Spatial Prediction of Fire Ignition Probabilities: Comparing Logistic Regression and Neural Networks

Maria José Perestrello de Vasconcelos, Sara Silva, Margarida Tomé, Margarida Alvim, and José Miguel Cardoso Pereira

Abstract

The objective of this work was to develop and validate models to predict spatially distributed probabilities of ignition of wildland fires in central Portugal. The models were constructed by exploring relationships between ignition location/cause and values of geographical and environmental variables using logistic regression and neural networks. The conclusions are that (1) the spatial patterns of fire ignition identified can be used for prediction, (2) the spatial patterns are different for the different causes, (3) the logistic models and the neural networks both reveal acceptable levels of predictive ability but the neural networks present better accuracy and robustness, (4) the maps produced by the two methods are similar, and (5) the information contained in the spatial position of ignition events can be used to gain predictive capability over an important phenomenon that is difficult to characterize and, for that reason, has not been included in most of the currently used fire danger estimation systems.

Introduction

Fire is the most severe threat to Portuguese forests. Wildfires are the cause of major losses in forest production, also affecting the quality of life in rural areas (Nogueira, 1990; Silva, 1990). It is widely accepted that the most cost-effective means for reducing fire incidence is by forest fire prevention (as opposed to fire suppression). In this context, several efforts have been made to develop fire danger rating systems that can predict zones of potentially severe fire behavior, due to a combination of vegetation and topographic situations and/or particular weather conditions (Deeming *et al.*, 1978; Stocks, 1989). However, fire behavior is only part of the problem, for there will be no fire without ignition. A complete fire danger estimation system needs to address the likelihood of ignition, which is determined by the same environmental conditions that favor fire spread and also by human activity.

Human activity is almost the only cause of fire ignition in Portugal, as in all countries of southern Europe (Commission Européene, 1996), but it is difficult to portray, quantify, and translate into procedural knowledge. Thus, human-induced ignition risk has been excluded from the fire danger models developed for this region. In our study, we show that, by analyzing historical data on fire ignition point locations, we can gain the necessary predictive capability, making it possible to quantify ignition probability in space. The analysis is performed using inductive approaches in a raster geographic information system (GIS), and it explores the information contained in the spatial attributes of the phenomenon.

The raster GIS database used in the study contains a layer with the location of ignition events and a set of layers corresponding to potentially explanatory variables. This data set is analyzed using genetic neural networks and logistic regression. Logistic regression is appropriate because we are trying to model a binary event (occurrence or absence of ignition) using multiple independent variables, and it has been used successfully in similar studies, such as habitat analysis (Pereira and Itami, 1991), archeological studies (Kvamme, 1985), and fire danger prediction (Chou *et al.*, 1993; Vasconcelos *et al.*, 1995). Here, we also use neural networks to test whether non-linear, non-parametric methods can improve upon the results obtained with traditional statistical methods.

The Study Area and the Geographic Database

The study area is located in central Portugal and corresponds to the five municipalities shown in Figure 1: Oliveira do Hospital, Arganil, Gois, Tábua, and Pampilhosa da Serra. Data on fire ignition locations are collected by several Forest Service field teams, where each team is responsible for a given set of municipalities. The data are not exhaustive and correspond to a sample of all the fires that occur in each fire season. Thus, it is important to consider the sampling strategy used by the Forest Service.

Data collection is oriented towards the maximization of the number of fires studied. Therefore, the field teams are more active in areas with a higher concentration of fire events. Although this is obviously far from an ideal random situation, it does reflect the spatial distribution of fire events because there are more ignition points sampled in areas with more fires and less in areas with fewer events. The ignition location data analyzed in this study, shown in Figure 2, were collected by a single field team during four fire seasons (June through September of years 1992 to 1995).

M.J.P. de Vasconcelos was with the Centro Nacional de Informação Geográfica, Lisboa, Portugal when this paper was prepared. She is currently at the Centro de Cartografia, Instituto de Investigação Científica Tropical, Travessa Conde da Ribeira, n° 9, 1300 Lisboa, Portugal (ccart@iict.pt).

S. Silva is with the Centro Nacional de Informação Geográfica, Tagus Park, Núcleo Central 301, 2780-920 Porto Salvo-Oeiras, Portugal.

M. Tomé, M. Alvim, and J.M.C. Pereira are with DEF, Instituto Superior de Agronomia, Tapada da Ajuda, 1300 Lisboa, Portugal

Photogrammetric Engineering & Remote Sensing
Vol. 67, No. 1, January 2001, pp. 73–81.

0099-1112/01/6701–73$3.00/0

Figure 1. The study area corresponds to five municipalities in Portugal: Tábua, Oliveira do Hospital, Arganil, Góis, and Pampilhosa da Serra.

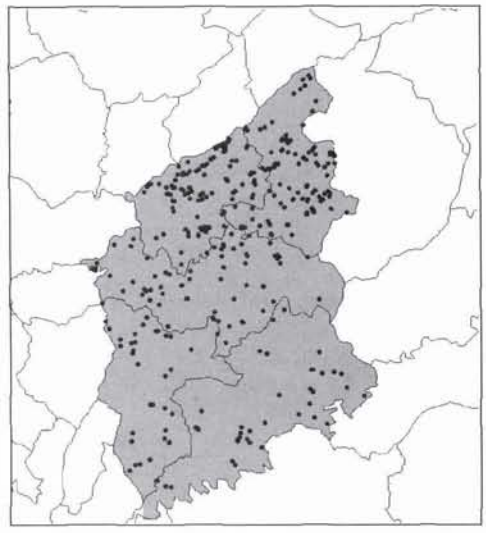

Figure 2. Distribution of ignition points in the study area.

Field data consist of records on fire ignition location and respective cause (arson, negligence, other) for a sample of all the fires that occurred in the area of concern. The locations of ignition points are marked on 1:25,000-scale topographic maps, and each point is associated with a field form where other information such as time of ignition, area burned, and land use are recorded. These point location data are digitized, converted to a point layer, and rasterized by cause for the analysis.

All base themes included in the GIS are obtained from 1:25,000-scale maps in a Gauss-Kruger projection and a plane coordinate system, and rasterized with a 25-m resolution. The study area has a total area of 142,700 ha. The geographical database includes raster base maps such as elevation, road network, and landuse; derived raster maps such as slope, distance to roads, and distance to agricultural fields; and vector layers for display and interpretation of results. Four layers with ignition locations were constructed based on the field data: (1) all causes, (2) arson, (3) negligence, and (4) other causes.

Methods

We used two different methods to obtain predictive models of ignition: genetic neural networks and logistic regression. Both methods were applied to a data set obtained by the following steps:

- Compilation of the information contained in the Forest Service field forms and construction of an attribute table for the ignition points, including information on date of occurrence, place, county, geographic coordinates, cause of ignition, land use, and area burned. The total number of ignition points is 405.
- Construction of the raster geographic database, including a thematic map layer for each of the variables considered and shown in Table 1, and four layers of ignition point locations as listed above.
- Extraction of a random sample of the environmental background, i.e., where no fires were ignited between 1992 and 1995 in the study area. The non-ignition point sample (n = 981) is larger than the ignition sample because the great majority of the database is the environmental background and it is expected that its variability is larger than that found in the site sample. The sampling process for non-ignition points was systematic with a random origin, and the distance between points is 11,750 m.
- Construction of four two-way tables, one for each cause of ignition and one for all ignition causes pooled together. There is a record for each sampled point (ignition and no ignition), and the fields correspond to the values of the mapped variables in the same position.

After a preliminary analysis with both methods, it became clear that the set of ignition points corresponding to "other causes" should be excluded from the analysis. This class has a small number of observations, a large proportion of which are of unknown cause, which most likely include observations of the other two causes. The set of points corresponding to "other causes" was therefore discarded from the analysis and the final ignition point sample size was 366 observations, with three separate data sets: pooled causes (with the 366 observations of ignition by arson and by negligence), arson (242 observations), and negligence (124 observations).

For validation purposes, each of the three initial data sets was randomly partitioned into two independent data sets for model training and validation, respectively. Neural networks require an additional testing set to ensure that generalization ability does not decrease. A classification accuracy assessment is performed with the validation sets, and the maps generated by applying the models in the GIS were compared.

Logistic Regression

Logistic regression is one of the most popular mathematical modeling approaches that can be used to describe the relationship of several variables to a dichotomous dependent variable (Hosmer and Lemeshow, 1989; Kleinbaum, 1994). As a first step, an analysis of the relationship of each individual independent variable with the response variable is performed to get an idea of the relative importance of each variable in explaining ignition. In the analysis, each categorical variable with k levels was substituted by a set of $k - 1$ dummy variables. The final multivariate model is obtained using stepwise regression on

TABLE 1. RASTER LAYERS OF THE GEOGRAPHIC DATABASE

Category	Layer	Information type
Topography	Elevation	Quantitative (m)—Base map
	Slope	Quantitative (%)—Derived map
	Aspect	Qualitative (8 classes)—Derived map
Land Use/Cover	Use	Qualitative (6 classes)—Base map
	Burned areas 92–95	Qualitative (2 classes)—Base map
Man-made Features	Road network	Qualitative (2 classes)—Base map
	Urban areas	Qualitative (2 classes)—Derived map
Spatial Relationships	Distance to roads (DistRoad)	Quantitative (m)—Derived map
	Distance to urban areas (DistUrb)	Quantitative (m)—Derived map
	Distance to agriculture (DistAg)	Quantitative (m)—Derived map
	Distance to shrublands (DistShrub)	Quantitative (m)—Derived map
	Distance to forest	Quantitative (m)—Derived map
Field sample	Ignition points (all causes)	Qualitative (2 classes)—Base map
	Ignition points arson	Qualitative (2 classes)—Base map
	Ignition points negligence	Qualitative (2 classes)—Base map
	Ignition points other	Qualitative (2 classes)—Base map
	Non-sites	Qualitative (2 classes)—Base map

the training set. In the analysis we use standard tests and statistics for logistic regression, i.e., the likelihood ratio test, Wald's test, the odds ratio, and concordance analysis (Hosmer and Lemeshow, 1989; Kleibaum, 1994).

The likelihood ratio test assesses the overall significance of the model, while the Wald's test refers to the significance of an individual variable in the presence of the other variables included in the model. The odds ratio is evaluated for each independent variable and has a different meaning depending on the type of independent variable, dichotomous or continuous. For a dichotomous variable, it indicates how much more likely (or unlikely) it is for the outcome (ignition) to be present among the individuals with $x = 1$ than among those with $x = 0$. For a continuous variable x, the estimated coefficient for x gives the variation in probability of success given the increase of one unit in the independent variable analyzed; in most cases, a variation of one unit has no environmental significance. For example, an increase of 1m in altitude is not likely to affect ignition; however, an increase of 500 m can represent a significant environmental change and be important for the probability of ignition. Therefore, the odds ratio is usually referred to n units of the independent variable. The units used for calculating the odds ratio of the continuous variables are as follows: altitude—500 m, slope—10 percent, and distances—1000 m. Additionally, and still using the training data set, we construct tables with classification error rates for varying cut-off points, for selection of the optimal cut-off point. Cut-off points are used to convert probability of ignition to dichotomous 0–1 data. Any cells with values below the cut-off are considered as non-ignition sites, while all above become predicted as ignition sites. Two statistics are computed for each cut-off point considered: sensitivity and specificity. Sensitivity is the proportion of true positives that were predicted as events and specificity is the proportion of true negatives that are predicted as non-events. Figure 3 is an example of the calculation of these statistics for varying cut-off points. The optimal cut-off point corresponds to the intersection of the two lines. The optimal cut-off points for all the models are determined in the same manner, both for the logistic regression and for the neural networks.

The validation of logistic models entails the quantification of their predictive ability and an assessment of the possibilities of extrapolation to independent data sets. The reason for considering this type of assessment of model performance is that the model always performs in an optimistic manner on the training data set (Hosmer and Lemeshow, 1989).

Quantification of predictive ability is done by comparing observed with predicted probability of ignition. Because the observed data are dichotomous (presence or absence of ignition), we have to use an artifact to generate observed probabilities: Data from the validation set are grouped into classes defined according to the range of each variable in the multivariate model. The proportion of ignition points falling in each class is considered as the observed probability of ignition. The corresponding values estimated by the model are computed for each class using, for the continuous variables, the central value of the class.

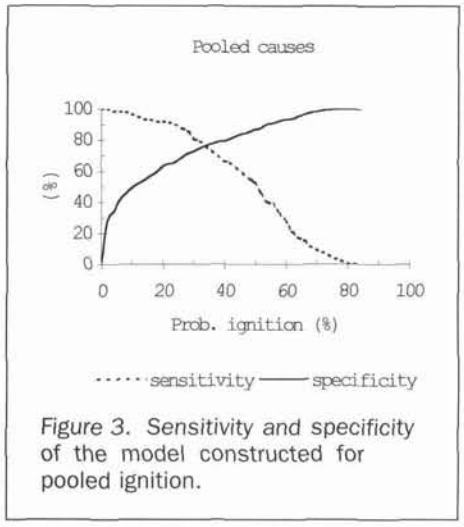

Figure 3. Sensitivity and specificity of the model constructed for pooled ignition.

Observed and predicted probabilities of ignition are graphically compared. Each one of the selected models is used to predict ignition occurrence in the respective data set, using the cut-off points previously found, and the following proportions are computed: (1) proportion of concordance in the points with ignition, (2) proportion of concordance in the points without ignition, and (3) proportion of concordance in the entire data set.

Neural Networks

Artificial neural networks have been used for classification of remotely sensed imagery in several studies (Paola and Schowengerdt, 1997; Jarvis and Stuart, 1996; German and Cahegan, 1996) and coupled with GIS for spatial analysis in various applications (Mann and Benwell, 1996; Kao, 1996; Murnion, 1996a; Murnion, 1996b). For a description of network architectures, algorithms, and applications see Fausett (1994). In this

study, we use a multilayer feedforward network architecture trained with a genetic algorithm.

Multilayer feedforward networks are neural architectures where neurons form several layers, fully connected to each other. Each neuron calculates the weighted sum of the activities coming from the neurons of the previous layer and applies a non-linear differentiable function to the result, usually a sigmoid. The stimuli presented to the neurons of the input layer are propagated this way through the hidden layers until the neurons of the output layer, where the network expresses the result.

Backpropagation (Werbos, 1974; Rumelhart *et al.*, 1986) is the algorithm typically used for training feedforward neural networks in a supervised manner. Using a gradient descent method, it finds the optimal values for the synaptic weights by minimizing an error function based on the difference between the desired response and the actual network output, for each stimulus. Although this procedure allows the training of arbitrarily large networks on highly complex classification problems, and in spite of the many improvements developed so far, it still cannot overcome a major handicap: the existence of local minima in the error landscape. This problem increases dramatically when training networks have more than one hidden layer.

A genetic algorithm (Holland, 1975) is a global optimization procedure inspired in natural evolution, where an initial set (population) of possible solutions to the problem (individuals) are encoded in the form of strings (chromosomes) and evolved by means of genetic operators, the most typical being crossover and mutation. Standard crossover takes two parents, splits their chromosomes at a random location, and produces two children whose chromosomes are composed of one part from each parent (`Parent1: 00101011; Parent2: 10010100; Child1: 00101100; Child2: 10010011`). Mutation takes one parent and produces a single child by randomly altering a small piece of the inherited chromosome (`Parent: 01101010; Child: 01111010`). The selection of individuals for reproduction is based on their fitness values, given by an evaluation function. Because better solutions receive larger values and, consequently, higher probabilities of being selected, the population tends to improve in quality. Hopefully, after a number of generations, the algorithm finds the optimal solution to the problem. For an introduction to genetic algorithms, see Goldberg (1989); for an overview of applications, see Mitchell (1996).

The first successful attempt to use genetic algorithms as learning rules for feedforward neural networks is due to Montana and Davis (1989). They present a simple method for expressing a neural network as a chromosome, and describe several new genetic operators specifically designed for this particular problem. Unlike backpropagation, genetic algorithms seldom get trapped in local minima, thanks to the high stochastic level introduced by the interaction of multiple solution trajectories.

Data Sets, Initialization, and Network Adjustment

To train a neural network and assess its performance in a reliable manner, the original data set is partitioned into three disjoint sets, for training, testing, and validation, each aimed at a different purpose.

The training set contains the only data points actually used for structuring the network. When using backpropagation, the algorithm must be interrupted before the network memorizes the train data points too well and loses its generalization ability. This phenomenon, known as overfitting, can be avoided by monitoring the network error on the test set, an independent set of data points, and stopping the algorithm when it starts increasing steadily, a technique suggested by Hecht-Nielsen (1990). Although the testing set is never used for training, it implicitly directs the training procedure, hence the need for yet another independent data set on which to perform unbiased accuracy measures—the validation set.

When using a genetic algorithm as the learning rule, the three data sets serve the same purpose. The selection of individuals for reproduction is based solely on the fitness values measured on the training set, but the evolution is halted when the fitness values measured on the testing set start decreasing, meaning the population is becoming too specialized. The validation set is then used to determine the unbiased fitness values.

The validation set contains 10 percent of the total number of data points; the testing set contains 20 percent of the remaining 90 percent; and the training set contains all the remaining data points. Very often, the number of data points available for each different class is far from being equivalent. Care must be taken to maintain similar proportions in both the validation and test sets. On the other hand, when using backpropagation, the training set must contain approximately the same number of data points for each class, to avoid a biased training, so that the less well represented classes may see their data points repeated. The genetic algorithm approach does not require this last procedure.

Three different partitions were formed randomly. Each experiment reported here was run once in each partition, and the results presented were determined as the average achieved on the three runs, except the maps, which were formed by only one network each, chosen at random.

In this work, all the neural networks used a bipolar activation function (the hyperbolic tangent) and contained two hidden layers, with nine and three neurons each in the cases of non-ignition versus arson ignition and non-ignition versus negligent ignition, and twelve and four neurons each in the case of pooled causes. The number of input neurons was determined by the number of variables used in each case, and the only output neuron was trained to obtain 1 when there was ignition and -1 otherwise. The number of hidden layers and their constitution was determined by a trial and error process based on the observation of the population's fitness curves measured on the training and testing sets.

In the genetic algorithm used as the learning rule, the population had a fixed dimension of 50 individuals. Smaller populations contained insufficient genetic variability, which resulted in premature convergence into sub-optimal solutions, and bigger ones were too expensive computationally. Each chromosome consisted of an ordered list of the synaptic weights, initialized as described in Montana and Davis (1989). The evaluation function was the opposite of the network error measured on the train set; the lower the error, the higher the fitness value. The selection for reproduction used the rank selection described by Montana and Davis (1989), but instead of the traditional Roulette Wheel (Goldberg, 1989), the ranking was followed by the more efficient Stochastic Universal Sampling (Baker, 1987). The genetic operators used were CROSSOVER-NODES and MUTATE-NODES (Montana and Davis, 1989), variations of standard crossover and mutation, with variable probabilities of occurrence (Davis, 1989), but dispensing with the initialization procedure and being set at 0.5 each. The algorithm always guaranteed survival of the best individual created so far, while replacing the others by the offspring; it was allowed to run for 250 generations, after which the population's fitness tended to stabilize.

Results

Logistic Regression

Table 2 shows the results of the analysis of relevance of each independent variable for the data sets corresponding to ignition caused by arson and by negligence. As can be seen, most of the single-variable models are significant in explaining the

TABLE 2. RESULTS OF THE SINGLE VARIABLE ANALYSIS. THE BOLD VALUES CORRESPOND TO SIGNIFICANT VARIABLES (p = 0.05)

		Arson ignition vs non ignition				Negligence ignition vs non ignition			
Original variables	Dummy variables	$Pr > \chi^2$ (−2LogL)	$Pr > \chi^2$ (Wald)	Odds ratio	Concord. (%)	$Pr > \chi^2$ (−2LogL)	$Pr > \chi^2$ (Wald)	Odds ratio	Concord. (%)
Elevation		**0.0001**	**0.0001**	0.996	67.0	**0.0062**	**0.0086**	0.998	57.8
Slope		**0.0001**	**0.0001**	0.941	61.4	0.3166	0.3126	0.986	50.4
DistRoad		**0.0001**	**0.0001**	0.999	78.9	**0.0001**	**0.0001**	1.000	76.8
DistUrb		**0.0001**	**0.0001**	0.999	62.4	**0.0001**	**0.0001**	0.999	66.7
DistAg		**0.0001**	**0.0001**	0.997	60.1	**0.0001**	**0.0006**	0.998	60.3
DistShrub		**0.0001**	**0.0001**	0.998	51.8	0.0533	0.0744	0.999	47.6
Land use	urban		0.1023	**7.765**			0.0644	**5.824**	
	agricult.	**0.0001**	**0.0025**	**22.195**	29.1	**0.0260**	**0.0164**	**6.133**	31.8
	forest		**0.0046**	**17.668**			0.0646	3.869	
Aspect	north		0.5265	2.480			**0.0324**	**15.500**	
	northeast		**0.0231**	0.401			0.6490	1.254	
	east		0.0747	0.511			0.8616	0.912	
	southeast		0.0734	0.522			0.563	0.714	
	south	0.2859	0.0531	0.512	49.6	0.0151	0.2268	0.505	55.4
	southwest		0.4162	0.763			0.2040	1.788	
	west		0.0703	0.542			0.1260	0.404	
	northwest		0.2903	0.719			0.9764	1.014	

TABLE 3. SUMMARY OF THE RESULTS OF THE STEPWISE REGRESSION FOR POOLED IGNITION. PROPORTION OF CONCORDANCE WITH THE TRAINING SET

Selected variables		$Pr > \chi^2$	Odds	Concord.
DistRoad	−0.00054	0.0001	0.583	
DistUrb	−0.00082	0.0006	0.440	
DistAg	−0.00239	0.0001	0.092	84.1
DistShrub	−0.00318	0.0001	0.0042	
Southwest	0.4991	0.0496	1.647	
intercept	1.5475			

TABLE 4. SUMMARY OF THE RESULTS OF THE STEPWISE REGRESSION FOR IGNITION DUE TO ARSON. PROPORTION OF CONCORDANCE WITH THE TRAINING SET

Selected variables		$Pr > \chi^2$	Odds	Concord.
Elevation	−0.00218	0.0028	0.336	
Slope	−0.03510	0.0130	0.703	
DistRoad	−0.00060	0.0001	0.549	86.1
DistAg	−0.00280	0.0001	0.061	
DistShrub	−0.00375	0.0001	0.024	
intercept	2.14790			

TABLE 5. SUMMARY OF THE RESULTS OF THE STEPWISE REGRESSION FOR IGNITION DUE TO NEGLIGENCE. PROPORTION OF CONCORDANCE WITH THE TRAINING SET

Selected variables		$Pr > \chi^2$	Odds	Concord.
DistRoad	−0.00042	0.0001	0.657	
DistUrb	−0.00148	0.0001	0.227	
DistShrub	−0.00238	0.0150	0.093	82.4
North	3.40370	0.0214	30.074	
Southwest	0.98440	0.0032	2.676	
intercept	0.06780			

probability of ignition caused by arson or negligence. Aspect is not significant in the data set corresponding to arson ignition, with most modalities non-significant and, jointly with land use, is the less significant variable in the data set corresponding to ignition by negligence. Tables 3 through 5 show a summary of the results obtained with the stepwise logistic regression for the three causes considered. The last columns of these tables show the values of concordance with the training set for each of the models.

According to the results shown, we have the following models in linear form ($g(x)$ is the logit):

(1) All data with ignition causes pooled $g(x) = 1.5475 - 0.00054\ droad - 0.00082\ durb - 0.00239\ dagri - 0.00318\ dshrub + 0.49\ southwest$ where *droad* is the distance to roads, *durb* is the distance to urban areas, *dagri* is the distance to agriculture, *dshrub* is the distance to shrublands (all in meters), and *southwest* is a dummy variable corresponding to one class of the eight directions of the compass of the categorical variable *aspect*.

(2) Ignition caused by arson $g(x) = 2.1479 - 0.00218\ altitude - 0.0351\ slope - 0.0006\ droad - 0.0028\ dagri - 0.00375\ dshrub$ where *altitude* is in meters, *slope* is in percent, and the remaining variables have the same meaning and units as above.

(3) Ignition caused by negligence $g(x) = 0.1705 - 0.00042\ droad - 0.0014\ durb - 0.00210\ dshrub + 3.4\ north + 0.98\ southwest$ where the variable *north* is another dummy variable corresponding to one class of the categorical variable *aspect*.

The probability of occurrence of ignition $\pi(x) = P(y = 0|x)$ can be easily obtained from $g(x)$ as

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}.$$

Cut-off points for each one of the models were selected as shown in Figure 3. The following cut-off points were selected for each one of the data sets: 0.34 for the pooled data, 0.26 for the data set corresponding to the ignition caused by arson, and 0.14 for the data set corresponding to ignition by negligence.

As explained above, observed and predicted probabilities of ignition were graphically compared, and the results are presented in Figure 4. The comparisons show that the observed versus predicted probabilities approximate a 1:1 straight line, thus indicating that the three models have good predictive abilities.

Table 6 shows the concordance levels obtained with the logistic regression for the training and the validation set. Even though the concordances obtained with the training set are of about 80 to 85 percent, those values drop when the model is tested with the independent validation set. In Table 7 we present the results of classification accuracy assessment. This table shows that the commission errors are high in the case of ignition when compared to omission errors, and that omission

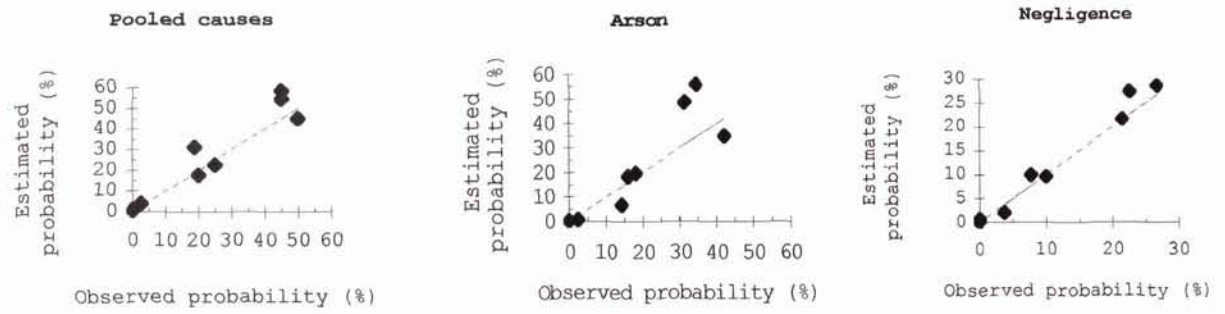

Figure 4. Observed versus estimated probabilities with the models constructed.

TABLE 6. CONCORDANCE LEVELS FOR THE LOGISTIC REGRESSION AND THE NEURAL NETWORKS

Concordance (%)	Training	Testing	Validation
		Logistic Model	
Pooled	84.1	—	73.9
Arson	86.1	—	74.9
Negligence	76.8	—	71.5
		Neural Network	
Pooled	77.2	77.0	77.0
Arson	80.3	79.2	77.9
Negligence	77.9	75.9	85.5

is slightly higher in the case of no ignition. The application of the models to the mapped variables generates the ignition probability maps shown in Figures 5a through 7a with corresponding frequencies shown in Table 8.

Neural Networks

Table 6 summarizes the results of the training, testing, and validation of the neural network. The variables used to fit each of the three models are the same as those selected in the logistic regression, and the outputs of the neural nets were rescaled from the range [−1 1] to the range [0 1] for comparison of results with the output of the logistic models. The concordance levels obtained with the training set are similar to those obtained with the testing and validation sets (77 to 80 percent). There is an exception in the case of negligence where the validation set shows a concordance level of 85.5 percent. The cut-off points were determined using the same procedure as that discussed for the logistic regression and are as follows: pooled causes -0.43, arson -0.39, and negligence -0.17.

Table 7 shows high commission errors for the ignition points, with omission errors significantly lower. In the neural net, both types of error are higher for the ignition points. The maps generated by applying the trained neural networks are shown in Figures 5b through 7b, and the corresponding frequencies are shown in Table 8.

Discussion and Conclusions

In this study, we verify that there are different spatial patterns of ignition for the causes considered and that the models constructed have a good fit and acceptable classification accuracies. Both the logistic regression and the neural networks produce models with good levels of concordance when compared to other models of the same type (e.g., Pereira and Itami, 1991; Chou *et al.*, 1993). Additionally, both methods achieve acceptable classification error rates for given cut-off points.

Of the variables shown as significant in Table 2, *Land use* is systematically not included in the final model. This may be related to the fact that its use results in a larger reduction of the degrees of freedom than alternative variables. In fact, when the variables distance to agriculture and distance to shrublands (quantitative variables) are excluded, the variable *Land use* is selected in the stepwise procedure.

The variables selected are not the same in the three sets considered, but there is always a strong contribution of the variables related to topography, distance to man-made features, and distance to two specific types of land cover. The linear models obtained with logistic regression are consistent with prior expectations regarding the variations of the explanatory environmental variables. The magnitude of the variation introduced in the models by the distance variables (*droad*, *durb*, *dshrub*, and *dagri*) can be directly intercompared because the scale is the same for all of them. It is interesting to note that the distance variables included in the models have a negative sign, indicating that the probability of ignition decreases with the increase of distance to the specified features. It is to be expected that the probability of ignition is higher closer to shrub patches due to their comparatively high flammability. Concurrently, it is more likely for ignition to occur close to areas that can function as ignition sources (roads, urban, and agricultural areas).

TABLE 7. FINAL CLASSIFICATION ERRORS

	Pooled				Arson				Negligence			
	Logistic		Neural Net		Logistic		Neural Net		Logistic		Neural Net	
	no ignition	ignition	no ignition	ignition	no ignition	ignition	no ignition	ignition	no ignition	ignition	no ignition	ignition
Producer's accuracy	72.4	77.8	77.6	75.7	74.0	78.8	81.6	62.5	70.6	78.6	87.8	66.7
User's accuracy	89.7	51.3	89.4	56.0	93.4	42.5	89.9	45.5	96.2	25.6	95.6	40.0
Omission error	27.5	22.1	22.4	24.3	25.9	21.3	18.4	37.5	29.4	21.4	12.2	33.3
Commission error	10.2	48.6	10.6	44.0	6.6	57.4	10.1	54.5	3.8	74.4	4.4	60.0

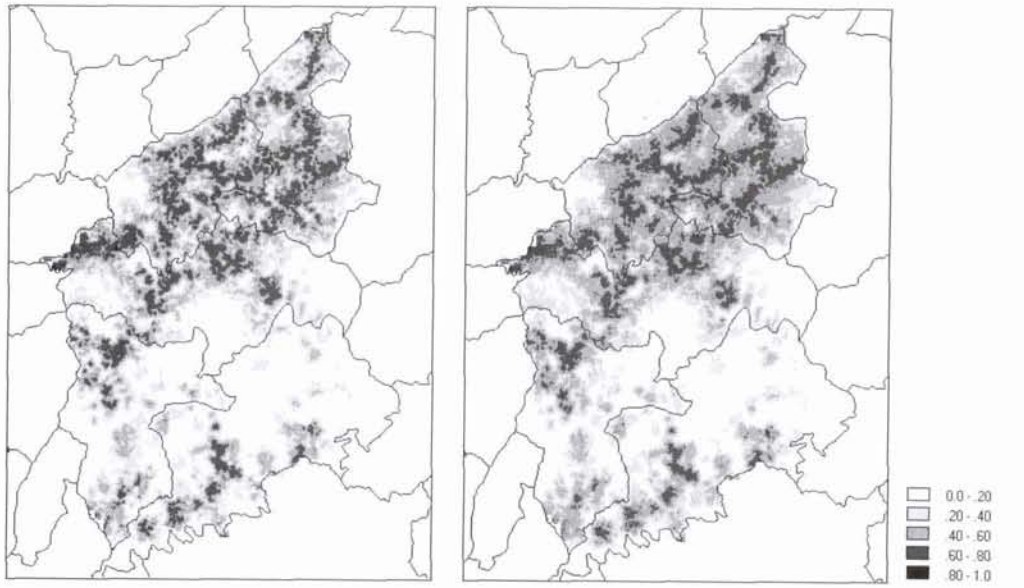

Figure 5. Maps of ignition probability for pooled causes for the entire study area. (a) Results of the application of the logistic model. (b) Results of the application of the neural network.

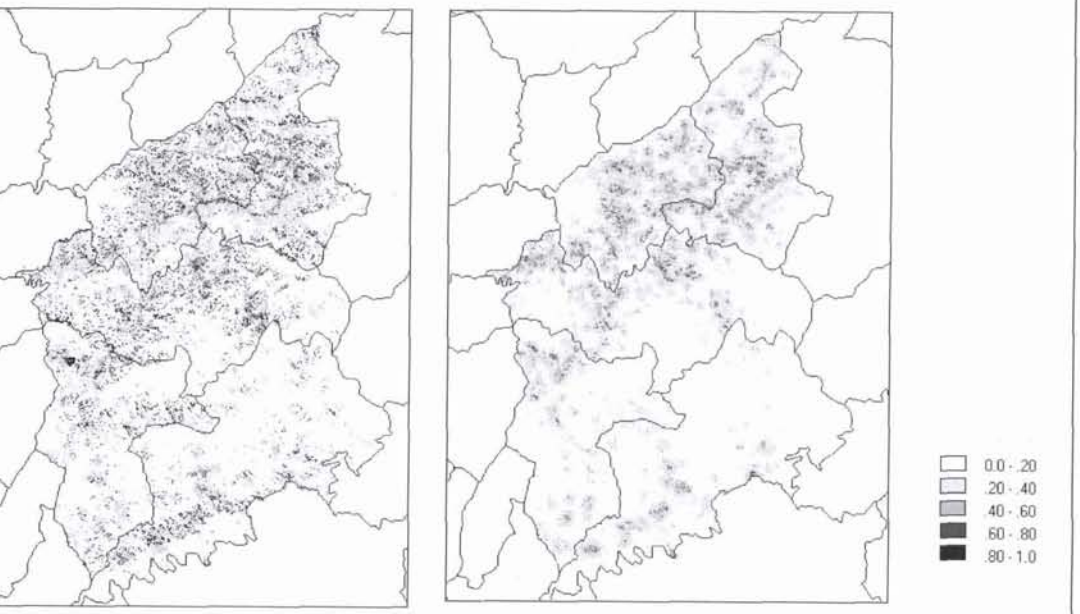

Figure 6. Maps of probability of ignition due to arson for the entire study area. (a) Results of the application of the logistic model. (b) Results of the application of the neural network.

In the case of arson, the variables associated with topography are selected before the distance variables and thus have higher importance in explaining ignition patterns. One possible interpretation is that, because fires spread mainly uphill, they are set at lower altitudes to increase the likelihood of larger burned areas. Another possibility is that arsonists are moved by a known economic motivation. Because burned wood does not loose its technological properties if used in a short period after burning, setting fires promotes the flooding of local markets with timber at lower prices. Arsonists want plantations to burn and these prevail on not too steep slopes. Another important difference between the model found for arson and those for negligence and pooled causes is that the variable *aspect* is not included in the model for arson.

The selection and positive influence of the variable *aspect-southwest* in the cases of pooled ignition and negligence can be explained. In fact, fires occurring on terrain facing southwest are more likely to spread due to higher solar incidence and comparatively higher fuel dryness. The selection of the variable *aspect-north* in the case of negligence is

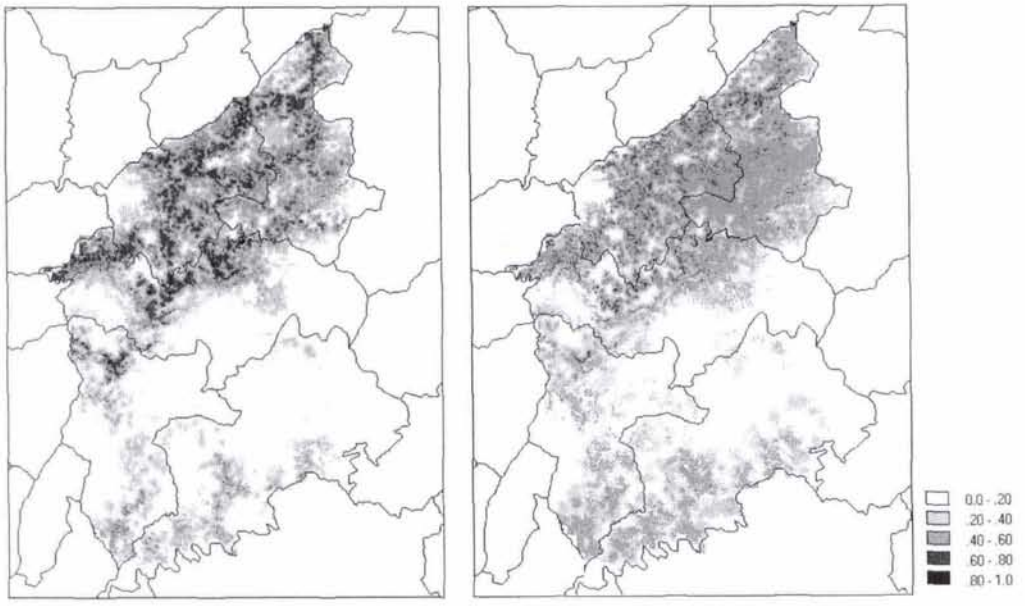

Figure 7. Maps of probability of ignition due to negligence for the entire study area. (a) Results of the application of the logistic model. (b) Results of the application of the neural network.

TABLE 8. FREQUENCY DISTRIBUTION OF PREDICTED IGNITION PROBABILITIES CORRESPONDING TO THE MAPS SHOWN IN FIGURES 5 THROUGH 7

Ignition Probability Class	Pooled		Arson		Negligence	
	Logistic	Neural Net	Logistic	Neural Net	Logistic	Neural Net
0–10%	0.7605	0.7553	0.8061	0.7935	0.8097	0.7935
11–20%	0.0473	0.0426	0.0452	0.0295	0.0782	0.0295
21–30%	0.0389	0.0359	0.0357	0.0346	0.0466	0.0346
31–40%	0.0378	0.0368	0.0333	0.0481	0.0248	0.0481
41–50%	0.0399	0.0437	0.0308	0.0547	0.0110	0.0547
51–60%	0.0374	0.0483	0.0257	0.0321	0.0054	0.0321
61–70%	0.0276	0.0354	0.0171	0.0073	0.0042	0.0073
71–80%	0.0106	0.0020	0.0059	0.0001	0.0045	
81–90%	0.0001		0.0001		0.0092	
91–100%					0.0064	

difficult to explain. However, it may be related to the fact that in the study region the frequency of *aspect-north* is the lowest of the eight compass directions possible, and that northern exposures have a cover of forest and shrubland fuels with very little agriculture.

The logistic models have better concordance than do the neural networks in the training phase; however, the neural networks show higher overall classification accuracies. From Table 6 it is apparent that the neural networks have stronger generalization ability, with consistent concordance levels in the training, testing, and validation sets. This is possible due to the use of the testing set during the training process, which guarantees that the network does not overfit to the training set. The logistic models, by having a stronger fit to the training set, loose generalization ability and have a markedly lower classification accuracy with the validation set. It is worth noting that in the case of ignition, which is the case we are most concerned about, the neural nets always have lower commission errors than the logistic regression, but they also consistently display slightly higher omission errors.

Table 7 shows that from a user's perspective it would be safer to use the maps resulting from the neural networks than those resulting from the logistic models. The results of the neural networks also meet the expectation that the models resulting from the separation of causes have better classification accuracies than the model for pooled causes. However, it is worthwhile noting that the diagnostic tests available for the logistic regression allow for more interpretation than the neural networks which provide no clues about the internal importance of each variable, and whose weights after training are not easily interpretable.

The maps resulting from the two model types have similar spatial structure, but have different distributions of probability values, especially in the case of ignition due to negligence. The maximum probabilities are always obtained with the logistic models. Concurrently, and with the exception of the case negligence, there are more cells with likelihood of ignition (intermediate probability values) on the maps resulting from the neural networks than on those resulting from the logistic models.

The maps resulting from the logistic models have smaller patches with higher internal heterogeneity, whereas the maps resulting from the neural networks have larger, more homogeneous ignition zones. In the case of ignition due to negligence,

the results from the logistic models present an unlikely spatial structure with small patches of very high ignition probability scattered throughout the entire study area. This is controlled by the variable *aspect*, with cells due north presenting very high probability of ignition. As can be seen in Table 2, *aspect-north* is a dichotomous variable with very strong explaining power. Although the same variable is used in the neural networks, the results do not show this effect and produce a smoother spatial structure. This fact may be related to the known robustness of neural networks relative to inconsistent, contradictory, and incomplete data sets.

The spatial patterns of ignition found are strongly associated with accessibility, and we do not know to what extent this portrays the real distribution of the phenomenon because it can also be a consequence of the sampling procedure. Fire starts are more likely to occur in areas of more intense human activity; however, it is also more likely that the ignition points close to the roads get investigated, in comparison to less accessible fire starts.

The feasibility of mapping fire ignition risk due to human causes is demonstrated in this study, and it can be an important step towards increased effectiveness in fire prevention and planning. The separation of ignition risk by cause allows the identification of areas with specific types of problems, and thus can help delineate the adequate prevention actions for different zones. For example, it can support the decision of allocating an educational campaign along the roads where negligence is the major source of concern or vigilance where arsonists are to be expected.

Many fire management decisions are based exclusively on the risk of fire propagation. However, it is important to differentiate priorities among several zones of equivalent propagation risk. The ability to quantify ignition risk can be the key to a more informed allocation of fire prevention resources. Additionally, ignition risk maps, when integrated with information on fire propagation risk, can support the optimization of silvicultural practices in specific areas.

Acknowledgments

Research dealing with the topics presented in this paper was funded by projects STORMS (AIR3-CT94-2392, DGXII) and MODERIS (PEAM/FF/475, CNEFF/JNICT). We are thankful to the Portuguese Forest Service (Direcção Geral das Florestas) for providing the data on ignition points and collaborating in this study, particularly to Eng. Manuela Baptista, Eng. Rui Natário, and Eng. Josefa Carvalho. Teresa Cardoso and João Melo compiled the data on ignition points and prepared the geographic database. Cristina Machado helped in the preparation of data for analysis and in the early stages of the analysis.

References

Baker, J.E., 1987. Reducing bias and inefficiency in the selection algorithm, *Genetic Algorithms and Their Applications: Proceedings of the Second International Conference on Genetic Algorithms* (J.J. Grefenstette, editor), 28–31 July, MIT, Cambridge, Massachusetts, Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 14–21.

Chou, Y.-H., R.A. Minnich, and R.A. Chase, 1993. Mapping probability of fire occurrence in San Jacinto Mountains, California USA, *Environmental Management*, 17(1):129–140.

Commission Europeénne, 1996. Les feux de forêt dans le sud le l'Union Européenne 1989–1993, Office des publications officielles des Communautés Européennes, Luxembourg, 52 p.

Davis, L.D., 1989. Adapting operator probabilities in genetic algorithms, *Proceedings of the Third International Conference on Genetic Algorithms* (J.D. Schaffer, editor), 04–07 June, George Mason University, Fairfax, Virginia, Morgan Kaufmann, San Mateo, California, pp. 61–69.

Deeming, J.E., R.E. Burgan, and J.D. Cohen, 1978. *The National Fire Danger Rating System—1978*, General Technical Report INT-39, USDA Forest Service, Intermountain Research Station, Ogden, Utah, 63 p.

Fausett, L., 1994. *Fundamentals of Neural Networks—Architectures, Algorithms and Applications*, Prentice-Hall, Englewood Cliffs, New Jersey, 461 p.

German, G.W.H., and M.N. Cahegan, 1996. Neural network architectures for the classification of temporal image sequences, *Computers and Geosciences*, 22(9):969–979.

Goldberg, D.E., 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, Massachusetts, 412 p.

Hosmer, D.W., and S. Lemeshow, 1989. *Applied Logistic Regression*. Wiley series in Probability and Mathematical Statistics, John Wiley and Sons, New York, N.Y., 307 p.

Jarvis, C.H., and N. Stuart, 1996. The sensitivity of a neural network for classifying remotely sensed imagery, *Computers and Geosciences*, 22(9):959–967.

Mann, S., and G.L. Benwell, 1996. The integration of ecological, neural and spatial modeling for monitoring and prediction for semi-arid landscapes, *Computers and Geosciences*, 22(9):1003–1012.

Murnion, S.D., 1996a. Comparison of back propagation and binary diamond neural networks in the classification of a Landsat TM image, *Computers and Geosciences*, 22(9):995–1001.

———, 1996b. Spatial analysis using unsupervised neural networks, *Computers and Geosciences*, 22(9):1027–1031.

Nogueira, C.D.S., 1990. A floresta Portuguesa, DGF Informação, 2:18–28.

Paola, J.D., and R.A. Showengerdt, 1997. The effect of neural network structure on a multispectral land-use/land-cover classification, *Photogrammetric Engineering & Remote Sensing*, 63(5):535–544.

Pereira, J.M.C., 1991. GIS based habitat modeling using logistic multiple regression: A study of the Mt. Graham red squirrel, *Photogrammetric Engineering & Remote Sensing*, 57(11):1475–1486.

Silva, J.M., 1990. La gestion forestiére et la silviculture de prevention des espaces forestiers menacés par les incendies au Portugal, *Revue Forestiére Française*, 40(n° spécial):337–345.

Stocks, B.J., B.D. Lawson, M.E. Alexander, C.E. van Wagner, R.S. McAlpine, T.J. Lynham, and D.E. Dube, 1989. The Canadian forest fire danger rating system: An overview, *The Forestry Chronicle*, 65:258–265.

Vasconcelos, Maria J. Perestrello, 1995. Integration of remote sensing and geographic information systems for fire risk management, *Proceedings of the EARSeL Workshop on Remote Sensing and GIS Applications to Forest Fires*, 07–09 September, Alcalá de Henares, Spain, pp. 129–147.

Vasconcelos, Maria J. Perestrello, Mário S. Caetano, and José M.C. Pereira, 1996. Application of GIS in forest fire prevention, *Remote Sensing and Computer Technology for Natural Resource Management* (J. Sramaki, B. Koch, and H. Gyde Lund, editors), Proceedings of the IUFRO XX World Congress, 06–12 August 1995, Tampere, Finland, The University of Joensuu, Research Notes 48, pp. 171–181.

(Received 18 August 1999; accepted 14 December 1999; revised 28 January 2000)