A NONPARAMETRIC FRAMEWORK FOR QUANTIFYING TEMPORAL TRENDS IN THE SEASONALITY OF FOREST FIRE RISK

by

Alisha Albert-Green

B.Sc. (Hons.), *The* University *of* Western Ontario, 2009

A Project submitted in partial fulfillment
of the requirements for the degree of
Master of Science
in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2011

APPROVAL

Name: Alisha Albert-Green

Degree: Master of Science

Title of Project: A Nonparametric Framework for Quantifying Temporal Trends in the Seasonality of Forest Fire Risk

Examining Committee: Dr. Tim Swartz
Chair

Dr. Charmaine Dean, Senior Supervisor
Simon Fraser University

Dr. Douglas Woolford, Supervisor
Wilfrid Laurier University

Dr. Steven Thompson, Internal Examiner
Simon Fraser University

Date Approved: _____

Abstract

Lightning-caused fires account for approximately 45% of ignitions and 80% of area burned annually in Canada. Investigating the seasonality of these fires and how this is changing over time is of interest to fire managers and researchers. In this project, we develop flexible models for describing the temporal variation in the risk of lightning-caused ignitions and fit these models to historical forest fire records from Alberta and Ontario, Canada. The generalized additive models we utilize provide smooth estimates of fire risk by day of year for each year. Inverse calculations are used to obtain interval estimates of the start and end of the fire season annually; these are defined by the crossing of a risk threshold. Permutation-based methods are employed to test for significant trends. Finally, trends from this complex approach are compared to those of simple empirical estimates. Results suggest changes to the timing of the fire season in Alberta, but not Ontario.

Acknowledgments

This project would not have not been completed without the generous support of my supervisors, Charmaine Dean and Doug Woolford. Working with them has been a highlight of my time at Simon Fraser University. I am sincerely thankful for their guidance and encouragement throughout all aspects of this degree.

I am also grateful to John Braun and Dave Martell for giving me the opportunity to work with them as an undergraduate research assistant for two summers and for their continued support throughout my M.Sc. Thanks to John for introducing me to this forestry project and for providing me with interesting and never-ending projects to work on during the latter half of my undergrad. I am grateful to Dave for teaching me the necessary forestry background and for answering any and all related questions.

Many thanks are due to my fellow graduate students who have made the past two year so enjoyable. Thanks also to the faculty and staff in the Department of Statistics and Actuarial Science for creating an encouraging and friendly atmosphere. In particular, thanks to Rick Routledge for introducing me to generalized additive models, Derek Bingham for his role as graduate chair, Steve Thompson for agreeing to be my internal examiner and Tim Swartz for his role as my examining committee chair.

Thanks to Patrick Brown for giving me the opportunity to work as a co-op student at Cancer Care Ontario and for introducing me to the area of infectious disease modelling. Also, to Tony Panzarella and my colleagues in the Biostatistics Department at the Princess Margaret Hospital, thank you for your patience and flexibility as I worked towards finishing this program.

Data for this project have been provided by Alberta Sustainable Resource Development and the Ontario Ministry of Natural Resources. The financial support from GEOIDE and NSERC is gratefully acknowledged.

On a more personal note, I would like to thank my family for their love, support and encouragement throughout every step of my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The goal of this project is to develop a method to test for changes in the timing of the forest fire season. We illustrate this methodology by analyzing Canadian historical forest fire records aggregated at the provincial level. The practical significance of such a study quickly becomes clear when examining the following summary statistics. On average each year in Canada, there are 8000 forest fires, resulting in approximately 2.1 million hectares (ha) being burned. Suppression resources for controlling and extinguishing these fires cost between $500 million and $1 billion annually (Natural Resources Canada, 2008). Under a warming climate, these statistics have the potential to substantially increase. The magnitude of these changes, however, may not be homogeneous throughout Canada.

The consequential impacts of climate change in forestry have been extensively studied in the literature. Bonsal et al. (2001) compared climate change effects in western and eastern Canada, demonstrating a difference in the effects between these two regions. In that paper, the authors showed how the upper and lower percentiles of daily temperature have increased during winter and spring from 1950 through 1998 in western Canada. Meanwhile, these quantities have decreased throughout eastern Canada. Projections for future trends was a topic considered by Williamson et al. (2009). Discussions therein centred around an expected increase in the frequency and severity of extreme weather events throughout Canada, such as fire, drought and severe storms. These researchers predicted a 74%-118% increase in annual area burned by the latter part of this century. Again, this is expected to be highest in the western half of Canada, which includes regions of western Ontario, a central province.

Wotton and Flannigan (1993) commented that projections of increased temperatures will likely result in drier forest fuels, conditional on rainfall patterns not being significantly altered. However, variability in rainfall scenarios from current general circulation models (GCMs) is relatively large. Under a warming climate, it is postulated that there could be an increase in the number of ignitions, an increase in the amount of severe fire-weather and an extended length of the fire season (Weber and Stocks, 1998; Williamson et al., 2009). In this project, we investigate the third effect listed above.

In terms of fire management operations, the fire season officially runs from April 1 to October 31 each year, in both Alberta and Ontario (Government of Alberta: Office of Statistics and Information, 2011; Ontario Ministry of Natural Resources, 2004). This defines a period where forest fires, under the right conditions, are likely to ignite and have the potential to rapidly spread. As outlined above, we might expect to experience longer periods each year that are conducive to forest fires as a result of a warming climate. Although much work has been done toward identifying potential future climate change trends in forestry through the analysis of data from GCM scenarios, relatively little has been done in terms of quantifying historical trends. Wotton and Flannigan (1993) used output from a GCM in a doubling of atmospheric carbon dioxide scenario and compared this to projections from a scenario where carbon dioxide levels remained constant. They defined the fire season as starting after three consecutive days with temperatures greater than twelve degrees Celsius and conversely for the end. Predictions based on their approach showed a 16%-17% extension in the length of the fire season in regions of western and central Canada.

With data aggregated at the provincial level, defining the length of the fire season from empirical estimates, as was done in Wotton and Flannigan (1993), can be troublesome due to its high variability which is reflected, in part, by the heterogeneity of fire-weather indices over such large spatial extents. We address this issue by adopting a modelling approach based on the use of a nonparametric generalized additive model. Flexible models such as these allow us to relax the linearity assumption imposed on linear and generalized linear models to accurately predict fire risk. This is an extension of an approach developed in an exploratory paper by Woolford et al. (2010). In that paper, a generalized additive mixed model framework was developed to estimate the probability of at least one fire being reported on a given day, termed a *fire day*. Probabilities were estimated from a bivariate smoother of time, where time was viewed as the ordered pair of day of year, and year. The local neighbourhood influencing the fit of that model at any time point contained observations

surrounding the date in question from the same year as well as those from neighbouring years. Our interest is in extending their model to accurately quantify trends in individual fire seasons. To do this, we construct a more flexible model where the local neighbourhood determining a fitted value is solely influenced by the surrounding days within a given year.

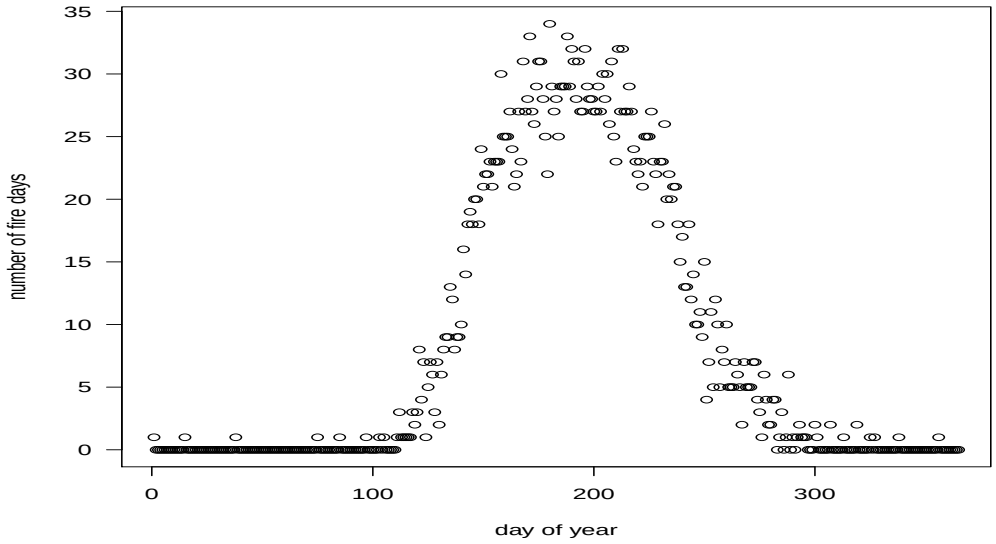
In this project, we develop a two-stage approach to test for possible climate change trends in the start and end of the forest fire season. We use a nonparametric generalized additive model to estimate fire day risk and the start and end of the fire season are defined as the crossing of a fire day risk threshold. Inverse techniques are employed to calculate confidence intervals associated with the point estimates of the start and end of the fire season. Linear models are then used to quantify these temporal trends and resampling techniques are employed to test for significance. Trends from this complex approach are contrasted with those of comparable empirical estimates of fire season length. We apply these approaches to historical forest fire records from Alberta and Ontario, Canada. It is important to note that although this project is motivated by climate change effects, much work is required before any trends may be definitively linked to climate change. We leave this discussion topic for Chapter 5.

1.2 Exploratory Analysis

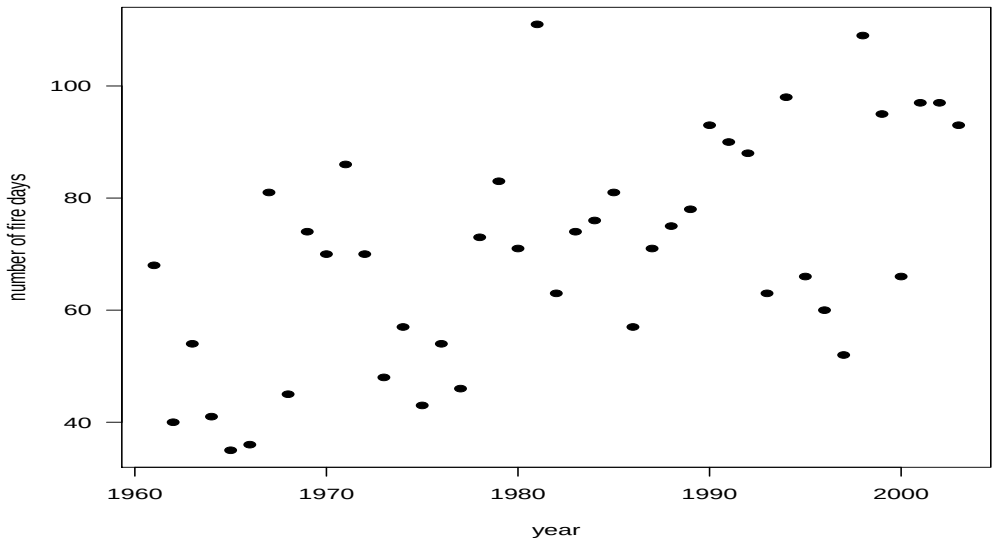
1.2.1 Analysis of Alberta Forest Fire Data

The historical forest fire data in Alberta, Canada considered here span forty-three years, from 1961-2003. During this period, there were roughly 37,000 fires, approximately half of which are attributed to lightning. These 18,000 lightning-caused fires resulted in about 6,000,000 ha being burned and corresponded to about 75% of the total area burned.

The panels in Figure 1.1 display the total number of fire days in Alberta, aggregated by day of year, and by year. The seasonality of the ignition process is clear in Figure 1.1(a); few fires occur at the beginning and end of each year. Annual trends are displayed below in Figure 1.1(b). In the Alberta forest fire data set, there appears to be an annual increase in the number of fire days. Figure 1.2 visually summarizes simple empirical estimates of the start and end of the fire season, displaying scatterplots of the day of year for the fifth, and fifth last, fire day by year. A decreasing trend in the start of the fire season can be observed from Figure 1.2(a). Informally, Figure 1.2(b) suggests a stronger trend in the end of the fire season, relative to the start.

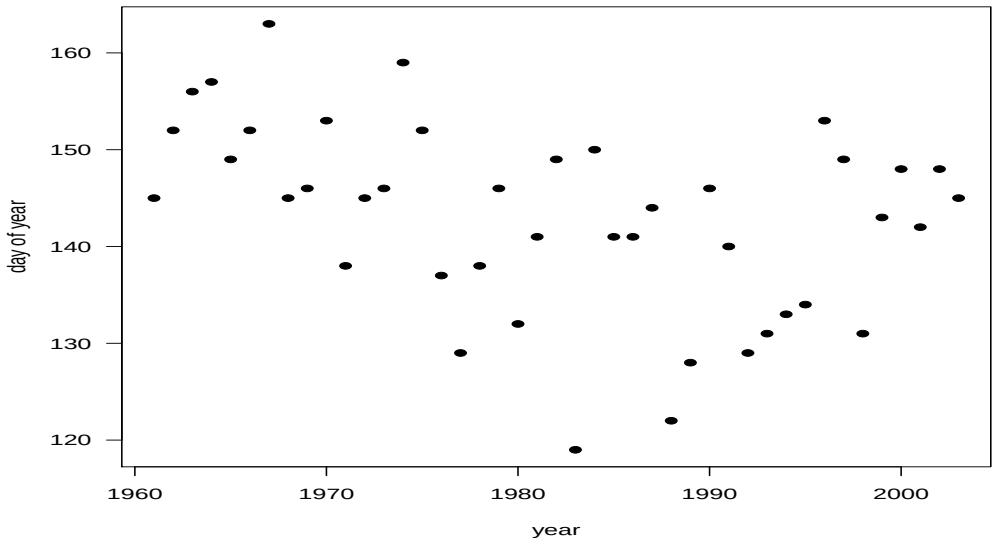


(a) Number of fire days aggregated by day of year.

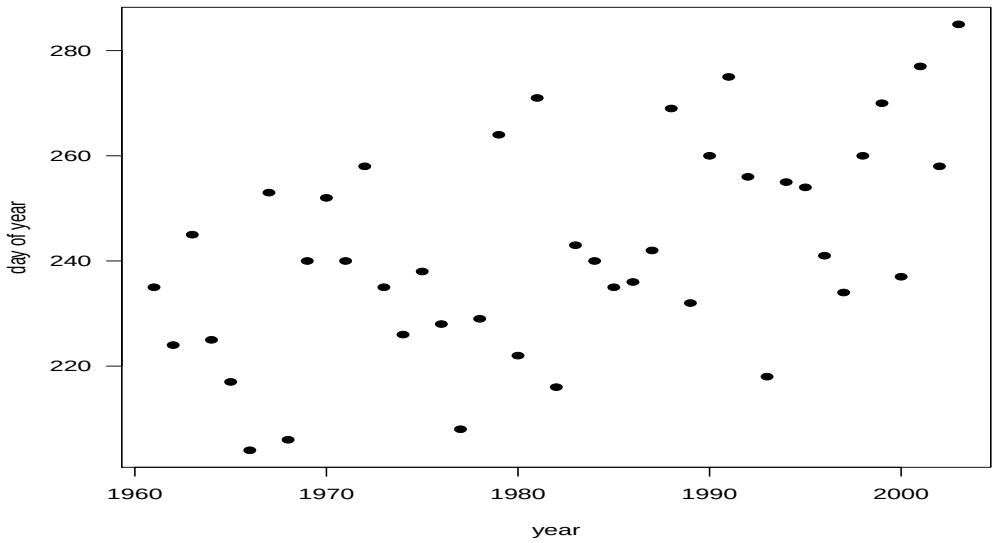


(b) Number of fire days aggregated by year.

Figure 1.1: Summary scatterplots of counts of fire days aggregated by day of year (top), and year (bottom) for lightning-caused fires in Alberta.



(a) Fifth fire day.



(b) Fifth last fire day.

Figure 1.2: Fifth (top), and fifth last (bottom) fire day each year for lightning-caused fires in Alberta.

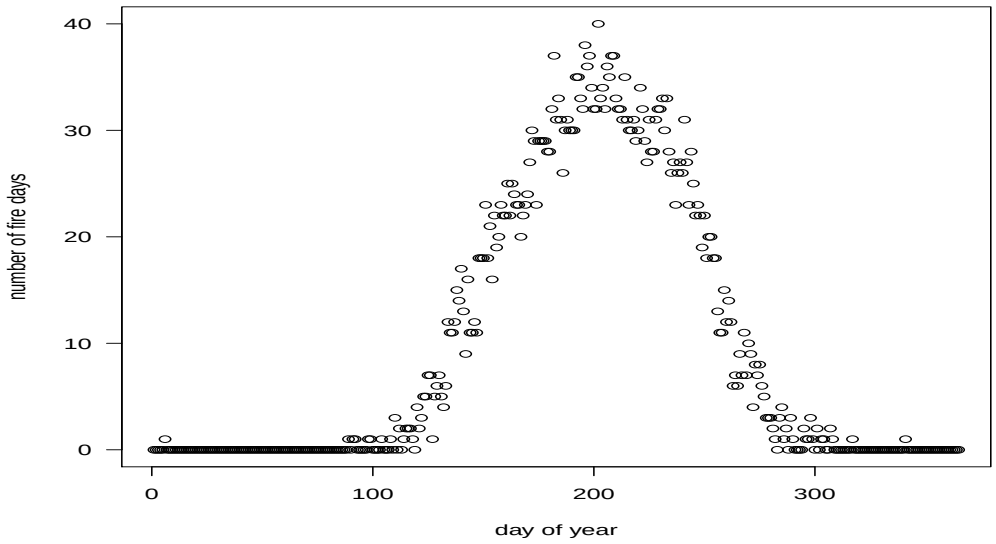
1.2.2 Analysis of Ontario Forest Fire Data

Our forest fire data for Ontario consists of all fires reported between 1963 and 2004, inclusive. This database contains approximately 61,000 fires, with about 22,000 caused by lightning. Approximately 7,000,000 ha were burned, 83% of which resulted from lightning-caused fires. In contrast to Alberta, Ontario experienced a larger number of fires, with a smaller proportion attributed to lightning.

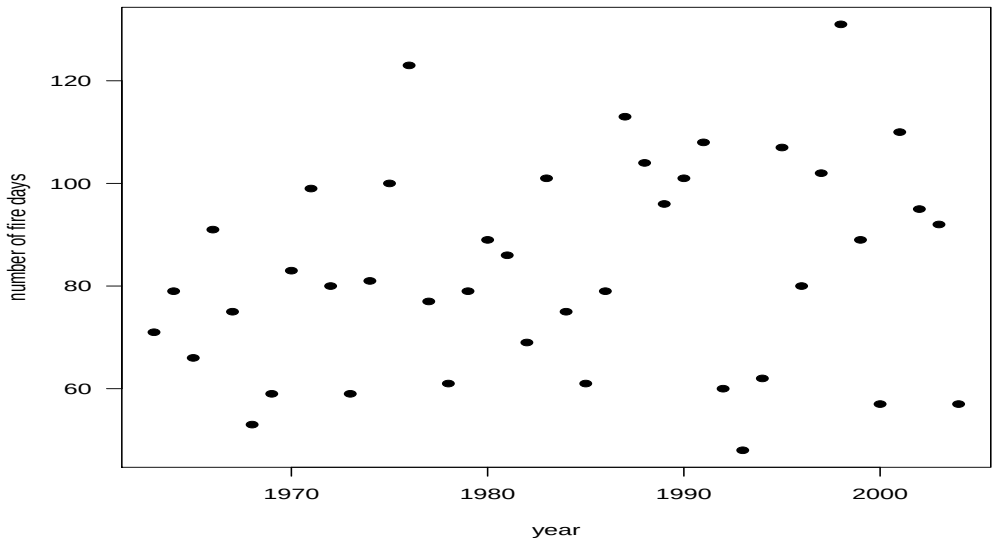
Exploratory figures, analogous to those from Section 1.2.1, are shown here for the Ontario data set. Figure 1.3 summarizes the number of fire days by day of year, and year. Although Figure 1.3(a) shows strong seasonality, no annual trends in the number of fire days are apparent in Figure 1.3(b). When using the fifth annual fire day and the fifth last annual fire day to define the start and end of the fire season, as illustrated in Figure 1.4, there is no clear trend in the length of the fire season.

1.3 Project Outline

The remainder of this project is organized as follows. Chapter 2 reviews the theory of generalized additive models with emphasis on thin plate regression splines. Trends in empirical estimates of the start and end of the fire season for Alberta and Ontario, including an outline of the approach used to estimate these trends, is the topic of Chapter 3. Chapter 4 describes our two stage nonparametric approach to test for temporal trends in the length of the fire season. Results from applying this to Alberta and Ontario historical forest fire records are also discussed and a comparison is made between these estimates and that of the empirical approaches. Chapter 5 concludes this project with a discussion and suggestions for future work.

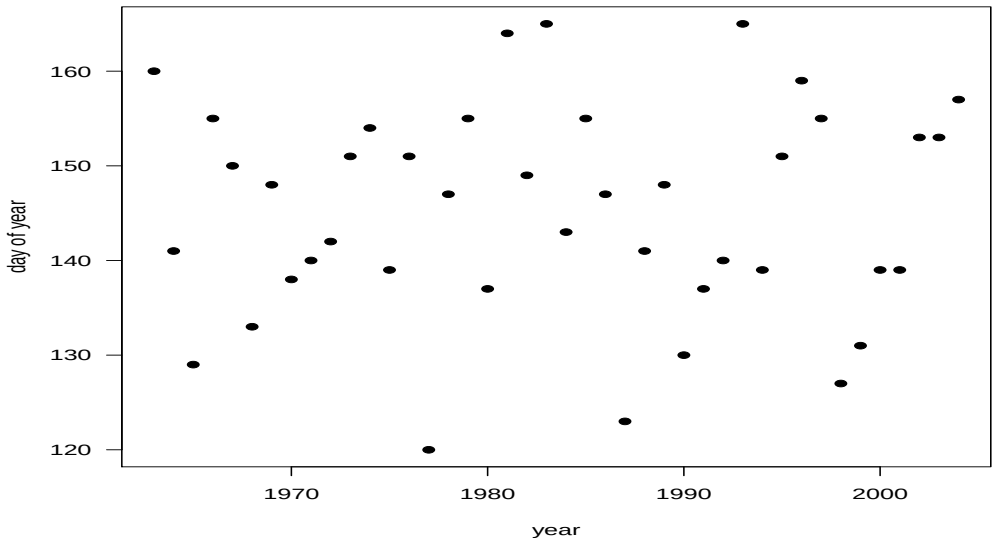


(a) Number of fire days aggregated by day of year.

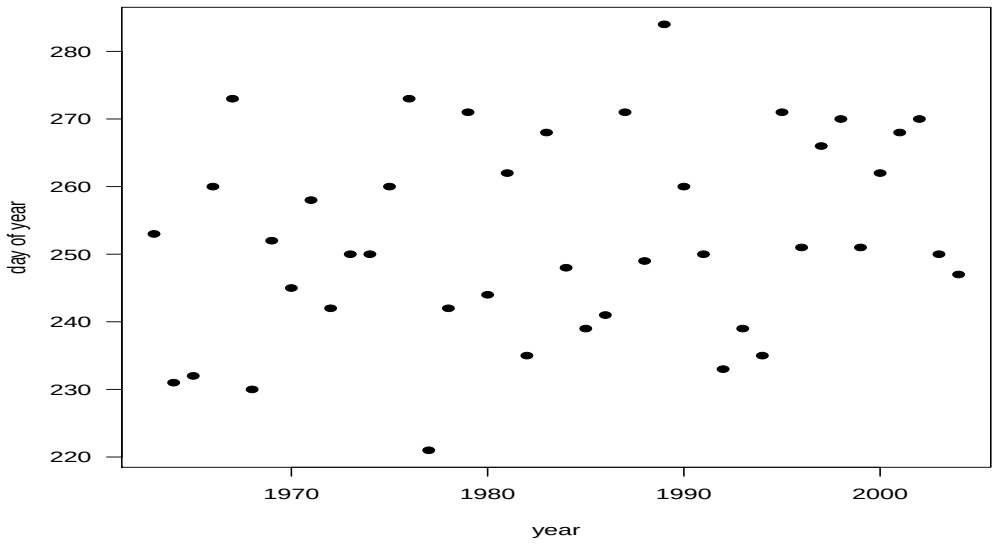


(b) Number of fire days aggregated by year.

Figure 1.3: Summary scatterplots of counts of fire days aggregated by day of year (top), and year (bottom) for lightning-caused fires in Ontario.



(a) Fifth fire day.



(b) Fifth last fire day.

Figure 1.4: Fifth (top), and fifth last (bottom) fire day each year for lightning-caused fires in Ontario.

Chapter 2

Generalized Additive Models

This chapter reviews theory underlying the models employed throughout this project. We begin with a brief review of generalized linear models in Section 2.1. Section 2.2 provides an introduction to generalized additive models and penalized spline smoothing. We also include an example of thin plate regression splines, the basis function we use for the models developed in Chapter 4.

2.1 Generalized Linear Models

Generalized linear models (GLMs) extend the familiar linear regression models by relaxing the normality assumption on the response, allowing it to follow any exponential family distribution including Poisson, binomial and gamma (McCullagh and Nelder, 1989). GLMs can be written as follows:

$$g(\mu_i) = \boldsymbol{X_i}\boldsymbol{\beta} \tag{2.1}$$

for $i = 1, \ldots, n$. Using the above notation, $g(.)$ is the link function, $\mu_i \equiv \mathrm{E}[Y_i|\boldsymbol{X}_i]$ is the conditional expectation of the ith observation, Y_i; $\boldsymbol{X}_i$ is the ith row of the model (or design) matrix which incorporates the covariates considered in the analysis, while $\boldsymbol{\beta}$ denotes the corresponding vector of parameters. On the scale of the link function, the conditional mean of the response is assumed to be a linear function of the parameters. Note that linear models are special cases of GLMs, where the link function is the identity.

The distribution of a random variable Y is from the exponential family if it has a probability density function that can be written in the form:

$$f(Y) = \exp\{[Y\theta - b(\theta)]/a(\phi) + c(y, \phi)\} \tag{2.2}$$

where a, b and c are known functions, ϕ is the scale parameter and θ is the canonical parameter. The mean and variance of any random variable belonging to an exponential family distribution are, respectively,

$$\mathrm{E}(Y) = b'(\theta)$$

and

$$\mathrm{Var}(Y) = b''(\theta)a(\phi)$$

where $\mathrm{V}(\mu) = b''(\theta)$ and $a(\phi) = \phi$.

Parameter estimation follows likelihood-based inference procedures. Specifically, an optimization method known as iteratively reweighted least squares (IRLS) is commonly employed. See McCullagh and Nelder (1989) for more details.

2.2 Generalized Additive Models

Generalized additive models (GAMs) are extensions of GLMs. They allow for nonlinear covariate effects by incorporating nonparametric smooth functions, referred to as smoothers or partial effects (Hastie and Tibshirani, 1990; Wood, 2006). In general, GAMs have the following structure:

$$g(\mu_i) = \boldsymbol{X}_i^*\boldsymbol{\beta} + \sum_{j=1}^{J} f_j(x_{ji}) \tag{2.3}$$

where i indexes the observation and j indexes the smoother, $i = 1, \ldots, n$ and $j = 1, \ldots, J$. As in GLMs, $g(.)$ denotes the link function and $\mu_i \equiv \mathrm{E}(Y_i|\boldsymbol{X}_i)$ where Y_i, the response variable, follows an exponential family distribution (see 2.2). In (2.3), the mean of the

response is a linear function of the coefficients, $\boldsymbol{\beta}$, on the scale of the link function. Hence, covariates $\boldsymbol{X}_i^*$ are incorporated into the model in a similar fashion as for GLMs. In contrast, $f_j(x_{ji})$ represent the nonparametric smooth functions of the covariates, x_{ji}, which affect the response nonlinearly. Note that x_{ji} need not be scalar, for example, it may have two components: $x_{ji} = (x_{1ji}, x_{2ji})$. As well, in the case of a multivariate smoother, x_{ji} may be a vector of covariates.

The two common concerns when fitting a GAM are related to determining the most appropriate basis function as well as determining the flexibility of each smoother. These two problems, along with inference for GAMs, are described in the remainder of this chapter.

2.2.1 Penalized Spline Smoothing in Generalized Additive Models

In GAMs, a smoother as a function of scalar x_j is represented as follows:

$$f_j(x_j) = \sum_{k=1}^{q_j} \theta_{jk} b_{jk}(x_j). \tag{2.4}$$

Here, k indexes the knot for the jth smoother, $k = 1, \ldots, q_j$, $b_{jk}(x_j)$ is the basis function for the jth covariate at the kth knot and θ_{jk} is the corresponding basis coefficient. Notice, in this context a smoother is simply a linear combination of basis functions. Let

$$\boldsymbol{b}_j(x_j) = \left[b_{j1}(x_j), b_{j2}(x_j), \ldots, b_{jq_j}(x_j)\right] \tag{2.5}$$

and

$$\boldsymbol{\theta}_j^T = \left[\theta_{j1}, \theta_{j2}, \ldots, \theta_{jq_j}\right], \tag{2.6}$$

with $\boldsymbol{\theta_j}^T$ denoting the transpose of $\boldsymbol{\theta_j}$. Then, $f_j(x_j)$ may be written in the form $\boldsymbol{b}_j(x_j)\boldsymbol{\theta}_j$, similar to a GLM. Throughout this project we make use of univariate thin plate regression splines. These are a class of flexible basis functions that can be used to smooth one or more covariates. An example of thin plate regression splines will be provided at the end of this section.

Controlling model smoothness using splines amounts to altering the number of knots for each smoother; the fewer the number of knots, the smoother the estimated function.

However, choosing the number and location of knots when fitting a regression spline is problematic as these factors strongly influence the resulting fit. The use of penalized splines circumvents this knot selection problem (Wood, 2006). In this approach, the basis dimension is fixed at a size larger than believed to be reasonable and overfitting is controlled by adding a penalty to the log likelihood, as will be discussed in detail below. This penalty determines the roughness of the spline and is often measured by the integral of the squared second derivative of the smoother:

$$\begin{aligned} J(f_j) &= \int [f_j''(x_j)]^2 dx_j \\ &= \int [\boldsymbol{b}_j''(x_j)\boldsymbol{\theta}_j]^2 dx_j \\ &= \int \boldsymbol{\theta}_j^T \boldsymbol{S}_j(x_j)\boldsymbol{\theta}_j dx_j, \end{aligned} \tag{2.7}$$

where $\boldsymbol{S}_j(x_j)$ is the $q_j \times q_j$ matrix:

$$\begin{bmatrix} b_{j1}''(x_j)^2 & b_{j1}''(x_j)b_{j2}''(x_j) & \cdots & b_{j1}''(x_j)b_{jq_j}''(x_j) \\ b_{j2}''(x_j)b_{j1}''(x_j) & b_{j2}''(x_j)^2 & \cdots & b_{j2}''(x_j)b_{jq_j}''(x_j) \\ \vdots & \vdots & \ddots & \vdots \\ b_{jq_j}''(x_j)b_{j1}''(x_j) & b_{jq_j}''(x_j)b_{j2}''(x_j) & \cdots & b_{jq_j}''(x_j)^2 \end{bmatrix}.$$

Other common measures of spline roughness are $\int [f_j'(x_j)]^2 dx_j$ and $\int [f_j'''(x_j)]^2 dx_j$ (Wood and Augustin, 2002). The only requirement is that they be continuous up to and including the order of the derivative of the penalty, as splines are joined at knot locations. The penalty, $J(f_j)$, will be large if $f_j(x_j)$ is rough and small for a smooth $f_j(x_j)$. Note that for this project, we use the penalty shown in (2.7).

Inference

As mentioned above, parameter estimation for a GAM is performed using a penalized likelihood-based approach. This is a reasonable remedy for overfitting as the penalties suppress estimates of smoothers that could lead to overfitting. To illustrate this estimation procedure, consider writing a GAM in the form of the following GLM:

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\theta} \tag{2.8}$$

where $\boldsymbol{X}$ is the full design matrix and $\boldsymbol{\theta}$ is the vector of coefficients. From Equation 2.8, $\boldsymbol{X} = [\boldsymbol{X}^*, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_J]$, where $\boldsymbol{X}^*$ is the model matrix with ith row $\boldsymbol{X}_i^*$ corresponding to the parametric model (i.e. linear) components and $\boldsymbol{X}_j$ is the $n \times q_j$ matrix with the ith row corresponding to $\boldsymbol{b}_j(x_{ji})$, $i = 1, \ldots, n$. Recall, $\boldsymbol{b}_j(x_j)$ is defined in (2.5). Additionally, $\boldsymbol{\theta}^T = [\boldsymbol{\beta}^T, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \ldots, \boldsymbol{\theta}_J^T]$, where $\boldsymbol{\theta}_j^T$ is defined in (2.6). Letting $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ denote the log likelihood of a GAM, the penalized log likelihood is then:

$$\ell_p(\boldsymbol{\theta}|\boldsymbol{y}) = \ell(\boldsymbol{\theta}|\boldsymbol{y}) - \frac{1}{2}\sum_{j=1}^{J} \lambda_j \int \boldsymbol{\theta}_j^T \boldsymbol{S}_j(x_j)\boldsymbol{\theta}_j dx_j. \tag{2.9}$$

When x_j is an n-vector, the elements in the matrix $\boldsymbol{S}_j(x_j)$ correspond to $\sum_{i=1}^{n} b''_{j\ell}(x_{ji}) b''_{j\ell'}(x_{ji})$ for $\ell = 1, \ldots, q_j$ and $\ell' = 1, \ldots, q_j$, where ℓ and ℓ' index rows and columns, respectively. In Equation 2.9, λ_j represents the smoothing parameter for the jth smoother. These smoothing parameters control the tradeoff between the two conflicting goals of a GAM: model fit and model smoothness. If $\lambda_j = 0$, then $f_j(x_j)$ is unpenalized and the resulting smoother is quite rough. Conversely, as $\lambda_j \to \infty$, $f_j(x_j)$ becomes increasingly smooth and closer to a straight line. Just as we discussed that IRLS is employed for parameter estimation in GLMs, penalized iteratively reweighted least squares (P-IRLS) is used to optimize the penalized log likelihood and estimate coefficients in GAMs. In P-IRLS, smoothing parameters are assumed to be known. Estimation of λ_j, the smoothing parameter, is performed using generalized cross validation, which is discussed in Section 2.2.2.

The degrees of freedom for the a fitted model would simply be the dimension of $\boldsymbol{\theta}$ if all smoothing parameters were zero. Conversely, if the smoothing parameters were large, the model would have relatively few degrees of freedom. Therefore, to measure the flexibility of a fitted model we use a quantity known as the effective degrees of freedom. This is defined as the trace of the influence matrix of a GAM, or $\text{tr}(\boldsymbol{A})$, where

$$\boldsymbol{A} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \sum_{j=1}^{J} \lambda_j \int \boldsymbol{S}_j(x_j)dx_j\right)^{-1} \boldsymbol{X}^T\boldsymbol{W} \tag{2.10}$$

and $\boldsymbol{W}$ is an $n \times n$ diagonal matrix of weights with diagonal elements equal to $\frac{1}{V(\hat{\mu}_i)[g'(\hat{\mu}_i)]^2}$. Note that $\hat{\mu}_i$ denotes the fitted values and $\mathrm{V}(\hat{\mu}_i)$ is a function of the fitted values. For logistic models, as we employ in Chapter 4, $\mathrm{V}(\mu_i) = \mu_i(1-\mu_i)$.

Based on the fitted model, we estimate the variance of the response in a manner analogous to that of a GLM. Specifically,

$$\widehat{\mathrm{Var}}(Y_i) = \mathrm{V}(\hat{\mu}_i)\hat{\phi}.$$

In GAMs, the scale parameter is estimated as:

$$\hat{\phi} = \frac{\sum_{i=1}^{n} \mathrm{V}(\hat{\mu}_i)^{-1}(y_i - \hat{\mu}_i)^2}{n - \mathrm{tr}(\boldsymbol{A})}.$$

Note that estimation and inference for GAMs is performed using existing software in the **mgcv** package (Wood, 2006) in R (R Development Core Team, 2008).

Example of Penalized Spline Smoothing: Thin Plate Regression Splines

Consider estimating the smoother $f(\boldsymbol{x}_i)$ from the following model:

$$g(\mu_i) = f(\boldsymbol{x}_i) \tag{2.11}$$

for $i = 1, \ldots, n$, where $\boldsymbol{x}$ consists of d covariates, each with n observations. Let $\boldsymbol{x}_j^*$, $j = 1, \ldots, q$, represent the knots. For thin plate regression splines, the estimated smoother has the form:

$$\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{q} \delta_j \eta_{wd}(||\boldsymbol{x} - \boldsymbol{x}_j^*||) + \sum_{j=1}^{M} \alpha_j \phi_j(\boldsymbol{x}) \tag{2.12}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ contain the unknown parameters that need to be estimated subject to the constraint $\boldsymbol{T}^T\boldsymbol{\delta} = 0$, and $\boldsymbol{T}$ is a $q \times M$ matrix with elements $T_{ij} = \phi_j(\boldsymbol{x}_i^*)$. Note that q represents the number of knots, while $M = \binom{w+d-1}{d}$, where w is the order of the derivative that measures the flexibility of the spline smoother, and d is the dimension of the smoother. Finally, ϕ_i are linearly independent polynomials and the basis function, $\eta_{wd}(r)$, has the following form:

$$\eta_{wd}(r) = \begin{cases} \frac{(-1)^{w+1+d/2}}{2^{2w-1}\pi^{d/2}(w-1)!(w-d/2)!} r^{2w-d}\log(r), & \text{if } d \text{ even} \\ \frac{\Gamma(d/2-w)}{2^{2w}\pi^{d/2}(w-1)!} r^{2w-d}, & \text{if } d \text{ odd.} \end{cases}$$

The penalty is:

$$J_{wd}(f) = \int \cdots \int_{\Re} \sum_{\upsilon_1+\cdots+\upsilon_d=w} \frac{w!}{\upsilon_1! \cdots \upsilon_d!} \left(\frac{\partial^w f}{\partial x_1^{\upsilon_1} \cdots \partial x_d^{\upsilon_d}} \right)^2 dx_1 \cdots dx_d \tag{2.13}$$

where $\upsilon_1, \ldots, \upsilon_d$ represents the order of the derivative for the respective covariate (Wood, 2003).

Take, for example, the case where $f(\boldsymbol{x})$ is a bivariate smoother and model flexibility is measured by the integral of the squared second derivative of the basis function (Wood and Augustin, 2002). This implies that $w = 2$ and $d = 2$. The penalty, $J_{22}(f)$, is then:

$$J_{22}(f) = \int \int \left[\left(\frac{\partial^2 f}{\partial x_1^2} \right) + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right) + \left(\frac{\partial^2 f}{\partial x_2^2} \right) \right] dx_1 dx_2. \tag{2.14}$$

The $M = 3$ linear polynomials are: $\phi_1(x_1, x_2) = 1$, $\phi_2(x_1, x_2) = x_1$ and $\phi_3(x_1, x_2) = x_2$ and the matrix, $\boldsymbol{T}$, is:

$$\boldsymbol{T} = \begin{bmatrix} 1 & x_{11}^* & x_{21}^* \\ 1 & x_{12}^* & x_{22}^* \\ \vdots & \vdots & \vdots \\ 1 & x_{1q}^* & x_{2q}^* \end{bmatrix}.$$

Finally, the basis function, $\eta_{wd}(r)$, is $\frac{1}{8\pi} r^2 \log(r)$. Generalizations to a univariate smoother, as we employ in this project, are straightforward.

2.2.2 Smoothing Parameter Selection

As mentioned above, P-IRLS is conditional on the estimated smoothing parameters, λ_j, $j = 1, \ldots, J$. Generalized cross validation (GCV) is employed to estimate these parameters (Wood, 2006; Wood and Augustin, 2002).

The idea behind cross validation techniques is to estimate smoothing parameters by minimizing the mean square prediction error, namely the average squared prediction error

that results from predicting a new observation using the fitted model. For the case of an additive model, where the response is approximately normally distributed, this amounts to minimizing:

$$\gamma_g = \sum_{i=1}^{n} \frac{n||\boldsymbol{Y} - \hat{\boldsymbol{\mu}}||^2}{[n - \mathrm{tr}(\boldsymbol{A})]^2} \tag{2.15}$$

where $\hat{\boldsymbol{\mu}} = \boldsymbol{Y}\boldsymbol{A}$ are the predicted values and $\boldsymbol{A}$ is the influence matrix.

In the case of GAMs, rather than minimizing the residual sum of squares, the GCV score is calculating using model deviance, $\mathrm{D}(\hat{\boldsymbol{\theta}})$, as follows:

$$\gamma_g = \frac{n\mathrm{D}(\hat{\boldsymbol{\theta}})}{[n - \mathrm{tr}(\boldsymbol{A})]^2}, \tag{2.16}$$

where $\mathrm{D}(\hat{\boldsymbol{\theta}})$ is simply $-2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_{\max})]$. Using this notation, $\ell(\hat{\boldsymbol{\theta}}_{\max})$ is the log likelihood from the saturated model, where we have one parameter estimate per observation and $\ell(\hat{\boldsymbol{\theta}})$ is the log likelihood from the model being fit. Note that GCV can also be used as a criteria for model selection, with the most appropriate model having the smallest GCV score.

Chapter 3

Investigating Trends in Empirical Estimates of Fire Season Length

The focus of this chapter is on quantifying trends in the start and end of the fire season from simple empirical estimates of fire season length. Section 3.1 discusses our methodology for investigating these trends, including explanations of our tests for significance. The results when testing for trends in the Alberta and Ontario forest fire data are summarized in Section 3.2. As well, model goodness of fit is assessed via residual analysis. Note that the intent of this chapter is to briefly discuss these results for use as a comparison to that of the preferred nonparametric approach we present in Chapter 4.

3.1 Empirical Approach to Estimating Trends in the Start and End of the Fire Season

As a preliminary investigation, we develop two empirical estimates of fire season length, one based on observed fire day counts and the other on the empirical cumulative distribution function (ECDF) of fire days. For the count-based empirical estimate, we define the start of the fire season using three different thresholds: the observed date of the first, fifth, and tenth fire day each year. The comparable quantity for the procedure based on the ECDF also considers three thresholds, defined as the first day the ECDF of fire days exceeds 1%, 5%, and 10%, annually. Analogous estimates are employed as definitions for the end of the fire season. That is, the last, fifth last, and tenth last fire day and correspondingly, the last

day the ECDF fails to exceed 99%, 95%, and 90%. Based on these results, we developed the following approach to estimate trends in the start and end of the fire season.

Let $\boldsymbol{X}^r = (X_1^r, X_2^r, \ldots, X_n^r)$ denote the vector of estimated dates for the start of the fire season for n years of data at the rth threshold and $\boldsymbol{Y}^r = (Y_1^r, Y_2^r, \ldots, Y_n^r)$ denote the end dates. We assume that both $\boldsymbol{X}^r$ and $\boldsymbol{Y}^r$ follow a correlated normal distribution. Trends in start of the fire season, can then be estimated using the following linear regression model:

$$X_i^r = \beta_{X0}^r + \beta_{X1}^r d_i + \varepsilon_i^r \tag{3.1}$$

and correspondingly

$$Y_i^r = \beta_{Y0}^r + \beta_{Y1}^r d_i + \varepsilon_i^r \tag{3.2}$$

for the end of the fire season where the covariate d represents year, and $i = 1, \ldots, n$. The subscripts on the slope and intercept are used to differentiate between parameters for the start and end of the fire season. We assume that the residuals, ε_i^r, follow a first order autoregressive process, denoted AR(1), to account for possible short term correlation between these estimates:

$$\varepsilon_i^r = \rho^r \varepsilon_{i-1}^r + a_i^r. \tag{3.3}$$

Here, a_i^r are independent and normally distributed random variables with mean 0 and constant variance, $\sigma_{a^r}^2$. For AR(1) errors, $\mathrm{E}(\varepsilon_i^r) = 0$, $\mathrm{Var}(\varepsilon_i^r) = \sigma_{a^r}^2 \left[\frac{1}{1-(\rho^r)^2}\right]$ and $\mathrm{Cov}(\varepsilon_i^r, \varepsilon_{i+u}^r) = \rho^{r|u|}\sigma_{a^r}^2 \left[\frac{1}{1-(\rho^r)^2}\right]$, $u \,\epsilon\, \mathbb{Z}$. Note that Montgomery et al. (2006) provides a thorough discussion of inference for linear models in this context.

Confidence intervals for the fitted values are obtained through a parametric bootstrap-based method. We construct $(1-\alpha)100\%$ pointwise confidence intervals using the following algorithm.

1. Let b index the replication. Set b to 1.

2. Simulate observations from a correlated normal distribution with parameters matching that of the estimated parametric model. For the bth permutation, these observations are denoted $\boldsymbol{X}^{rb} = (X_1^{rb}, X_2^{rb}, \ldots, X_n^{rb})/\boldsymbol{Y}^{rb} = (Y_1^{rb}, Y_2^{rb}, \ldots, Y_n^{rb})$ for the start/end of the fire season.

3. Using the simulated observations from step 2, refit the AR(1) linear regression model (3.1)/(3.2).

4. Set b to $b+1$.

5. Repeat steps 2 to 4 a large number of times (e.g. $b = 1, \ldots, 1000$) to obtain the bootstrap distribution of the fitted values.

The $(\alpha/2)$th and $(1-\alpha/2)$th quantiles of the ECDF of the fitted values provide the desired $(1-\alpha)100\%$ bootstrap pointwise confidence intervals.

We test for significant trends in the start and end of the fire season via a Wald test (Casella and Berger, 2002) and a permutation test (Davison and Hinkley, 2006). Regardless of the estimation method, we test the null hypothesis $H_0 : \beta^r_{X1} = 0$ versus the one-sided alternative $H_1 : \beta^r_{X1} < 0$ for trends in the start of the fire season. That is, the fire season is starting significantly earlier over the span of our data. Testing whether the fire season is ending later corresponds to $H_0 : \beta^r_{Y1} = 0$ versus $H_1 : \beta^r_{Y1} > 0$. The remainder of this section explains these tests in detail.

Wald Test

If we let t_0 denote the test statistic for the Wald test, then:

$$t_0 = \frac{\hat{\beta}^r_{X1} - 0}{\sigma_{\hat{\beta}^r_{X1}}} \sim t_{n-2} \tag{3.4}$$

where $\hat{\beta}^r_{X1}$ is the estimated slope in models for the start of the fire season and $\sigma_{\hat{\beta}^r_{X1}}$ is its standard error. The p-value can then be calculated as $\Pr(t > t_0 | \beta^r_{X1} = 0)$. For trends in the end of the fire season, we replace $\hat{\beta}^r_{X1}$ with $\hat{\beta}^r_{Y1}$ in (3.4) and the corresponding p-value is $\Pr(t < t_0 | \beta^r_{Y1} = 0)$. This test relies on asymptotic-based parametric inference and is appealing because of its simplicity and computational ease.

Permutation Test

Throughout this project, we also make use of resampling methods for inference rather than solely relying on maximum likelihood asymptotics. A class of such tests are permutation tests. These are nonparametric resampling techniques used to numerically estimate the

sampling distribution of a statistic under the null hypothesis. To test for significant trends we employ the following permutation algorithm.

1. Fit (3.1)/(3.2) to the observed data to model the estimates of the start/end of the fire season. From the AR(1) linear regression model, we obtain an estimate of the slope, $\hat{\beta}^r_{X1}/\hat{\beta}^r_{Y1}$.

2. Let b index the replication. Set b to 1.

3. Randomly permute the responses from the model in step 1 and let $\boldsymbol{X}^{rb} = (X^{rb}_1, X^{rb}_2, \ldots, X^{rb}_n)/\boldsymbol{Y}^{rb} = (Y^{rb}_1, Y^{rb}_2, \ldots, Y^{rb}_n)$ denote the bth random permutation of the response.

4. Refit (3.1)/(3.2) using the response vector from step 3. From the model, we obtain an estimate of the slope, $\hat{\beta}^{rb}_{X1}/\hat{\beta}^{rb}_{Y1}$.

5. Set b to $b + 1$.

6. Repeat steps 3 to 5 a large number of times (e.g. $b = 1, \ldots, 1000$) to obtain an accurate estimate of the sampling distribution of the estimated slope under the null hypothesis of no trend.

Letting I represent the indicator function, the p-value for trends in the start of the fire season is calculated as:

$$\frac{1}{B}\sum_{b=1}^{B} \mathrm{I}\left\{\hat{\beta}^{rb}_{X1} \leq \hat{\beta}^{r}_{X1}\right\}$$

and as:

$$\frac{1}{B}\sum_{b=1}^{B} \mathrm{I}\left\{\hat{\beta}^{rb}_{Y1} \geq \hat{\beta}^{r}_{Y1}\right\}$$

for trends in the end, where B represents the total number of permutations.

3.2 Results

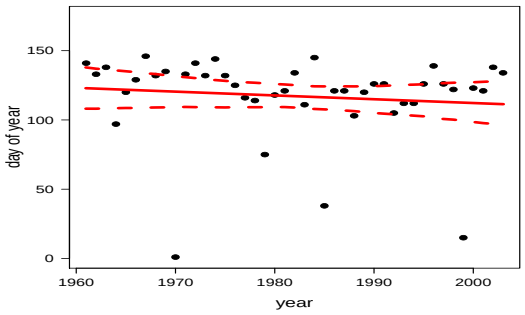
3.2.1 Analysis of Alberta Forest Fire Data

Figure 3.1 displays the empirical estimates of the start and end of the fire season when using annual counts of fire days to define the length. Overlayed are the estimated trends in the timing of the fire season with 95% boostrap-based confidence intervals of fitted values, as discussed in Section 3.1. The magnitudes of these trends, including standard errors and p-values from the significance tests are displayed in Table 3.1. The trend in the first annual fire day is not significant because of its high variability. Meanwhile, those of the fifth and tenth annual fire days are quite strong and both the Wald and permutation tests indicate that they are significantly negative. For the end of the fire season, all trends are all highly significant and quite close in magnitude, regardless of the threshold. Note that, in general, trends based on the first and last fire day have wider confidence intervals than those at other thresholds.

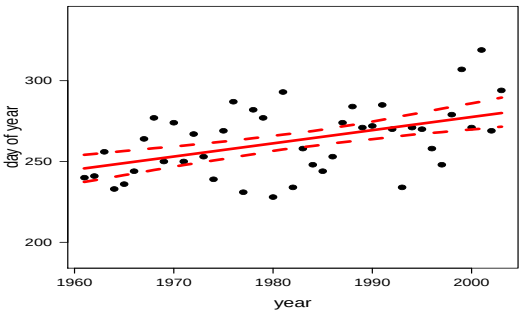
Analogous summaries estimated when using the first and last day a threshold in the ECDF of fire days is crossed each year are visually presented in Figure 3.2 with corresponding summary statistics in Table 3.2. Once again, the trend in the start of the fire season constructed from the first day the ECDF exceeds the 1% threshold each year is not significant. Despite being relatively strong in magnitude, the associated standard error is quite large. The trend estimate based on an annual crossing of a 5% ECDF threshold for the start of the fire season is marginally significant, while at the 10% threshold is not significant. For the analysis in the end of the fire season, all trends are highly significant, regardless of the threshold employed. Trends in the first day the ECDF exceeds 1% each year and the last day the ECDF fails to exceed 99% each year have relatively large standard errors.

Goodness of Fit

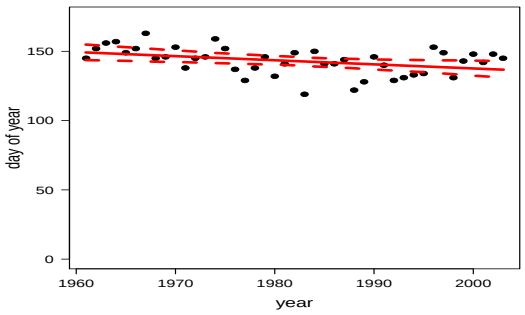
The goodness of fit for these regression models is examined using standard residual analysis techniques. Figure 3.3 displays QQ plots of the residuals for each fitted model. Note that these panels do not indicate any striking discrepancies, except for those of the first fire day and the first day the ECDF of fire days exceeds 1%.



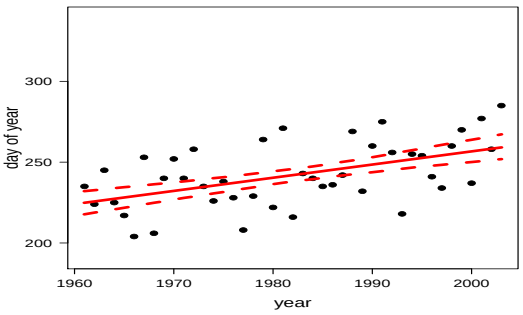
(a) First fire day.



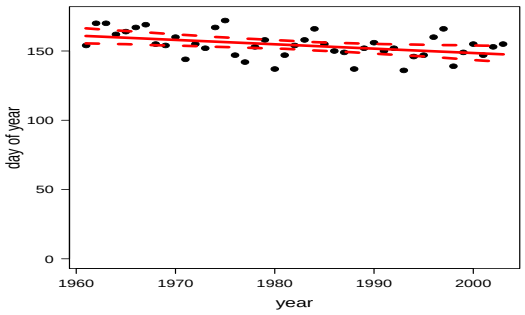
(b) Last fire day.



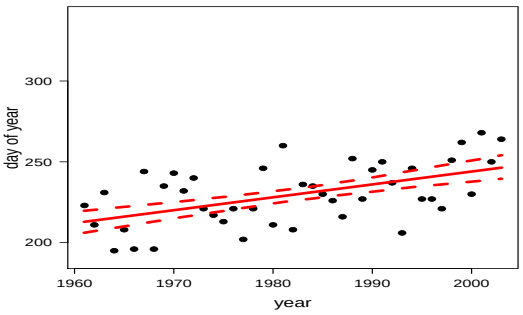
(c) Fifth fire day.



(d) Fifth last fire day.



(e) Tenth fire day.

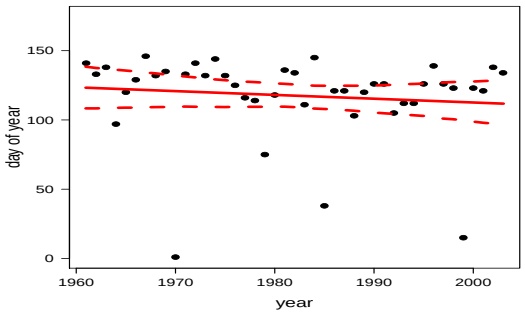


(f) Tenth last fire day.

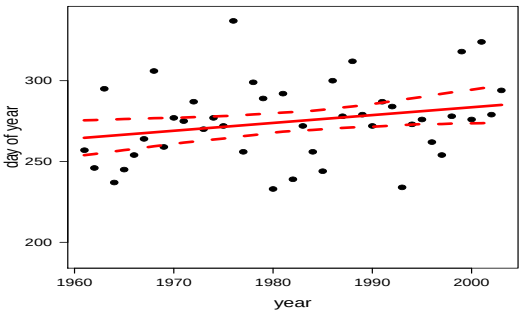
Figure 3.1: Empirical estimates of the start (left column) and end (right column) of the fire season for each year in Alberta (black points). Estimates are based on observed fire day counts. Overlayed are the fitted values from the linear models of the annual start and end of the fire season (solid red line), as appropriate, with 95% confidence intervals (dashed red lines).

Quantity	Threshold	Slope	SE	WT	PT
Start	1	-0.273	0.359	0.226	0.226
	5	-0.293	0.137	0.019	0.004
	10	-0.311	0.133	0.012	0
End	1	0.818	0.202	<0.001	0
	5	0.817	0.172	<0.001	0
	10	0.798	0.163	<0.001	0

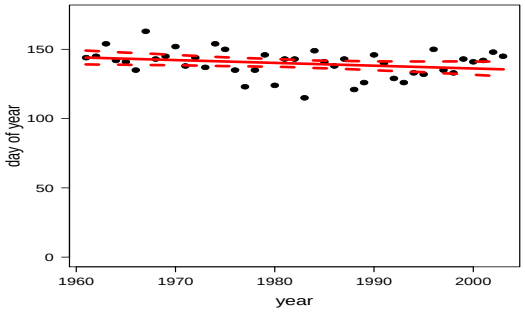
Table 3.1: Summary of trends, including the estimated slope and its standard error (SE), along with p-values from the Wald (WT) and permutation (PT) significance tests when using observed fire day counts to define the start and end of the fire season each year in Alberta.



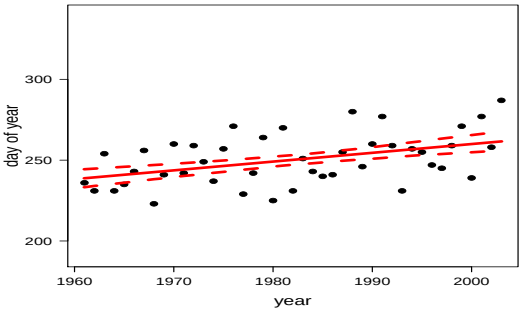
(a) First day ECDF exceeds 1%.



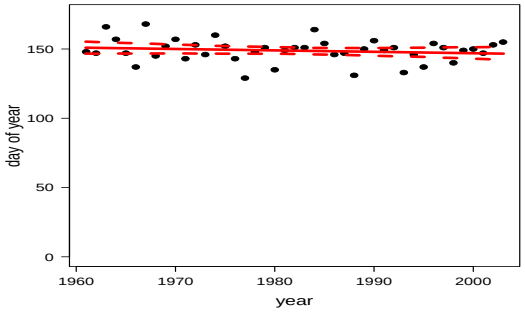
(b) Last day ECDF fails to exceed 99%.



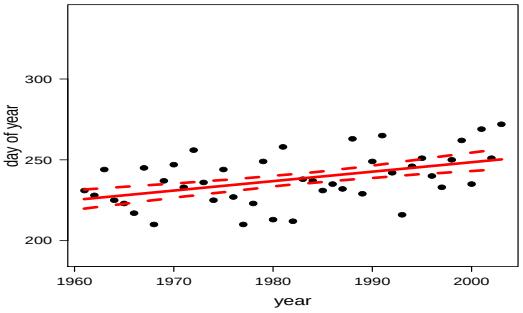
(c) First day ECDF exceed 5%.



(d) Last day ECDF fails to exceed 95%.



(e) First day ECDF exceeds 10%.

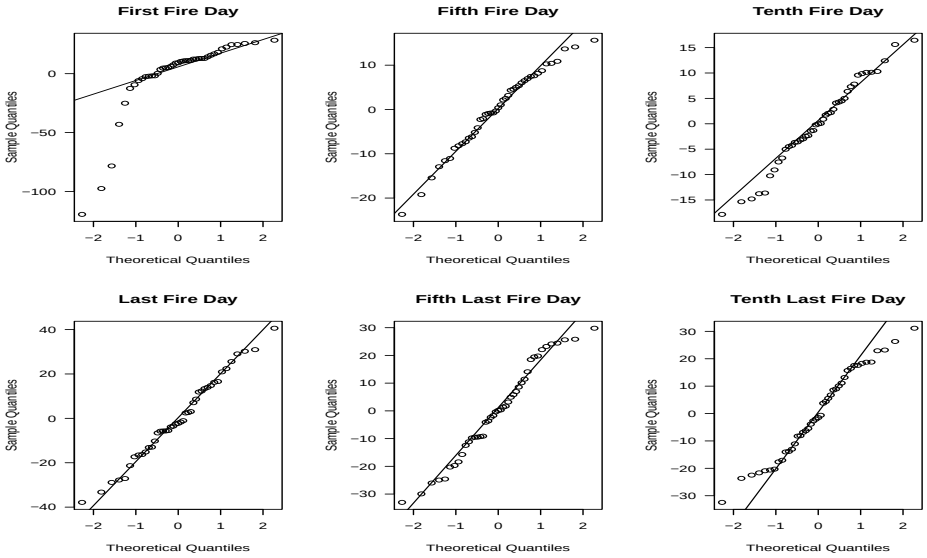


(f) Last day ECDF fails to exceed 90%.

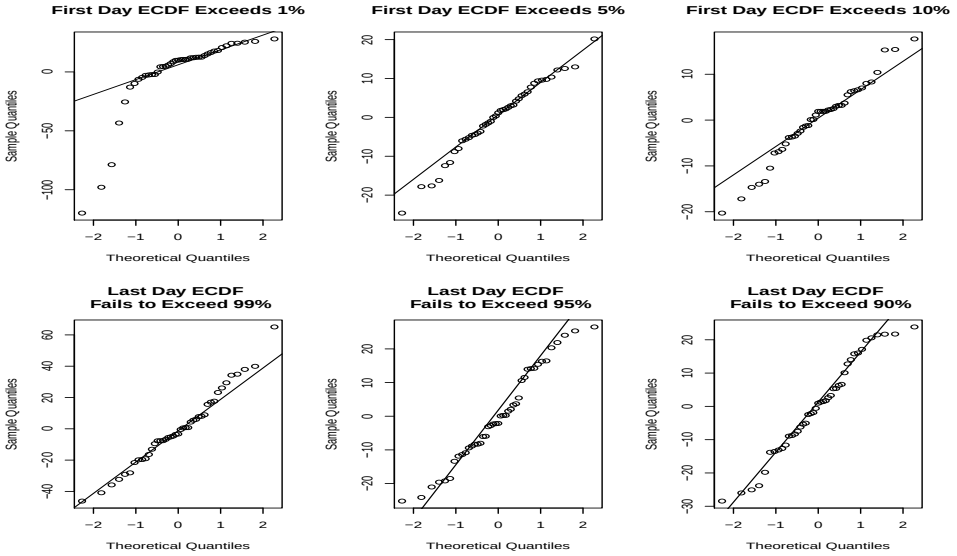
Figure 3.2: Empirical estimates of the start (left column) and end (right column) of the fire season for each year in Alberta (black points). Estimates are based on the ECDF of fire days. Overlayed are the fitted values from the linear models of the annual start and end of the fire season (solid red line), as appropriate, with 95% confidence intervals (dashed red lines).

Quantity	Threshold	Slope	SE	WT	PT
Start	1	-0.273	0.363	0.228	0.256
	5	-0.200	0.120	0.052	0.048
	10	-0.100	0.104	0.170	0.164
End	99	0.486	0.260	0.034	0.056
	95	0.545	0.132	<0.001	0.002
	90	0.586	0.142	<0.001	0

Table 3.2: Summary of trends, including the estimated slope and its standard error (SE), along with p-values from the Wald (WT) and permutation (PT) significance tests when using the ECDF of fire days to define the start and end of the fire season each year in Alberta.



(a) Residuals from the linear models of the observed fire day counts for the start and end of the fire season.



(b) Residuals from the linear models of the ECDF of fire days for the start and end of the fire season.

Figure 3.3: Residual QQ plots from the linear models of the observed fire day counts (upper panels) and the ECDF of fire days (lower panels) for the start and end of the fire season in Alberta.

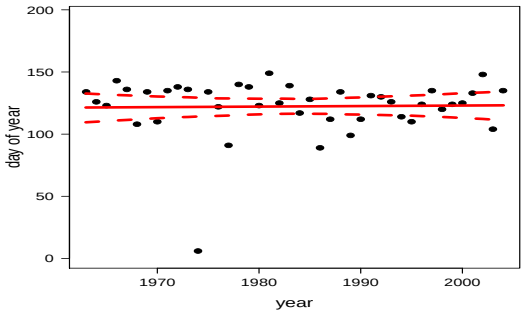
3.2.2 Analysis of Ontario Forest Fire Data

Trends in the Ontario fire data, when defining the start of the fire season by the count of fire days crossing a threshold each year, are displayed in Figure 3.4 and summarized in Table 3.3. No significance is found when defining the start of the fire season as the first, fifth, or tenth fire day each year. Corresponding trends in the end of the fire season are marginally significant when using the fifth last and tenth last annual fire day. The trend in the timing of the last fire day each year is not significant.

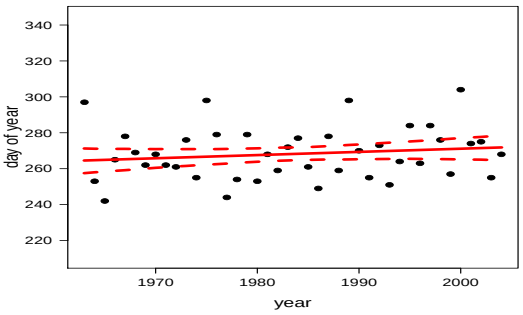
We now turn our attention to estimating trends by defining the start of the fire season as the crossing of 1%, 5%, and 10% thresholds in the ECDF of fire days and conversely for the end. Consider Figure 3.5 and Table 3.4. As with the fire day counts discussed previously, no significance is found in the start of the Ontario fire season. However, trends in the end of the fire season are relatively strong when looking at the last day the ECDF of fire days fails to exceed 95% and 90% annually; the p-values from both the Wald and permutation tests are quite small. Trends based on the last day each year the ECDF fails to exceed 99% are quite weak and neither the p-value from the Wald test nor the p-value from the permutation test indicate significance.

Goodness of Fit

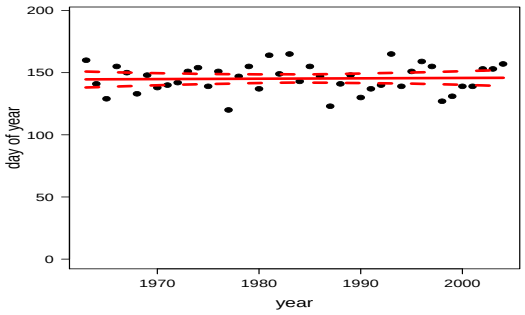
Once again, model goodness of fit is assessed via residual QQ plots. Figure 3.6 displays the QQ plots separately for each model constructed. These panels only indicate discrepancies at smaller thresholds.



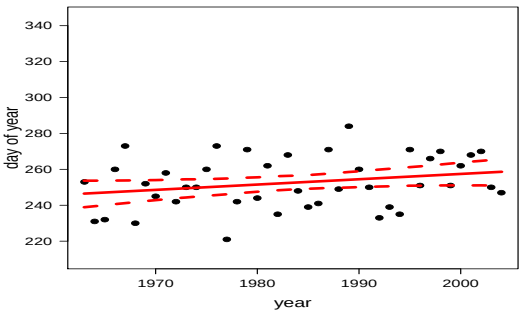
(a) First fire day.



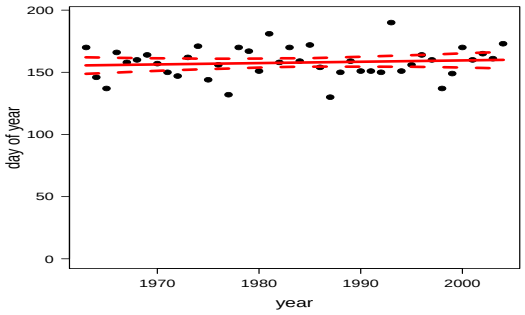
(b) Last fire day.



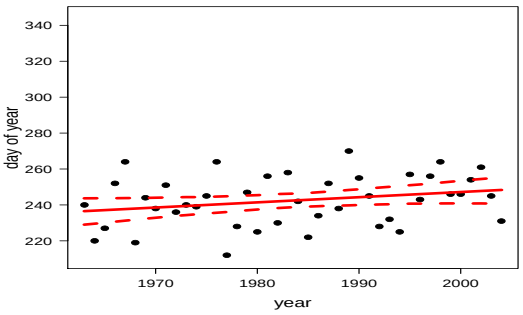
(c) Fifth fire day.



(d) Fifth last fire day.



(e) Tenth fire day.

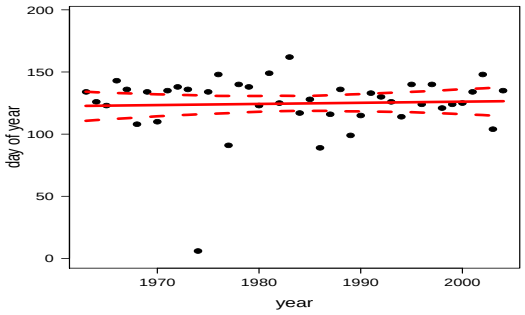


(f) Tenth last fire day.

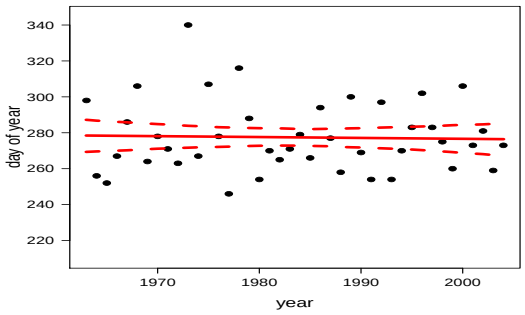
Figure 3.4: Empirical estimates of the start (left column) and end (right column) of the fire season for each year in Ontario (black points). Estimates are based on observed fire day counts. Overlayed are the fitted values from the linear models of the annual start and end of the fire season (solid red line), as appropriate, with 95% confidence intervals (dashed red lines).

Quantity	Threshold	Slope	SE	WT	PT
Start	1	0.043	0.266	0.563	0.522
	5	0.030	0.149	0.578	0.610
	10	0.107	0.154	0.755	0.774
End	1	0.174	0.159	0.141	0.154
	5	0.294	0.171	0.046	0.062
	10	0.285	0.170	0.051	0.060

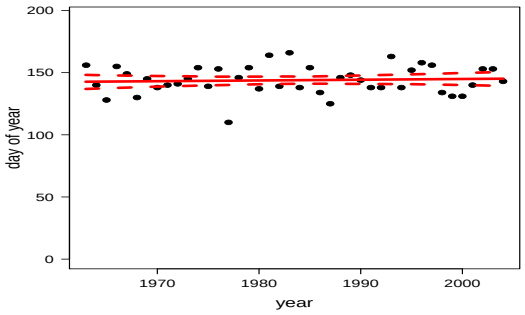
Table 3.3: Summary of trends, including the estimated slope and its standard error (SE), along with p-values from the Wald (WT) and permutation (PT) significance tests when using observed fire day counts to define the start and end of the fire season each year in Ontario.



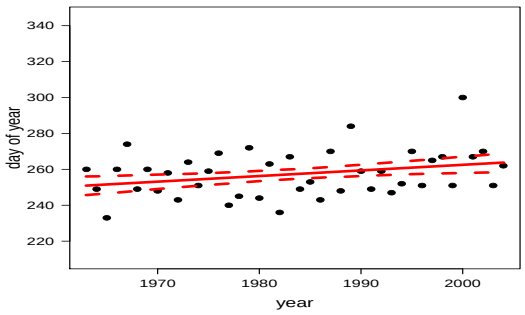
(a) First day ECDF exceeds 1%.



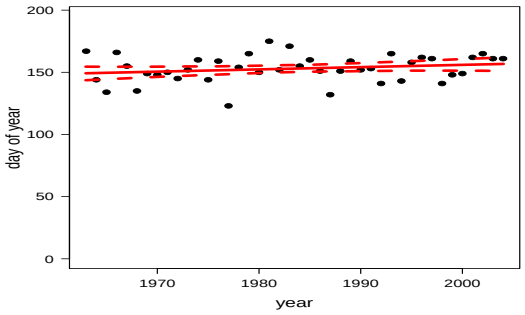
(b) Last day ECDF fails to exceed 99%.



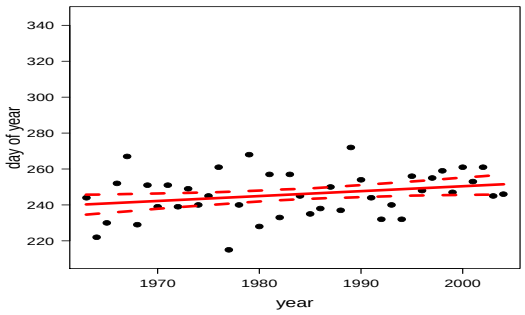
(c) First day ECDF exceeds 5%.



(d) Last day ECDF fails to exceed 95%.



(e) First day ECDF exceeds 10%.

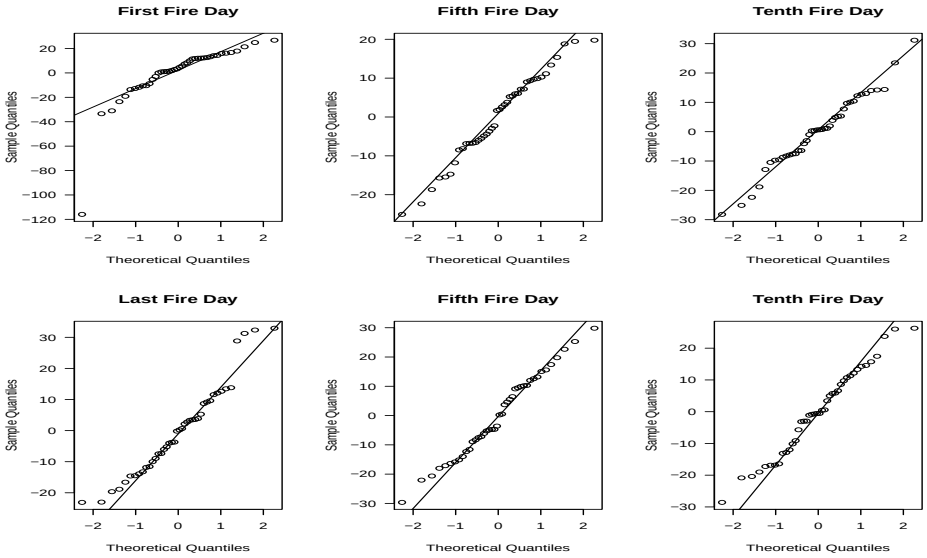


(f) Last day ECDF fails to exceed 90%.

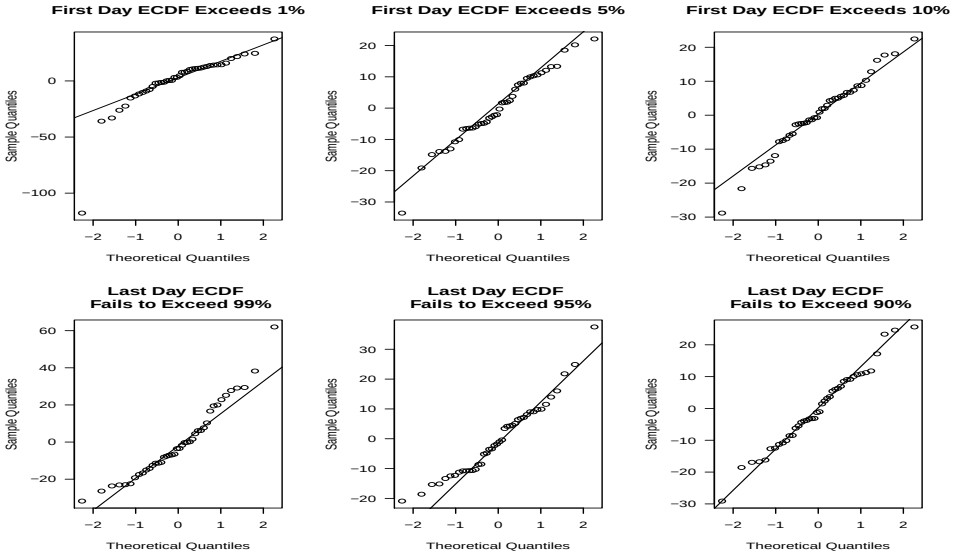
Figure 3.5: Empirical estimates of the start (left column) and end (right column) of the fire season for each year in Ontario (black points). Estimates are based on the ECDF of fire days. Overlayed are the fitted values from the linear models of the annual start and end of the fire season (solid red line), as appropriate, with 95% confidence intervals (dashed red lines).

Quantity	Threshold	Slope	SE	WT	PT
Start	1	0.091	0.268	0.632	0.580
	5	0.058	0.130	0.670	0.680
	10	0.181	0.126	0.921	0.910
End	99	-0.051	0.207	0.597	0.560
	95	0.312	0.121	0.007	0.032
	90	0.273	0.129	0.022	0.042

Table 3.4: Summary of trends, including the estimated slope and its standard error (SE), along with p-values from the Wald (WT) and permutation (PT) significance tests when using the ECDF of fire days to define the start and end of the fire season each year in Ontario.



(a) Residuals from the linear models of the observed fire day counts for the start and end of the fire season.



(b) Residuals from the linear models of the ECDF of fire days for the start and end of the fire season.

Figure 3.6: Residual QQ plots from the linear models of the observed fire day counts (upper panels) and the ECDF of fire days (lower panels) for the start and end of the fire season in Ontario.

Chapter 4

Investigating Trends in Nonparametric Estimates of Fire Season Length

In this chapter, we focus on a two stage approach for estimating trends in the start and end of the fire season. Section 4.1 outlines this method, detailing the models used along with the significance tests employed. The results, when applied to the Alberta and Ontario forest fire data sets, are the subject of Section 4.2, where trends are discussed and model goodness of fit is assessed. Finally, in Section 4.3, comparisons are made between the trends from our two stage nonparametric approach to those of the empirical estimates presented in the previous chapter.

4.1 Nonparametric Approach to Estimating Trends in the Start and End of the Fire Season

4.1.1 Stage 1: Estimating the Start and End of the Fire Season

From Chapter 1, recall the definition of a fire day as a day where one or more fires were reported. Let Z_t represent the fire day indicator:

$$Z_t = \begin{cases} 1, & \text{if day } t \text{ is a fire day} \\ 0, & \text{otherwise.} \end{cases}$$

The time index is denoted t, where $t = 1, \ldots, T$, with T being the number of observations in our data set. In vector notation for year i, let $\boldsymbol{t}_i = (\boldsymbol{t}_{1+(i-1)365}, \boldsymbol{t}_{2+(i-1)365}, \ldots, \boldsymbol{t}_{365+(i-1)365})$ for $i = 1, \ldots, n$. Then, $\boldsymbol{t} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n)$. If we let $\mathrm{E}(Z_t|t) = p_t$, then p_t represents the probability of day t being a fire day. Estimates of p_t are termed the estimated risk of a fire day at time t. This probability is modelled using the following logistic GAM:

$$\mathrm{logit}(p_t) = \beta_0 + f(t) \tag{4.1}$$

with β_0 begin an intercept parameter and $f(t)$ being a thin plate regression spline. The flexibility inherent in this basis function was required to adequately characterize the shape of individual fire seasons as we expect there to be one mode within each season, dropping to periods of nil risk between fire seasons. To visualize this, refer to Figure 4.1 for fitted values and 95% confidence intervals from model 4.1 for the first three years of the Alberta fire data. In this data set, $t = 1$ corresponds to January 1, 1961, while $t = T$ is December 31, 2003, with $T = 15,695$. For the Ontario data set, $T = 15,330$ corresponding to forty-two years of fire records. Note that this model can be extended to include other covariates, such as fire-weather and climate indices. We return to this discussion in Chapter 5.

As mentioned previously, the fire season, in both Alberta and Ontario, officially runs from April 1 - October 31 in terms of fire management operations. However, in order to test for significant temporal trends we redefine the start and end of the fire season, allowing these to change dynamically over years. Specifically, we define these quantities as the annual crossing of a fixed risk threshold: the start of the fire season is the first day the estimated risk of a fire day exceeds a predefined threshold, and the end of the fire season is defined as the last day the estimated risk exceeds that threshold. Figure 4.2 illustrates this concept using a 5% risk threshold for the 1965 subset of the estimated smoother for Alberta.

Figure 4.3 displays the estimated risk during 1965 in Alberta as well as 95% pointwise confidence bands for such risks, shaded in gray. These confidence bands can be inverted and used to identify two-sided confidence intervals for the start and end of the fire season by finding all values of t for which we may not reject the hypothesis that the risk is 5%. The confidence limits for the start of the fire season correspond to the first day the upper and lower confidence bands in Figure 4.3 cross the threshold of 5%. For the end of the fire season, the confidence interval corresponds to the last day the lower and upper confidence bands exceed 5%. These confidence limits are highlighted by the red vertical lines in Figure 4.3. In

general, such confidence intervals may consist of a union of disjoint intervals depending on the form of the function, $f(t)$. However, with the smoother employed in our context, they consist of a single connected interval. We compute the standard error of the estimated start and end of the fire season by dividing the width of the corresponding confidence interval obtained through this inversion procedure by 2×1.96, or 3.92, as would be appropriate for confidence intervals derived from a normal distribution. The normality assumption permits the use of linear models at the second stage of this analysis, discussed in Section 4.1.2. These procedures employ likelihood-based parameter estimation and are therefore robust to departures from the normality assumption (Montgomery et al., 2006). We also comment that these confidence intervals are not guaranteed to be symmetric. However, as our results show in Section 4.2, they are approximately symmetric, with those calculated at higher fire day risk thresholds being closer to symmetric.

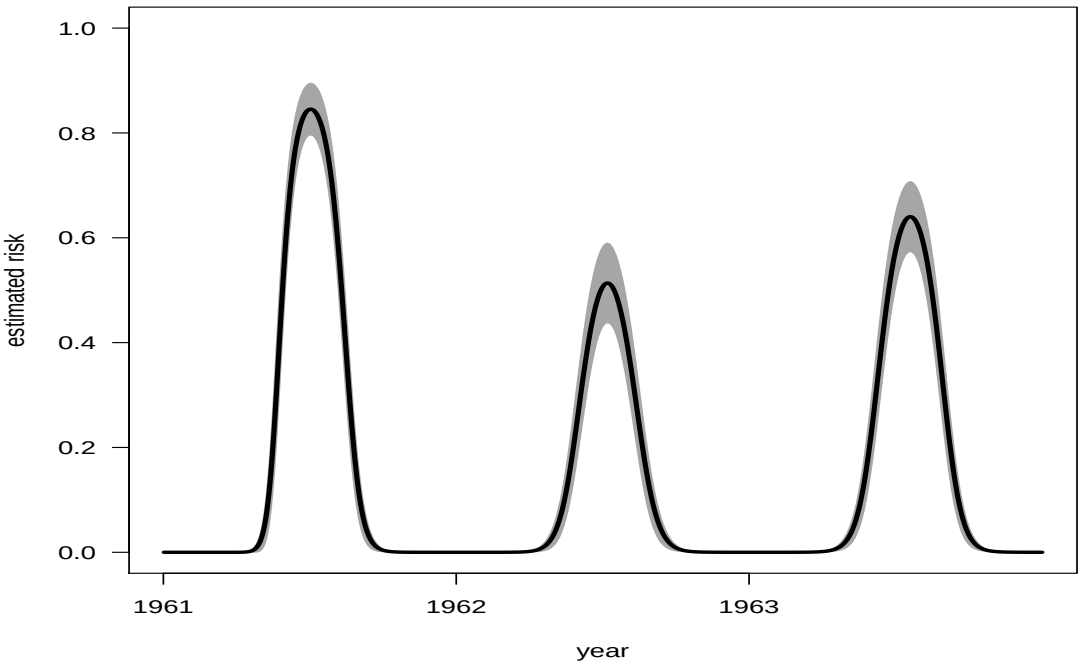


Figure 4.1: Estimated fire day risk (black curve) and 95% confidence intervals (gray shaded region) from the GAM for the first three years of Alberta fire data.

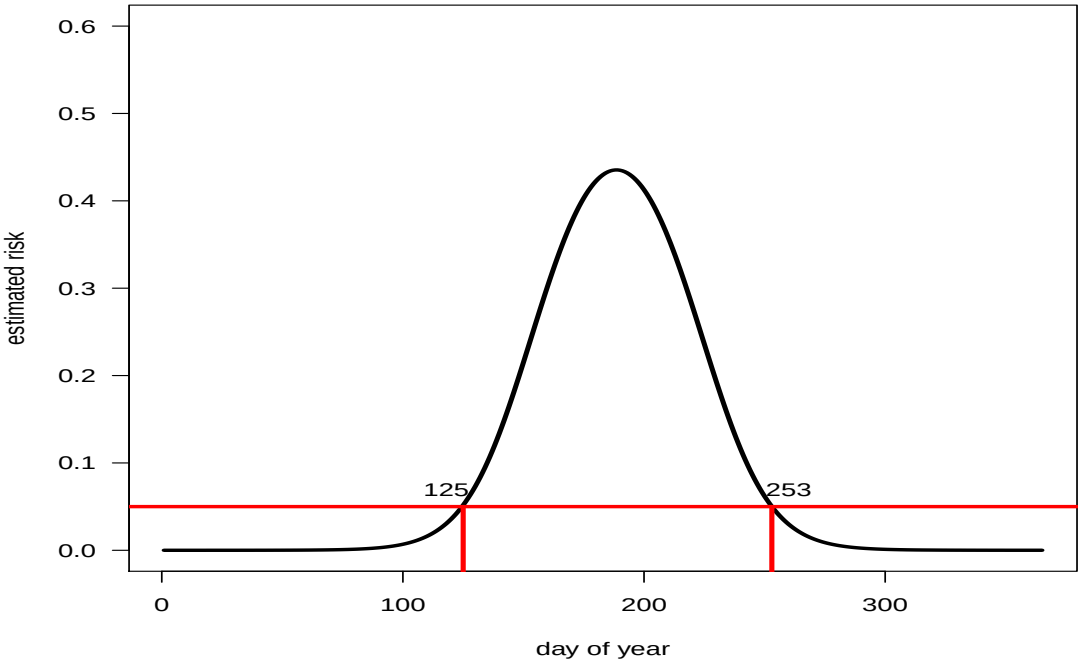


Figure 4.2: Estimated fire day risk (black curve) from the GAM for the period of 1965 in Alberta. Point estimates for the start and end of the fire season are identified (vertical red lines) for a 5% fire day risk threshold (red horizontal line).

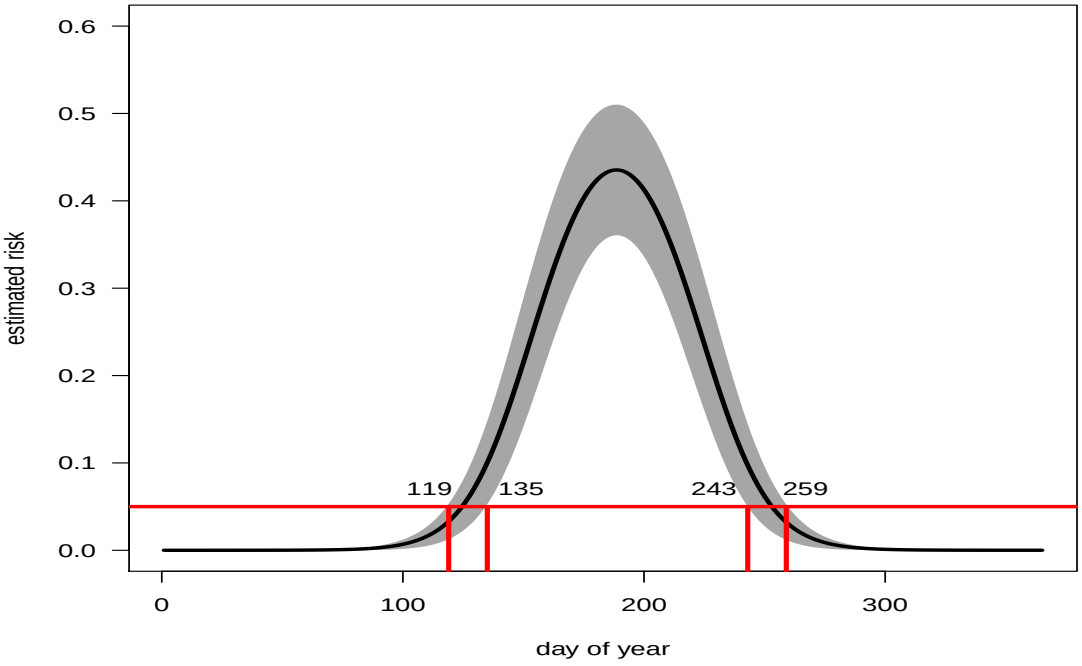


Figure 4.3: Estimated fire day risk (black curve) from the GAM with 95% confidence intervals (grey shaded region) for the period of 1965 in Alberta. Overlayed are the inverse confidence intervals for the start and end of the fire season (red vertical lines) for a 5% fire day risk threshold (red horizontal line).

4.1.2 Stage 2: Quantifying Trends in the Start and End of the Fire Season

Using the notation defined in Chapter 3, let X_i^r and Y_i^r denote the estimated start and end of the fire season, defined by the rth threshold for year i, $i = 1, \ldots, n$. Although a threshold of 5% is used throughout the examples in this section, note that any threshold can be employed in practice; we also consider thresholds of 1% and 10% for comparative purposes in Section 4.2. In Section 4.1.1, we described methods for estimating X_i^r and Y_i^r for specific values of r and approximating their standard errors, denoted $V_{X_i^r}^{-1/2}$ and $V_{Y_i^r}^{-1/2}$. In order to quantify trends, we model X_i^r and Y_i^r using linear regression with the inverse of their standard errors as known weights incorporated to handle the heteroscedasticity of the response and an AR(1) error to account for possible short-term correlation in the behaviour of fire risk from year to year. Let $\alpha_{X1}^r/\alpha_{Y1}^r$ denote the trend parameter (slope) in such models for the start/end of the fire season.

We contrast three different techniques to test for significant trends: a Wald test and two different permutation techniques. As in Chapter 3, we are testing the null hypothesis $H_0 : \alpha_{X1}^r = 0$ versus the one-sided alternative $H_1 : \alpha_{X1}^r < 0$ for trends in the start of the fire season. For trends in the end of the fire season, we test $H_0 : \alpha_{Y1}^r = 0$ versus $H_1 : \alpha_{Y1}^r > 0$. We omit a discussion of the Wald test from this section, as it was previously discussed in Section 3.1. The remainder of this section details the two permutation test algorithms.

Stage 1 Block Permutation Test

The first permutation test block permutes the data, using years as blocks, and repeats the full two stage analysis to develop a sampling distribution for the estimated slope. The algorithm for this test follows.

1. Perform the full two stage analysis on the observed data (c.f. 4.1) to estimate the start/end of the fire season each year and subsequently the trend in the start/end of the fire season. We denote the slope estimate $\hat{\alpha}_{X1}^r/\hat{\alpha}_{Y1}^r$.

2. Let b index the replication. Set b to 1.

3. Randomly permute the fire day indicator, grouping together blocks of observations within each year so that the order of the observations within each year remains unchanged. Let i^b index the resampled years, where $i^b = 1^b, 2^b, \ldots, n^b$, a permutation of the vector $1, 2, \ldots, n$. The vector of fire day indicators then becomes $\boldsymbol{Z}_t^b =$

$\left(\boldsymbol{Z}_{1+(i^b-1)365}, \boldsymbol{Z}_{2+(i^b-1)365}, \ldots, \boldsymbol{Z}_{365+(i^b-1)365}\right)$.

4. Model the permuted responses using (4.1). Then, use the approach described in Section 4.1.1 to obtain point estimates and confidence intervals for the annual start/end of the fire season based on the output from (4.1).

5. Fit a weighted AR(1) linear regression model (c.f. Section 4.1.2) to estimate the slope, denoted $\hat{\alpha}^{rb}_{X1}/\hat{\alpha}^{rb}_{Y1}$.

6. Set b to $b+1$.

7. Repeat steps 3 to 6 a large number of times (e.g. $b = 1, \ldots, 1000$) to obtain an accurate estimate of the sampling distribution of the estimated slope under the null hypothesis of no trend.

The p-value for trends in the start of the fire season is calculated as the proportion of replications where $\hat{\alpha}^{rp}_{X1}$ is less than $\hat{\alpha}^{r}_{X1}$. Correspondingly, for the end of the fire season, the p-value is the proportion of replications where $\hat{\alpha}^{rp}_{Y1}$ is greater than $\hat{\alpha}^{r}_{Y1}$.

Refitting the GAM during each replication is computationally intensive. We therefore consider a second approach which permutes the data only at the second stage. It is likely that X^r_i and Y^r_i are primarily influenced by data within year i rather than the full data series across years, as fire risk drops to zero at the end of each year. Therefore, this approach is justified.

Stage 2 Permutation Test

The second permutation method offers computational simplicities, with resampling occurring only at the second stage of the analysis, as described in the algorithm below.

1. Perform the full two stage analysis on the observed data (c.f. 4.1) to estimate the start/end of the fire season each year and subsequently the trend in the start/end of the fire season. We denote the slope estimate $\hat{\alpha}^{r}_{X1}/\hat{\alpha}^{r}_{Y1}$.

2. Let b index the replication. Set b to 1.

3. Randomly permute the response and weight pairs from the weighted AR(1) linear regression model in step 1, letting $\boldsymbol{X}^{rb} = (X^{rb}_1, X^{rb}_2, \ldots, X^{rb}_n)/\boldsymbol{Y}^{rb} = (Y^{rb}_1, Y^{rb}_2, \ldots, Y^{rb}_n)$

denote the bth permutation of the response vector and $\boldsymbol{V}_{X^r}^{-1/2} = (V_{X_1^r}^{-1/2}, V_{X_2^r}^{-1/2}, \ldots, V_{X_n^r}^{-1/2})/\boldsymbol{V}_{Y^r}^{-1/2} = (V_{Y_1^r}^{-1/2}, V_{Y_2^r}^{-1/2}, \ldots, V_{Y_n^r}^{-1/2})$ denote the bth permutation of weight vector.

4. Fit a weighted AR(1) linear regression model (c.f. Section 4.1.2) to estimate the slope, denoted $\hat{\alpha}_{X1}^{rb}/\hat{\alpha}_{Y1}^{rb}$.

5. Set b to $b+1$.

6. Repeat steps 3 to 5 a large number of times (e.g. $b = 1, \ldots, 1000$) to obtain an accurate estimate of the sampling distribution of the estimated slope under the null hypothesis of no trend.

Again, the p-value for trends in the start of the fire season corresponds to the proportion of replications where $\hat{\alpha}_{X1}^{rp}$ is less than $\hat{\alpha}_{X1}^{r}$. The proportion of replications where $\hat{\alpha}_{Y1}^{rp}$ is greater than $\hat{\alpha}_{Y1}^{r}$ corresponds to the p-value for trends in the end of the fire season.

4.2 Results

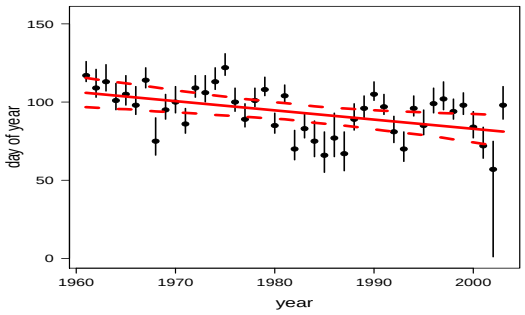
4.2.1 Analysis of Alberta Forest Fire Data

Figure 4.4 summarizes the point estimates and inverse confidence intervals for the start and end of the fire season using fire day risk thresholds of 1%, 5%, and 10%. Overlayed on these panels are the fitted values from the linear models and 95% bootstrap-based confidence intervals for the response (c.f. Section 3.1). Table 4.1 quantifies these trends, summarizing the slopes from the linear models and their corresponding standard errors as well as p-values from the tests for significant trends. Panels 4.4(a), 4.4(c), and 4.4(e) in Figure 4.4, show strong, negative trends, suggesting that the fire season is starting earlier regardless of the threshold used to define it. The steepest trend occurs when utilizing a threshold of 1%, where we estimate that the fire season is starting approximately 0.6 days earlier per year. This corresponds to twenty-five days over the forty-three years of observed Alberta data. At the 5% and 10% thresholds, Alberta's fire season is estimated to start eighteen and fifteen days earlier, respectively. The Wald and permutation tests indicate that these trends are statistically significant. As well, there are only minor discrepancies between the p-values from these three significance tests. Notice, from Table 4.1, that trends are

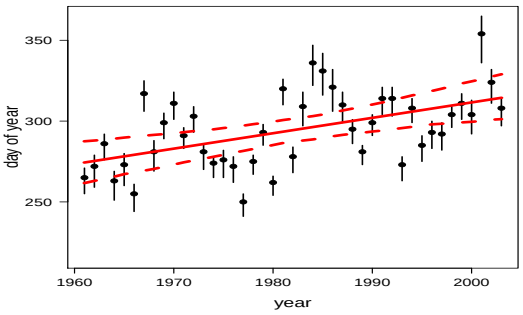
more pronounced for the end of the fire season. Our models predict that Alberta's fire season is ending forty-one, thirty-five and thirty-three days later in 2003 than it was in 1961, at risk thresholds of 1%, 5%, and 10%, respectively. All estimated trends are highly significant. Therefore, regardless of the threshold employed, our approach suggests that there is a significant lengthening of the forest fire season in Alberta due significant negative trends in the start of the fire season and positive trends for the end of the season.

Goodness of Fit

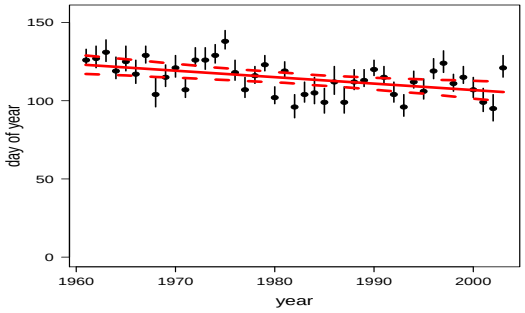
Goodness of fit for our two stage approach is conducted at both the first stage, on the GAM, and the second stage, on the linear models. For the GAM, goodness of fit consists of a comparison between the expected and the observed number of fire days, aggregated by day of year and by year. These summaries are displayed in Figure 4.5 and show close agreement between these two quantities. In particular, note how closely the expected and observed number of fire days in Figure 4.5(b) agree; the GAM essentially interpolates the annual total number of fire days across years. This result is expected, as a large number of knots are required in order to ensure that the model is flexible enough to drop to a period of nil risk between each fire season. We also note that, in addition to the correlated residuals in the linear models, correlation is present between day to day observations in the logistic GAM. However, this has not yet been incorporated into the model and we return to this discussion in Chapter 5. Goodness of fit in the second stage linear models is again assessed via residual QQ plots, shown in Figure 4.6. These diagnostics do not indicate any glaring problems with model fit.



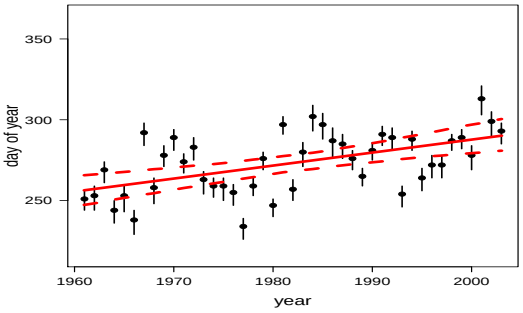
(a) First day 1% threshold is exceeded.



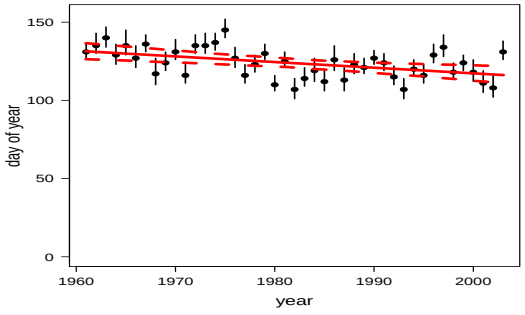
(b) Last day 1% threshold is exceeded.



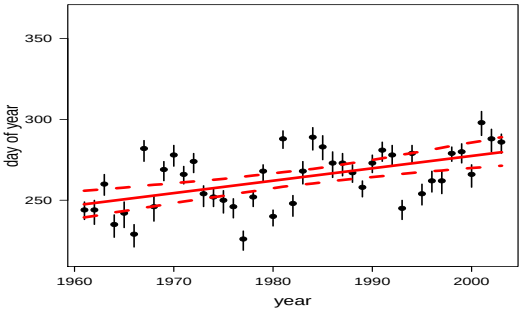
(c) First day 5% threshold is exceeded.



(d) Last day 5% threshold is exceeded.



(e) First day 10% threshold is exceeded.

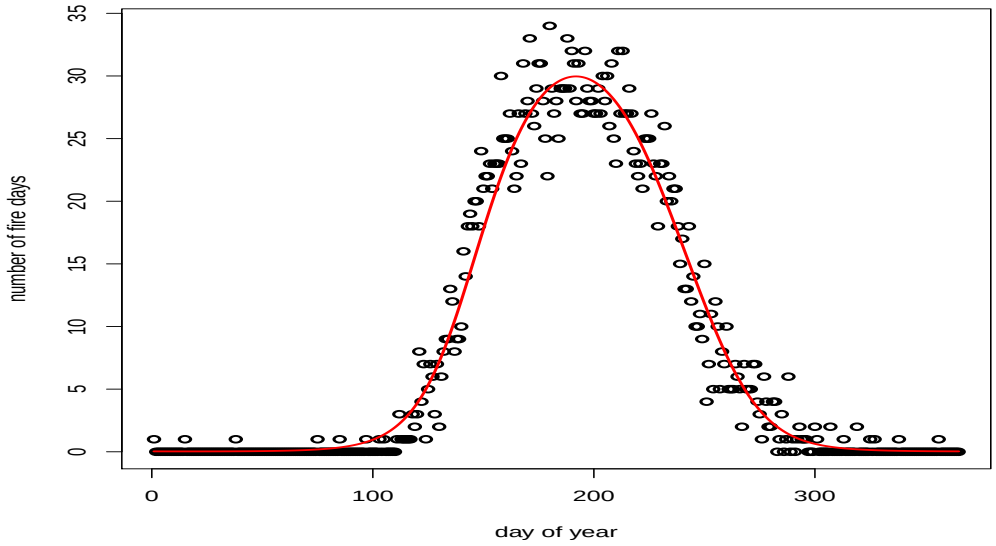


(f) Last day 10% threshold is exceeded.

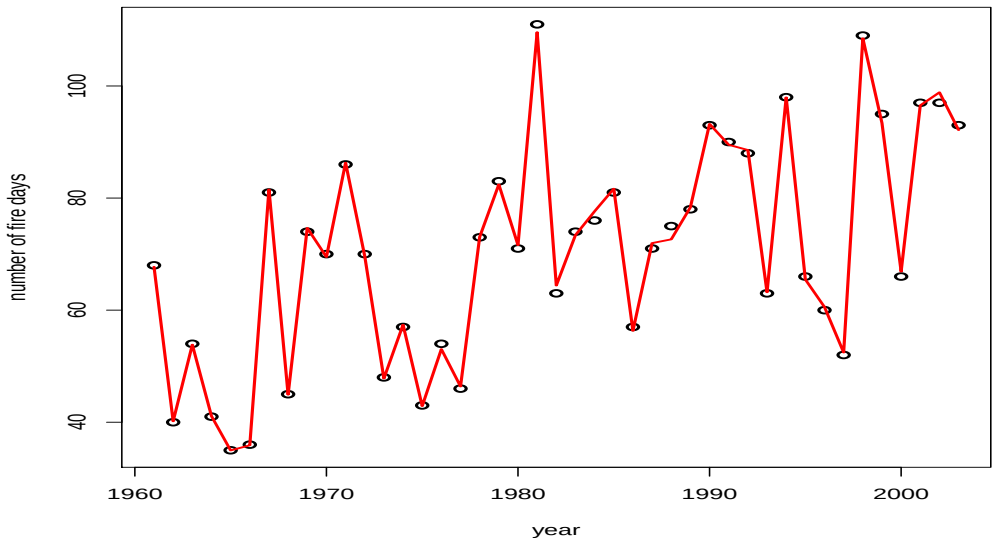
Figure 4.4: Estimates (black points) and confidence intervals (black lines) for the start (left column) and end (right column) of the fire season in Alberta. Overlayed are the trends (red solid line) and 95% confidence intervals for the trends (red dashed lines).

	Threshold	Slope	SE	WT	1BPT	2PT
Start	1	-0.590	0.215	0.004	0.020	0
	5	-0.413	0.135	0.002	0.002	0
	10	-0.361	0.117	0.002	0	0
End	1	0.953	0.296	0.001	0	0
	5	0.804	0.210	<0.001	0	0
	10	0.767	0.189	<0.001	0	0

Table 4.1: Summary of trends, including the estimated slope and its standard error (SE), along with p-values from the Wald (WT), stage 1 block permutation (1BPT) and stage 2 permutation (2PT) significance tests for trends in the start and end of the Alberta fire season.



(a) Number of fire days aggregated by day of year.



(b) Number of fire days aggregated by year.

Figure 4.5: Observed (black points) and expected (red solid lines) number of fire days aggregated by day of year, and year from the GAM when applied to the Alberta fire data.

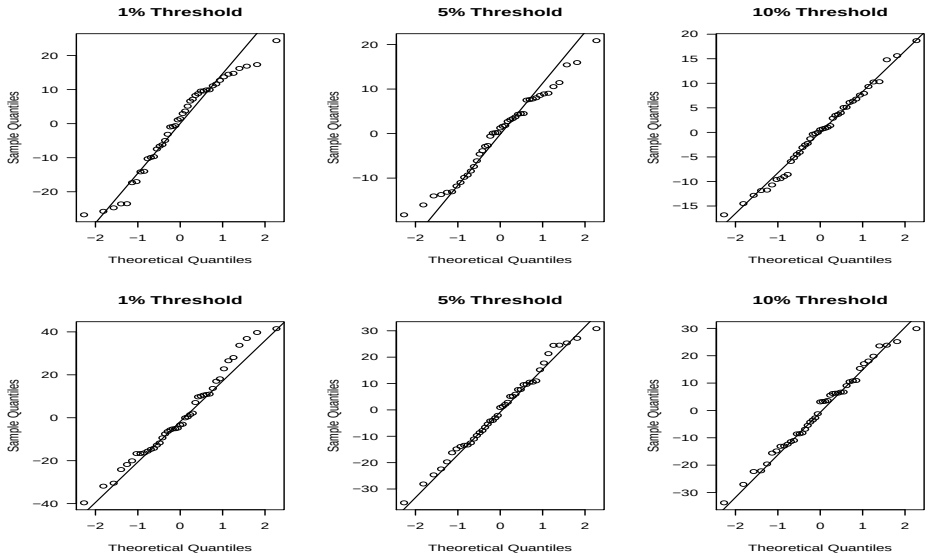


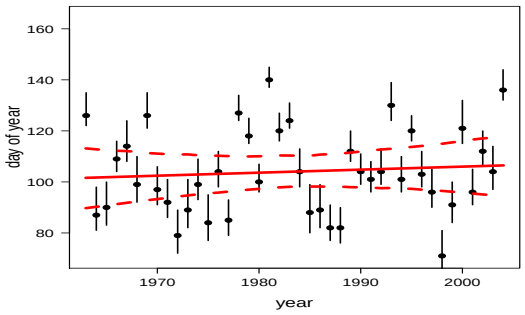
Figure 4.6: Residual QQ plots from the linear models fit to the nonparametric estimates of the start (top row) and end (bottom row) of the fire season in Alberta.

4.2.2 Analysis of Ontario Forest Fire Data

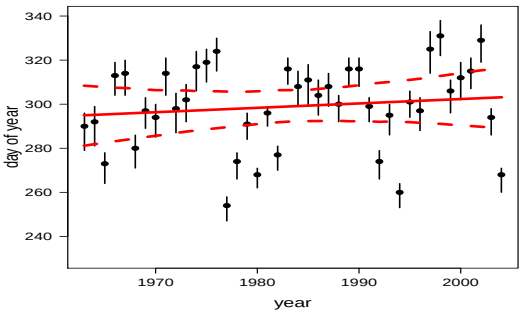
Figure 4.7, displays the fitted values and 95% bootstrap-based confidence intervals from the linear models. Again, these trends are overlayed on the point estimates and inverse confidence intervals for the start or end of the fire season, as indicated. No overwhelming trends are observed for the beginning of the fire season, as displayed in the right column of Figure 4.7. Table 4.2 summarizes the trends in the start of the fire season; the estimated slopes are quite small in comparison to their standard errors. Additionally, the corresponding p-values displayed in this table are large. Hence, there are no significant trends detected in the start of the fire season from 1963-2004 in Ontario. Trends in the end of the fire season are also displayed in Figure 4.7 and summarized in Table 4.2. The slope estimates are largest for the 10% threshold. As well, it is only for this threshold that there is significant evidence against the hypothesis of no trend. Marginal significance is detected for the 5% threshold. The p-values in Table 4.2 decrease as the thresholds increase. Note that for the Ontario data, there are inconsistencies between the results for the three significance tests, with the Wald test being the most conservative. This is in contrast to the results for Alberta, where trends in both the start and end of the fire season were quite strong, and the p-values from all three tests closely agreed.

Goodness of Fit

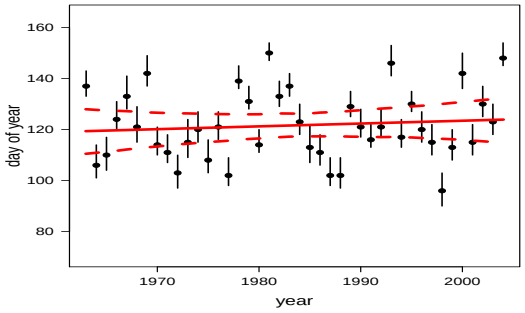
Goodness of fit is again assessed for the Ontario GAM via a comparison of the observed and expected number of fire days, and for the corresponding linear models through the use of residual QQ plots. Figure 4.8 summarizes the goodness of fit results in the former model and the residuals in the latter are displayed in Figure 4.9. The expected and observed number of fire days, when aggregated by both day of year, and year show close agreement. The residual QQ plots do not indicate any substantial problems with model fit.



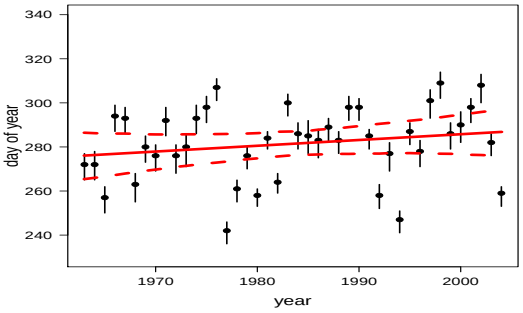
(a) First day 1% threshold is exceeded.



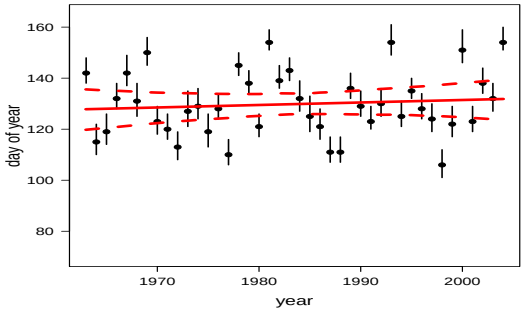
(b) Last day 1% threshold is exceeded.



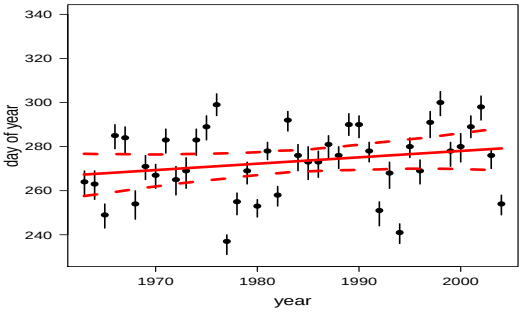
(c) First day 5% threshold is exceeded.



(d) Last day 5% threshold is exceeded.



(e) First day 10% threshold is exceeded.

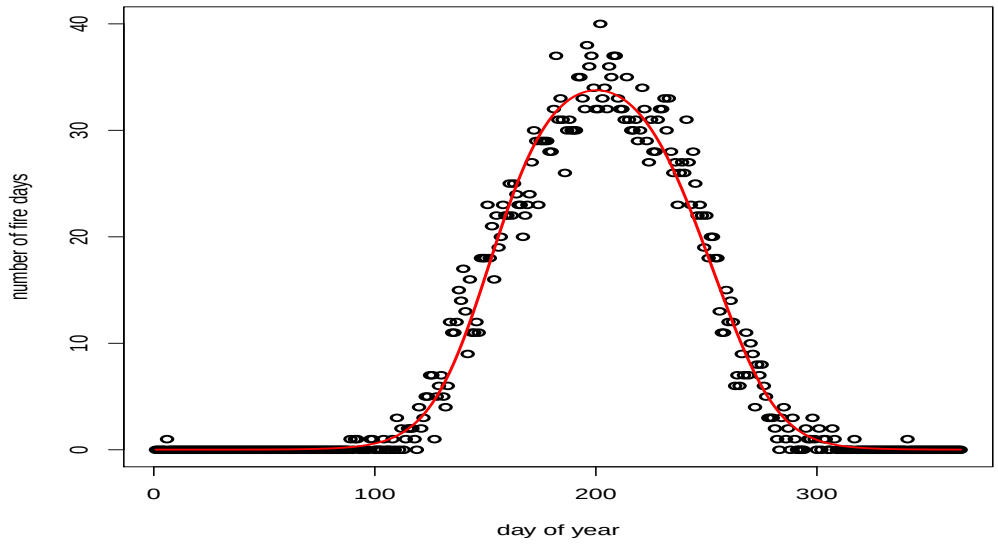


(f) Last day 10% threshold is exceeded.

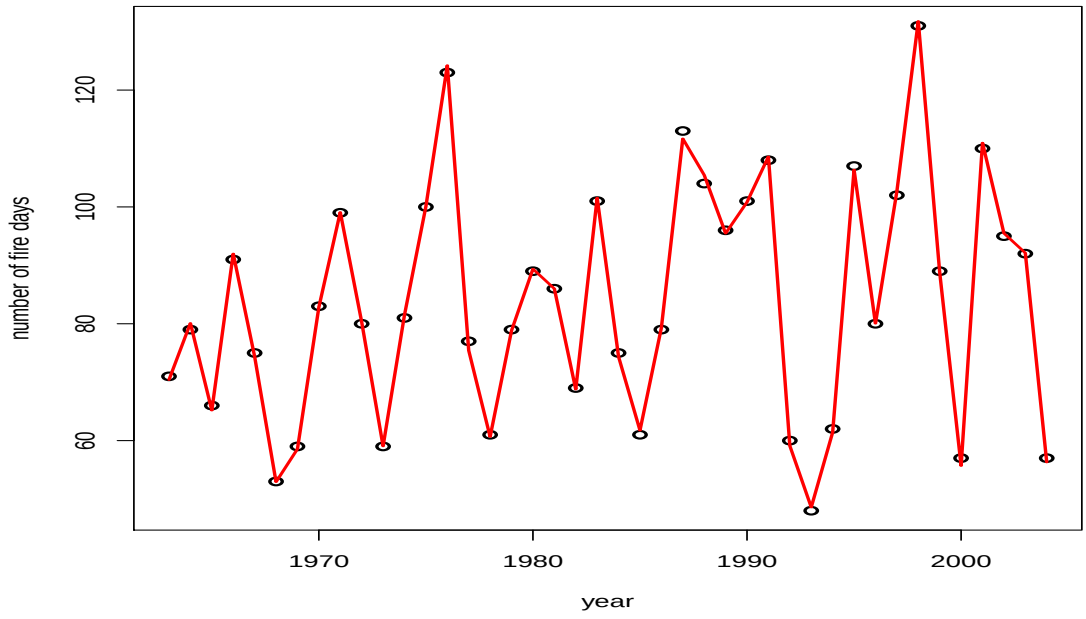
Figure 4.7: Estimates (black points) and confidence intervals (black lines) for the start (left column) and end (right column) of the fire season in Ontario. Overlayed are the trends (red solid line) and 95% confidence intervals for the trends (red dashed lines).

	Threshold	Slope	SE	WT	1BPT	2PT
Start	1	0.119	0.257	0.676	0.766	0.760
	5	0.110	0.192	0.716	0.764	0.782
	10	0.098	0.174	0.712	0.726	0.76
End	1	0.199	0.300	0.255	0.172	0.212
	5	0.262	0.231	0.255	0.078	0.104
	10	0.289	0.210	0.089	0.050	0.06

Table 4.2: Summary of trends, including the estimated slope and its standard error (SE), along with p-values from the Wald (WT), stage 1 block permutation (1BPT) and stage 2 permutation (2PT) significance tests for trends in the start and end of the Ontario fire season.



(a) Number of fire days aggregated by day of year.



(b) Number of days fires aggregated by year.

Figure 4.8: Observed (black points) and expected (red solid lines) number of fire days aggregated by day of year, and year from the GAM when applied to the Ontario fire data.

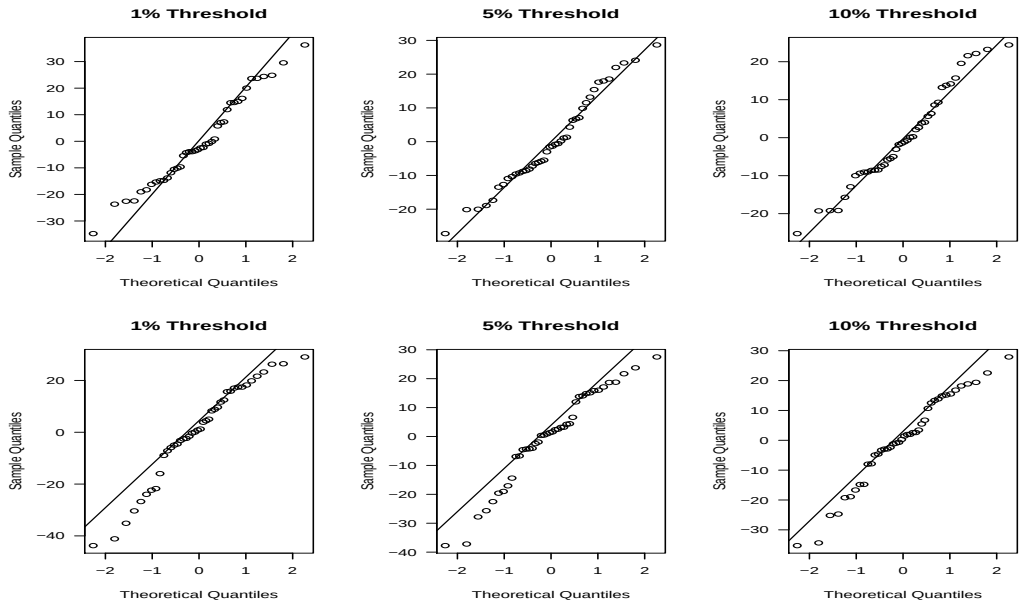


Figure 4.9: Residual QQ plots from the linear models fit to the nonparametric estimates of the start (top row) and end (bottom row) of the fire season in Ontario.

4.3 A Comparison of Trends from the Nonparametric and Empirical Approaches

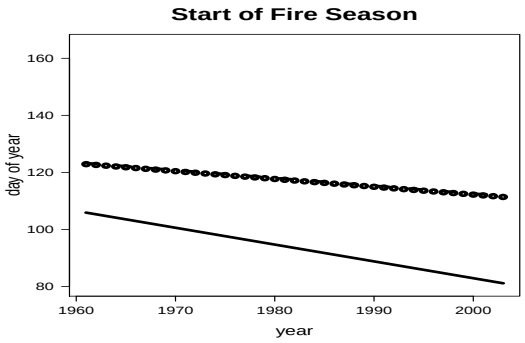
This section is devoted to contrasting the results from our three approaches for estimating trends in the start and end of the forest fire season. Specifically, we focus on the magnitudes of the trends. Recall that the three approaches to estimating fire season length were based on the estimated fire day probabilities from the logistic GAM, the counts of fire days and the ECDF of fire days within each year. In this section, we refer to the three fire season estimation methods as the GAM approach, the count-based approach and the ECDF approach, respectively. Note that for all figures referenced in this section, "threshold" refers to the percentage used to define the fire season length for the GAM and ECDF methods, while for fire day counts refers to the fire day number. Also, in this discussion, we refer to the last day the ECDF of fire days fails to exceed 99%, 95% and 90% as thresholds of one, five and ten, respectively.

4.3.1 Analysis of Alberta Forest Fire Data

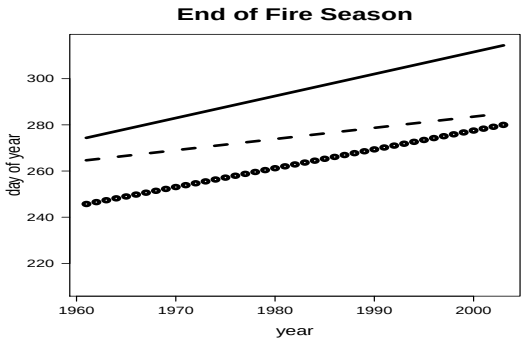
Figure 4.10 displays the trends from all three methods, summarized by threshold. Panel 4.10(a) displays the fitted values from the linear models based on all three approaches at the first threshold. Trends based on the GAM are relatively strong in comparison to those of the two empirical approaches. For thresholds of five and ten, summarized in 4.10(c) and 4.10(e), the magnitudes of these trends from the GAM and count-based methods are quite similar, with weaker trends present in the ECDF of fire days. Panels 4.10(b), 4.10(d) and 4.10(f) compare fitted values from the models for the end of the fire season. From these figures, the trends based on the GAM estimates and counts-based estimates of the end of the fire season are approximately parallel, once again indicating that they are of roughly the same strength. Meanwhile, those of the ECDF of fire days are relatively weak. Note that trends based on counts of fire days and the ECDF of fire days are conservative estimates of fire season length. That is, the start and end of the fire season based on the GAM is estimated to be earlier and later during the year than the corresponding estimates from the two empirical approaches.

4.3.2 Analysis of Ontario Forest Fire Data

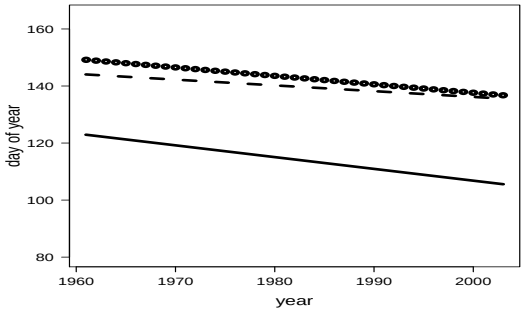
Similar patterns can be seen in Figure 4.11 for the Ontario forest fire data set. Panels 4.11(a), 4.11(c) and 4.11(e) display the estimated trends for the start of the fire season. The magnitudes of the trends are similar in all panels. For thresholds of five and ten at the end of the fire season, the strength of the estimated trends are quite similar. This is indicated by the three lines being approximately parallel. However, in panel 4.11(b), trends based on fire day risk from the GAM and counts of fire days are much stronger than that of the ECDF of fire days. Again, the start and end of the fire season from our nonparametric approach is estimated to be earlier and later in the year, respectively, when compared to the two empirical approaches.



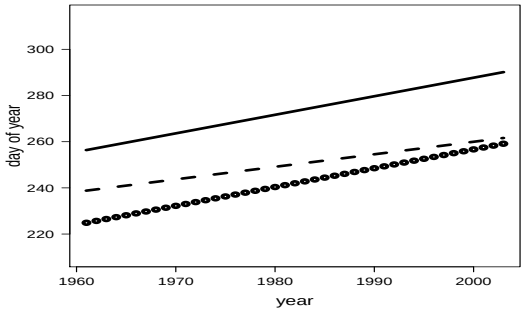
(a) Threshold 1.



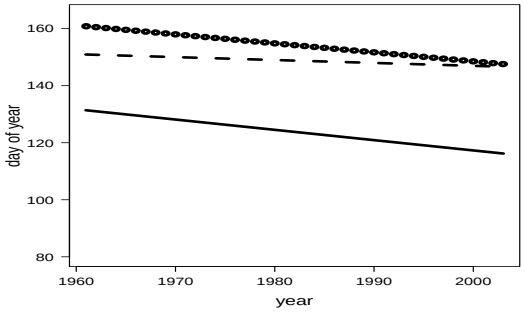
(b) Threshold 1.



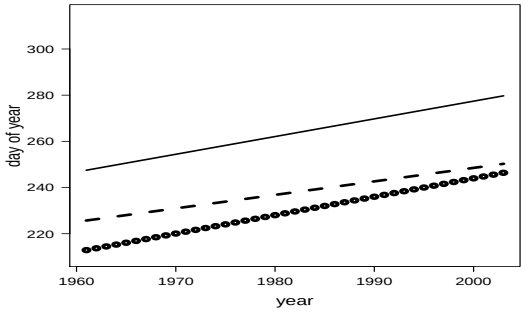
(c) Threshold 5.



(d) Threshold 5.

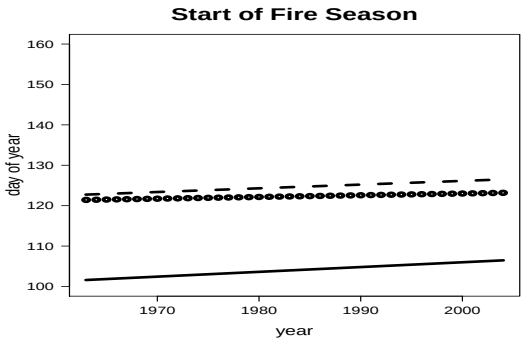


(e) Threshold 10.

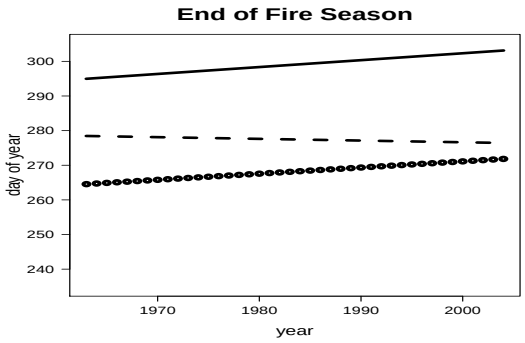


(f) Threshold 10.

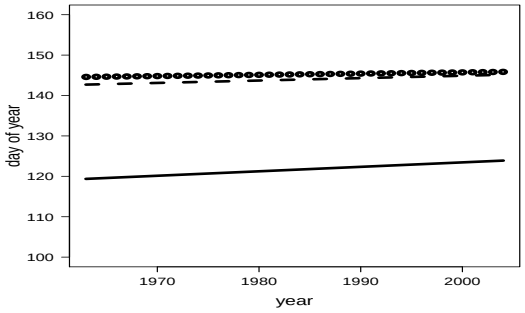
Figure 4.10: A comparison of trends in the start (left column) and end (right column) of the Alberta fire season when estimating its length from GAM estimates of fire day probabilities (solid lines), observed fire day counts (points) and the ECDF of fire days (dashed lines).



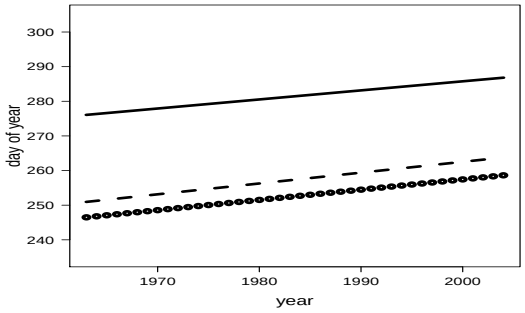
(a) Threshold 1.



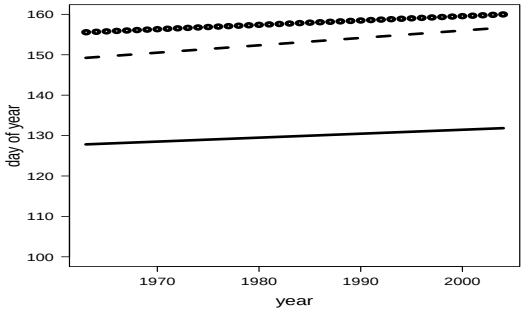
(b) Threshold 1.



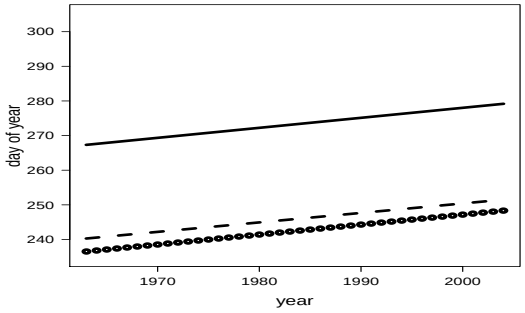
(c) Threshold 5.



(d) Threshold 5.



(e) Threshold 10.



(f) Threshold 10.

Figure 4.11: A comparison of trends in the start (left column) and end (right column) of the Ontario fire season when estimating its length from GAM estimates of fire day probabilities (solid lines), observed fire day counts (points) and the ECDF of fire days (dashed lines).

Chapter 5

Discussion

In this project, we developed a two stage approach to quantify and test for trends in the start and end of the fire season. A logistic GAM, using a thin plate regression spline basis, estimated the daily probability of a fire day within and across years. The fire season was defined as the crossing of a risk threshold within each year and inverse confidence intervals were calculated for the point estimates of the start and end of the fire season. Linear models were then fit to estimate trends, the significance of which were examined though the use of a Wald test as well as two bootstrap-based resampling tests. We then contrasted these trends to comparable empirical estimates of fire season length based on counts of and the ECDF of fire days.

Emphasis throughout this project was on applications of this approach to historical lightning-caused forest fire data from Alberta and Ontario, Canada. We noted significant trends in both the start and end of Alberta's fire season. The strongest trends were detected when using the output of the GAM to estimate the length of the fire season, while the weakest were those based on the ECDF of fire days. Meanwhile, no significant trends were found in the start of the Ontario fire season, but marginal significance was observed in the end of the fire season at the two highest thresholds. In contrast to Alberta, there were small differences in the magnitudes of the trends from the three estimates of fire season length. For both provinces, we noted stronger trends in the end of the fire season, in comparison to the start. We postulate that this result is due to a lingering effect of dry sub-litter fuels causing higher fire risk to last later into the fall.

Future work for this project includes incorporating spatial and temporal correlation into our GAM. Day-to-day temporal correlation is present in the residuals, but not yet

incorporated into the GAM. The addition of an AR(1) correlation structure to account for this was investigated, but resulted in oversmoothed estimates of fire day risk. This is likely due to the long runs of zeros between fire seasons leading to an inflated estimate of the autocorrelation parameter. Additionally, large smoothing parameters resulted in unusually low effective degrees of freedom. The presence of residual autocorrelation is known to result in underestimated standard errors. This occurs because the model constructed here from n observations assumes these are independent. However, when there is correlation present, the effective number of observations is fewer than the actual number of observations. Ideally, autocorrelation should be incorporated in the model to only account for correlation within the fire season. Future work could investigate the use of a time-varying autocorrelation component, which is currently not feasible using existing software. Spatial correlation should also be incorporated. In this project, we examine trends at the provincial level. However, both Alberta and Ontario can be partitioned into a set of sub-regions, where the area within each sub-region can be considered approximately homogenous with respect to fuel, weather and fire management strategy. Neighbouring regions may be correlated in terms of their forest fire activity, with decreasing correlation proportional to the distance between sub-regions.

Recall that this work is motivated by climate change. However, the trends discussed in this project have not yet been directly linked to changes in climate variables. Covariates, such as fire-weather indices should be considered. The inclusion of relevant teleconnections as additional covariates, such as El-Niño Southern Oscillation, could also be incorporated. Significant confounding factors, such as changes in detection efficiency, are also a concern.

A power study could be undertaken to allow for more objective comparisons between trends in the fire season when using the output of a GAM, empirical estimates of fire days and the ECDF of fire days to define the length of the fire season each year. This could be conducted via simulation and could take into account factors such as the number of years of fire records available for study and the strength of the trend.

Finally, constructing a model that incorporates counts of fires, rather than simply absence versus presence of a fire day, will likely result in more powerful tests for trends. Additionally, it has the potential to reveal further information. For example, rather than just investigating trends in the length of the fire season, we may be able to simultaneously detect trends in the number of fires. Hence, investigating the use of a Poisson-based GAM is of interest. Lightning caused fires occur in clusters, with the potential to ignite as many

as fifty to one hundred fires in a day (Canadian Forest Service, 2011). Woolford and Braun (2007) identified spatio-temporal centers of lightning activity and explored its relation to fire ignitions. Therefore, there is a need to account for the clustering that arrises in the data once counts are incorporated. As well, the data are zero-heavy and a mixture model framework would likely be necessary to handle this overdispersion.

Bibliography

Bonsal, B. R., Zhang, X., Vincent, L. A., and Hogg, W. D. (2001). Characteristics of Daily and Extreme Temperatures over Canada. *Journal of Climate*, 14:1959–1976.

Canadian Forest Service (2011). A lightning fire prediction system. *GLFC e-Bulletin Issue 13* .

Casella, G. C. and Berger, R. L. (2002). *Statistical Inference*. Duxbury, Pacific Grove, California, second edition.

Davison, A. C. and Hinkley, D. V. (2006). *Bootstrap Methods and their Applications*. Cambridge, Cambridge, eighth edition.

Government of Alberta: Office of Statistics and Information (2011). *Number of Hectares Burned from Wildfire in Alberta During the Fire Season, 2006 to 2010*. www.osi.alberta.ca.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall/CRC, Boca Raton, Florida.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall/CRC, Boca Raton, Florida, second edition.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2006). *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics, New York, New York, fourth edition.

Natural Resources Canada (2008). *Forest fire facts and questions: General facts about forest fires in Canada*. www.fire.cfs.nrcan.gc.ca.

Ontario Ministry of Natural Resources (2004). *Ontario's Rules and Laws Controlling the Use of Fire*. www.mnr.gov.on.ca.

R Development Core Team (2008). *R: A language and enviroment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL www.R-project.org.

Weber, M. G. and Stocks, B. J. (1998). Forest fires in the boreal forests of Canada. In J.M. Moreno (ed). *Large Forest Fires*, pages 215–233.

Williamson, T. B., Colombo, S. J., Duinker, P. N., Gray, P. A., Hennessey, R. J., Houle, D., Johnston, M. H., Ogden, A. E., and Spittlehouse, D. L. (2009). Climate Change and Canada's Forests: From Impact to Adaption.

Wood, S. N. (2003). Thin plate regresion splines. *Journal of the Royal Statistical Society Series B*, 65:95–114.

Wood, S. N. (2006). *Generalized Additive Models: An introduction with R*. Chapman and Hall/CRC, Boca Raton, Florida.

Wood, S. N. and Augustin, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 157(2–3):157–177.

Woolford, D. G. and Braun, W. J. (2007). Convergent data sharpening for the identification and tracking of spatial temporal centers of lightning activity. *Environmetrics*, 18:461–479.

Woolford, D. G., Cao, J., Dean, C. B., and Martell, D. L. (2010). Characterizing temporal changes in forest fire ignitions: looking for climate change signals in a region of the Canadian boreal forest. *Environmetrics*, 21:789–800.

Wotton, B. M. and Flannigan, M. D. (1993). Length of the fire season in a changing climate. *The Forestry Chronicle*, 69(2):197–192.