Burned area prediction with semiparametric models

Miguel Boubeta[1]*, María José Lombardía[1],

Wenceslao González-Manteiga[2] and Manuel Francisco Marey-Pérez[2]

[1] Universidade da Coruña, Spain,

[2]Universidad de Santiago de Compostela, Spain,

Abstract

Wildfires are one of the main causes of forest destruction, specially in Galicia (north-west of Spain), where the burned area by forest fires in spring and summer is quite high. This work uses two semiparametric time series models to describe and predict the weekly burned area in a year: ARMA modellling after smoothing and smoothing after ARMA modelling. These models can be described as a sum of a parametric component modelled by an autoregressive moving average process and nonparametric one. To estimate the nonparametric component local linear and kernel regression, B-Splines and P-Splines have been considered. The methodology and software have been applied to a real data set of burned area in Galicia in the period $1999-2008$. The burned area in Galicia increases strongly during summer periods. Forest managers are interested in knowing the burned area (in advance) to manage resources more efficiently. The two semiparametric models are analysed and compared against a purely parametric model. In terms of error, the most successful results are provided by the first semiparametric time series model.

Table of contents: Semiparametric models provide a tool to predict the burned area in a particular time. The predictions obtained are competitive. The two bootstrap prediction intervals given are fast computationally. The methodology used can be applicable to other hazard risk.

Key words: Bootstrap, burned area, forest fires, prediction, semiparametric models, time series.

*email: miguel.boubeta@udc.es

Introduction

A wildfire is any uncontrolled fire that affects forested and wooded areas. According to Schmuck et al (2014), in Mediterranean Europe there has been an increase in the number of fires and in the burned area. Figure 1 shows the distribution of annual wildfires every 10 sqkm of wildland. Galicia, the region where this study was carried out, is in the Northwest of the Iberian Peninsula. It is one of the regions most seriously affected by wildfires in Europe.

Recent literature shows that 95% arson fires are related to human activities, see among others, Chuvieco et al (2010); González-Olabarria et al (2011); Juan et al (2012); San Miguel-Ayanz et al (2013); Rodrigues et al (2013); Fuentes-Santos et al (2013) and Rodrigues and de la Riva (2014). The main causes are: 1) Changes in agricultural land use caused by increased activity or abandonment (Moreira et al, 2009; Verde and Zêzere, 2010; Badia et al, 2011; Ricotta et al, 2012; Pausas and Fernández-Muñoz, 2012; Ganteaume and Jappiot, 2013). 2) Increase of forest area (Seidl et al, 2011; Rego et al, 2013). 3) Expansion of shrubland surface (Koutsias et al, 2009; Fernandes et al, 2012; Curt et al, 2013). 4) Forestry operations, garbage dumps, power line accidents (Rodrigues et al, 2014; Cardil and Molina, 2015). 5) Behaviour-related factors (delinquency, smoking,...) (Sebastián-López et al, 2008; Ganteaume and Jappiot, 2013; Reis and Domingues, 2014) and 6) Ineffective forest policies (Rego et al, 2010; Montiel-Molina, 2013; Galiana et al, 2013). Lack of prevention is a fundamental problem. Medium and long-term solutions need to be found considering different aspects.

Numerous methods have been applied to assess the likelihood of fires for each forest, town or region. Generalized linear regression models (GLMs) are used in Viedma et al (2009) to analyse wildfire ignition and in Ordóñez et al (2012) to identify variables that have a significant influence on fire occurrence. Among the most used GLMs, logistic regression models were applied to socioeconomic variables at municipal or provincial level (Vasconcelos et al, 2001; Martínez et al, 2008, 2009; Catry et al, 2009; Chuvieco et al, 2010; Padilla and Vega-García, 2011; Verdú et al, 2012; Vicente López and Crespo Abril, 2012; Chas-Amil et al, 2015; Gudmundsson et al, 2014; Preisler and Westerling, 2007). Mandallaz and Ye (1997) use Poisson regression models to predict forest fires and Boubeta et al (2015) consider an extension of the classical Poisson regression models including forest areas as random effects. Serra et al (2008) used multiple linear regression models to explain the main driving forces of land-cover and land-use, and the relationship with wildfire occurrence in an area of the coast of Catalonia (Spain). Other commonly used methods in the analysis of forest fires are: generalized additive models (Vilar et al, 2010; Loepfe et al, 2012), generalized linear spatial models (Ordóñez et al, 2012) and weights of evidence (WofE) using Bayes probability theorem to measure the spatial association between evidence variable maps and ignition wildfire maps (Stoyan and Penttinen, 2000; Romero-Calcerrada et al, 2010; Rohde et al, 2010; Vega-Orozco et al, 2012; Penman et al, 2015). In the non-parametric context, we cite (Schoenberg et al, 2009) among others, who explore the kernel smoothing and parametric estimation of the relation between incidence and meteorological variables.

In this study, we focused on the application of time-series analysis (Box and Jenkins, 1970), specifically ARIMA processes (Hamilton, 1994) to assess forest fire behaviour. Schroeder (1969) first introduced time-series models between climate antecedents and wildfire variability. Podur et al (2002) employed time-series methods to look for cycles or trends in series of annual fire occurrence and burned area in Ontario (Canada) between 1918 and 2000. Crimmins and Comrie (2004) identified the importance of antecedent climatic conditions for wildfire variability (total burned area and total number of fires) in Arizona. Miller et al (2009) studied fire severity using Autoregressive Integrated Moving Average (ARIMA) in Sierra Nevada (California). They calculate trends in the percentage of burned area at high severity per year. Also for Sierra Nevada, Taylor and Scholl (2012) identify the influence of interannual and interdecadal climate variation and changes in land use on fire regimes using autocorrelation functions (ACF) and auto-regressive moving average (ARMA).

The aim of the present study is to implement and validate a temporal analysis model with one-year predictive capacity for variables related to forest fires. In this case, we take the weekly burned surface by forest fires in Galicia from 1999 until 2008.

In section *Study area and fire database* we describe the study area, the criteria for this choice and its specific features, and the wildfire database. The techniques of temporal process analysis used are introduced in section *Semiparametric models in time series.* Finally, in section *Application to real data* we present the results of wildfire dataset analyses. Conclusions and future lines of investigation outlined are shown in section *Conclusions.*

Materials and methods

Study area and fire database

The forest area of Galicia is $2,060,453$ ha, 69% of the region, which makes it one of the Spanish communities with more woodland. Several authors have studied the characteristics of forest owners and the productive capacity of the region (Marey-Pérez et al, 2006; Marey-Pérez and Rodríguez-Vicente, 2008; Rodríguez-Vicente and Marey-Pérez, 2009, 2010; Marey-Pérez et al, 2012). They concluded that there are many owners with small and very productive plots and a significant presence of collective forest land and no presence of public forest ownership. The first problem of the forestry sector is forest fires: there were $249,387$ wildfires registered since 1968, the year in which forest fire statistics started, until December 2012. These fires swept an area of $1,794,578$ ha, equivalent to 63% of the geographical area of the region.

We have a total of $85,134$ fire events registered in the database for the period 1999-2008, regardless of their size. In addition to the spatial location and the date of occurrence of the ignition points, we use two marks: burned area and cause. We denote by S the burned area and consider small fires when $S < 1$ ha (72.66%), regular fires when $1 \leq S < 25$ ha (25.41%) and large fires when $S \geq 25$ ha (1.90%). Regarding the causes, the main one is arson fires (82.5%), with unknown cause (8.49%) and other causes (9.02%).

The alphanumerical information about the wildfires registered in the study area corresponded to the ignition point coordinates, which were translated to the actual land area with the aid of GIS. Subsequent data quality control confirmed the information about the attributes of the burned area (forest species composition, parish and land use) with existing data from the area in the year of fire. Therefore, Ignition Point UTM (Universal Transverse Mercator) coordinates were available for each fire and other measures of interest attached to these coordinates. These measures are related to burned material, vegetation type, fire behaviour, fire extinction, fire damage and possible fire causes.

Semiparametric models in time series

Although in recent years the development of nonlinear time series models has made great progress, Box-Jenkins methodology is still the most parametric model family employed nowadays. The importance of this class of models is in its generality and its good performance as they provide optimal linear predictions in some contexts. When parametric assumptions about the noise are not flexible, semiparametric models has particular interest since they can capture the variability not processed by parametric models. In this section we propose a prediction interval based on a bootstrap mechanism to study the burned areas in time t, that we denote by Y_t in general notation in the methodology section, and by S_t relative to burned area in the real application section for easy understanding of the sections. We show two alternatives to construct the prediction interval based on nonparametric and parametric components: ARMA modelling after smoothing and smoothing after ARMA modelling.

ARMA modelling after smoothing

Let $(\boldsymbol{Z}_t, Y_t)$, $t = 1, \ldots, T$ a time series, where $\boldsymbol{Z}_t$ is an r-dimensional series and Y_t is an one-dimensional response series. In $\boldsymbol{Z}_t$ one can consider intrinsic or exogenous information of the response time series Y_t. In our case of study, we consider as $\boldsymbol{Z}_t$ the burned area in the previous week, Y_{t-1}. Let be the model

$$Y_t = \varphi(\boldsymbol{Z}_t) + e_t,$$

where e_t has an ARMA(p, q) structure independent of $\boldsymbol{Z}_t$, and

$$\varphi\left(\boldsymbol{z}_t^0\right) = \mathbb{E}\left[Y_t | \boldsymbol{Z}_t = \boldsymbol{z}_t^0\right]$$

is the dynamic regression function. First we estimate the nonparametric part φ consistently, and second we estimate the error e_t. Thus, $\varphi\left(\boldsymbol{z}_t^0\right)$ is approximated nonparametrically from a sample of size n at time t as

$$\widehat{\varphi}_n\left(\boldsymbol{z}_t^0\right) = \sum_{i=1}^{n} W_{ni}\left(\boldsymbol{z}_t^0, \left(\boldsymbol{Z}_1^t, Y_1^t\right), \ldots, \left(\boldsymbol{Z}_n^t, Y_n^t\right)\right) Y_i. \tag{1}$$

The succession of weights $\{W_{ni}\}$ can be obtained by kernel smoothers or splines between others. The choice of random weights $\{W_{ni}\}$ allows us to incorporate various semiparametric models, for example kernel and local linear regressions, B-Splines and P-Splines. There is much literature about these smoothers, see the monographs Hastie and Tibshirani (1990); Wand and Jones (1995); Ruppert et al (2003), among others. To facilitate understanding of the proposed methodology and for simplicity, we consider the kernels weights

$$W_{ni}\left(\boldsymbol{z}_t^0, (\boldsymbol{Z}_1^t, Y_1^t), \ldots, (\boldsymbol{Z}_n^t, Y_n^t)\right) = \frac{K\left(\dfrac{\boldsymbol{z}_t^0 - \boldsymbol{Z}_i^t}{h_n}\right)}{\displaystyle\sum_{j=1}^{n} K\left(\dfrac{\boldsymbol{z}_t^0 - \boldsymbol{Z}_j^t}{h_n}\right)}, \tag{2}$$

being K ($K \geq 0$) the kernel function and h_n the bandwidth parameter.

Then, assuming both time series have been observed up to time $t-k$, the prediction $\widehat{Y}_t$ of Y_t at time t is

$$\widehat{Y}_t = \widehat{\varphi}_n(\boldsymbol{Z}_t) + \widehat{e}_t, \tag{3}$$

where $\widehat{\varphi}_n$ is the nonparametric estimator of φ given in (1), considering for instance weights of type (2), and $\widehat{e}_t$ is the Box-Jenkins prediction to k lags built from the estimated ARMA component of the series

$$\widehat{e}_t = Y_t - \widehat{\varphi}_n(\boldsymbol{Z}_t).$$

Usually, predictions (3) obtained with the semiparametric model are generally better than those obtained by only nonparametric models because the first can capture dependence structures not modelled by the nonparametric component. García Jurado et al (1995) propose a prediction interval for Y_t based on the bootstrap distribution obtained from the ARMA component,

$$\left(\widehat{\varphi}_n(\boldsymbol{Z}_t) + \widehat{q}_t^{*(\alpha/2)}, \widehat{\varphi}_n(\boldsymbol{Z}_t) + \widehat{q}_t^{*(1-\alpha/2)}\right), \tag{4}$$

where $\widehat{q}_t^{*(\alpha/2)}$ and $\widehat{q}_t^{*(1-\alpha/2)}$ denote the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap distribution of $\widehat{e}_t^*$, respectively. Here, we consider an adaptation of the bootstrap method proposed in Cao et al (1997) to get the bootstrap quantiles. The steps are the following:

Bootstrap algorithm for prediction with dependent data

1. Fit one ARMA(p,q) model to the error time series $\{e_t\}_{t=1,\ldots,T}$,

$$e_t = c + \phi_1 e_{t-1} + \phi_2 e_{t-2} + \cdots + \phi_p e_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q},$$

where $\{a_t\}$ is white noise and calculate the estimates of the model parameters $\widehat{c}, \widehat{\phi}_1, \ldots, \widehat{\phi}_p, \widehat{\theta}_1, \ldots, \widehat{\theta}_q$.

2. Construct the empirical distribution of the corrected forward residuals, $F_T^{\widehat{a}'}$, being $\widehat{a}' = \{\widehat{a}_t - \overline{a}\}$ and $\overline{a} = \frac{1}{T-(p+q)} \sum_{t=p+q+1}^{T} \widehat{a}_t$.

3. Repeat B times $(b = 1, \ldots, B)$

   (a) Generate $\widehat{a}_i^*$ with distribution $F_T^{\widehat{a}'}$, $i = T - q + k, \ldots, T + k$.

   (b) For each bootstrap sample, construct the future bootstrap replications to k lags,

$$\widehat{e}^*_{T+k} = \widehat{c} + \widehat{\phi}_1 e_{T+k-1} + \widehat{\phi}_2 e_{T+k-2} + \cdots + \widehat{\phi}_p e_{T+k-p} + \widehat{a}^*_{T+k} + \widehat{\theta}_1 \widehat{a}^*_{T+k-1} + \widehat{\theta}_2 \widehat{a}^*_{T+k-2} + \cdots + \widehat{\theta}_q \widehat{a}^*_{T+k-q}.$$

Given that the residual time series is not observed in practice, the bootstrap algorithm for prediction with dependent data is applied to the estimated residual time series $\{e_t = \widehat{e}_t\}_{t=1,\ldots,T}$.

Smoothing after ARMA modelling

Given an ARMA stationary process $\{Y_t\}_{t=1,\ldots T}$, its optimal linear predictor can be expressed as a linear combination of past values obtained from the ARMA model

$$\mathbb{E}\mathrm{L}\left[Y_t \middle| (Y_u, u < t)\right] = \sum_{i=1}^{\infty} a_i Y_{t-i},$$

but sometimes this optimal linear predictor does not match the optimal predictor,

$$\mathbb{E}\left[Y_t \middle| (Y_u, u < t)\right] = \phi\left(Y_{t-1}, Y_{t-2}, \ldots\right). \tag{5}$$

To estimate the optimal predictor, Dabo Niang et al (2010) propose a semiparametric model using the ARMA residuals as regressors in the nonparametric approach. More precisely, they use the decomposition of the optimal predictor given in (5) as the sum of the optimal linear predictor and the (nonlinear) optimal predictor of innovation process

$$\mathbb{E}\left[Y_t \middle| (Y_u, u < t)\right] = \mathbb{E}\mathrm{L}\left[Y_t \middle| (Y_u, u < t)\right] + \mathbb{E}\left[\varepsilon_t \middle| (\varepsilon_u, u < t)\right],$$

where

$$\varepsilon_t = Y_t - \mathbb{E}\mathrm{L}\left[Y_t \middle| (Y_u, u < t)\right].$$

Since innovations ε_t are unknown, these are replaced by the ARMA model residuals. Therefore, the prediction of Y_t at time t is given by

$$\widehat{Y}_t = \widehat{Y}_t^L + \widehat{\varphi}\left(\widehat{\varepsilon}_u, u < t\right), \tag{6}$$

where $\widehat{Y}_t^L$ denotes the ARMA prediction at time t, $\widehat{\varepsilon}_u$ the estimated ARMA residual at time u and $\widehat{\varphi}$ the nonparametric estimator of $\varphi\left(x\right) = \mathbb{E}\left[\varepsilon_t\right|\left(\varepsilon_u,\, u < t\right) = x]$. In this case, the error structure is predicted nonparametrically. In this framework, we propose a prediction interval for Y_{t+1} following the idea of (4), i.e. adjustment nonparametric plus parametric bootstrap:

$$\left(\widehat{q}_t^{*(\alpha/2)} + \widehat{\varphi}_n(\varepsilon_t), \widehat{q}_t^{*(1-\alpha/2)} + \widehat{\varphi}_n(\varepsilon_t)\right). \tag{7}$$

Where $\widehat{q}_t^{*(\alpha/2)}$ and $\widehat{q}_t^{*(1-\alpha/2)}$ denote the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap distribution of $\widehat{Y}_{t+1}^*$, respectively. This bootstrap distribution is obtained by using the previous bootstrap algorithm for prediction with dependent data in Section *ARMA modelling after smoothing.* In this case we replace the time series $\{e_t\}$ by $\{Y_t\}$.

Data analysis was performed using the statistical software R 3.1.1, using the *mgcv* package to obtain the B-Splines and P-Splines estimators, and *np* package for kernel and linear local smoothers.

Application to real data

A reliable prediction of the weekly burned area in forest fires of Galicia has vital importance to forest managers, as this allows us to anticipate in time and enable prevention systems for fire fighting. To predict this variable, the semiparametric models seen above are used. Figure 2 shows the sequential graph of weekly burned area in Galicia for $1999 - 2008$. We denote by S_t the burned areas at time t. Its behaviour is approximately constant, incorporating sharp increases in summers mainly.

To set the model we consider the training sample $\left(S_1, \ldots S_T\right)$, with $T = 468$, and we reserve the remaining portion (year 2008) for further validation. The optimal nonparametric estimation chosen to incorporate into the semiparametric models will be chosen among: kernel and local linear regressions, B-Splines and P-Splines. To do this, we use the following steps:

General algorithm for the nonparametric component selection

i) Given the training sample with size T, $\{X_1,\dots,X_T\}$, we consider the historical matrix generated by this sample,

$$\mathrm{HM}=\{(X_t,X_{t+1})\},\ t=1,\dots,T-1.$$

ii) From the above historical matrix, we randomly selected 75% of its observations (historical training matrix), leaving the remaining 25% for subsequent validation (historical validation matrix).

iii) With the historical training matrix we construct the corresponding nonparametric estimates of the regression function.

iv) We calculate the empirical MSE in the historical validation matrix,

$$MSE=\frac{1}{K}\sum_{i=1}^{K}\left(X_i-\widehat{X}_i\right)^2,$$

where K is the dimension of this matrix.

v) We repeat steps $ii)-iv)$ M = 1000 times.

We choose the nonparametric regression model that provides a lower empirical MSE for more times. This algorithm will be used to choose the nonparametric component in both semiparametric models, taking $\{X_t\}=\{S_t\}$ for the first semiparametric model and $\{X_t\}=\{\widehat{\varepsilon}_t\}$ for the second one. In the following sections we study the prediction interval of burned areas at time S_t, for each procedure explained in Section *Semiparametric models in time series.*

ARMA modelling after smoothing

For adjustment of the ARMA modelling after smoothing (hereinafter we will be denoted by SP1), we take a particular expression of (3), where $Y_t=S_{t+1}$ and $\boldsymbol{Z}_t=S_t$. Thus, the model can be rewritten as

$$\widehat{S}_{t+1}=\widehat{\varphi}(S_t)+\widehat{e}_t. \tag{8}$$

Since the historical matrix suggests the presence of heteroscedasticity, we propose a logarithmic transformation to stabilize the variability (see Figure 3). Thus, the regression function is

$$\varphi(x) = \mathbb{E}\left[log(S_{t+1}+1) \mid log(S_t+1) = x\right],$$

that is estimated by kernel and local linear regressions, B-Splines and P-Splines. In the first two cases, we take the Gaussian density function as kernel function and the bandwidth parameter is selected using cross validation (CV). For both B-Splines and P-Splines, degrees of freedom are chosen by GCV criterion.

Table 1 presents the number of times in the general algorithm for the nonparametric component selection described on previous page that each nonparametric estimator has a lower empirical MSE. The results show that the SP1 with P-Splines provides a lower MSE in 486 of M = 1000 times.

Figure 4 shows the residual series obtained applying P-Splines. For modelling this series, we consider the ARMA$(p,q) \times (P,Q)_s$ class, that is a combination of $AR(p) \times AR(P)_s$ and $MA(q) \times MA(Q)_s$ with stationary patterns in s, and we select the proposed model by Bayesian Information Criteria (BIC) taking a grid of values with maximum orders of 10 for (p,q) and 3 for (P,Q).

Table 2 presents the significant parameter estimates for the optimal model proposed by BIC approach and that has successfully passed the diagnosis test, ARMA$(9,9) \times (2,0)_{50}$.

Figure 5 (left) displays the predictions to one lag of weekly burned area in wildfires of Galicia by model SP1 (8) on a *log* scale. These are reasonably good as predicted values are close to observed real values. Before getting a new prediction, the series is updated by incorporating the observation of the previous week once it is already known and we repeat the process as many times as instants we want to predict. In this case we want to obtain the predictions of 2008, so during 52 weeks. Figure 5 (right) compares the obtained values by the model and the future observed values (validation sample) in the original scale. The predictions given by SP1 capture the dynamics of the observed series, except in week 7 where there is a clear outlier data. It was due to the high concentration of fires in this week, with 474 forest fires when the weekly average for the rest of the year is approximately 43.

In addition, we have incorporated the bootstrap prediction interval (4) on a *log* scale (see Figure 6). We consider the bootstrap algorithm proposed in Section *Semiparametric models in time series* as an alternative to Tombs and Schucany (1990) for being faster computationally and also consistent.

Smoothing after ARMA modelling

We consider a particular expression for the second semiparametric model, denoted by SP2, given in (6),

$$\widehat{S}_{T+1} = \widehat{S}^{L}_{T+1} + \widehat{\varphi}(\widehat{\varepsilon}_{T}), \tag{9}$$

where $\widehat{S}^{L}_{T+1}$ denote the optimal linear predictor (Box-Jenkins) and $\varphi(\widehat{\varepsilon}_T) = \mathbb{E}\left[\varepsilon_{T+1}|\widehat{\varepsilon}_T\right]$. The parametric component of the model is calculated using a *log* transformation

$$\left\{log(S_1+1),\ldots,log(S_T+1)\right\}, \quad \text{with} \;\; T = 468$$

and we apply BIC approach with maximum orders of 3 for p and q. Table 3 presents the parameter estimates of optimal model, $AR(1)$, given by this criterion.

The model residuals $(\widehat{\varepsilon}_1,\ldots,\widehat{\varepsilon}_T)$ are used for calculating the nonparametric component in (9). By reasoning analogous to the first case, we estimate the regression function of ε_{t+1} on ε_t, using the four considered nonparametric estimators (kernel, linear local, B-splines, P-Splines). Table 4 suggests the use of P-Splines as the nonparametric estimator of the regression function, because it provides the best results in 526 of $M = 1000$ times.

Figure 7 shows the predictions for 2008 given by SP2 to one lag on a *log* (left) and original (right) scale. The observations of the series are automatically updated once we know what happened in the previous instant. As occurred with the previous model, the predictions obtained for 2008 are reasonably good, except for the outlier value observed in week 7.

Figure 8 shows the bootstrap prediction interval proposed in (7).

Comparative analysis

Adjusted the two semiparametric models, one can perform a comparative analysis between both approaches for studying which one provides better predictions. For this, we use the mean squared error (MSE), the relative root-MSE (RRMSE), the absolute error (AE) and the relative absolute error (RAE). Furthermore, as reference we take the purely parametric model (Box-Jenkins, BJ), that it has been an $ARIMA(3,0,0) \times (2,1,3)_{52}$.

Table 5 suggests that the first semiparametric model (SP1) provides better results because their errors are smaller in all cases. Thus, for this data, we recommend first to estimate the regression function by P-Splines and then fit the ARMA (p, q) model to the residual series.

Conclusions

Forest fires are a major environmental, economic and social threat. In Southern Europe they have become the main problem for environmental authorities. In the case of Galicia, they have a remarkable impact on certain areas of the autonomous community and pose a challenge for the future of the region. The development of different methodologies, especially those contrasted by the evidence of the data, allow a more efficient organization and planning of fire fighting, which will result in a lower burned area and a lower risk for lives.

In this work two semiparametric models are reviewed for time series which divide the prediction into two components. One is estimated by nonparametric regression techniques, while the other is performed with Box-Jenkins models. The behaviour of these prediction approaches is competitive in comparison with other temporal models, such as nonparametric models or Box-Jenkins methodology. In addition, the SP1 model allows the inclusion in the r-dimensional time series $\boldsymbol{Z}_t$ of all the auxiliary information available to explain the response time series Y_t since this approach starts with a regression model. By contrast, the SP2 model focuses on time series approach and does not allow the inclusion of auxiliary information outside of the studied time series. The predictions obtained by SP1 are better since their errors are lower. In this framework we give two bootstrap prediction intervals, which are easy to implement, faster computationally and consistent.

Both semiparamtric time series models are used taking the historical burned area data in the autonomous community of Galicia. Another interesting alternative approach is to study an integrated model that calculates the number of ignitions and the burned area jointly or just the burned area by subregions. It requires incorporating geographic information of fires. This issue will be addressed in a future study.

Acknowledgements

This work is supported by the grants MTM2014-52876-R, MTM2011-22392 and MTM2013-41383-P of the Spanish “Ministerio de Economía y Competitividad”, by “Xunta de Galicia” CN2012/130 and 07MRU035291PR, by “Ministerio del Medio Ambiente, Rural y Marino” PSE-310000-2009-4 and by COST Action/UE COST-OC-2008-1-2124.

References

Badia A, Serra P, Modugno S (2011) Indentifying dynamics of fire ignition probabilities in two representative mediterranean wildland-urban interface areas. Applied Geography 31:930–940

Boubeta M, Lombardía MJ, Marey-Pérez M, Morales D (2015) Prediction of forest fires occurrences with area-level poisson mixed models. Journal of Environmental Management 154:151–158

Box GEP, Jenkins G (1970) Time series analysis: forecasting and control. Holden Day, San Francisco

Cao R, Febrero-Bande M, González Manteiga W, Prada-Sánchez J, García-Jurado I (1997) Saving computer time in constructing consistent bootstrap prediction intervals for autoregressive processes. Commun Statist Simula 26:961–978

Cardil A, Molina DM (2015) Factors causing victims of wildland fires in Spain (1980-2010). Human and Ecological Risk Assesment 21:67–80

Catry F, Rego F, Bação F, Moreira F (2009) Modeling and mapping wildfire ignition risk in Portugal. International Journal of Wildland Fire 18:921–931

Chas-Amil ML, Prestemon JP, McClean CJ, Touza J (2015) Human-ignited wildfire patterns and responses to policy shifts. Applied Geography 56:164–176

Chuvieco E, Aguado I, Yebra M, Nieto H, Salas J, Martín M, Vilar L, Martínez J, Martin S, Ibarra P, de la Riva J, Baeza J, Rodríguez F, Molina J, MA H, Zamora R (2010) Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. Ecological Modelling 221:46–58

Crimmins M, Comrie A (2004) Interactions between antecedent climate and wildfire variability across south-eastern Arizona. International Journal of Wildland Fire 13:455–466

Curt T, Bourgniet L, Boullion C (2013) Wildfire frequency varies with the size and shape of fuel types in southeastern France: Implications for environmental management. Journal of Environmental Management 117:150–161

Dabo Niang S, Francq C, Zakoian JM (2010) Combining nonparametric and optimal linear time series predictions. Journal of the American Statistical Association 105:1554–1565

Fernandes P, Loureiro C, Magalhães M, Ferreira P, Fernandes M (2012) Fuel age, weather and burn probability in Portugal. International Journal of Wildlan Fire 21:380–384

Fuentes-Santos I, Marey-Pérez M, González-Manteiga W (2013) Forest fire spatial pattern analysis in Galicia (NW Spain). Journal of Environmental Management 128:30–42

Galiana L, Aguilar S, Lázaro A (2013) An assessment of the effects of forest-related policies upon wildland fires in the European Union: Applying the subsidiarity principle. Forest Policy and Economics 29:36–44

Ganteaume A, Jappiot M (2013) What causes large fires in southern France. Forest Ecology and Management 294:76–85

García Jurado I, González Manteiga W, Prada Sánchez JM, Febrero Bande M, Cao R (1995) Predicting using box-jenkins, nonparametric, and bootstrap techniques. Technometrics 37:303–310

González-Olabarria J, Mola-Yudego B, Pukkala T, Palahi M (2011) Using multiscale spatial analysis to assess fire ignition density in Catalonia, Spain. Annals of Forest Science 68:861–871

Gudmundsson L, Rego FC, Rocha M, Seneviratne SI (2014) Predicting above normal wildfire activity in southern Europe as a function of meteorological drought. Environmental Modelling & Software 29:44–50

Hamilton J (1994) Time series analysis. Princeton University Press, New Jersey

Hastie TJ, Tibshirani RJ (1990) Generalized Additive Models. Chapman and Hall

Juan P, Mateu J, Saenz M (2012) Pinpointing spatio-temporal interactions in wildfire patterns. Stoch Environ Res Risk Assess 26:1131–1150

Koutsias N, Arianoutsou M, Kallimanis A, Dimopoulos P (2009) Is there any special pattern of the extreme wildland fires occurred in Greece in the summer of 2007? In 52 nd International Symposium of the International Association for Vegetation Science, Vegetation Processes and Human Impact in a Changing World, Chania, Greece

Loepfe L, Rodrigo A, Lloret F (2012) Two thresholds determine climatic control of forest-fire size in Europe. Biogeosciences Discuss 9:9065–9089

Mandallaz D, Ye R (1997) Prediction of forest fires with Poisson models. Canadian Journal of Forest Research 27:1685–1694

Marey-Pérez M, Rodríguez-Vicente V (2008) Forest transition in northern Spain: Local responses on large-scale programmes of field-afforestation. Land Use Policy 26:139–156

Marey-Pérez M, Rodríguez-Vicente V, Crecente-Maseda R (2006) Using GIS to measure changes in the temporal and spatial dynamics of forestland: experiences from north-west Spain. Forestry 79:409–423

Marey-Pérez M, Rodríguez-Vicente V, Álvarez-López C (2012) Practical application of multivariant analysis techniques to the forest management of active farmers in the northwest of Spain. Small scale forestry DOI 10.1007/s11842-012-9195-1

Martínez J, Chuvieco E, Martín P (2008) Estimation of risk factors of human ignition of fires in Spain by means of logistic regression. Proceedings of the Second International Symposium on Fire Economics, Planning, and Policy: A Global View pp 265–278

Martínez J, Vega-Garcia C, Chuvieco E (2009) Human-caused wildfire risk rating for prevention planning in Spain. Journal of Environmental Management 90:1241–1252

Miller J, Safford H, Crimmins M, Thode A (2009) Quantitative evidence for increasing forest fire severity in the Sierra Nevada and southern Cascade Mountains, California and Nevada, USA. Ecosystems 12:16–32

Montiel-Molina C (2013) Comparative assessment of wildland fire legislation and policies in the European Union: Towards a fire framework directive. Forest Policy and Economics 29:1–6

Moreira F, Vaz P, Catry F, Silva J (2009) Regional variations in wildfire susceptibility of land-cover types in Portugal: implications for landscape management to minimize fire hazard. International Journal of Wildland Fire 18:563–574

Ordóñez C, Saavedra A, Rodríguez-Pérez JR, Castedo-Dorado F, Covián E (2012) Using model-based geostatistics to predict lightning-caused wildfire 29:44–50

Padilla M, Vega-García C (2011) On the comparative importance of fire danger rating indices and their integration with spatial and temporal variables for predicting daily human-caused fire ocurrences in Spain. International Journal of Wildland Fire 20:46–58

Pausas JG, Fernández-Muñoz S (2012) Fire regime changes in the western Mediterranean basin: from fuel-limited to drought-driven fire regime. Climatic Change 110:215–222

Penman TD, Nicholson AE, Bradstock RA, Collins L, Penman SH, Price OF (2015) Reducing the risk of house loss due to wildfires. Environmental Modelling & Software 67:12–25

Podur J, Martell D, Knight K (2002) Statistical quality control analysis of forest fire activity in Canada. Canadian Journal of Forest Research 32:195–205

Preisler HK, Westerling AL (2007) Statistical model for forecasting monthly large wildfire events in western united states. J Appl Meteorol Climatol 46:1020–1030

Rego F, Rigolot E, Fernandes P, Montiel C, Sande-Silva J (2010) Towards integrated fire management. EFI Policy Brief 4

Rego F, Louro G, Constantino L (2013) The impact of changing wildfire regimes on wood availability from Portuguese forests. Forest Policy and Economics 29:56–61

Reis PM, Domingues VM (2014) The choices of the fire - debating socioeconomic determinants of the fires observed at Portuguese municipalities. Forest Policy and Economics 43:29–40

Ricotta C, Guglietta D, Migliozzi A (2012) No evidence of increased fire risk due to agricultural land abandonement in Sardinia (Italy). Natural Hazards and Earth System Sciences 12:1333–1336

Rodrigues M, de la Riva J (2014) An insight into machine-learning algorithms to model human-caused wildfire ocurrence. Environmental Modelling & Softwar 57:192–201

Rodrigues M, San Miguel J, Oliveira S, Moreira F, Camia A (2013) An insight into spatial-temporal trends of fire ignitions and burned areas in the European Mediterranean countries. Journal of Earth Science and Engineering 3:497–505

Rodrigues M, de la Riva J, Fotheringham S (2014) Modeling the spatial variation of the explanatory factors of human-caused wildfires in Spain using geographically weighted logistic regression. Applied Geography 48:52–63

Rodríguez-Vicente V, Marey-Pérez M (2009) Land-use and land-base patterns in non-industrial private forests: Factors affecting forest management in northern Spain. Forest Policy and Economics 11:475–490

Rodríguez-Vicente V, Marey-Pérez M (2010) Analysis of individual private forestry in northern Spain according to economic factors related to management. Journal of Forest Economics 16:269–295

Rohde D, Corcoran J, Chhetri P (2010) Spatial forecasting of residential urban fires: A bayesian approach. Computers, Environment and Urban Systems 34:58–69

Romero-Calcerrada R, Barrio-Parra F, Millington J, Novillo C (2010) Spatial modelling of socioeconomic data to understand patterns of human-caused wildfire ignition risk in the SW of Madrid (central Spain). Ecological Modelling 221:34–45

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press

San Miguel-Ayanz J, Moreno JM, Camia A (2013) Analysis of large fires in European Mediterranean landscapes: Lessons learned and perspectives. Forest Ecology and Management 294:11–22

Schmuck G, San-Miguel-Ayanz J, Camia A, Durrant T, Boca R, Libertà G, Petroliagkis T, Di Leo M, Rodrigues D, Boccacci F, Schulte E (2014) Forest fires in Europe, middle east and north Africa 2013. Tech. rep., Joint report of JRC and Directorate-General Environment

Schoenberg FP, Pompa JL, Chang C (2009) A note on non-parametric and semi-parametric modeling of wildfire hazard in los angeles county, california. Environmental and Ecological Statistics 16 (2):251–269

Schroeder M (1969) Critical fire weather patterns in the conterminous United States. Silver Spring, Office of Meteorological Operations Washington

Sebastián-López A, Salvador-Civil R, Gonzalo-Jimenez J, San Miguel-Ayanz J (2008) Integration of socio-economic and environmental variables for modelling long-term fire danger in southern europe. European Journal of Forest Research 127:149–163

Seidl R, Schelhaas M, Lexer M (2011) Unraveling the drivers of intensifying forest disturbance regimes in Europe. Global Change Biology 17:2842–2852

Serra P, Pons X, Saurí D (2008) Land-cover and land-use change in a Mediterranean landscape: A spatial analysis of driving forces integrating biophysical and human factors. Applied Geography 28:189–209

Stoyan D, Penttinen A (2000) Recent applications of point process methods in forestry statistics. Statistical Science 15:61–78

Taylor A, Scholl A (2012) Climatic and human influences on fire regimes in mixed conifer forests in Yosemite National Park, USA. Forest Ecology and Management 267:114–156

Tombs LA, Schucany WR (1990) Bootstrap prediction intervals for autoregression. Journal of the American Statistical Association 85:486–492

Vasconcelos MJP, Silva S, Tomé M, Alvim M, Pereira JMC (2001) Spatial prediction of fire ignition probabilities: Comparing logistic regression ans nueral network. Photogrametric Engineering & Remote Sensing 67:73–81

Vega-Orozco C, Tonini M, Conedera M, Kanveski M (2012) Cluster recognition in spatial-temporal sequences: the case of forest fires. Geoinformatic 16:653–673

Verde J, Zêzere J (2010) Assessment and validation of wildfire susceptibility and hazard in Portugal. Nat Hazards Earth Syst Sci Natural Hazards and Earth System Sciences 10:485–497

Verdú F, Salas J, Vega-García C (2012) A multivariate analysis of biophysical factors and forest fires in Spain, 1991-2005. International Journal of Wildland Fire 21:498–509

Vicente López FJD, Crespo Abril F (2012) A new wildland fire danger index for a Mediterranean region and some validation aspects. International Journal of Wildland Fire 21(8):1030–1041

Viedma O, Angeler DG, Moreno JM (2009) Landscape structural features control fire size in a Mediterranean forested area of central Spain. International Journal ofWildland Fire 18:575–583

Vilar L, Woolford D, Martell D, Pilar-Martín M (2010) A model for predicting human-caused wildfire occurrence in the region of Madrid, Spain. International Journal of Wildland Fire 19:325 –337

Wand MP, Jones MC (1995) Kernel Smoothing. Monographs on Statistics and Applied Probability. Chapman & Hall

Tables and figures

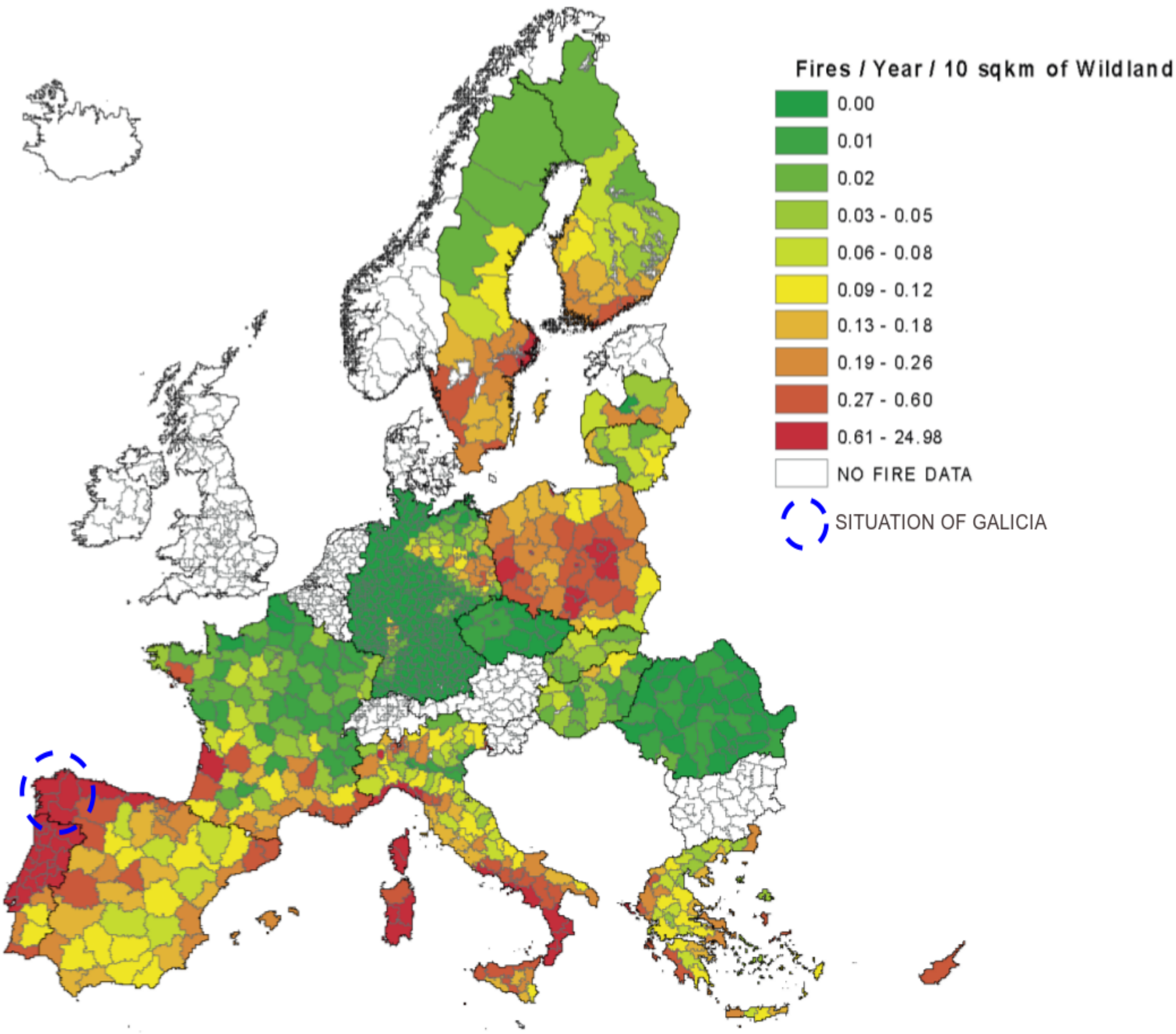

Figure 1: Situation of Galicia in Europe and wildfire annual density for province. Source: Birot, 2009.

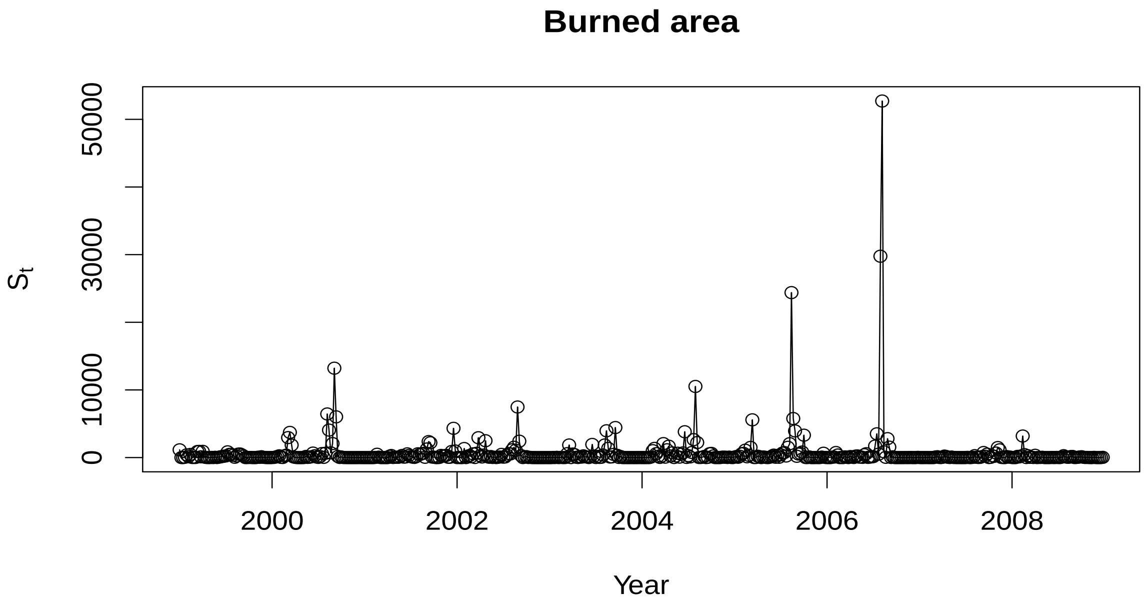

Figure 2: Sequential graph of weekly burned area in Galicia (1999 − 2008).

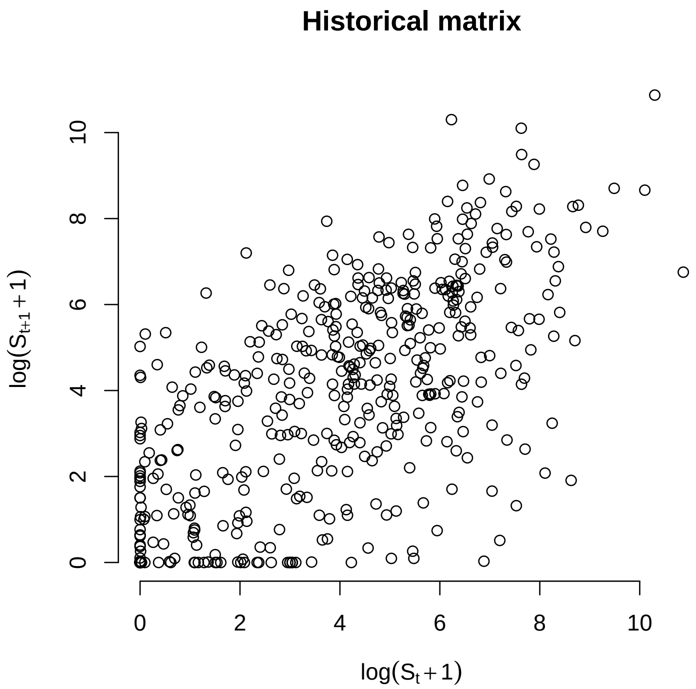

Figure 3: Transformed historical matrix (using log) associated with the time series of weekly burned area in forest fires of Galicia (1999-2007).

Table 1: Accountants of best nonparametric estimator in M = 1000 iterations.

Model	Kernel	Linear local	B-Splines	P-Splines
SP1	209	207	98	486

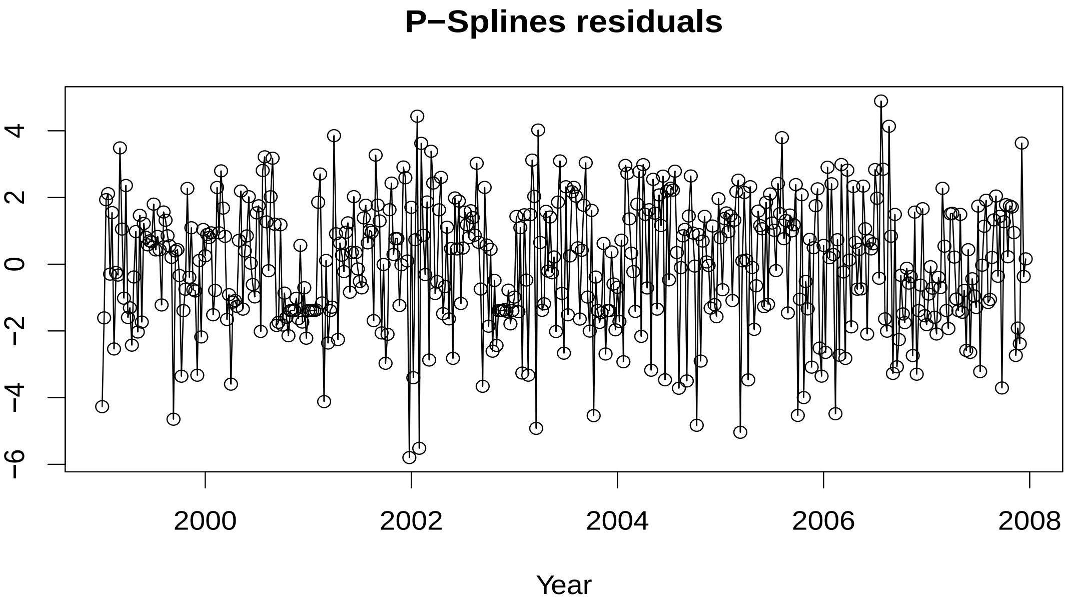

Figure 4: Sequential graph of residuals time series (P-Splines).

Table 2: Significant parameter estimates of ARMA$(9,9) \times (2,0)_{50}$ model estimated by least squares.

	ϕ_1	ϕ_3	ϕ_9	θ_3	θ_5	θ_7	θ_8	θ_9	Φ_1	Φ_2
coef.	-0.094	-0.350	0.494	0.414	0.111	0.193	0.156	-0.512	0.102	0.096
s.e.	0.040	0.069	0.074	0.065	0.039	0.034	0.039	0.063	0.046	0.046

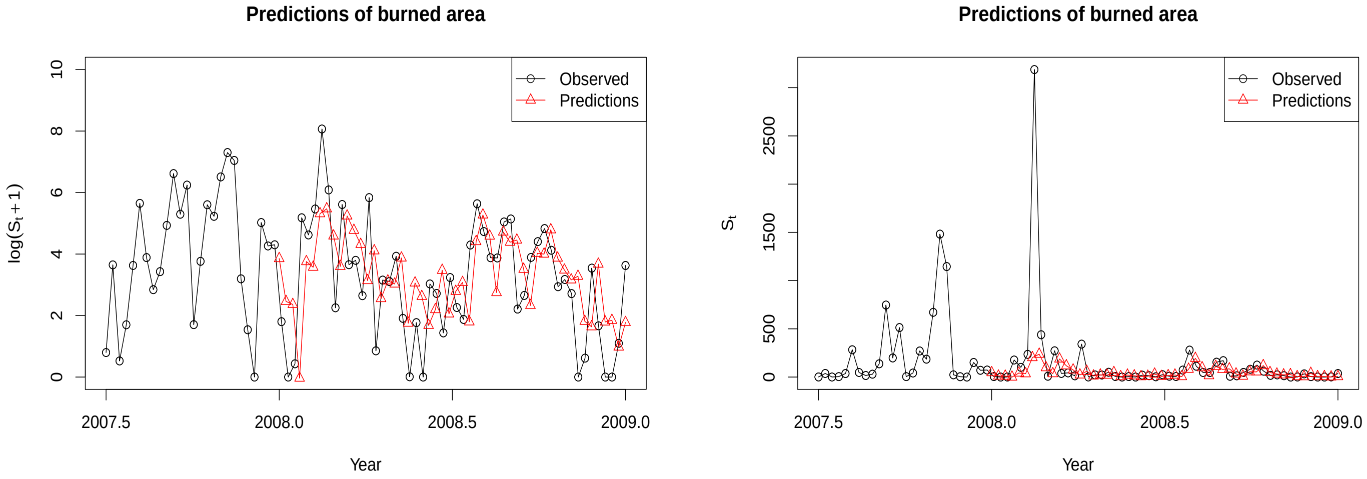

Figure 5: Predicted values by SP1 model (8) on a *log* (left) and original (right) scale.

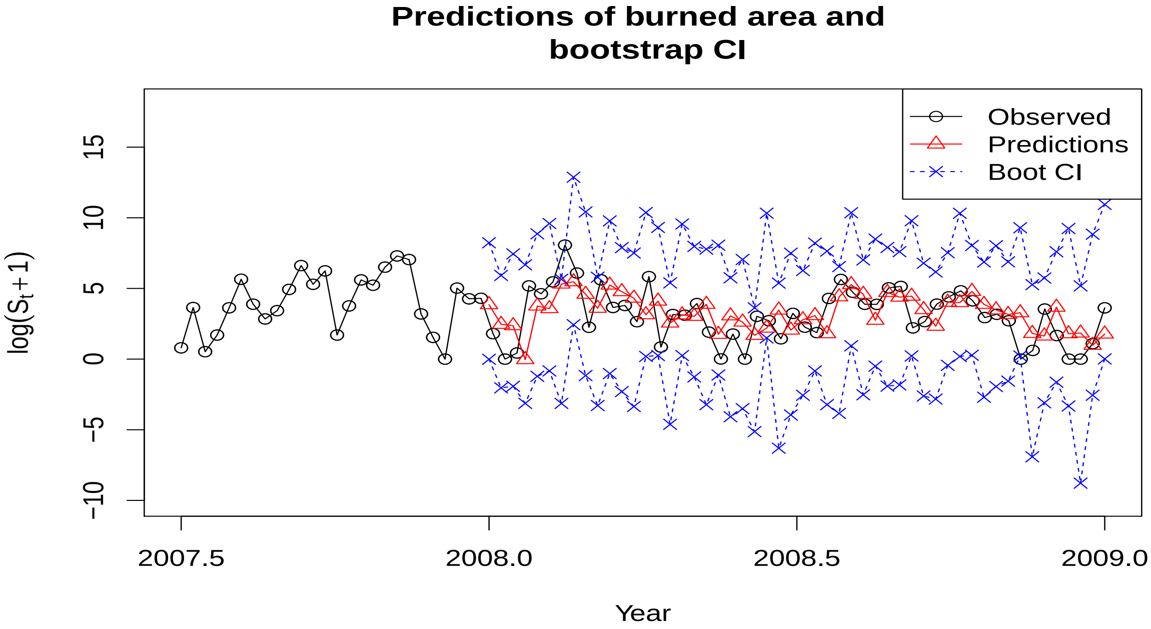

Figure 6: Bootstrap prediction interval for SP1 model ($\alpha = 0.05$).

Table 3: Parameter estimates of AR(1) model by least squares.

	ϕ_1	c
coef.	0.644	3.917
s.e.	0.035	0.247

Table 4: Accountants of best nonparametric estimator in M = 1000 iterations.

Model	Kernel	Linear local	B-Splines	P-Splines
SP2	231	179	64	526

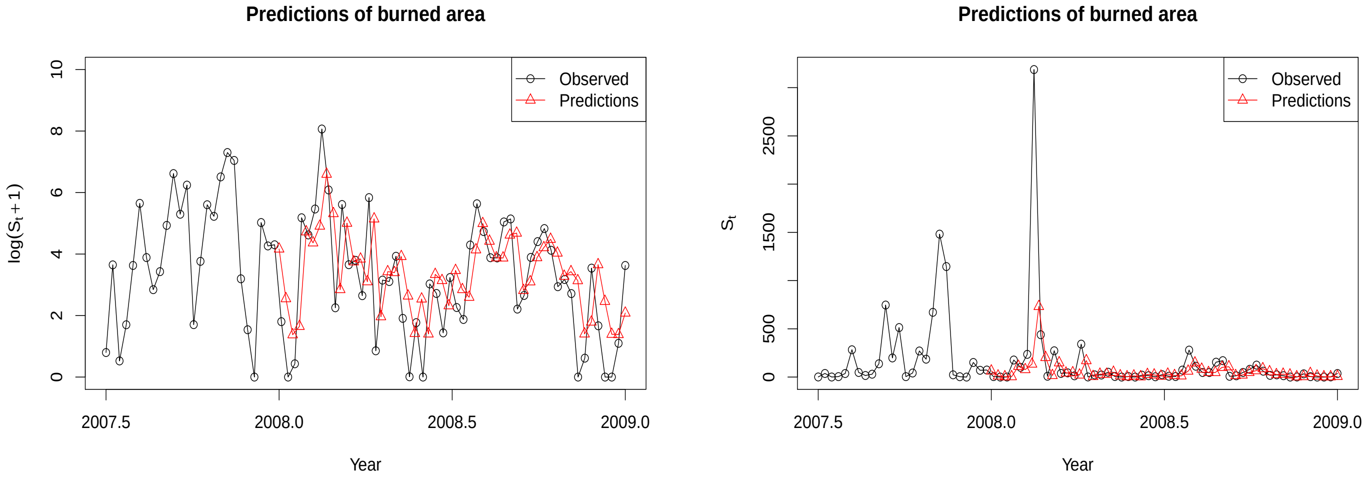

Figure 7: Predicted values by SP2 model (9) on a *log* (left) and original (right) scale.

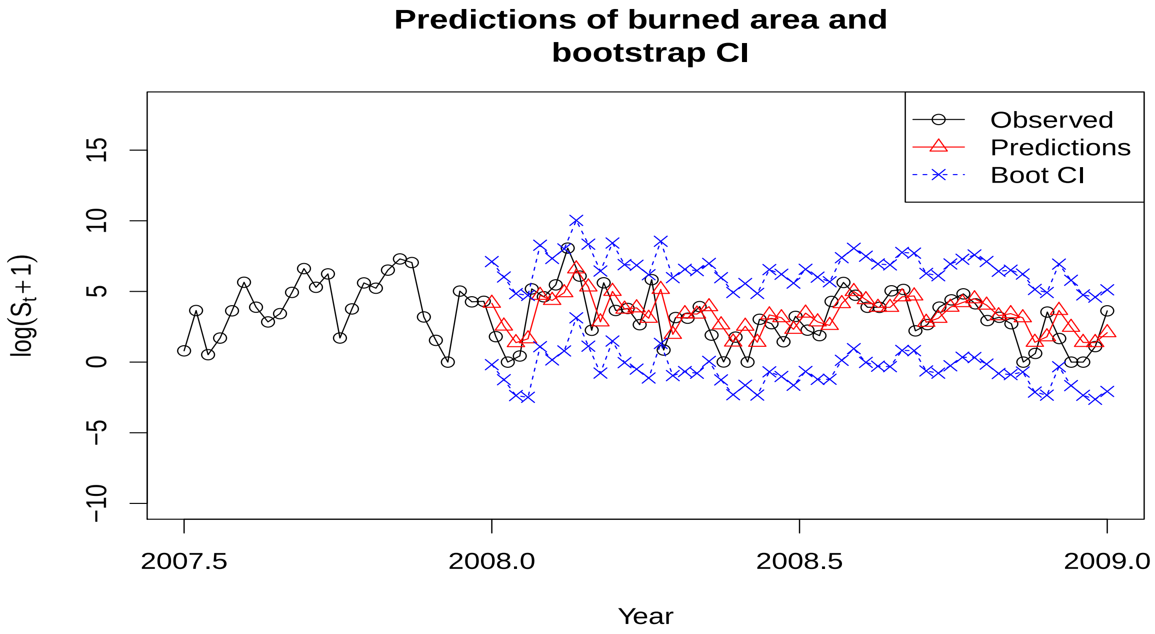

Figure 8: Bootstrap prediction interval for SP2 model ($\alpha = 0.05$).

Table 5: Mean squared error (MSE), relative root-MSE (RRMSE), absolute error (AE) and relative absolute error (RAE).

Model	MSE	RRMSE	AE	RAE
BJ	190913.60	4.22	111.94	4.22
SP1	144796.60	3.92	104.20	3.92
SP2	188610.40	4.70	115.62	4.70