Using model-based geostatistics to predict lightning-caused wildﬁres

C. Ordóñez a,*, A. Saavedra b, J.R. Rodríguez-Pérez c, F. Castedo-Dorado d, E. Covían a

a Polytechnic School of Mieres, University of Oviedo, Campus de Mieres, C/ Gonzalo Gutiérrez Quirós s/n, 33600 Mieres, Spain
b Department of Statistics, University of Vigo, Campus Lagoas-Marcosende, 36310 Vigo, Spain
c Geomatics Engineering Group, University of León, Campus de Ponferrada, Avda. de Astorga s/n, 24400 Ponferrada, Spain
d Department of Engineering and Agricultural Sciences, University of León, ETSIA, Avda. Astorga s/n, 24400 Ponferrada, Spain

ARTICLE INFO

Article history:
Received 11 January 2011
Received in revised form 28 July 2011
Accepted 14 October 2011
Available online 17 November 2011

Keywords:
Generalized linear spatial model
Markov-chain Monte Carlo
Spatial correlation
geostatistics
wildﬁre

ABSTRACT

The probability of ﬁre in a particular area depends on a range of environmental and geographic variables. Fire prevention planning can be assisted by the construction of models to identify the variables that have a signiﬁcant inﬂuence on the occurrence of ﬁres and by building maps showing the spatial probability distribution for ﬁres occurring in speciﬁc geographic areas. We used generalized spatial linear models to predict spatially distributed probabilities for ﬁre occurrence in locations where storms featuring lightning occurred, on the basis of a set of variables related to climatology, orography, vegetation and lighting characteristics, and to assess the relative importance of these variables. A comparison of this model with simple logistic regression models used by other researchers to resolve similar problems demonstrates the importance of bearing in mind spatial correlation between variables.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Lightning is the main natural cause of ignition in woodlands all over the world (Pyne et al., 1996) and is even the main cause of ignition, whether natural or otherwise, in boreal forests (Johnson, 1992). In Spain and other Mediterranean countries ﬁres caused by humans are, by far, the most common kind of ﬁre (Vázquez and Moreno, 1998; Loepe et al., 2011) and this is probably the reason why not much attention has been paid to forest ﬁres caused by lightning. Nonetheless, lightning-caused ﬁres can burn larger areas of forest than human-caused ﬁres because of remoteness and aggregation in time and space (Podur et al., 2003).

The probability of ﬁre depends on a range of environmental and geographic variables. Díaz-Avalos et al. (2001) have indicated that topographic variables such as terrain altitude, gradient and orientation have a bearing on the probability of ignition. The type of vegetation is another factor that affects ignition (Dissing and Verbyla, 2003), Dlamini (2008) used a Bayesian belief network to predict wildﬁre from variables such as land cover, temperature, elevation, rainfall, aspect, slope, among others. Land cover, elevation, and climate (mean annual rainfall and mean annual temperature) were found to be strong predictors of wildﬁre occurrence, while aspect had the least inﬂuence on the wildﬁre occurrence.

Lightning characteristics (frequency of strikes, polarity, multiplicity, etc.) also affect the occurrence of ﬁres in wooded areas (Flannigan and Wotton, 1991). Weather characteristics such as rainfall, relative humidity and temperature are other factors that have a bearing on ignition (Hely et al., 2001). Most of the variables that affect the occurrence of lightning-ignited ﬁres are spatially correlated; consequently, the application of models reﬂecting the spatial component is likely to improve not only interpretation of the inﬂuence of known variables but also the prediction of the probability of a ﬁre. However, most research into wildﬁre prediction relies on models that fail to take into account the spatial nature of the explanatory variables (Martell et al., 1989; Vega-García et al., 1995; Perestrello de Vasconcelos et al., 2001; Wotton and Martell, 2005; Catry et al., 2008; Martínez et al., 2009).

In this article we analyze the applicability of model-based geostatistics to assessing the probability of ﬁres using the generalized linear spatial model (GLSM). The term ‘model-based geostatistics’ was coined by Diggle et al. (1998), who extended this geostatistical approach to situations in which stochastic variation in the data is known to be non-Gaussian. These authors described an approach to geostatistical problems based on formal statistical models and inferential procedures. Other authors that have studied
denote a Gaussian random chain are analytically manageable. By performing a sufficiently large number of Markov chain iterations—called burn-in—we obtain a realization of the equilibrium distribution of interest. Continuing with the iterations, we can obtain a sample that is almost independent of the distribution, provided that the realizations of this sample are separated by a sufficiently large number of Markov chain iterations. The analysis described in this article is based on the work of Christensen (2004) and was performed using the geospatial package that is freely available within the framework of the open-source R statistical system (R Development Core Team, 2010).

2.2. Model fit evaluation: ROC curves

A binary classification system can be used to calculate receiver operating characteristic (ROC) curves and determine the precision of a diagnostic test. ROC curves represent true positive versus false positive ratios for different values of the discrimination threshold (Zhou et al., 2002). By calculating the area under the ROC curve (AUC) associated with a binary classifier, we can compare models and select explanatory variables to be included in the model. This approach was used by Rodríguez-Álvarez et al. (2010), who described a bootstrap-based method for testing for a significant factor effect on the ROC curve. ROC curves can therefore be used to select an optimal model from among several GLMs with a binomial error distribution. The AUC can also be used to calculate the goodness-of-fit of the GLSM compared to similar models without a spatial component (i.e., GLMs).

3. Data description and model formulation

3.1. The study area and the geographic database

The study area, located in the province of León in northwest Spain (Fig. 1), measured 15,590 km². It was divided into cells measuring 3 × 3 km, resulting in a total of 1882 cells. The cell size was chosen on the basis of accuracy in the locations of the ignition and on the assumption that weather and fuel conditions would be homogeneous within a cell.

Data on the ignition locations (approximate UTM coordinates) for the years 2002–2007 were supplied by the Spanish Ministry of the Environment. The 1882 cells into which the province of León was divided were coded for occurrence of lightning-caused ignitions during the study period: 0 for non-occurrence and 1 for occurrence. Only major fires, i.e. those requiring intervention, were considered.

Lightning data and data on intensity, polarity and date of occurrence were provided by the Spanish Meteorology Agency. A total of 78,256 flashes were recorded; fewer than 5% had negative polarity.

Topographic variables such as average slope, altitude and aspect were obtained from a digital elevation model with a pixel resolution of 50 m, obtained from the Spanish.

Fig. 1. Location of the study area: province of León (Spain).
Cartographic Database using the ArcGIS 9.2 spatial analysis module, a computer program developed by ESRI (Environmental System Research Institute). Data on land cover was obtained from the forestry map for 2003 for the province of Leon, supplied by the Spanish Ministry of the Environment. The different types of land cover were categorized in 7 classes.

Daily meteorological data for each cell were derived from raw data collected at 344 weather stations located in Leon and nearby provinces. The data, provided by the Spanish Meteorology Agency, consisted of daily observations of temperature, relative humidity and rainfall. Rainfall and relative humidity for each 3 x 3 km cell were estimated using universal kriging. Temperature for each cell was estimated by cokriging interpolation, with elevation used as a secondary variable given that it is highly correlated with temperature.

Spanish Meteorology Agency data were considered error free since this agency does not provide any additional information concerning the quality of the data. This is obviously a problem because observation errors are propagated in the interpolation process. A possible solution would be to perform a sensitivity analysis (Yang, 2011) assuming that the explanatory variables have associated some kind of uncertainty which can be model using random distribution functions. However, this is not affordable from a computational point of view given the time required to run the algorithm each time.

Kriging interpolation has been carried out in a classic way but in future research it will be done using a hierarchical spatial model (Banerjee et al., 2004). In this model, the spatial error component is added to the media, leaving in the covariance matrix the residual error (nugget).

3.2. Model formulation

The response variable, namely, the occurrence of fire, follows a binomial distribution with a spatial dependence structure that can be modeled using a GLSM. According to this model, it is assumed that conditionally on a stationary Gaussian process, \( S(x) \), the response variables \( Y_l, l = 1, \ldots, n \), are modeled as independent binomial variables. The parameters involved in this binary GLSM are \( \theta = \{ \alpha, \phi, \tau^2, \beta \} \), with \( \beta = (\beta_0, \ldots, \beta_p) \), where \( \beta_0 \) is the intercept and \( \beta_1, \ldots, \beta_p \) are the regression coefficient corresponding to the \( i \) variables or factors.

It is assumed that \( S(\cdot) \) is a stationary Gaussian process that determines the spatial variation of the probability of a fire occurring in a specific location \( x \), in such a way that \( P(X) = E[S(X)|S(x)] \). A logit function was used as the link function, as this kind of function is typical in binary GLSMs, as it establishes a relationship between \( S(x) \) and the conditional mean \( P(X) \); thus:

\[
g(P(X)) = \ln\left( \frac{P(X)}{1-P(X)} \right) = S(x) \tag{3}
\]

To infer the parameters of the model we implemented a method based on MCMC. Within the Bayesian framework it can be considered that the vector of parameters, \( \theta \), is conditionally independent of \( Y \) given \( S \). Considering a realization \( y = (y_1, \ldots, y_n)^T \) of \( (Y_1, \ldots, Y_n)^T \), the likelihood function to be maximized is as follows:

\[
L(\theta) = f(y|\theta) = \int f(y|s)f(s|\theta)\,ds = \int f(y|s)f(s|\theta)\,ds = f(y|s)f(s|\theta) \tag{4}
\]

where \( f(y|s) = f(y|s)f(s|\theta) \), \( f(s) = f(s|\theta) \), \( \theta \) is a vector of initial parameters and \( f(s|y) = f(y|s)f(s) \) and \( E[f|y] \) denote expectation with respect to \( f(\cdot|y) \).

Briefly, the algorithm functions as follows:

1. Select an initial value for \( \theta_0 \), if \( S(0) = 0 \) is assumed in this first step, the parameters estimated for the GLM can be used as \( \theta_0 \).
2. Obtain values for \( s(1), \ldots, s(m) \), sampled by an MCMC from the distribution \( f(\cdot|y) \). The initial \( K \) samples are discarded as burn-in until it is judged that the equilibrium distribution of interest has been achieved. The subsequent realizations are used to obtain the \( s^{(j)}(y) \). Furthermore, to build a sample that is almost independent of the distribution, rather than storing the first \( m \) samples of the simulated chain, each sample is taken at every \( 1^{st} \) iteration.
3. Select the \( \theta_1 \) that maximizes the following function:

\[
L(m)(\theta) = \frac{1}{m} \sum_{j=1}^{m} f(s^{(j)}(y)|\theta) \tag{5}
\]

4. Repeat steps 2 and 3 until \( \theta_1 = \theta_{1+1} \).

The likelihood function \( L(\theta) \) can be approximated by using the MCMC to simulate samples from \( f(\cdot|y) \), departing from an initial \( \theta_0 \). The estimator for \( \theta \) will be that which maximizes this approximation.

4. Results and discussion

Table 1 summarizes all the variables considered in regard to the construction of the model. The large number of explanatory variables considered prioritized finding a GLM with a high goodness-of-fit using a reduced subset of factors.

Taking the occurrence of fire as the response variable, binary GLM models were fitted to different groups of explanatory...
variables to locate the minimum number of variables providing a high AUC value. Given the correlation between the explanatory variables, the solution to this problem was not a single subset of factors; rather, several combinations of factors obtained similar AUC values. Finally, only 6 explanatory variables those with an AUC above 0.5 were considered, variables that were highly correlated with these 6 variables were discarded. These variables were:

1. Percentage of agricultural land.
2. Number of dry-storm days (days with accumulated rainfall below 2.5 mm).
3. Mean altitude of the cell (m).
4. Number of positive strikes in the cell.
5. Percentage of woodland.
6. Number of strikes in broadleaf woodland.

Fig. 2 shows the ROC curves and corresponding AUC values for the binary GLM fitted for a single explanatory variable; Fig. 3 includes the ROC curve and AUC value of 0.73 for the GLM fitted for the 6 variables referred to above.

Once the explanatory variables were selected, a binary non-spatial GLM model was fitted using Bayesian methods and the MCMClogit function from the MCMCpack (R language). Table 2 shows the estimated coefficients, and the 2.5% and 97.5% quantiles for the fitted model. It can be observed that, for a significance level of α = 0.05, only 3 variables are significant: number of dry-storm days, mean altitude and percentage of woodland area.

Bearing in mind the special nature of the variables, a GLSM was fitted keeping the 5 explanatory variables. The correlation function considered was of the type \( \text{cov}(S(x), S(x')) = \sigma^2 \rho(x, x'; \varphi) + \sigma^2 \gamma(x - x') \), with \( \rho(x, x'; \varphi) = \exp \left( -\|x - x'\|/\varphi \right) \) as the isotropic exponential model. The fit was performed using Bayesian inference implemented via MCMC algorithms. First, it is necessary to define options for the simulation of the samples using the MCMC algorithm. The first 10,000 samples, where convergence was judged to have occurred, were discarded, and all the subsequent samples were used to obtain the posterior distributions of the parameters of interest. To avoid bias, not every observation for the MCMC algorithm was taken into account, but the chain was sampled for each 100th iteration of 50,000 iterations, yielding a sample of 500 values (see Christensen (2004) for more details). Subsequently, we used the binom.krige.bayes function of the geoRglm package to obtain the parameters that maximize the likelihood approximation to function given in Equation (5).

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>2.5% quantile</th>
<th>97.5% quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 ) (intercept)</td>
<td>-3.336</td>
<td>-1.435</td>
</tr>
<tr>
<td>( \beta_1 ) (agriculture land)</td>
<td>5.858E-04</td>
<td>2.90E-03</td>
</tr>
<tr>
<td>( \beta_2 ) (dry-storm days)</td>
<td>1.111E-01</td>
<td>1.88E-01</td>
</tr>
<tr>
<td>( \beta_3 ) (mean altitude)</td>
<td>-2.352E-03</td>
<td>-1.31E-03</td>
</tr>
<tr>
<td>( \beta_4 ) (positive strike)</td>
<td>4.742E-02</td>
<td>1.95E-01</td>
</tr>
<tr>
<td>( \beta_5 ) (woodland area)</td>
<td>3.443E-03</td>
<td>5.59E-03</td>
</tr>
<tr>
<td>( \beta_6 ) (broadleaf wood)</td>
<td>8.460E-03</td>
<td>2.64E-02</td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>2.5% quantile</th>
<th>97.5% quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 ) (intercept)</td>
<td>-3.336</td>
<td>-1.435</td>
</tr>
<tr>
<td>( \beta_1 ) (agriculture land)</td>
<td>5.858E-04</td>
<td>2.90E-03</td>
</tr>
<tr>
<td>( \beta_2 ) (dry-storm days)</td>
<td>1.111E-01</td>
<td>1.88E-01</td>
</tr>
<tr>
<td>( \beta_3 ) (mean altitude)</td>
<td>-2.352E-03</td>
<td>-1.31E-03</td>
</tr>
<tr>
<td>( \beta_4 ) (positive strike)</td>
<td>4.742E-02</td>
<td>1.95E-01</td>
</tr>
<tr>
<td>( \beta_5 ) (woodland area)</td>
<td>3.443E-03</td>
<td>5.59E-03</td>
</tr>
<tr>
<td>( \beta_6 ) (broadleaf wood)</td>
<td>8.460E-03</td>
<td>2.64E-02</td>
</tr>
</tbody>
</table>
The estimated parameters were as follows: \( \varphi = 6490, \sigma^2 = 0.173, \) \( r^2 = 0.732 \) and \( \beta = (-3.336, 5.858e-04, 1.111e-01, -2.352e-03, 4.742e-02, 3.443e-03, 8.460e-03). \)

It is important to remember that \( \beta \) has a conditional rather than a marginal interpretation; furthermore, since the estimation approaches are essentially different, a direct comparison of the spatial and non-spatial GLM is not recommended.

Table 3 shows the values for the coefficients estimated for the GLSM and the 2.5% and 97.5% quantiles. Comparing Tables 2 and 3 it can be observed that, for a significance level of \( \alpha = 0.05 \), the same significant variables are maintained in the 2 models, namely, number of dry-storm days, mean altitude and percentage of woodland area.

In Fig. 4 we overplotted the density plots for the coefficients estimated using GLM and GLSM. It can be observed that the inclusion of the spatial component in the GLSM method seems to have little influence on the distribution of most of the coefficients.

The ROC curve and AUC value were calculated for the GLSM. The AUC value for the spatial binary model was 0.99, indicating a substantial improvement in precision for the GLSM. This improvement is reflected in Fig. 5, which shows ROC curves and the corresponding AUC values for the binary GLM and GLSM.

Fig. 6 shows the fire probability map obtained using the GLSM. The highest probabilities can be observed to occur in the east and north of the province.

Referring to the sign and significance of the coefficients for each explanatory variable obtained in the GLM and GLSM, it can be deduced from the data in Table 2 and Table 3 that altitude plays a negative role in the probability of lightning igniting a fire. This contrasts with the results obtained by other authors (Rivas et al., 2005). The other topographic variables considered were not associated with the occurrence of fires. The presence of wooded areas was positively related to fire occurrence in both models. This corroborates the results obtained by other researchers (Vázquez and Moreno, 1998).

The number of dry-storm days, significant in both models, was another variable that favored the occurrence of fire. This result corroborates those of other authors (Rorig et al., 2007). The explanation is likely to be that the occurrence of storms on low-rainfall days favors the occurrence and propagation of fires.

The presence of positive lightning strikes or of strikes in broadleaf woodland was not significant in either of the models. This contrasts with the hypothesis of authors such as Latham and Williams (2001), who consider that the presence of fine fuels, such as dead leaves, on the forest floor favor ignition.
5. Conclusions

Most research to date into mathematical models that relate the occurrence of lightning-caused fires with physiographic and environmental variables fails to take into account the spatial structure of these variables. A general interpretation for the GLSM used in this work is that the spatial term represents the cumulative effect of unidentified and unobserved spatially structured covariates. In non-spatial GLM, the spatial correlation of residuals is not considered and this can significantly affect the quality of the statistical results.

It was shown that the GLSM fitted the data better than the standard GLM. It appears, therefore, that the GLSM is necessary to take account of any unexplained spatial variation not considered. This statement is justified by a comparison between the ROC curves and AUC values. In the specific problem considered in this article, the AUC for the GSLM is 35% larger than in the GLM model.

The analysis carried out above demonstrates the great potential of GLSM for analyzing spatial data. It also indicates the significant risk of drawing erroneous conclusions when non-spatial models are used for spatially structured data.

Acknowledgments

This work was partly funded by project 10PXIB300068PR of the Xunta de Galicia and by projects MTM2008-03129 and MTM2011-23204 of the Spanish Ministry of Science and Innovation.

References


Fig. 6. Locations of fires (left) and fire probability map for lightning-caused fires in the province of Leon in Spain (right).


