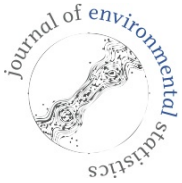

Journal of Environmental Statistics

December 20011, Volume 2, Issue 1. http://www.jenvstat.org

A Spatio-temporal Model for People-Caused Forest Fire Occurrence in the Romeo Malette Forest

Douglas G. Woolford
Mathematics, WLU

David R. Bellhouse
Statistical and Actuarial Sci., UWO

W. John Braun
Statistical and Actuarial Sci., UWO

Charmaine B. Dean
Statistics and Actuarial Sci., SFU

David L. Martell
Forestry, UT

Jessica Sun
Statistical and Actuarial Sci., UWO

Abstract

This paper describes the development and an assessment of a spatio-temporal model for people-caused forest fires in a portion of boreal forest in northeastern Ontario, a central province in Canada. Space and time along with location-specific weather-based fire danger rating indices and anthropogenic effects are included in the modelling we present, which parallels the structure of recent methodology for assessing fire risk using logistic generalized additive models (GAMs) introduced in Brillinger et al. (Institute of Mathematical Statistics Lecture Notes, 2003). In these models, the data consist of observations on a very fine set of space-time cells, where fires are rare and the complete data set is too large to analyze. Consequently, the non-fire observations are sampled. This induces an offset in the additive structure, which we connect to the analysis of case-control studies. The model's fit and estimated partial effects are shown to be sensitive to large reductions in this inclusion probability. We also make comparisons between a model with an additive decomposition of spatial and temporal effects to one with a spatio-temporal interaction, and we investigate the impact of restricting fire-weather and anthropogenic effects to be linear. Our results suggest that, when using logistic GAMs to model our wildland fire occurrence data on this scale, there is no advantage to including space and time interaction effects, and that models with linear terms, which have dominated the fire risk literature, are inadequate.

Keywords: case control studies, goodness of fit, logistic generalized additive model, model assessment, space-time interaction, spline smoothing, stratified sampling, wildland fire ignition.

1. Introduction

The locations, times and number of forest fire ignitions arising from human activity can seem to appear "randomly" in a region over a given period of time. Consequently, a point process, or more specifically, a Poisson process with a covariate-dependent nonhomogeneous spatio-temporal intensity function is likely an appropriate modelling framework for empirical investigations of such ignitions. This connection to Poisson processes was identified early on in the literature by Cunningham and Martell (1973) who constructed a Poisson model for the number of people-caused fires in their roughly 1.8 million ha study area near Sioux Lookout in northwestern Ontario during the summer as a function of fuel moisture.

Yet, logistic generalized linear models dominate the historical literature on fire occurrence modelling, where presence/absence of fires on a landscape is considered. In the central province of Ontario, Canada, the use of logistic regression to model fire risk dates back over 20 years. Martell et al. (1987) constructed season and ignition source specific logistic models for daily people-caused occurrence in northern Ontario. Martell and Belivacqua (1989) extended this by incorporating periodic functions to account for the seasonal variation inherent to forest fire occurrence rates. Recently, a set of site-specific logistic models for the ignition and subsequent detection of lightning-caused forest fires were proposed by Wotton and Martell (2005). The use of logistic models for fire ignitions, rather than Poisson process based models, likely stems from the well known fact that a Bernoulli process is the discrete-time analogue of a Poisson process, and logistic generalized linear modelling techniques are easily implemented in statistical software packages (see e.g., Berman and Turner 1992). Moreover, overdispersion, a common concern when fitting Poisson-based models is not a concern when logistic binary models are employed. However, it is important to note that Turner (2009) illustrated recent advances in software for modelling point patterns that have led to new and interesting ways of visualizing and modelling forest fire ignitions on a landscape.

Brillinger et al. (2003) reminded the fire science community of the connection between wildfire ignition risk and Poisson processes, providing a thorough review of both the underlying spatio-temporal conditional intensity function as well as suggestions on how the corresponding likelihood can be approximated. In particular, they advocated partitioning the space-time domain into a fine set of 1 km $\times$ 1 km $\times$ 1 day "voxels". Then, the likelihood for the underlying spatio-temporal process is approximated by a Bernoulli process on this lattice, whose response variable is an indicator of the presence of a fire event in a voxel. Moreover, by incorporating spline smoothers in an additive model structure, their methodology permits the quantification of spatial and temporal partial effects, components that were not present or as easily incorporated into the earlier linear models. An additional consideration in their modelling framework, is that the subset of non-fire voxels must be sampled in order to obtain a computationally feasible data set; this induces a deterministic offset term into the additive structure when a logit link is employed. However, given that fire ignitions are rare events on a daily, 1 km^2 scale, such a sample provides adequate covariate information for parameter estimation and inference.

Brillinger et al. (2003) focused on obtaining baseline spatial and temporal effect estimates for federal lands in Oregon, U.S.A. Since that initial publication, a series of related articles have appeared. Preisler et al. (2004) extended the first model by incorporating the partial effects of fire-weather variables and proposed a conditional probability framework for estimating the probability of a large fire event given an ignition. Then Brillinger et al. (2006) presented sim-

ilar models for California, assessing the incorporation of random effects. Recently, this work has culminated with Presiler and Westerling (2007) and Preisler et al. (2008) demonstrating how their framework could be employed to produce one month ahead forecasts for large fire events.

The modelling we present builds upon the earlier work discussed above. We incorporate and estimate nonlinear relationships between forest fire risk and spatially referenced anthropogenic variables, demonstrating a clear link between people-caused ignitions and human land-use patterns. Using a variety of visual and diagnostic techniques, we also investigate several sensitivity concerns including the advantages and disadvantages of assuming an additive structure for spatial and temporal effects, the impact of replacing nonlinear fire-weather and anthropogenic covariate effects with linear ones, and the effect of varying the inclusion probability when sampling the zero-fire voxels. Finally, we draw attention to a link between the bias that results from response-based sampling and case-control studies, a connection which does not appear to have been made explicit in the literature. Our application broadens the use of these models, expanding their geographic scope to an international context by constructing a model for a region of boreal forest in Canada.

The next section describes the data and study area. Section 3 provides the necessary background on generalized additive models and spline smoothing, and discusses a connection between the biased sampling we employ and biostatistical case-control studies. Section 4 outlines the modelling framework, discusses model selection, presents and visualizes a spatio-temporal model for people-caused forest fire occurrence in the Romeo Malette forest, and assesses its fit. Our article ends with a brief concluding discussion.

2. The data and study area

We analyze records of people-caused forest fires and weather in a rectangular region encapsulating the Romeo Malette forest. This region is located in the province of Ontario, Canada between approximately $[-82.7072, -80.7037]$ longitude $\times$ $[47.6364, 48.8806]$ latitude as illustrated in Figure 1, and was partitioned into a regular 138×147 grid of 1 km^2 cells for our study. We construct a model for the active "fire season" in this area, namely April 1 to September 30, for the years 1976 through 1999. During this 24-year period there were a total of 890 fires, of which 560 were people-caused. The remaining fires were ignited by lightning. The people-caused fires can be further subdivided by ignition source, with categories and relative frequencies as follows: recreation (35.2%), railway (15.4%), residents (12.6%), forest industry (6.5%), arson (5.4%), non-forest industry (5.2%), and unknown or miscellaneous ignition (19.8%).

The forest fire data set contains both dynamic fire-weather variables, and static spatially-referenced variables. The weather variables we used were the daily solar noon observations of the local temperature, relative humidity, precipitation, and wind speed. These weather observations are used to compute the following daily fire weather variables which are components of the Canadian Forest Fire Weather Index system (Van Wagner 1987), and can be used to help describe relative forest fire danger at its mid-afternoon peak: the fine fuel moisture code (FFMC), which represents the moisture content of dead fine litter fuels on the forest floor; the duff moisture code (DMC), which represents the moisture content in the compacted and decaying dead organic layer below the forest floor; the drought code (DC), a

measure of long-term drought conditions; the initial spread index (ISI), which represents how fast an ignited fire will spread; the build-up index (BUI), which represents approximately how much fuel is available for consumption by a fire; and, the fire-weather index (FWI), which is an index of fire intensity. The Fire Weather Index system is a subsystem of the Canadian Forest Fire Danger Rating system (Natural Resources Canada 2006), which is used to assess forest fire danger in Canada. The static spatially-referenced variables we investigated include longitude and latitude, the population density in each grid cell, and the respective distances from each grid cell to the nearest railway, road and town. In what follows, data from 1976 through 1996 were used for model fitting and data from 1997 through 1999 were reserved for cross-validation. Details regarding the construction of our data set appear in Morgenroth (2003).

3. Methods

3.1. Generalized additive models and penalized smoothing

Generalized additive models (Hastie and Tibshirani 1986) extend the well known generalized linear models (GLMs) by allowing for non-linear covariate effects through the incorporation of additive non-parametric smooth functions. In the univariate case, assume the distribution of the response variable Y_i belongs to the exponential family and that its mean $\mu_i \equiv \mathrm{E}[Y_i]$ is related to explanatory variables $x_{1i},\ x_{2i}, \ldots$ through the use of a "link function" $h(\cdot)$. Letting $\eta_i \equiv h(\mu_i)$, a GAM has the following general structure

$$\eta_i = \mathbf{X}_i\boldsymbol{\beta} + \sum_{m=1}^{M} f_m(x_{mi}) \ , \tag{1}$$

where $f_m(\cdot)$ are smooth functions (commonly referred to as "partial effects") of the covariates x_m; $\mathbf{X}_i$ is the i^{th} row of the design matrix for any fixed linear effects, whose p coefficients are contained in the vector $\boldsymbol{\beta}$, which usually includes an intercept term.

In our work, the smoothers are modelled using linear expansions of basis functions. For example, a univariate smooth function $f(x)$ can be represented as

$$f(x) = \sum_{k=1}^{K} \phi_k b_k(x) \ , \tag{2}$$

where ϕ_k are unknown coefficients for the basis functions $b_k(x)$ over a partition of the range of the covariate x defined by the set of "knots" $k = 1, \ldots, K$. In other words, $f(x)$ can be represented as a linear combination of known functions, and hence, estimating f is equivalent to estimating the coefficients ϕ_k. This linear representation facilitates estimation via the likelihood-based inference used for GLMs.

An additional consideration arises due to the set of knots, because the number and location of these knots can influence the resulting fit. However, an alternative methodology is to use a relatively large number of knots and incorporate penalty component(s) in the likelihood to modulate the amount of smoothing. Given the additive characteristic of the basis expansion,

one commonly penalizes the total amount of "wiggliness" in f, measured by integrating its second derivative, denoted by $f''(\cdot)$. This yields the penalty term

$$\lambda \int [f''(x)]^2 dx \ , \tag{3}$$

where the penalty parameter, λ, is to be estimated. Given λ, the model fitting objective function becomes the minimization of

$$\sum_{i=1}^{n} \left[y_i - \mathbf{X}_i \boldsymbol{\beta} - \sum_{m=1}^{M} f_m(x_{mi}) \right]^2 + \sum_{m=1}^{M} \lambda_m \int [f_m''(x)]^2 dx_m \tag{4}$$

with respect to the fixed linear parameters $\boldsymbol{\beta}$ and the set of basis coefficients $\{\phi_k\}_{k=1,\ldots,K}$. The penalty terms $\{\lambda_m\}_{m=1,\ldots,M}$, are chosen so that the Generalized Cross Validation (GCV) score (Craven and Wahba 1979; Golub et al. 1979) is minimized. The GCV score is essentially a numerically efficient modification of the ordinary cross validation score, where the average squared prediction error is estimated over all data sets where a single observation is omitted when model fitting, and then predicted. Technical details on modelling and GCV estimation for GAMs with multiple smoothing parameters appears in Wood (2000, 2004), a very informative summary of which appears in Wood (2006).

Smooth functions of multiple variables are handled similarly: they can be expressed as a linear combination of basis functions with over/under-fitting controlled by a penalty term. The multivariate partial effects presented herein are modelled using tensor product smooths. This method may be preferred when smoothing interactions of variables that do not have the same units of measurements (Wood 2006, Table 5.2), such as our trivariate smoother of space and time.

A tensor product smooth of two variables, say $f(x_1, x_2)$, can be constructed by modifying a univariate smooth of x_1 in the form of (2). This is achieved by allowing the corresponding basis coefficients for x_1 to vary smoothly over x_2, yielding

$$f(x_1, x_2) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \phi_{k_1 k_2} b_{k_2}(x_2) b_{k_1}(x_1) \ . \tag{5}$$

Notice that the expression above is still a linear expansion in the parameters $\phi_{k_1 k_2}$ and hence estimation remains in the framework of linear models. The generalization to higher dimensions is then straightforward–a trivariate smoother $f(x_1, x_2, x_3)$ would be constructed by allowing the parameters in the above function to fluctuate smoothly over x_3, and so on.

The penalty term for a tensor product smooth is developed by considering the variation of the surface in each of the component dimensions of the smoother. For the bivariate scenario presented in (5), the "wiggliness" of $f(x_1, x_2)$ in the x_1 direction is calculated by summing $(\partial^2 f / \partial x_1^2)^2$ over x_2, and vice-versa in the x_2 direction, leading to the penalty term

$$\int_{x_1, x_2} \lambda_{x_1} \left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + \lambda_{x_2} \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 dx_1 dx_2 \ . \tag{6}$$

For further technical details on tensor product bases and smoothers, including the generalization to higher dimensions and the numerical integration of the penalty terms, the interested reader is directed to Wood (2006, section 4.1.8).

3.2. Biased sampling and a connection to case-control studies

The complete data set of 1 km × 1 km × 1 day space-time voxels would consist of nearly 90 million records. However, fire ignitions are very rare events at this space-time scale. Consequently, a sample of the data corresponding to the zero-fire voxels yields a computationally manageable data set containing sufficient covariate information for model building and inference. We employ a stratified sampling scheme, including data from all voxels with fire ignitions and a simple random sample of ten-percent of the zero-fire voxels. This sampling induces an offset of $\log(1/\pi_{st})$ in the generalized linear/additive model, where π_{st} denotes the inclusion probability for site s at time t. The first use of such a sampling scheme in a logistic model for fire ignitions appears to be Vega Garcia et al. (1995). As previously indicated, it has appeared in a recent series of papers using logistic generalized additive models to assess wildland fire ignition risk (Brillinger et al. 2003, 2006; Preisler et al. 2004).

The logistic model considered here relates fire occurrence to several explanatory variables. More generally, the model relates whether ($y = 1$) or not ($y = 0$) a 1 km^2 × 1 day space/time interval contains a fire event via the equation

$$\mathsf{P}(y = 1|\mathbf{x}) = \frac{\exp\left(\alpha + \mathbf{x}\boldsymbol{\beta} + \sum_{m=1}^{M} f_m(x_{mi})\right)}{1 + \exp\left(\alpha + \mathbf{x}\boldsymbol{\beta} + \sum_{m=1}^{M} f_m(x_{mi})\right)}$$

which is equivalent to the specification (1) where $h(\cdot)$ is the logit function. The model implies that the relative risk for two space-time intervals having two sets of explanatory variables $\mathbf{x}_1$ and $\mathbf{x}_2$ is

$$\frac{\mathsf{P}(y = 1|\mathbf{x}_1)\ \{1 - \mathsf{P}(y = 1|\mathbf{x}_2)\}}{\{1 - \mathsf{P}(y = 1|\mathbf{x}_1)\}\ \mathsf{P}(y = 1|\mathbf{x}_2)} = \exp\big((\mathbf{x}_1 - \mathbf{x}_2)\boldsymbol{\beta}\big)$$

when there are no additive components. In this case α represents the log odds of a fire event for a standard set of regressor variables ($\mathbf{x} = 0$) and $\exp(\beta_k)$ is the change in this risk for a unit change in x_k. When there are additive components, these interpretations still hold, and additionally $\exp\{f_m(x_{m1}) - f_m(x_{m2})\}$ represents the change in risk when the covariate in the m^{th} additive component changes from x_{m1} to x_{m2}. This model is identical to what is termed a *prospective* analysis in medical studies when it is not known in advance whether or when an individual will develop a disease and individuals with varying covariate values (some exposed to a pollutant, others not; some treated, others not) are followed in time and their responses ($y = 1$ indicating disease is developed; $y = 0$ indicating no disease develops) after some period of time is measured.

In contrast, with the case-control approach, subjects are selected on the basis of their disease status, and their history of covariate values (exposures, treatments) determined *retrospectively*. Hence, it is the covariate values which should be regarded as random. In case-control studies, however, it is well known that inferences about relative risk are obtained using the identical logistic model as for *prospective* studies (Breslow and Powers 1978). Let δ denote whether or not an individual is sampled ($\delta = 1$ if sampled, 0 otherwise) and let $\pi_1 = \mathsf{P}(\delta = 1|y = 1)$ and $\pi_0 = \mathsf{P}(\delta = 1|y = 0)$. Typically, π_1 is close to 1 (many cases are included), while π_0 is typically fairly small. Consider the probability that a person is diseased, given that they have

covariate values $\mathbf{x}$, and was sampled for the study; using Bayes theorem we have

$$\begin{aligned}
\mathsf{P}(y=1|\delta=1,\mathbf{x}) &= \frac{\mathsf{P}(\delta=1|y=1,\mathbf{x})\mathsf{P}(y=1|\mathbf{x})}{\mathsf{P}(\delta=1|y=0,\mathbf{x})\mathsf{P}(y=0|\mathbf{x})+\mathsf{P}(\delta=1|y=1,\mathbf{x})\mathsf{P}(y=1|\mathbf{x})} \\
&= \frac{\pi_1 \exp\left(\alpha+\mathbf{x}\boldsymbol{\beta}+\sum_{m=1}^{M} f_m(x_{mi})\right)}{\pi_0+\pi_1 \exp\left(\alpha+\mathbf{x}\boldsymbol{\beta}+\sum_{m=1}^{M} f_m(x_{mi})\right)} \\
&= \frac{\exp\left(\alpha^*+\mathbf{x}\boldsymbol{\beta}+\sum_{m=1}^{M} f_m(x_{mi})\right)}{1+\exp\left(\alpha^*+\mathbf{x}\boldsymbol{\beta}+\sum_{m=1}^{M} f_m(x_{mi})\right)}
\end{aligned}$$

where $\alpha^* = \alpha + \log(\pi_1/\pi_0)$. Note that sampling probabilities depend only on disease status and not on covariate values, and that apart from this intercept term, the covariate effects are identical to the logistic model from the prospective analysis. The analysis of the fire occurrence data is identical to the formulation above with π_1 representing the probability of including space-time intervals with fire events, which here is 1, and π_0 representing the probability of including space-time intervals without fire events.

Extensions to this parallel with case-control studies offers a design perspective. For example, if we were specifically interested in the effects of the distance to a railway a matched case-control study may be of interest, since these have been shown to provide more efficient designs. In matched case-control studies each case ($y = 1$) is matched to one or more, usually several, controls ($y = 0$) with similar covariate values on the matching variables. In such studies relative risk is obtained on the matching variables but a highly efficient estimate of the unmatched ones are also obtained. Also, better design schemes over space-time might be employed, such as equi-spaced sampling over a space-time grid, as long as model assumptions discussed above are not violated.

4. Results

4.1. A spatio-temporal model for people caused fire occurrence

Let Y_{st} be the random indicator variable for whether or not a fire ignition occurred at location s at time t. Here $s = (s_1, s_2)$ is a location index for the spatial grid of cells, where s_1 denotes longitude and s_2 denotes latitude, and time is indexed by $t = (t_1, t_2)$, where t_1 denotes the day of year and t_2 denotes the year. We assume this is a Bernoulli variable taking on a value of 1 if an ignition occurs at (s, t) and 0 otherwise. Then, the corresponding ignition probability $p(s, t) \equiv \mathrm{E}[Y_{st}]$, can be modelled using the following logistic generalized additive model framework

$$\text{logit}\left\{p(s,t)\right\} = \beta_0 + f_1(s,t) + \sum_{j=1} f_{j+1}(x_{jst}) \tag{7}$$

where β_0 is an intercept parameter, $f_1(s, t)$ is a smooth function of space and time, and the $f_j(\cdot)$'s are smoothing splines that account for non-linear relationships between the probability of ignition and the explanatory variables x_{jst}.

Prior to model building we performed some exploratory analyses to identify variables to be considered. Covariates which appeared to have empirical associations with fire ignitions

were included in candidate models. The model we present herein was selected by examining the AIC scores (Akaike 1973) of a sequence of nested models and choosing the model with the minimum observed AIC. Hypothesis testing via a Chi-squared analysis of deviance test confirms a significant improvement in fit when the selected model is compared to the sequence of nested models (see Table 1). The addition of other covariate effects to our model did not lead to a significant improvement in deviance. This led to a model with an intercept, a temporal (within year) effect, a bivariate spatial effect, partial effects of FFMC, DMC, BUI and of the respective distances to the nearest railway, road and town. Simple linear effects were found to best describe the associations with BUI and distance to the nearest town. The estimated coefficients for the linear effects of BUI and distance to town were 3.32×10^{-2} and -3.26×10^{-5}, respectively. These two linear relationships were significant at the 5% level (p-values of 0.011 and 0.016, respectively), and they had very wide confidence intervals that nearly enveloped zero along their entire range. Consequently, they were dropped from our model. Such a decision is in agreement with the comments on model selection for GAMs made by Wood and Augustin (2002).

The estimated partial effects are plotted in Figure 2. The corresponding coefficient estimates associated with each partial effect are listed in Table 2, while p-values inclusion of these components in the model are provided in Table 3. The spatial surface suggests there is a lower risk of ignition in the northwest and a region of higher risk centred near -81.60 longitude $\times$ 48.4 latitude. Otherwise, the spatial effect is relatively flat, indicating that the location-specific fire-weather and anthropogenic partial effects in the model are adequately capturing what influences changes in fire risk. The univariate partial effects estimates are all intuitively sensible. The temporal effect is bi-modal, possibly reflecting the fact that much of Ontario experiences two peaks in people-caused fire occurrence during the fire season, each year. The first peak occurs usually early in the fire season when there is ample dead grass and other cured fine fuel that can dry quickly. This supports the ignition and spread of accidental fires caused by rural residents burning debris near their homes and cottages, and railway operations that result in the ignition of dead grass along railway corridors. Camping and berry-picker fires tend to occur later during the summer when such recreation activities peak, if the moss or duff layer is dry enough to support ignition and fire spread. Increasing values of FFMC and DMC represent decreasing fuel moisture content and low fuel moisture and is associated with increased fire risk (see e.g., Martell et al. 1987; Martell and Belivacqua 1989; Wotton and Martell 2005). In addition, ignition risk decreases as the proximity to a railway, road or town increases, which is to be expected with people-caused fires. The estimated intercept is -6.25, has a standard error of 0.54, and is highly significant (p-value ≈ 0. Its sign and magnitude, relative to the range of the estimated partial effects, reflects the fact that fire ignitions are rare events on the fine spatio-temporal scale of our analysis.

4.2. Model assessment

A common simplifying assumption when modelling spatio-temporal processes is of no interaction between space and time. We incorporated this into our model via an additive decomposition, parallelling the structure of previous similar models (Brillinger et al. 2003, 2006; Preisler et al. 2004). There are several advantages to this framework. For example, a trivariate smoother uses up more degrees of freedom, which can lead to increases in standard error estimates for other components in the model. Moreover, it is of interest to fire management agencies to quantify how fire risk varies throughout the fire season and to identify

which areas may experience higher ignition rates, on average. A univariate temporal effect can be interpreted as a baseline for the within year seasonality, which in the past had to be estimated by incorporating less flexible periodic functions in logistic GLMs (e.g., Martell and Bevilacqua, 1989). A bivariate spatial component yields a map which highlights areas that, on average, have a higher ignition risk once relevant fire-weather and land-use covariates have been accounted for. Such seasonality characteristics and high-risk areas are not as easily identified in a trivariate spatio-temporal interaction effect.

Given the aforementioned advantages, it is of interest to assess the impact of ignoring space-time interaction. Note that such an interaction in these data was identified and explored in a preliminary study by Sun (2007). Her work suggested that the space-time interaction arose because ignition rates for each category of people-caused fires were not constant throughout the fire season. For example, she found that railway ignitions occurred more frequently earlier on in the fire season, which resulted in a ridge in the spatial surface around the main railway line in this region during that period. Since not all ignitions for the remaining people-caused categories occur in isolated regions and/or occur more often during specific periods, fully explaining what causes the space-time interaction is challenging and remains the subject for a future study.

To assess the impact of ignoring any space-time interaction we fit a second model, replacing the additive spatial and temporal components with a single trivariate spatio-temporal interaction term. We compare the fit of these two models by contrasting respective observed and expected counts of fire ignitions when aggregating at different spatial and temporal scales. We also compare observed fire rates versus those predicted for the subset of data which was reserved for cross validation. Here expected counts are easily obtained by summing fitted (or predicted values) that have been transformed to probabilities using the inverse of the logistic function. Four examples of this appear in Figure 3. In these plots the observed number of fires appear as points, while those expected under the models appear as lines. A solid red line represents the counts for the model with the space-time interaction, while a dashed blue line is used for the model with space and time as additive effects. Panel (a) compares the fit when counts are aggregated annually and panel (b) when counts are aggregated monthly. The fit appears reasonable, although the annual fit could be improved. At these scales, there does not appear to be any advantage to having a space time interaction. To explore this further, we compared monthly counts of observed and expected fires under each of these models based on proximity to a railway. Specifically, we classified a location as "close" to a railway if the distance to the nearest railway line was less than 5 km. This threshold was chosen because 95% of all observed railway-caused fires occurred within this distance. The results appear in panels (c) and (d), and again are suggesting there is no advantage to modelling space and time as an interaction here. In addition, Figure 4 compares the fit of these models for each of four spatial quadrats that partition the study area into a regular 2×2 lattice. The figure compares observed and expected counts of fires, aggregated monthly for each quadrat, and demonstrates once again a reasonable fit of the model with additive spatial and temporal effects.

Given their dominance in the historical fire risk literature, we also fit a logistic GLM, where the spatial and temporal partial effects were excluded and the remaining covariate effects were linear. In general, this model performed reasonably well when compared to the spatio-temporal models. For example, we compared observed versus expected when counts were aggregated using the Ontario Ministry of Natural Resource's classification scheme for the Fire Weather Index (FWI). The FWI is a risk index used to represent the potential frontal intensity

of a forest fire based on relative risk of spread and vegetation available for combustion. It is computed as a function of the fine fuel moisture code, the duff moisture code, a drought code and windspeed (see e.g., Van Wagner 1987). These results appear in Table 4. The assumption of an additive decomposition of space and time again appears adequate. Several similar comparisons of observed versus expected aggregating over different variables in the data set showed similar results: little difference was observed when comparing the impact of assuming space and time were additive effects.

Although from Table 4 one might postulate that linear effects may be sufficient, it is easy to demonstrate that using nonlinear effects (rather than assuming linear associations) lead to an improvement in fit. To do so, we examine the model with some linear components on temporal or spatial scales. Some of these results are also illustrated in Figure 3, where the model with linear components is represented by a green dotted line. Annual counts are comparable across models (panel a), but on a monthly basis (panel b) the model with the linear components does not fit as well as the other two completely nonparametric models. This apparent lack of fit is amplified, when observed and expected counts are compared across categories of distances to the nearest railway (panels c and d): the model with a linear effect of distance to the nearest railway clearly underpredicts for locations close to a rail road, and mostly overpredicts for locations further away.

4.3. The effects of changing the inclusion probability

Here, we describe the results of a small sensitivity study to investigate the impact of varying the zero-fire voxel inclusion probability. Specifically, we assess how the estimates of the partial effects change when this sampling rate reduces. Recall an inclusion probability of 10% was used to construct the original data set.

Decreasing the sampling rate for the zero-fire voxels can influence both the mean and standard error estimates for the partial effects. Small reductions in the sampling rate result in partial effect curves which are similar to those in the model, but the confidence regions become wider. If the sampling probability gets small enough, the estimated effect can change dramatically. These sensitivities are illustrated in Figure 5 where the partial effect of the duff moisture code is plotted for three scenarios. The solid line and shaded area represent the estimated effect and confidence region for the original data set, while the red lines illustrate how these change when the sampling rate is reduced to 5% (panel a) or to 2.5% (panel b). Similar changes are seen in the other partial effects. Changes in mean effects were observed to be be more pronounced for covariates that have sparse sections of data, such as the illustrated duff moisture code partial effect, which does not have many observations above 100.

5. Discussion

In this paper we have demonstrated the flexibility that logistic generalized additive models provide for understanding the relationship between fire risk and relevant explanatory variables. As discussed previously in the literature (Brillinger et al. 2003), logistic models approximate the corresponding point process' likelihood. This approximation is based on the discrete analogue of the Poisson process: a set of Bernoulli observations on a set of fine spatio-temporal cells partitioning the study area, where each cell is assigned a 1 or a 0, depending on the presence or absence of a fire event. A consequence of this setup, is the requirement to sample

the 0-fire cells in order to obtain a data set for which estimation is computationally feasible. When the response is the log-odds of a fire event, this sampling induces a deterministic offset into the additive modelling structure. We show here that such a scheme is directly connected to response-based sampling and case-control studies.

Our investigations demonstrated that both the mean and standard errors of partial effects are sensitive to decreases in the inclusion probability for the 0-fire cells. Stratification or matched case-control sampling methods would likely lead to increased precision. These sorts of design considerations will be considered in future work drawing on optimality properties of designs in case-control studies. Additionally, other measures of exposures developed in case-control studies point to the need to consider better measures of exposure to ignition sources in fire risk analyses beyond distance to the nearest railway, for example. More appropriate measures may include how much railway is within small neighourhoods of the ignition event, and the frequency of use of the railway, or some combination of such variables to create more relevant measures of exposure.

The methodology we employ provides a reasonable starting point for modelling the ignition risk of large fires, although some modifications may be required. One possibility was demonstrated by (Preisler et al. 2004), who employed a conditional framework to estimate both the conditional and unconditional probability of a large fire. The former is calculated by constructing a model where the observed fire ignition events are re-classified to a dichotomous response: given final size information, an ignition event is classified as 1 if it grew to a large fire, and 0 if not. Then, a logistic GAM can be fitted to estimate the probability of a fire becoming large, given ignition. The unconditional probability of a large fire is simply the product of this conditional probability and the corresponding estimated ignition probability based on the modelling framework we employ herein.

In our study of fire risk for this region of Canadian boreal forest, spatial, temporal, fire-weather and human land-use characteristics were found to affect the risk of ignition. Using a variety of visual and diagnostic techniques, we found that nonlinear covariate effects are superior to linear modelling, and that an additive decomposition was adequate for incorporating space and time. An intra-annual trend component was clearly necessary. The spline-based smoothers used in GAMs provide a more flexible approach than those using periodic functions which have appeared in the historical literature. Besides seasonal trends within each fire season, trends across years are of also of interest to the fire science community. Previously, Brillinger et al. (2006) examined the incorporation of a random effect component to account for inter-annual variability. Other possibilities for future studies of annual trends could include a nonlinear partial effect smoothing across years, or a multivariate trend surface that smoothed within and across years. The latter would permit investigations into not only average annual trends, but also into assessing whether the length of the active fire season is changing over time. In such studies, there are significant sources of confounding, such as changes in management strategy and detection efficiency as discussed by Woolford et al. (2010). Accounting for such changes and developing appropriate methodology for hypothesis testing is ongoing and will be discussed in future work.

To conclude, we note that there exists an entirely different class of models that can also be used to model wildfire behaviour and ignition processes: process models (which are sometimes referred to as mechanistic models). Process models explicitly model the underlying physical processes and contain model parameters that are interpretable in terms of specific physical aspects, for example, conduction and radiation heat transfer parameters, fuel mois-

ture content and ignition temperature. Such physically-grounded models contrast with our statiscially modelling framework, which is motivated by our understanding of the social and physical processes that produce fire ignitions–for example, that people engage in specific type of land use activities (e.g., blueberry picking) in some areas during specific times of the year, that they sometimes carelessly discard cigarettes butts without properly extinguishing them, and that when they do so in flammable fuel complexes, the probability that their actions will ignite a fire depends upon many physical parameters including the moisture content of the fuel. It would be very difficult to develop and couple together all pertinent social and physical and social models of fire ignition processes. In their comprehensive classification of landscape fire succession models (LFSM) Keane et al. (2004) identified ignition as an important process and reported that "The physical approach attempts to explicitly simulate the physical processes that govern fire initiation using driving variables including weather, fuel moisture, and lightning events. This is an extremely difficult challenge that is filled with scale, data, and knowledge limitations. We know of no LFSM that simulates fire ignition using this approach." More recently, Martell and Sun (2008) sketched out a conceptual stochastic process model for lightning fire occurrence and used it to motivate their empirical model but they made no attempt to estimate the physical parameters of their conceptual model.

Our primary objective was to estimate the probability of people-caused forest fire ignitions. Our empirical approach is a very flexible framework that is broadly applicable; we can account for uncertainties and estimate the effects of specific covariates without the need to estimate underlying social and physical process parameters. We did find that our estimated covariate effects were intuitively sensible. We also demonstrated in our simulation study, the precision of our approach will increase as our sample size increases. Consequently, models such as ours could be employed in a calibration scenario where an empirical model or methods could be used to calibrate a physical model. This was illustrated in the forest fire context by Garcia et al. (2008), who demonstrated that nonparametric smoothing methods could be used to reduce computational difficulties in a deterministic fire growth model used operationally by Canadian forest fire management agencies. Finally, we note that output from a process model could be used as input for an empirical model. For example, weather output from a climate simulation model could be used as input to our model to forecast changes in fire ignition risk across a landscape. This has been used for short-term forecasts (e.g., Preisler et al. 2008), and to quantify potential changes in fire risk under climate change scenarios (e.g., Wotton et al. 2003). Our model could be employed in a similar fashion, using output from appropriate mechanistic weather/climate, forest succession, and human-land use dynamics models.

Acknowledgements

The support of Geomatics for Informed Decisions (GEOIDE), the Natural Sciences and Engineering Research Council of Canada (NSERC) and the National Institute for Complex Data Structures (NICDS) is gratefully acknowledged. We also thank the Ontario Ministry of Natural Resources for the use of their fire data, and Justin Morgenroth for pre-processing that fire and weather data, developing the non-fire sampling software and sampling the non fire days. Models were fit in R (R Development Core Team 2008) using the **mgcv** package (Wood 2006).

References

Akaike H (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In B Petran, F Csaaki (eds.), "International Symposium on Information Theory, Akadeemiai Kiadi, Budapest, Hungary," pp. 267–281.

Berman M, Turner TR (1992). "Approximating Point Process Likelihoods with GLIM." *Applied Statistics*, **41**, 31–38.

Breslow N, Powers W (1978). "Are There Two Logistic Regressions for Retrospective Studies?" *Biometrics*, **34**, 100–105.

Brillinger DR, Preisler HK, Benoit JW (2003). "Risk Assessment: A Forest Fire Example." In DR Goldstein (ed.) "Science and Statistics: A Festschrift for Terry Speed." *Institute of Mathematical Statistics Lecture Notes*, **40**, 177–196. Beechwood, OH.

Brillinger DR, Preisler HK, Benoit JW (2006). "Probabilistic Risk Assessment for Wildfires." *Environmetrics*, **17**, 622–633. DOI: 10.1002/env.768.

Craven P, Wahba G (1979). "Smoothing Noisy Data with Spline Functions." *Numerische Mathematik*, **31**, 377–403.

Cunningham AA, Martell DL (1973). "A Stochastic Model for the Occurrence of Man-Caused Forest Fires." *Canadian Journal of Forest Research*, **3**, 282–287.

Garcia T, Braun J, Bryce R, Tymstra C (2008). "Smoothing and Bootstrapping the PROMETHEUS Fire Growth Model." *Environmetrics*, **19**, 836–848. doi:10.1002/env.907

Golub GH, Heath M, Wahba G (1979). "Generalized Cross Validation as a Method for Choosing a Good Ridge Parameter." *Technometrics*, **21**, 215–223.

Hastie T, Tibshirani R (1986). "Generalized Additive Models (With Discussion)". *Statistical Science*, **1**, 297–318.

Keane RE, Cary GJ, Davies ID, Flannigan MD, Gardner RH, Lavorel S, Lenihan JM, Li C, Rupp TS (2004). "A Classification of Landscape Fire Succession Models: Spatial Simulations of Fire and Vegetation Dynamics." *Ecological Modelling*, **179**: 3–27.

Martell DL, Belivacqua E (1989). "Modelling Seasonal Variation in Daily People-Caused Forest Fire Occurrence." *Canadian Journal of Forest Research*, **19**, 1555–1563.

Martell DL, Otukol S, Stocks BJ (1987). "A Logistic Model for Predicting Daily People-Caused Forest Fire Occurrence in Ontario." *Canadian Journal of Forest Research*, **17**, 394–401.

Martell DL, Sun H (2008). "The Impact of Forest Fire Suppression, Vegetation and Weather on Burned Area in Ontario." *Canadian Journal of Forest Research*, **38**, 1547–1563.

Morgenroth J (2003). "New Fire Occurrence Prediction Model for Tembec Area." Technical report. Fire Management Systems Laboratory, Faculty of Forestry, University of Toronto. Toronto, Canada.

Natural Resources Canada (NRC) (2006). "Canadian Forest Fire Danger Rating System." http://www.nofc.forestry.ca/fire/research/environment/cffdrs/cffdrs_e.htm:NRC.

Preisler HK, Brillinger DR, Burgan RE, Benoit JW (2004). "Probability Based Models for Estimation of Wildfire Risk." *International Journal of Wildland Fire*, **13**, 133–142.

Preisler HK, Westerling AL (2007). "Statistical Model for Forecasting Monthly Large Wildfire Events in Western United States." *J. Appl. Meteorology and Climatology*, **46**, 1020–1030.

Preisler HK, Chen S, Fujioka F, Benoit JW, Westerling AL (2008). "Wildland Fire Probabilities Estimated from Weather Model-Deduced Monthly Mean Fire Danger Indices." *International Journal of Wildland Fire*, **17**, 305–316.

R Development Core Team (2008). "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sun J (2007). "Human-Caused Fires in the Romeo Malette Forest." M.Sc. project report. Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Canada.

Turner R (2009). "Point Patterns of Forest Fire Locations." *Ecological and Environmental Statistics*, **16**, 197Ű-223. doi:10.1007/s10651-007-0085-1.

Van Wagner CE (1987). "Development and Structure of the Canadian Forest Fire Weather Index System." Canadian Forest Service, Ottawa, Ontario. Canada. Forestry Technical Report 35.

Vega Garcia C, Woodard PM, Titus SJ, Adamowicz WL, Lee BS (1995). "A Logit Model for Predicting the Daily Occurrence of Human Caused Forest Fires." *International Journal of Wildland Fire*, **5**, 101–111.

Wood SN, Augustin NH (2002). "GAMs with Integrated Model Selection using Penalized Regression Splines and Applications to Environmental Modelling." *Ecological Modelling*, **157**, 157–177.

Wood SN (2000). "Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties." *Journal of the Royal Statistical Society B*, **62**, 413–428.

Wood SN (2003). "Thin Plate Regression Splines." *Journal of the Royal Statistical Society B*, **65**, 95–114.

Wood SN (2004). "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models." *Journal of the American Statistical Association*, **99**, 673–686. doi:10.1198/016214504000000980.

Wood SN (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC Press: Boca Raton, FL.

Woolford DG, Cao J, Dean CB, Martell DL (2010). "Characterizing Temporal Changes in Forest Fire Ignitions: Looking for Climate Change Signals in a Region of the Canadian Boreal Forest." *Submitted*.

Wotton BM, Martell DL, Logan KA (2003). "Climate Change and People-Caused Forest Fire Occurrence in Ontario." *Climatic Change*, **60**, 275–295.

Wotton BM, Martell DL (2005). "A Lightning Fire Occurrence Model for Ontario." *Canadian Journal of Forest Research*, **35**, 1389–1401. DOI: 10.1139/X05-071.

A. Figures

Figure 1: Location of the study region (shaded area) encapsulating the Romeo Malette forest in Ontario, Canada (shaded area in insert).

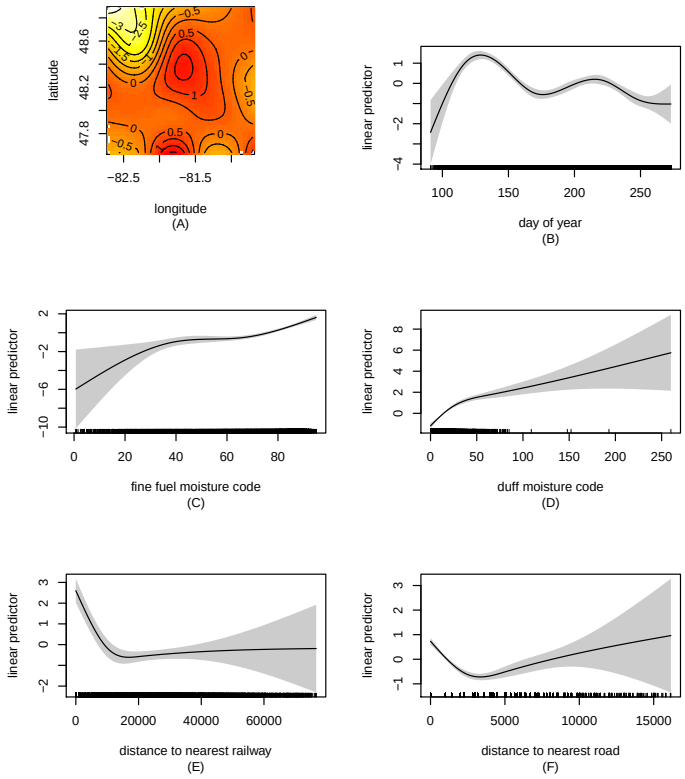

Figure 2: Plots of the partial effects in the fitted model. Plots are on the logit scale.

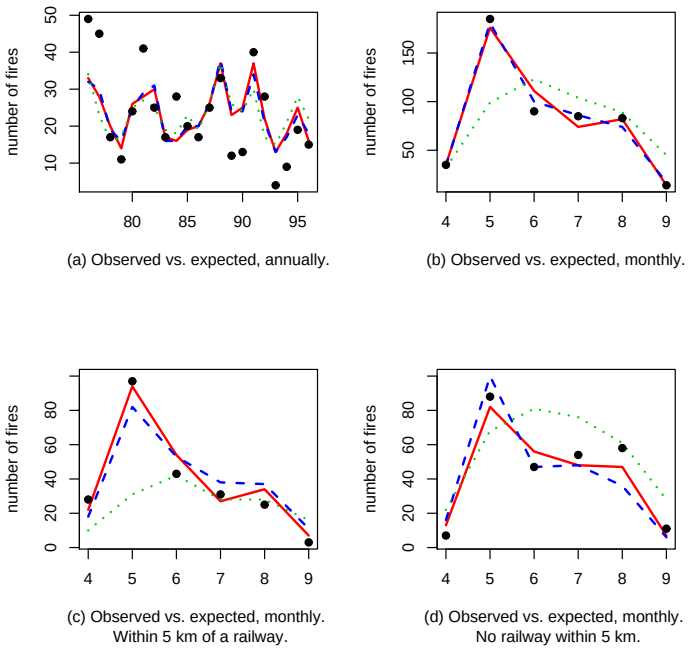

Figure 3: Comparison of observed number of fires (points) versus those predicted by a model with a space-time interaction term (red line), a model where space and time are separate additive effects (blue dashed line) and a model with additive space and time effects and linear effects for the remaining components (green dotted line). Counts are aggregated (a) annually, (b) monthly, (c) monthly for locations within 5 km of a railway, and (d) monthly for locations further than 5 km away from a railway.

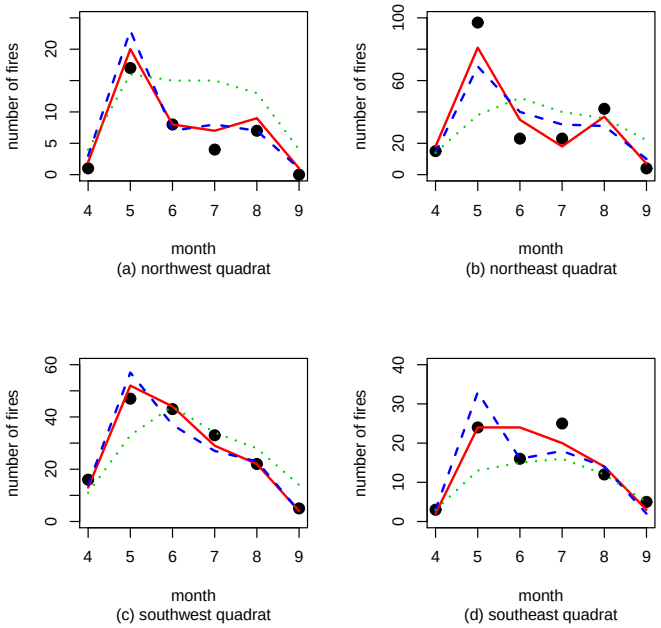

Figure 4: Comparison of observed number of fires (points) versus those predicted by a model with a space-time interaction term (red line), a model where space and time are separate additive effects (blue dashed line) and a model with additive space and time effects and linear effects for the remaining components (green dotted line). Counts are aggregated monthly for each of four quadrats partitioning the study region: (a) northwest quadrat, (b) northeast quadrat, (c) southwest quadrat, and (d) southeast quadrat.

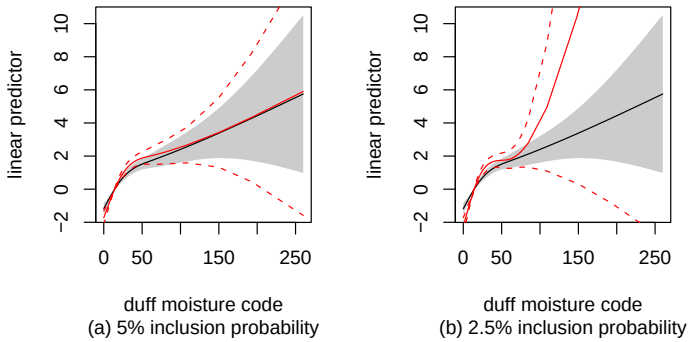

Figure 5: Comparing the estimated partial effect and its confidence region for the duff moisture code when the sampling rate for the non-fire voxels is decreased from 10% (the black line and grey shaded region) to (a) 5% and (b) 2.5% (red lines).

B. Tables

Model Terms	Resid. Df	Resid. Dev	ΔDf	ΔDeviance	p-value
intercept	7683.00	5861.8			
+ space-time	7654.31	5110.4	28.79	751.4	≈ 0
+ FFMC	7651.31	4728.0	2.99	382.4	≈ 0
+ distance to nearest railway	7651.14	4598.6	0.18	129.4	≈ 0
+ DMC	7647.76	4503.3	3.37	95.3	≈ 0
+ distance to nearest road	7647.39	4450.8	0.38	52.5	≈ 0
+ BUI	7646.65	4446.4	0.74	4.4	0.02
+ distance to nearest town	7645.92	4444.8	0.73	1.6	0.10

Table 1: Analysis of deviance table for the series of nested models. The first column outlines the order in which the partial effects entered at each step in the model building process. P-values were calculated according to a Chi-squared test statistic.

Predictor	Basis Dimension	Est. Coefficients
bivariate effect		
(longitude, latitude)	25	{0.22, 0.55, -1.14, -3.82, -0.21, 0.52, 0.26, -2.11, -2.60, 2.12, 0.34, 1.46 1.11, -0.93, -0.43, 0.92, 1.36, 1.12 -0.33, 1.14, -0.09, -0.78, 0.34, 0.60}
univariate effects		
day of year	8	{ 1.79, 2.01, 0.48, 0.78, 1.10, 0.28, -0.66}
FFMC	4	{-0.04, 4.28, 4.61}
DMC	4	{ 1.16, 1.75, 5.75}
dist. to railway	4	{ 0.47, -2.61, -0.38}
dist. to road	4	{-0.87, -0.68, 1.04}

Table 2: The estimated coefficients for the penalized spline-based partial effects in the model, using cubic regression spline basis functions (see Wood 2006).

Partial Effect	edf	Obs. χ^2	P-value	Est. Smoothing Parameter(s)
longitude, latitude	16.99	76.91	9.9×10^{-9}	7.77×10^{-4} & 5.97×10^{-3}
day of year	6.33	126.63	≈ 0	3.37×10^{-3}
FFMC	2.79	95.25	≈ 0	2.03×10^{-4}
DMC	2.75	94.99	3.9×10^{-4}	1.59×10^{-3}
dist. to railway	2.99	174.70	≈ 0	1.05×10^{-5}
dist. to road	2.87	80.55	≈ 0	1.23×10^{-3}

Table 3: The estimated degrees of freedom (edf), observed Chi-squared test statistic and corresponding p-values, as well as the estimated smoothing parameters for the penalized spline-based partial effects in the model.

FWI Range	FWI Class	Observed	Interaction Model	Additive Model	Linear Model
0	Nil	12	10	10	8
(0, 4]	Low	61	74	76	84
(4, 11]	Moderate	183	174	170	181
(11, 23]	High	192	201	204	185
> 23	Extreme	44	32	32	34

Table 4: Comparison of observed number of fires versus those predicted by models with a space-time interaction term, where space and time are separate additive effects, and where the partial effects of the fire-weather and anthropogenic variables are linear. Counts are aggregated according to the Fire Weather Index classification scheme used in Ontario.

Affiliation:

Douglas G. Woolford
Department of Mathematics
Wilfrid Laurier University
Waterloo, ON, CANADA, N2L 3C5
E-mail: dwoolford@wlu.ca
URL: http://www.wlu.ca/math

David R. Bellhouse
Statistical and Actuarial Sciences
University of Western Ontario
London, ON, CANADA, N6A 5B7
E-mail: bellhouse@stats.uwo.ca

W. John Braun
Statistical and Actuarial Sciences
University of Western Ontario
London, ON, CANADA, N6A 5B7
E-mail: braun@stats.uwo.ca

Charmaine B. Dean
Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC, CANADA, V5A 1S6
E-mail: dean@stat.sfu.ca

David L. Martell
Faculty of Forestry
University of Toronto
Toronto, ON, CANADA, M5S 3B3
E-mail: david.martell@utoronto.ca

Jialin Sun
Statistical and Actuarial Sciences
University of Western Ontario
London, ON, CANADA, N6A 5B7
E-mail: jsun5@uwo.ca

http://www.jenvstat.org
Submitted: 2009-07-20
Accepted: 2010-05-03