Predicting the occurrence of wildfires with binary structured additive regression models

Laura Ríos-Pena, a, * Thomas Kneib, b Carmen Cadarso-Suárez, c Manuel Marey-Pérez d

a Conservation Biology Department, Doñana Biological Station-CSIC, C/Américo Vespucio s/n, 41092 Seville, Spain.

b Chair of Statistics, Georg-August-University Göttingen, Humboldtalle 3, 37073 Göttingen, Germany.

c Unit of Biostatistics, Department of Statistics and Operations Research, School of Medicine, University of Santiago de Compostela, C/San Francisco s/n, 15782 Santiago de Compostela, Spain.

d PROEPLA Research Group, Polytechnic Institute, University of Santiago de Compostela, Campus Universitario s/n, 27002 Lugo, Spain.

*Corresponding author. E-mail: laura.rios@ebd.csic.es

Abstract

Wildfires are one of the main environmental problems facing societies today, and in the case of Galicia (north-west Spain), they are the main cause of forest destruction. This paper used binary structured additive regression (STAR) for modelling the occurrence of wildfires in Galicia. Binary STAR models are a recent contribution to the classical logistic regression and binary generalized additive models. Their main advantage lies in their flexibility for modelling non-linear effects, while simultaneously incorporating spatial and temporal variables directly, thereby making it possible to reveal possible relationships among the variables considered. The results showed that the occurrence of
wildfires depend on many covariates which display variable behaviour across space and time, and which largely determine the likelihood of ignition of a fire. The joint possibility of working on spatial scales with a resolution of $1 \times 1$ km cells and mapping predictions in a colour range makes STAR models a useful tool for plotting and predicting wildfire occurrence. Lastly, it will facilitate the development of fire behaviour models, which can be invaluable when it comes to drawing up fire-prevention and firefighting plans.

**Key words**: wildfires, structured additive regression models, covariates, voxel, penalized splines, Markov random fields.

1. Introduction

The concept of wildfire risk, defined as the probability of a wildfire occurring in an area within a given time interval, is well suited to strategic planning in fire and land management, since it integrates the probability of fire with its consequences (Bachmann and Allgower, 2000; Hardy 2005; Fairbrother and Turnley 2005; Finney et al. 2011; Scott 2006, Calkin et al. 2010). Quantitative approaches to risk assessment are still relatively new in wildland fire studies (Fairbrother and Turnley 2005; Finney et al. 2011; Scott 2006, Calkin et al. 2010). According to Miller and Ager (2013), two decision environments could be supported by fire risk analyses in the future, namely, those of fuel management and ignition management.

The 21st century has seen an increase in the number of wildfires (Plucinski 2013). Recent studies based on global burnt area estimate that 3.5 million km$^2$ are burnt every year (Chang and Song 2009; Giglio et al. 2010). The impact of climate change on
ignition occurrence has been documented by several authors (Kasischke and Turetsky
2006; Westerling et al. 2006), something that heightens the relevance of establishing
future scenarios of fire risk conditions. Forest fires are the leading environmental
problem in the Mediterranean region of Europe (Boubeta et al. 2015; Rodrigues et al.
2013; Fernandes et al. 2011; Fuentes-Santos et al. 2015; Rodríguez y Silva and
In the study of wildfires, different concerns have been considered, such as the increase
in forest area (Catry et al. 2009), overlap between urban development and forest
(Lampin-Maillet et al. 2010; Mitsopoulos et al. 2014; Moreno 2014), new agrarian
landscape patterns (Amatulli et al. 2007; Catry et al. 2009; Corcoran et al. 2007;
Fernandes 2009; Kalabokidis et al. 2007; Martínez et al. 2009; Moreira et al. 2009,
2011; Verde and Zézere 2010; Rego and Silva 2014), changes in traditional agrarian
landscape patterns (Ruíz-Mirazo et al. 2012; Chas-Amil et al. 2013), agricultural
abandonment (Acosta et al. 2005; Aranzabal et al. 2008; Falcucci et al. 2007),
exansion of shrubland surface (Bajocco and Ricotta 2008; Díaz-Delgado et al. 2004;
González et al. 2006; González and Pukkala 2007; Koutsias et al. 2012; Montané et al.
2009; Moreira et al. 2009; Mouillot el al. 2005; Numes et al. 2005; Sebastian-Lopez et
al. 2008), socio-economic changes (Arianoutsou 2001; Badia-Perpinya and Pallares-
Barbera 2006; Cardille et al. 2001; Curt and Delcros 2010; Loboda and Csizar 2007;
Martínez et al. 2009; Romero-Calcerrada et al. 2010; Vasconcelos et al. 2001; Vilar et
al. 2010; Wittenberg and Malkinson 2009; De Zea Bermudez et al. 2009; Chas-Amil et
al. 2013; Calviño et al. 2015), post-fire reactions (Conedera et al. 2011), and changes in
income distribution and higher unemployment levels (González-Olabarria et al. 2015).
Different ways of modelling the risk of wildfires have been explored, depending upon
their origin (e.g., whether caused by lightning or arson) (Albert-Green et al. 2013). In recent years, a variety of statistical methods have been used to predict the occurrence of wildfires based on the presence/absence of fire, in addition to the logistic generalized linear models (logistic GLM, McGullagh and Nelder 1989) that have traditionally dominated the literature on ignition occurrence modelling (Serra et al. 2008; Woolford et al. 2011). In particular, models have been developed which explain the temporal variability in the likelihood of fires occurring over time. Rodrigues et al. (2014) applied the Mann-Kendall non-parametric test as well as the Spearman and Cox-Stuart tests for data at the Nomenclature des Unités Territoriales Statistiques (NUTS) 3 level. As it discriminates between clustering and stochastic sequences, Benavent-Corai et al. (2007) used the coefficient of variation (CV) to highlight the presence of clusters in the temporal process. The dataset that they studied consisted of forest fires recorded from 1 January 1989 to 31 December 2003 in the Valencian Region in Spain. Levin and Heimowitz (2012) examined statistical relationships (using Pearson’s correlation coefficient) between weather variables and annual burnt areas. They also examined the effects of land-use and vegetation cover on fire frequency in Israel for all wildfires caused by humans in 2006 and 2010.

Many studies addressing optimal spatial management of forests under risk (Meilby et al. 2001) underscore the need to include spatial aspects in risk assessment (Koutsias and Karteris 1998; Koutsias 2003; Preisler et al. 2004). Kalabokidis et al. (2007) applied logistic regression and correspondence analysis to a spatial database which had been developed and managed in a GIS for the Sithonia peninsula in northern Greece. Artificial neural networks (ANNs, Zhanqing et al. 2001) have a great potential for modelling these events, since they can approximate mathematical relationships to which statistical models cannot be applied and in which relationships are not linear. Non-
normal distributions are common in wildfire data, and very high correlations are to be expected among landscape structure indices. To investigate these patterns, a study was made of the landscape structure and all pastoral wildfires recorded from 1988 to 2000 in 24 natural parks in Andalusia (Spain) (Ruiz-Mirazo et al. 2012). Viedma et al. (2009) used bootstrapping for validation of generalized linear models (GLMs) of the Sierra de Gredos range (Spain) for the period 1975-1990. Kens and Ager (2007) found quantitative risk assessment in the inclusion of stochastic spatial processes a major challenge. Tuia et al. (2008a) applied spatial clustering to fire data recorded from 1997 to 2003 in the Tuscany region of central Italy, using several independent indices (Voronoi polygon area index, Morishita index, and fractal dimension). In recent years, flexible regression models, such as generalized additive models (GAMs) (Hastie and Tibshirani 1990; Wood 2006), have been proposed for the purpose of studying possible non-linear effects of continuous covariates on the response variable (Roccaforte et al. 2012; Boer et al. 2009), i.e., these models support the fact that the response variable is modelled as a function of the smoothed predictor. Nevertheless, they have certain limitation when a spatial effect is included in the model.

Accordingly, this paper proposes the use of a structured additive regression (STAR) model with binary distribution, ultimately aimed at developing consistent, spatial-temporal models that can be easily integrated into a fire ignition-point prediction system.

2. Material and methods

2.1. Study Area

In terms of total wooded area, Spain is the third most forested state in Europe (with 15 million hectares), after Sweden and Finland (Forest Europe 2011). The Autonomous
Region (*comunidad autónoma*) of Galicia, where this study was conducted, is located in the north-west of the Iberian Peninsula. Galicia has a surface area of 29,574 km² (5.8% of mainland Spain) and a population of 2,795,422 (5.9% of the Spanish population), amounting to a population density of 99 persons/km², similar to the mean population density of Spain, i.e., 92 persons/km² (INE, 2014). To improve forest management, in 2008 Galicia was statutorily divided into 19 forest districts (Ley 43/2003, Decreto 43/2008).

The ratio of forested surface is notably higher than the national average (55.2%) (MAGRAMA 2014). As regards the local property regime, i.e., the basis on which holders are entitled to the land, most Galician forests are privately owned (97.2%) and are in the hands of a large number of individuals possessing an average of 2 ha per capita, which is in turn divided into an average of four plots (Marey-Pérez et al. 2012; Picos 2009). From a commercial point of view, Galicia is currently the leading timber forest product producer in Spain and the ninth in Europe (Balsa-Barreiro and Hermosilla, 2013). Hence, the timber industry is of significant economic importance. With annual revenues of €2,259,000,000 the forestry sector generates 26,000 direct and 50,000 indirect jobs and represents 4.5% of the regional GDP (FEARMAGA 2011).

Figure 1 shows the average annual distribution of the number of wildfires in Europe by province (Birot, 2009), and indicates that Galicia is one of the regions severely affected by wildfires (Comas et al. 2014). Indeed, between 1961 and 2014, over 251,106 wildfires were recorded in Galicia, affecting an estimated area of 1,830,000 ha, which accounts for approximately 66% of the region's surface area, being the accumulated burned area over the years higher than the current wooded forester’s area of Galicia. Furthermore, it should be noted that many rural areas in the north-western part of Spain
have been abandoned and, as a consequence, agro-pastoral activities linked to the traditional rural lifestyle have vanished (Stellmes et al., 2013; Balsa-Barreiro and Hermosilla, 2013). Gómez-Vázquez et al. (2009) and Marey-Perez et al. (2010, 2014a, 2014b) analysed wildfires in relation to property rights, and found that collective private property rights, in particular, might be the cause of many fires. Caballero (2015) examined the process of change of ownership and land use management, role of government and causes of conflict as potential explanatory variables for wildfires. Given the importance of the problem, different authors have developed and validated methods designed to account for and predict fires in Galicia in recent years. In this study, we used binary STAR for modelling the occurrence of wildfires in Galicia.

2.2. Source and description of the datasets
For analysis purposes, we used the wildfire database of the Galician Regional Rural and Maritime Authority (Consellería de Medio Rural y Mar) (mediorural.xunta.gal). This database contains information on the site (geographical co-ordinates) and time (hour and date) at which a fire occurs. We selected wildfires recorded during the first half of August 2006 (see Figure 2). A total of 2,060 wildfires were recorded in that period: of these, 1,553 were due to arson, 222 to unknown causes, 198 to resurgence of a previous wildfire, 58 to other causes, 9 to power lines, 5 to smoking, 5 to bonfires, 3 to landfill leaks, 3 to agricultural burning, 3 to engines and machinery, and 1 to burning brush (Xunta de Galicia, 2014). The spatial component was made up of 30,685 cells, derived by dividing Galicia into a 1 × 1 km cells. For analysis and presentation of results, we used a map of Galicia divided into municipal areas, which are the closest to basic local administrative units and consist of a total of 315 municipalities (IGN 2011). The centroid of each cell is associated with a municipal area, so that each grid pertains to a municipality. This map enabled the effect of spatial correlation of municipalities to be
interpreted with respect to wildfire occurrence. Moreover, working on this scale made for easier interpretation of results and lower computational costs.

Daily weather covariates were obtained from meteorological monitoring stations distributed throughout the study area. A total of 71 meteorological monitoring stations were operational during the study period. The daily weather covariates selected in this study were: (i) average temperature; (ii) average relative humidity; and, (iii) rainfall. We further restricted our analysis to meteorological stations situated at a maximum distance of 30 km from any given wildfire. Due to differences in altitude between fires and the nearest meteorological stations, it was necessary to make a correction of the temperature and rainfall variables with respect to the difference in altitude. The variation in daily average temperature with height was 0.65°C/100m to 11000 m (Nuñez and Colhoun 1986), i.e., \( tH = 15°C - 0.0065°C/m \cdot H \) where \( tH \) is the air temperature at height \( H \) and \( H \) is the difference between the wildfire and the nearest meteorological station. Likewise, a simple method for correcting rainfall (\( P \)) in relation to differences in altimetry is as follows, \( P_{(corrected)}(\text{monthly}) = P_{(\text{monthly})} \pm 0.08 \), i.e., the rainfall gradient amounted to an 8% variation per 100 m difference in height between any given ignition point and the meteorological station selected.

Rothermel fuel models (Rothermel, 1983) are a classification of different types of vegetation and their behaviour in fire. The combustibility of a forest system is defined as its capacity to burn, release enough material to be consumed and cause burning of neighbouring vegetation, thereby increasing the distribution and direction of wildfire. Each type of vegetation corresponds to a certain flammability and combustibility, which vary depending on the type and quantity of biomass and its spatial distribution or stratification (Galiana et al. 2009). The Rothermel fuel-model classification considered
13 types of fuel models, divided into 4 major groups based on the primary means of spreading fire, i.e., pasture, shrub, litter under trees and wood residuals. These models were drawn from the series of National Forest Inventories (MAGRAMA 2014). In addition, we included a fifth group, denoted as "unallocated", since Galician forest is extremely heterogeneous and this often makes it difficult to assign a forest area to one specific category. Table 1 provides a brief description of the response variable, explanatory covariates and spatial aggregation used for the purposes of our study.

2.3 Statistical methods

2.3.1. Structured Additive Regression

Structured additive regression (Fahrmeir et al. 2004) provides a generalization of generalized linear and additive models for a response from an exponential family, where a more general type of predictor including non-linear, spatial or random effects can be used to study the regression effects of a variety of covariates. In our analysis, we combined potentially non-linear effects of continuous covariates with a spatial effect based on the municipalities, as well as a random effect for such municipalities. The latter two effects capture any spatially structured and unstructured heterogeneity remaining after accounting for covariates.

In generic notation, the predictor of a STAR model for grid cell k at time t is given by

\[ \eta_{kt} = \beta_0 + f_1(v_{kt}) + \cdots + f_r(v_{kt}) \]  

where \( \beta_0 \) is the overall intercept, and the functions \( f_1, \ldots, f_r \), represent regression effects of different types of covariates generically called \( v_{kt} \). In our specific case, let \( p_{kt} \) denote the probability of occurrence of a wildfire at grid cell k at time t, with \( \text{km}^2 \) per day
being the unit measured. We then assume a logistic structured additive regression model where

$$
\log \left( \frac{p_{kt}}{1 - p_{kt}} \right) = \eta_{kt} = \beta_0 + f_1(v_{kt}) + \cdots + f_r(v_{kt}) + f_{struc}(v_{kt}) + f_{unstruc}(v_{kt})
$$

In this model specification, the functions $f_1, \ldots, f_r$ represent potentially non-linear effects of continuous covariates, while $f_{str}$ and $f_{unstr}$ represent spatially structured and unstructured heterogeneity for each municipality. The basic idea of STAR models is to approximate any of the covariate effects $f(v)$ by a linear combination of basis functions, such that

$$
f(v) = \sum_{j=1}^{J} \beta_j B_j(v)
$$

where $B_j(v)$ are the basis functions, while $\beta_j$ denotes the corresponding basis coefficients. Specific choices relevant to our modelling exercise will be discussed in the next paragraph. In matrix notation, the STAR predictor (1) can now be rewritten as

$$
\eta = \beta_0 1 + Z_1 \beta_1 + \cdots + Z_r \beta_r
$$

where $\eta$ is obtained by stacking all individual observations, 1 refers to a vector of ones, $Z_j$ arises from evaluating the basis functions at the observed covariates, and $\beta_j$ is the vector of corresponding basis coefficients.

To ensure specific properties of the function estimates, such as smoothness in the case of non-linear, and spatial effects or shrinkage in the case of cell-specific random effects,
we relied on a Bayesian formulation where prior distributions are assigned to the vectors of regression coefficients. Generically, these are given by multivariate normal distributions.

\[ p(\beta_j | \tau_j^2) \propto \exp \left( -\frac{1}{2\tau_j^2} \beta_j' K_j \beta_j \right) \]

where \( \tau_j^2 \) is a prior variance determining the impact of the prior distribution on the function estimates, and \( K_j \) is the prior precision matrix that implements the smoothness/shrinkage assumptions. From a frequentist perspective, these prior distributions correspond to penalties of the form \( \lambda_j \beta_j' K_j \beta_j \), where \( \lambda_j \geq 0 \) is the smoothing parameter that determines the impact of the penalty on the penalized likelihood estimate.

2.3.2. Model Components

For the non-linear effects of continuous covariates, we considered penalized splines (Eilers and Marx 1996, Brezger and Lang 2006), where a specific function \( f(x) \) is approximated by a linear combination of B-spline basis functions. A common choice is cubic B-splines with 20 equidistant knots that yield a twice continuously differentiable (i.e., visually smooth) function estimate, while allowing for enough flexibility to capture a typical non-linear pattern. As a roughness penalty, squared first- or second-order differences are used, which correspond to the assumption of a first- or second-order random walk prior (with the latter being the default choice).

For the spatially structured effect, we used a Markov random field approach, where
each grid cell is associated with a separate regression effect (such that the basis
functions correspond to indicator functions for the grid cells), while spatial smoothness
is achieved by assuming an adjacency matrix structure for the penalty matrix $K$. In
contrast, the same basis functions are used for the spatially unstructured part but in this
case, the penalty matrix is reduced to an identity matrix which corresponds to the
assumption of an independent identically distributed (i.i.d.) random effect. A more
detailed discussion of potential model components and how they fit into the generic
framework outlined above can be found in Fahrmeir et al. (2013).

2.3.3. Empirical Bayes Inference

In our analyses, we relied on an empirical Bayes approach that treats smoothness
variances $\tau_j^2$ as fixed unknowns to be estimated from the corresponding marginal
posterior. Estimation is thus basically reduced to iterating between optimising a
penalized logistic likelihood and numerical determination of the smoothing variances by
approximate restricted maximum likelihood estimation (see Fahrmeir et. al. 2004, for
details). The penalized likelihood estimates for the regression coefficients can be
obtained by denoting the complete model as $\eta = Z\beta$, where $Z = [1, Z_1, ..., Z_r]$ and
$\beta' = (\beta_0, \beta'_1, ..., \beta'_r)$, such that we iteratively compute

$$\hat{\beta}[t] = (Z'W^{[t-1]}Z + K)^{-1}Z'W^{[t-1]}\tilde{y}^{[t-1]}$$

where $K = \text{blockdiag}(0, \lambda_1 K_1, \ldots, \lambda_r K_r)$ and $W^{[t-1]}$ and $\tilde{y}^{[t-1]}$ are working weights
and working observations as in standard generalized linear models. To derive the
estimated model complexity of a specific term $f_j(\nu)$ in the structured additive predictor,
one can consider the equivalent degrees of freedom $df_j$ obtained from the trace of the
hat matrix

\[ H = Z(Z'WZ + K)^{-1}Z'W \]

upon convergence of the penalized likelihood iterations. These degrees of freedom can be interpreted as an effective parameter count, which can be used, for instance, to quantify the deviation of a non-linear effect from the simple linear case, with one degree of freedom for the slope parameter.

3. Results

We generated a model that studies the relationship between wildfires and daily weather covariates, types of fuel model on the ground, and any possible spatial effects corresponding to all the municipalities located in the study area. This model is expressed as follows,

\[
\log \left( \frac{p_{kt}}{1-p_{kt}} \right) = \beta_0 + f_1(altitude_k) + f_2(ta_mean_k) + f_3(hr_k) + f_4(pp_k) \\
+ f_5(unallocated_k) + f_6(pastures_k) + f_7(shrub_k) + f_8(litter_k) \\
+ f_9(wood_resid_k) + f_{struc}(cdconc_k) + f_{unstruc}(cdconc_k)
\]

where \(k\) is the index of the set of grid cells in the study area and period, \(p_{kt}\) denotes the probability of ignition point in the voxel \(k\) at time \(t\), given the observed covariate set.

The functions \(f_j\) are smooth effects capturing the potentially non-linear relationship between the explanatory variables and the logit probability of the response variable. These functions were estimated using Bayesian cubic P-splines with a second-order random walk prior and 20 inner knots. The spatial effects were estimated using Markov Random Fields (MRF, \(f_{str}\)) and using i.i.d. Gaussian random effects (\(f_{unstr}\)).
To demonstrate that the structured additive regression with binary response improves predictions of wildfires, as compared to classical logistic regression models and generalized additive models, a table was drawn up showing the results yielded by the following models: GLM; GLM with spatial effect; GAM; and lastly, STAR. The distribution of the response variable was binary in all cases. The selection criteria used to choose the model that produced the best estimate were the Akaike Information Criterion (AIC) and Generalized Cross-Validation (GCV), which are shown in Table 2 below, along with the $-2\times\log$-likelihood and degrees of freedom (df).

As can be seen from Table 2, the model that best estimated the occurrence of wildfires was the STAR model with binary response. Accordingly, a breakdown of the results obtained from the latter is now given in Figure 3 below.

The results of the smoothed effects of the weather and fuel model covariates are shown in Figure 3. The effect of altitude (Figure 3a) indicates a decrease in wildfire probability as altitude increases. This is reasonable, since most of the wildfires occurred in coastal areas, where altitudes are lower. The estimated degrees of freedom for this covariate were 2,502 which indicates slight non-linearity, with a gradual downward trend at low altitudes and a sharp decrease as altitude increases. Across the study period, the occurrence of fires in the region would thus appear to have been related to lower altitudes. It is noteworthy, however, that the credible bands increased in width from an altitude of approximately 1000 m, meaning that these results should not be overinterpreted, since there were few data showing altitudes above 1000 m in the study area.
Average daily temperature (Figure 3b) appeared to have a non-linear effect, implying strongly increasing occurrence probabilities up to a level of around 20°C. From this temperature upwards, the probability of wildfire occurrence remained approximately constant up to 30°C. From 30°C upwards, the probability started to increase again, though not as sharply as at low temperatures (<20°C). The degrees of freedom took a value of 4,069, which is in accordance with the visual impression of a non-linear effect.

The effect of relative humidity (Figure 3c) was rising for low relative humidity values up to about 35%; for higher values, the overall effect was downward, with the sharpest decline being observed for high humidity of around 75%-80%. Wildfire occurrence was thus associated with low relative humidity, which is reasonable in the light of the fact that most of the wildfires occurred in coastal areas, where relative humidity below 50% is common. Precipitation (Figure 3d) had an almost linearly decreasing effect.

In terms of type of fuel model, the effect of the unallocated category (Figure 3e) was non-linear and increasing, i.e., the greater the surface area with no known type of plant cover, the greater the likelihood of wildfire. The effect of pasture (Figure 3f) was linear and increasing. The effect of shrub (Figure 3g) showed that increasing the amount of area in this category also increased the probability of wildfires occurring; this category had a strong presence throughout the study area (see Appendix). The effect of litter (Figure 3h) was non-linear, with an increasing effect up to values of around 400,000 m², after which the effect remained stable up to 1 million m². The effect of wood residuals (Figure 3h) was linear and almost constant. However, since there was a low overall concentration of wood residuals/m², the results for this covariate should not be overinterpreted.
The effect of the spatial covariate (cdconc, corresponding to municipalities) was divided into spatially structured (Figure 4a and 4b) and unstructured effects (Figure 4c and 4d), as mentioned above. The plot with structured spatial effects (Figure 4a) shows that municipalities situated in the north and north-east had a low occurrence of fires as compared to municipalities situated in the west and south, after taking covariates into account. While 95% credible intervals (Figure 4b) present in black municipalities denote that the effect was significantly negative, white corresponds to significantly positive effects, and grey indicates a lack of significance.

The results of the unstructured spatial effects (Figures 4c and 4d) were mostly non-significant. Municipalities shown in pink had a greater possibility of a local effect than did those drawn in green. The Chandrexa de Queixa municipal area (depicted in the south-east in black) calls for a more detailed study to ascertain the possible causes of the singular pattern of wildfire in this area.

4. Discussion and conclusions

For over 40 years, researchers and forest services offices have been developing models that would be capable of predicting fire occurrence zones (Koutsias et al. 2015; Martínez-Fernández et al. 2013; Koutsias et al. 2010). This process has faced two major difficulties relating, firstly, to the randomness of the phenomenon, especially in the case of arson, and secondly, to the availability of statistics and computer tools that would enable accurate models to be developed for a given spatial and temporal context.

The designated goal of our study was to evaluate the predictability of ignition points by
using flexible regression techniques based on STAR models, a methodology that has been shown to provide greater accuracy than the simpler special cases of logistic GLMs and GAMs. Additive or geo-additive models with non-linear interaction effects, which consider spatial information in the form of a location variable, or models with random effects, can accelerate the development of fire behaviour models and prove most useful for drawing up fire prevention and firefighting plans. Their principal advantage lies in their flexibility, in that they can include spatial and temporal covariates along with the potentially non-linear effects of other continuous covariates. Moreover, these models generate maps of both structured and unstructured effects that can be plotted separately (Fahrmeir et al. 2013).

The validation process performed with ignition points across Galicia, a region with an arson problem of major dimensions, working at spatial scales with a pixel resolution of 1 × 1 km, a temporal resolution of one day, and physiographic, climatic and land-use covariates, combined with the possibility of mapping the predictions in a colour range, shows binary STAR models to be an important tool for predicting and managing wildfires in this region.

With respect to the model variables and results, several authors have studied wildfires from an environmental perspective for many years (Vega et al. 2009b; Lombao et al. 2014), and there are also various studies on the causes and consequences of emergency situations (González-Alonso and Merino-de-Miguel 2009; Balsa-Barreiro and Hermosilla 2013). In addition, while many authors have considered physiographic variables in their analyses, only a few have analysed the influence of altitude on fire ignition points. For instance, Vilar et al. (2010) reported a decrease in fire risk
probability with increasing altitude in the Madrid Region. This effect could be explained by less human land-use and activity at high elevations. Ordóñez et al. (2012) also found that the increase in altitude minimized wildfire risk in the Region of Leon. Our results appear to agree with these results, showing that increased human presence in near-shore areas also increases the risk of arson-induced wildfires.

Weather variables have been considered by most studies addressing all fire hazards. While certain temperature and humidity conditions are essential for wildfire to spread, in the case of arson-induced fires, these logically fail to explain the cause. Studies conducted in Galicia, such as that by Molina-Terrén and Cardil (2015), have studied the influence of temperature on large fires. Chas-Amil et al. (2015) considered precipitation and temperature in an attempt to understand the reasons underlying wildfire occurrences in 3600 parishes (the parróquia or administrative parish being the lowest tier of territorial designation in Galicia). Their results show that, whereas average temperature and precipitation are important when it comes to explaining the presence of fire, mean precipitation is not significant in the case of arson-induced fires. Likewise focusing on Galicia, Barreal and Loureiro (2015) used average summer rainfall and maximum summer temperature in Moran's and LISA statistics to analyse the spatial patterns of wildfires. Boubeta et al. (2015) introduced days without rain as a meteorological variable in a Poisson regression model for the entire region, and it was this model which best explained the concentration of wildfires in some areas. Our results appear to be in line with those of other authors, inasmuch as the presence of fire depends on certain weather conditions that often occur in Galicia in summer and facilitate incendiary activity.
Land-use variables, analysed statically or dynamically, have been commonly used to explain fire risk. Two main types of analysis have traditionally been used: one seeks to ascertain how changes in land-use affect the risk of fire; the other seeks to ascertain which type of land-use can explain an incendiary crisis. We adopted the latter approach. Much like us, when studying the occurrence of large fires in the south of France, Gantaume and Jappiot (2013) found that shrubland and pasture tended to appear in areas with fire risk. As explained by Vélez (2000) and Marino et al. (2014), shrubland and pasture are often associated with the practice of burning to rejuvenate pasture land in Spain. This means that the presence of shrub could well be a consequence of the form of land management rather than being genuinely associated with the occurrence of wildfires. Bisquert et al. (2014) constructed a fire danger index for Galicia and the neighbouring Region of Asturias. In the results, they note that increases in shrub and grassland areas were found to be significant in the accuracy of the index. Information relating to litter under trees and wood residuals are integrated into fuel-model types. Aguado et al. (2007) studied the characteristics and moisture content of forest residuals and considered them critical to fire risk. Chuvieco et al. (2010) analysed which variables influenced risk of fire in three regions of Spain (Aragon, Madrid and Valencia), and found dead-fuel moisture content to be an important risk factor. In their fire risk index for the Valencian Region, Vicente and Crespo (2012) introduced fuel conditions in the group of variables denoted as "ignition danger" and increased the accuracy of the index.

In terms of spatial distribution of ignition points, Comas et al. (2014) identified the south-western corner of Galicia as being especially important, a finding that is in good agreement with the area shown by our model to contain spatial aggregation points and
spatio-temporal structures with distances in a range of 0-12 km. In the north-west of Galicia, Fuentes-Santos et al. (2013) established spatial dependence between ignition points and an interaction radius of 3-4 km, corresponding to a lower concentration than in our results. Considering the region as a whole, Chas-Amil et al. (2015) noted that the highest concentrations were produced both in areas of higher population density, a finding that coincides with our results, and areas of greater depopulation. The difference in results might, at least in part, be related to the different observation times: August 2006 was an exceptional case in the history of fires in Galicia.

In summary, we believe logistic STAR models to be an important contribution to the toolbox of forestry managers and statisticians. Future challenges to the improvement of STAR models and their application to wildfire prediction are adding new variables at the scale of pixels used here. It should be noted that, in some cases, the estimates of the effect of certain covariates, such as rainfall, pasture group, litter-under-trees group and wood-residuals group, were seen to be perfectly linear. Improved data computation, the incorporation of variables linked to forest fires and a new statistical methodology would thus constitute a solid line of future research.

Appendix

1. Dataset

The total of nine variables studied can be classified into two types:

- Meteorological covariates. Data were collected via meteorological stations (see Figure 5). During the first half of August 2006, a total of 71 meteorological stations were in operation. Data were only taken into account where the distance between the
meteorological station and the wildfire was less than 30 km. The area from which data were extracted is shown in Figure 4.

- Rothermel fuel-model groups. Figure 6 shows the spatial distribution of the five fuel-model types in Galicia.

2. Software

All statistical analyses were performed using the R statistical software environment, and the following packages in particular:

- BayesX. This package provides functionality for the purpose of exploring and visualising the results of structured additive regression models obtained with BayesX. It also provides functions that read, write, and manipulate objects required to visualize spatial effects obtained with BayesX.

- R2BayesX. This package provides an interface between R and the BayesX statistical software (http://www.BayesX.org). To analyse the occurrence of wildfires, we used the bayesx function, which is the main function for fitting models in the R2BayesX package.

Acknowledgements

T. Kneib and C. Cadarso-Suárez would like to thank the Spanish Ministry of Innovation & Competitiveness for support from the MTM2014-52975-C2-1-R project; and M. Marey-Pérez and C. Cadarso-Suárez would like thank the Galician Regional Authority (Xunta de Galicia) for supporting research of the INBIOEST network (2014-PG141) and Competitive Research Reference Group (PROePLA CN2012/015).
References


Calviño, M., Chas-Amil, M., García-Martínez, E., Touza, J., 2015. Wildfire risk associate with land cover types, topography and human activities within and outside


configurations of fire ignition points through spatiotemporal point processes. Natural Hazards and Earth System Sciences 2, 2891-2911.


Fahrmeir, L., Kneib, T., Lang, S., 2004. Penalized additive regression for spacetime
data: a Bayesian perspective. Statistica Sinica 14, 731-761.


of Probabilistic Wildfire Risk Components for the Continental United States. Stochastic
Environmental Research and Risk Assessment 25(7): 973–1000.


Gómez-Vázquez, I., Álvarez-Álvarez, P., Marey-Pérez, M., 2009. Conflicts as enhancers or barriers to the management of privately owned common land: a method to
analyze the role of conflicts on a regional basis. For. Policy Econ., 11 (2009), pp. 617-627.


Ministry of Agriculture, Food and Environment of the Spain Government (MAGRAMA), 2014. IV National Forest Inventory.


Mouillot, F., Ratte, J., Joffre, R., Mouillot, D., Rambal, S., 2005. Long-term forest dynamic after land abandonment in a fire prone Mediterranean landscape (central


modelling of socioeconomic data to understand patterns of human-caused wildfire ignition risk in the SW of Madrid (central Spain). Ecological Modelling 221, 34-45.


Figures and tables

Figure 1. Location of Galicia in Europe and average annual distribution of the number of wildfires in the EU by province (Source: Birot, 2009)
Figure 2. Spatial distribution of wildfires in the study area. Red points mark the centroid of the wildfire.
Figure 3. Estimated non-linear effects for weather and fuel model covariates, with their corresponding 80% and 95% pointwise credible intervals.
Figure 4. Structured and unstructured spatial effects on wildfires.
Figure 5. Area distribution in cases where weather covariate data were obtained from weather stations.
Figure 6. Grid distribution in m$^2$ for the Rothermel fuel-model groups.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wildfires</td>
<td>&quot;occurrence&quot; (=1) and &quot;non-occurrence&quot; (=0) of wildfires</td>
</tr>
<tr>
<td>Altitude</td>
<td>Altitude recorded for each grid in m</td>
</tr>
<tr>
<td>Ta_mean</td>
<td>Average daily temperature recorded for each grid in ºC</td>
</tr>
<tr>
<td>Hr</td>
<td>Daily average relative humidity for each grid in %</td>
</tr>
<tr>
<td>Pp</td>
<td>Daily rainfall recorded for each grid in l/m$^2$</td>
</tr>
<tr>
<td>Unallocated_model</td>
<td>Heterogeneous formations not assigned to any model for each grid in m$^2$</td>
</tr>
<tr>
<td>Pastures</td>
<td>Grassland formations for each grid in m$^2$</td>
</tr>
<tr>
<td>Shrub</td>
<td>Shrub formations for each grid in m$^2$</td>
</tr>
<tr>
<td>Litter under trees</td>
<td>Litter-under-tree formations for each grid in m$^2$</td>
</tr>
</tbody>
</table>
Wood_residuals  Forest formations with considerable quantities of dead woody residuals for each grid in m²
Cdconc    Municipalities in Galicia

Table 1. Variables generated from the baseline data

<table>
<thead>
<tr>
<th>Model</th>
<th>-2*log-likelihood</th>
<th>df</th>
<th>AIC</th>
<th>GCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM</td>
<td>15137.4</td>
<td>10</td>
<td>15137.4</td>
<td>0.1527</td>
</tr>
<tr>
<td>GLM plus spatial</td>
<td>14088.1</td>
<td>155.74</td>
<td>14399.6</td>
<td>0.0491</td>
</tr>
<tr>
<td>GAM</td>
<td>14992.3</td>
<td>25.35</td>
<td>15137.4</td>
<td>0.0527</td>
</tr>
<tr>
<td>STAR</td>
<td>11059.0</td>
<td>141.54</td>
<td>11342.0</td>
<td>0.0390</td>
</tr>
</tbody>
</table>

Table 2. Models fitted