS. 2003. Evaluation of fire danger rating indexes using logistic regression and percentile analysis. Internat J Wildland Fire 12(2):213–26

Bonazountas M, Kallidromitou D, Kassomenos PA, *et al.* 2005. Forest fire risk analysis. Hum Ecol Risk Assess 11(3):617–26

Bond WJ and van Wilgen BW. 1996. Fire and Plants. Chapman & Hall, London, UK

Castedo-Dorado F, Rodríguez-Pérez JR, Marcos Menéndez JL, *et al.* 2011. Modelling the probability of lightning-induced forest fires occurrence in the province of León (NW Spain). For Syst 20(1):95–107

Conedera M, Cesti G, Pezzatti GB, *et al.* 2006. Lightning-induced fires in the alpine region: An increasing problem. For Ecol Manage 234(Supplement 1):S68

Díaz-Avalos C, Peterson DL, Alvarado E, *et al.* 2001. Space-time modelling of lightning-caused ignitions in the Blue Mountains, Oregon. Can J For Res 31(9):1579–93

Dissing D and Verbyla D. 2003. Spatial patterns of lightning strikes in interior Alaska and their relations to elevation and vegetation. Can J For Res 33(5):770–82

Efron B. 1979. Bootstrap methods—Another look at the jackknife. Ann Stat 7(1):1–26

Evett RR, Mohrle CR, Hall BL, *et al.* 2008. The effect of monsoonal atmospheric moisture on lightning fire ignitions in southwestern North America. Agric For Meteorol 148(10):1478–87

Finney MA. 1998. FARSITE: Fire Area Simulator–Model Development and Evaluation. Research Paper RMRS-RP-4. USDA Forest Service, Rocky Mountain Research Station, Ft. Collins, CO, USA

Flannigan MD and Wotton BM. 1991. Lightning-ignited forest-fires in northwestern Ontario. Can J For Res 21(3):277–87

García CV, Woodard PM, Titus SJ, *et al.* 1995. A logit model for predicting the daily occurrence of human caused forest-fires. Internat J Wildland Fire 5(2):101–11

Granstrom A. 1993. Spatial and temporal variation in lightning ignitions in Sweden. J Veg Sci 4(6):737–44

Hosmer D and Lemeshow S. 2000. Applied Logistic Regression, 2nd edit. Wiley-Interscience, New York, NY, USA

Junta de Castilla y León. 2005. Castilla y León crece con el bosque. Consejería de Medio Ambiente, Junta de Castilla y León. Serie Divulgativa

Kourtz PH and Todd B. 1992. Predicting the Daily Occurrence of Lightning-Caused Forest Fires. Inf Rep PI-X-112. Canadian Forest Service, Petawawa, Canada

Krawchuk M, Cumming S, Flannigan M, *et al.* 2006. Biotic and abiotic regulation of lightning fire initiation in the mixedwood boreal forest. Ecology 87(2):458–68

Latham D and Williams E. 2001. Lightning and forest fires. In: Johnson EA and Miyanishi K (eds), Forest Fires. Academic Press, San Diego, CA, USA

Manry D and Knight R. 1986. Lightning density and burning frequency in South-African vegetation. Vegetatio 66(2):67–76

Martín RE, and Means JE. 1982. Fire History and Its Role In Succession. Forest Succession and Stand Development Research in the Northwest. Oregon State University, Forest Research Laboratory, Corvallis, OR, USA

Martínez J, Vega-García C, and Chuvieco E. 2009. Human-caused wildfire risk rating for prevention planning in Spain. J Environ Manage 90(2):1241–52

McCullagh P and Nelder JA. 1989. Generalized Linear Models, 2nd edit. Chapman and Hall, London, UK

McRae R. 1992. Prediction of areas prone to lightning ignition. Internat J Wildland Fire 2(3):123–30

Nieto H, Aguado I, and Chuvieco E. 2006. Estimation of lightning-caused fires occurrence probability in Central Spain. Proc 5th International conference on forest fire research. Coimbra, Portugal. Nov 27–30

Ogilcie CJ. 1989. Lightning fires in Saskatchewan forests. Fire Manage Notes 50(1):31–6

Opsomer JD. 2000. Asymptotic properties of backfitting estimators. J Multivar Anal 73(2):166–79

Pacheco CE, Aguado I, and Nieto H. 2009. Análisis de ocurrencia de incendios forestales causados por rayo en la España peninsular. Geofocus 9:232–49.

Podur J, Martell DL, and Csillag F. 2003. Spatial patterns of lightning-caused forest fires in Ontario, 1976–1998. Ecol Model 164(1):1–20

Pyne SJ, Andrews PL, and Laven RD. 1996. Introduction to Wildland Fire, 2nd edit. John Wiley & Sons, New York, NY, USA

Rivas-Soriano L, de Pablo F, and Diez E. 2001. Cloud-to-ground lightning activity in the Iberian Peninsula: 1992–1994. J Geophysical Research-Atmospheres 106(D11):11891–901

Rivas-Soriano L, de Pablo F, and Tomas C. 2005. Ten-year study of cloud-to-ground lightning activity in the Iberian Peninsula. J Atmos Solar Terr Phys 67(16):1632–9

Roca-Pardinas J, Cadarso-Suarez C, Tahoces PG *et al.* 2009. Selecting variables in non-parametric regression models for binary response. An application to the computerized detection of breast cancer. Stat Med 28(2):240–59

Rorig ML and Ferguson SA. 1999. Characteristics of lightning and wildland fire ignition in the Pacific Northwest. J Appl Meteorol 38(11):1565–75

Rorig ML and Ferguson SA. 2002. The 2000 fire season: Lightning-caused fires. J Appl Meteorol 41(7):786–91

Saveland JM and Neuenschwander LF. 1990. A signal-detection framework to evaluate models of tree mortality following fire damage. For Sci 36(1):66–76

Spanish Ministry of the Environment. 2003. Mapa Forestal de España, escala 1:50.000 (MFE50) de la provincia de León. Organismo Autónomo Parques Nacionales, Madrid, Spain

Swets JA and Pickett RM. 1982. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory, 1st edit. Academic Press, New York, NY, USA

Vankat JL. 1985. General patterns of lightning ignitions in Sequoia National Park, California. Proceedings Symposium and Workshop on Wilderness Fire. USDA Forest Service General Technical Report INT-182; Nov. 15/18, 1983; Missoula, MT, USA

Vasconcelos MJP, Silva S, Tomé M, *et al.* 2001. Spatial prediction of fire ignition probabilities: Comparing logistic regression and neural networks. Photogramm Eng Remote Sens 67(1):73–81

Vázquez A and Moreno JM. 1993. Sensitivity of fire occurrence to meteorological variables in Mediterranean and Atlantic areas of Spain. Land Urb Plan 24:129–42

Vázquez A and Moreno JM. 1998. Patterns of lightning-, and people-caused fires in peninsular Spain. Internat J Wildland Fire 8(2):103–15

Vélez R. 1990. Mediterranean forest fires: A regional perspective. Unasylva 162:3–9

Vilar L, Nieto H, and Martin MP. 2010. Integration of lightning- and human-caused wildfire occurrence models. Hum Ecol Risk Assess 16(2):340–64

Wierzchowski J, Heathcott M, and Flannigan MD. 2002. Lightning and lightning fire, Central Cordillera, Canada. Internat J Wildland Fire 11(1):41–51

Wotton BM and Martell DL. 2005. A lightning fire occurrence model for Ontario. Can J For Res 35(6):1389–401