Fire Research Report

The application of data mining and statistical techniques to identify patterns and changes in fire events

University of Auckland

May 2009

This study explores the extent to which data mining and statistical techniques might assist the Fire Service in detecting threshold and pattern changes in its spatio-temporal fire data. Three entirely different scenarios are investigated. A post-hoc search for patterns was made of fires of suspicious or unknown cause in an area where a subsequently convicted arsonist was known to be operating. The spatio-temporal occurrence of chimney fires was compared with local climate data looking for any threshold conditions which might trigger the seasonal changes in occurrence. Finally an attempt is made to measure the effectiveness of the Firewise programme, which involves fire fighters visiting schools to instruct students in fire safety. The before and after incidence of residential fires in proximity to schools visited is assessed to determine whether the programme has had any measurable effect. Different data mining techniques are applied to each scenario.

The literature on change-point detection is reviewed and the applicability of identified techniques to real time fire data is discussed. Software options are discussed. The results suggest that fire data, especially time and location data, would be adequate for the purposes of detecting change-points but the problem under investigation must be clearly defined. Interpreted data about the fire must be accurate and unambiguous if it is to be of assistance in identifying change-points.

New Zealand Fire Service Commission Research Report Number 95
ISBN Number 978-1-877349-98-0 (on-line)
ISBN Number 978-1-877349-97-3 (paperback)

The application of data mining tools and statistical techniques to identify patterns and changes in fire events

Mark Holmes, Yong Wang, Ilze Ziedins
The University of Auckland

Final Report: May 2009

1 Introduction

The ability to identify patterns and changes in fire events as they occur is highly desirable, particularly if they are unexpected. The frequency of different kinds of fire events may change with the season, day of the week, and time of day. Guy Fawkes day always sees the greatest number of fire events of any day of the year. These are changes that can be modelled and are therefore predictable to some extent. Others, however, are less predictable, such as an arsonist laying a series of fires. It may nevertheless be important to detect such changes, in order to prevent further occurrences. It may also be desirable to know whether an intervention, such as a fire safety campaign, has led to a decline in fire events of particular types. The aim of this study is to explore the extent to which data mining tools and statistical techniques might assist the Fire Service in detecting such changes, and to illustrate the kinds of results that can be obtained from the data currently available.

This study explores three different scenarios in detail. Each has different questions of interest, and requires different analyses. The three scenarios of interest are:

[1] Chimney fires – fluctuations in the rate of fires with changes in temperature and weather.

[2] Changes in fire rates in proximity to schools where the New Zealand Fire Service has run its Firewise programme.

[3] Detecting increases in the frequency of fires due to fire laying by individuals (arson) – the particular case considered here is that of a known arsonist in Blenheim.

Data from 2004–2007 was used for the analyses.

In the period 2004–2007, 2752 chimney fires were recorded in New Zealand. Of these, 874 were in the major cities, with Dunedin and Invercargill having the highest recorded numbers (294 and 196 respectively). Chimney fires are clearly seasonal, and the question of interest here was how the frequency changes with changes in temperature and weather, and, more particularly, whether there is a threshold temperature below which the frequency of chimney fires increases markedly. There was also interest in investigating whether the *onset* of colder weather leads to a higher frequency of chimney fires, or whether a higher frequency at lower temperatures is sustained throughout the winter.

The second scenario that was considered was the Firewise programme. The Fire Service delivers fire safety and prevention programmes to early childhood education centres and primary schools throughout New Zealand. A total of 5864 Firewise programmes were delivered between June 2004 and July 2008. Here the interest is in whether it is possible to detect any changes in the frequency of fires after the delivery of these programmes.

The third and final scenario that was considered was of a known case of arson in Blenheim over the period of the study. The question of interest here was whether the fires attributed to this arsonist could be identified,

and whether they could then be identified in real time. That is, how soon after the onset of the arson events could the pattern be detected?

Since all of the above three scenarios concern the frequencies of events, using a Poisson-distributed response variable appears to be most suitable. Four different types of models are used below: generalised linear models (Dobson(1990), McCullagh and Nelder (1989), Lindsey (1997)), generalized linear mixed-effects models (Pinheiro and Bates (2000)), generalised additive models (Hastie and Tibshirani (1999)), and decision tree models (Breiman et al. (1984), Quinlan (1993), Witten and Frank (2005)). These models are not applied to all three scenarios nor are they described in the order given above. We have conducted the investigation in a problem-oriented manner and thus the results are presented below in the order of increasing relevancy of the models to the problem studied. These models are briefly described in Section 4, with further details given later with the results of their application to the fire data. We have not included all intermediate models that have been fitted during our search for the best fit.

Although each of the three scenarios is concerned with possible changes in the frequency of fires, each of the scenarios is different in nature, has different kinds of questions associated with it and responds best to different statistical or data mining techniques. One of the major conclusions of this study is therefore that while it may be possible to set up an automatic detection system, it would need to be tailored to particular scenarios of interest, and would need to be supplemented with statistical analyses of patterns that might be detected.

Section 2 gives a description of the data with some discussion. Section 3 gives a short literature review of related work. Section 4 gives some background for the models used in the analysis. Section 5 discusses the chimney data in detail, Section 6 the Firewise programme and Section 7 the suspicious fires data. We conclude with discussion and recommendations in Section 8.

We are most grateful to Neil Challands of the New Zealand Fire Service for assistance in providing the data and help with our understanding of the variables and the data collection process.

2 The data

The New Zealand Fire Service collects data for every fire event. This study used data from 2004 to 2007 inclusive. A large number of data fields is available, with the data fields recorded depending on the type of incident. We have used the following fields in this study:-

- CAD number
- Date and time of the incident
- Location:- gridpoint location, street, suburb and town
- Whether the incident was rural or urban
- Incident type
- Cause of fire (group name and more detailed description)
- Object ignited

The CAD number is a unique identifying number assigned to each event. The time of the incident was recorded to the nearest minute.

In addition to the Fire Service data, we also have weather data available, recorded at weather stations in the major cities. This was particularly useful for the analysis of the chimney data. The weather data was recorded at midday for each day in the period of the study, and included the fire weather index, temperature, relative humidity, wind speed and wind direction.

Data on the Firewise programme included the locations of all schools where the Firewise programme was delivered from 2004 onwards and the location of all early childhood centres where the Fire Prevention Programme was delivered over the same period. Other fire safety/prevention advice sessions were also included. The date of completion of each programme was given as well as the number of children to whom it was delivered.

3 Literature review

3.1 Data mining

As a new technology, data mining has proved to be valuable in numerous practical applications (Vapnik (1998), Hastie et al. (2001), Witten and Frank (2005), Felici and Vercellis (2007)). The majority of the problems studied in the data mining community can be categorized as regression, classification, clustering, or link analysis. The data sets involved are often very large in scale, with a mixture of both numerical and categorical variables. Popular models include decision trees, kernel methods, support vector machines, nearest-neighbours methods, and neural networks. However, the problem of change point detection does not belong in the above categories, thus making popular data mining methods difficult to apply directly. In fact, there are not many studies on change or change-point detection in the literature of data mining. Among the few studies, Guralnik and Srivastava (1999) investigated how to detect events from time series data; Zeira et al. (2004) used classification models for change detection; Takeuchi and Yamanishi (2006) considered simultaneous detection of both outliers and change points; Ide and Tsuda (2007) proposed a new algorithm for change-point detection based on principal component analysis using Krylov subspace. Furthermore, the book edited by Roddick and Hornsby (2001) contains several papers on change-point detection in different situations.

3.2 Statistical models and change-point detection

Change-point detection has a long history of investigation in statistics. Statistical models and hypothesis tests usually have a narrower focus on specific problems than do data mining techniques, see e.g. Shewhart (1931), Page (1954), Shiryaev (1963), Roberts (1966), Crowder (1987), Carlstein (1988), Basseville and Nikiforov (1993), Lai (1995), Bai (1997), Mason et al. (1997), Molnau et al. (2001), Jones (2002), Frisen (2003), Hawkins et al. (2003), Reynolds and Stoumbos (2004), Chiu et al. (2006) and Akakpo (2008).

In this Fire Service project, we have studied a number of identified scenarios of interest. Since many of these scenarios involve detecting changes in fire rates, generalized linear models, in particular Poisson regression, have proved useful (McCullagh and Nelder (1989), McCulloch (1997)) as has their extension to nonparametric generalized additive models (Hastie and Tibshirani (1990), Lin and Zhang (1999), Wood (2008)). Embedding mixture models in these models may also provide a better fit to the data, as discussed in Heckman and Singer (1984), Aitkin (1996), Tsodikov (2003) and Zeng and Lin (2007).

There are several other established techniques for detecting change points in data, not of all of which are appropriate for this data set. Since the time of callout for a fire event is recorded accurately, we could have considered models in continuous time. However, less complex models based just on a count of the number of fire events in a day were sufficient here. Simpler count models also obviated the need to model diurnal variation. The review below considers methods for both discrete and continuous data. We consider both the detection of change points in historic data, and then, more briefly, the detection of change points in real time.

The problem of detecting change points in historic data has long been a question of interest. The simplest problem is to identify a single change point from one known parameter value to another. There is a considerable step in complexity from identifying the location (whether in space or time) of a single change in parameter values, to the problem of identifying both the number of change points and their location. The possibility that the parameters (such as the rates of a process) may themselves be random variables adds yet further complexity.

Most processes in time can exhibit change points of one kind or another, so there are many and various underlying models where the change point detection problem is of interest. We illustrate this by citing some recent works. Akakpo (2008) considers the problem of detecting multiple change points in sequences of independent categorical data (such as, for example, DNA sequences). Polansky (2007) considers the problem of detecting change-points in Markov chains (which can be thought of as sequences of dependent random variables) and Ge and Smyth (2000) use a (semi)-Markov hidden model, fitted using the EM algorithm, to detect change points in semiconductor manufacturing. The bent-cable regression models in Chiu et al. (2006) join linear segments (where rates are constant) with quadratic bends, which give a smoother transition than piecewise linear models. They also permit the abruptness of the transition to be assessed. Bai (1994), Zeira (2004) and Takeuchi and Yamanishi (2006) use classical time series models as the underlying model. Stochastic processes such as Poisson processes (Akman and Raftery (1986), Kennett and Pollak (1996)), Levy processes and Gaussian processes have also been the underlying models in change point detection problems.

A wide range of methods have also been used to detect change points (see also Basseville and Nikiforov (1993)). These include hypothesis testing (e.g. Akman and Raftery (1986) and Siegmund (1988)), maximum likelihood (e.g. Hinkley (1970), Worsley (1986) and Hawkins (2001)) and quasi-likelihood (Reed (1998)), Bayesian methods (e.g. Raftery and Akman (1996) and Carlin et al. (1992)), least squares (Bai (1994) and Bai (1997)), nonparametric methods (Carlstein (1998)) and cumulative sum approaches (see e.g. Galeano (2007)).

The papers cited above all apply the methods to historical sets of data. However, it may also be of interest to detect changes in real time. This has long been important for industrial process control and quality control (see e.g. Roberts (1966), Gan (1994), Lai (1995), Kenett and Pollak (1996), Reynolds and Stoumbos (2004)), but has also arisen in other contexts, such as real time detection of attacks on computer networks, increases in disease incidence, changes in stock and share prices. This is a difficult problem. There is a tradeoff between the delay in detecting the change and generating too many false alarms. The problem of detecting the change can also be viewed as an optimal stopping problem (Shiryaev (1963)). Gal'chuk and Rozovskii (1971) and Davis (1976) partially solved this problem for the Poisson process, when the aim is to detect a change from one known rate to another known rate, and it was then fully solved by Peskir and Shiryaev (2002). Bayraktar and Dayanik (2006) solved the problem when there is an exponential penalty for non-detection. In applications the new rate is not usually a known constant, however, but a random variable, and may need to be estimated from the data. This problem has only recently been solved by Beibel (1997) and Beibel and Lerche (2003) for Brownian motion and Bayraktar et al. (2006) for the Poisson process. More recently, Dayanik and Goulding (2007) have addressed this problem using Bayesian methods in a far more general context with an underlying hidden Markov model.

Change point methods are used in many other applications than modelling of fire events, some of which have already been mentioned. Three areas of application have seen considerable research development recently. The first, and possibly most relevant, is the problem of biosurveillance and the need for early detection of outbreaks of disease associated with bioterrorism, so that preventive measures can be undertaken. Sebastiani and Mandl (2004) give a good introductory discussion of this area with an extensive bibliography. The problem is to detect unusual clusters of disease, preferably in real time, rather than post the event. The combined spatial and temporal elements here are similar to those found in the problem of detecting a fire starter. A second area of application is in the detection of change points in Internet data, and particularly detection in real time of attacks on service. However, the spatial aspect is not always considered important in Internet data, and the time scales of interest are orders of magnitude smaller than for the Fire Service data. The third area of application is in DNA sequencing, where the change points occur in space (the DNA sequence) rather than time. Again, compared with the Fire Service data, this does not possess both spatial and temporal dimensions. There are many other important areas of application which we do not discuss further here – detecting change points in financial data is a very obvious one.

We have not found much work on applying change point methods to fire data, and it is mostly found in the context of modelling forest fires. Of particular interest to researchers in North America has been the modelling of the time since the last fire. Reed (1998) uses quasi-likelihood methods and backwards selection

to identify change points. A more recent paper (Reed (2000)) uses the Bayes Information Criterion to select an appropriate model. However, the question addressed here, and the data available are very different from the New Zealand Fire Service data. The aim is to estimate the date and size of fires (whereas this data is already known to us).

More generally, there is a very large body of literature on historical fire frequency in the context of forest fire management. Estimation of wildfire risk is important for forest fire management and assessing insurance risk. Brillinger (2003) and Preisler et al. (2004) construct probabilistic models to estimate wildfire risk. Peng et al. (2005) use a spatial point process model to evaluate the effectiveness of the burning index in predicting wildfire occurrence. Ryden and Rychlik (2006), on the other hand, use point process models to model the occurrence of urban fires.

4 Models

In this section we describe the models used in the following sections.

4.1 Poisson distribution

The Poisson distribution with rate λ has probability function

$$f(y;\lambda) = e^{-\lambda}\frac{\lambda^y}{y!},$$

for $y = 0, 1, \ldots$ and $\lambda > 0$. It is commonly used for modelling the number of occurrences of a rare event in, say, a unit of time.

4.2 Poisson regression

The Poisson regression model is known as the log-linear model, where the logarithm of the rate of the Poisson response variable is assumed to have a linear relationship with the explanatory variables, that is,

$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

where the response y for count data has the Poisson distribution with rate λ.

4.3 Generalised additive models

Generalised additive models are an extension of generalised linear models, in the sense that a term in the latter is replaced by a function of variables. A very basic extension is to replace an explanatory variable with a spline function of the variable, and hence sophisticated nonlinearity can be used to describe the relationship between the response and an explanatory variable. For Poisson regression, the model looks something like

$$\log(\lambda) = \beta_0 + s_1(x_1) + \cdots + s_p(x_p),$$

where $s_1, \ldots, s_p$ are spline functions that need to be fitted to the data.

The complexity of a spline function is controlled by the parameter of the degrees of freedom. A spline function with one degree of freedom is simply the variable itself and its associated coefficient.

4.4 Mixed-effects Poisson regression

A mixed-effects model treats some coefficients as random variables, which thus have distributions, instead of being held fixed as in a conventional statistical model. This is usually because observations may belong to distinct groups and those in the same group are likely to share the same coefficient. Taking the Poisson regression model as an example, the log-rate still has a linear relationship with other predictor variables, but the intercept is conditional on the group membership of an observation:

$$\log(\lambda_j) = \beta_{0j} + \beta_1 x_1 + \cdots + \beta_p x_p,$$

where β_{0j} is the intercept for group j. All β_{0j}'s together follow a distribution, whose family is usually pre-chosen with a few unknown parameters.

A mixed-effects Poisson regression model can be fitted by using the R package `lme4`. In this package, a normal distribution is always used for modelling a random-effects variable, with its mean and variance determined from the data.

4.5 Decision trees

Decision trees are popular in the data mining community for providing solutions to difficult nonlinear regression and classification problems. They can be used to identify not just existing patterns, but also changes in patterns. They appear to be particularly useful for the Fire Service dataset.

The basic idea of decision trees is to group observations based on their neighborhood and response values. A decision tree recursively partitions the entire feature space into many (hyper-)rectangular regions, and possesses a tree-like structure. The average or majority of the response variable values in each region is used to make predictions for new observations that fall in the same region. Decision trees can be used for a range of problems categorized by the type of response variable – categorical (classification), continuous (regression), and count data (Poisson regression) – but the structure is similar for all of them.

Decision trees have several advantages over other data mining tools, as well as conventional statistical techniques. Their training is typically very fast and can be done repeatedly. They deal well with different types of covariates and nicely with missing values. Their results are easily comprehensible and often shed light on the problems studied. However, their accuracy of prediction is sometimes not as good as some other data mining tools, such as neural networks and support vector machines. They also have the disadvantage of rapid segmentation of a data set, that is, a rapid decrease in the number of observations in each region, and each pattern found is usually described by only a few most influential variables. It should also be noted that a decision tree produced by an algorithm is usually not optimal in the sense of a performance measure (e.g., log-likelihood, squared errors, etc.), but finding the "optimal tree", if one ever exists, is computationally intractable (or NP-hard, technically speaking).

5 Chimney fires

5.1 Problem and data

In this scenario, we study chimney fires, in particular the extent to which the frequency of chimney fires fluctuates with changes in temperature and weather conditions, and whether there is a temperature threshold below which the frequency increases markedly.

A total of 2752 chimney fires occurred between 1 Jan 2004 and 31 Dec 2007 in New Zealand, as plotted in Figure 1. Six cities are included in the study: Auckland (Auckland, Manukau, North Shore, Waitakere), Hamilton, Wellington (Wellington, Lower Hutt), Christchurch, Dunedin and Invercargill. They have, respectively, 106, 11, 109, 158, 294 and 196 chimney fires over the four year period. In addition to the data collected

by the New Zealand Fire Service, weather data collected at the nearest weather station to each city was made available to us.

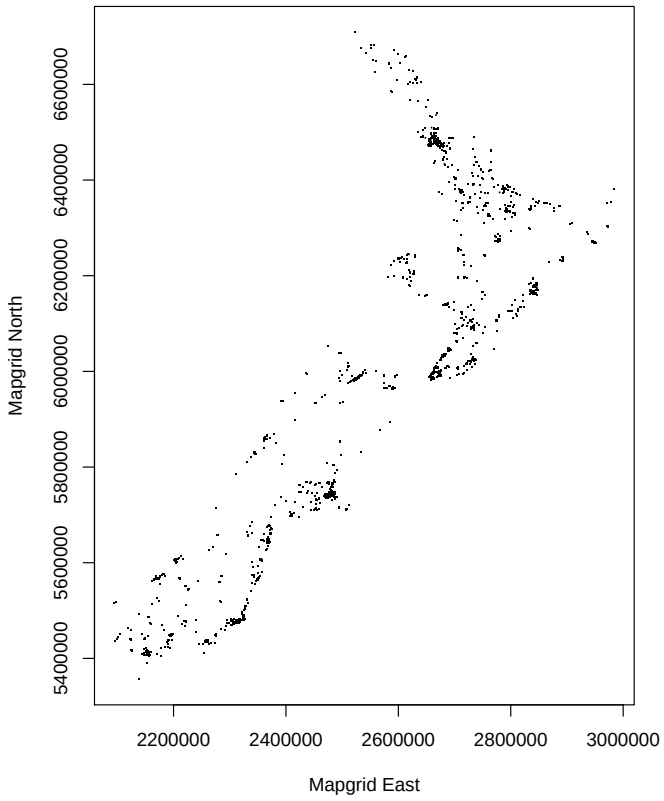

Figure 1: Location of all chimney fires between 2004–2007 in New Zealand.

The daily frequencies of all the chimney fires occurring in these six cities are plotted in Figure 2, from day 1 (1 Jan 2004) to day 1461 (31 Dec 2007) and the strong dependence on temperature is obvious: there are more chimney fires during the winters than during the summers. The obvious reason is that more fires are lit in fireplaces on cold days, which increases the size of the fire population that may potentially lead to chimney fires. Figure 3 plots the weekly frequencies of chimney fires for the three main South Island cities. A more detailed plot of the seasonal variation is given in Figure 4 which plots the frequencies per week, summed over the four years, for each of the South Island cities. Figure 5 shows the number of chimney fires by time of day for each of the South Island cities. They occur most commonly in the evenings, with a slight increase around midmorning. Figure 6 shows the rate of chimney fires per day vs. midday temperature for each of the three cities considered, with Dunedin having the highest rate, and Christchurch the lowest. At least in part this reflects the fireplace fire ban in Christchurch and the likely higher number of chimneys in Dunedin than Invercargill. We have attempted to accommodate this second effect in Figure 7 where we have adjusted the rates according to population. After this adjustment it seems that Invercargill has a slightly higher chimney fire rate. The fitted curves in these plots are so-called `lowess` smoothers, which use locally-weighted polynomial regression. They have the important feature of not being overly affected by outliers, but at the same time they can underestimate, or be slow to respond to important trends in the data. The slope of the Invercargill fitted curve appears to be largest at a temperature of around 11 degrees Celsius. However, in these plots there is no clear "critical temperature" at which there is a significant jump in the chimney fire rate.

Since there is no direct information or indeed any information in the data about the total number of fires

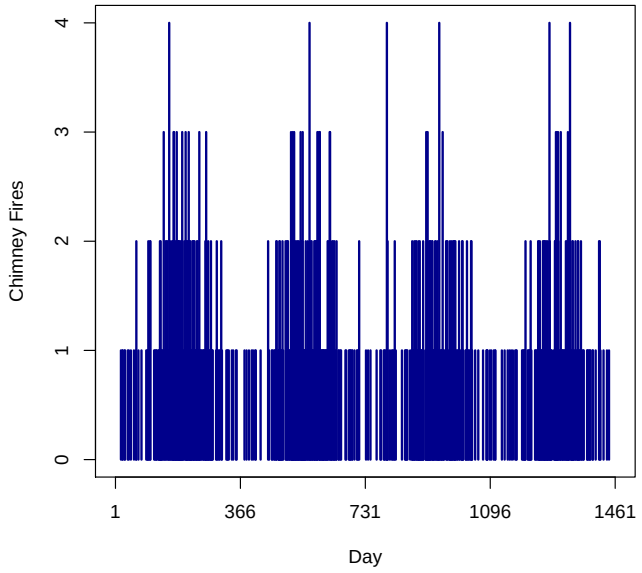

Figure 2: Daily frequencies of chimney fires in the six cities between 2004–2007

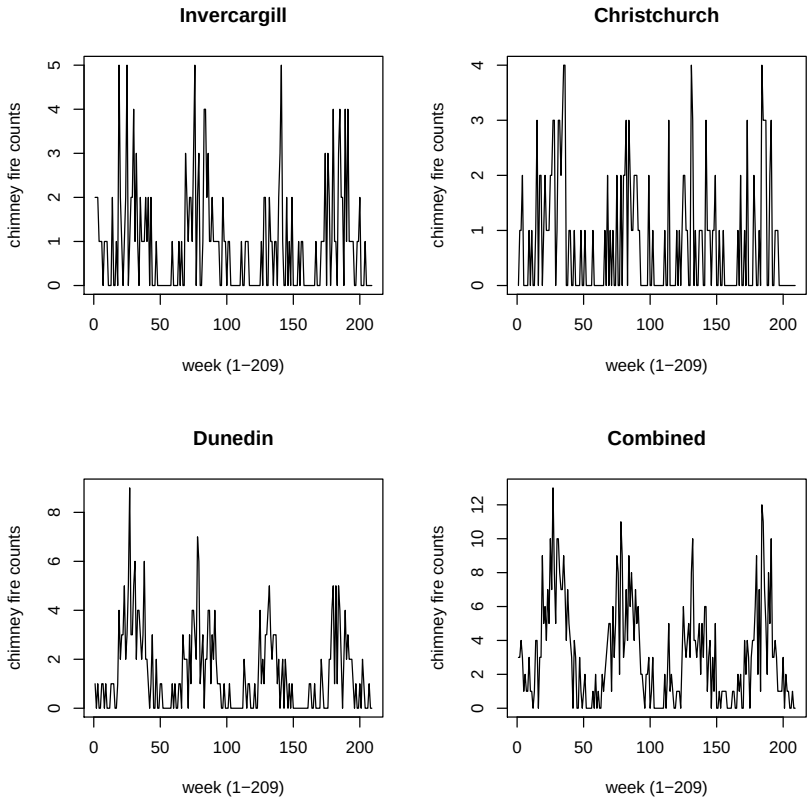

Figure 3: Frequencies of chimney fires in the South Island cities between 2004–2007.

made in fireplaces in a day, it is very difficult to answer questions regarding the change of the proportion of chimney fires in the population of fireplace fires, namely the risk of a chimney fire in a household. Temperature

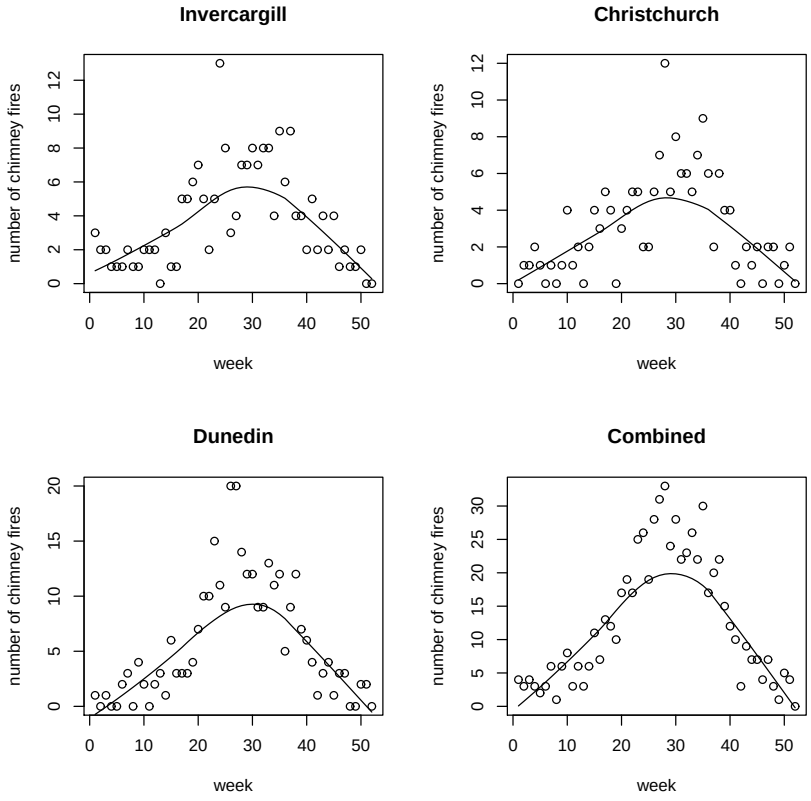

Figure 4: Number of chimney fires per week, summed over 2004-2007.

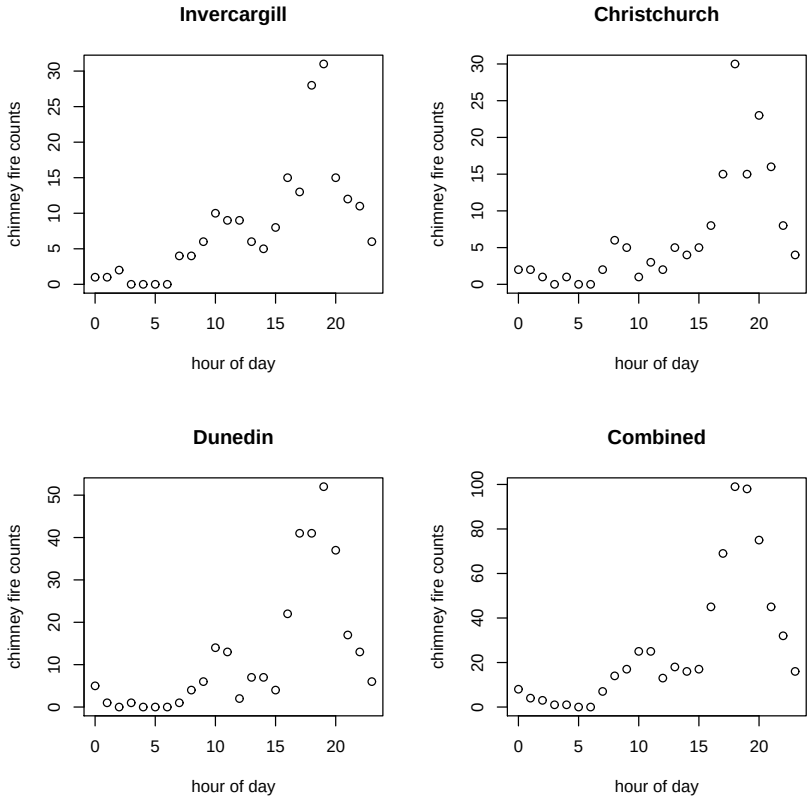

Figure 5: Number of chimney fires vs. hour of the day.

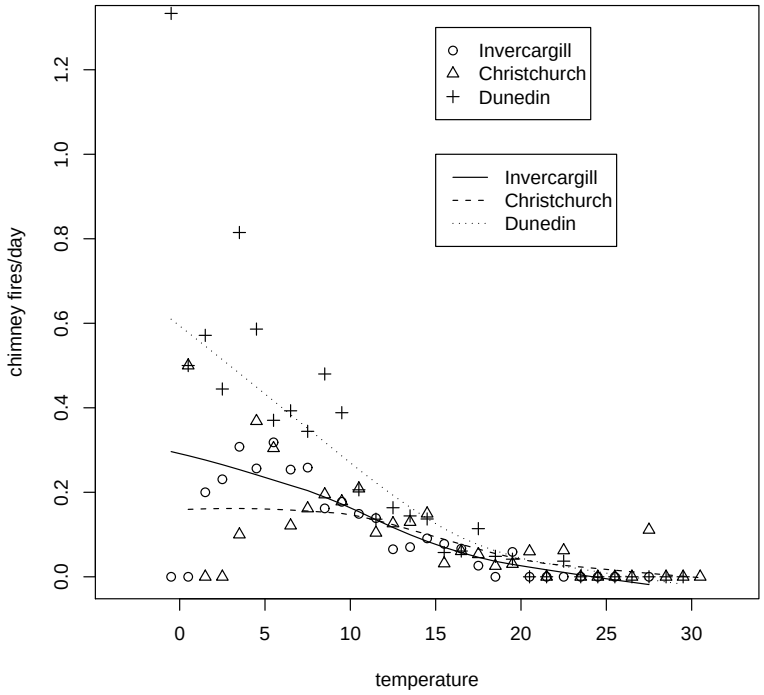

Figure 6: Chimney fires per day vs. temperature.

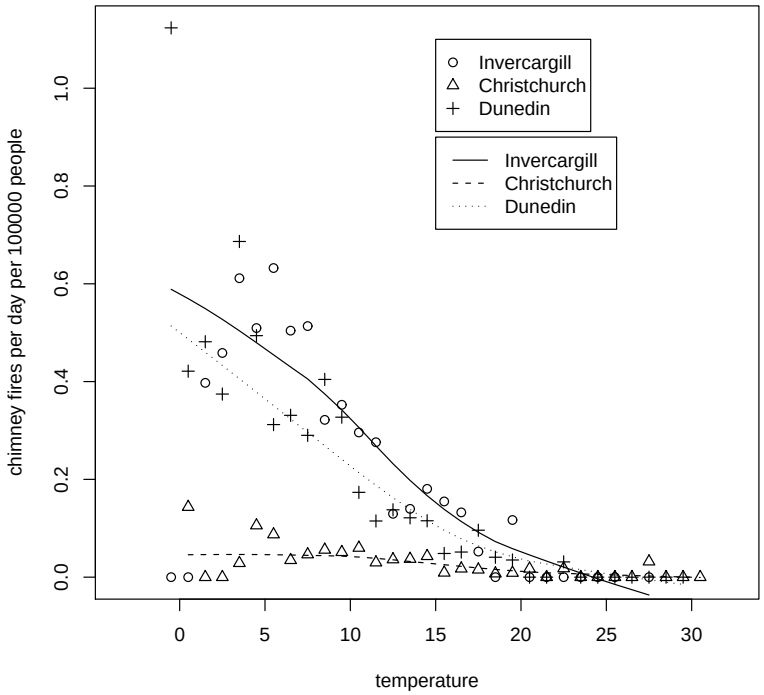

Figure 7: Chimney fires per day per 100,000 population vs. temperature.

is closely correlated with both, and it appears impossible to distinguish them based purely on the information that the data provides. We note also that it may be more appropriate to use the evening temperature, rather than

the midday temperature which was used here.

The data set we created contains the following variables for storing the potentially relevant information:

`city`	one of `Akl`, `Ham`, `Wel`, `Chr`, `Dun` and `Inv`
`pop`	population of the city
`day`	day of the four years (ranging from 1 to 1461)
`month`	month (ranging from 1 to 12)
`year`	year (ranging from 1 to 4)
`weekend`	indicator whether the day is a weekend day
`fwi`	calculated fire weather index (FWI) at midday for each day
`humid`	relative humidity at midday for each day
`T`	temperature at midday for each day
`D1`	$=$ `T` $-$ `T1`, where `Ti` is the midday temperature on `day i`
`D2`	$=$ `T1` $-$ `T2`
`count`	number of chimney fires on a given day in a given city

The inclusion of these variables (apart from `day`) is an attempt to account for the daily frequency of chimney fires as much as possible. In particular, by including `D1` and `D2`, we hope to detect whether the change of temperature contributes to the occurrence of chimney fires, while `month`, defined as having 30.4375 days per month, can help estimate seasonal effects. The variable `pop` stores the population of each city, downloaded from `http://www.citypopulation.de/NewZealand-UA.html` and rounded to the nearest thousand. These population data were collected in the census conducted on 7 March 2006 by Statistics New Zealand. The inclusion of `pop` is to provide an offset term in a Poisson model so that the fire rate can be defined relative to a unit of the population (a thousand here) and hence different-sized cities can be treated equally. A small random subset of the data set is given in Appendix A.1.

We have applied decision tree, generalised linear, and generalised additive models to this data set.

5.2 Poisson decision trees

A plot of the Poisson decision tree built from the data is shown in Figure 8; see also Appendix A.2 for a printed version. A decision tree has two different types of nodes: internal and terminal nodes. At each internal node, there is a splitting criterion, where an observation is sent down the left branch if the criterion is satisfied, or down the right branch if otherwise. The prediction for a new observation is made at the terminal node it eventually reaches. For the Poisson decision tree shown in Figure 8, there are 8 terminal nodes. The three figures at each terminal node obtained from the training observations reaching the node are, respectively, the estimated daily rate of chimney fires, the number of chimney fires, and the number of observations (or days here). The prediction for a new observation is made on the basis of these figures.

Decision tree models can outperform others in the situation when relationships are irregular (e.g., non-monotonic), but they are also known to be inefficient when relationships can be well approximated by simpler models, e.g., Poisson regression models. The chimney fire scenario appears to be a reasonably regular situation, so we expect the Poisson linear regression model to provide sufficiently accurate estimates. On the other hand, the decision tree model does provide a certain level of information about the most relevant variables, such as `T`, `city` and `month`. It also provides a different, and perhaps more intuitive, way of interpreting the relationship between the daily rate of chimney fires and the explanatory variables.

The distinction of Hamilton from other cities is not really a problem of the decision tree constructor or the data quality. The reason is that the rate of chimney fires defined here is for a city, not per capita or per fireplace fire. Hamilton is similar to Auckland in terms of climatic conditions, but has a much smaller population. The other cities also have smaller populations than Auckland, but they have colder weather and thus larger fireplace fire populations. This distinction can also be seen in the other two models, if the population is not taken into account. The function `rpart` that we use for building a Poisson regression tree does not appear to deal well with an offset term included to account for the effect of the population.

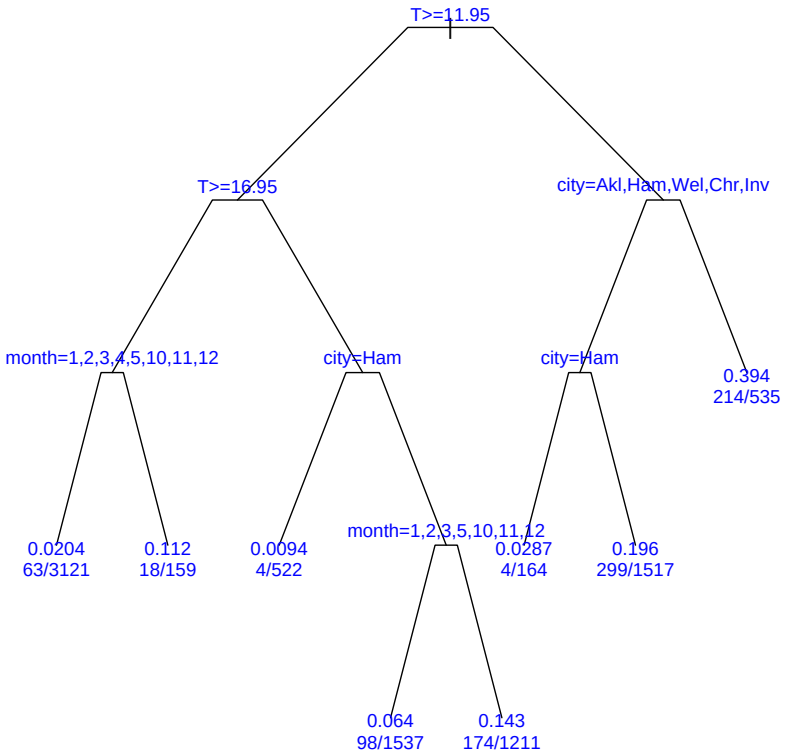

Figure 8: Poisson decision tree for chimney fires in six cities

5.3 Poisson regression

The fitted Poisson regression model that includes all variables (except `day`) is given in Appendix A.3. The variable `pop` is used as an offset term, so the rate of chimney fires is defined per thousand people in the population. Of these variables, `fwi` and `humid` are not significant; nor are `D1` and `D2`. The highly significant `month` suggests that there exist seasonal effects, most likely due to more fires lit in fireplaces during the winters. Both `year` and `weekend` are only weakly significant. Unsurprisingly, `T` is the most significant variable. Also highly significant is `city`, where, using the z-values, the six cities can be grouped into three pairs due to their similarity: Auckland and Hamilton; Wellington and Christchurch; Dunedin and Invercargill. This means that, under the same weather conditions that have been explained by the model, the rates of chimney fires per thousand population are still different between these cities. This appears to suggest that people living in colder areas tend to light more fires in fireplaces under the same weather conditions.

In order to produce a model with better fit, the Akaike information criterion (AIC) (Akaike (1974)) is used for variable selection; see Appendix A.3. The final selected model has 5 variables and an AIC value of 5053, reduced from 5058 of the above model. The AIC is known to be conservative, in the sense that it tends to produce a model with a greater number of parameters than necessary. However, no model selection criterion is entirely satisfactory. Here we use the AIC as a general guide, which should be fine, especially when a reduction in its value is large (say, ≥ 5).

5.4 Generalised additive models

It is possible that the relationship between the log-rate of chimney fires and other variables is not linear. To find such a nonlinear relationship, generalised additive models are fitted to the data; see Appendix A.4. By using spline functions with 3 degrees of freedom, `humid` is only very weakly significant to have a nonlinear

relationship but it is extremely significant that T has a nonlinear effect. After taking these into account, the value of AIC is further reduced to 5047 from 5053 of the Poisson linear regression model.

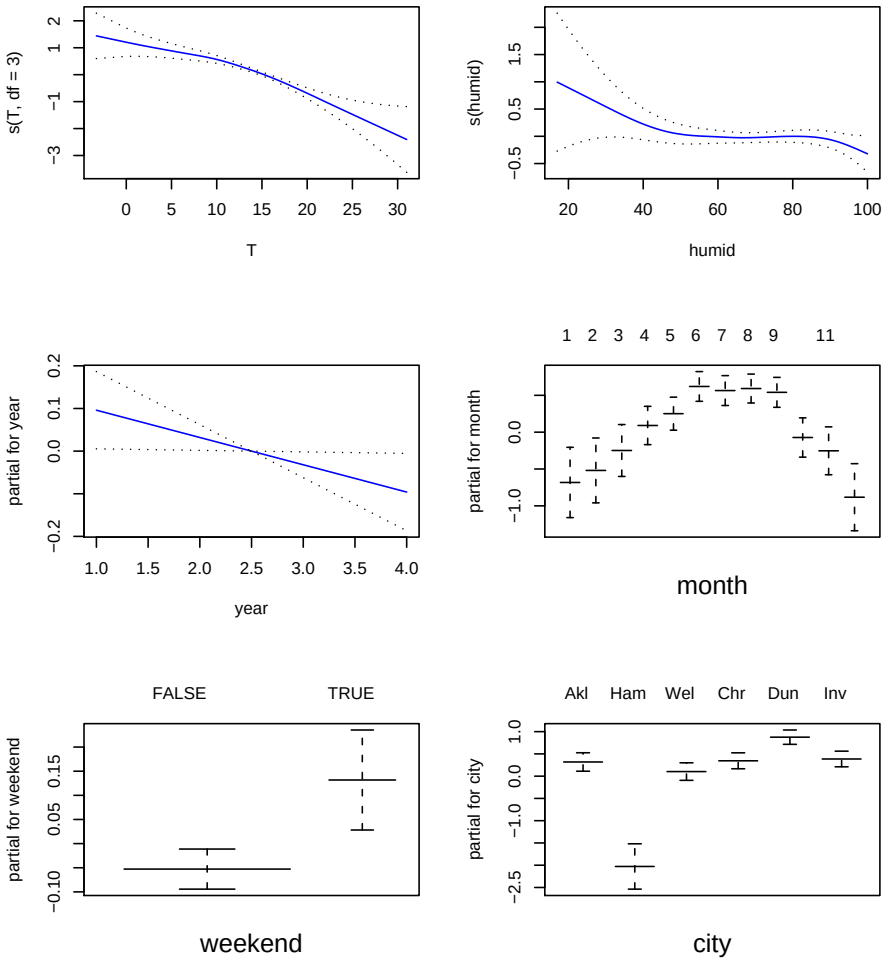

Figure 9: Partial effects of the variables used in the generalised additive model on the log-rate of chimney fires.

The partial effects of the explanatory variables and their twice-standard-error bands are plotted in Figure 9. From the plot for `month`, seasonal effects on the rate of chimney fires are quite obvious—there are simply more chimney fires during the winters, which is not surprising. In addition, at a temperature of about 13 degrees Celsius there is clearly a change in the trend of the partial effect of temperature on the log-rate of chimney fires.

5.5 Summary and remarks

The main conclusions that we have drawn from the analysis presented above are as follows. The log-rate of chimney fires per day per city is related to `T`, `city`, `month`, `weekend`, `year` and `humid`. It has a strong nonlinear relationship with `T` and possibly a weak nonlinear relationship with `humid`. In general, the rate of chimney fires increases as `T` decreases and there is clearly a seasonal effect. By introducing nonlinear effects through the generalised additive model, a change can be clearly found in the trend of the partial effect of temperature on the log-rate of chimney fires.

Our main intent here is to demonstrate the application of appropriate statistical models for answering practical questions of interest. Decision tree models are suitable for describing highly irregular relationships, but this does not appear to be the case here. In contrast, Poisson regression models and generalised additive models are more appropriate here; both are able to find more variables that are relevant and provide more precise descriptions of the relationship. Generalised additive models are more powerful than Poisson regression models, in that they can deal with nonlinear effects of individual variables and can thus provide answers one level deeper.

The analysis above is unlikely to be exhaustive. For example, more informative variables can perhaps be included initially or constructed from others.

6 Firewise programme

6.1 Problem and data

In this scenario, we examine whether there are any changes of the fire rates in proximity to schools where the New Zealand Fire Service has delivered the Firewise programme.

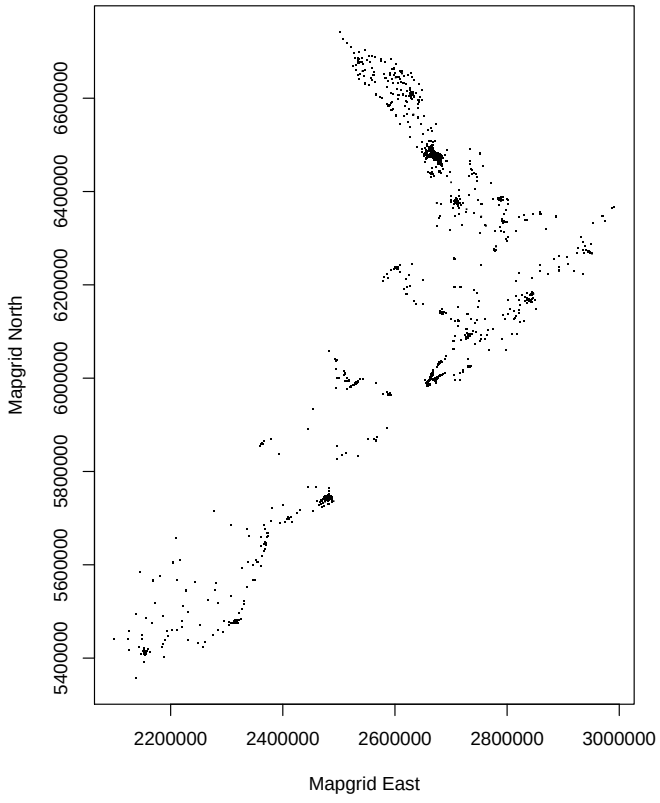

Figure 10: Locations of schools with Firewise programme delivered

We have removed early childhood education centres from the data, because it seems harder to define a possible catchment area for these. Parents base their choice of early childhood centre on many factors, including but not limited to proximity (for instance, some parents choose to use centres that are close to their workplaces rather than their homes). This left 1680 schools that had at least one Firewise programme delivered between 1 Jan 2004 and 31 Dec 2007; the locations of these schools are plotted in Figure 10.

Since the four-year data is unlikely to reveal any long-term effect, we focus our attention on a short-term effect: whether there is a difference in the rates of structure fires (of incident types 1101 and 1102) in proximity to a school, between the 52-week period after the completion of a Firewise programme at the school and any other period. The reason for using 52 weeks is to fully remove any potential correlation between the delivered Firewise programme and the seasonal effects that have not been accounted for by other variables. It also appears to make sense that the one-year period is perhaps where a short-term effect of the Firewise programme, if any, may hopefully exist. We treat the circular area that is centered about a school with radius 1 kilometer as its approximate proximity. The number of structure fires inside this circle is then counted for each of the $4 \times 52 = 208$ weeks (from 1 Jan 2004 to 26/Dec/2007). If a fire incident is located inside more than one circle, it is counted only once for its nearest school and hence there is no duplicated counting. This is equivalent to using the Voronoi tessellation inside the overlapping areas and the Voronoi tessellation is independent of the fire incidents. Over the four year period, there are in total 10761 structure fires, out of 22440 across New Zealand, that occurred in the above-defined proximities of the schools. This definition of proximity is somewhat abritrary and the results may differ if this definition is changed.

For 1680 schools and 208 weeks, the data set that is created for this scenario is quite large and has $1680 \times 208 = 349440$ observations; see a small random subset given in Appendix B.1. It contains the following variables:

`school`	a unique number for each school
`year`	year (ranging from 1 to 4)
`season`	one of `Spr`, `Sum`, `Aut` and `Win`
`gf`	indicator whether the week contains the Guy Fawkes day
`stathol`	indicator whether the week contains any statutory holiday
`firewise`	indicator whether a Firewise programme has been delivered to the school and completed within the last 52 weeks
`count`	number of structure fires occurred in the proximity of the school in the week

We have applied both the usual and the mixed-effects Poisson regression models to this data set.

6.2 Poisson regression

The fitted Poisson regression model is given in Appendix B.2. For the variables that are of less interest, `gf`, `season` and `year` are highly significant, while `stathol` is insignificant. Of central interest is `firewise`, which is not significant. This suggests that overall there is no statistical evidence that there is a difference in the rate of house fires in the proximities of schools within the 52-week period after the completion of a Firewise programme at a school, as compared with the other periods. This does not neccessarily mean that the Firewise programme does not have any effect, since there could be some effects that are too weak to be detectable or there may exist some long-term effects which can not be examined here. It may also be because of the comparison we made, between the 52-week period after the completion of the programme and other time periods, which is perhaps not the most appropriate way of detecting a short-term effect.

6.3 Mixed-effects Poisson regression

There is, however, a potential problem in applying the Poisson regression model to the data here. By using the Poisson regression model above we assume that the weekly rates of house fires for all school regions are the same, if all the other variables take the same values. It is clear that this assumption does not really hold in practice: schools can be different in many aspects, such as catchment area, rural vs. urban, decile rating, number of students, weather, geographic location and natural environment. It would be very costly, even if it were possible, to collect all these and other differentiating data and, even if they are available, there is still no guarantee that all differences among the schools have been properly addressed. If one treats `school`

as a categorical variable, there will be $1680 - 1 = 1679$ new parameters adding to the model. Not only is fitting a model with so many parameters almost computationally infeasible, but the resulting model will also be dramatically over-fitted (namely too many parameters used, thus giving a bad fit).

A better alternative is to use the mixed-effects Poisson regression model, where a statistical distribution is introduced to describe the effects of these schools on the log-rate of house fires. With the effects of different schools accounted for in this way, a much better fitted model can be obtained and the significance levels of the other variables more accurately assessed.

The fitted mixed-effects Poisson regression model is given in Appendix B.3. The variable `firewise` is again insignificant. Note that the value of AIC is greatly reduced from 97083 to 67634, suggesting that the new fit is considerably better and that using a mixed-effects model is a much better choice.

6.4 Summary and remarks

The models fitted to the data show no evidence of a decline in structure fires in a 1 kilometre radius around a school during the 52 week period following the completion of a Firewise programme at that school. As mentioned above, this does not rule out the possibility of such an effect. We were only able to consider a short term effect and there may be insufficient data to detect such an effect if it is weak. We used a simple approximation for the catchment area of a school, but this does not allow for differences in catchment areas between rural and urban areas, for instance.

If any such decline were apparent, we would need to take great care with the interpretation. A decline associated with the Firewise programme would not imply causation. Ideally, a comparison should be made between those schools that did not have the Firewise programme delivered (the controls) and schools that did. For instance, if there were evidence of a decline, it might be simply due to a general decline in structure fires in proximity to schools.

Earlier data from a period before the Firewise programme was delivered could perhaps be used to study both the short- and long-term effects. However, as noted above, it would be necessary to consider schools both with and without the programme. A practical difficulty is that there have been changes in the way in which data has been gathered over the years, so that earlier data may not be easily comparable in any case.

7 Blenheim suspicious fires

7.1 Problem and data

In this scenario, we examine methods for detecting increases in the frequency of fires possibly due to fire laying by individuals (arson). The goal is to detect and isolate such events automatically from the collected data of fire incidents.

The data we use here for illustrative purposes are the fire incidents that occurred in the area of Blenheim between 1 Jan 2004 and 31 Dec 2007. It is known that a convicted arsonist had been laying vegetation fires during the period between late October 2006 to early January 2007 in this area. It appears to us that these intentionally-laid fires are most likely correlated to those labeled “suspicious” by fire fighters on the spot. Thus we have turned the original problem into one that detects changes in the frequency of suspicious fires.

Unlike in the previous two scenarios, this problem is very “irregular”. There may exist a large number of potentially relevant variables, which may be of numerical or categorical types and have missing values. The frequency of suspicious fires is most unlikely to exhibit a nice, monotonic relationship with the variables used. An arsonist may operate in certain time periods and in certain neighborhoods and light fires of certain types. Such an irregularity makes it very hard for traditional statistical techniques, such as generalised linear or additive models or univariate hypothesis testing, to provide nice solutions. Modern data mining techniques, on the other hand, can be quite suitable for solving problems like this. We demonstrate below how decision tree models can be applied in this scenario.

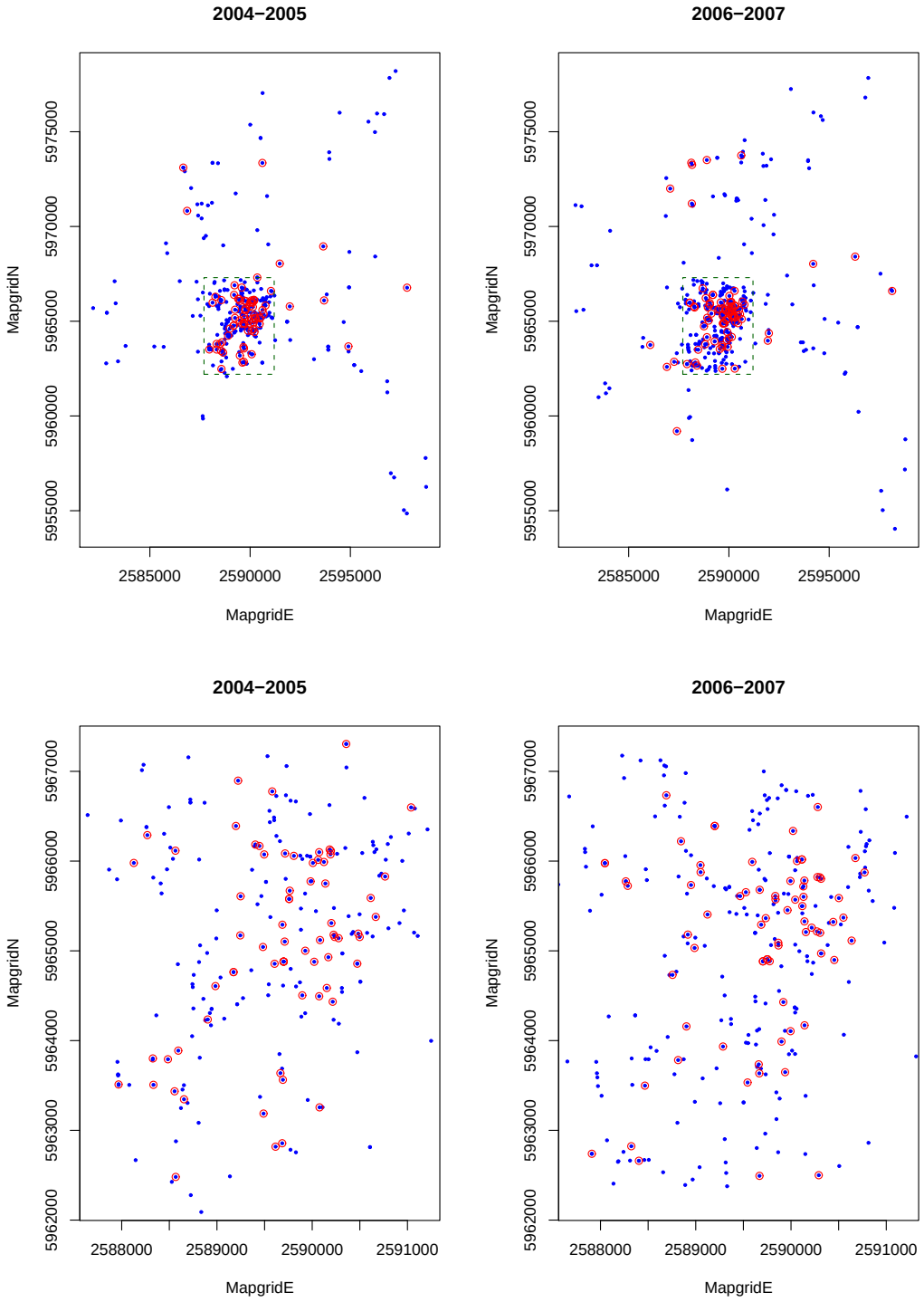

Figure 11: All fires in the area of Blenheim in 2004–2007 are plotted, as shown by dots, while suspicious fires are also marked by circles.

The analysis below includes all the fire incidents in 2004–2007 in the rectangular region specified by

$$(\texttt{MapgridE}, \texttt{MapgridN}) \in (2582000, 2600000) \times (5954000, 5979000).$$

This gives a total of 704 fire incidents, 171 of which are suspicious. To facilitate the analysis, they are further divided into two data sets, one for those events that occurred in 2004–2005 and the other for those in 2006–

2007. These data sets contain, respectively, 318 and 386 fire incidents, of which 80 and 91 are suspicious. These fires are plotted in Figure 11.

Both data sets thus created contain the following variables:

`MapgridE`	map grid east (as given in the original data set)
`MapgridN`	map grid north (as given)
`CurrentUrbanRural`	binary variable, indicating whether an urban (1) or rural (0) area (as given)
`AlarmMethodCode`	alarm method code (as given)
`IncidentType`	type of incident (as given)
`Heatsource`	heat source (as given)
`Obj1ignited`	object ignited (as given)
`time`	time of the day $\in [0, 24)$,
`day`	day of the two-year period $\in \{1, 2, \ldots, 731\}$
`dayweek`	day of the week ($1 =$ Monday, $\ldots$, $7 =$ Sunday)
`type`	binary variable (`Susp` = suspicious; `Other` = other type)

A small random subset from both data sets is given in Appendix C.1. Note that both `Heatsource` and `Obj1ignited` contain missing values, which are represented by `NA` (originally coded `0`).

In the following, we demonstrate how decision trees can be applied to detect changes in the frequencies of suspicious fires between 2004–2005 and 2006–2007. Two approaches are adopted below. One is to build a classification tree from one data set. Since patterns (rules) found from one data set are not necessarily changes, all found rules are tested against the observations from the other data set that are covered by the same rules. As a result, significant rules correspond to the differences between the two data sets. The other approach is to build a Poisson regression tree that directly describes the differences between the two data sets. The second approach is more efficient.

7.2 Pattern discovery

We begin by building a classification tree from the first data set and test all the found rules against the second data set; and then build another classification tree from the second data set and test the found rules against the first data set. This should help us find the differences between the two data sets. It is also possible to use Poisson regression trees, where daily or weekly counts of fires could be used as the response.

Classification trees are perhaps the most popular type of decision tree to be used in practice. The goal is to maximise the separation of observations that belong to different classes (here `Other` *vs.* `Susp`). The resulting classification tree should then tend to have different proportions for a class at the terminal nodes, where predictions are made for new observations based on the proportions achieved from the training data.

The classification tree built from the first two years of data is shown in Figure 12, with its more detailed text version given in Appendix C.2. The tree identifies 7 situations (or rules) where the proportions of suspicious fires should be distinguished. Each rule here corresponds to a path from the root of the tree to a terminal node, where the proportions of different types of fires are estimated from the training data. These seven rules are listed in Table 1, in the ascending order of their estimated proportions of suspicious fires. These rules by themselves are different, not only in terms of their estimated proportions of suspicious fires, but also in terms of the splitting criteria used along their paths down to a terminal node. While rules built from one data set shed light on how the proportions of suspicious fires are related to other variables, they are not directly indicating differences from another data set, of course. Some rules may be produced due to effects that are not of direct interest here, e.g., seasonal effects. However, we can compare these rules with the second two years of data. Then patterns that are due to factors such as seasonal effects should turn out to be insignificant. Any significant rules that may remain can only be attibuted to the differences between the two data sets.

The test we use is the likelihood ratio test for Poisson distributions; see Appendix D for details of the test. As shown in Table 1, there is only one rule, Rule 6, that is remarkably significant, with a p-value of

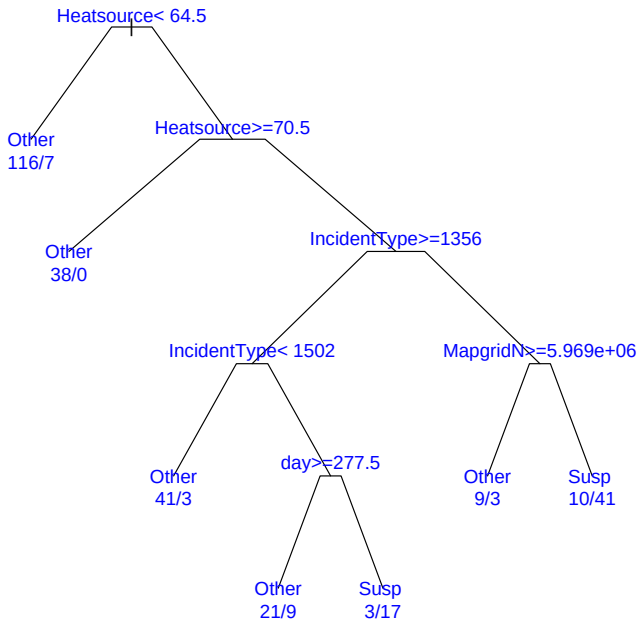

Figure 12: Classification tree built from fire incidents occurred between 1 Jan 2004 and 31 Dec 2005 in the Blenheim area

		2004–2005		2006–2007		
	Proportion	Other	Suspicious	Other	Suspicious	P-value
Rule 1	0.000	38	0	27	2	0.098
Rule 2	0.057	116	7	102	3	0.28
Rule 3	0.068	41	3	50	3	0.64
Rule 4	0.250	9	3	11	6	0.54
Rule 5	0.300	21	9	39	11	0.058
Rule 6	0.804	10	41	55	53	0.000000017
Rule 7	0.850	3	17	11	13	0.067

Table 1: Rules built from data between 2004–2005 and tested by data between 2006–2007

1.7×10^{-8}. Nonetheless, the frequencies suggest that the significance is largely due to more fires of type "Other" in 2006–2007, rather than a change in the frequency of suspicious fires.

		2004–2005		2006–2007		
	Proportion	Other	Suspicious	Other	Suspicious	P-value
Rule 1	0.000	13	0	9	0	0.69
Rule 2	0.040	106	9	97	4	0.31
Rule 3	0.116	69	37	129	17	0.0000022
Rule 4	0.200	30	9	44	11	0.24
Rule 5	0.682	16	8	14	30	0.0011
Rule 6	0.935	4	17	2	29	0.15

Table 2: Rules built from data between 2006–2007 and tested against data between 2004–2005

To be complete, we also need to find rules from the second data set and test them against the first data set.

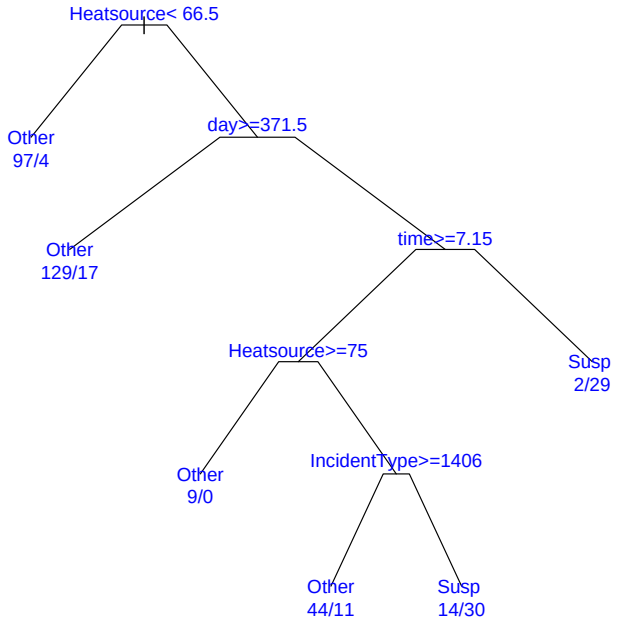

Figure 13: Classification tree built from fire incidents occurred between 1 Jan 2006 and 31 Dec 2007 in the Blenheim area

The decision tree built from the second two years of data is shown in Figure 13, with its text version given in Appendix C.2. There are in total 6 rules (namly 6 terminal nodes) that are found from the data set, and are listed in Table 2. Two of the rules are highly significant and can be reformatted as follows:

```
Rule 3:  IF     66  < Heatsource
                371 < day
         THEN   type = Other

Rule 5:  IF     66  < Heatsource < 75
                day <= 371
                7.15 <= time
                IncidentType < 1406
         THEN   type = Susp
```

Rule 5 corresponds to a significant increase (about 22) in the frequency of suspicious fires and appears to be directly related to the arson case. In particular, it indicates that a change, in terms of the proportion of suspicious fires, has occurred during the day time (after 7:09 am), before day 371 (7/Jan/2007), with heat source covering a range with "Cigarettes, Matches and Candles", and with incident types covering a range with "Vegetation fires". It possesses a very different characteristic from Rule 6, which has the highest proportion of suspicious fires but specifies that the fires occurred between 12:00am and 7:09am and is not significant as a change. Rule 3 is also significant but corresponds to a decrease in the frequncy of suspicious fires, as well as a substantial increase in the frequency of other fires, which occurred after day 371 (7/Jan/2007).

7.3 Direct change detection

Alternatively, we can build a single Poisson regression tree for detecting changes directly between the two data sets. Since there does not appear to be an implementation available in R (or in other data mining software packages), we have partially implemented this method and applied it to the Blenheim data sets; see Appendix C.3 for the tree produced.

The main idea of this method is to search for the optimal splitting point in the value of each of the given variables so that two given data sets differ most, in the sense of the likelihood ratio test (Appendix D). This process is recursive, which divides the two data sets according to the found splitting criteria into smaller and smaller subsets. It proceeds until some stopping criterion is satisfied, such as when there are too few observations left (less than 10 here) or when the minimal p-value among all candidate splits is larger than a user-chosen threshold value (0.05 here). This gives a tree-structured model, which can be further pruned back by using the AIC, for example, to avoid overfitting.

	2004–2005		2006–2007		
	Other	Suspicious	Other	Suspicious	P-value
Rule 1	4	0	18	28	0.000000000030
Rule 2	23	14	63	1	0.000000076
Rule 3	1	7	12	3	0.0018
Rule 4	4	0	9	6	0.0058
Rule 5	90	14	66	4	0.0083
Rule 6	11	5	3	14	0.0096
Rule 7	22	9	23	1	0.025
Rule 8	10	9	5	2	0.038
Rule 9	4	0	6	4	0.051
Rule 10	4	0	6	3	0.10
Rule 11	42	13	59	14	0.23
Rule 12	17	9	21	11	0.73
Rule 13	6	0	4	0	0.82

Table 3: Rules found by direct detection of changes between the data in 2004–2005 and in 2006–2007

All 13 rules that are found by the Poisson regression tree are listed in the ascending order of their p-values in Table 3. The two most significant rules are as follows:

```
Rule 1:  IF     1200 < IncidentType <= 1500
                284 < day <= 384
                MapgridE <= 2592500
                MapgridN <= 5966004
         THEN   pvalue = 3.01e-11 [(4 0) (18 28)]

Rule 2:  IF     1200 < IncidentType
                434 < day <= 586
         THEN   pvalue = 7.59e-08 [(23 14) (63 1)]
```

Being the most significant, Rule 1 appears to directly relate to the arson case. In particular, it suggests that a change, in terms of fire frequencies, has occurred between day 284 (11/Oct/2006) and day 384 (19/Jan/2007), with incident types covering a range with "Vegetation fires", and in a particular neighbourhood with `MapgridE` ≤ 2592500 and `MapgridN` ≤ 5966004. The change is due to a substantial increase in suspicious fires (from 0 to 28), as well as an increase in other fires (from 4 to 18). The 28 suspicious fires in 2006–2007 that are covered by this rule are shown in Figure 14.

Rule 2 specifies a situation where there is a decrease in the number of suspicious fires and yet an increase in the number of other fires. This change took place between day 434 (10/Mar/2007) and day 586 (9/Aug/2007), for incident types less than 1200 (thus excluding "Vegetation fires").

The general conclusions that are drawn here are similar to those in Section 7.2, but the rules found by detecting changes directly both provide more detail about the differences between the two data sets and are supported by much stronger statistical evidence.

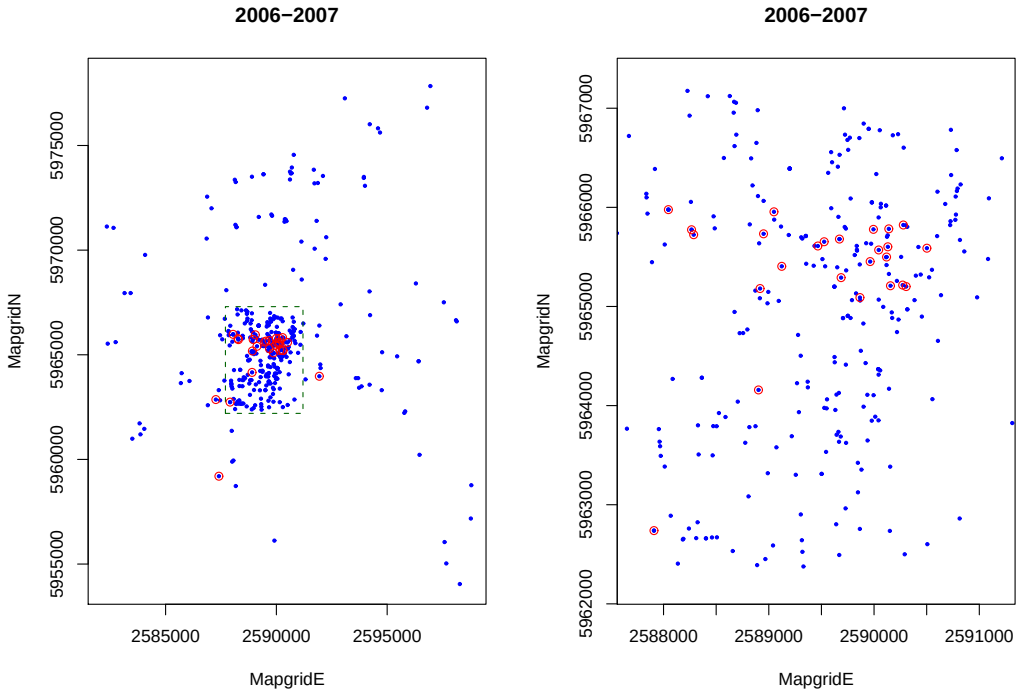

Figure 14: All fires in the area of Blenheim in 2006–2007 are plotted, as shown by dots, while suspicious fires covered by Rule 1 are also marked by circles.

7.4 Summary and remarks

In this scenario, we have demonstrated how irregular changes of fire frequencies can be detected by using decision tree methodology. Two specific methods are developed and applied to the data, with mutually supporting conclusions obtained.

The second method can also be applied to build a tree for suspicious fires only or for any number of types of fires. Further improvements on the method and its implementation are possible. We have not found similar methods in the literature to those presented above.

The direct change detection method makes use of the data more efficiently. It is possible that there may be cases where the first method may fail to detect any significant change, but the second may still succeed in doing so. The increased sensitivity of this method could be critically important for providing early warning through an online monitoring system.

8 Discussion and recommendations

8.1 Expertise required

A good knowledge of both the subject matter and statistics is needed to both analyze the data and interpret the results. Knowledge of the subject matter is needed to determine the questions of interest and is useful in selecting the most relevant variables for the problems under investigation. It also helps reduce substantially the total number of variables that will be used in a statistical model and increases the efficiency of statistical estimation. The choice of an appropriate model depends on the question of interest, and requires some statistical expertise. Some computer programming is always needed, if only to format the data correctly. Beyond that some familiarity with computer packages is needed, and it may even be necessary to implement the method that appears to be most suitable.

8.2 Statistical and data mining methodology

Both conventional statistical methods and modern data mining tools can be useful for solving practical problems. Which sets of tools are more suitable depends on the specific problem under consideration. Conventional statistical methods usually work better if their assumptions are satisfied reasonably well, while data mining methodology is often applied in settings where the data exhibit some irregularity.

Although each of the three scenarios we considered is concerned with possible changes in the frequency of fires, each scenario is different in nature, has different kinds of questions associated with it and responds best to different statistical or data mining techniques. The first two scenarios were best addressed using conventional statistical techniques while the third responded well to a data mining technique.

8.3 Software

Our investigation for this Fire Service project has been carried out exclusively in R (Ihaka and Gentleman, 1996; R Development Core Team, 2006). We believe it is the most appropriate environment for solving these problems. R is a free, open-source software package and is well supported by the international R development core team. It contains an extensive collection of built-in functionalities and tools for data analysis and modelling, as well as many add-on packages that are contributed by researchers around the world to implement their latest research. Indeed, implementations in R can be found for almost all of the methods mentioned above. It provides a nice programming environment, facilities for implementing new ideas quickly, and interfaces to other programming languages and software, if these are required. It also has many elegant graphical functionalities that can help understand data better, discover hidden relationships and present results nicely. An R programme can be easily run online at regular times, in an automatic manner, and/or with options specified for different queries. R is very widely used, and is the preferred computing environment for professional statisticians.

Although we did not use them in this project, there are also some other data mining software packages that may be potentially useful, depending on the problems under investigation. Among them, WEKA (Witten and Frank (2005)) is an internationally-known, freely-available data mining benchmarking package. Although it is implemented in the JAVA language, using the implemented methods does not require knowledge of JAVA and can be fully done through a user-friendly graphical interface. A large number of data mining methods have been implemented in WEKA, perhaps more than any other single data mining software package.

There are also a number of commercial data mining software packages, including the better known ones: Enterprise Miner of SAS, Insightful Miner of S-Plus, Data Mining Suite of Salford Systems, and RuleQuest. While these systems usually contain implementatons of commonly-used methods such as decision trees, neural networks and support vector machines, they may also provide specialised methods that have been developed by well-known researchers, such as MARS and RandomForest of Salford Systems, and Cubist and GritBot in RuleQuest.

Nevertheless, the major research effort in the data mining community is on classification and regression problems, where the response variable is either categorical or continuous. For count data, as in the three scenarios studied above, there appear to be very few implemented methods that can be used off the shelf to answer questions that may arise from the Fire Service perspective. It is therefore not entirely clear to us what additional benefits these commercial packages can bring to the analysis of these data.

8.4 Data

The data collected on fire events is extensive and very detailed – this is a very rich data set. We experienced some difficulties with the large numbers of categories available for some fields and, the possible overlap between them, and the fact that different types of fire events require different fields to be entered.

The numbers of categories for some types of incident are large, and the categories are not always mutually exclusive. While some categories of fire appear easy to identify (e.g. house fires), fires lit as a result of deliberate fire laying are harder to identify, since there are several categories that could cover this. Thus under fire cause, deliberately lit fires can fall into several categories including “unlawful”, “legality not known”, “suspicious”, and “not classified”.

Not all fire events have data recorded for each field. In particular, miscellaneous fires have missing entries in many fields, including Heat Source and Object Ignited. Miscellaneous fires constituted 37.1% of the fire events in 2003-2007 (35,345 of a total 95,303 events). It would be desirable to have greater consistency in the data fields recorded for each event. It may on occasions be tempting to classify fires as miscellaneous, since that then requires fewer additional fields to be entered.

8.5 Recommendations

In summary, we make the following recommendations:-

- If the New Zealand Fire Service were to implement an automatic detection system, it would need to be tailored to detect particular scenarios of interest, and supplemented with statistical analyses of patterns that might be detected.
- The decision tree analysis developed here has potential for real time detection of changes in the incidence of fire events such as those associated with arson. We recommend that further development work be undertaken on this.
- We recommend using the statistical environment R for data analysis. It is the preferred tool for professional statisticians, has the flexibility to allow non-standard analyses and can easily be integrated with data servers. In addition, it is free, open-source software.
- The temporal and spatial data originating from the computerised dispatch system is effective for pattern detection, but the large number of response options, many of which overlap or are potentially ambiguous, may lead to loss of information. We recommend that further consideration be given to data collection design.

A Chimney fires – R output

A.1 Data

```
> city.chmn.fires[sort(sample(nrow(city.chmn.fires), 30)),]
```

	city	pop	day	month	year	weekend	fwi	humid	T	D1	D2	count
320	Akl	1208	320	11	1	FALSE	1.78	53	18.8	-1.5	0.0	0
463	Akl	1208	463	4	2	FALSE	10.80	51	25.5	3.5	-0.7	0
1043	Akl	1208	1043	11	3	FALSE	0.38	79	16.4	-6.7	3.5	0
1421	Akl	1208	1421	11	4	FALSE	5.08	64	20.1	1.1	-1.4	0
1859	Ham	155	398	2	2	FALSE	21.14	56	26.0	-1.0	2.0	0
2008	Ham	155	547	6	2	FALSE	3.12	56	10.0	-5.0	4.0	0
2071	Ham	155	610	9	2	FALSE	1.35	75	16.0	0.0	1.0	0
2161	Ham	155	700	11	2	FALSE	4.95	53	21.0	1.0	5.0	0
2329	Ham	155	868	5	3	FALSE	0.32	76	10.3	-2.3	1.5	0
2361	Ham	155	900	6	3	TRUE	0.00	95	10.0	2.2	-5.6	0
2554	Ham	155	1093	12	3	FALSE	12.12	64	20.0	0.0	-1.0	0
2613	Ham	155	1152	2	4	TRUE	15.70	54	22.0	-4.0	3.0	0
2884	Ham	155	1423	11	4	FALSE	5.86	61	20.0	1.0	-1.0	0
2902	Ham	155	1441	12	4	FALSE	3.48	77	20.0	0.0	0.0	0
3328	Wel	276	406	2	2	FALSE	21.69	61	25.0	2.0	-2.0	0
4228	Wel	276	1306	7	4	TRUE	1.34	69	14.0	0.0	0.0	0
4384	Chr	361	1	1	1	FALSE	45.77	31	29.0	6.0	5.0	0
4419	Chr	361	36	2	1	FALSE	5.52	53	16.0	-8.0	3.0	0
4934	Chr	361	551	7	2	FALSE	2.57	89	10.3	-3.9	5.2	0
4988	Chr	361	605	8	2	TRUE	2.99	85	5.7	-6.0	1.0	0
5527	Chr	361	1144	2	4	TRUE	9.83	56	20.0	4.0	-1.0	0
5607	Chr	361	1224	5	4	FALSE	2.42	62	16.9	-2.1	4.0	0
5644	Chr	361	1261	6	4	FALSE	8.37	77	8.0	1.0	-3.0	0
5959	Dun	111	115	4	1	TRUE	1.68	74	13.0	-1.0	0.0	0
6346	Dun	111	502	5	2	FALSE	0.47	64	9.3	-2.1	2.0	0
6760	Dun	111	916	7	3	FALSE	0.07	71	6.7	-5.6	-0.6	1
7438	Inv	47	133	5	1	FALSE	2.06	93	10.0	-2.0	1.0	2
8155	Inv	47	850	4	3	TRUE	0.76	77	13.0	-2.0	1.0	0
8320	Inv	47	1015	10	3	FALSE	2.55	73	12.0	5.0	-4.0	1
8587	Inv	47	1282	7	4	FALSE	0.00	86	5.0	-1.0	-2.0	1

A.2 Poisson decision trees

```
> library(rpart)
> rpart(count ~ ., data=city.chmn.fires, method="poisson", cp=0.005)

n= 8766

node), split, n, deviance, yval
      * denotes terminal node

 1) root 8766 4300.0 0.0997
   2) T>=11.9 6550 2130.0 0.0546
     4) T>=16.9 3280  611.0 0.0249
       8) month=1,2,3,4,5,10,11,12 3121  492.0 0.0204 *
       9) month=6,7,8,9 159   89.5 0.1120 *
     5) T< 16.9 3270 1410.0 0.0845
      10) city=Ham 522   39.1 0.0094 *
      11) city=Akl,Wel,Chr,Dun,Inv 2748 1300.0 0.0990
        22) month=1,2,3,5,10,11,12 1537  545.0 0.0640 *
        23) month=4,6,7,8,9 1211  712.0 0.1430 *
   3) T< 11.9 2216 1720.0 0.2330
     6) city=Akl,Ham,Wel,Chr,Inv 1681 1140.0 0.1800
      12) city=Ham 164   29.8 0.0287 *
      13) city=Akl,Wel,Chr,Inv 1517 1080.0 0.1960 *
     7) city=Dun 535  501.0 0.3940 *
```

A.3 Poisson regression

Using all covariates

```
> r = glm(count ~ . - day - pop, offset=log(pop), data=city.chmn.fires,
          family="poisson")
> summary(r)

Call:
glm(formula = count ~ . - day - pop, family = "poisson", data = city.chmn.fires,
    offset = log(pop))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.297  -0.481  -0.291  -0.166   3.998

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.78670    0.52215  -16.83  < 2e-16 ***
cityHam     -0.26029    0.31688   -0.82  0.41141
cityWel      1.32943    0.14136    9.40  < 2e-16 ***
cityChr      1.27437    0.14033    9.08  < 2e-16 ***
cityDun      2.97482    0.13398   22.20  < 2e-16 ***
cityInv      3.36725    0.14180   23.75  < 2e-16 ***
month2       0.16314    0.33938    0.48  0.63073
month3       0.48466    0.31047    1.56  0.11851
month4       0.89715    0.28883    3.11  0.00190 **
month5       1.07054    0.28805    3.72  0.00020 ***
month6       1.40721    0.29454    4.78  1.8e-06 ***
month7       1.36316    0.29527    4.62  3.9e-06 ***
month8       1.41084    0.28980    4.87  1.1e-06 ***
month9       1.36305    0.28162    4.84  1.3e-06 ***
month10      0.75024    0.29363    2.56  0.01062 *
month11      0.51057    0.30497    1.67  0.09410 .
month12     -0.13101    0.35034   -0.37  0.70843
year        -0.07111    0.03023   -2.35  0.01865 *
weekendTRUE  0.17980    0.07257    2.48  0.01323 *
fwi          0.00361    0.00810    0.45  0.65563
humid       -0.00349    0.00323   -1.08  0.28055
T           -0.09735    0.01534   -6.35  2.2e-10 ***
D1           0.00219    0.01371    0.16  0.87287
D2           0.01085    0.01164    0.93  0.35103
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5605.1  on 8765  degrees of freedom
Residual deviance: 3390.8  on 8742  degrees of freedom
AIC: 5058

Number of Fisher Scoring iterations: 7
```

Model selection using AIC

```
> step(r)
Start:  AIC= 5058
 count ~ (city + pop + day + month + year + weekend + fwi + humid +
    T + D1 + D2) - day - pop

          Df Deviance  AIC
- D1       1     3391 5056
- fwi      1     3391 5056
- D2       1     3392 5057
- humid    1     3392 5057
<none>           3391 5058
- year     1     3396 5062
- weekend  1     3397 5062
- T        1     3431 5096
- month   11     3474 5119
- city     5     4287 5945

Step:  AIC= 5056
 count ~ city + month + year + weekend + fwi + humid + T + D2

          Df Deviance  AIC
- fwi      1     3391 5054
- D2       1     3392 5055
- humid    1     3392 5055
<none>           3391 5056
- year     1     3396 5060
- weekend  1     3397 5060
- T        1     3444 5107
- month   11     3490 5133
- city     5     4391 6046

Step:  AIC= 5054
 count ~ city + month + year + weekend + humid + T + D2

          Df Deviance  AIC
- D2       1     3392 5053
- humid    1     3393 5054
<none>           3391 5054
- year     1     3396 5058
- weekend  1     3397 5058
- T        1     3445 5106
- month   11     3490 5132
- city     5     4398 6052

Step:  AIC= 5053
 count ~ city + month + year + weekend + humid + T

          Df Deviance  AIC
- humid    1     3393 5053
<none>           3392 5053
- year     1     3398 5057
- weekend  1     3398 5057
- T        1     3445 5105
- month   11     3495 5135
- city     5     4404 6056

Step:  AIC= 5053
```

```
 count ~ city + month + year + weekend + T

          Df Deviance  AIC
<none>           3393 5053
- year     1     3399 5056
- weekend  1     3399 5057
- T        1     3459 5116
- month   11     3506 5143
- city     5     4467 6116
```

A.4 Generalised additive models

```
> library(gam)
> r = gam(count ~ s(T,df=3) + year + s(humid,df=3) + month + weekend + city,
          offset=log(pop), data=city.chmn.fires, family="poisson")
> summary(r)

Call: gam(formula = count ~ s(T, df = 3) + year + s(humid, df = 3) +
    month + weekend + city, family = "poisson", data = city.chmn.fires,
    offset = log(pop))
Deviance Residuals:
   Min     1Q Median     3Q    Max
-1.208 -0.488 -0.287 -0.154  4.036

(Dispersion Parameter for poisson family taken to be 1)

    Null Deviance: 4298 on 8765 degrees of freedom
Residual Deviance: 3378 on 8741 degrees of freedom
AIC: 5047

Number of Local Scoring Iterations: 8

DF for Terms and Chi-squares for Nonparametric Effects

                 Df Npar Df Npar Chisq P(Chi)
(Intercept)       1
s(T, df = 3)      1       2      10.03   0.01
year              1
s(humid, df = 3)  1       2       4.36   0.11
month            11
weekend           1
city              5
```

B Firewise programme – R output

B.1 Data

For structure fires of incident types 1101 and 1102:

```
> all.fw[sort(sample(nrow(all.fw), 30)),]
```

	school	year	season	gf	stathol	firewise	count
17408	84	3	Spr	FALSE	FALSE	FALSE	0
28872	139	4	Aut	FALSE	FALSE	FALSE	0
43153	208	2	Spr	TRUE	FALSE	TRUE	0
45326	218	4	Win	FALSE	FALSE	FALSE	0
64910	313	1	Aut	FALSE	FALSE	FALSE	0
72497	349	3	Sum	FALSE	FALSE	TRUE	0
75011	361	3	Win	FALSE	FALSE	TRUE	0
75790	365	2	Win	FALSE	FALSE	FALSE	0
90108	434	1	Spr	FALSE	FALSE	FALSE	0
98911	476	3	Sum	FALSE	FALSE	TRUE	0
101031	486	3	Spr	FALSE	FALSE	FALSE	1
110672	533	1	Aut	FALSE	FALSE	FALSE	0
112558	542	1	Win	FALSE	FALSE	FALSE	0
131782	634	3	Aut	FALSE	FALSE	TRUE	0
141140	679	3	Aut	FALSE	FALSE	TRUE	1
154543	743	4	Sum	FALSE	FALSE	TRUE	0
197876	952	2	Aut	FALSE	FALSE	FALSE	0
202442	974	2	Sum	FALSE	FALSE	FALSE	0
227749	1095	4	Spr	FALSE	FALSE	FALSE	0
259896	1250	2	Sum	FALSE	TRUE	TRUE	0
273073	1313	4	Aut	FALSE	FALSE	FALSE	0
278375	1339	2	Aut	FALSE	FALSE	FALSE	0
280147	1347	4	Win	FALSE	TRUE	FALSE	0
290841	1399	2	Sum	FALSE	FALSE	FALSE	1
295974	1423	4	Spr	FALSE	FALSE	FALSE	0
299630	1441	3	Sum	FALSE	TRUE	TRUE	0
313801	1509	3	Win	FALSE	FALSE	TRUE	0
329736	1586	2	Sum	FALSE	FALSE	FALSE	0
334978	1611	2	Spr	FALSE	FALSE	TRUE	0
348651	1677	1	Spr	FALSE	TRUE	FALSE	0

B.2 Poisson regression

```
> r = glm(count ~ . - school, family=poisson, data=all.fw)
> summary(r)

Call:
glm(formula = count ~ . - school, family = poisson, data = all.fw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2894  -0.2546  -0.2469  -0.2407   6.4326

Coefficients:
              Estimate Std. Error  z value Pr(>|z|)
(Intercept)  -3.547513   0.030003 -118.240  < 2e-16 ***
year          0.023451   0.008686    2.700  0.00694 **
seasonSum    -0.084339   0.028973   -2.911  0.00360 **
seasonAut     0.010104   0.028087    0.360  0.71906
seasonWin     0.120909   0.027312    4.427 9.56e-06 ***
gfTRUE        0.280670   0.064187    4.373 1.23e-05 ***
statholTRUE  -0.051231   0.030561   -1.676  0.09368 .
firewiseTRUE -0.021838   0.021578   -1.012  0.31152
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 76335  on 349439  degrees of freedom
Residual deviance: 76246  on 349432  degrees of freedom
AIC: 97083

Number of Fisher Scoring iterations: 6
```

B.3 Mixed-effects Poisson regression

```
> library(lme4)
> r = glmer(count ~ . - school + (1 | school), family=poisson, data=all.fw)
> summary(r)

Generalized linear mixed model fit by the Laplace approximation
Formula: count ~ . - school + (1 | school)
   Data: all.fw
   AIC   BIC logLik deviance
 67634 67731 -33808    67616
Random effects:
 Groups Name        Variance Std.Dev.
 school (Intercept) 1.57     1.25
Number of obs: 349440, groups: school, 1680

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.21789    0.04448   -94.8  < 2e-16 ***
year          0.02434    0.00873     2.8   0.0053 **
seasonSum    -0.08470    0.02913    -2.9   0.0036 **
seasonAut     0.00947    0.02824     0.3   0.7373
seasonWin     0.12091    0.02746     4.4  1.1e-05 ***
gfTRUE        0.28071    0.06453     4.4  1.4e-05 ***
statholTRUE  -0.05111    0.03072    -1.7   0.0962 .
firewiseTRUE -0.03734    0.02221    -1.7   0.0927 .
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Correlation of Fixed Effects:
            (Intr) year   sesnSm sesnAt sesnWn gfTRUE stTRUE
year        -0.475
seasonSum   -0.317  0.003
seasonAut   -0.332  0.000  0.512
seasonWin   -0.341  0.003  0.518  0.534
gfTRUE      -0.146 -0.001  0.216  0.226  0.234
statholTRUE -0.042 -0.024 -0.131 -0.054  0.007  0.038
firewisTRUE -0.078 -0.140  0.017  0.029 -0.003 -0.001 -0.006
```

C Blenheim suspicious fires - R output

C.1 Data

```
> blenheim1[sort(sample(nrow(blenheim1), 5)),]

      MapgridE MapgridN CurrentUrbanRural AlarmMethodCode IncidentType
19828  2588409  5965751                 1              11         1501
22427  2586740  5972911                 0              11         1201
28762  2590169  5964929                 1              11         1502
28821  2593940  5973917                 0              11         1312
43544  2589277  5971743                 0              52         1201
      Heatsource Obj1ignited day   time dayweek  type
19828         NA          NA 108 17.500       6 Other
22427         67         911 245 15.300       3 Other
28762         NA          NA 676  1.683       7  Susp
28821         99         723 704  8.617       7 Other
43544         67         911 555  9.050       5 Other

> blenheim2[sort(sample(nrow(blenheim2), 5)),]

      MapgridE MapgridN CurrentUrbanRural AlarmMethodCode IncidentType
64734  2587836  5966137                 1              11         1301
64900  2588460  5962670                 1              11         1312
66143  2589938  5963648                 1              11         1502
73062  2589623  5965200                 1              11         1102
75205  2588260  5966055                 1              11         1201
      Heatsource Obj1ignited day  time dayweek  type
64734         61         721 287 13.37       6 Other
64900         NA         811 672 18.38       6 Other
66143         NA          NA 595  0.75       6  Susp
73062         45         116 578 19.30       3 Other
75205         NA         912 650 21.28       5 Other
```

C.2 Pattern discovery and testing

```
> rpart(type ~ ., blenheim1, cp=0.05, method="class")

n= 318

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 318 80 Other (0.74843 0.25157)
   2) Heatsource< 64.5 123  7 Other (0.94309 0.05691) *
   3) Heatsource>=64.5 195 73 Other (0.62564 0.37436)
     6) Heatsource>=70.5 38  0 Other (1.00000 0.00000) *
     7) Heatsource< 70.5 157 73 Other (0.53503 0.46497)
      14) IncidentType>=1356 94 29 Other (0.69149 0.30851)
        28) IncidentType< 1502 44  3 Other (0.93182 0.06818) *
        29) IncidentType>=1502 50 24 Susp (0.48000 0.52000)
          58) day>=277.5 30  9 Other (0.70000 0.30000) *
          59) day< 277.5 20  3 Susp (0.15000 0.85000) *
      15) IncidentType< 1356 63 19 Susp (0.30159 0.69841)
        30) MapgridN>=5.969e+06 12  3 Other (0.75000 0.25000) *
        31) MapgridN< 5.969e+06 51 10 Susp (0.19608 0.80392) *

> rpart(type ~ ., blenheim2, cp=0.05, method="class")

n= 386

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 386 91 Other (0.76425 0.23575)
   2) Heatsource< 66.5 101  4 Other (0.96040 0.03960) *
   3) Heatsource>=66.5 285 87 Other (0.69474 0.30526)
     6) day>=371.5 146 17 Other (0.88356 0.11644) *
     7) day< 371.5 139 69 Susp (0.49640 0.50360)
      14) time>=7.15 108 41 Other (0.62037 0.37963)
        28) Heatsource>=75 9  0 Other (1.00000 0.00000) *
        29) Heatsource< 75 99 41 Other (0.58586 0.41414)
          58) IncidentType>=1406 55 11 Other (0.80000 0.20000) *
          59) IncidentType< 1406 44 14 Susp (0.31818 0.68182) *
      15) time< 7.15 31  2 Susp (0.06452 0.93548) *
```

C.3 Direct change detection

The following is the R output for the Poisson regression tree built for detecting directly the differences between two data sets. Each terminal node is marked by an asterisk and has five additional values printed. The pair in the first parenthesis are the frequencies of the events (`Other` and `Susp` here) from the first data set, while the pair in the second parenthesis are those from the second data set. The last value is the p-value of the likelihood ratio test for heterogeneity between the rates of the fire events in the two data sets under the circumstance specified by the splitting criteria along the path.

```
> dtd(blenheim1, blenheim2)

IncidentType <= 1200: (90 14) (66 4) 0.00827 *
IncidentType >  1200:
|  day <= 384:
|  |  day <= 284:
|  |  |  IncidentType <= 1502:
|  |  |  |  day <= 60.5: (10 9) (5 2) 0.0384 *
|  |  |  |  day >  60.5:
|  |  |  |  |  IncidentType <= 1312: (11 5) (3 14) 0.00957 *
|  |  |  |  |  IncidentType >  1312: (17 9) (21 11) 0.733 *
|  |  |  IncidentType >  1502: (1 7) (12 3) 0.00182 *
|  |  day >  284:
|  |  |  MapgridE <= 2592500:
|  |  |  |  MapgridN <= 5966984:
|  |  |  |  |  MapgridN <= 5966004:
|  |  |  |  |  |  IncidentType <= 1500: (4 0) (18 28) 3.01e-11 *
|  |  |  |  |  |  IncidentType >  1500: (4 0) (9 6) 0.00582 *
|  |  |  |  |  MapgridN >  5966004: (4 0) (6 4) 0.0511 *
|  |  |  |  MapgridN >  5966984: (4 0) (6 3) 0.102 *
|  |  |  MapgridE >  2592500: (6 0) (4 0) 0.818 *
|  day >  384:
|  |  day <= 586:
|  |  |  day <= 434: (22 9) (23 1) 0.0249 *
|  |  |  day >  434: (23 14) (63 1) 7.59e-08 *
|  |  day >  586: (42 13) (59 14) 0.233 *
```

D Likelihood ratio test

Let x_1 be the number of events in the first data set and x_2 be that in the second data set with the same exposure. Assume that x_i, $i = 1, 2$, has the Poisson distribution with rate λ_i, with density $f(x_i; \lambda_i)$. For testing homogeneity

$$H_0 : \lambda_1 = \lambda_2,$$

the likelihood ratio test statistic is given by

$$W = 2\left\{\log f(x_1; x_1) + \log f(x_2; x_2) - \log f(x_1; \frac{x_1 + x_2}{2}) - \log f(x_2; \frac{x_1 + x_2}{2})\right\}.$$

The statistic W has approximately χ^2_1, the chi-square distribution with 1 degree of freedom.

If there are k types of events in both data sets, then the likelihood ratio test statistic W for testing homogeneity is the sum of the individual ones. It thus has approximately χ^2_k, the chi-square distribution with k degrees of freedom. For example, the most significant rule produced by the Poisson regression tree that detects differences directly, as given in Appendix C.3, has $(4 \;\; 0)$ for the frequencies of other and suspicious fires in the first data set and $(18 \;\; 28)$ for those in the second data set. The test statistic value is

$$\begin{aligned} W &= 2\left\{\log f(4; 4) + \log f(18; 18) - \log f(4; 11) - \log f(18; 11)\right\} + \\ &\quad 2\left\{\log f(0; 0) + \log f(28; 28) - \log f(0; 14) - \log f(28; 14)\right\} \\ &\approx 48.45, \end{aligned}$$

which has χ^2_2 and hence gives the p-value 3.01×10^{-11}.

References

[1] Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalised linear models. *Statistics and Computing*, 6, 251–262.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716–723.

[3] Akakpo, N. (2008). Detecting change-points in a discrete distribution via model selection. *Electronic Journal of Statistics*, 2, (in press).

[4] Akman, V. E. and Raftery, A. E. (1986) Asymptotic inference for a change-point Poisson process. *Annals of Statistics* 14, 1583 – 1590.

[5] Bai, J. (1994) Least squares estimation of a shift in linear processes. *Journal of Time Series Analysis* 15, 453 – 472.

[6] Bai, J. (1997). Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, 79, 551–856.

[7] Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Applications*. Englewood Cliffs, Prentice Hall.

[8] Bayraktar, E. and Dayanik, S. (2006). Poisson disorder problem with exponential penalty for delay. *Mathematics of Operations Research* 31, 217 - -233.

[9] Beibel, M. (1997) Sequential change-point detection in continuous time when the postchange drift is unknown *Bernoulli* 3, 457 – 478.

[10] Beibel, M. and Lerche, H. R. (2003) Sequential Bayes detection of trend changes, Foundations of statistical inference (Shoresh, 2000), Contrib. Statist., Physica, Heidelberg, pp. 117???130.

[11] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont CA.

[12] Brillinger, D. (2003) Three environmental probabilistic risk problems. *Statistical Science* 18, 412-421

[13] Carlin, B. P., Gelfand, A. E., Smith, A. F. M. (1992) Hierarchical Bayesian analysis of changepoint problems. *Journal of the Royal Statistical Society. Series C* 41, 389 – 405.

[14] Carlstein, E. (1988). Nonparametric change-point estimation. *The Annals of Statistics*, 16, 188–197.

[15] Chiu, G., Lockhart, R. and Routledge, R. (2006). Bent-cable regression theory and applications. *Journal of the American Statistical Association*, 101, 542–553.

[16] Crowder, S. V. (1987). A simple method for studying run-length distributions of exponentially weighted moving average charts. *Technometrics*, 29, 401–407.

[17] Davis, M. H. A. (1976) A note on the Poisson disorder problem *Banach Center Publications* 1, 65 – 72.

[18] Dayanik, S. and Goulding, C. (2007) Detection and identification of an unobservable change in the distribution of a Markov-modulated random sequence. Submitted.

[19] Dobson, A. J. (1990). *An Introduction to Generalized Linear Models*. London: Chapman and Hall.

[20] Felici, G. and Vercellis, C. (2007). *Mathematical Methods for Knowledge Discovery and Data Mining*. Idea Group Reference.

[21] Frisen, M. (2003). Statistical surveillance. Optimality and methods. *International Statistical Review*, 71, 403–434.

[22] Gal'chuk, L. I. and Rozovskii, B. L. (1971) The disorder problem for a Poisson process, *Theory of Probability and its Applications* 16, 712 – 716.

[23] Galeano, P. (2007) The use of cumulative sums for detection of changepoints in the rate parameter of a Poisson Process. *Computational Statistics and Data Analysis* 51, 6151 – 6165.

[24] Gan, F. F. (1994) Design of optimal exponential CUSUM control charts. *Journal of Quality Technology* 26, 109 – 124.

[25] Guralnik, V. and Srivastava, J. (1999). Event detection from time series data. *Proc. ACM-SIGKDD Intational Conference on Knowledge Discovery and Data Mining*, 33–42.

[26] Hastie, T. and Tibshirani, R. (1999). *Generalized Additive Models* (2nd Ed.). Chapman & Hall/CRC.

[27] Hastie, T. J., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

[28] Hawkins, D.M. (2001) Fitting multiple change-point models to data. *Computational Statistics and Data Analysis* 37, 323 – 341.

[29] Hawkins, D. M., Qiu, P. H. and Kang, C. W. (2003). The changepoint model for statistical process control. *Journal of Quality Technology*, 35, 355–366.

[30] Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributions assumptions in econometric models for duration data. *Econometrica*, 52, 271–320.

[31] Hinkley, D. V. (1970) Inference about the change-point in a sequence of random variables. *Biometrika* 57, 1 – 17.

[32] Ide, T. and Tsuda, K. (2007). Change-point detection using Krylov subspace learning. In *Proceedings of the 2007 SIAM International Conference on Data Mining. Minneapolis*, Minnesota: SIAM.

[33] Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299–314.

[34] Jones, L. A. (2002). The statistical design of EWMA control charts with estimated parameters. *Journal of Quality Technology*, 34, 277–288.

[35] Kenett, R. S. and Pollak, M. (1996) Data-analytic aspects of the Shiryayev-Roberts control chart: Surveillance of a non-homogeneous Poisson process. *Journal of Applied Statistics* 23, 125 – 137.

[36] Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society*, Series B, 613–658.

[37] Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing spline. *Journal of the Royal Statistical Society*, Series B, 61, 381–400.

[38] Lindsey, J. K. (1997). *Applying Generalized Linear Models*. New York: Springer.

[39] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd Ed.). Chapman & Hall.

[40] McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92, 162–170.

[41] Mason, R. L., Champ, C. W., Tracy, N. D., Wierda, S. J. and Young, J. C. (1997). Assessment of multivariate process control techniques. *Journal of Quality Technology*, 29, 140–143.

[42] Molnau, W. E., Runger, G. C., Montgomery, D. C., Skinner, K. R., Loredo, E. N. and Prabhu, S. S. (2001). A program of ARL calculation for multivariate EWMA charts. *Journal of Quality Technology*, 33, 515–521.

[43] Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41, 100–114.

[44] Peng, R. D., Schoenberg, F. P., Woods, J. (2005) A space-time conditional intensity model for evaluating a wildfire hazard index. *Journal of the American Statistical Association* 100, 26-35.

[45] Peskir, G. and Shiryaev, A. N. (2002). Solving the Poisson disorder problem. *Advances in Finance and Stochastics* Springer, Berlin, pp. 295???312.

[46] Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects Models in S and S-Plus*. Springer.

[47] Polansky, A. M. (2007) Detecting change-points in Markov chains *Computational Statistics and Data Analysis* 51, 6013 – 6026.

[48] Preisler, H. K., Brillinger, D. R., Burgan, R. E. and Benoit, J. W. (2004) Probability based models for estimation of wildfire risk. *International Journal of Wildland Fire* 13, 133-142.

[49] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Mateo, Calif.: Morgan Kaufmann Publishers.

[50] R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

[51] Raftery, A. E. and Akman, V. E. (1986) Bayesian analysis of a Poisson Process with a change-point. *Biometrika* 73, 85 – 89.

[52] Reed, W. J. (1998) Determining changes in historical forest fire frequency from a time-since-fire map. *Journal of Agricultural, Biological, and Environmental Statistics* 3, 430 – 450

[53] Reed W. J. (2000) Reconstructing the history of forest fire frequency: identifying hazard rate change points using the Bayes information criterion. *Canadian Journal of Statistics - Revue Canadienne de Statistique* 28, 353 – 365.

[54] Reynolds, M. R. and Stoumbos, Z. (2004). Control charts and the efficient allocation of sampling resources. *Technometrics*, 46, 200–214.

[55] Roberts, S. W. (1966). A comparison of some control chart procedures. *Technometrics*, 8, 411–430.

[56] Roddick, J. F. and Hornsby, K. (Eds.). (2001). *Temporal, Spatial, and Spatio-Temporal Data Mining*. Springer.

[57] Ryden, J. and Rychlik, I. (2006) A note on estimation of intensities of fire ignitions with incomplete data. *Fire Safety Journal* 41, 399-405.

[58] Sebastiani, P. and Mandl, K. (2004) Biosurveillance and Outbreak Detection. In Data Mining: Next Generation Challenges and Future Directions, H. Kargupta A. Joshi K.Sivakumar and Y. Yesha (Eds). MIT Press. 185 – 198.

[59] Shewhart, W. A. (1931). *Economic Control of Manufactured Products.* Van Nostrand Reinhold, New York.

[60] Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability and its Applications*, 8, 22–46.

[61] Siegmund, D. (1988) Confidence sets in change points problems. *International Statistical Review* 56, 31 – 48.

[62] Takeuchi, J. and Yamanishi, K. (2006). A unifying framework for detecting outliers and change points from time series. *IEEE Transactions on Knowledge and Data Engineering*, 18, 482 – 492.

[63] Tsodikov, A. (2003). Semiparametric models: a generalized self-consistency approach. *Journal of the Royal Statistical Society*, Series B, 65, 759–774.

[64] Vapnik, V. (1998). *Statistical Learning Theory*. Wiley-Interscience, New York.

[65] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations* (2nd Ed.). Morgan Kaufmann, San Francisco.

[66] Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society*, Series B, 70, 495–518.

[67] Worsley, K. J. (1986) Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika* 73, 91 – 104.

[68] Zeira, G., Maimon, O., Last, M. and Rokach, L. (2004). Change detection in classification models induced from time series data. In M. Last, A. Kandel and H. Bunke (eds.), *Data Mining in Time Series Databases*, 101–125. World Scientific.

[69] Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society*, Series B, 69, 507–564.