Point-process based Bayesian modeling of space-time structures of forest fire occurrences in Mediterranean France

Thomas Opitz, Florent Bonneu, Edith Gabriel

To cite this version:

Thomas Opitz, Florent Bonneu, Edith Gabriel. Point-process based Bayesian modeling of space-time structures of forest fire occurrences in Mediterranean France. Spatial Statistics, 2020, 40 (100429), 10.1016/j.spasta.2020.100429 . hal-02098412

HAL Id: hal-02098412
https://univ-avignon.hal.science/hal-02098412

Submitted on 12 Apr 2019

Point-process based Bayesian modeling of space-time structures of forest fire occurrences in Mediterranean France

Thomas Opitz

BioSP, INRA, F-84914 Avignon, France

Florent Bonneu, Edith Gabriel

LMA EA2151, Avignon University, F-84000 Avignon, France

Abstract

Due to climate change and human activity, wildfires tend to become more frequent and extreme, causing economic and ecological disasters. The deployment of preventive measures and operational forecasts can be aided by stochastic modeling that helps to understand and quantify the mechanisms governing the occurrence intensity. We here use a point process framework for wildfire ignition points in the French Mediterranean basin since 1995, and we fit a spatio-temporal log-Gaussian Cox process with monthly temporal resolution in a Bayesian framework using the integrated nested Laplace approximation (INLA). Human activity is the main direct cause of wildfires and is indirectly measured through a number of appropriately defined proxies related to land-use covariates (urbanization, road network) in our model, and we further integrate covariates of climatic and environmental conditions to explain wildfire occurrences. We include spatial random effects with Matérn covariance and temporal autoregression at yearly resolution. Two major methodological challenges are tackled : first, handling and unifying multi-scale structures in data is achieved through computer-intensive preprocessing steps with GIS software and kriging techniques; second, INLA-based estimation with high-dimensional response vec-

[☆] Fully documented templates are available in the elsarticle package on CTAN.

tors and latent models is facilitated through intra-year subsampling, taking into account the occurrence structure of wildfires.

Keywords: Integrated Nested Laplace Approximation; log-Gaussian Cox process; SPDE approach; wildfires

1. Introduction

In the French territories close to the Mediterranean, the number of wildfire occurrences and their burnt surfaces have reached alarmingly high levels due to evolutions such as the increase in human activity and their interfaces with forest areas, and climate change. For instance in 2017, $2,300$ forest fire occurrences with $20,000$ hectares of burnt surface have been reported. For efficient risk prevention, we first have to understand the stochastic mechanisms governing the spatial and temporal variability in the intensity of wildfire occurrences. In this context, statistical modeling of forest fires is crucial to identify the drivers of fire occurrences and to develop operational forecasting tools. With the point-process based modeling of space-time structures in wildfire occurrences proposed in this work, we take a major step in this direction by taking into account climatic and environmental trends and complex interaction in the spatio-temporal distribution.

Wildfires have already been studied with point process approaches to assess how the spatial heterogeneity of wildfires at a fixed time depends on the spatial distribution of land use like vegetation, urban zones or wetlands (Juan et al., 2012; Møller & Diaz-Avalos, 2010; Pereira et al., 2013; Serra et al., 2014). In practice, raw data of environmental covariates is often available at different spatial and temporal scales, often with very high resolution, and comes in different numerical formats. Appropriate preprocessing of such data is important to obtain good predictive models. Raw data may also have a low signal-to-noise ratio with respect to forest fire prediction, and one can obtain better predictive relevance by generating artificial covariates that summarize relevant conditions. While such preprocessing steps are often not presented or done in the literature, we will here put focus on how we preprocess raw data to obtain meaningful covariates for driving the point process intensity.

The high spatio-temporal dimension of wildfire data (observed occurrences and control cases without occurrences) has often been coped with by separating the data into subsets or by strongly aggregating them, by year or by spatial areas (e.g., Genton et al., 2006; Serra et al., 2014; Turner, 2009; Xu & Schoenberg, 2011). Some recent approaches have concentrated more strongly on studying the interplay of the spatial and temporal structures (Gabriel et al., 2017). In the following, we choose a monthly temporal resolution to be able to capture the intra-year seasonality of weather influences and other effects.

We here use log-Gaussian Cox process models, which have already been identified as useful models for wildfires since they allow capturing spatio-temporal aggregation structures through random effects (Gabriel et al., 2017; Pereira et al., 2013; Serra et al., 2014). Since we consider a monthly time scale over 24 years with a spatial resolution given by the $2km$-grid for wildfire occurrence numbers and covariates, established by the the French authority responsible for forest fires ("Défense de la forêt contre les incendies", DFCI in short), a very high-dimensional response vector and covariate matrix arises in our model. Bayesian inference for log-Gaussian Cox processes using the integrated nested Laplace approximation (INLA, see Rue et al., 2009; Illian et al., 2012) is now well-established, but the high dimension of our regression model remains extremely challenging. To overcome this difficulty, we here develop a subsampling technique to reduce the length of the covariate vector. This method is designed to "trade space for time" and relies on the strong small-scale spatial dependence in covariates and seasonal behavior, and at each DFCI grid cell it subsamples the months of the year where no fire occurred. This technique allows us to divide by 11 approximately the size of the data matrix used inside INLA.

The remainder of the paper is organized as follows. Section 2 describes data, related to forest fire occurrences, environmental and climatic covariates. It details the GIS-based preprocessing steps, and for weather variables we use daily station data and interpolate them by spatio-temporal kriging. Section 3 develops our log-Gaussian Cox process models with different spatial-temporal structures for the random effets. Next, Section 4 is dedicated to the Bayesian estimation method where we detail our choices for prior distributions for INLA and present the subsampling technique of intra-year effects. We report and interpret parameter estimation results and intensity prediction of our most complex model in Section 5. Finally, we conclude with a discussion of our approach and further perspectives in Section 6.

2. Data on fire occurrences and covariates

2.1. Prométhée fire occurrence dataset

Since 1973, the French Government maintains the database Prométhée of forest fire occurrences in Southern France to allow for the development of statistical tools for a better knowledge of the spatial and temporal distribution of wildfires and their causes. We consider a subset of this database by selecting all wildfire occurrences, with burnt areas more than 1 hectare, in the years from 1995 to 2018 in the French Mediterranean basin, which is composed of seven "départements"[1] with overall surface of around $40,000\ km^2$. We focus on the years 1995 to 2018 since they correspond to the time period where our purely spatial environmental covariate dataset has been established. The spatial resolution of wildfire reports is given by the DFCI coordinates spanning a grid in the Lambert93 projection with quadratic cells covering approximatively $4km^2$ each. The study area is covered by $9,562$ DFCI cells. Therefore, there is some positional uncertainty since we do not know the exact locations of wildfire occurrences inside the cells. Figure 1 shows a map of the $23,309$ wildfire occurrence positions during the study period.

2.2. Land use and land cover

Land use and land cover information is very useful to model the probability of wildfire occurrences. The concomitance of environmental conditions (water,

[1] Pyrénées-orientales, Aude, Hérault, Gard, Bouches-du-Rhône, Var, Alpes-maritimes.

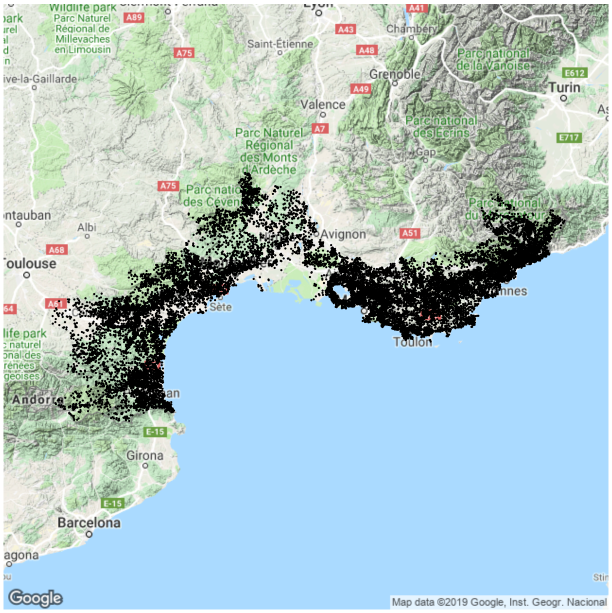

Figure 1: Map showing the $23,309$ wildfire locations observed during $1995-2018$ over the $9,562$ DFCI grid cells in the Mediterranean basin, Southern France.

vegetation type) and human infrastructure (buildings, road networks) is often crucial to explain wildfire outbreaks, which occur most of the time due to human negligence. We exploit several land use databases[2], freely provided by the French National Institute of Geographic and forest information (IGN) for research purposes, to generate land use covariates. IGN maps are available for each "département" in several formats (shapefiles or .tiff images) with different spatial resolution (2m, 25m) and geometry types (lines, polygons). We chose 14 covariates, subdivided into 4 groups: 1) water, elevation and slope; 2) protected zones: special regulatory rules for tourism; 3) vegetation: open forest, coniferous forest, shrubland, moorland, coniferous and coppices; 4) urbanization with building cover, primary roads, secondary roads, roads and paths.

The first goal is to unify these datasets with different spatial resolution towards the DFCI grid. By computer-intensive preprocessing steps, these raw

[2] BD FORET V1, BD TOPO and BD ALTI

datasets are transformed to summary covariates of land cover into the DFCI grid. For each covariate, we provide a single map by using GIS software (QGIS) to aggregate the databases available for the seven "départements". Next, we use the recent R package `sf`, allowing us to handle spatial geometric objects, to compute the proportion of millions of polygon areas of forests and buildings, and the overall road lengths for the cells of a regular grid with 200 meter resolution. Finally, for each DFCI cell at $2km$ resolution and each covariate, we compute the mean and standard deviation of the 100 values in the 200m-subgrid inside each DFCI cell. This two-step approach has the advantage to keep a maximum of information concerning the fine-scale structure covariate values, even if we have to aggregate them to the DFCI grid. The mean summary shows the overall trend inside each DFCI cell, and the standard deviation measures the variability around this trend.

Another important objective is to generate synthetic covariates that highlight the interface of forested areas to human activity. Indeed, wildland-to-urban interfaces are heterogeneous areas where the two main drivers of wildfires, vegetation and human activity, are present and favor outbreaks. We compute two additional covariates to take into account the interface between open forests and urbanized areas, and between open forests and paths. For this purpose, we multiply the percentage of open forests by respectively the percentage of buildings and the total length of paths inside each DFCI cell.

Land use is of course a dynamic process evolving over time and some minor to moderate changes may have taken place during the study period, but we have only static spatial data for the covariates described above. Indeed, the census process used by IGN is long and spans over several years to provide complete maps for regions such as the French Mediterranean area. We point out that the impact of environmental changes may be relatively minor in our case since our study period corresponds to the census period for the IGN databases that we use here; they provide a summary of partial measurements over the time period $1995-2000$. Maps for several of the land cover covariates are shown in Figure 2.

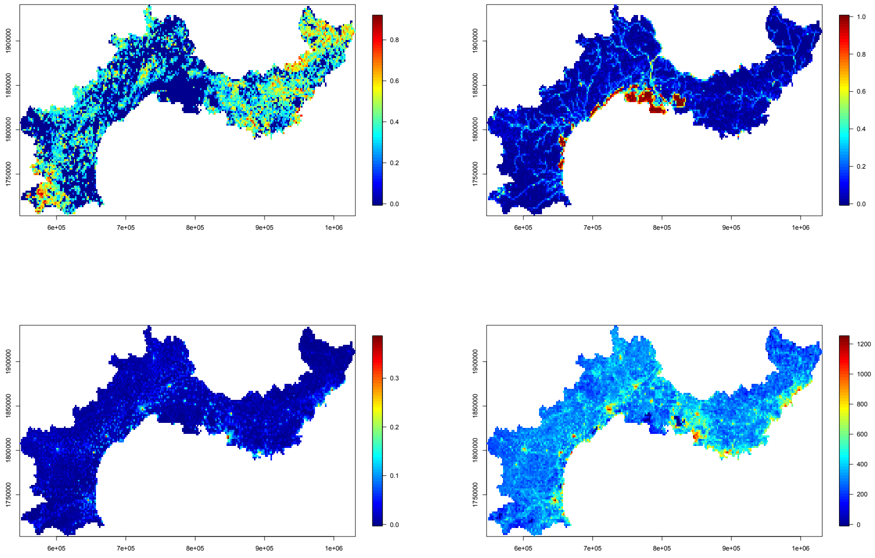

Figure 2: Land cover covariates. Top left: coverage of coniferous trees; right: water coverage; bottom left: building coverage; bottom right: total length of roads (in kilometers).

2.3. Monthly weather data

Weather plays an important role for forest fire outbreaks. Often, favorable conditions such as high temperature and low precipitation leading to very dry soil and vegetation build up over several months and increase the probability of wildfire occurrences. We here consider daily temperature and precipitation height data given by the National Oceanic and Atmospheric Administration (NOAA), with units Celsius and inches, respectively. These data have been recorded at 17 weather stations in the French Mediterranean area and contain the minimum, maximum and average temperature, and the total precipitation for each month. To interpolate the local weather condition for each DFCI cell, we apply spatio-temporal kriging to these data. After some preliminary analyses, we decide to implement product-sum variograms Iaco et al. (2001) for the temperature and a space-time separable variogram for the square-root of the precipitation, all with exponential submodels for the spatial and temporal components. The product-sum variogram model is defined as

$$\gamma(h,u) = (k \cdot \mathrm{sill}_t + 1)\,\gamma_s(h) + (k \cdot \mathrm{sill}_s + 1)\,\gamma_t(u) - k\gamma_s(h)\gamma_t(u), \qquad (1)$$

where γ_s and γ_t are spatial and temporal variograms (with sill sill_s and sill_t respectively), and k is a positive parameter. The separable variogram model is

$$\gamma(h,u) = \mathrm{sill}\left(\bar{\gamma}_s(h) + \bar{\gamma}_t(u) - \bar{\gamma}_s(h)\bar{\gamma}_t(u)\right), \qquad (2)$$

where $\bar{\gamma}_s$ and $\bar{\gamma}_t$ are standardized spatial and temporal variograms with separate nugget effects and (joint) sill 1. The overall sill parameter is denoted by “sill”. The models are fitted using the `gstat R` package Pebesma (2004). Table 1 provides the estimated parameters of models (1) and (2) for the four weather variables. For illustration, Figure 3 shows interpolated maximum temperature (Celsius) and precipitation in inches over the French Mediterranean region for the months of February and July 2014.

In the construction of our models in Section 3.1, we aim to separate a seasonal effect of fire intensity (expressed on a monthly scale) from weather-related effects. To better decorrelate the weather observations from seasonal behavior

	Product-sum model with exponential components				
	Spatial component		Temporal component		
Variable	$sill_s$	range	$sill_t$	range	k
TAVG	46.29	60000	99.98	3.97	$1.49\ 10^-8$
TMIN	46.28	60000	149.95	11.73	$1.49\ 10^-8$
TMAX	61.72	60000	61.72	8.75	$1.49\ 10^-8$

	Separable model with exponential component				
	Spatial component		Temporal component		
Variable	(nugget, $sill_s$)	range	(nugget, $sill_t$)	range	$sill$
PRCP	(0.562, 0.438)	60000	(0.562, 0.438)	5.33	0.019

Table 1: Spatio-temporal variogram models fitted.

and to focus on the influence of the weather anomalies with respect to typical climatic conditions, we also calculate averages of each of the 12 months taken over the full study area and period, and we then subtract these monthly averages from the monthly kriging interpolations to obtain our final weather (anomaly) covariates. The estimated global average monthly effects, used for calculating monthly weather anomalies, are shown in Figure 4. Further analyses have shown that the three monthly temperature variables TAVG, TMIN and TMAX are highly correlated, with linear correlation coefficients larger than 0.99 for kriging predictions and larger than 0.97 for anomalies. Therefore, we avoid identifiability issues in our model by keeping only the anomalies of TAVG as a covariate in the models proposed in Section 3.1. Moreover, to appropriately capture the influence of precipitation on wildfire occurrences, we use two covariates: anomalies of monthly precipitation, and anomalies in the square root of monthly precipitation. The latter has different structure, e.g. higher amplitudes, and the combination of both variables allows to capture certain nonlinearities in the precipitation effect, even if we estimate only linear regression coefficients in our models.

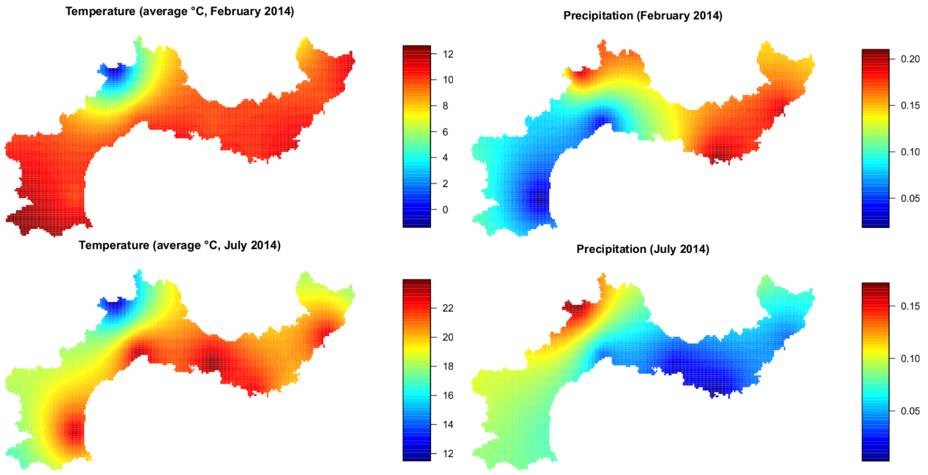
Temperature (average °C, February 2014)
12
10
8
6
4
2
0
Precipitation (February 2014)
0.20
0.15
0.10
0.05
Temperature (average °C, July 2014)
22
20
18
16
14
12
Precipitation (July 2014)
0.15
0.10
0.05

Figure 3: Examples of monthly weather covariates obtained by kriging interpolation. Left: temperature; right: precipitation.

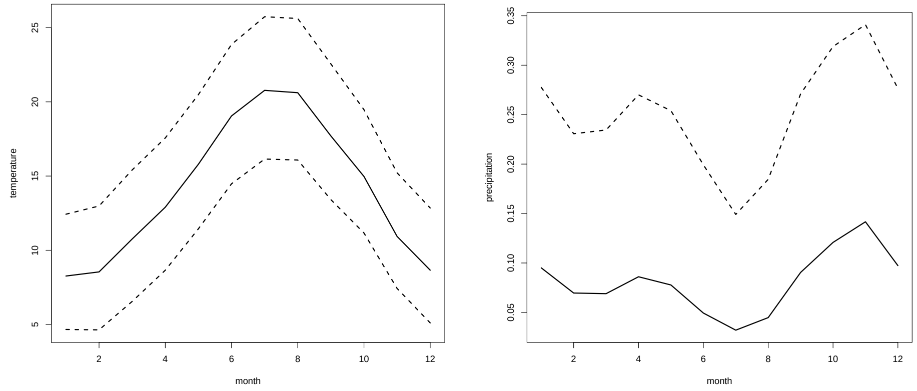

Figure 4: Seasonal (monthly) trends in kriged weather data. Left: TAVG (solid curve), and TMIN, TMAX (dashed curves). right: precipitation (solid curve); square root of precipitation (dashed curve).

3. Spatio-temporal log-Gaussian Cox process models

Conceptually, we can think of our model as defined over continuous space and time. In practice, we here choose a discretization of space (DFCI grid cells) and of time (months), such that the random intensity $\Lambda(s,t)$ does not vary within one DFCI cell during one month when estimating the model. The center of the DFCI grid cell is used as a representative coordinate when working with random spatial effects.

3.1. Model structure

The log-intensity of the point process in our model has the following additive structure, including 30 covariates z_j^{land} related to land use and 3 covariates $\hat{z}_j^{\text{clim}}$ related to weather conditions:

$$\log \Lambda(s,t) = \beta_0 + \beta^{\text{time}}\tilde{t} + \sum_{j=1}^{30} \beta_j^{\text{land}} z_j^{\text{land}}(s) + \sum_{j=1}^{3} \beta_j^{\text{clim}} \hat{z}_j^{\text{clim}}(s,t) \tag{3}$$

$$+ f(\text{month}(t)) + W(s, a(t)) \tag{4}$$

where the first line shows fixed effects with coefficients $\beta_{\cdot}^{\cdot}$ to estimate, and the second line shows random effects. A Gaussian space-time random effect $W(s,a(t))$ is defined at the level of the year $a(t)$ associated with the month t (overall 282 months), i.e. $a(t) \in \{1995, \dots, 2018\}$. Moreover, based on a transformation $\tilde{t} = \frac{t - t_{\min}}{t_{\max} - t_{\min}}$ of the observation period to the interval $[0,1]$, a seasonal effect $f(\text{month}(t))$ is defined at monthly resolution with 12 levels. We here use the hat-notation $\hat{z}_j^{\text{clim}}$ to underline that climate covariates have been estimated beforehand through kriging of observations on an irregular grid of 17 weather stations. Therefore, the influence of the kriging uncertainty on landslide intensity predictions could be studied in practice, although we will not pursue this idea here.

For each of the spatial fields $W(s,a(t))$ for years $a(t)$, we use the well-known SPDE approximation of the Matérn covariance function based on a triangulation of space and with regularity parameter fixed to 1 (see Lindgren et al., 2011; Lindgren & Rue, 2015; Krainski et al., 2018, for theory and practice of the SPDE approach). The grid of DFCI cells and the triangulation mesh constructed for the spatial effect are presented in Figure 5. For spatio-temporal structure in $W(s, a(t))$, we consider four choices:

1. no effect (i.e., $W(s,t) \equiv 0$), or
2. perfect temporal dependence (i.e., a single spatial random effect $W(s, a(t)) \equiv W(s, 1)$), or
3. independent spatial fields, or
4. time-stationary autoregression defined as follows:

$$W(s, a(t)) = \rho W(s, a(t) - 1) + \sqrt{1 - \rho^2}\varepsilon_{a(t)}(s), \quad \rho \in (-1, 1), \qquad (5)$$

where $\varepsilon_{a(t)}(s)$ are the spatial Matérn-SPDE innovation fields.

By adding temporal autocorrelation in model 4, we can capture the persistence of spatial effects through time, such as subregions that tend to be systematically stronger affected by forest fires than others. Moreover, the temporal dependence could help to appropriately correct relatively smooth nonlinearities with respect to the linear time trend captured through the coefficient β^{time} (if any of such nonlinearities exist). We will estimate the following four hyperparameters: the Matérn range parameter (equal to the approximate spatial distance at which correlation 0.1 is reached); the Matérn variance; the autoregression coefficient ρ (if part of the model); a smoothness parameter for the seasonal effect $f(\text{month}(t))$.

4. Bayesian inference using INLA

4.1. Formulation as a Bayesian regression problem for INLA

The spatial-temporal resolution of our model corresponds to DFCI grid cell centers s_i, $i = 1, \dots, 9562$ and the months t_j of the 24-year study period (with only 6 observed months in 2018), $j = 23 \times 12 + 6 = 282$. Consequently, we can consider the counts of fire occurrences $N_{ij} \in \{0, 1, 2, \dots\}$ for each space-time cell

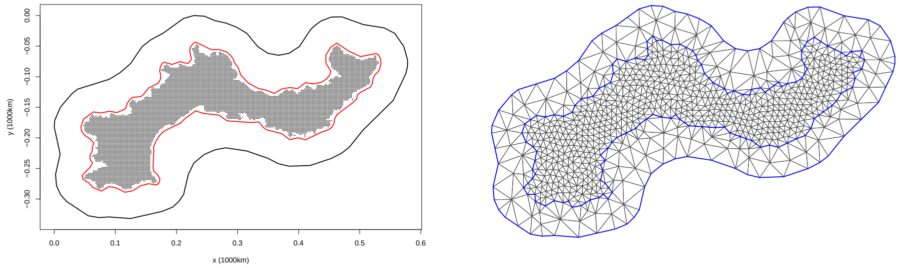

Figure 5: Triangulation mesh. Left: DFCI grid cells and internal and external boundary for the mesh construction (distance units are in 1, 000km). Right: constructed mesh with 816 nodes.

in the Cartesian grid of (s_i, t_j), $i = 1, \ldots, 9562$, $j = 1, \ldots, 282$, and reformulate the model as

$$N_{ij} \mid \Lambda(s_i, t_j) \sim \text{Poisson}(4\Lambda(s_i, t_j)), \quad i = 1, \ldots, 9562, \, j = 1, \ldots, 282, \tag{6}$$

where the constant 4 means that we use a spatial unit of km^2 for Λ (note that one DFCI cell covers $4km^2$). Therefore, the log-linear structure of $\Lambda(s, t)$ with random effects in (3) necessitates the estimation of a generalized additive mixed model. In our dataset, more than 99% of observed counts N_{ij} are 0, and the overall number of observed wildfires is 23, 309. If wildfires occur, the occurrence is single (i.e., $N_{ij} = 1$) in more than 91% of cases, and less than 0.25% of positive wildfire counts are larger than 4; only 2 counts are larger than 9, and the maximum observed count is 12.

The Integrated Nested Laplace Approximation (Rue et al., 2009) and its implementation `R-INLA` in the `R` statistical software (see Rue et al., 2017; Opitz, 2017, for instance) is a fast and accurate inference tool for high-dimensional generalized additive models where conditional independence of the response distribution arises with respect to latent Gaussian processes, as it is the case in our Poisson regression model with log-link function in (6). In the calculation of posterior estimations, it uses an astute combination of analytical Laplace approximations to perform numerical integration with respect to the latent Gaussian variables, and of numerical discretization to approximate the integrals with respect to the hyperparameters. Specifically, using INLA to fit log-Gaussian Cox process models with complex nonlinear effects of space, time and covariates in the intensity function Λ has become common practice in recent years (Illian et al., 2012; Serra et al., 2014; Gómez-Rubio et al., 2015; Lombardo et al., 2018, for instance).

4.2. Choice of prior distributions

For fixed effect coefficients $\beta_{\cdot}^{\cdot}$, we use independent Gaussian prior distributions centered at 0 with fixed precision 0.1 (i.e., variance 10). For the seasonal effect $f(\text{month}(t))$ defined over $1, 2, \ldots, 12$, we fix a periodic first-order random walk prior such that months 12 and 1 are linked.

We fix moderately informative prior distributions for hyperparameters, which helps stabilizing the estimation procedure by penalizing excessively complex models with unstructured random effects and by facilitating convergence of Laplace approximations. In particular, we make systematic use of penalized complexity (PC) priors (Simpson et al., 2017), which penalize the distance of model components with respect to a relatively simple baseline specification. We refer to Fuglstad et al. (2018) for the derivation of PC priors for the Matérn covariance where the baseline is a deterministic field with value 0 everywhere. Effectively, we fix prior distributions such that the prior probability of observing a covariance range below $50km$ and of observing a variance larger than 1 is 50% in each case. Next, we refer the reader to Sorbye & Rue (2014) for the definition of prior distributions for the marginal variance of the seasonal first-order random walk $f(\text{month}(t))$; here we fix priors that set the probability of having marginal variance larger than 0.25 to 50%. Finally, for the temporal autoregression coefficient ρ in (5), we consider the value 0 as baseline model, such that a priori no strong temporal persistence of spatial wildfire occurrence patterns is assumed, and the prior distribution puts 50% of its mass to absolute values $|\rho| > 0.5$.

4.3. Subsampling of intra-year effects

When estimating the Bayesian regression model (6) with INLA, the response vector of counts has approximately $2,7$ million entries, which has turned out to lead to excessively high memory requirements and to numerical instabilities when running INLA, even on machines with $128Gb$ or more of memory. Therefore, we have devised a subsampling procedure as follows. Positive counts $N_{ij} > 0$ and their covariate configurations are kept in the model without modification. Next, for each year and each DFCI grid cell i_0, we consider the months without fire occurrences. If there are $1 \leq k \leq 12$ of such months, we sample at random one of these months j_0 and keep only the value $N_{i_0 j_0} = 0$ and the covariate configuration of this month, while we discard the $k-1$ other months from the dataset. Moreover, in the regression model (6), we set $N_{i_0 j_0} \mid \Lambda(s_{i_0}, t_{j_0}) \sim \text{Poisson}(k\Lambda(s_{i_0}, t_{j_0}))$ to account for the $k-1$ removed response values. This subsampling scheme reduces the length of the response vector to approximately $250,000$, and has led to a stable estimation procedure with `R-INLA`. We can here argue that the loss of information from the original data is limited since the sampled grid cells i_0 still provide a good coverage of the study region for each month, with very similar weather conditions between close DFCI grid cells. Moreover, the set of covariate configurations associated to observations with no fire occurrence, as determined through the subsampling procedure, remains very large and can still be considered as representative for the set of original covariate configurations.

5. Results

For the results reported in the following, we focus on the most “complex” model 4 presented in Section 3.1, where the yearly spatial random effects are linked through temporal autocorrelation. Estimations of fixed covariate effects and of the seasonal effect with monthly resolution have been very similar in terms of magnitude and significance between the models 1 to 4 with different specifications of the space-time Matérn random effect. For a quantitative assessment of relative improvement in goodness-of-fit of model 4 with respect to the models 1 to 3, we report the difference in marginal likelihood values between models 1 to 3 with respect to model 4; these values are 4.2×10^3 (model 1), 3.0×10^3 (model 2) and 1.2×10^3 (model 3). Therefore, we see that each of the three modeling extensions (single spatial random effect; independently replicated spatial random effect; spatial random effect with temporal autoregression) with respect to the baseline model (fixed effects, seasonal monthly effect) leads to a substantial augmentation of marginal likelihood by more than $1,000$.

Regarding the results of model 4, we first study the posterior mean estimates of the four hyperparameters (with estimated standard errors indicated in parentheses). The spatial range is estimated at $20km(1km)$, such that fires tend to ignite in a spatially clustered way, although at relatively small scales. In the construction of the spatial triangulation for discretizing the Matérn SPDE field, we have set a maximum edge length of $12,5km$; therefore, we can suppose that our model can still differentiate between the situation of no spatial dependence and weak but significant spatial dependence within the estimated range of $20km$. The estimate of the temporal autocorrelation coefficient ρ in (5) is 0.89(0.009), such that spatial clustering patterns seem to persist quite strongly over consecutive years. Finally, the precision (i.e., 1/variance) of the seasonal effect with monthly resolution is estimated at 2.4(1.0).

Moreover, we underline a significant negative linear time trend in the log-intensity; see Table 2. Indeed, it corresponds to an almost 40% drop in the point process intensity when comparing the beginning and the end of the 24-year study period. In the context of a warming climate with a tendency towards higher positive temperature anomalies and more arid climate in the Mediterranean basin, this contrary effect can be interpreted as a consequence of the combination of increased vigilance and a set preventive measures coordinated by the competent local authorities, firefighters and forest wardens. Nevertheless, observing extremely high yearly fire counts is not excluded for recent years; for instance, we observe very large numbers of ildfires in 2017.

In the following subsection, we present detailed results and interpretations for the significant fixed effects with respect to the other 33 covariates.

5.1. Influence of season, weather and interfaces of forest to human activity

Recall that our model 4 has monthly temporal resolution over the spatial DFCI grid, with covariates for weather anomalies obtained through spatiotemporal kriging of temperature and precipitation. Land cover covariates at the DFCI grid resolution are included through means and standard deviations of high-resolution land cover data. In particular, including the variability of the coverage inside DFCI cells allows us to identify the effect of the heterogeneity of the land cover distribution on wildfire outbreak. The standard deviation can also be interpreted as a measure that approximately quantifies the length of the interface between a given land cover type and the other types. Moreover, to study wildland-to-urban interfaces, we have introduced two additional synthetic covariates, one for joint forest and building coverage, the other for joint forest cover and road length.

Table 2 presents the model coefficients significantly different from 0 associated with each covariate, ranked in descending order of their level of significance. In the following analysis,we refer to the rank of the level of significance by numbers in parentheses, (1) representing the highest significance rank and (23) the weakest one. Model coefficients higher (respectively lower) than 0 indicate that a high covariate value tends to lead to an increase (resp. decrease) of the fire occurrence probability.

The land cover covariates expressed through the mean over the DFCI cell which contribute to increase the probability of occurrence of a forest fire are the following: temperature anomaly (2), road length (4), proportion of conifers (14), slope (15), and area of regulated tourist areas (17). On the other hand, high values of the following factors significantly decrease this probability: altitude (1), precipitation anomaly (3), water coverage (5), the building covarage (7), length of secondary roads (10), and length of paths (20).

For the French Mediterranean area, our results precisely quantify the current knowledge about factors that tend to favor or limit the occurrence of a forest fire. We believe that altitude, with the most significant coefficient, already summarizes part of such global information. Indeed, low altitude levels can be found near the coastline where the climate is the most Mediterranean, hot and dry, with highly flammable plant species (conifers) and a strong human presence (buildings, roads, tourism), whereas at higher altitudes temperatures are lower, precipitation is higher, vegetation is less present and less favorable to ignition, and human activity is lower or more strongly supervised in the case of tourism. The total length of roads can be seen as a proxy for human presence, while conifers are very present along the coast in the Provences-Alpes-Côte d'Azur region and reepresent a highly flammable tree species. The slope is an important factor in the mountains and provides complementary information with respect to altitude; indeed, on the Mediterranean coast we can find areas with low elevation but with steep slopes (creeks, valleys) and a lot of tourmism. Slope is known to be a factor of propagation of forest fires, therefore it is logical to find this covariate significantly positive. The significance of the regulated tourist zone covariate might be explained by the fact that, despite the efforts of conservation and prevention of the forested areas, touristic pressure is so high in this region of the South of France that it increases the risk of wildfire occurrences. DFCI areas with a very high proportion of buildings (urban areas) or water naturally have a much lower level of forest fire exposure. According to our model, the presence of many secondary roads and paths tends to limit the occurrence of a forest fire. This characteristic contrasts with an opposite effect observed for the total length of all roads, highlighting the major impact of primary roads.

Next, we analyze the interface effects of forest-to-building and forest-to-paths, both of which are significant. The forest-building covariate (13) increases the fire occurrence intensity, while the forest-to-paths factor (16) leads to a decrease. Forest-to-building interfaces concentrate the main cause of wildfire outbreaks: human activity in a forest environment. The risk reduction owing to the forest-to-paths factor is more difficult to explain, but may be due to the fact that the presence of small dirt roads and paths does not necessarily coincide with fire-hazard-prone human activities.

We now study the coefficients of the standard deviation covariates that are significantly different from 0. A first group of variables increases the probability of fire occurrence: secondary road length (6), forest cover (9), path length (11), cover (12), shrubland (19), conifers (21) and moorland (23); while a second group of variables are associated to a decrease of wildfire intensity: road length (8), altitude (18). Overall, these effects are in line with our general understanding that locally heterogeneous environments, where human activity often coincides with presence of combustible material, favor the outbreak of wildfires. The effects of the different road types are not always easy to disentangle, but our model shows that allowing for the interplay of average road lengths and variances for different types considerably improves the goodness-of-fit.

Finally, we show the residual seasonal effect in Figure 6. An explication of its two-peak structure comes from the vegetation cycle and the flammability of vegetation during different seasons of the year. At the end of winter and the beginning of spring (months 2 to 4 approximately), vegetation is still relatively dry and new plant shoots and leaves only start to appear, such that fires break out relatively easily. Fire outbreak risk then drops during the following spring months with relatively high precipitation and relatively “green” vegetation. Finally, the highest peak arrives during the summer months 7 and 8 where extremely dry conditions, dry vegetation and very high touristic activity coincide.

Overall, our Bayesian modeling approach allows for a global and simultaneous consideration of all these factors in space and time, it leads to a prioritization of such factors, and it provides a methodology that can easily be reproduced or updated for other covariates, other regions or more recent environmental and fire occurrence data.

5.2. Intensity mapping

In Figure 7, we illustrate the posterior mean of the latent log-intensity $\log(\Lambda(s,t))$ estimated from our model 4 by showing maps for the 12 months

covariate	**estimate**	**CI**
altitude (average)	-1.48	[-1.64,-1.33]
temperature anomaly	0.09	[0.08,0.1]
precipitation (square root)	-3.15	[-3.66,-2.65]
road length (average)	2.45	[2,2.91]
water (average coverage)	-1	[-1.21,-0.8]
secondary road length (standard deviation)	2.69	[2.11,3.27]
building cover (average)	-5.21	[-6.71,-3.7]
road length (standard deviation)	-1.87	[-2.45,-1.29]
forest cover (standard deviation)	0.77	[0.49,1.04]
secondary road length (average)	-1.28	[-1.81,-0.76]
path length (standard deviation)	1.49	[0.83,2.15]
building cover (standard deviation)	2.71	[1.38,4.04]
forest cover+building cover	4.53	[2.27,6.79]
coniferous cover (average)	0.36	[0.17,0.55]
slope (average)	1.2	[0.5,1.9]
forest cover+paths	-2.54	[-4.06,-1.02]
protected zone cover (average)	0.14	[0.05,0.22]
altitude (standard deviation)	-1.56	[-2.66,-0.46]
shrubland (standard devation)	0.33	[0.05,0.6]
path length (average)	-0.89	[-1.64,-0.14]
coniferous cover (standard deviation)	0.29	[0.04,0.54]
time	-0.48	[-0.91,-0.05]
moorland (standard deviation)	0.21	[0,0.43]

Table 2: Significant fixed covariate effects at the 95% level (based on the credible intervals indicated in brackets in the right column) for the model 4 in Section 3.1. The effects are presented through their posterior mean, and they are ordered from most significant (in terms of high values of $|\hat{\beta}/\mathrm{sd}(\hat{\beta})|$) to less significant from top to bottom. Non-significant fixed covariate effects are not reported. In the covariate names, "average" indicates the use of average values over covariate pixels for each DFCI cell, while "standard deviation" refers to the use of the standard devation of the covariate values; see Section 2 for details.

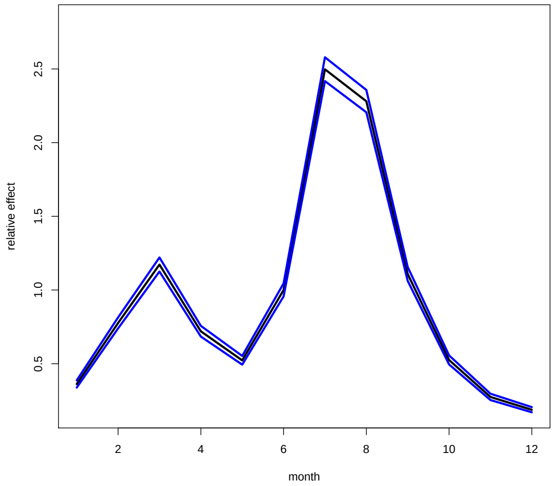

Figure 6: Estimated seasonal effect. The plot shows the posterior mean (black curve) of the odds ratio with the month of June as reference (i.e., with its value scaled to 1). Blue curves show the symmetric 95% credible interval.

of the year 2017, which has shown the highest occurrence numbers $(1, 263)$ in the period after 2003. Clearly, the combination of weather influence and the "residual" seasonal effect leads to substantial differences between the 12 maps, with rather high values for sommer months (7,8,9) and rather low values for winter months (12,1,2).

6. Discussion and conclusion

In the context of a changing climate, wildfires will remain a major challenge for the human societies and natural ecosystems around the Mediterranean. The significant covariate effects revealed by our model for mainland Southern France point out the important contribution of unusually high temperatures and low precipitation amounts to increased fire occurrence risk. Moreover, we observe a strong effect of land use, especially of areas where human activity (agriculture, recreation, tourism) takes place in the presence of forest cover. In particular, DFCI cells with high average forest cover alone were not found to be exposed more strongly to fire risk in a significant way; rather, the presence of buildings

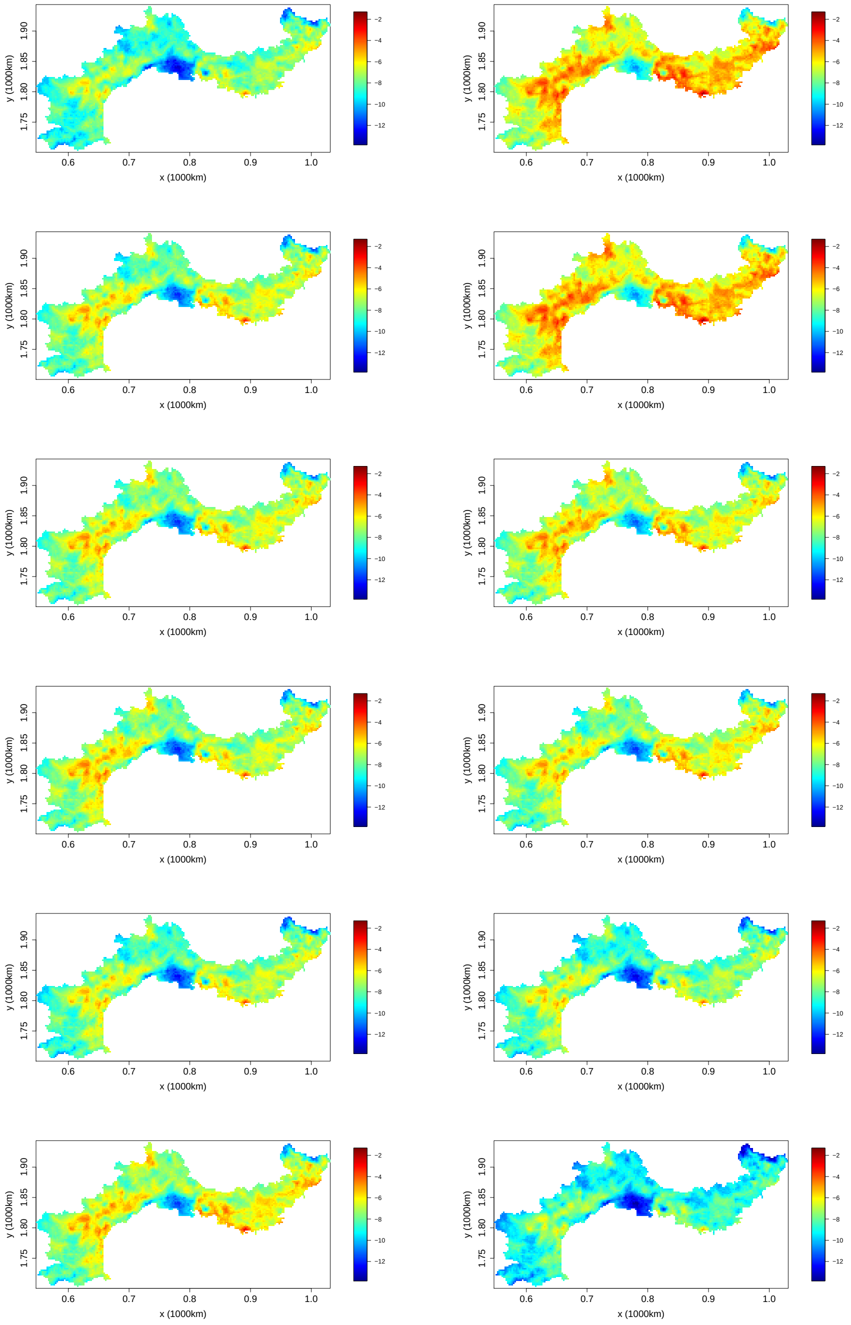

Figure 7: Estimated log-intensity functions $log(\Lambda(s,t))$ for the months of January to December in 2017, with spatial unit $1/(km^2)$. First column: months 1 to 6 from top to bottom. Second column: months 7 to 12 from top to bottom.

and forest together, the dominance of coniferous trees, and a fragmented forest cover have been identified as important factors (among others) contributing to increased fire occurrences.

Our model can be seen as a first important step towards short-term operational forecasting of fire occurrence intensity, for instance at a weekly scale. Indeed, if data reaching until the month of December of the preceding year n have been used for fitting our model, and if weather scenarios are available for the current year, it is possible to provide an intensity forecast for the current year. With `R-INLA`, this solution is simple to implement by specifying the covariates (without missing data) and the count values (missing) for the current year. Statistical inference with even higher-dimensional temporal resolution (such as weeks or days) of spatial and temporal random effects could become feasible by using frequentist inference techniques for estimating fixed effects and random effect hyperparameters (Waagepetersen & Guan, 2009), followed by INLA-based prediction of the intensity function for moderately large temporal subwindows of the observation and prediction period. Indeed, the fully Bayesian approach implemented in this paper is challenging but rewarding since complex fixed and random effects and associated uncertainties are estimated simultaneously in a very precise manner, but certain compromises with respect to the choice of inferential tools may be necessary when going towards even higher dimensions of data and model components.

Finally, we point out that the use of static land use data and of wildfire ignition locations known only at a $2km$ resolution may have introduced some slight biases in the effects estimated through our models. With the increasing availability of high-resolution high-frequency monitoring tools for land use and wildfire activity, such biases may be eliminated in future models. However, such increasingly massive datasets also call for new methodological developments at the interface of spatio-temporal statistical modeling, which provides a sound basis for uncertainty assessment in the context of rare events, and of powerful data mining and learning tools for “big” data.

References

Fuglstad, G.-A., Simpson, D., Lindgren, F., & Rue, H. (2018). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, (pp. 1–8).

Gabriel, E., Opitz, T., & Bonneu, F. (2017). Detecting and modeling multi-scale space-time structures: the case of wildfire occurrences. *Journal of the French Statistical Society (Special Issue on Space-Time Statistics)*, *158*, 86–105.

Genton, M., Butry, D., Gumpertz, M., & Prestemon, J. (2006). Spatio-temporal analysis of wildfire ignitions in the St Johns River water management district, Florida. *International Journal of Wildland Fire*, *15*, 87–97.

Gómez-Rubio, V., Cameletti, M., & Finazzi, F. (2015). Analysis of massive marked point patterns with stochastic partial differential equations. *Spatial Statistics*, *14*, 179–196.

Iaco, S., Myers, D., & Posa, D. (2001). Spacetime analysis using a general productsum model. *Statistics Probability Letters*, *52*, 21 – 28.

Illian, J. B., Sørbye, S. H., & Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics*, (pp. 1499–1530).

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

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., & Rue, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC.

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, *63*.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society (Series B)*, *73*, 423–498.

Lombardo, L., Opitz, T., & Huser, R. (2018). Point process-based modeling of multiple debris flow landslides using INLA: an application to the 2009 Messina disaster. *Stochastic environmental research and risk assessment*, *32*, 2179–2198.

Møller, J., & Diaz-Avalos, C. (2010). Structured spatio-temporal shot-noise cox point process models, with a view to modelling forest fires. *Scandinavian Journal of Statistics*, *37*, 2–25.

Opitz, T. (2017). Latent Gaussian modeling and INLA: A review with focus on space-time applications. *Journal of the French Statistical Society (Special Issue on Space-Time Statistics)*, *158*.

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, *30*, 683 – 691.

Pereira, P., Turkman, K., Amaral-Turkman, M., Sa, A., & Pereira, J. (2013). Quantification of annual wildfire risk: A spatio-temporal point process approach. *Statistica*, *73*, 55–68.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society (Series B)*, *71*, 319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, *4*, 395–421.

Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Diaz-Avalos, C., & Rue, H. (2014). Spatio-temporal log-Gaussian Cox processes for modelling wildfire occurrence: the case of Catalonia, 19942008. *Environmental and Ecological Statistics*, *21*, 531–563.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H. et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, *32*, 1–28.

Sorbye, S. H., & Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, *8*, 39–51.

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

Waagepetersen, R., & Guan, Y. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society (Series B)*, *71*, 685–702.

Xu, H., & Schoenberg, F. (2011). Point process modelling of wildfire hazard in Los Angeles county, California. *The Annals of Applied Statistics*, *5*, 684–704.