A spatio-temporal Poisson Hurdle point process to model wildfires

Laura Serra[1,2*], Marc Saez[2,1] , Pablo Juan[3] , Diego Varga[2,4], Jorge Mateu[3]

Abstract. Wildfires have been studied in many ways, for instance as a spatial point pattern or through modelling the size of fires or the relative risk of big fires. Lately a large variety of complex statistical models can be fitted routinely to complex data sets, in particular wildfires, as a result of widely accessible high-level statistical software, such as R. The objective in this paper is to model the occurrence of big wildfires (greater than a given extension of hectares) using an adapted two-part econometric model, specifically a hurdle model. The methodology used in this paper is useful to determine those factors that help any fire to become a big wildfire. Our proposal and methodology can be routinely used to contribute to the management of big wildfires.

Key words and phrases. Hurdle model, INLA, Spatio-temporal point processes, SPDE, Wildfire.

1. Introduction

Fire risk can be defined as a product of fire occurrence probability and expected impacts [3]. An area can be considered to have high wildfire risk if the probability of fire is high and the expected impacts of fire are large. Furthermore, fires are getting larger, more destructive, and more economically expensive due to fuel accumulations, shifting land management practices, and climate change. Wildfires have negative effects on human life and health, human property and wellbeing, cultural and natural heritage, employment, recreation, economic and social infrastructures and activities. It is worth noting that some fire episodes have caused catastrophic damages as loss of human lives and very significant economic and environmental losses.

The European Mediterranean is a highly populated region. Approximately 65,000 fires occur in the European Mediterranean region every year. Wildfires destroy around 500,000 hectares every year in the European Union, 0.7 to 1 million hectares in the Mediterranean basin. This has a serious impact on the environment and on socio-economic activities, especially in southern Europe. Over 95% of the fires in Europe are due to human causes. An analysis of fire causes show that the most common cause of fires comes from agricultural practices, followed by

Serra L, Saez M, Mateu J, Varga D, Juan P, Diaz-Ávalos C, Rue H. Spatio-temporal log-Gaussian Cox processes for modelling wildfire occurrence: the case of Catalonia, 1994-2008. Environmental and Ecological Statistics 2013.

negligence and arson ([34]). These wildfires are relatively frequent events with recurrence time of 23 years ([42]).

Wildfires also destroy biodiversity, increase desertification, affect air quality, the balance of greenhouse gases and water resources. During recent years the increasing extension of urban areas mixed with rural or forest areas associated with a marked increase of fire activity make this impact even greater. The intense urbanization of our societies, the abandonment of rural lands and rural activities such as forest management along with the rapidly expanding of urban/forest interface are key drivers for wildfires in Europe and in the Mediterranean region.

Weather is a fundamental component of the fire environment. The prolonged drought and high temperatures of the summer period in the Mediterranean climate are the typical drivers that demarcate the temporal and spatial boundaries of the main fire season. Future trends of wildfire risks in the Mediterranean region, as a consequence of climate change, will lead to the increase of temperature in the East and West of the Mediterranean, with more frequent dryness periods and heat waves facilitating the development of very large fires. Future scenarios of climate change should affect locally fire regimes, and therefore local analyses need to be performed by adapting global climatic models to regional conditions. Many factors have been considered to explain the temporal variation in fire regime in recent decades in Spain: Climate change is one factor, with a clear relationship between increasing number of days with extreme fire hazard weather and the number and size of fires in the Mediterranean coast of Spain.

Earlier detection often leads to smaller fire size, and therefore reduces the probability of fire escape ([21]), final fire size, cost and risks to fire response crews. Wildfire prevention should be considered as an important part of sustainable forest management and should integrate a landscape approach taking into account different land uses. Knowledge of short and long-term impacts of wildfire is essential for effective risk assessment, policy formulation, and wildfire management.

Spain is one of the most affected countries in Europe, both considering number of fires and area burned. Between 1980 and 2004 nearly 380.000 fires have occurred in Spain, and more than 4.7 millions hectares have been burned (roughly 10% of the country). Extreme fires (>500ha) are relatively frequent events with recurrence time of 2-3 years, causing large human, economic and environmental damage altogether. Their ignition and spread occur under favorable weather conditions, often following drought periods, in areas where fuel accumulation helps quick fire spread and high fire intensity, they usually burn out of control and can only be stopped when meteorological conditions support aerial and ground fire fighting ([39]). In Catalonia these fires only represent 1.4% of all fires and 79% of burned area. In this study we have included wildfires larger than 50ha because in the Mediterranean region represent more

than 75% of the area burned, although they represent only 2.6% of the total number of wildfires ([19] and [30]). Over the last few years, the occurrence of large wildfire episodis with extreme fire behavior has affected different regions of Europe: Portugal, south-eastern France, Spain and Greece.

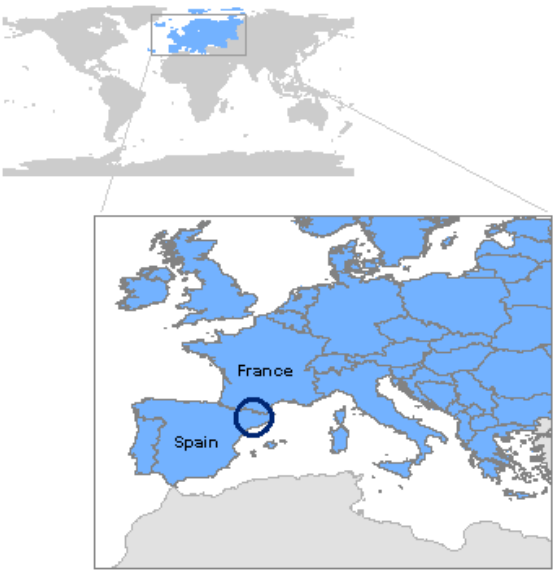

Figure 1.Catalonia location in Europe.

Wildfires have been studied in many ways, for instance as a spatial point pattern ([8], [9], [24], [42] and [44]) or through modelling the size of fires ([1]) or the relative risk of the big fires ([45]). Lately a large variety of complex statistical models can be fitted routinely to complex data sets, in particular wildfires, as a result of widely accessible high-level statistical software, such as R ([32]). Researchers from many different disciplines are now able to analyse their data with sufficiently complex methods rather than resorting to simpler yet non-appropriate methods. In this case, the objective in this paper is to model the occurrence of big wildfires, and to determine those factors which are significative in helping any fire to become a big wildfire.

We analyse the occurrence of big wildfires in Catalonia between 1994 and 2011, and consider a big wildfire to be a fire that burns areas larger than a fixed extension of hectares. Specifically we consider three sizes of areas; 50ha, 100ha and 150ha. Moreover, we distinguish between the numerous potential causes of wildfire ignition. In particular, we consider: (i) natural causes; (ii) negligence and accidents; (iii) intentional fires or arson; and (iv) unknown causes and rekindled. The study area encompasses 32,000 square kilometers and represents about 6.4% of the total Spanish national territory (1).

In addition to the locations of the fire centroids, several marks and covariates are considered. The year the wildfire occurred is the unique mark considered. The spatial covariates are also

considered, specifically, eight continuous covariates (i.e. topographic variables – slope, aspect, hill shade and altitude, proximity to anthropic areas – roads, urban areas and railways, and meteorological variables – maximum and minimum temperatures) and one categorical variable (land use).

The methodology for fitting spatial point process models to complex data sets has seen previous advances in facilitating routine model fitting for spatial point processes. For instance, the work by [4] has facilitated the routine fitting of point processes based on an approximation of the pseudolikelihood to avoid the issue of intractable normalizing constants ([5]) through the use of the library spatstat for R ([4]). In the same way, ([22]) consider hierarchical models able to analyse a wide variety of point process models, for example those appearing in fire problems.

In our case, spatio-temporal data can be idealised as realizations of a stochastic process indexed by spatial and temporal coordinates. Spatio-temporal clustering of wildfires might indicate the presence of risk factors which are not evenly distributed in space and time. In fact, what is usually of interest is to assess the association of clustering of wildfires to spatial and seasonal covariates ([42]). Covariate information usually comes in the form of spatial patterns in regular lattices or as regular vector polygons that may be rasterised into lattice images using GIS ([41]). The right methodological context able to deal with these pieces of information comes from spatio-temporal point processes. To bypass the problem of inefficiency in the estimation under a general integrated nested Laplace approximation (INLA)([36]), we have tried a computationally tractable approach based on stochastic partial differential equation (SPDE) models ([25]). On one hand, we use SPDE to transform the initial Gaussian Field (GF) to a Gaussian Markov Random Field (GMRF). GMRFs are defined by sparse matrices that allow for computationally effective numerical methods. Furthermore, by using Bayesian inference for GMRFs in combination to the INLA algorithm, we take advantage of the many significant computational improvements ([36]). If, in addition, we follow the approach suggested by Simpson et al. (2011), in which the specification of the Gaussian random field is completely separated from the approximation of the Cox process likelihood, we gain far greater flexibility.

The proposed method in this paper is an adapted two-part econometric model, specifically a Hurdle model. It consists of two stages and it is specified in such a way as to gather together the two processes theoretically involved in the presence of wildfires, that is, the fact to be a big wildfire (greater than a given extension of hectares) and the frequency of big wildfires per spatial unit. Specifically, the Poisson hurdle model consists of a point mass at zero followed by a truncated Poisson distribution for the non-zero observations.

This paper addresses two issues. We develop complex joint models for big wildfires and, at the same time, we provide methods facilitating the routine for the fitting of these models, using a Bayesian approach. The approach is based on the INLA, which speeds up parameter estimation substantially so that particular models can be fitted within feasible time.

This paper is organised as follows: the following section describes the data. Section 3 presents the methodology used, including the statistical framework, the description of the Poisson Hurdle model and the statistical inference explanation. Section 4 presents the results. Finally, the paper ends with a discussion and future coming steps.

2. DATA SETTING

In this paper we analyse the occurrence of big wildfires in Catalonia between 1994 and 2011. The total number of fires recorded in the analysis is 3,283, which are distributed as follows: 206 wildfires bigger than 50ha, 141 wildfires bigger than 100ha, and 112 wildfires bigger than 150ha. In Figure 2, on the left, we can see all wildfires and wildfires bigger that 50ha.

In Catalonia, the agency responsible for identifying the coordinates of the origin of the fire, the starting time and the cause of the fire is the Forest Fire Prevention Service (Government of Catalonia). In addition, they record the ending time of the fire, the hectares (and their type) affected, and the perimeter of the fire. The data used in this article are provided directly by the Service, and have been tested and polished before handling.

We distinguish between the numerous potential causes of wildfire ignition. In particular, we consider: (i) natural causes; (ii) negligence and accidents; (iii) intentional fires or arson; and (iv) unknown causes and rekindled. The first category includes lightning strikes or heat from the sun. The second takes into account that human carelessness can also start a wildfire, for instance, with campfires, smoking, fireworks or improper burning of trash. Negligence and accidents also includes those wildfires caused purely by chance. The third cause considers those wildfires that are started deliberately. Finally, the fourth set includes unknown causes and rekindled fires. In Figure 2, on the right, we show the spatial distribution of wildfires bigger than 50ha distinguishing by causes.

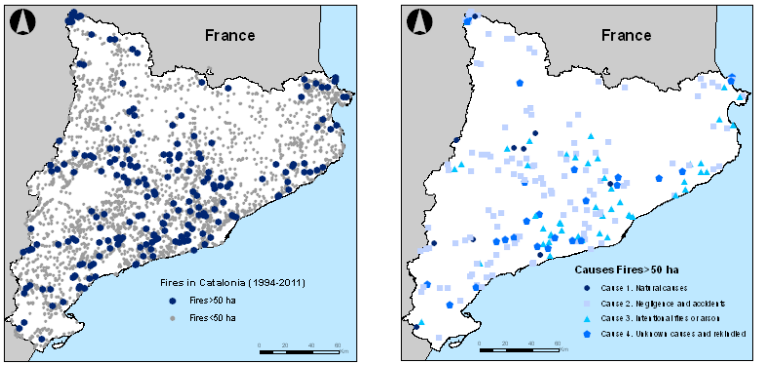

Figure 2. Left: All wildfires (1994-2011) and big wildfires. Right: Big wildfires distinguishing by causes.

In addition to the locations of the fire centroids, measured in Cartesian coordinates (Mercator transversal projections, UTM, Datum ETRS89, zone 31-N), several covariates are considered. Specifically, eight continuous covariates (i.e. topographic variables – slope, aspect, hill shade and altitude; proximity to anthropic areas – roads, urban areas and railways; and meteorological variables – maximum and minimum temperatures) and one categorical variable (land use).

Land use will obviously affect fire incidence, but moreover, topographic variables (slope, aspect and hill shade) affect not only fuel and its availability for combustion ([29]), but also the weather, inducing diverse local wind conditions, which include slope and valley winds. In fact, [15] point out that those topographic variables are relatively more important predictors of severe fire occurrence, than either climate or weather variables. The proximity to anthropic areas can be considered a factor explaining not only the incidence of fires in the intentional fires and arson category, but also why natural cause fires do not occur. As climatic variables are feasibly important for natural cause fires and perhaps rekindled fires, we use the maximum and minimum temperatures (further details can be found in [42]).

In this paper, slope is the steepness or degree of incline of a surface. Slope cannot be directly computed from elevation points; one must first create either a raster or a TIN surface. In this article, the slope for a particular location is computed as the maximum rate of change in elevation between the location and its surroundings. Slope is expressed in degrees. Aspect is the orientation of the slope and it is measured clockwise in degrees from 0 to 360, where 0 is north-facing, 90 is east-facing, 180 is south-facing, and 270 is west-facing. Hill shading is a technique used to visualise terrain as shaded relief by illuminating it with a hypothetical light source. Here, the illumination value for each raster cell is determined by its orientation to the light source, which, in turn, is based on slope and aspect and is also measured in degrees, from 0 to 360. Finally, altitude is considered as elevation above sea level and it is expressed in meters. To obtain topographic variables (DTM) we use the MET-15 model, which is a regular grid containing orthometric heights distributed according to a metricconverterProductID15 m15 m grid side, and is created for the Cartographic Institute of Catalonia. We also use the surface analysis tools included in the ArcGis10 application Spatial Analyst ([42]).

The distances, in meters, from the location of the wildfire to urban areas, roads and railroads, are constructed by considering a geographical layer in each case. The urban area and road layers are obtained from the Department of Territory and Sustainability of the Catalan Government, through the Cartographic Institute of Catalonia (ICC) (http://www.icc.cat). To obtain the two new raster layers we use the Euclidean distance function, included in the ArcGis10 application Spatial Analyst. Then, we use the merge function of ArcGis10 Geoprocessing module, to combine those two layers (urban areas and roads and railroads) into one single layer. The layers are continuous and defined as a raster layer (details can be found in [42]).

We also use the land use in Catalonia maps (1:250,000), with classification techniques applied on existing LANDSAT MSS images for 1992, 1997 and 2002 ([7], [17] and [35]). Additionally, we use orthophotomaps (1:5000) 2005-2007, to create the land use map for 2010. Specifically, we assign the land use map just before the date of each wildfire. We assign, as the land use, only the percentage value corresponding to the principal land use of the spatial units. In this paper, we transform the twenty-two categories, obtained from the Catalonian Cartographic Institute (ICC) cover map of Catalonia, into eight categories: coniferous forests; dense forests; fruit trees and berries; artificial non-agricultural vegetated areas; transitional woodland scrub; natural grassland; mixed forests; and urban, i.e., beaches, sand, bare rocks, burnt areas, and water bodies.

We also consider the temperatures (maximum and minimum) and up to seven days before the occurrence of the fire, at the location of the wildfire (note that meteorological data are provided by the Area of Climatology and Meteorological Service of Catalonia). The temperatures at the point of the occurrence of the wildfire, along with the temperatures from the previous day and up to a week before, are estimated by means of a two-step Bayesian model. Further details can be found in [37].

3. METHODS

3.1. **Statistical framework**. Spatio-temporal data can be idealised as realizations of a stochastic process indexed by a spatial and a temporal dimension

$$Y(s,t) \equiv \{y(s,t) | (s,t) \in D \times T \in \mathbb{R}^2 \times \mathbb{R}\} \tag{3.1}$$

where$D$ is a (fixed) subset of $\mathbb{R}^2$ and T is a temporal subset of $\mathbb{R}$. The data can then be represented by a collection of observations $y = \{y(s_1, t_1), \dots, y(s_n, t_n)\}$, where the set$(s_1, \dots, s_n)$ indicates the spatial locations, at which the measurements are taken, and $(t_1, \dots, t_n)$ the temporal instants.

In our case we assume separability in the sense that we model the spatial correlation by the Matérn spatial covariance function defined in (3.7) and the temporal correlation using a Random Walk model of order 1 (RW1). We introduce also the interaction effect between the space and time using another RW1 structure. Nevertheless, this inclusion does not change the separability structure. This temporal structure can be justified by the apparent randomness as shown in Figure 3. In fact, the dispersion of big wildfires varies between the periods considered. In particular, there is a reduction considering the number of them, specifically in the period 2008-2011.

3.2. **The Poisson hurdle model**. The model used in this paper is an adapted two-stage econometric model proposed by [13], specifically a hurdle model. It consists of two stages and specified in a way to gather together the two processes theoretically involved in the presence of wildfires, that is, the occurrence of being a big wildfire (greater than a given extension of hectares) and the frequency of big wildfires per spatial unit ([28]). Specifically, the Poisson hurdle model consists of a point mass at zero followed by a truncated Poisson distribution for the non-zero observations.

In the first stage, we predict the probability that any wildfire becomes larger than 50ha, 100ha and 150ha. In the second part, we model the number of these big wildfires per spatial unit.

The first part of the process can be modeled using a logistic regression that models the probability that any wildfire becomes larger than a fixed area

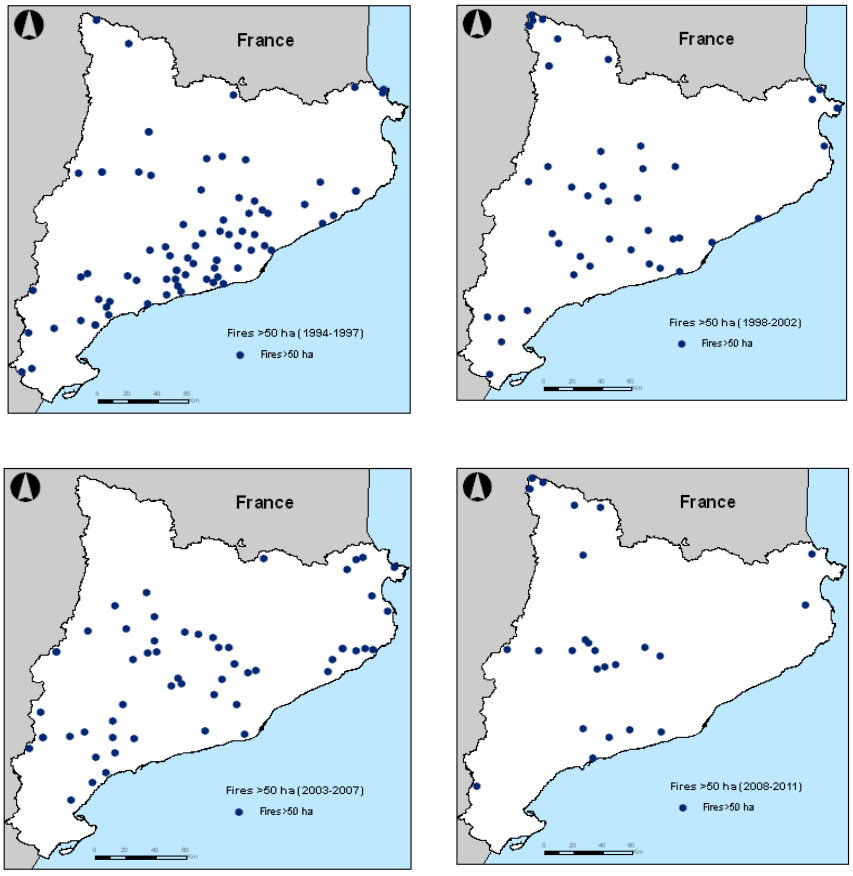

Figure 3.Big wildfires in Catalonia in 1994 to 2011. Left-Up: 1994-1997; Right-Up: 1998-2002; Left-Down: 2003-2007 and Right-Down: 2008-2011.

$$p_{itk} = Prob(y_{itk} > A|Z, \beta)$$

$$log\left(\frac{p_{itk}}{1 - p_{itk}}\right) = Z'\beta + S_i + \tau_t + \upsilon_{it} \quad (3.2)$$

where A denotes one of the fixed area's values (50ha, 100ha or 150ha), y is the response variable (in this case, each wildfire), Z a matrix of explanatory spatial covariates (containing the intercept), β is the vector of unknown parameters associated with the covariates, the subscript i denotes the wildfire, the subscript t (t=1994,..., 2011) the year of occurrence of the wildfire, and the subscript k (k =1,..., 4) the cause of occurrence. We also introduced three random effects: (i) spatial dependence, S_i, (ii) temporal dependence, τ_t and (iii) spatio-temporal interaction, υ_{it}.

In accordance with that proposed by [27], in the second stage of the model the distribution of being a big wildfire follows a truncated Poisson that models the number of big wildfires per spatial unit, introducing covariates and spatial random effects ([28])

$$p(y_{itk}|S_i) = (1-p_{itk})1_{(y_{itk}<A)} + p_{itk}\,Tpois(y_{itk};\mu_{itk})1_{(y_{itk}>A)}$$
$$\log(\mu_{itk}) = \eta(p_{itk}) \tag{3.3}$$
$$\eta(p_{itk}) = \sum_m \beta_m Z_{m,it} + S_i + \tau_t + \upsilon_{it}$$

where $Tpois(y_{itk};\mu_{itk})$ denotes a truncated Poisson distribution with parameter μ_{itk}, η denotes a link function such as the logit link, $Z_{m,it}$ represents the same spatial covariates used in the first stage, and β_m denotes the parameters associated with these covariates.

The particular estimation process has two steps. In the first step we use a binomial link in order to estimate the occurrence of a big wildfire. The probabilities of occurrence obtained from this first step are used in the second step as interim priors. In the second step the link is a truncated Poisson distribution. In any case, the likelihood of each part is introduced multiplicatively in only one equation.

3.3. Statistical inference.

3.3.1. *SPDE approach.* The SPDE approach allows to represent a Gaussian Field with the Matérn covariance function defined in (3.7) as a discretely indexed spatial random process which produces significant computational advantages ([25]). Gaussian Fields are defined directly by their first and second order moments and their implementation is highly time consuming and provokes the so-called "big n problem". This is due to the computational costs of $O(n^3)$ to perform a matrix àlgebra operation with $n \times n$dense covariance matrices, which is notably bigger when the data increases in space and time. To solve this problem, we analyse an approximation that relates a continuously indexed Gaussian field with Matérn covariance functions, to a discretely indexed spatial random process, i.e., a Gaussian Markov random field (GMRF). The idea is to construct a finite representation of a Matérn field by using a linear

combination of basis functions defined in a triangulation of a given domain D. This representation gives rise to the stochastic partial differential equation (SPDE) approach given by (3.8), which is a link between the GF and the GMRF. This link allows replacement of the spatio-temporal covariance function and the dense covariance matrix of a GF with a neighbourhood structure and a sparse precision matrix, respectively, typical elements that define a GMRF. This, in turn, produces substantial computational advantages ([25]).

In particular the SPDE approach consists in defining the continuously indexed Matérn GF X(s) as a discrete indexed GMRF by means of a basis function representation defined on a triangulation of the domain D,

(3.4) $$X(s) = \sum_{l=1}^{n} \varphi_l(s)\omega_l$$

where n is the total number of vertices in the triangulation, $\{\varphi_l(s)\}$ is the set of basis function and $\{\omega_l\}$ are zero-mean Gaussian distributed weights. The basis funcions are not random, but rather are chosen to be piecewise linear on each triangle

$$\varphi_l(s) = \begin{cases} 1 \; at \; vertix \; 1 \\ 0 \; elsewhere \end{cases}$$

The key is to calculate the weights $\{\omega_l\}$, which reports on the value of the spatial field at each vertex of the triangle. The values inside the triangle will be determined by linear interpolation ([41]).

Thus, expression (3.4) defines an explicit link between the Gaussian field $X(s)$ and the Gaussian Markov random field, and it is defined by the Gaussian weights $\{\omega_l\}$ that can be given by a Markovian structure.

Both the temporal dependence (on t) and the spatio-temporal interaction (on j and t) are assumed smoothed functions, in particular RW1 ([33]). Thus, RW1 for the Gaussian vector $x = (x_1, \dots, x_n)$ is constructed assuming independent increments

(3.5) $$\Delta x_i = x_i - x_{i-1} \sim N(0, \tau^{-1})$$

The density for x is derived from its n−1 increments as

(3.6) $$\pi(x|\tau) \propto \tau^{(n-1)/2} exp\left\{-\frac{\tau}{2}\sum(\Delta x_i)^2\right\} = \tau^{(n-1)/2} exp\left\{-\frac{1}{2}x^T Q x\right\}$$

where $Q = \tau R$ and R is the structure matrix reflecting the neighbourhood structure of the model ([33]).

Considering a spatio-temporal geostatistical data we need to specify a valid spatio-temporal covariance function defined by $Cov(y_{it}, y_{jq}) = \sigma_C^2 M(s_i, s_j|t, q)$where $\sigma_C^2 > 0$ is the variance component and $M(s_i, s_j|t, q)$ is the Matérn spatio-temporal covariance function. Depending on our assumptions the spatio-temporal covariance function can be adapted to each situation. In

the case of stationarity in space and time, the spatio-temporal covariance function can be specified as a function of the spatial Euclidean distance Δ_{ij}, and of the temporal lag $\Delta_{tq} = |t - q|$so it is defined by $Cov(y_{it}, y_{jq}) = \sigma_C^2 M(\Delta_{ij}; \Delta_{tq})$. If we assume separability, the spatio-temporal covariance function is given by $Cov(y_{it}, y_{jq}) = \sigma_C^2 M_1(\Delta_{ij}) M_2(\Delta_{tq})$, with M_1 and M_2 being the spatial and temporal correlation functions, respectively. Alternatively it is possible to consider a purely spatial covariance function given by $Cov(y_{it}, y_{jq}) = \sigma_C^2 M(\Delta_{ij})$ when t=q and 0 otherwise. In this last case, the temporal evolution could be introduced assuming that the spatial process evolves in time following an autoregressive dynamics ([20]).

Assuming separability we need to define the Matérn spatial covariance function which controls the spatial correlation at distance $\|h\| = \|s_i - s_j\|$ and this covariance is given by

(3.7) $$M(h|\nu, k) = \frac{2^{1-\nu}}{\Gamma(\nu)} (k\|h\|)^{\nu} K_{\nu}(k\|h\|)$$

whereK_{ν} is a modified Bessel function of the second kind and k>0 is a spatial scale parameter whose inverse, 1/ k, is sometimes referred to as a correlation length. The smoothness parameter ν>0 defines the Hausdorff dimension and the differentiability of the sample paths ([18]). Specifically, we tried ν=1,2,3 ([31]). Using the expression defined in (3.7), when $\nu + d/2$ is an integer, a computationally efficient piecewise linear representation can be constructed by using a different representation of the Matérn field x (s), namely as the stationary solution to the stochastic partial differential equation (SPDE) ([41])

(3.8) $$(k^2 - \Delta)^{\alpha/2} x(s) = W(s)$$

A $\alpha = \nu + d/2$ is a integer, $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial s_i^2}$ is the Laplacian operator and $W(s)$ is spatial white noise.

In the general spatial point process context, intensity stands for the number of events (fires in our case) per unit area. When considering the total intensity in each cell, we refer to the number of fires per cell area. A particular problem in our wildfire dataset is that the total intensity in each cell, Λjt is difficult to compute, and so we use instead the approximation, Λjt ≈ |sj | exp(ηjt (sj)), where ηjt (sj) is a 'representative value' (i.e., it represents the intensity or number of fires in a particular cell given by a linear predictor of covariates and other terms) ([41]), within the cell and |sj | is the area of the cell sj. To treat this kind of problems, Cox processes are widely used. In particular, Log Gaussian Cox processes (LGCP), which define a class of flexible models are particularly useful in the context of modelling aggregation relative to some underlying unobserved environmental field ([22]; [41]) and they are characterised by their intensity surface being modeled as

(3.9) $$\log(\lambda(s)) = Z(s)$$

where$Z(s)$ is a Gaussian random field.

3.3.2. *LGCP*. Conditional on a realization of $Z(s)$, a log-Gaussian Cox process is an inhomogeneous Poisson process. Considering a bounded region $\Omega \subset \mathbb{R}^2$ and given the intensity surface and a point pattern Y, the likelihood for a LGCP is of the form

(3.10) $$\pi(Y|\lambda) = exp\left(|\Omega| - \int_{\Omega} \lambda(s)ds \prod_{s_i \in Y} \lambda(s_i)\right)$$

where the integral is complicated by the stochastic nature of $\lambda(s)$. We note that, the log-Gaussian Cox process fits naturally within the Bayesian hierarchical modelling framework. Furthermore, it is a latent Gaussian model, which allows to embed it within the INLA framework. This embedding paves the way for extending the LGCP to include covariates, marks and non-standard observation processes, while still allowing for computationally efficient inference ([23]).

The basic idea is that, as we have explained in previous paragraphs, from a Gaussian Field (GF) with a Matérn covariance function, we use a SPDE approach to transform the initial Gaussian Field to a Gaussian Markov Random Field (GMRF), which, in turn, has very good computational properties. In fact, GMRFs are defined by sparse matrices that allow for computationally effective numerical methods. Furthermore, by using Bayesian inference for GMRFs, it is possible to adopt the Integrated Nested Laplace Approximation (INLA) algorithm which, subsequently, provides significant computational advantages.

Because our data is potentially zero inflated, as not all our events will become big fires, in this paper we present a spatial Poisson hurdle model to address these particular aspects of the data.

3.3.3. *Bayesian computation*. In a statistical analysis, to estimate a general model it is useful to model the mean for the i-th unit by means of an additive linear predictor, defined on a suitable scale

(3.11) $$\eta_i = \alpha + \sum_{m=1}^{M} \beta_m z_{mi} + \sum_{l=1}^{L} f_l(v_{li})$$

where α is a scalar which represents the intercept, $\beta = (\beta_1, \dots, \beta_M)$ are the coefficients which quantify the effect of some covariates $z = (z_1, \dots, z_M)$ on the response, and $f = \{f_1(.), \dots, f_L(.))$ is a collection of functions defined in terms of a set of covariates $v = (v_1, \dots, v_L)$. From this definition, varying the form of the functions $f_l(.)$ we can estimate different kind of models, from standard and hierarchical regression, to spatial and spatio-temporal models ([36])

Given the specification in (3.8), the vector of parameters is represented by θ ={ α, β, f}.

In our case, assuming that the subscript i denotes the wildfire, the subscript j the municipal district and the subscript t (t=1994... 2011) the year of occurrence of the wildfire, for each cause, we specify the log-intensity of the Poisson process by a linear predictor ([23]) of the form

(3.12) $$\eta_{ijt}(s_j) = \alpha_{0j} + \beta_1 G_{ijt} + \beta_2 Z_{jt} + \beta_3 W_j + S_j + \tau_t + \upsilon_{jt}$$

where α_{0j} represents the heterogeneity accounting for variation in relative risk across different municipals districts,G_{ijt} represents those covariates which depend on the wildfire, the municipal district and the time, Z_{jt} represents those covariates which depend onthe municipal district and the time, W_j corresponds to those covariates which only depend on the municipal district,S_j is the spatial dependence, τ_t is the temporal dependence, and υ_{jt} is the spatio-temporal interaction.

Note that, we assume separability between spatial and temporal patterns and allow interaction between the two components.

Following the Bayesian paradigm we can obtain the marginal posterior distributions for each of the elements of the parameters vector

(3.13) $$p(\theta_i|y) = \int p(\psi|y)p(\theta_i|\psi, y)d\psi$$

and (possibly) for each element of the hyper-parameters vector

(3.14) $$p(\psi_k|y) = \int p(\psi|y)pd\psi_{-k}$$

Thus, we need to compute: (i) $p(\psi|y)$, from which all the relevant marginals $p(\psi_k|y)$ can be obtained, and (ii) $p(\theta_i|\psi, y)$, which is needed to compute the marginal posterior for the parameters. The INLA approach exploits the assumptions of the model to produce a numerical approximation to the posteriors of interest, based on the Laplace approximation ([43]).

Operationally, INLA proceeds by first exploring the marginal joint posterior for the hyper-parameters $\hat{p}(\psi|y)$ in order to locate the mode; a grid search is then performed and produces a set G of "relevant" points $\{\psi^*\}$ together with a corresponding set of weights, $\{w_{\psi^*}\}$to give the approximation to this distribution. Each marginal posterior $\hat{p}(\psi^*|y)$ can be obtained using interpolation based on the computed values and correcting for (probable) skewness, e.g. by using log-splines. For each ψ^*, the conditional posteriors $\hat{p}(\theta_i|\psi^*, y)$ are then evaluated on a grid of selected values for θ_i and the marginal posteriors $\hat{p}(\theta_i|y)$ are obtained by numerical integration ([6])

(3.15) $$\hat{p}(\theta_i|y) \approx \sum_{\psi^* \in G} \hat{p}(\theta_i|\psi^*, y)\hat{p}(\psi^*|y)w_{\psi^*}$$

Given the specification in (3.12), the vector of parameters is represented by $\theta_j = \{\beta, \beta_\alpha, S, \tau_t, \upsilon_{jt}\}$ where we can consider $X_i = (S, \tau_t, \upsilon_{jt})$ as the i-th realization of the latent GF X(s) with the Matérn spatial covariance function defined in (3.7). We can assume a GMRF prior on θ, with mean 0 and a precision matrix Q. In addition, because of the conditional independence

relationship implied by the GMRF, the vector of the hyper-parameters $\psi = (\psi_S, \psi_\tau, \psi_v)$ will typically have a dimension of order 4 and thus will be much smaller than θ.

Note that in both parts of the model we control for heterogeneity, spatial dependence and spatio-temporal extra variability. Models are estimated using Bayesian inference for Gaussian Markov Random Field (GMRF) through the Integrated Nested Laplace Approximation (INLA).

The use of INLA and the SPDE algorithms produce massive savings in computational times and allow the user to work with relatively complex models in an efficient way. All analyses are carried out using the R freeware statistical package (version 2.15.2) ([32]) and the R-INLA package ([33]).

4. RESULTS

We note that, in general, wildfires caused by natural causes are not larger than 50ha. The same happens for those fires caused by unknown causes or for those rekindled. For this reason, even if we have analysed the forth causes we focus our results only on big wildfires caused by negligence and accidents and on those caused intentionally or arson.

4.1. **First stage results**.

We first consider a logistic regression to model the probability of a wildfire becoming larger than a particular area. Table 1 shows the significant factors of the logistic model distinguishing by the three sizes (50ha, 100ha and 150ha) and considering wildfires occurred by negligence and accidents (cause 2) and those caused by intention or arson (cause 3). The main factors that have an influence in the presence of wildfires (larger than a given extension of hectares) are the orientation and the land use. Taking into account the rest of the covariates considered we can see that the hill shade, the distance to anthropic areas and the maximum temperature have no influence in the probability of a fire to become larger than a specific area. Table 2 shows the means of the posterior distributions for the hyper-parameters of the first stage considering the three sizes of area analysed. The heterogeneity, the time and the interaction have a small impact and moreover, their values decrease when the extension of the wildfires increases. We can also appreciate that there are not big differences between the two causes. On the other hand, the values of the spatial component show that there is an important spatial dependence, especially for wildfires occurred by negligence and accidents.

In Figures 4 and 5, we show the marginal distribution of hyper-parameters κ, τ, ρ, heterogeneity, time and interaction for Causes 2 and 3. In all of them, the distribution is Gamma, the distributions are similar for both causes. Finally, Figure 6 shows the prediction of the probability of a fire to become larger than 50ha as well as the standard deviation of this prediction. Looking

at the wildfires occurred by negligence and accidents we can see that higher probabilities are concentrated around the main urban areas of Catalonia: Girona (in the north-east), Barcelona (in the middle of the coast), Tarragona (in the south along the coast) and Lleida (in the centre west). There are also high probabilities in the north-west, corresponding to a large forest area. With respect to intentional and arson wildfires the probabilities are less concentrated than in wildfires occurred by negligence and accidents but are also higher in the same areas. Regarding the standard deviation we do not appreciate alarming values. On the second cause higher values are found where the probabilities are also higher. The third cause presents lower values of deviation than wildfires occurred by negligence and accidents meaning that the model works better with wildfires occurred by intention or arson.

	Cause 2			Cause 3		
	50	100	150	50	100	150
(Intercept)	X	X	X	X	X	X
factor(Aspect)2						
factor(Aspect)3		X				
factor(Aspect)4				X	X	X
factor(Slope)2						
factor(Slope)4						
factor(Slope)5			X			
factor(Altitude)3	X					
factor(Land use)1	X					
factor(Land use)3		X	X			
factor(Land use)4					X	X
factor(Land use)6						X
ftmin 3					X	
ftmin 5		X				

Table 1. Significative factors for the logistic model in the first stage of the analysis.

	50ha		**100ha**		**150ha**	
	Cause 2	**Cause 3**	**Cause 2**	**Cause 3**	**Cause 2**	**Cause 3**
Heterogeneity	0.000054	0.000054	5.212E-09	5.192E-09	3.959E-09	5.247E-09
Space	0.246900	0.148810	0.3908300	0.0520790	0.0884000	0.0131780
Interaction	0.000043	0.000043	3.885E-09	3.827E-09	3.408E-09	3.762E-09
Time (year)	0.000053	0.000049	5.187E-09	5.135E-09	4.444E-09	4.759E-09

Table 2. Means of the posterior distributions for the hyper-parameters of the first stage.

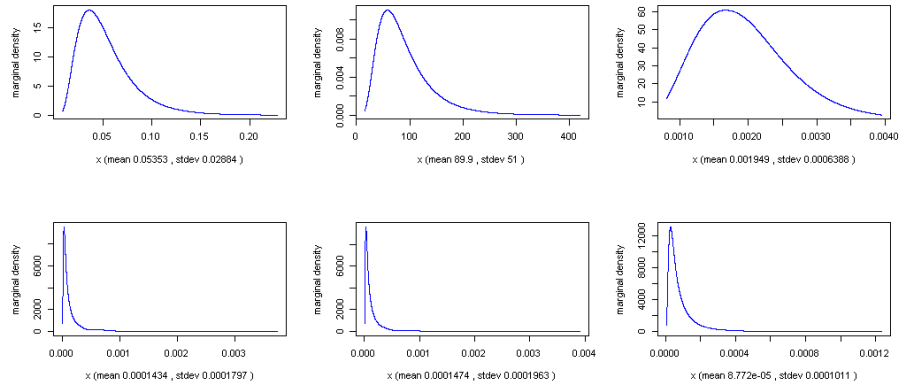

Figure 4. From Top-Left to Bottom-Right: Marginal posterior distribution for κ, τ, ρ, heterogeneity, time and interaction, respectively, for Cause 2.

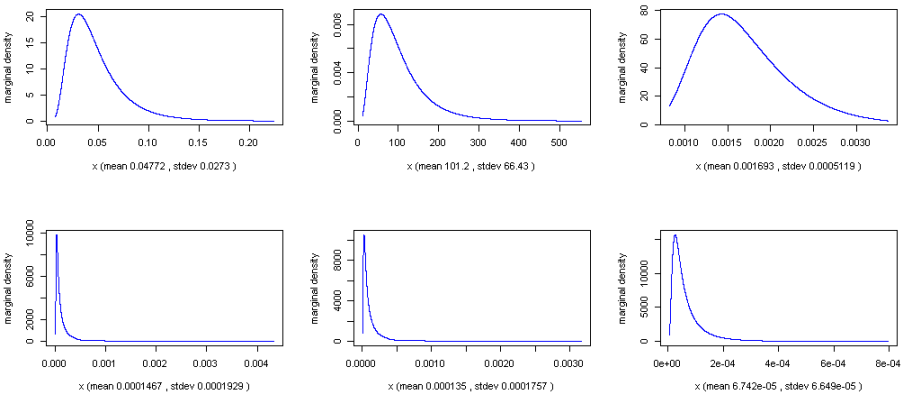

Figure 5. From Top-Left to Bottom-Right: Marginal posterior distribution for κ, τ, ρ, heterogeneity, time and interaction, respectively for Cause 3.

4.2. **Second stage results**. In the second stage we model the frequencies of wildfires (larger than a specific area) per spatial unit. Table 3 shows the values of the hyper-parameters. It is important to note that in this second stage the spatial values are not included. The reason is because there is a too high correlation between the spatial dependence component, S_i, and the spatio-temporal interaction, υ_{jt}, that prevents the model from working properly. Therefore, we introduce the spatial random effect through the interaction. The heterogeneity is quite much significant than in the first stage, especially for intentional wildfires and arson. Something similar happens with the interaction. It is much larger than in the first stage and it is also more representative for wildfires occurred by intention and arson. Finally, with respect to the temporal dependence, this is also larger than in the first stage but it has almost no variation between the two causes. In addition there are not relevant differences between the three extensions of hectares in any of the three hyper-parameters analysed. In Figure 7, we show the marginal

posterior distribution of hyper-parameters for heterogeneity, time and interaction for Causes 2 and 3. In all of them, the distribution is Gamma. Finally, Figure 8 shows the predicted number of wildfires larger than 50ha per spatial unit. Wildfires occurred by negligence and accidents and those caused by intention or arson present the same pattern of distribution according to the probabilities obtained in the first stage of the model. In general, big wildfires are concentrated along the coast being denser around the metropolitan area of Barcelona. Looking at the standard deviations we point out that intention wildfires and arson have very low values so, again, we note that the model correctly fits wildfires occurred intentionally or arson.

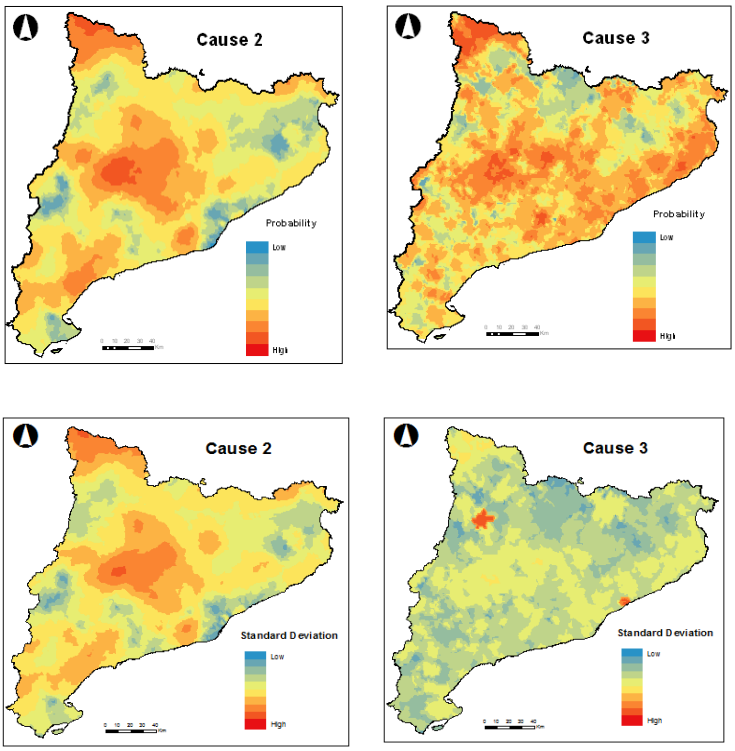

Figure 6. Top: Prediction maps for Cause 2 and Cause 3. Bottom: Standard Deviation for the prediction under Cause 2 and Cause 3.

	50ha		**100ha**		**150ha**	
	Cause 2	**Cause 3**	**Cause 2**	**Cause 3**	**Cause 2**	**Cause 3**
Heterogeneity	0.116645	1.083424	0.116918	1.088495	0.116836	1.089681
Interaction	0.000181	0.010143	0.000177	0.010101	0.000180	0.009634
Time (year)	0.000048	0.000048	0.000047	0.000048	0.000048	0.000040

Table 3. Hyper-parameters for the model in the second stage.

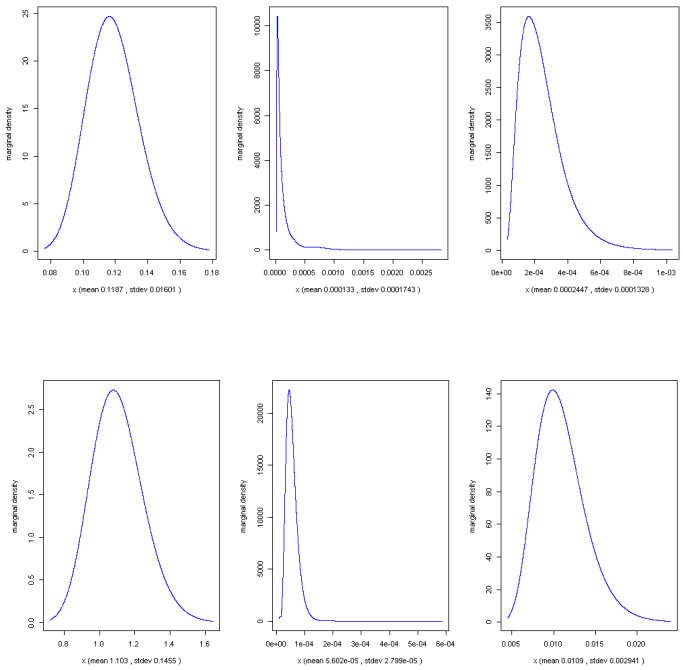

Figure 7.Posterior distribution of the hyper-parameters for the second stage. Left: heterogeneity, Middle: time and Right: interaction. First line: Cause 2, second line: Cause 3.

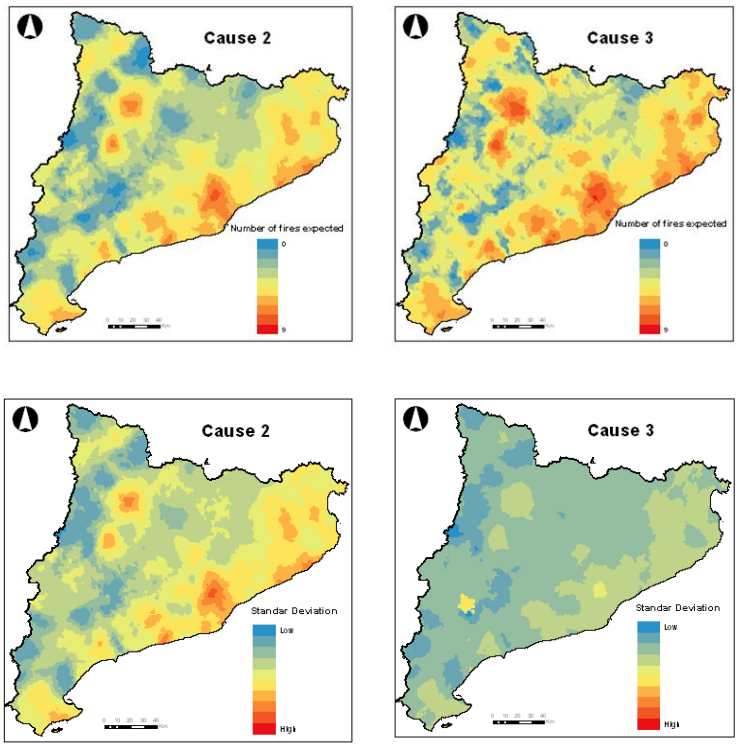

Figure 8. Number of fires expected Maps: On the Top: Cause 2 and Cause 3 and on the Bottom: Cause 2-sd and Cause 3-sd.

5. DISCUSSION

The main finding of this study is that big wildfires are mostly caused by human actions either by negligence and accidents or by intention or arson. These results make sense with what the bibliography shows and what we have commented in the introduction; over 95% of the fires in Europe are due to human causes.

Normally a natural wildfire does not spread as much as an intentional wildfire and so, the number of wildfires which are larger than a big extension, is not enough to obtain results. Analyzing the four causes separately we noticed no significant results for wildfires caused by natural causes and for those caused by unknown causes or rekindled. In fact separating wildfires by cause and by its extension we almost did not have wildfires caused by natural causes nor unknown causes or rekindled. In particular in our data there are only 15 wildfires bigger than 50ha occurred by natural causes compared to 180 caused by negligence or accidents. Our model does not work properly with such a limited small number of data so, even if we have studied the four causes, we have restricted the study to the second and the third causes. To analyse and estimate the number of zeros in a dataset there are different statistical alternatives. On one hand we have the ZIP model, which is employed to estimate event count models in which the data result in a larger number of zero counts than would be expected. The hurdle Poisson model [27] is a modified count model with two processes, one generating the zeros and one generating the positive values. The two models are not constrained to be the same.

The concept underlying the hurdle model is that a binomial probability model governs the binary outcome of whether a count variable has a zero or a positive value. If the value is positive, the ”Hurdle is crossed,” and the conditional distribution of the positive values is governed by a zero-truncated count model. In the ZIP models, unlike the hurdle model, there are thought to be two kinds of zeros, ”true zeros” and ”excess zeros”. Although the practical results are very similar in both approaches, hurdle models are most appropriate in our case, since every wildfire can turn into a big wildfire and therefore, every point is susceptible to become larger than a specific number of hectares.

ACKNOWLEDGEMENTS

Work partially funded by grant MTM2010-14961 from the Spanish Ministry of Science and Education.

REFERENCES

[1] Amaral-Turkman, M.A., Turkman, K.F., Le Page. Y., Pereira, J.M.C. (2011). Hierarchical space-time models for fire ignition and percentage of land burned by wildfires. *Environmental and Ecological Statistics*, **18**, 601-617.

[2] Alan A. Ager, Nicole M. Vaillant, Mark A. Finney, Haiganoush K. Preisler (2012). Analyzing wildfire exposure and source-ink relationships on a fire prone forest landscape *Forest Ecology and Management*, **267**, 271-283.

[3] Bachmann A. and Allgower, B. (2001). A consistent wildland fire risk terminology is needed. *Fire Management Today*, **61** (4), 28-33

[4] Baddeley, A. and Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, **12**, 1-42.

[5] Berman, M. and Turner, T.R. (1992). Approximanting point process likelihoods with GLM. *Journal of Applied Statistics*, **41**, 31-38.

[6] Blangiardo, M., Cameletti, M., Baio, G. and Rue, H. (2013). Spatial and Spatio-temporal models with R-INLA.*Spatial and Spatio-temporal Epidemiology*, **4**, 33-49.

[7] Chuvieco, E. (2009). *Earth Observation of Wildland Fires in Mediterranean Ecosystems*.Springer-Verlag Berlin Heidelberg.

[8] Comas, C., Palahi, M., Pukkala, T. and Mateu, J. (2009). Characterising forest spatial structure through inhomogeneous second order characteristics.*Stochastic Environmental Research and Risk Assessment*, **23**, 387-397

[9] Comas, C. and Mateu, J. (2011). Statistical inference for Gibbs point processes based on field observations. *Stochastic Environmental Research and Risk Assessment*, **25**, 287-300.

[10] Cressie, N.A.C. (1993). *Statistics for Spatial Data(Revised ed.).* New York: Wiley.

[11] Díaz-Avalos, C., Peterson, D. L., Alvarado C. E., Ferguson, Sue A. and Besag, J.E. (2001). Space-time modelling of lightning-caused forest fires in the Blue Mountains, Oregon. *Canadian Journal of Forest Research*, **31**(9), 1579-1593.

[12] Daley, D. and Vere-Jones, D. (2003). An Introduction to the Theory of Point Processes, 2nd edition. Volume I. Springer, New York.

[13] Deb, D.P. and Trivedi, P.K. (2002). The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics*, **21**, 601-625.

[14] Diggle, PJ. (2003). *Statistical Analysis of Spatial Point Patterns*. (2nd ed). Arnold, London UK.

[15] Dillon, G.K., Holden, Z.A., Morgan, P., Crimmins, M.A., Heyerdahl, E.K. and Luce, C.H. (2011). Both topography and climate affected forest and woodland burn severity in two regions of the western US, 1984 to 2006. *Ecosphere*, **2** (12), 1-33.

[16] Flannigan, M. and Wotton, B. (2001). *Climate, weather, and area burned*. In Forest Fires: Behavior and Ecological Effects E. Johnson and K. Miyanishi (eds), 351-373. San Diego: Academic Press.

[17] García, M., Chuvieco, E., Nieto, H. and Aguado, I (2008). Combining AVHRR and meteorological data for estimating live fuel moisture. *Remote Sensing of Environment*, **112** (9), 3618-3627.

[18] Gneiting, T., Kleiber, W. and Schlather, M. (2010). Matérn Cross-Covariance functions for multivariate random fields. *Journal of the American Statistical Association*, **105** (491), 1167-1177.

[19] Gonzalez, J.R. and Pukkala, T. (2007). Characterization of forest fires in Catalonia (north-east Spain). *European Journal of Forest Research*, **126**, Issue 3, 421-429.

[20] Harvill, J. L. (2010). Spatio-temporal processes. *WIREs Computational Statistics*, **2**, 375-382.

[21] Hirsch, R.E., Lewis, B.D., Spalding, E.P., and Sussman, M.R. (1998). A role for the AKT1 potassium channel in plant nutrition.*Science*, **8**, 918-921.

[22] Illian, J.B. and Hendrichsen, D.K. (2010). Gibbs point process models with mixed effects. *Environmetrics*, **21**, 341-353.

[23] Illian, J.B., Sorbye, S.H. and Rue, H. (2012). A toolbox for fitting complex spatial point processes models using integreted nested Laplace approximations (INLA). *The Annals of Applied Statistics*, **6**, Issue 4, 1499-1530.

[24] Juan, P., Mateu, J. and Saez, M. (2012). Pinpointing spatio-temporal interactions in wildfire patterns.*Stochastic Environmental Research and Risk Assessment*, **26**, Issue 8, 1131-1150.

[25] Lindgren, F., Rue, H. and Lindstrom, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields the SPDE approach. *Journal of the Royal Statistical Society, Series B*, **73**, 423-498.

[26] Møller, J. and Díaz-Avalos, C. (2010). Structured spatio-temporal shot-noise Cox point process models, with a view to modelling forest fires.*Scandinavian Journal of Statistics*, **37**, 2-15.

[27] Mullahy, J. (1986). Specification and Testing of Some Modified Count Data Models.*Journal of Econometrics*, **33**, 341-365.

[28] Neelon, B., Ghosh, P. and LoebsMullahy, P.F. (2013). A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society. Series A*, **176** (2), 389-413.

[29] Ordóñez, C., Saavedra, A., Rodríguez-Pérez, J.R., Castedo-Dorado, F. and Covin, E. (2012). Using model-based geostatistics to predict lightning-caused wildfires.*Environmental Modelling and Software*, **29** (1), 44-50.

[30] Piñol, J., Terradas, J. and Lloret, R. (1998). Climate warming, wildfire hazard, and wildfire occurrence in coastal eastern Spain. *Climate Change*, **38**, 345-357.

[31] Plummer, M., Penalized, L. (2008). Functions for Bayesian model Comparation. *Biostatistics*, **9** (3), 523-539.

[32] R Development Core Team. (2011). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, http://www.r-project.org/.

[33] R-INLA project [available in: http://www.r-inla.org/home, accessed on August 13, 2012].

[34] Reus Dolz, M.L. and Irastorza, F. (2003). Estado del Conocimiento de causas sobre los incendios forestales en España. *APAS and IDEM Estudio sociologico sobre la percepcion de la población española hacia los incendios forestales*. www.idem21.com/descargas/pdfs/IncediosForestales.pdf.

[35] Røder, A., Hill, J., Duguy, B., Alloza, J.A. and Vallejo, R. (2008). Using long time series of Landsat data to monitor fire events and post-fire dynamics and identify driving actors. A case study in the Ayora region (eastern Spain).*Remote Sensing of Environment*, **112** (1), 259-273.

[36] Rue H., Martino S. and Chopin N. (2009). Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion). *Journal of the Royal Statistical Society*, Series B, **71**, 319-392.

[37] Saez, M., Barcelo, M.A., Tobias, A., Varga, D., Ocaña-Riola, R., Juan, P. and Mateu, J. (2012). Space-time interpolation of daily air temperatures.*Journal of Environmental Statistics*, **3** (5).

[38] San-Miguel-Ayanza, J., Rodrigues, M., Santos de Oliveira, S., Kemper Pacheco, C., Moreira, F., Duguy, B. and Camia, A. (2012). Land cover change and fire regime in the European Mediterranean region. *Post-Fire Mangement and Restoration of Southern European Forests Managing Forest Ecosystems*, F. Moreira, M. Arianoustsou, P. Corona, J. de las Heras (Eds.), Springer, Berlin, Heidelberg. 21-43.

[39] San-Miguel-Ayanza, J., Moreno, J. M. and Camia, A. (2013). Analysis of large fires in European Mediterranean landscapes: Lessons learned and perspectives. *Forest Ecology and Management*, **294**, 11-22.

[40] Silva, J.S., Vaz, P., Moreira, F., Catry, F. and Rego, F.C. (2011). Wildfires as a major driver of landscape dynamics in three fire-prone areas of Portugal.*Landscape and Urban Planning*, **101**, Issue 4, 349-358.

[41] Simpson, D., Illian, J., Lindgren, F., Sorbye, S.H. and Rue, H. (2011). Going off grid: computationally efficient inference for log-Gaussian Cox processes.Technical Report, Trondheim University.

[42] Serra, L., Juan, P., Varga D., Mateu, J. and Saez, M. (2012). Spatial pattern modelling of wildfires in Catalonia, Spain 2004-2008.*Environmental Modelling and Software*, **40**, 235-244.

[43] Tiernery, L. and Kadane, J.B. (1986). Accurate Approximations for posterior moments and marginal densities.*Journal of the American Statistical Association*, **81**, 82-86.

[44] Turner, R. (2009). Point patterns of forest fire locations. *Environmental and Ecological Statistics*, **16**, 197-223.

[45] Wang, Z., Ma, R. and Li, S. (2012). Assessing area-specific relative risks from large forest fire in Canada. *Environmental and Ecological Statistics*, To appear.

[46] Wisdom M. Dlamini, A. (2010). Bayesian belief network analysis of factor influencing wildfire occurrence in Swaziland.*Environmental Modelling and Software*, **25**, Issue 2, 199-208.

[1] CIBER of Epidemiology and Public Health (CIBERESP), [2] Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Spain, [3] Department of Mathematics, Campus Riu Sec, University Jaume I of Castellon, Spain, [4] Geographic Information Technologies and Environmental Research Group, University of Girona, Spain.