PEOPLE-CAUSED FOREST FIRE PREDICTION USING POISSON AND LOGISTIC REGRESSION

by

Melanie Poulin-Costello
B.Math., University of Waterloo, 1990

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

We accept this thesis as conforming
to the required standard

[redacted]

Dr. W. J. Reed, Supervisor (Department of Mathematics/Statistics)

[redacted]

Dr. B. R. Johnson, Departmental Member (Department of Mathematics/Statistics)

[redacted]

Dr. M. Lesperance, Departmental Member (Department of Mathematics/Statistics)

[redacted]

Dr. R.E. Odeh, Additional Member (Department of Mathematics/Statistics)

[redacted]

W. Bergerud, External Examiner (Ministry of Forests)

Supervisor: Dr. Bill Reed

ABSTRACT

The objective of this thesis is to develop a statistical predictive model of people-caused forest fire occurrence that could be implemented in the province of British Columbia to assist fire managers in fire-fighting preparedness. Forest fires occur at an average rate of 2820 per fire season in B.C. Of these, fifty-five percent are classified as *people-caused forest fires.* The fire suppression program of B.C. is managed and run by the Ministry of Forests' Protection Branch.

People-caused forest fire occurrence depends on weather conditions, and people's use of the forests. The weather conditions are measured not only by *rain*, *temperature*, *wind speed*, and *relative humidity*, but also by six fire weather indices:

FFMC: Fine Fuel Moisture Code,

DC: Drought Code,

DMC: Duff Moisture Code,

BUI: Build Up Index,

ISI: Initial Spread Index, and

FWI: Fire Weather Index.

The FFMC, DC, BUI and wind speed are found to be good predictors of people-caused forest fires.

An indication of people's use of the forests includes locations of campsites, populated areas, logging and industrial operations. These locations are not yet available in electronic form; populated areas such as lakes, parks, cities, and highways, were determined from a map. However, campsite and logging road and camp locations were not available at all. Surrogate indicators of people's use of the forest such as day of the week, day of the fire season, and cause of fire were investigated but not successful in forest fire prediction. Indicator variables for whether or not a lake, park, city or road are present in the location of prediction were included in the final predictive models.

Generalized linear regression methods were used to develop statistical models for the forest fire data. Specifically, Poisson and logistic regression

were used. Poisson regression, with mean λ_i, assumes there is a multiplicative effect on the covariates,$\mathbf{x}_i^T$, and the regression equation is

$$\eta_i = \ln \lambda_i = \mathbf{x}_i^T \boldsymbol{\beta} .$$

The $\boldsymbol{\beta}$'s are estimated by weighted least squares using the interative generalized linear regression software GLIM. The η_i is called the linear predictor. Logistic regression, with parameter p_i, applies a logistic transformation to the response so that the regression equation is

$$\eta_i = \ln \left\{ \frac{p_i}{1 - p_i} \right\} = \mathbf{x}_i^T \boldsymbol{\beta} .$$

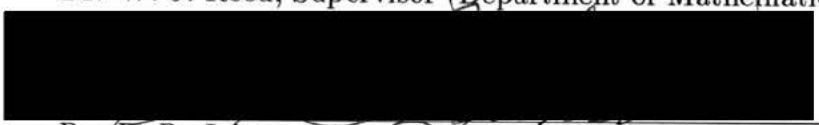
The $\boldsymbol{\beta}$'s are again estimated by GLIM. The "best" fitted models were selected by backwards elimination.

Forest fires are rare events, sometimes making the regression analysis difficult. In an effort to have forest fire data with fewer zeroes, the data were aggregated. The Ministry of Forests has divided the province into fifteen by fifteen kilometer cells. In this thesis, these cells were grouped into *zones* according to *major climate type* and *fuel type*. Three models were fitted to one particular zone: a Poisson model fitted to fire weather data averaged over the zone; a Poisson model fitted to all data in a zone without averaging; and a logistic model fitted to the unaveraged data common to the whole zone. One of the aims of this thesis is to investigate the merits of fitting a Poisson model to the data aggregated across the zone, over fitting a logistic model to unaggregated data.

The predictive ability of the three models developed here is compared to several established forest fire models. Firstly, they are compared to the current prediction model used by the Protection Branch, and developed by Peter Kourtz. As well, the models were compared to two logistic models, developed by Dr. David Martell of the University of Toronto, and applied to both the aggregated and common British Columbia zonal data. All models developed in this thesis provide better predictions than those provided by Kourtz's model. Martell's models do fit the B.C. data better than the models from this thesis in damper locations, with fewer fires. The Poisson model fitted to the unaveraged common data within a zone provides the best predictions of the three models developed here.

Examiners:

Dr. W. J. Reed, Supervisor (Department of Mathematics/Statistics)



Dr. B. R. Johnson, Departmental Member (Department of Mathematics/Statistics)

Dr. M. Lesperance, Departmental Member (Department of Mathematics/Statistics)

Dr. R. E. Odeh, Additional Member (Department of Mathematics/Statistics)

W. Bergerud, External Examiner (Ministry of Forests)

ACKNOWLEDGEMENT

I would like to thank Dr. Bill Reed, my supervisor, for his guidance throughout the research and writing of this thesis. I would like to thank Dr. Mary Lesperance, also, for her suggestions and helpfulness during the research for this thesis. I also thank the members of my commitee, Dr. Bruce Johnson and Dr. Bob Odeh from Mathematics and Statistics, and Wendy Bergerud from the Ministry of Forests. I thank the Protection Branch at the Ministry of Forests for supplying data, computer access, funding, and motivation for this thesis. In particular, from Ministry of Forests, I thank Peter Fuglem, Phil Symington, and Val Fletcher for their time and patience with my endless questions. I also thank Kevin Haukaas without whose programming efforts the data for this thesis would not have been made available. I also thank Widijanto Nugroho for his ever present assistance with LaTeX. Lastly, I thank my husband Peter for his constant patience and encouragement.

I dedicate this thesis to the two men in my life who made a difference: my father, Frank Poulin and my friend, Peter Costello.

Contents

List of Figures

List of Tables

Part I

Introduction, Definitions, and Preliminary Analysis

Chapter 1
Introduction

The forest fire suppression program costs the province of British Columbia on average 67.2 million dollars annually to fight an average of 2820 fires per year. The costs increase each year. This program protects over 290 billion dollars worth of standing timber and property. British Columbia experiences 31 percent of the total forest fires in Canada, but because of the B.C¿ Ministry of Forests' efforts in suppression, B.C. suffers only 7 percent of the total area burned in Canada. In B.C., without forest fire suppression, the timber volume available to support forest and related industries would be 25 to 50 percent less than the current average of 75 million cubic meters annually, costing the local economy 3 billion dollars, annually. Thousands of jobs would also be affected by forest fire destruction. Clearly, the forest fire suppression program is necessary but expensive.

Currently, B.C. has the objective of controlling fires by 10 a.m. of the morning following the reporting of the fire. B.C. has a 92% success rate in meeting this objective, allowing only an average of 216 fires to escape per year. However, the direct cost of each escaped fire is approximately 113,000 dollars. For each percentage increase in their objective, the B.C. government would save 3.2 million dollars.

Forest fires occur throughout the year; however, the *fire season* for which systematic records are kept extends from April fifteenth to October

fifteenth. Of all forest fires in B.C., 55 percent are caused by people; the remainder are lightning-caused. If the time and location of people-caused forest fires could be predicted with some accuracy, then fire fighters would be better prepared and, therefore more cost-effective. The above facts, costs and quotes are from personal communication with the Protection Branch of the B.C. Ministry of Forests and from [25].

The objective of this thesis is to develop a statistical predictive model for the probability of people-caused forest fires in a small geographic location of the Kamloops Forest Region, British Columbia. Such predictions, for small areas, should assist fire managers to better prepare for a day's activities of fire suppression. The model is to subsequently be tested on further small locations within the Kamloops Region. This thesis will examine the benfits of using Poisson regression over logistic regression and of using Poisson regression over models fitted by previous researchers Martell and Kourtz.

Dr. D.L. Martell, of the University of Toronto's Faculty of Forestry, and co-workers have conducted considerable research on predicting people-caused forest fires in Northern Ontario. Cunningham and Martell [9] began their research into prediction by examining ten years of data from Sioux Lookout, Ontario. They showed that a Poisson process is acceptable for modelling the probability of a people-caused forest fire on a given day.

Martell, Otukol, and Stocks in [18] use the Poisson distribution to predict the probability of people-caused forest fires. They use a three step process to obtain predictions of the probability of 0, 1, 2,... fires in a day. Their first step was to estimate the probability of one or more fires occur-

ring on a given day, p_{est}, using *logistic regression*. They assumed that p_{est} is a logistic function of the Fine Fuel Moisture Code (FFMC), a numerical rating of the moisture content of the litter and fine fuels in a forest. Under the assumption that one or more fires follows the Poisson distribution, the probability of one or more fires is

$$P(X \geq 1) = 1 - e^{-\lambda} .$$

The second step of the process was to estimate λ by rearranging the above equation and substituting in p_{est} for the probability of one or more fires:

$$\lambda_{est} = -\ln(1 - p_{est}) .$$

The final step to obtain the probabilities was to substitute λ_{est} for λ into the Poisson probability mass function,

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} , x = 0, 1, 2, \ldots .$$

Martell did not use Poisson regression directly because of a lack of software for Poisson regression at the time he developed his models.

Martell developed a separate logistic model for different seasons, general fire causes and administrative regions (as defined by the Ontario Ministry of Natural Resources). He concluded that the logistic model was suitable for predicting the probability of a fire in the administrative districts of northern Ontario. He has not to date developed a successful prediction model for many fires on a given day. He also concluded that a Poisson model is acceptable for making decisions about forest fire probability and that the Fine Fuel Moisture Code (FFMC) has predictive value.

Martell, Bevilacqua and Stocks [17] improved upon Martell's earlier work by incorporating Fourier terms into the logistic regression model to account for seasonal variation, instead of dividing the data by season. They found the logistic regression model predicted better with the periodic terms included. The two models of Martell's from [17] used for comparison in this thesis are (1) a logistic model with covariates *FFMC* and *BUI* (Build Up Index); and (2) this model with the following Fourier series terms added;

$$\sum_{k=1}^{4} \{\cos(2k\pi\, day/140) + \sin(2k\pi\, day/140)\} \times \text{FFMC}$$

where *day* is the day of the fire season, and the length of the fire season is 140[1] days.

Todd and Kourtz [27] also model forest fires using the Poisson distribution. They assume that the Poisson parameter λ is a random variable with a gamma distribution. Like Martell, they also divided the fire season into subseasons. They found weather indices FFMC and Duff Moisture Code (DMC) and wind speed to be good predictors of fire occurrence. They used Bayesian estimation to estimate the gamma parameters. This model is currently being used by the Protection Branch of the British Columbia Ministry of Forests and is apparently somewhat successful in Québec. This model does not predict many fires per day well and indeed, as implemented in B.C., does not predict fire probabilities, only the number of fires on a given day and in a given region. Most of these single fire predictions are, not surprisingly, zero.

Haines, Main, Frost and Simard [14] evaluate the ability to predict

[1]This denominator is adjusted to 186 to account for the length of the B.C. fire season when this model is fitted to B.C. data.

forest fires of the National Fire-Danger Rating System (NFDRS) of the United States, the Fosberg Fire Weather Index (FFWI) and the Canadian Fire Weather Indices FFMC and ISI (Initial Spread Index). In their conclusions the FFMC and ISI were included in their list of best predictors of fire occurrence. They, too, found difficulties predicting the probability of many fires per day.

1.1 Definition of Weather Terms

Weather affects the probability of a fire. The term *weather* here is a general statement for rain, temperature, relative humidity, wind speed, wind direction and six fire weather indices from the Canadian Forest Fire Weather Index System (CFFWIS) developed by Van Wagner in [30] and calculated from solar noon weather observations. See Figure 1.1 for the structure and relationship of these indices as given in [29]. See [5], [20], and [29] for a description of the calculation of these indices.

Three of these six indices, *FFMC*, *DMC* and *DC*, are moisture codes and follow daily changes in moisture content of three respective classes of forest fuels. DC (Drought Code) is a measure of the moisture content of the deep compact organic layers of the forest floor (10–20 cm depth); DMC (Duff Moisture Code) is a measure of the moisture content of the loosely compacted organic layer (5–10 cm depth); and FFMC (Fine Fuel Moisture Code) is a measure of the moisture content of surface litter and other fine fuels (needles, leaves). FFMC is used as a measure of the ease of ignition of the forest.

The three other indices, *ISI*, *BUI* and *FWI*, are fire behaviour indices.

ISI (Initial Spread Index) represents the relative fire spread expected after ignition; BUI (Build Up Index) represents the total amount of fuel available for combustion; and FWI (Fire Weather Index) represents the potential fire intensity. FWI is a summary index of all the components of the CFFWIS.

The FWI and BUI together with topographic and weather factors are used to determine the *Danger Class* rating scale: I–V, where I denotes difficult ignition conditions and V very easy ignition conditions. The Danger Class ratings are useful in heightening public awareness to forest fire potential but are not useful in forest fire prediction. It is the Danger class that determines the reading on the roadside fire danger rating signs (green corresponding to danger class I; red to danger class V).

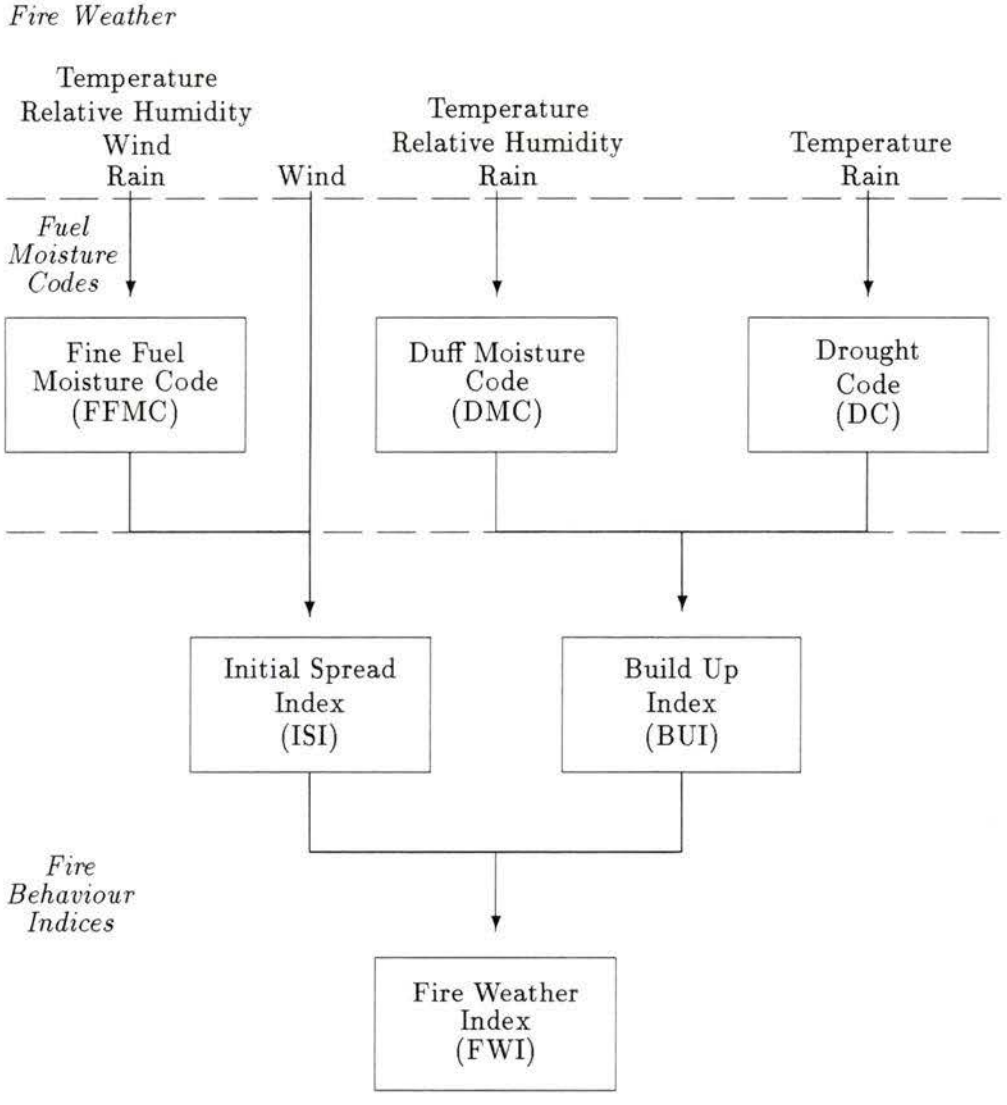


Figure 1.1: Canadian Forest Fire Weather Index System Structure

1.2 The Statistical Theory

In this thesis, *generalized linear models* are fitted to the data. Two instances of generalized linear models, with Poisson and binary response variables are used throughout this thesis. In particular for binary data, the *logistic* model is used.

The Poisson distribution is the distribution of the number of events in a finite time interval when events occur independently, with a homogeneous hazard rate. The Poisson distribution can be used to model counted data when there is conceptually no upper limit to the frequency of occurrence. People-caused forest fires in similar locations on a given day of the fire season can be assumed to occur independently and with a constant hazard rate for that day. Hence, it is reasonable to assume the number Y_{ij} of fires in location i on day j is Poisson distributed with hazard rate λ_{ij}. The probability mass function (p.m.f.) of Y_{ij} is

$$P(Y_{ij} = y) = \frac{e^{-\lambda_{ij}} \lambda_{ij}^{y}}{y!} \quad \lambda_{ij} > 0; \quad y = 0, 1, \ldots . \tag{1.1}$$

The hazard rate λ_{ij} is assumed to depend on weather, moisture codes and other explanatory variables giving rise to a *Poisson regression* model (a specific instance of *generalized linear regression*). From [19], the assumption for Poisson regression is that $V(Y) = \sigma^2 E(Y)$ for a random variable Y and dispersion parameter σ^2 which is assumed constant over the data. For Poisson regression, the further assumption is made that the dispersion parameter σ^2 is 1 and the mean and variance of the Poisson random variable Y are equal. For $\sigma^2 > 1$, there is *overdispersion* in the Poisson model.

In logistic regression the response data are assumed to be binary taking values 1 or 0 according as to whether or not any people-caused forest fires occur. Thus, $Y_{ij} = 1$ when one or more fires occur and $Y_{ij} = 0$ otherwise. When the data comprise only zeroes and ones, a Bernoulli distribution[2] for the response can be used in place of the Poisson distribution. The Bernoulli probability mass function is

$$P(Y_{ij} = y) = p_{ij}^{y}(1 - p_{ij})^{1-y} \text{ for } y = 0, 1 . \tag{1.2}$$

Thus, $P(Y_{ij} = 1) = p_{ij}$ and $P(Y_{ij} = 0) = 1 - p_{ij}$. In this model p_{ij} is the probability of a fire in location i on day j; it is assumed to depend on the values of weather, moisture codes and other explanatory variables.

1.3 The Data

Most forests in British Columbia are under the jurisdiction of the B.C. Ministry of Forests. The Protection Branch of this ministry is responsible for forest fire detection and suppression in B.C. For administrative purposes the Ministry of Forests has divided the province into six forest regions all of which are further divided into forest districts. The Protection Branch has divided the province into cells that are fifteen by fifteen kilometers square. The *Advanced Fire Management* (AFM) computer software manages a database of information regarding lightning-caused and people-caused forest fires and weather at the cell level for the six forest regions. This thesis arose from the desire of the Protection Branch to examine historical fire and weather data over a small area, possibly by cell or

[2]The Bernoulli distribution is a binomial distribution with parameters $n = 1$ and p *i.e.* a single trial with binary outcomes, $P(1) = p$, $P(0) = 1 - p$.

over a group of cells. To narrow the scope of the analysis to a manageable amount of data, only data from the Kamloops Forest Region was analysed.

The Kamloops Forest Region is located in the southern interior of B.C. with the city of Kamloops at its centre and the Canada-U.S.A. border at its southern limit. The Kamloops Region also contains some of the dryest areas in all of Canada, specifically parts of the Okanagan Valley. The Okanagan and the Shuswap Valleys (also in the Kamloops Region) are fairly heavily populated, have some industry and are popular tourist locales. The dry conditions and relatively high population make the Kamloops Region a good candidate for the study of people-caused forest fires.

Protection Branch supplied weather and fire data for 284 cells in the Kamloops Region giving cell row (east-west) and column (north-south) coordinates. The Kamloops Region has 289 cells in total. The five cells for which no data were supplied are at the westernmost tip of the region and are contiguous and so were not included in any analysis.

The weather is measured at solar noon each day and weather indices are then calculated for that day. See [20] and [29] for explanations of collection of weather data and weather indices calculations. The weather observations are made from a weather station in the nearby vicinity. Some cells in the Kamloops Region have the same weather station so will have identical weather data for a given day. It is assumed that the weather is measured accurately and recorded without error.

One ASCII file of data was supplied for each fire season from 1985 to 1991. The ASCII files had one record of weather measurements and number of fires per day, per cell in the Kamloops Region. Data were also available

from Datatrieve[3] which has weather information only on days that a fire occurred and is organized by latitude and longitude not by cells. The data provided from the AFM source did not agree with data from Datatrieve in terms of the dates and numbers of fires; hence, Datatrieve was ignored for model fitting. However, there was a striking discrepancy reported by the two systems for the number of fires in the Kamloops Forest Region for 1989: from Datatrieve 276 fires were reported but the AFM database reported an unusually low ten fires! The parameter estimates from models fitted without 1989 data did not differ greatly from the parameter estimates fitted with 1989 data included. Hence, the final models reported in this thesis were fitted without the 1989 data. Otherwise, it is assumed that the data from the AFM database are correct.

As the AFM database is relatively new, data in cell format is not available prior to the 1985 fire season. The lack of data for earlier years is not a major concern since patterns in forest fires especially people-caused fires are changing. See [25]

1.3.1 Aggregating the Data

People-caused forest fires are rare events; even so, the aim of this thesis is to predict people-caused forest fires on a daily basis within a small area. Predictions over too large of an area are not helpful to the suppression program. Predictions over too small of an area are difficult to obtain because forest fires are rare events. Within a cell, most days have zero

[3]Datatrieve is a VAX/VMS database management system with its own sequential query language.

fires, even in cells that could be described as "high fire risk" cells. This abundance of zeroes leaves little information for analysis. Hence, fitting a regression model for an individual cell may not be meaningful. The model fitting process over a small area would benefit from fewer zeroes and days on which several fires occurred. The paradox is, however, that in reality it is desirable to have as few fires as possible. One approach to this problem is to aggregate the data across some cells. This decreases the number of zero fire days, and also increases the number of multiple fire days.

Aggregating weather and moisture code data across cells with minimal information loss requires aggregation across cells which are relatively homogeneous with respect to weather, moisture codes, and fuel types. Rather than the overwhelming task of examining 284 cells for similarities on the various weather variables, the cells were examined by climate and fuel type. The Ministry of Forests has classified B.C. into *major climate types* which classify geographic regions according to weather patterns and topographic characteristics. A map of the Kamloops Region displaying the major climate types was overlayed with a grid to mark by hand the locations of cells throughout the region. The climate type which takes up the majority of a cell by area was determined and is called the *primary major climate type* of a cell. The climate types per cell need to be accurately digitalized by a geographer and added to the AFM database.

Cells with the same primary major climate type were grouped together into relatively large groups of cells. It is also necessary to incorporate fuel type into forest fire models, so these groups of cells based on climate were broken down according to fuel type. *Primary fuel type* of a cell is the type

of forest that occurs in the largest percentage area within a cell. Primary fuel type of a cell is available from the AFM database. Hence, cells are grouped into *zones* of contiguous cells with the same primary major climate type and primary fuel type. The Kamloops Region, then, was divided into forty-five zones as depicted in Figure 1.2.

The FFMC appears to be homogeneous within a zone. Figure 1.3 shows FFMC plotted by day for a cell in zone 63-C7. Graphs for all cells in zone 63-C7 of the FFMC by day for each cell showed little difference from Figure 1.3 by eye.

1.3.2 The Data used in Model Fitting

Zone 63-C7 (southern-very dry — ponderosa pine and douglas fir) was selected for analysis. Other zones were selected to test the portability of the models developed in 63-C7. The basic idea for analysis is to fit generalized linear regression models, using the number of fires as the response variable and the weather, moisture codes and other influential variables as "covariates" or explanatory variables. Two methods of formatting the data within a zone and the analysis used for each are explained here. The first is referred to as *Method A* or the *Aggregated Method* which has weather data averaged, and people-caused forest fires totalled across cells in the zone. Each day yields one data point for *Method A*. The Poisson generalized linear model is applied to the *aggregated* data to result in *Method AP* the *aggregated Poisson model.* The two logistic models developed by Martell were also fitted to the *aggregated* data and compared to the Poisson model. Martell's models are referred to by letters M for the model with

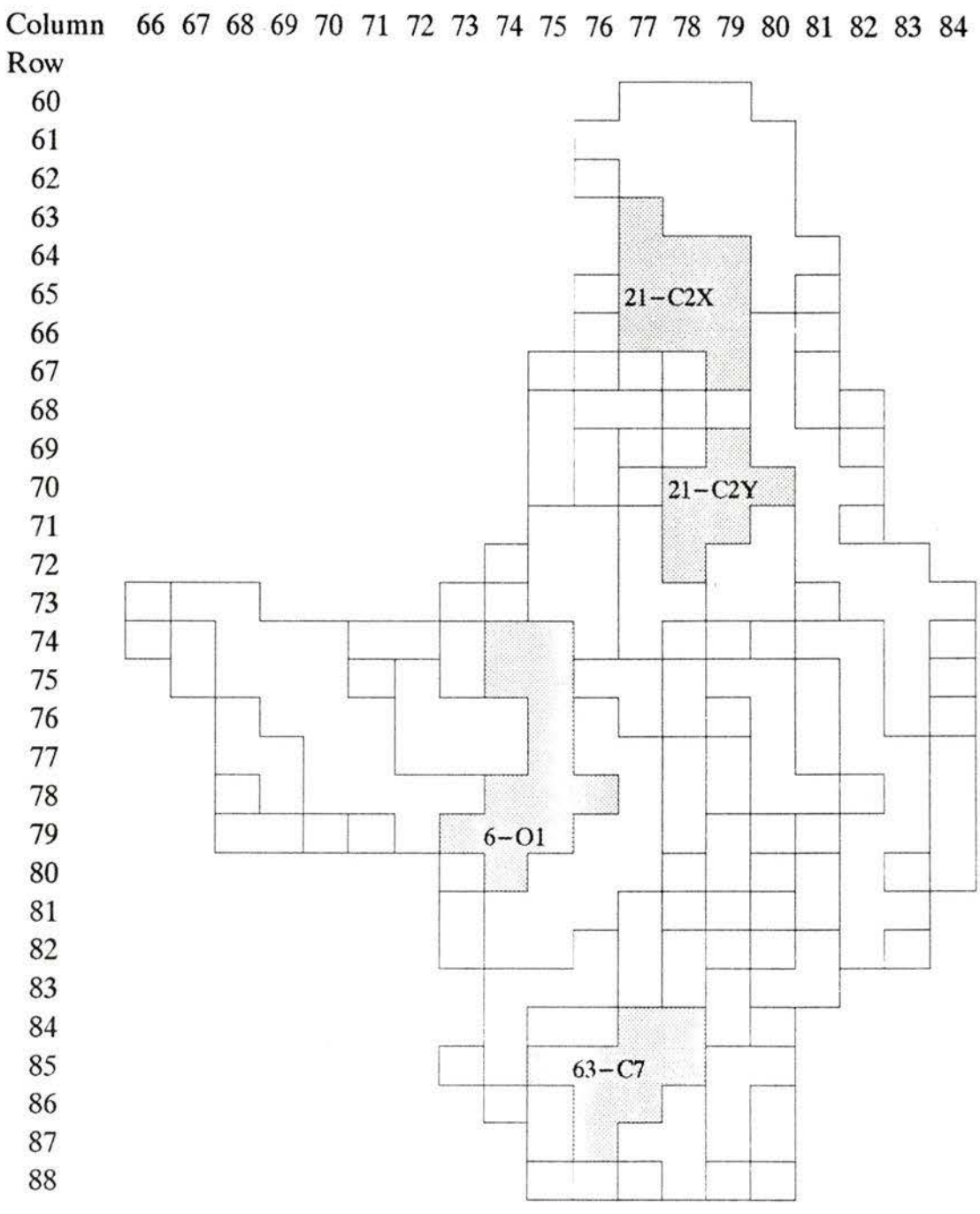


Figure 1.2: Zones in the Kamloops Forest Region as defined by Climate and Fuel Type.

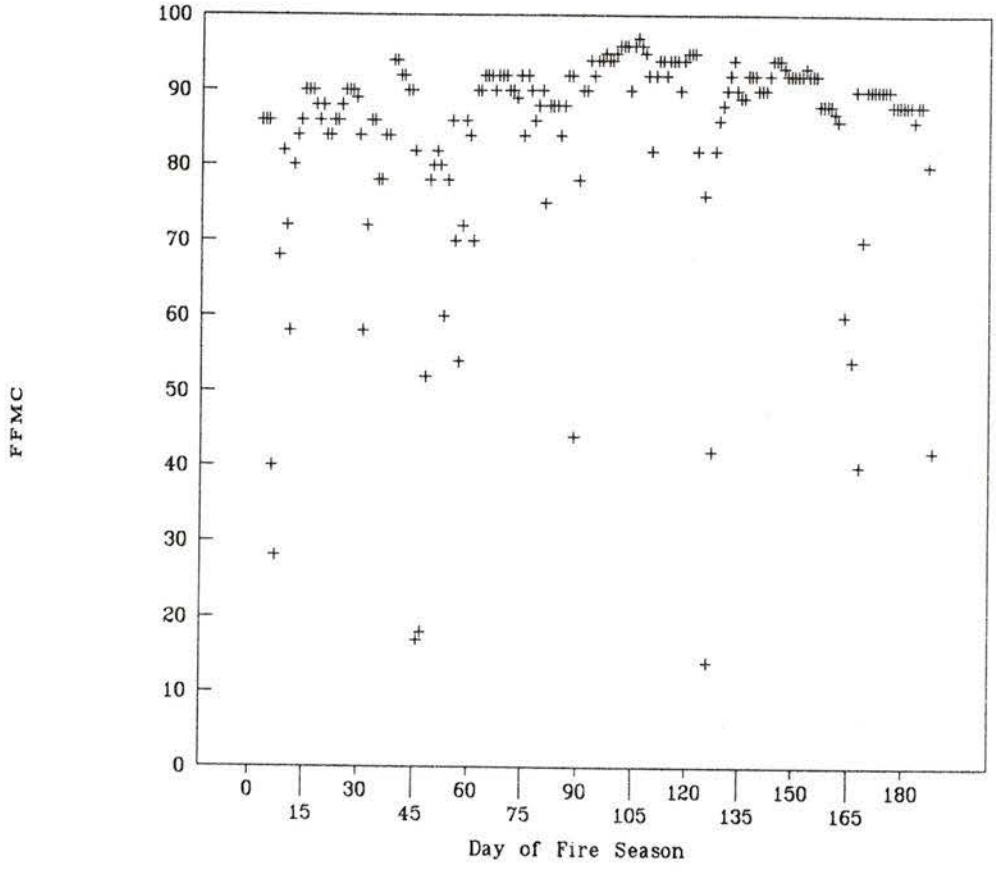


Figure 1.3: FFMC by Day of Fire Season for one Cell in Zone 63-C7

FFMC+BUI, and F for the model with Fourier terms. Hence, there are three *Method A* models discussed in this thesis: AP, AM, and AF.

For the second method, which will be referred to as *Method C* or *common*, the weather data were not averaged and fires were not totalled within a zone. One model is fitted to all cells in the zone. In this method, each (day $\times$ cell) combination yields one data point. *Method C* allows for an indication as to whether a cell contains any areas that may be prone to people-caused fires such as roadways, or recreational lakes. The *common* data did have days with multiple fires so a *common Poisson regression* was fitted; however, there were fewer multiple fire data points than with the aggregated data, and many more zeroes. Hence, the question as to whether a logistic regression model might fit such data better than a Poisson was addressed by selecting and fitting a *common logistic regression model.* Martell's models were also fitted under the *common* method. There are four *Method C* models discussed in this thesis, referred to by letters as follows: the Poisson and logistic models are CP and CL, respectively, and Martell's models are CM and CF, similar to above.

See Appendix A for a table of the data formats and analysis methods discussed here.

One of the main aims of this thesis is to investigate the merits of using Poisson regression on aggregated data rather than the logistic regression as used by Martell.

Chapter 2

Statistical Theory

2.1 Generalized Linear Regression

Generalized linear models (GLM) extend ordinary linear regression models to models with non-normal responses. The methodology relies on the theory of likelihood-based inference. This Section, 2.1, discusses generalized linear models and their estimation. The following Section, 2.2, focuses on residual analysis for GLM.

As in classical linear models, a GLM consists of an n-dimensional vector $\mathbf{Y}$ of independently distributed random variables with means $\boldsymbol{\mu}$ (representing the responses), and an $n \times p$ matrix X with values of p covariates. The classical model assumes that

$$\boldsymbol{\mu} = X\boldsymbol{\beta} ,$$

where $\boldsymbol{\beta}^T = [\beta_1, \ldots, \beta_p]$ is a vector of unknown parameters to be estimated, and that $\mathbf{Y}$ has a normal distribution with known covariance structure.

The generalization from the classical model to the set of generalized linear models is in the error distribution and in the relationship between the random and systematic components of the model, called the *link*.

The error distribution in generalized linear modelling is restricted not to the normal distribution (as in classical regression), but rather to the *exponential family of distributions.* In general, the probability density or

probability mass function $f(y;\theta)$ of an exponential family distribution can be written as

$$f(y;\theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\} . \tag{2.1}$$

When $a(y) = y$, $f(y;\theta)$ is in *canonical form* and $b(\theta)$ is the *canonical parameter* of the distribution.

The generalized linear model assumes that there is a function g, known as the *link function*,

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta} \quad i = 1, \ldots, n \tag{2.2}$$

relating $\mu_i = E(Y_i)$ to the *linear predictor* $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$. The link function is required to be a monotone, differentiable function. *Canonical links* occur when the canonical parameter of the distribution is used as the link function.

The classical model is a specific instance of a GLM with the normal error distribution, which belongs to the exponential family, and with the identity link function.

For a random variable Y that follows the Poisson distribution with parameter λ (so that $E(Y) = \lambda$), the probability mass function of Y written in exponential form is

$$f(y;\lambda) = \exp\{y \ln \lambda - \lambda - \ln y!\} .$$

Its canonical parameter is $\ln \lambda$. For n independent Poisson responses, Y_i, with means λ_i, the *log link* function could be used so that Model 2.2 becomes

$$\ln \lambda_i = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta} . \tag{2.3}$$

The application of the log link to the model results from the assumption that λ relates to the covariates $\mathbf{x}_i$ multiplicatively so that

$$E(Y_i) = \lambda_i = e^{\eta_i} = e^{\mathbf{x}_i^T \boldsymbol{\beta}} . \tag{2.4}$$

Also it conveniently maps the range of the covariates, $\mathbf{R}^p$, onto the range of the Poisson parameter λ, $(0, \infty)$.

The Bernoulli distribution of Equation 1.2 belongs to the exponential family and the probability mass function of a Bernoulli random variable Y written in exponential form is

$$f(y;p) = \exp\left\{ y \ln\left[\frac{p}{1-p}\right] + \ln(1-p) \right\} .$$

Its canonical parameter is $\ln\left\{\frac{p}{1-p}\right\}$ which gives rise to the *logit* link function. For n independent Bernoulli random variables, Y_i, $i = 1, \ldots, n$ with mean $E(Y_i) = p_i$ (p_i is the probability that Y_i assumes the value 1), and the logit link, Model 2.2 becomes

$$\ln\left\{\frac{p_i}{1-p_i}\right\} = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} .$$

Inverting this link gives

$$E(Y_i) = p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$$

which maps the range $\mathbf{R}^p$ onto $(0,1)$, the range of the Bernoulli parameter p.

We now discuss estimation in generalized linear models focusing on Poisson regression using Model 2.3. The estimates are achieved in the same way for the Bernoulli distribution with the logit link in which case the

regression is called *logistic regression.* The p parameters $\boldsymbol{\beta}^T = [\beta_1 \ldots \beta_p]$ can be estimated by maximum likelihood, where the Poisson regression log-likelihood is

$$\begin{aligned} l(\boldsymbol{\lambda}; \mathbf{y}) &= \sum_{i=1}^{n} (y_i \ln \lambda_i - \lambda_i - \ln y_i!) \\ &= \sum_{i=1}^{n} (y_i \ln \mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{x}_i^T \boldsymbol{\beta} - \ln y_i!) \, . \end{aligned}$$

Since the first derivative of the log-likelihood is non-linear in the $\boldsymbol{\beta}$'s, the maximum likelihood estimates, $\hat{\boldsymbol{\beta}}$, must be obtained numerically by an iterative procedure. Newton-Raphson iteration relates the m^{th} approximation, $\hat{\boldsymbol{\beta}}^{(m)}$, to the $(m-1)^{th}$, $\hat{\boldsymbol{\beta}}^{(m-1)}$, by

$$\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m-1)} - \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]^{-1}_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(m-1)}} \left[\frac{\partial l}{\partial \beta_j} \right]_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(m-1)}} \tag{2.5}$$

where

$$\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(m-1)}}$$

is the matrix of second derivatives of the log-likelihood l evaluated at the previous estimate $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(m-1)}$ and $\left[\frac{\partial l}{\partial \beta_j} \right]_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(m-1)}}$ is the vector of first derivatives evaluated at the previous estimate $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(m-1)}$. The *method of scoring* simplifies Equation 2.5 by replacing the matrix of second derivatives with the matrix of expected values, that is, the variance-covariance matrix of the vector of first derivatives. The maximum likelihood estimates can then be obtained by *iteratively weighted least squares.* The iterative equations solved by the method of scoring are:

$$X^T W X \hat{\boldsymbol{\beta}}^{(m)} = X^T W \mathbf{z} \tag{2.6}$$

where X is the $n \times p$ matrix of covariates; W is the $n \times n$ diagonal matrix of weights, $V(\mathbf{z}) = W^{-1}$, where $w_{ii} = \frac{1}{\lambda_i^2}\left(\frac{\partial \lambda_i}{\partial \eta_i}\right)^2 = \lambda_i$; $\hat{\boldsymbol{\beta}}^{(m)}$ is the p-dimensional vector of parameter estimates on the m^{th} iteration; and $\mathbf{z}$ is the p-dimensional vector of the adjusted dependent variable where,

$$\begin{aligned} z_i &= \sum_{j=1}^{p} x_{ij}\hat{\beta}_j^{(m-1)} + (y_i - \hat{\lambda}_i)\left[\frac{\partial \eta_i}{\partial \lambda_i}\right]_{\hat{\lambda}^{m-1}}, \\ z_i &= \mathbf{x}_i\hat{\boldsymbol{\beta}}^{(m-1)} + \frac{y_i - \hat{\lambda}_i^{(m-1)}}{\hat{\lambda}_i^{(m-1)}} \end{aligned} \tag{2.7}$$

where $\hat{\lambda}_i^{(m-1)} = e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{(m-1)}}$ from Equation 2.4. (See [13] Chapter 4, [19] Section 2.5, or [2] Part 1 Chapter 4 for further details of this method.)

2.2 Statistics used to Assess fitted Models

The goodness of fit of a model is measured by its *deviance*, the logarithm of a ratio of likelihoods. The deviance is explained and its uses for goodness of fit and for hypothesis testing are outlined in Section 2.2.1. Discussion on residuals, analysis-of-deviance, and overdispersion are presented in the remaining sections of 2.2. Methods used in this thesis for assessing the predictive value of a model are given in Section 2.2.5

2.2.1 Assessing Goodness of Fit

The deviance statistic is a measure of the goodness of fit of a model relative to the saturated model. The saturated model has the same number, n, of parameters as observations *i.e.* $\boldsymbol{\beta}_{MAX}^T = [\beta_1, \ldots, \beta_n]$. The saturated model describes the data completely so that the maximum likelihood estimates of the parameter vector result in a $\hat{\boldsymbol{\mu}}$ equal to the n observations, *i.e.* $\hat{\boldsymbol{\mu}} = \mathbf{y}$

(see [2]). The saturated model is compared to a reduced model which has the same distribution and link function but only p parameters, where $p < n$ so that $\boldsymbol{\beta}^T = [\beta_1, \ldots, \beta_p]$. The reduced model describes the data well if the *likelihood ratio* is close to 1. The *log-likelihood ratio* is

$$\ln \Lambda \;=\; \ln\left[\frac{L(\hat{\boldsymbol{\beta}};\mathbf{y})}{L(\hat{\boldsymbol{\beta}}_{MAX};\mathbf{y})}\right] \;=\; l(\hat{\boldsymbol{\beta}};\mathbf{y}) \;-\; l(\hat{\boldsymbol{\beta}}_{MAX};\mathbf{y})\,. \qquad (2.8)$$

The *deviance* is defined as minus twice the log-likelihood ratio,

$$\begin{aligned} D &= -2\left\{l(\hat{\boldsymbol{\beta}};\mathbf{y}) \;-\; l(\hat{\boldsymbol{\beta}}_{MAX};\mathbf{y})\right\} \\ &= -2\sum_{i=1}^{n}\left\{l(\hat{\boldsymbol{\beta}};y_i) \;-\; l(\hat{\boldsymbol{\beta}}_{MAX};y_i)\right\} \end{aligned}$$

for independent observations. The *deviance residuals* d_i are the terms of the deviance, $d_i = \left\{l(\hat{\boldsymbol{\beta}};y_i) - l(\hat{\boldsymbol{\beta}}_{MAX};y_i)\right\}$. For the Poisson distribution with the log link the deviance is

$$\begin{aligned} D &= 2\sum_{i=1}^{n}\left\{l(\hat{\boldsymbol{\beta}}_{MAX};y_i) - l(\hat{\boldsymbol{\beta}};y_i)\right\} \\ &= 2\sum_{i=1}^{n}\left\{(y_i \ln y_i - y_i) \;-\; (y_i \ln \hat{\lambda}_i - \hat{\lambda}_i)\right\}, \text{ where } \hat{\lambda}_i = e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}} \\ &= 2\sum_{i=1}^{n}\left\{y_i \ln \frac{y_i}{\hat{\lambda}_i} \;-\; (y_i - \hat{\lambda}_i)\right\} \\ &= 2\sum_{i=1}^{n} d_i \end{aligned}$$

and $d_i = y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)$. The deviance reduces to the residual sum of squares for a normal distribution with identity link.

For the Poisson distribution, or any non-normal exponential family distribution, under the hypothesis that the reduced model is correct, the deviance D is asymptotically distributed as χ^2_{n-p}. See [13] for a proof of

this result. Thus, for a well fitting model, D should have approximately a χ^2_{n-p} distribution. The approximation is not good, however, in small samples according to [2]. Hence, the p-value of an observed deviance should not be interpreted rigorously [2], [19]. The deviance should be used as a guideline of goodness of fit: *the smaller the deviance (relative to its degrees of freedom) the better the model fits the data.*

The null hypothesis to test the significance of t of the p parameters in a model, is $H_0 : \boldsymbol{\beta}_t^T = [\beta_1, \ldots, \beta_t] = \mathbf{0}^T$ with alternative $H_1 : \boldsymbol{\beta}_t \neq \mathbf{0}$, (where $\boldsymbol{\beta}_t$ is a vector of the t parameters, and $\mathbf{0}$ is a t-vector of zeroes). To test H_0 against H_1 two models with the same response distribution and link function are fitted to the data: the *full model* with p parameters (the model under to H_1) and the *reduced model* with $q = p - t$ parameters (the model under H_0). The statistic used to perform this test is the change in the deviances of the full and reduced models,

$$\begin{aligned}
\Delta D &= D_0 - D_1 \\
&= -2\left\{l(\hat{\boldsymbol{\beta}}_0; \mathbf{y}_0) - l(\hat{\boldsymbol{\beta}}_{MAX}; \mathbf{y})\right\} + 2\left\{l(\hat{\boldsymbol{\beta}}_1; \mathbf{y}_1) - l(\hat{\boldsymbol{\beta}}_{MAX}; \mathbf{y})\right\} \\
&= 2\left\{l(\hat{\boldsymbol{\beta}}_1; \mathbf{y}_1) - l(\hat{\boldsymbol{\beta}}_0; \mathbf{y}_0)\right\}
\end{aligned}$$

where the deviances D, the responses $\mathbf{y}$, and parameter estimates $\hat{\boldsymbol{\beta}}$ are subscripted $0/1$ according as to whether they represent the reduced model under H_0 or the full model under H_1. Under H_0, ΔD is asymptotically χ^2_t, since $D_0 \sim \chi^2_{n-q}$ and $D_1 \sim \chi^2_{n-p}$ as discussed in [2] and [13]. Large values of ΔD_{obs} yield small p-values and evidence against H_0. When the p-value is large and there is no evidence against H_0, the reduced model is preferred on the grounds of parsimony.

2.2.2 Residuals

The deviance residuals, given above, are one type of residual used in generalized linear model analysis. The deviance residual is the amount contributed by each observation to the deviance statistic. The *Anscombe residual* is similar to the deviance residual. To calculate the Anscombe residual, the observed y's and fitted $\hat{y}$'s are first transformed by,

$$\int \frac{\mathrm{d}\mu}{V^{\frac{1}{3}}(Y)}$$

where $V(Y)$ is the variance of Y as a function of the parameter μ. The residuals from the transformed y and $\hat{y}$ are scaled and standardized by dividing by

$$V^{\frac{1}{6}}(Y)\sqrt{(1-h_{ii})}$$

to stabilize the variance where h_{ii} are the diagonal entries of the matrix $H = X(X^TX)^{-1}X^T$. The standardized Anscombe residual for the Poisson distribution, then, is

$$r_A = \frac{\frac{3}{2}(y^{\frac{2}{3}} - \hat{\lambda}^{\frac{2}{3}})}{\hat{\lambda}^{\frac{1}{6}}\sqrt{(1-h_{ii})}}.$$

The Anscombe residuals have properties for non-normal distributions similar to the properties of normal residuals. The Anscombe and deviance residuals give very similar values. See [19] for a discussion on Anscombe residuals.

McCullagh and Nelder [19] state that for the Poisson distribution, the Anscombe and deviance residuals are smaller and more stable than the *Pearson residuals*,

$$r_P = \frac{y-\mu}{\sqrt{V(\mu)}} = \frac{y-\hat{\lambda}}{\sqrt{\hat{\lambda}}}.$$

The Pearson residuals may have a skewed rather than normal distribution.

See [19] Section 12.6 for a discussion on residual plots. The residual plots used in this thesis for the Poisson regression models are plots of the Anscombe residuals against $2\sqrt{\hat{\lambda}_i}$, the variance-stabilizing transformation of the fitted y_i's. The residuals used in this thesis for the logistic regression residual plots are the Pearson residuals

$$\frac{(y - \hat{\mu})}{\sqrt{(y(1-y))}} .$$

If these plots show some pattern, then the model may not fit well, or the link function may be incorrect.

Plots of the residuals against cell row and column co-ordinates were also constructed for the *Method A* aggregated models to determine if there is any spatial component in the residuals that could be incorporated into the model. One other residual plot, referred to by Aitken [1] as a quantile plot, was constructed to check whether the correct distribution was used in Poisson model fitting. For this residual plot, the Anscombe residuals are sorted then plotted against the normal quantiles $\Phi^{-1}\{(i - .5)/n\}$ where $i = 1, \ldots, n$ and $\Phi(x)$ is the normal cumulative distribution function. The plot should be close to a straight line when the Poisson distribution is correct; otherwise, a pattern in the plot suggests the Poisson distribution may not be appropriate for the data.

2.2.3 Analysis-of-deviance and Variable Selection

The generalized linear regression analogue of analysis-of-variance is *analysis-of-deviance* or ANODEV. To construct an ANOVA table for, say, a model

with two factors, the sums of squares for these two factors and their interaction can be obtained by fitting the successive models 1, A, $A+B$, and $A+B+A\cdot B$ and taking the first differences of the residual sums of squares. To construct an ANODEV table, a sequence of generalized linear models are fitted and the first differences of their deviances will give the values for the table. An entry in an ANODEV table for a covariate is interpreted as representing the variation attributed to that covariate eliminating the effects of the covariates above it and ignoring covariates below it in the table.

As in classical regression, if the covariates in a generalized linear model are not orthogonal, different sequential models will produce different ANODEV tables. ANODEV is most effective when two ANODEV tables are constructed for a single model using two different sequences of nesting of the parameters. The variable with the larger change in deviance regardless of its sequence in the model is selected over variables with a smaller change in deviance in at least one ANODEV table. Table 2.1 shows the two ANODEV tables generated for the fire weather indices FWI and DC on data in zone 21-C2Y. The interchange of deviance when the fitting order is reversed for FWI and DC indicates that there is some correlation between these two indices. The preferred of these two indices is FWI since FWI is significant whether or not DC is in the model and, when FWI is in the model, DC is not significant. Let it be reiterated here, that because of the approximation of the Chi-square distribution of the deviance, no p-values are attached to the deviance and hence no p-values are given in the ANODEV table. ANODEV is used in this thesis to assist in the selection

Model	D	d.f.	Source	ΔD	d.f.
1	236.9	889			
FWI	214.9	888	FWI	22.0	1
FWI+DC	212.7	887	DC	2.2	1

Model	D	d.f.	Source	ΔD	d.f.
1	236.9	889			
DC	231.0	888	DC	5.9	1
FWI+DC	212.7	887	FWI	18.3	1

Table 2.1: Example of Analysis-of-deviance Tables

of correlated covariates, but the primary technique for model selection is backwards elimination. See [1] and [19] for an explanation of backwards elimination.

2.2.4 Overdispersion

Another measure of the goodness of fit of a Poisson regression model is to test the Poisson characteristic of having equal mean and variance. *Overdispersion* occurs when the variance is significantly larger than the mean of the data. A Poisson overdispersion test statistic is

$$v = \frac{(n-1)S^2}{\bar{X}}$$

and can be calculated for a set of data. Statistic v follows a Chi-square distribution when calculated for a set of Poisson data. The statistic to test

for overdispersion from Dean and Lawless in [12] used here is the *adjusted T test for overdispersion*

$$T = \frac{\sum_{i=1}^{n} \left\{(y_i - \hat{\lambda}_i)^2 - y_i + \hat{h}_{ii}\hat{\lambda}_i\right\}}{\sqrt{2\sum_{i=1}^{n} \hat{\lambda}_i^2}}$$

where λ_i is defined in Equation 2.4 and h_{ii} are the diagonal entries of the matrix $H = X(X^TX)^{-1}X^T$. Dean and Lawless show that T converges to $N(0,1)$ for large sample sizes. If the Poisson model fits the data well, the adjusted T statistic will be small in absolute value. Large values will indicate evidence of overdispersion.

If there is overdispersion in the model, the selected model should be reconsidered. If the overdispersion cannot be explained by the covariates in the model, or by any other means, then McCullagh and Nelder [19] suggest estimating the dispersion parameter $\hat{\sigma}^2$ = deviance/d.f. and dividing the deviance and the standard errors of the parameter estimates by $\hat{\sigma}^2$. The parameter estimates themselves do not change, however, the parameters that are selected by the backwards elimination technique may differ under the scaled deviance. This process is a result of *quasi-likelihood* methods and is also discussed in [1]. In this thesis, the models are selected to be used to predict fires. Since quasi-likelihood does not change the parameter estimates, the predicted values will not change when adjusting for overdispersion, unless the selected covariates change. Hence, when overdispersion is noted, models are selected using the scaled deviance and predictions are calculated as usual.

2.2.5 Assessing a Model's Predictive Ability

To further assess the goodness of fit of a model, it is necessary to assess how well the model predicts forest fires. The remainder of this section will present and discuss methods to assess the predictive ability of a model. The three statistics used to evaluate the predictions are the *deviance*, the *PRESS* statistic and *cross validation* statistics. Bonneu explains in [4] that the *deviance* not only is a measure of the goodness of fit of a model but also supplies a measure of the predictive value of a model.

The *Predicted Sum of Squares* or PRESS statistic is obtained by fitting the model to the data while omitting the i^{th} observation, then predicting the omitted y_i value from the resulting model. The notation to refer to the model fitted without y_i uses the subscript (i). The actual value y_i is compared to the predicted $\hat{y}_{(i)}$ by the statistic

$$p_{(i)} = \frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{1 + h_{(i)}}},$$

where

$$s_{(i)} = \sqrt{\frac{(n-p)\hat{y}_i^2 - e_i/(1-h_{ii})}{n-p-1}} \; and e_i = y_i - \hat{y}_i \; .$$

This procedure is performed by omitting all $i = 1, \ldots, n$ observations in turn where i indexes the day of the fire season. The PRESS statistic is the sum of the $p_{(i)}$'s

$$P = \sum_{i=1}^{n} p_{(i)} \; .$$

The model does not have to be fitted n times for the omission of all n observations, but can be calculated from the residuals of the fitted model. See [1], [19], [23], and Appendix B for the derivation of the PRESS statistic

used for calculation. The PRESS becomes

$$p_{(i)} = \frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{(1 - h_{ii})}} \tag{2.9}$$

where n is the number of observations, p is the number of parameters, and h_{ii} are the diagonal entries of the matrix $H = X(X^TX)^{-1}X^T$.

The third type of prediction statistics are based on *cross-validation* which is appropriate for large data sets, such as six years of forest fire data. These cross-validation prediction statistics are given more emphasis in assessing the prediction ability of the models than the PRESS statistic. The main reason the PRESS statistic is not emphasized in this thesis is that it is not useful in comparing aggregated models to common models because of the difference in the number of data points, n.

Cross-validation of the model involves omitting a whole fire season of data and fitting a model on the remaining data. The omitted observations are predicted and a *predicted deviance statistic* and two *sum of squares statistics* in Equations 2.10 and 2.11 are calculated.

The predicted deviance is simply the deviance with the fitted values replaced by predicted values. The predicted deviance for a Poisson model is

$$D_p = 2\sum_{i=1}^{186}\left\{y_i \ln\frac{y_i}{\hat{y}_i} - (y_i - \hat{y}_i)\right\}$$

where i indexes the day of the fire season, and the $\hat{y}_i$s are the predicted values for a zone. For *Method A* aggregated data, the $\hat{y}_i$'s are calculated from the fitted model. For *Method C* common data, predictions are made for each cell and are summed to give a prediction for the whole zone so that $\hat{y}_i = \sum_{j=1}^{m}\hat{y}_{ij}$ where m is the number of cells in the zone. The predicted

deviances for aggregated and common models, then, can be compared. The predicted deviance from Poisson models can be compared with the predicted deviance from Martell's logistic regression since he assumes that fires occur in a Poisson process and the Poisson distribution is used to make predictions from his models. In addition, the common logistic *CL* model can be compared using the above predicted deviance since the logistic predictions for each cell are summed to give a prediction for the zone and the distribution of a sum of a large number of Bernoulli random variables each with a small parameter is approximately Poisson. The predicted deviance is used to compare the predictive ability of all models.

The cross-validation predicted deviance is a better indicator of the predictive value of a model than the PRESS statistic because it can be used to compare aggregated to common models.

The above measures of the predictive value of the models are useful only when models fitted with the same underlying distribution and link function are compared (or are approximately comparable as with the predicted deviance for the *Method CL* model). In this thesis it is also desirable to compare the predictive power of the models used here namely the aggregated *AP* and common *CP* and *CL* models with the models of Kourtz and Martell applied to the same data. Kourtz's predictions were provided by day and cell for the Kamloops Region by Protection Branch. They are in the form of the predicted number of fires per cell and, of course, are mostly zeroes. Predictions were obtained from Martell's aggregated *AM* and *AF* and common *CM* and *CF* models fitted to B.C. data. The *sum*

of squares statistic

$$S_1 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.10}$$

is used to compare all models. A *"standardized" sum of squares* version (similar in form to the Pearson goodness-of-fit statistic),

$$S_2 = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}\,, \tag{2.11}$$

is also used for comparing all but Kourtz's model (since his predictions are mostly zeroes.) Neither of these statistics are used in a formal hypothesis testing fashion, rather their values are simply compared for the various methods with smaller values indicating a "better" predictor. Since the predictions for the common models are summed to give a prediction for the whole zone, and since the sum of squares statistics do not depend on any distribution, they can be used to compare all common and aggregated models. Both statistics S_1 and S_2 will be small if predictions are low on days when fires do not actually occur. However, statistic S_2 is more sensitive than S_1 to having high predictions on days when fires do occur. The sensitivities of the statistics can be seen by expanding the sums of squares S_1 and S_2 as follows:

$$\begin{aligned}
S_1 &= \sum_i y_i^2 - 2\sum_i y_i\hat{y}_i + \sum_i \hat{y}_i^2 \\
&= (N+3) - 2\sum_i^{*} \hat{y}i + \sum_i \hat{y}_i^2 \\
S_2 &= \sum_i \frac{y_i^2}{\hat{y}_i} - 2\sum y_i + \sum \hat{y}_i \\
&= \sum_i^{*} \frac{y_i^2}{\hat{y}_i} - 2N + \sum_i \hat{y}_i
\end{aligned}$$

where N is the total number of fires and $\sum_i^*$ represents the sum only over days with fires. The statistic S_2 will do well (*i.e.* be small) when the $\hat{y}_i$'s are large on days when fires occur, but is otherwise small. The statistic S_1 will do well in similar situations (since $\sum_i \hat{y}_i^2$ will be small but the $\sum_i^* \hat{y}_i$ will be large). However, S_2 will place a greater importance on having $\hat{y}_i$ large on days when fires occur than having it small on days with no fire.

2.3 Software

The software *Generalized Linear Interactive Modelling* or *GLIM* (release 3) developed by R.J. Baker, M.R.B. Clarke and J.A. Nelder of the Royal Statistical Society is a FORTRAN based program and solves Equation 2.6 iteratively. See [2]. By specifying a link function and error term (`$LINK L $ERR P` assign the log link and Poisson error structure) the specified regression model can be fitted to the data. The data are read in from an ASCII file and output is written to a listing file. GLIM provides the deviance, parameter estimates and their standard errors, Pearson residuals, and the fitted values of the model as standard output. Plots and other residuals and statistics can be programmed.

A problem arises in GLIM for a data set which has no fires because the iterative procedure may not converge. The response variable, the number of fires, then, is a column of zeroes so the linear predictor $\eta = \ln \lambda$ tends to minus infinity. The deviance and parameter estimates will be approximately correct so model selection can carry on as usual. In fact, this problem arose when analysing forest fires in individual cells as mentioned in Section 4.1.1.

Chapter 3

Preliminaries

3.1 Initial Data Analysis

This section describes the data analysis that was performed prior to fitting any models. The initial data analysis includes examining the data for weekly, seasonal, and yearly trends, and for geographic trends. An examination of the data's conformity to the Poisson characteristic of equal mean and variance is included. This section ends with some comments on missing observations and errors in the data.

Following Martell in [9] the data were firstly examined for any trends in fire occurrence by day of the week. The Pearson Chi-square statistic was calculated for nine forest fire causes to test the hypothesis that fires occur uniformly with probability $\frac{1}{7}$ on each day of the week. The p-values were small ($< .05$) for many of the cause classes. Forest fires in the Kamloops Forest Region classified as caused by recreationalists (*e.g.* campfire) occur in larger proportion on weekends while logging fires occur primarily on weekdays as seen in Figure 3.1. In fact, the difference in day of the week occurrence of fires may be due to a weekday versus weekend[1] trend. Table 3.1 gives details of data and testing of the hypothesis of uniformity: *i.e.* that with probability five sevenths a fire will occur on a weekday and

[1] A weekday is one of Monday, Tuesday, Wednesday, Thursday, or Friday; a weekend is a Saturday or Sunday.

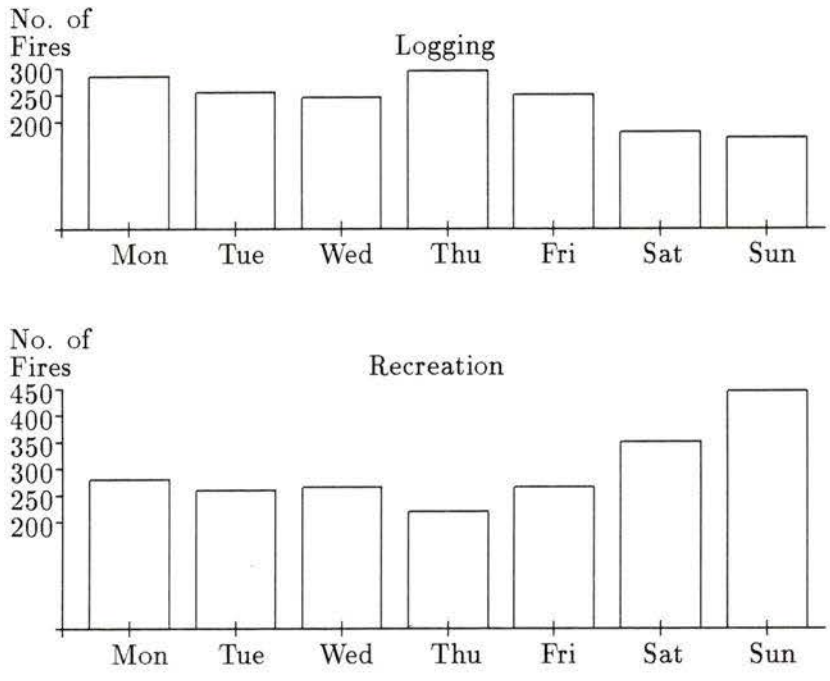


Figure 3.1: Number of People-Caused Forest Fires by Day of the Week for the Kamloops Forest Region, for General Causes 'Recreation' and 'Logging' from 1985 to 1988 and 1990.

with probability two sevenths on the weekend for each of the nine general causes. The Pearson Chi-Square statistic is significant ($p \leq .05$) indicating a difference in forest fire probability from weekday to weekend for general causes 'unknown', 'recreation', 'logging', and 'miscellaneous known'.

The data were examined for trends in the number of people-caused forest fires over the fire season. Figure 3.2 shows the number of people-caused fires per week from 1985 to 1990 in zone 63-C7. The number of fires appears to begin increasing in June and to peak in late July and

General Cause	Frequency of Fires (Expected Freq. under H_0)		χ^2_{obs}	p-value
	weekday	weekend		
Unknown	109	60		
	(121)	(48)	3.98	0.046
Recreation	1283	795		
	(1484)	(594)	95.54	0.0000
Railroad	250	79		
	(235)	(94)	3.35	0.067
Logging	1333	395		
	(1209)	(83)	44.84	0.0000
Right-of-	69	37		
Way Constr.	(76)	(30)	2.08	0.15
Other	161	52		
Industrial	(152)	(61)	1.81	0.18
Not	29	16		
Assigned	(32)	(13)	1.08	0.3
Land	810	326		
Clearing	(811)	(325)	0.01	0.9
Misc.	2325	1052		
	(2412)	(965)	11.02	0.0009

Table 3.1: Statistics and p-values for testing H_0: Probability of a people-caused fire being on a weekday is $\frac{5}{7}$ and probability of being on a weekend is $\frac{2}{7}$ for the Kamloops Forest Region, 1985-1990 data.

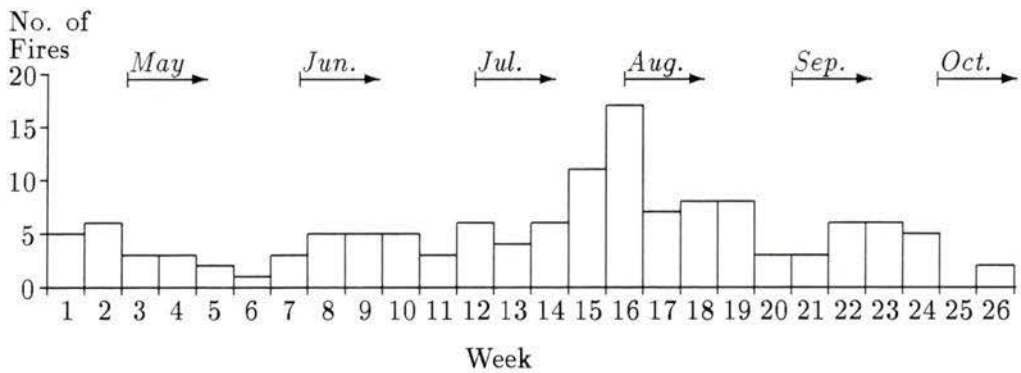


Figure 3.2: Number of People-Caused Forest Fires per week of the Fire Season in Kamloops Forest Region, 1985-1990.

early August. Although the number of fires decreases as the return to school in September approaches, there is a slight increase in late September due possibly to the hunting season and increased industrial activity. The seasonal trend in the number of fires seems to differ somewhat from year to year, even though in the grouped data there appears to be an overall trend.

The number of all forest fires (including lightning caused) in the whole province of B.C. from 1970 to 1989 was examined for trends from year to year. These data were obtained from Datatrieve datafiles, a different source than that used for the data for model fitting. See Section 1.3.2 for an explanation of the datafiles. The Ministry of Forests' document [25] states that the number of forest fires in B.C. has increased over the last seventy-five years. This increase may be attributed to improvements in weather station detection of lightning strikes and fires, to increased aerial detection, and possibly to an increase in logging and recreation activities

in the forests and parks. A simple least squares linear regression model fitted to forest fire data from 1970 to 1989 had a positive slope; however, the slope is not significantly positive ($p = .8$). The fitted least squares line is plotted in Figure 3.3 along with the number of forest fires. The fitted regression line, (with standard errors in brackets) is

$$\begin{aligned} \mathrm{E}(\textit{Forest Fires}) &= 1941.2 + 8.3 \times (\textit{YEAR} - 1900) \\ &\quad (2610) \qquad (32.4) \end{aligned}$$

with an R^2 of 0.0036. A simple linear regression line is fitted to the number of people-caused forest fires from 1970 to 1991. The number of people-caused forest fires and the fitted regression line are shown in Figure 3.4. The fitted regression line (with standard errors in brackets) is as follows:

$$\begin{aligned} \mathrm{E}(\textit{People-Caused Forest Fires}) &= 2456.5 - 12.7 \times (\textit{YEAR} - 1900) \\ &\quad (1225) \qquad (15.4) \end{aligned}$$

with an R^2 of only 0.0367. There was no significant ($p = 0.42$) difference in the number of people-caused fires from year to year. Thus, it would appear that the claimed increase in fire incidence over the last seventy-five years is due mainly to an increase over the years 1912 to 1970 (this period saw improvements in fire detection technology) rather than any change over the last twenty years.

A Poisson regression model was fitted to the number of people-caused forest fires from 1970 to 1991 in B.C. resulting in the estimated model,

$$\begin{aligned} \lambda &= exp(\ 8.02 - 0.009 \times (\textit{YEAR} - 1900)) \\ &\quad (0.08) \qquad (0.001) \end{aligned}$$

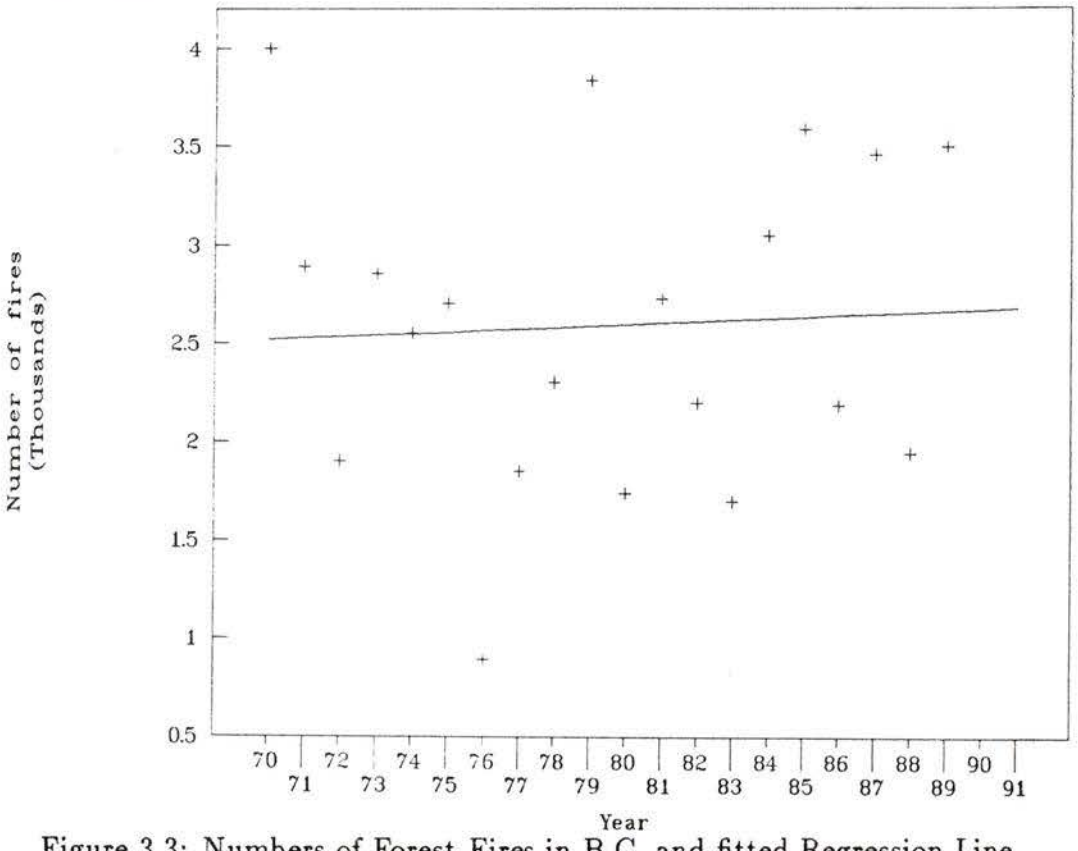


Figure 3.3: Numbers of Forest Fires in B.C. and fitted Regression Line

with a deviance of 1913 on 18 degrees of freedom. The parameter for *YEAR* is significant ($p < 0.01$). This suggests that the number of people-caused forest fires is a function of *YEAR*. Figure 3.5 shows the plot of the fitted line. From the normal linear regressions, the number of fires is not increasing or decreasing linearly. The Poisson regression models the variation from year to year in the number of fires. Part of the aim of this thesis is to determine how to explain this variation by causes such as weather.

The number of people-caused fires and the average of the six fire weather indices over a fire season per cell over the whole Kamloops Region were

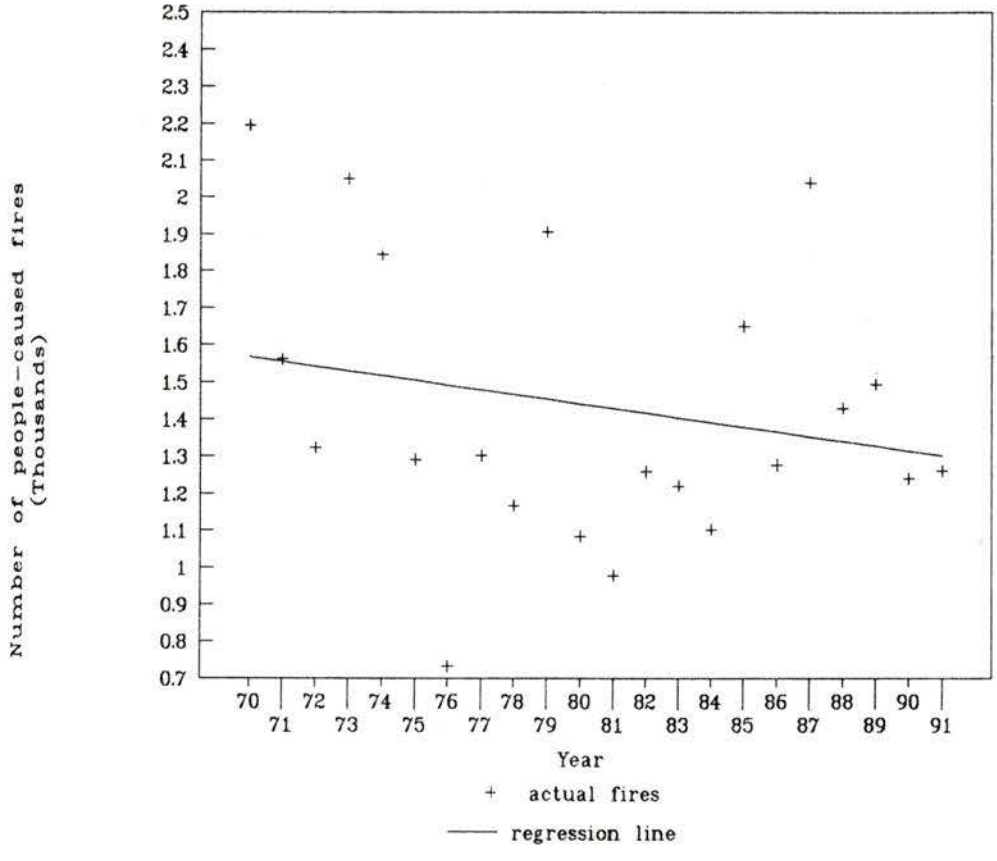


Figure 3.4: Numbers of People-Caused Forest Fires in B.C. and fitted Regression Line

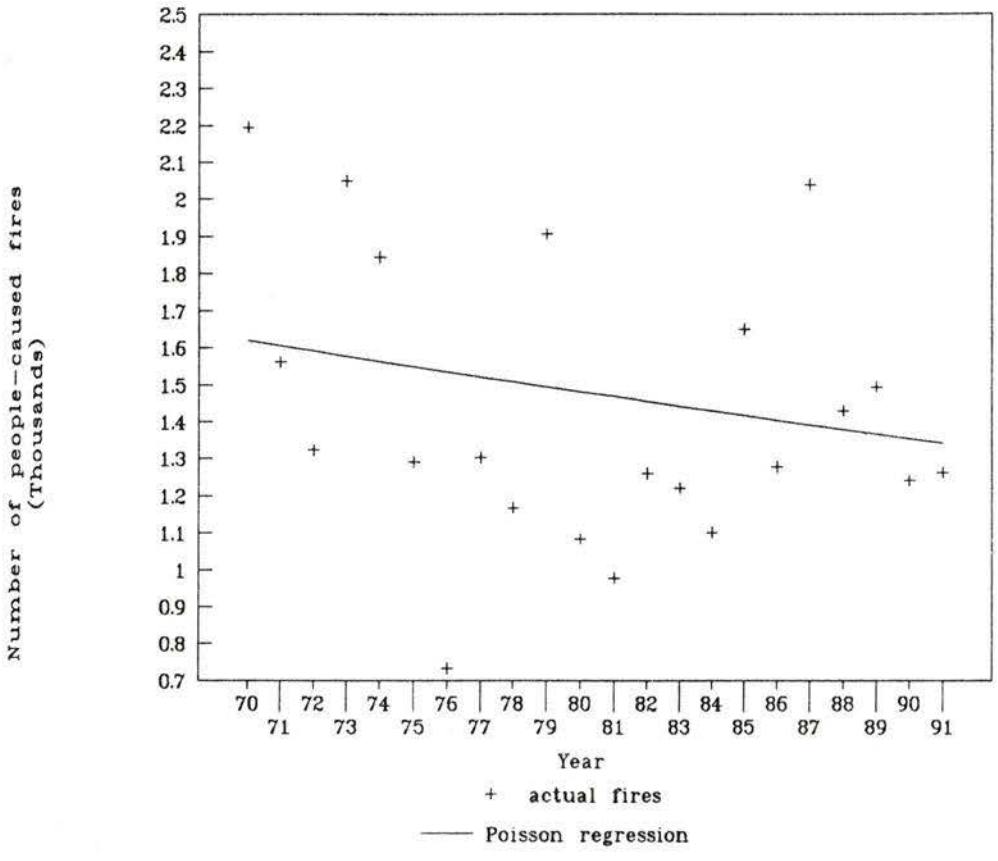


Figure 3.5: Poisson Regression of Number of People-Caused Fires per Year

examined. The number of people-caused fires per cell for the 1988 fire season are shown in Figure 3.6. People-caused forest fires tend to be located near roads and highways which are usually also highly populated areas. Hence, proximity to a road may be important in people-caused forest fire prediction.

The geographic trends for the six weather indices for the 1988 fire season are shown in Figures 3.7 through 3.12. The locations with dry measures of these indices tend to overlap with the locations of higher numbers of people-caused fires. Therefore, dryer conditions and people-caused forest fire occurrence appear to be correlated.

The Poisson assumption that the mean and variance of the data are equal is tested to assure that Poisson regression is reasonable in view of the fact that no other people-caused forest fire research has employed Poisson regression. The sample means and variances of the number of fires for each cell in zone 63-C7 and for the whole zone from 1985 to 1990 were calculated and compared. The Poisson overdispersion test statistic

$$v = \frac{(n-1)S^2}{\bar{X}}$$

was calculated for each cell. The means and variances were not significantly different for any cells in zone 63-C7 except for cell with row and column co-ordinates (84×78) where the p-value is .0597. The *adjusted T statistic* from Section 2.2.4 is used as an indication of overdispersion once a model is in place.

An outlying observation was found in zone 63-C7 where five fires were recorded on May thirty-first, 1988 in one cell. Upon investigation of these

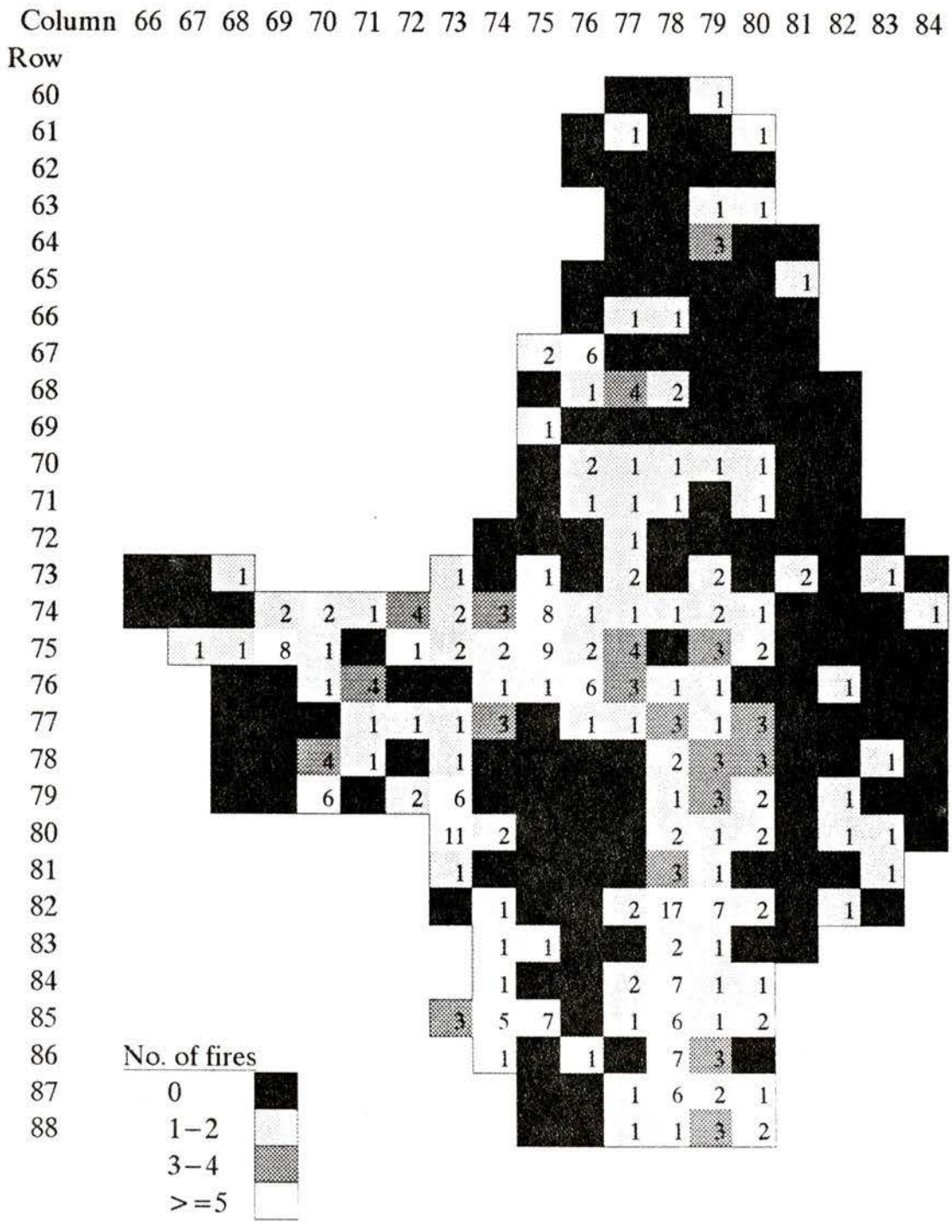


Figure 3.6: The number of people–caused Forest Fires in 1988 in the Kamloops Region, by cell.

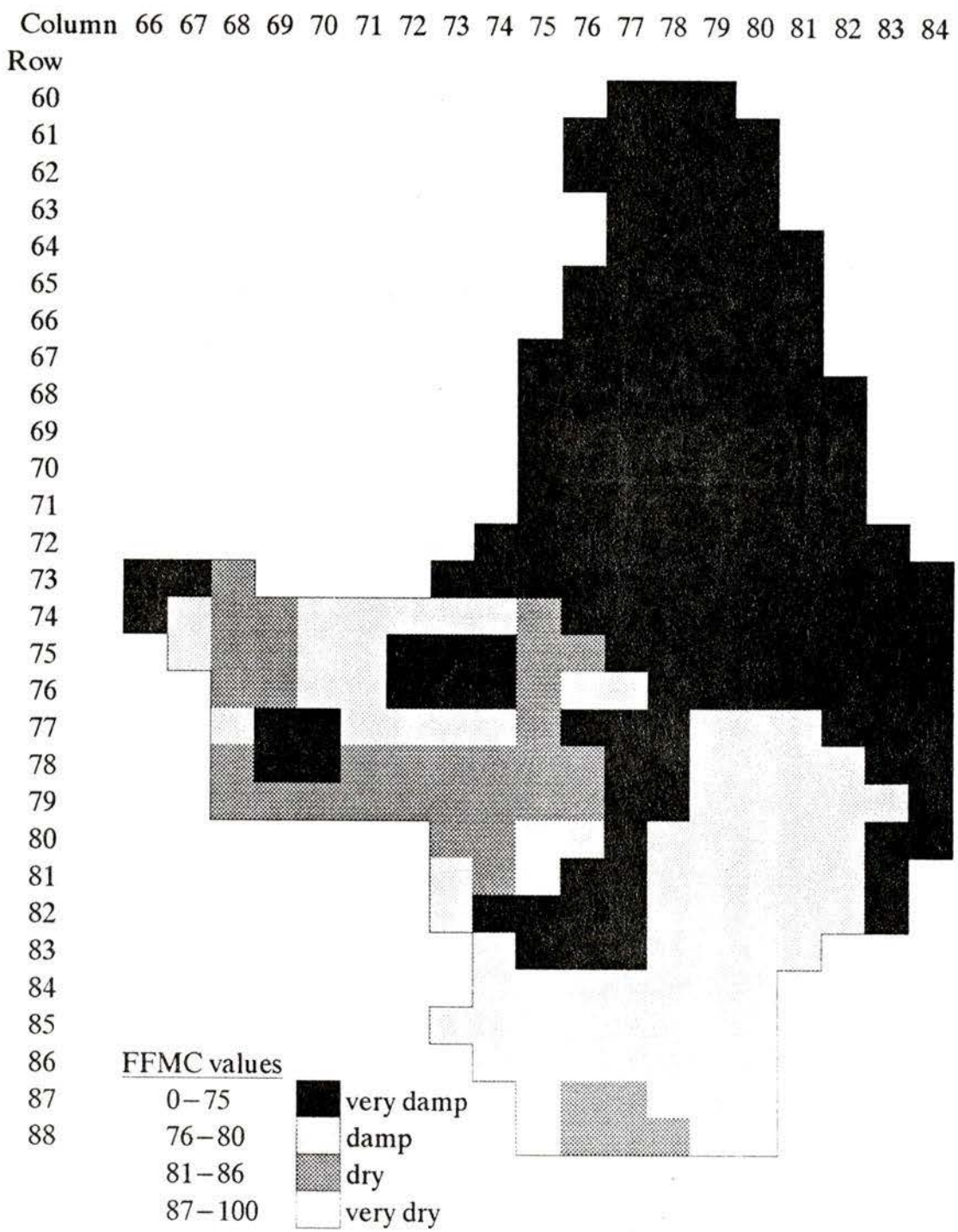


Figure 3.7: The average FFMC for the 1988 fire season in the Kamloops Region, by cell. The FFMC is the Fine Fuel Moisture Code.

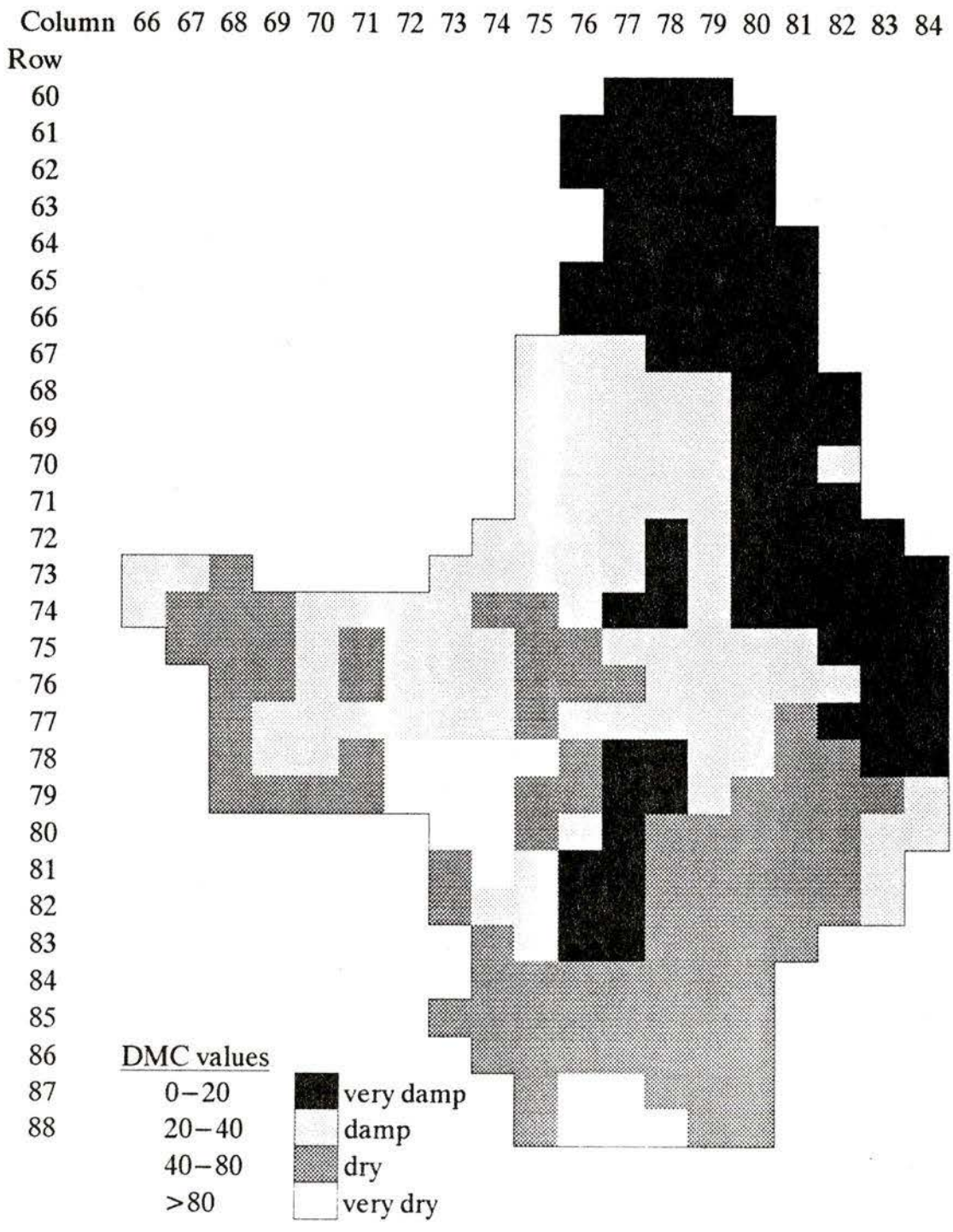


Figure 3.8: The average DMC for the 1988 fire season in the Kamloops Region, by cell. The DMC is the Drought Moisture Code.

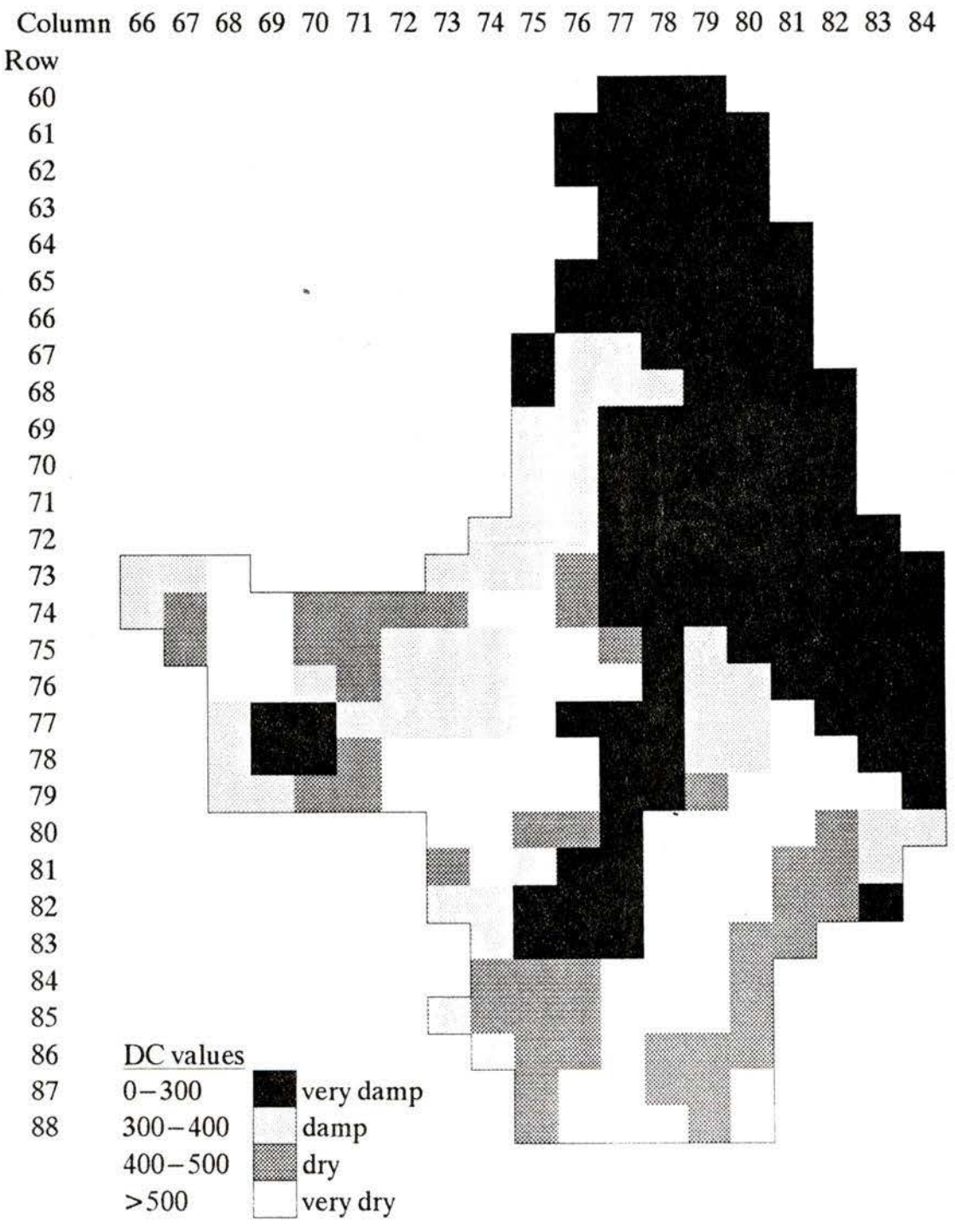


Figure 3.9: The average DC for the 1988 fire season in the Kamloops Region, by cell. The DC is the Drought Code.

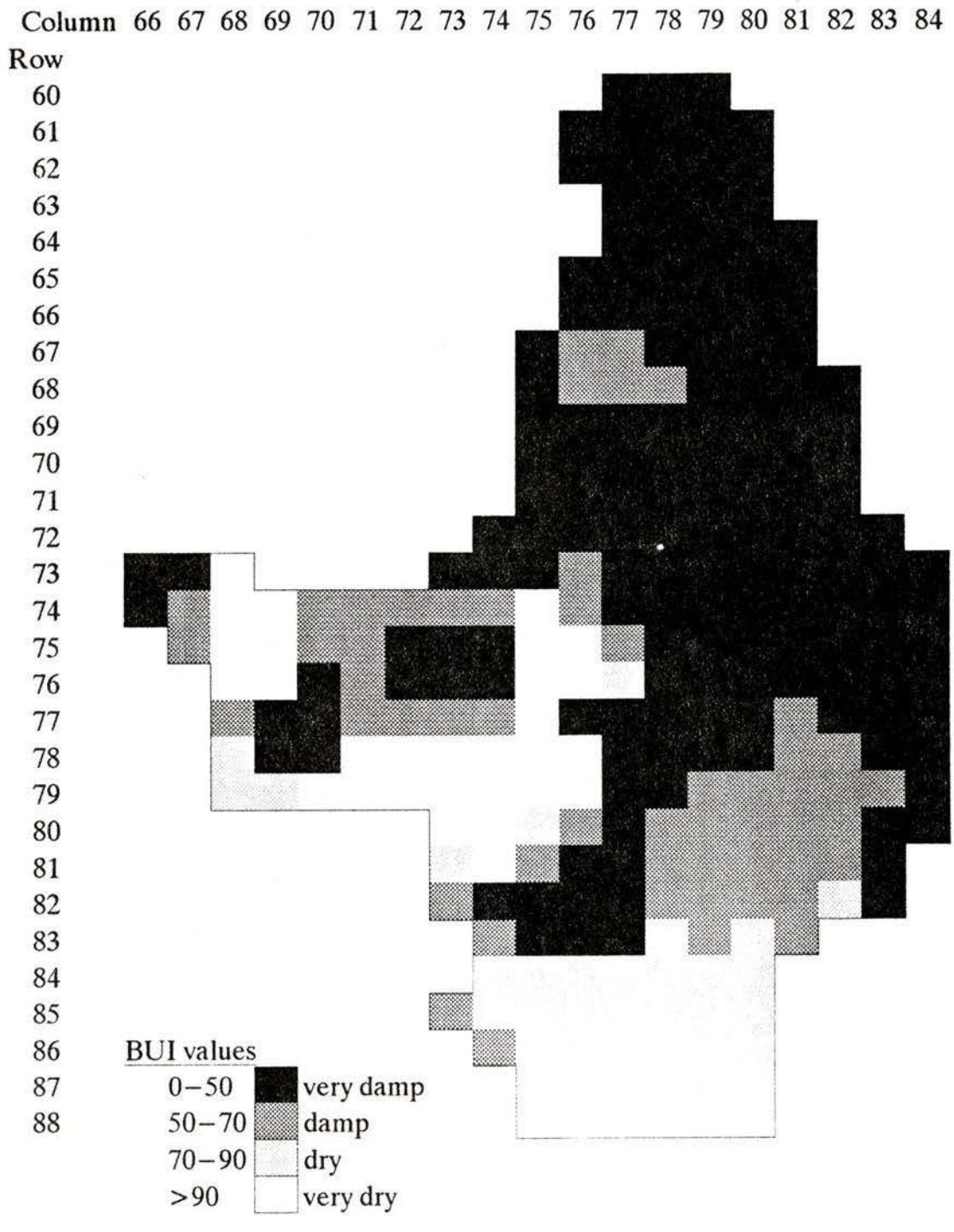


Figure 3.10: The average BUI for the 1988 fire season in the Kamloops Region, by cell. The BUI is the Buildup Index.

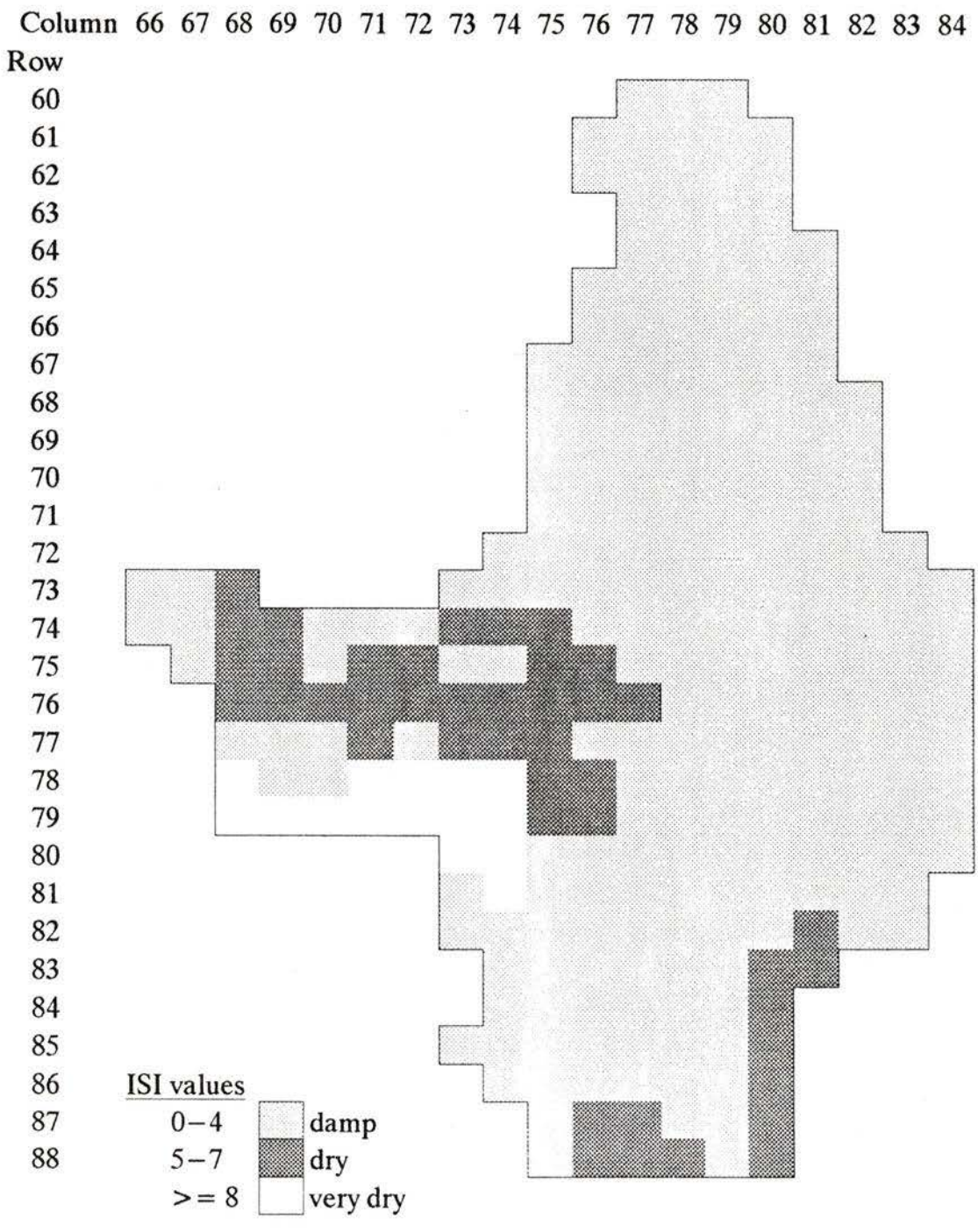


Figure 3.11: The average ISI for the 1988 fire season in the Kamloops Region, by cell. The ISI is the Initial Spread Index.

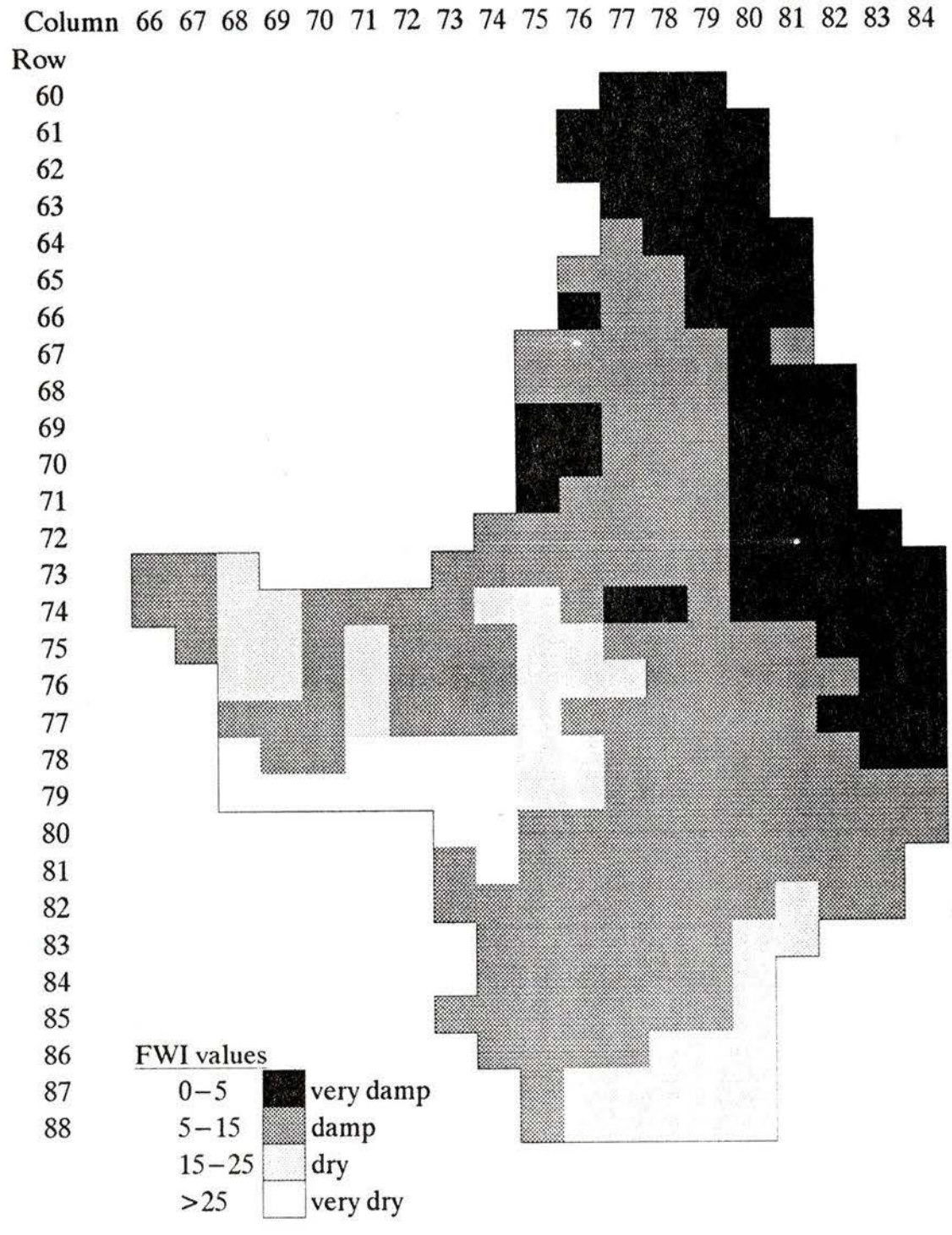


Figure 3.12: The average FWI for the 1988 fire season in the Kamloops Region, by cell. The FWI is the Fire Weather Index.

fires in Datatrieve, it was found that all five of these fires were recorded as having been ignited and detected at the same time, having the same general and specific cause classes of no spark arrestor on a train, and occurring at the same latitude and longitude. In addition, only one of the five fires was recorded to have burned any area. For the purpose of regression analysis, these five fires were regarded as one.

3.2 Covariates

This section lists and discusses possible explanatory variables useful in people-caused forest fire occurrence predictions. The process of covariate selection discussed in Appendix 2 of [8] was followed here. *Backwards elimination* was used in this thesis to select covariates for the final models. See [1] and [19] for an explanation of backwards elimination. This section discusses which variables were included in the initial stage of backwards elimination as well as some preliminary analysis of certain covariates. Also, the initial data analysis of Section 3.1 is referred to for some of the variables.

The possible explanatory variables are (numbered for future reference),

1. weather data, including FFMC, DMC, DC, BUI, ISI, and FWI,
2. fuel type, such as lodgepole pine, Douglas fir, *etc.*,
3. topography; elevation,
4. public awareness of fire danger,
5. land use patterns.

The following covariates (6-11) may not only have predictive value but may also be considered as surrogate explanatory variables for land use pattern

(5.) information:

6. day of the week,
7. day of the fire season,
8. proximity to a campground, lake, road, city, logging operation,
9. when and where logging road and logging camp closures occur,
10. when and where restrictions are put on campfires, and
11. cause of fire.

(1.) **Weather** affects the moisture content of the forest fuels and so affects people-caused forest fire occurrence. See [20] and [28] for some discussion on the influence of weather on forest fire occurrence. The FFMC is the most sensitive index to daily changes in weather and has been used for forest fire prediction by others. (See Chapter 1.) The proportion of days with fires occurring in zone 63-C7 of the Kamloops Forest Region for three ignition classes of the FFMC index are shown in Table 3.2. These proportions suggest that, as might be expected, more people-caused forest fires occur per day of dry, hot conditions (for FFMC ≥ 92) than per day of damp conditions (for FFMC ≤ 87). In fact, the Chi-square test of independence of the number of people-caused forest fires and FFMC was highly significant ($p = 1.4 \times 10^{-9}$); hence, it appears that the proportion of days with fires increases as the FFMC index increases.

Todd and Kourtz, [27], suggest using an indicator variable for the three ease of ignition classes of FFMC in Table 3.2 rather than actual FFMC values. A similar idea, from Dr. M. Lesperance, is to use as a covariate a

FFMC	Number of Days	Number of Fires	Proportion of days with Fires	Ease of Ignition
≤ 87	5831	31	0.0053	difficult
88-91	3932	61	0.0155	moderate
≥ 92	1911	40	0.021	very easy

Table 3.2: Proportions of fires per day for zone 63-C7 from 1986-1991 for the three FFMC ease of ignition classes.

logistic transformation of FFMC, namely

$$\text{FFMC}^* = \frac{2\exp\{0.1(\text{FFMC} - 100)\}}{1 + \exp\{0.1(\text{FFMC} - 100)\}}, \tag{3.1}$$

rather than FFMC itself. The graph of this transformation is shown in Figure 3.13. It was chosen in a purely ad hoc way to reflect the steep change in fire probability that seems to occur between values 80 and 100 of the FFMC.

Ideally, historic *forecasted* weather data should be used to develop predictive models since in practice forest fires will be predicted from weather forecasts. However, only observed weather data are kept on record at the Ministry of Forests. Furthermore, historic weather forecasts are not kept electronically at Environment Canada. Since weather forecasting techniques are improving, the foresters at B.C. Ministry of Forests expressed no concern that the predictive models would be developed using observed weather data rather than weather forecasts.

The correlation of raw weather observations with the six weather indices

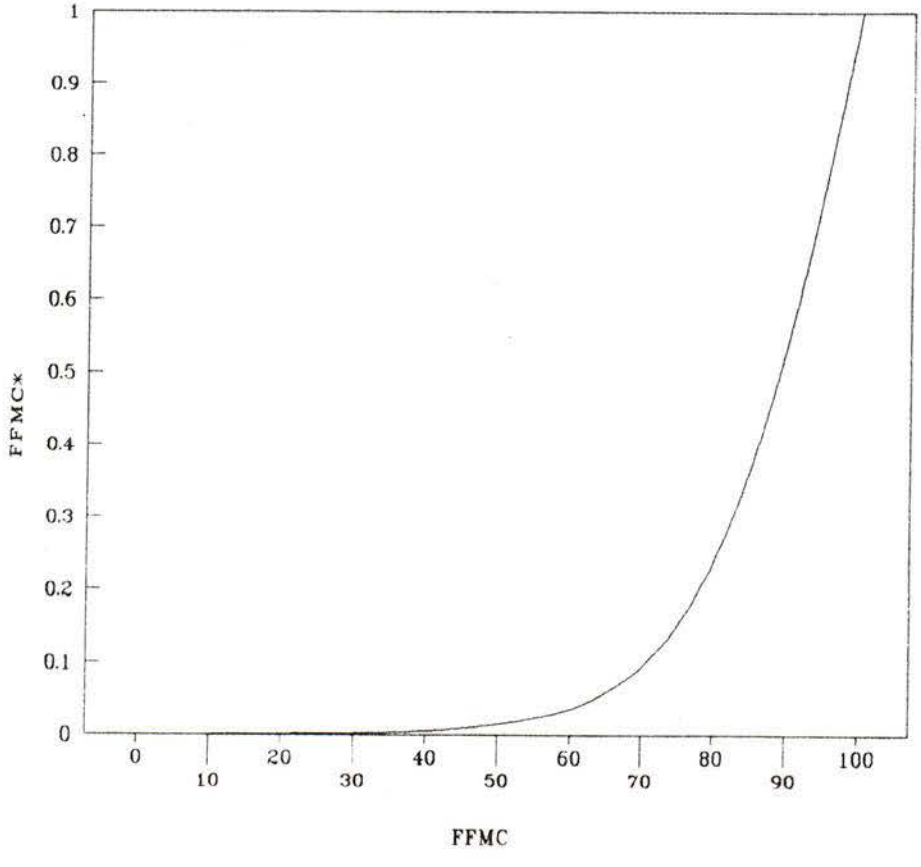


Figure 3.13: Logistic Transformation of FFMC

	Temp.	Humidity	Rain	Wind: Speed	Direction
FFMC	0.6320	-0.7181	-0.7139	0.2043	0.0553
DMC	0.4669	-0.4320	-0.2398	0.2355	0.1226
DC	0.3258	-0.2035	-0.1668	0.1360	0.2887
BUI	0.4754	-0.4255	-0.2414	0.2262	0.1492
ISI	0.5641	-0.5830	-0.3012	0.5599	0.1119
FWI	0.6119	-0.6052	-0.3022	0.4810	0.1369

Table 3.3: Correlation Coefficients for weather and weather indices for the Kamloops Forest Region

is shown in Table 3.3. There appears to be no problem with collinearity in model fitting. In addition, initial model fitting indicated that the weather indices give a smaller deviance than the raw weather data with the exception of wind speed. The process of model selection began with the six indices and wind speed.

(2.) **Fuel type** is a necessary explanatory variable for forest fire prediction since the ease of ignition varies with fuel type (see [28]). For example, a mature Douglas Fir forest would require drier conditions for a longer period of time than an immature Pine forest for a fire to start. Fuel type within a cell is incorporated into the model by using primary fuel type to define a zone. Since models were fitted by zone, the difference in models from zone to zone may account for the different effect of various fuel types.

(3.) Changes in elevation result in changes in weather conditions, and vegetation. The effect that elevation has on forest fire occurrence is some-

what accounted for by the weather data. **Topography** and **elevation** are taken into consideration by the definition of major climate type which was used to define a zone. See Section 1.3.1. Pairing fuel type and major climate type to define a zone eliminates the confounding effect of fuel type with topography and elevation. As with fuel type above, the models from zone to zone account for changes in the effect of major climate type of fire occurrence.

(4.) **Public awareness of fire danger** is difficult to assess. No measure of the influence of advertisement of forest fire danger during high risk periods or of campfire safety campaigns is included in the model. The Danger Class rating is used to display roadside fire danger signs the public sees but the Danger Class is only a vague connection to public awareness. The Danger Class has no predictive value, as noted in Section 1.1. Danger classes III and IV were combined and a four level indicator was included in the Poisson and logistic models. Combining levels of an indicator variable is discussed in [8]. The Danger Class did not affect the deviance significantly ($p \gg 0.1$).

(5.) **Land use patterns** refer to any way in which the forests are used by people and how that usage changes over time and space. No information on land use patterns is maintained, so it is not known where, when, or how many recreationalists, loggers, sportsmen, *etc.* are in a forest. Land use pattern information would be very difficult to obtain accurately. The phenomenon of the presence of people in a forest increasing the likelihood of a fire could possibly be accounted for by including variables 6.-11.

(6.) Martell, [9], suggested accounting for land use patterns by incorpo-

rating **day of the week** or weekday and weekend trends in a people-caused fire occurrence model. Section 3.1 discusses the presence of these trends for some general causes. An indicator variable for weekday and weekend was included in the model to account for these differences. However, for reasons explained below, cause was not included in the models so the weekday/weekend indicator variable did not produce a significant reduction in the deviance ($\Delta D \simeq 0.1$) and was not included in the final models.

(7.) Changes in patterns in recreation over time and in forest vegetation may be reflected by the trends in the number of people-caused forest fires over the fire season. For example, many people-caused fires occur in the summer months when recreation activity is increased and the forest vegetation is dry and past the green budding stage. **Day of the fire season** may model the seasonal variation in vegetation in terms of when the vegetation is dry and when it is green and fresh. This variation in forest vegetation over the season may be accounted for by including the weather data in the models. Day of the fire season may also act as a proxy for land use patterns such as recreation and logging activity when included in a prediction model. See the discussion on variation in the number of fires throughout the fire season in Section 3.1.

Martell accounted for the trend in the number of forest fires over the fire season in two ways. Firstly, he fitted a different logistic model to different stages of the fire season called subseasons. This approach was also recommended by Kourtz. (See [17] and [27].) From Figure 3.2 and knowledge of events that will change recreational activity such as the end and beginning of the school year and hunting season, the subseasons in

Subseason	Days	Dates
Early Spring	1-44	Apr. 15 - May 27
Late Spring	45-79	May 28 - Jul. 1
Spring	1-79	Apr. 15 - Jul. 1
Summer	80-135	Jul. 2 - Aug. 26
Fall	136-186	Aug. 27 - Oct. 15

Table 3.4: Division of Fire Season into Subseasons

Table 3.4 were selected for Kamloops Region data. The subseason models can then be combined to give one model for the whole fire season. Martell included sinusoidal Fourier series terms in one logistic regression model for the whole season [18]. The number of people-caused forest fires in the Kamloops Region does not appear to exhibit symmetric or periodic sinusoidal properties as seen in Figure 3.2. The term

$$\sin\left[\frac{1.5\pi(k-190)}{93}\right] - \cos\left[\frac{1.5\pi(k-190)}{186}\right] \tag{3.2}$$

where k is the day of the fire season which is 186 days long mimics the pattern in Figure 3.2 and was fitted to zone 63-C7 data, but with no significance. See Chapter 4.2 for a brief discussion on the subseason and Fourier term approaches for zone 63-C7 *Methods A* and *C* models. See [3] for a discussion on including Fourier series terms in a regression model to account for periodic trends.

(8.) Whether or not a particular cell within a zone contains a **road, lake, campground or city** were included as indicator variables in the

common *Method C* backwards elimination process. The existence of a road, lake, campground or city within a cell is used as a surrogate for daily land use data. The Ministry of Forests does not have a geographic database to provide this information. The existence of a road, lake, campground or city was determined primitively by laying a grid of the cells of the Kamloops Region over a roadmap. For this information to meaningfully account for landuse patterns it is necessary that the location of roads, lakes, campgrounds and cities be digitalized accurately by cell in the Kamloops Region by a cartographer.

Similarly, the existence of a logging road or logging operation within a cell may be useful to predict people-caused forest fires. This information, both current and historic, needs to be digitalized by cell so it can be incorporated into a people-caused forest fire model. The location of logging operations changes frequently over the fire season and information on historical locations and changes is not available within the Kamloops Region in any format.

(9.), (10.) No information is maintained on **logging closures** or **campfire restrictions** by the Ministry of Forests or by Parks and Recreation. This information may be very useful in predicting people-caused forest fires. The decision for closure is made by a forester on site at a weather station and does not follow a reproduceable algorithm. In order to aid in future predictions and to assess the effectiveness of closures *it is suggested that records of closures and campfire restrictions be kept in the future.*

The Protection Branch classifies a fire by general cause and specific cause. See Tables 3.5 and 3.6 for lists of these causes. All fires with

general cause not *1* are assumed people-caused (as recommended by Protection Branch). Martell [18] suggests breaking up fires by general cause because land use patterns are different for recreational and industrial fires. Table 3.1 show that fires tend to occur at different times of the week depending on their cause classification. Different models for different causes may prove useful. Most people-caused forest fires in the Kamloops Forest Region, unfortunately, are given general cause classification *miscellaneous known* and are then given a specific cause. The specific causes are not representative of landuse patterns, for example a specific cause of smoking could be under any general cause class. Tithecott [26] does not recommend fitting a separate model for each cause. A problem that would arise if models were broken up by specific cause is that, because there are thirty-one specific causes, most would have zero fires, leaving little or no information for model fitting. One of the problems with fitting statistical models to forest fire data is that fires are rare events. Dividing the fires into subclasses would make the events even more rare. Aggregation of cells into zones is used in this thesis as a way of reducing the rarity of fires. Disaggregating fires into causes would have a contrary effect and has thus been avoided. The inclusion of an indicator variable for weekday/weekend as a candidate covariate should cover any significant differences that might exist in the probabilities of industrial and recreational fires. In fact, this variable turned out to be not significant.

0	Unknown
1	Lightning
2	Recreational
3	Railroad
4	Logging and Lumbering
5	Right-of-way Construction
6	Other Industrial Operations
7	Not assigned
8	Land Clearing and Brush Burning
9	Miscellaneous Known

Table 3.5: General Cause Classes

01 Faulty Spark Arrestor	02 Hot Ashes
04 Blasting Operation	05 Brakeshoe Fragments
06 Brush Burn Industrial	07 Brush Burn Domestic
09 Burning Building	10 Burning Vehicle
11 Campfire	12 Camp Stove Exploding
13 Child with Match	14 Garbage Dump Fire
15 Friction Log Line	18 Friction Log Skid
19 Fusee	20 Hang Sawdust Fire
21 Hang Log Slash	22 Hang Millsite Landing
23 Hang Other	24 Incendiary
25 Lightning	26 Logging Machine Operation
27 Mill Burner	28 Mill Burner Open
29 Millsite Debris Fire	30 Powersaw Exhaust
32 Powersaw Other	34 R/R Tie & Brush Burn
35 Range Burning	36 Refuse Burn Domestic
37 Refuse Burn Industrial	38 Short Circuit Power Line
40 Slash Burn	41 Slash Burn
42 Smoker	45 No Spark Arrestor
46 Spont. Combustion	47 Sun Rays Focusing
50 Sparks Other	51 Permit Escape
53 Welding Operations	54 Sparks Other
55 Landing Debris Fire	96 Agency Known Only
97 Misc. Known	99 Unkown

Table 3.6: Specific Cause Classes

Part II

Modelling and Predicting

Chapter 4

Models

The models fitted to zone 63-C7 are discussed in this chapter. Models were fitted to data from 1985 to 1990 in zone 63-C7. The covariates to include were selected by backwards elimination. All models were fitted using GLIM [2]. As described in Section 1.3.2, the data were in formats *aggregated* where weather data is averaged and forest fires are totalled across the zone, and *common* where one model is fitted to all cells in the zone but the data are not averaged. The Poisson model fitted to the aggregated data is called *Method AP* and the Poisson and logistic common models are called *Methods CP* and *CL*. See Appendix A to review the three descriptions of the model types developed in this thesis. The resulting models were then used to predict 1991 fires. Chapter 5 discusses these predictions.

Some results from model fitting that were not retained in the final models are presented in Section 4.1. The final models are given in Section 4.2.

4.1 Modelling Attempts

4.1.1 Models for Individual Cells

As well as fitting the models to aggregated and common data, a separate logistic model was fitted to each cell in zone 63-C7. Some cells had very few or no fires and in these cases the iterative procedure for parameter estimation did not always converge. The reason for this is, of course,

that with zero fires the maximum likelihood estimate of λ_{ij} is zero which requires at least some of the estimates of the β's to assume a value of negative infinity. In practise GLIM produces larger negative values for the estimates at each iteration until the maximum number of iterations is reached. The resulting deviance is approximately correct. See [2].

An alternative approach to backwards elimination was to fit only FFMC and only FFMC* to each cell with some fires and select a "best" model from these two for each cell based on smallest deviance. Parameter estimates differed for each cell.

Fitting a model to a single cell was tedious and did not appear to provide any benefits. With so few fires occurring in an individual cell there is too little information to adequately fit a regression model. This opinion concurs with that of Martell who recommended not analysing data at the cell level (personal communication). Another drawback to fitting models to single cells is that there are approximately 4500 cells in the whole province of B.C. for which model selection would have to be performed. It was decided in view of these statistical and practical difficulties not to pursue further the modelling of people-caused forest fires on the cell by cell level; rather it was decided to use the aggregated and common Poisson models *AP* and *CP* and the common logistic model *CL* described in Section 1.3.2 (and Appendix A) and discussed further in Section 4.2.

4.1.2 Models for Seasonal Variation

This section describes the results of attempts made to follow Martell's lead in incorporating day of the fire season in the fire occurrence models (see

Chapter 1) on aggregated and common data. The subseason and Fourier term approaches given by Martell were tried here to see whether there was any seasonal effect, although none was apparent from a preliminary analysis of the B.C. data. The subseason and Fourier models were tried for reasons of completeness in mimicking Martell's work on B.C. forest fire data. Since no seasonal effect was apparent in the prelimanary data analysis, it was thought that the seasonal models would only be pursued if they did indeed outperform models with no effect of seasonality included.

The fire season was divided into subseasons as shown in Table 3.4 in Section 3.1. The backwards elimination process starting with the six weather indices and wind speed resulted in aggregated and common Poisson models for each of the five subseasons. For both aggregated (A) and common (C) data, models were fitted to the subseasons *Early Spring*, *Late Spring*, *Spring*, and *Summer*, and *Fall*. The four aggregated subseason models were incorporated to provide an aggregated model for the whole season. Similarly, a common model was constructed from the incorporation of the four common subseason models. Again, on the same rationalization, incorporating subseason models for *Spring*, *Summer* and *Fall* provided yet another *Method AP* model and *CP* model for the whole fire season. The deviances of the subseason models were added together to give a deviance for the whole fire season. The subseason models were compared to the deviances of the *Methods AP* and *CP* final models fitted to the whole fire season in Section 4.2. The degrees of freedom for the comparison were the change in the number of parameters from the subseason to the whole season model. The subseason models had significantly (p-values smaller than .05) smaller

deviances except for the *Method AP* Spring-Summer-Fall subseason model which had a larger deviance. However, the prediction statistics, predicted deviance and the two prediction sums of squares statistics, were larger for the subseason aggregated and common Poisson models (*AP, CP*) than for the aggregated and common Poisson models for the whole season of Section 4.2.

A second attempt to account for the difference in the number of people-caused forest fires over the fire season was made by adding Fourier series terms similar to Martell's to the models in Section 4.2 for the aggregated and common data. The Fourier terms added to the models are as follows:

$$\sin\left[\frac{2j\pi}{93}\right] \quad \text{and} \quad \cos\left[\frac{2j\pi}{93}\right]$$
$$\sin\left[\frac{2k\pi}{186}\right] \quad \text{and} \quad \cos\left[\frac{2k\pi}{186}\right]$$

where k is the day of the fire season and j is the day of the fire season for the first 93 days and is the day of the season minus 93 beyond that. These terms are similar to the terms used by Martell to model any sinusoidal trend in the number of fires over the whole fire season (hence denominator of 186 in the sine and cosine arguments) or over half of the fire season (hence denominator of 93 in the sine and cosine arguments). The sine and cosine terms in Equation 3.2 in Section 3.2 were also added. None of these sine and cosine terms were significant when added to the aggregated *AP* model or the common *CP* model in Section 4.2. No further attempts were made to account for the variation in number of fires over the fire season.

4.2 Fitted Models

In the initial attempts at model fitting, before the question of how to divide the Kamloops Region into zones was resolved, it was found, by backwards elimination, that the fire weather indices were selected over the raw weather variables. The wind speed measure was, however, retained in some instances. All research by Martell, Kourtz, and Haines has been done using only the fire weather indices [14], [18], [27]. Therefore, once the present zones were decided upon, the process of backwards elimination was begun with only the six fire weather indices and wind speed as candidate regressors.

The fitted models are presented in Table 4.1 and the deviances of these models are listed in Table 4.2. The "best" covariates for the aggregated Poisson regression model were the indices FFMC*, DC, BUI, and wind speed. These weather measures were also the selected covariates for both the common Poisson and logistic methods. Indicator variables for whether or not a cell contained a *road* or a *lake* were also significant in the common *Method CP* and *CL* models. There are no parks in zone 63-C7. The presence of a *city* (Penticton) was included but was not significant. The adjusted T overdispersion statistics for the *Methods AP* and *CP* regressions were 1.212 and 1.428, respectively with corresponding p-values of 0.11 and 0.08. Thus there is weak but not conclusive evidence of overdispersion. The normal quantile plot for *Method CP* is shown in Figure 4.1 and adds to the evidence of overdispersion. The points scattered off to the far right on the graph are residuals from days where fires occurred. The normal quantile plots for the other models are similar to that in Figure 4.1 so are

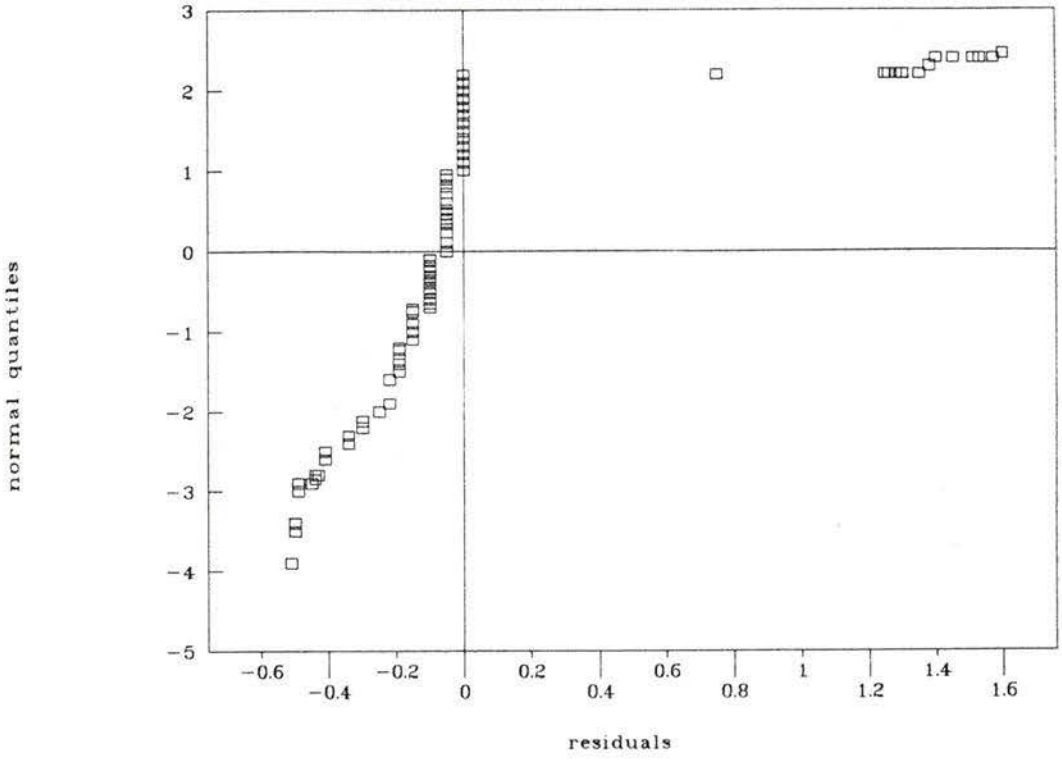


Figure 4.1: Normal Quantile Plot of *Method CP* model

not all shown here. The PRESS statistic is smallest for the aggregated data at 448 for the *Method AP* models compared to 2943 and 3632 for Martell's aggregated models. The PRESS statistic for *Method CP* is 4638 and for *CL* and Martell's common models is much higher.

From the residual plots of the Anscombe residuals against $2\sqrt{\hat{\lambda}_i}$, where $\hat{\lambda}_i$ are the fitted values, there appears to be a slight pattern indicating residuals are decreasing as $2\sqrt{\hat{\lambda}_i}$ increases. These is no evidence in the residual plots of instability in the variance. No examples of Anscombe or Pearson residual plots are shown as these plots are not interesting with only zero s and ones as responses.

Method	AP	CP	CL
Intercept	-3.263	-6.392	-6.456
	(.350)	(.405)	(.413)
FFMC*	1.760	1.888	1.887
	(.561)	(.538)	(.556)
DC	-.001	-.001	-.001
	(.0007)	(.0007)	(.0007)
BUI	.009	.008	.008
	(.003)	(.002)	(.003)
wind speed	.038	.024	.022
	(.018)	(.013)	(.013)
road	–	.580	.561
		(.299)	(.301)
lake	–	1.448	1.461
		(.200)	(.207)

Table 4.1: Fitted Models for zone 63-C7: parameter estimates (with standard errors in brackets).

Method	Deviance	Degrees of Freedom
AP	499.7	887
CP	940.7	7972
CL	1152	7972

Table 4.2: Deviances for Fitted Models in zone 63-C7.

Chapter 5

People-caused Forest Fire Predictions

This chapter discusses the predictions made for the 1991 fire season in zone 63-C7 using the final models from Chapter 4. These predictions are compared to predictions made by Martell's models fitted to zone 63-C7 data and to Kourtz's predictions. Data in the form of *Method A* (aggregated) and *Method C* (common) were both used to fit Martell's models. The fewer number of multiple fire days in *Method C* data may accomodate Martell's logistic regression better than the *Method A* data. The deviances of Martell's models with covariates FFMC and BUI (referred to by the letter M) and his models with the four sine and four cosine terms added (referred to by the letter F) are given in Table 5.1. The Fourier terms are not significant ($p = .12$ and $p = 1.0$) in either of his aggregated or common logistic models. However, Martell's models with Fourier terms were included in the comparisons of prediction performance, for the sake of completeness.

Predictions were calculated from the three models AP, CP and CL from Chapter 4 for the number of people-caused forest fires in zone 63-C7 per day of the 1991 fire season. The predictions were calculated by first substituting the 1991 weather data into the fitted models, then taking the inverse of the link function. The predicted expected number of fires on day i in zone 63-C7 from the aggregated Poisson model, AP, and in a cell of

Model	deviance	Degrees of Freedom
AM	649.7	889
AF	636.8	881
CM	1235	7976
CF	1235	7968

Table 5.1: Deviances for Martell's models fitted to zone 63-C7

the zone for the common Poisson model, CP, is then

$$\hat{\lambda}_i = \exp\left(\sum_{j=1}^{p} x_{ij}\hat{\beta}_j\right)$$

where p is the number of regressors (4 for *Method AP* and 6 for *Method CP*). The predicted expected number of fires on day i for a cell in zone 63-C7 from the common logistic regression CL is

$$\hat{p}_i = \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\hat{\beta}_j\right)}{1+\exp\left(\sum_{j=1}^{p} x_{ij}\hat{\beta}_j\right)}.$$

The predictions from all *Method C* models are summed over all nine cells to give a daily prediction for the whole zone.

The predictions, ideally, should be small on days when a fire did not occur and large on days when fires occurred. The sum of squares statistic S_1 in Equation 2.10 and its standardized version S_2 in Equation 2.11 from Section 2.2.5 will be relatively large if the prediction is large on a day when a fire did not occur. These statistics will be relatively large if the prediction is small on a day when a fire occurred. The statistic S_2 will be a bit larger than S_1 for small predictions on days when a fire occurrs. Small values

of these two statistics will indicate a prediction model that not only has small predictions on non-fire days, but also has increased predictions on fire days.

The predictions generated from the *Methods AP, CP,* and *CL* models and Martell's four models *AM, AF, CM,* and *CF* are compared with Kourtz's predictions, currently in use by Protection Branch. The prediction statistics are given in Table 5.2. All of the models developed in this thesis predict better than the predictions provided by Kourtz's model, according to statistic S_1. All of Kourtz's predictions are either zero or one so the standardized S_2 statistic can not be calculated. The models developed here, *AP, CP,* and *CL*, have S_1 values similar to one another as well as to Martell's four *AM, AF, CM,* and *CF* models; so we compare prediction deviance and S_2 values to compare prediction performance. The aggregated Poisson model (*AP*) predicts better than Martell's models on the aggregated and the common data according to the standardized sum of squares S_2 criterion. The predicted deviance for Martell's common model with Fourier terms (*CF*) is slightly smaller than for the aggregated Poisson *AP* model, but since S_2 is considerably larger for *CF* (and also because the Fourier terms were not significant) we cannot draw any conclusions in favour of model *CF*.

The common *Method C* models also predict better than Martell's four models, according to the predition statistics. The *CP* model does somewhat better than the logistic model *CL* according to all three criteria. Thus the *Method CP* common Poisson model appears to be the best of the eight models in Table 5.2 in terms of its prediction performance, although the

Method	Pred.Dev.	S_1	S_2
AP	96.5	23.0	186.1
AM	119.2	23.23	214.2
AF	171.1	24.26	406.4
CP	93.9	22.49	174.0
CL	94.8	22.67	189.2
CM	97.2	23.25	269.3
CF	94.8	23.04	233.5
Kourtz	–	36	–

Table 5.2: Prediction Statistics

Method CL common logistic model follows closely in performance. The fact that there is not strong evidence of overdispersion for the Poisson model lends to its credibility (see Section 4.2).

Graphs of the actual number of fires and the predictions for each week of the fire season from the seven models and Kourtz's predictions are given in Figures 5.1 through 5.8. Weekly, as opposed to daily, predictions are shown for purposes of illustration. All fitted models provide a similar general trend in the shape of the graph of predictions. Note that *Method AP* aggregated Poisson (and to a lesser extent the common Poisson model and Martell's aggregated model without the Fourier terms AM) shows an increase in fire probability at the end of the fire season when two fires occurred after twenty or so days of no fires. The graphs for Martell's models are smoother than those for the Poisson aggregated and common

Methods AP and *CP*. Hence, the Poisson models developed in this thesis are more sensitive to changes in the weather than Martell's three stage procedure that employs logistic regression.

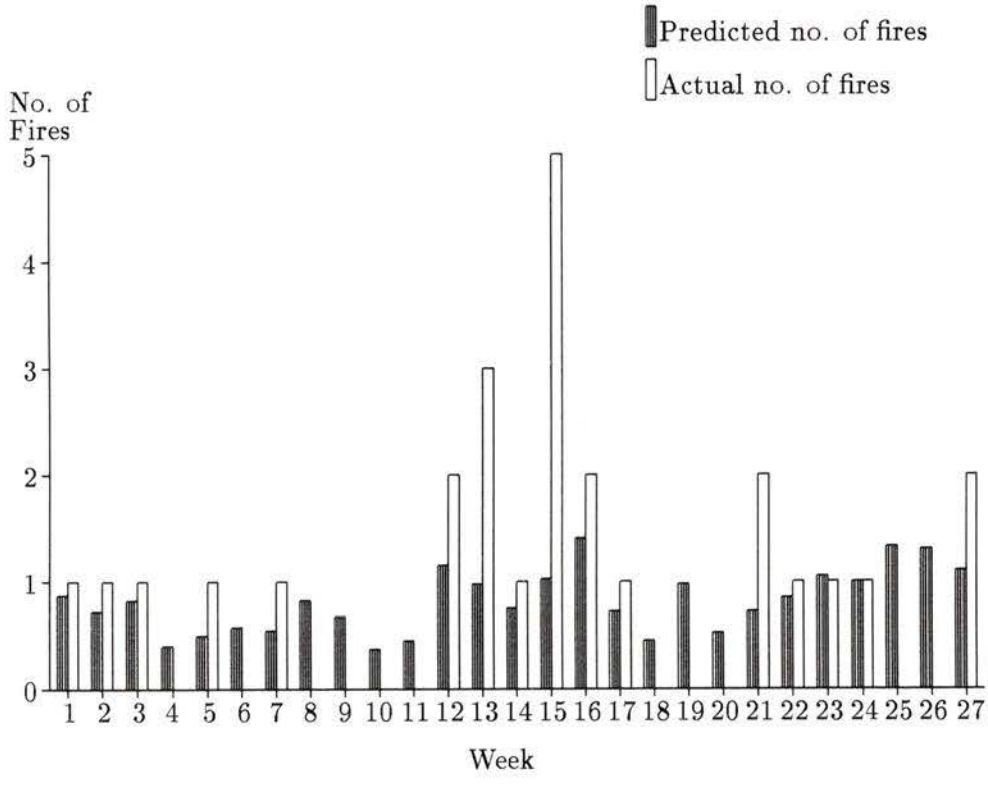


Figure 5.1: *Method AP* aggregated Poisson, zone 63-C7 Weekly Predictions for 1991.

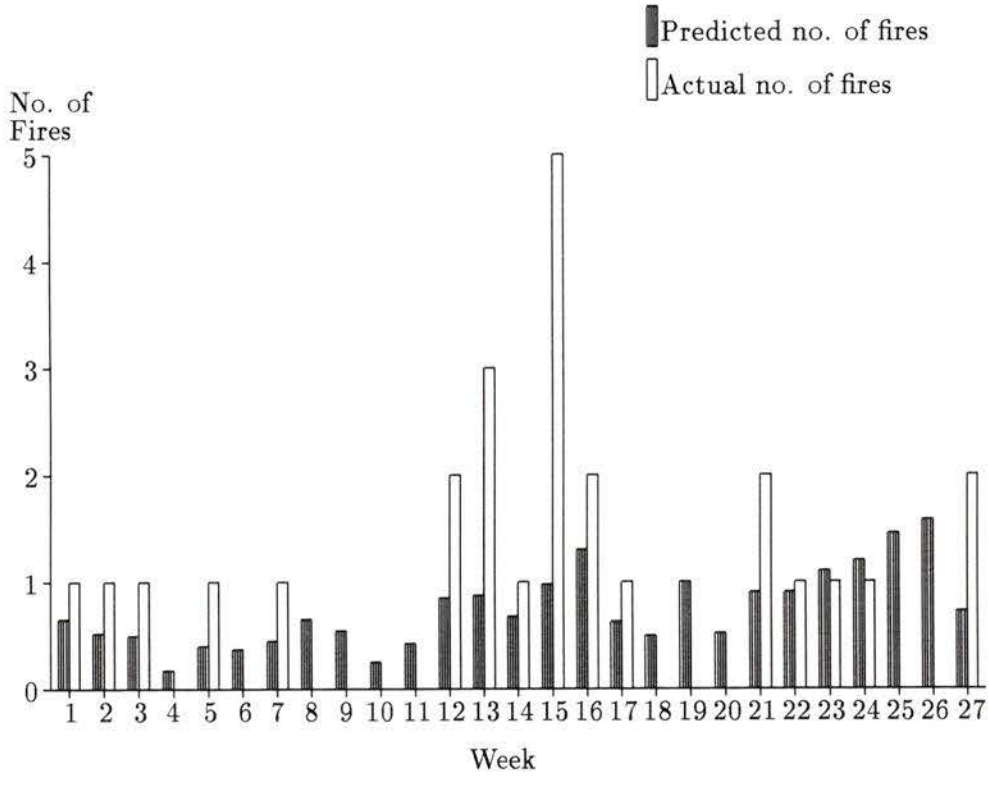


Figure 5.2: Martell's *AM* aggregated FFMC+BUI, zone 63-C7 Weekly Predictions for 1991.

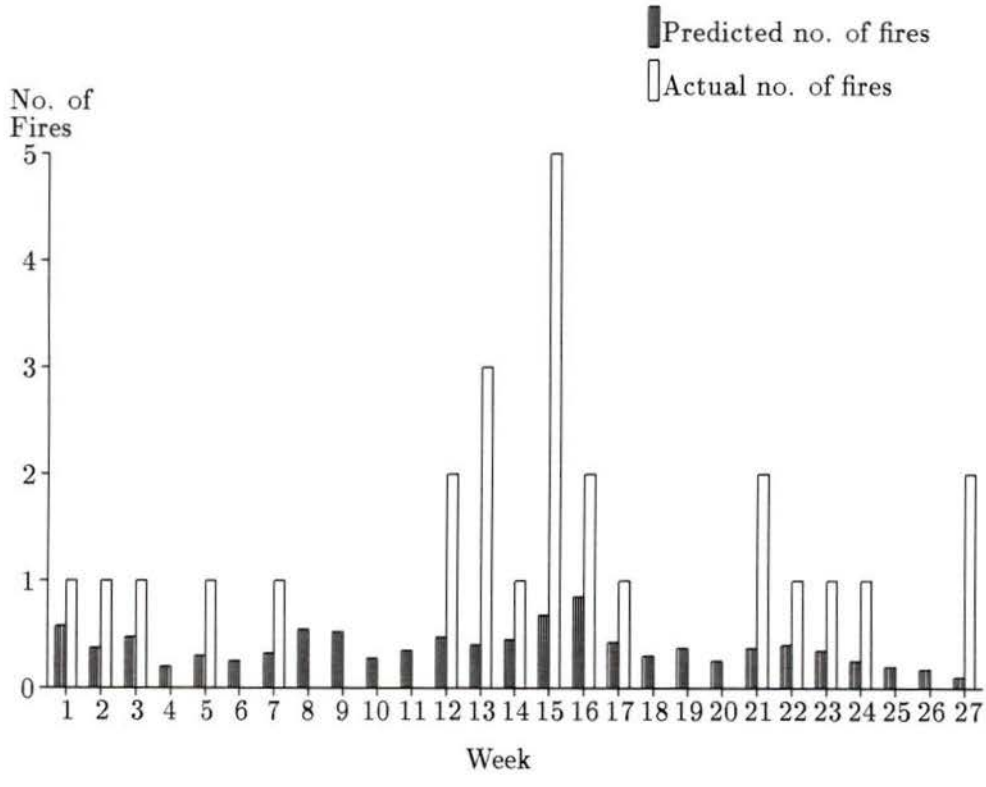


Figure 5.3: Martell's AF aggregated Fourier, zone 63-C7 Weekly Predictions for 1991.

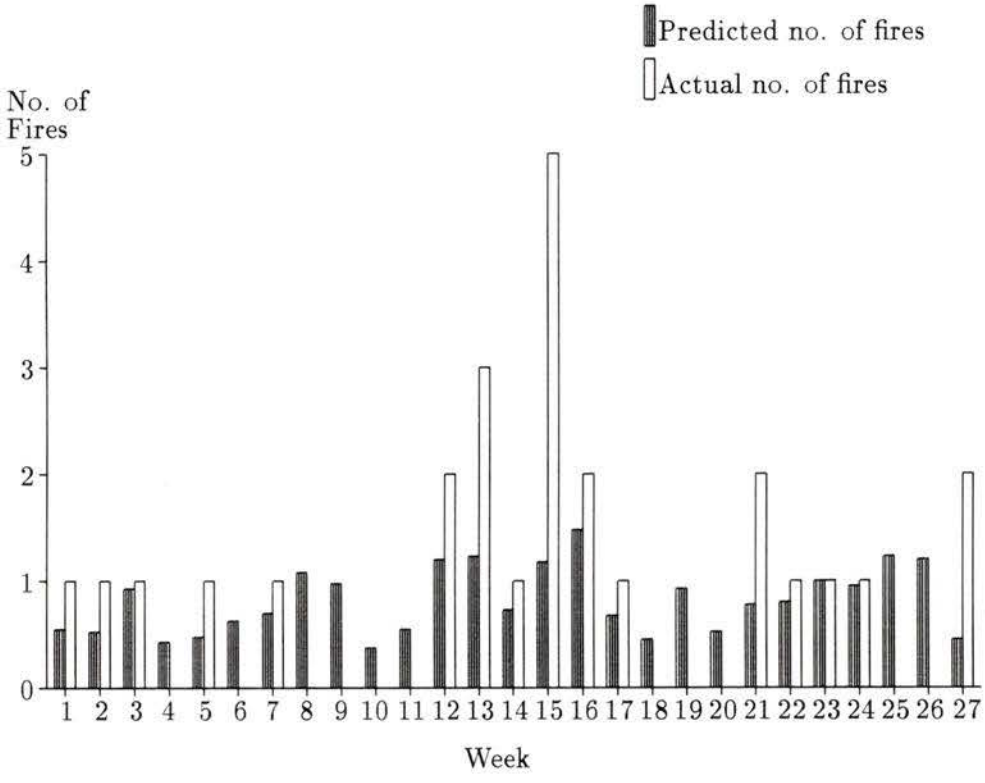


Figure 5.4: *Method CP* common Poisson, zone 63-C7 Weekly Predictions for 1991.

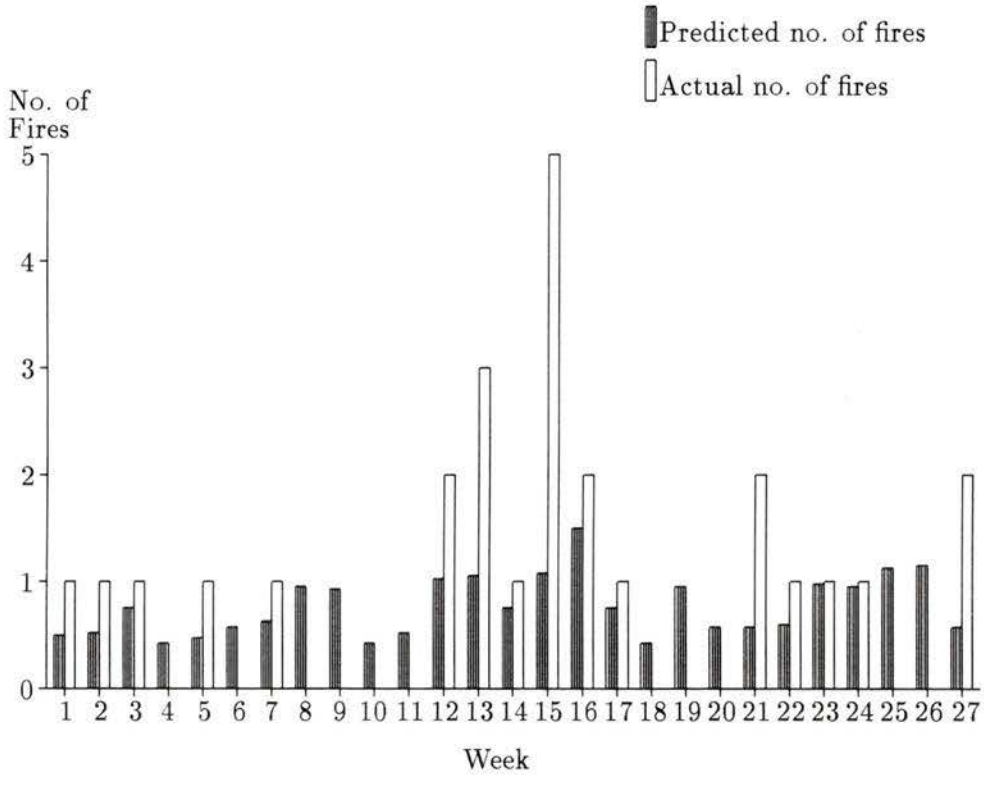


Figure 5.5: *Method CL* common logistic, zone 63-C7 Weekly Predictions for 1991.

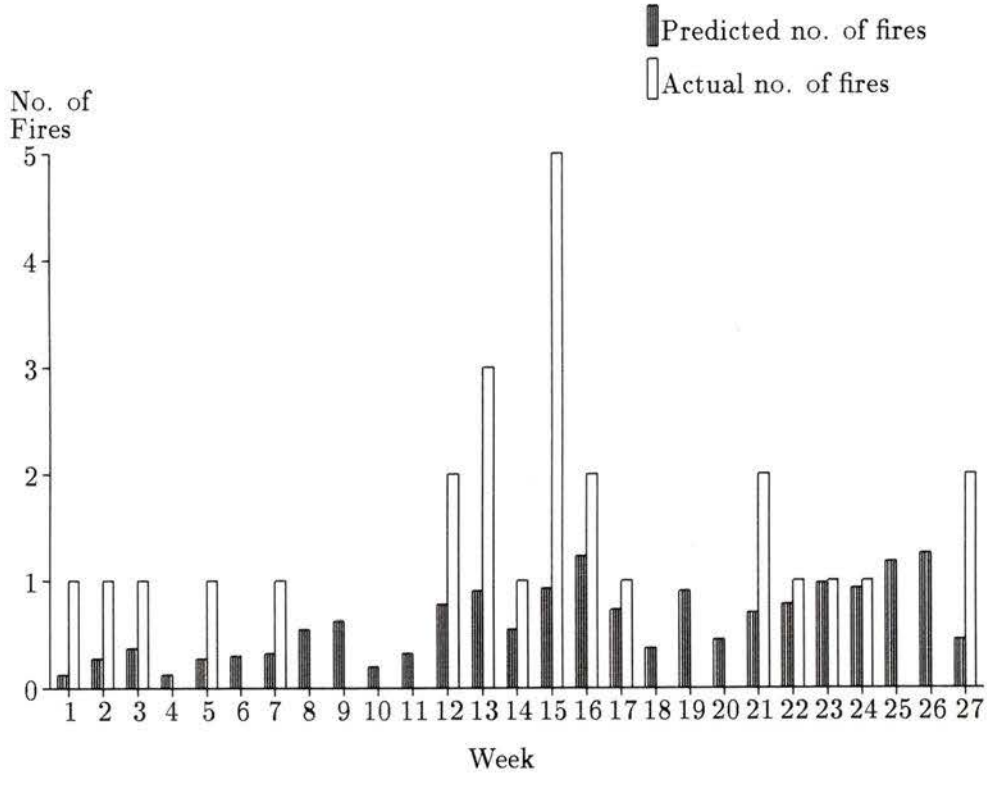


Figure 5.6: Martell's *CM* common FFMC+BUI, zone 63-C7 Weekly Predictions for 1991.

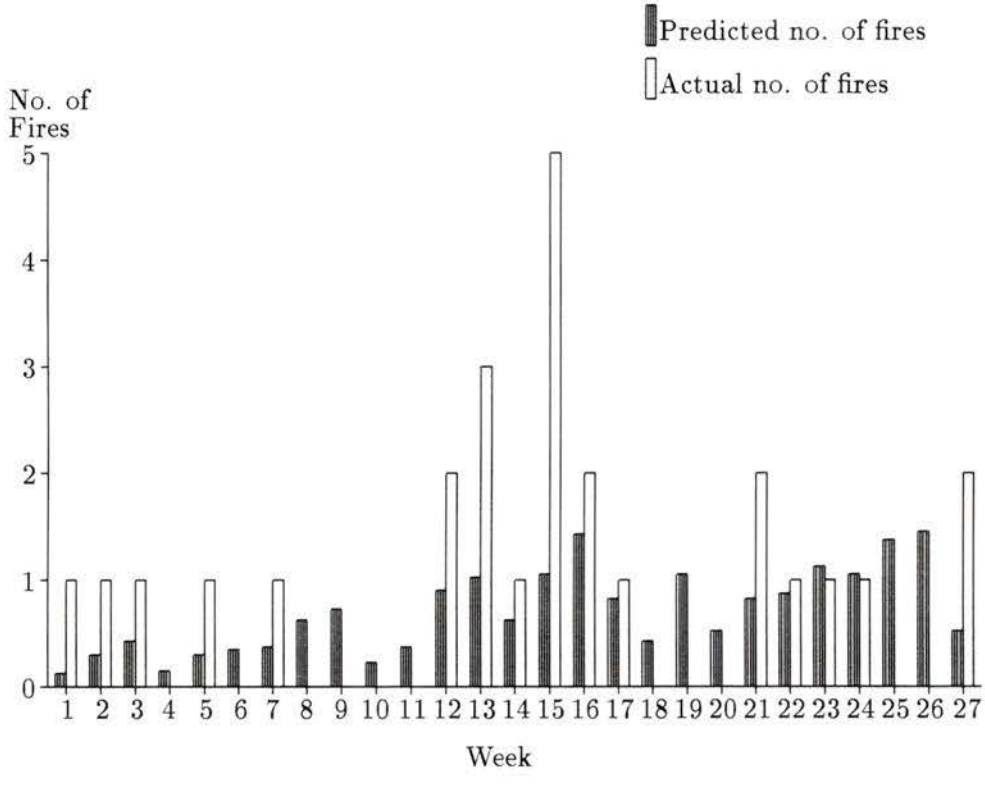


Figure 5.7: Martell's *CF* common Fourier, zone 63-C7 Weekly Predictions for 1991.

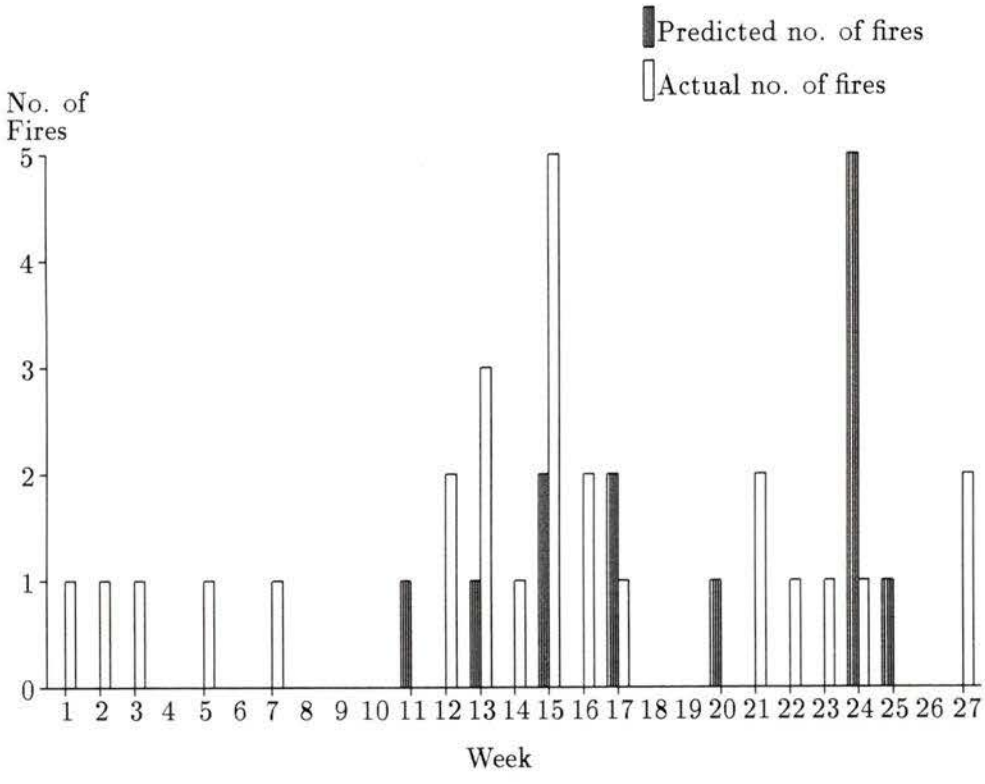


Figure 5.8: *Kourtz's*, zone 63-C7 Weekly Predictions for 1991.

Method	S_1	S_2
AP	32.026	33.623
AM	37.618	46.555
AF	38.23	53.756
CP	31.173	31.563
CL	22.668	189.2
CM	23.256	269.3
CF	23.039	233.5
Kourtz	39	–

Table 5.3: S_1 and S_2 values for Weekly Forest Fire Predictions

The predictions for the Poisson models *AP* and *CP*, logistic model *CL*, Martell's aggregated and common models with and without Fourier terms, *AM, AF, CM,* and *CF,* and that for Kourtz were aggregated into weekly forest fire predictions for zone 63-C7. The two sum of squares statistics, S_1 and S_2, were calculated for the weekly predictions and are shown in Table 5.3 Again, the *Method C* Poisson regression appears to give the best prediction performance since it gives the smallest value for S_2 and one of the smallest values for S_1.

Ideally, most of the predictions should be small since most days have zero fires; in addition, the prediction when a fire actually does occur should be relatively large. Figures 5.1 through 5.8 show that relative to Martell's models, the aggregated and common Poisson models produce many small predictions with fires occurring more when predictions are relatively large.

These graphs and statistics suggest that the Poisson models developed here perform better than Martell's models applied to B.C. data, and than Kourtz's prediction method currently implemented by Protection Branch.

The graph in Figure 5.9 shows the boxplots of the predictions for the different modelling methods *only on days when fires actually occurred.* The endpoints of the plots represent the minimum and maximum predicted values and the "box" represents the interquartile range of the predictions. The minimum and endpoints of the interquartile range for Kourtz's predictions are all zero with a maximum of one. The actual number of fires also has a minimum and endpoints of its interquartile range equal to zero with a maximum of five.

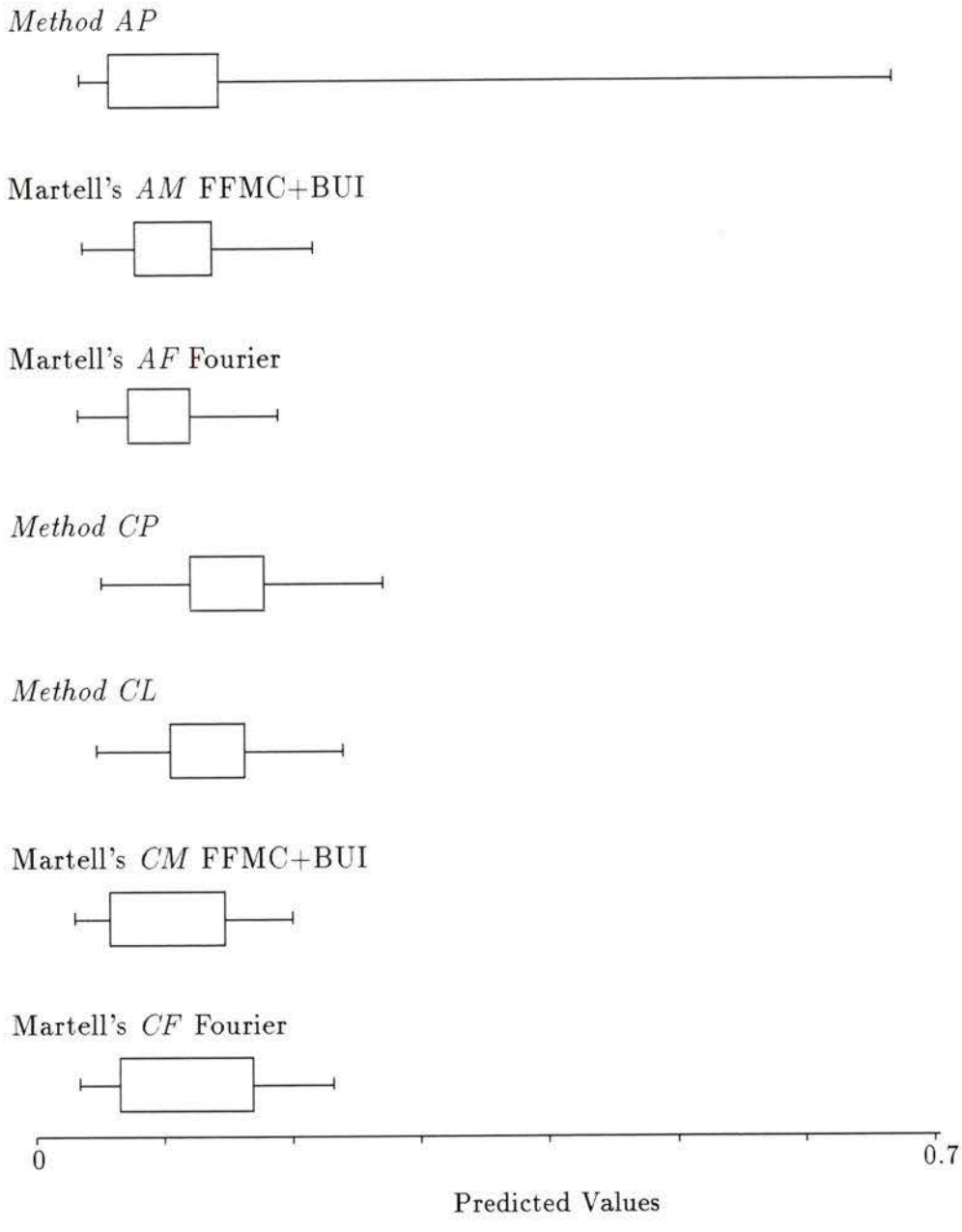


Figure 5.9: Boxplots of Daily Predictions for zone 63-C7 1991.

Chapter 6
Model Portability

The portability of the three models, aggregated Poisson *AP*, common Poisson *CP* and common logistic *CL* fitted to zone 63-C7 was investigated by fitting the covariates that were selected for zone 63-C7 from Chapter 4 to three other zones in the Kamloops Region. See Figure 1.2 for a display of zone 63-C7 and the three other zones. These three other zones differ from zone 63-C7 in climate, fuel type, and land use patterns. The covariates from Chapter 4 were fitted and any parameters that were not significant were removed by backwards elimination for each of the three zones. Secondly, models were also selected for each zone by the backwards elimination process starting with all six weather indeces and wind speed. In addition, indicator variables for each of *road*, *city*, *lake*, or *park* were included in the common *Method C* models. The final *Methods AP, CP,* and *CL* models were chosen on the basis of smallest deviance. The final models for the three new zones have some combination of the weather indices FFMC, BUI, DC, the same indices used in the zone 63-C7 models.

6.1 Predictions in Zone 6-O1

Zone 6-O1 is in the centre of the Kamloops Region and contains the city of Kamloops. It is located in major climate type *central – very dry* and has primary fuel type *O1* which is grass. The zone comprises thirteen cells.

There were one hundred and thirty-nine fires in this zone from 1985 to 1990. The mean number of people-caused fires in this zone per day was 0.125 with a variance of 0.148.

The final fitted models are shown in Table 6.1. The deviances of the models are in Table 6.2. When the zone 63-C7 models were fitted to zone 6-O1 data from 1985 to 1990, wind speed was no longer significant in any model ($p = 0.17$ for *Method AP* and $p = 1.0$ for *Methods CP* and *CL*). Neither indeces BUI nor DC were significant with both in the model yet each were significant when included in the model individually. Hence, analysis-of-deviance was used and selected BUI over DC. The analysis-of-deviance table is shown in Table 6.3. Indicator variables for *road*, *city*, and *lake* were included in the common models and were all significant; there is no park in this zone. Note that the estimated coefficients for *lake* and *city* are negative. The transformation FFMC* was chosen over FFMC in all models since the FFMC* gave a smaller deviance (a difference in deviance of 21 for the common models and 8 for the aggregated model). The overdispersion statistics for the final Poisson models fitted to aggregated AP and common CP are 1.754 and 9.277 respectively, both giving evidence of overdispersion ($p = .039715$ and $p < 1 \times 10^{-4}$). Martell's aggregated models AM and AF and common models CM and CF were also fitted to zone 6-O1. The PRESS statistic is smaller for the aggregated Poisson model (448.6) than for Martell's aggregated models (625.6 and 644.4).

The models were then used to predict the 1991 people-caused forest fires in zone 6-O1. The prediction statistics are in Table 6.4. There were fifty fires in this zone in 1991. The predictions generated from these models

	Method AP	Method CP	Method CL
Intercept	-3.278	-5.652	-5.638
	(.254)	(.269)	(.275)
FFMC*	2.540	2.413	2.348
	(.513)	(.430)	(.445)
BUI	.0028	–	–
	(.0019)		
road	–	.453	.383
		(.200)	(.207)
lake	–	-1.434	-1.334
		(.358)	(.358)
city	–	-.728	-.626
		(.366)	(.370)

Table 6.1: Fitted Models for zone 6-O1 (standard errors in brackets).

Method	Deviance	d.f.
AP	521	889
AM	665.3	889
AF	660.5	881
CP	1234	11334
CL	1405	11334
CM	1435.9	11336
CF	1422.3	11328

Table 6.2: Deviances of All Models in zone 6-O1.

Source	ΔD	d.f.
DC	0.4658	1
BUI	20.22	1
Source	ΔD	d.f.
BUI	12.74	1
DC	7.948	1

Table 6.3: Analysis-of-deviance Table for DC and BUI in zone 6-O1

were compared with the predictions made by Kourtz. Kourtz predicted twenty-five fires, three of which are on days when a fire occurred in the zone. (Although, Kourtz's predictions were not in the same cell as the one in which the fire occurred.) All models developed in this thesis and Martell's models give smaller prediction statistics than Kourtz's model does. Martell's models for *Method C* common data do not predict well relative to the other models here. However, the common Poisson and logistic models perform very well. See recommendations in Chapter 7.

Based on predicted deviance and the standardized sum of squares statistic, S_2, Martell's aggregated AM and AF models predict better than the AP aggregated Poisson model. However, for the (non-standardized) sum of squares statistic, S_1, Martell's $AM$and AF models do no better than the *Method AP* model.

The common method predictions give smaller statistics than all of the aggregated models and Kourtz's model. The common logistic CL model gives small predicted deviance relative to the common Poisson CP and a similar sum of squares indicating that the logistic model might provide better predictions. However, based on the statistic S_2, the common Poisson model provides better predictions than the common logistic. In summary, the *Method C* Poisson and logistic models appear to perform better than any other models for these data.

6.2 Predictions in Zone 21-C2X

Zone 21-C2X is in the north of the Kamloops Region and overlaps onto Wells Gray Park. The zone is located in major climate type *northern wet*

Method	Pred.Dev.	S_1	S_2
AP	107.2	86.66	932.0
AM	104.5	86.28	842.0
AF	105.1	86.41	874.8
CP	102.1	85.38	751.1
CL	88.9	85.46	760.0
CM	208.4	94.96	1052
CF	212.36	95.03	1149
Kourtz	–	190	–

Table 6.4: Prediction Statistics for 1991 Predictions in zone 6-O1.

and has a primary fuel type of Boreal Spruce with code *C2*. The zone has eleven contiguous cells. The mean number of people-caused forest fires per day in this zone from 1985 to 1990 was 0.032 with a variance of .038. There were thiry-six people-caused fires reported in this time period.

There is an agreement with Wells Gray Park that the Ministry of Forests will fight forest fires in the park and maintain data on these fires. The Ministry of Forests does not fight all fires in Wells Gray Park, and only maintains data on those fires in which they are involved. The models developed here, therefore, may not be accurate, because of this limitation in the data.

The fitted models are shown in Table 6.5. The deviances of the fitted models, including Martell's models are all given in Table 6.6. From the covariates FFMC*, BUI, DC, and wind speed selected for zone 63-C7, it

	Method AP	Method CP	Method CL
Intercept	-5.680	-7.152	-7.058
	(.554)	(.510)	(.505)
FFMC*	3.986	2.800	2.659
	(.786)	(.773)	(.771)
DC	.005	.005	.004
	(.001)	(.001)	(.001)
park	–	-1.689	-1.643
		(.377)	(.377)

Table 6.5: Fitted Models for zone 21-C2X (standard errors in brackets).

was found that BUI did not contribute significantly to the models ($p = 1.0$) for zone 21-C2X. An indicator variable was added for *park* and for *road* in the common models. The indicator for *road* was not significant ($p = .9$) and that for *park* was significant ($p < 1 \times 10^{-4}$). The negative coefficient for *park* may be due to the fire fighting agreement with Wells Gray or simply that more people-caused fires do occur outside the park than inside though this seems somewhat counter-intuitive. The PRESS statistic is smaller for the aggregated Poisson model (452) than for Martell's aggregated AM and AF models (2428 and 2254).

The overdispersion statistic for the Poisson models AP and CP respectively are 1.933 and 1.814 suggesting some overdispersion ($p = .02668$ and .0348). The normal quantile plot of the residuals is slightly curved adding to the evidence of overdispersion.

Method	Deviance	d.f.
AP	209.4	881
AM	253.6	882
AF	227.3	874
CP	328.6	9625
CL	385.2	9625
CM	439.1	9625
CF	412.9	9617

Table 6.6: Deviances of All Models in zone 21-C2X.

Predictions of people-caused forest fires in zone 21-C2X for 1991 were calculated for the aggregated Poisson *AP*, common Poisson *CP*, common logistic *CL* and Martell's models The prediction statistics are given in Table 6.7 for the seven models and for Kourtz's predictions. There were no fires in this zone in 1991. Kourtz predicted four fires in this zone in 1991. All models developed in this thesis and Martell's models provide better predictions than Kourtz, according to the prediction statistics. However, according to all prediction statistics, Martell's models provide the best predictions in this zone. Of the models developed in this thesis, the aggregated Poisson model gives the best predictions.

6.3 Predictions in Zone 21-C2Y

Zone 21-C2Y is just south of zone 21-C2X and has the same major climate type and primary fuel type as zone 21-C2X. Zones 21-C2X and Y can be

Method	Pred.Dev.	S_1	S_2
AP	5.602	.349	5.602
AM	4.463	.178	4.463
AF	6.776	.699	6.776
CP	6.661	.390	6.661
CL	6.661	.390	6.661
CM	4.612	.002	.502
CF	5.007	.0006	.271
Kourtz	–	4	–

Table 6.7: Prediction Statistics for 1991 Predictions in zone 21-C2X.

thought of as two distinct parts of the same zone as shown in Figure 1.2. Zone 21-C2Y has seven cells. It is located in a fairly uninhabited area. There were thirty-four fires in this zone from 1985 to 1990. The mean number of fires per day was .03 with a variance of .04.

Analysis-of-deviance and backwards elimination were used to determine the models. The fitted models are shown in Table 6.8. The deviances of the fitted models and of Martell's models are listed in Table 6.9. For the common Poisson (CP) regression, the FFMC index gives a smaller deviance ($D = 241.4$) when fitted to the data than the transformation FFMC* ($D = 246.6$) used in other zones. Indicator variables for *lake* (the northern tip of Adams Lake is in zone 21-C2Y), and for *road* were included in the commmon models but neither were signficant ($p > .2$ for *lake* and *road* indicators for *Methods CP* and *CL*). The PRESS statistic for the aggregat-

	Method AP	Method CP	Method CL
Intercept	-5.151	-13.30	-7.254
	(.583)	(2.55)	(.537)
FFMC*	3.513	–	3.965
	(.854)		(.989)
FFMC	–	.093	–
		(.029)	
DC	.002	–	–
	(.002)		

Table 6.8: Fitted Models for zone 21-C2Y (standard errors in brackets).

ed Poisson *Method AP* model is 494.1 but is as high as 8774 for Martell's FFMC+BUI aggregated *Method AM* model. The PRESS statistics for the other models are all much larger than that for the aggregated Poisson *AP*. The PRESS statistic is smaller for the common Poisson (967.2) than for each of Martell's (minimum 6783).

The overdispersion statistics for the Poisson aggregated and common *AP* and *CP*, 4.512 and 10.4 respectively, give evidence for overdispersion (p-values both less than 1×10^{-4}). Section 2.2.4 mentions scaling the deviance and standard errors of Poisson models with overdispersion. The selection of covariates was not affected when the deviance was scaled by the dispersion parameter, σ^2 = deviance/d.f. . Since the covariates in the model and their parameter estimates did not change, the final model for prediction did not change in view of the overdispersion.

Method	Deviance	d.f.
AP	212.3	888
AM	226.3	887
AF	216.3	879
CP	241.4	5974
CL	261.7	5974
CM	256.7	5973
CF	251.0	5966

Table 6.9: Deviances of All Models in zone 21-C2Y.

Predictions for people-caused forest fires in zone 21-C2Y for the 1991 fire season were calculated for the seven models; their prediction statistics were calculated and are given in Table 6.10 along with the S_1 value for Kourtz's predictions. Kourtz predicted two fires in 1991 but only one fire occurred. Neither of Kourtz's predictions fall on the same day as the day the fire occurred (August third). All models developed in this thesis and Martell's models give better predictions than Kourtz, according to the prediction statistics.

For all models other than Kourtz's the values of the S_1 statistics are similar. The statistics give mixed results with no one model giving consistently smallest values of all the statistics. Martell's models for both *Methods A* and *C* have somewhat smaller prediction deviances. However, these values are outweighed by the large values of S_2. These large values are a result of small predictions on the day when the single fire occurred.

Method	Pred.Dev.	S_1	S_2
AP	6.17	1.072	13.99
AM	6.745	1.085	17.47
AF	5.99	1.097	22.50
CP	6.20	.965	20.04
CL	6.44	.997	4.44
CM	4.43	.988	138.33
CF	5.63	.999	461.95
Kourtz	–	3	–

Table 6.10: Prediction Statistics for 1991 Predictions in zone 21-C2Y.

The predictions for days without fires are small since the sum of squares statistic is small for Martell's and *Method AP* models. *Method CL* appears to predict well according to the statistic S_2. By criterion of statistic S_2, the *Method CP* common Poisson predictions are somewhat better than Martell's since the two sum of squares statistics are either similar to or smaller than Martell's.

Part III

Conclusions

Chapter 7

Recommendations and Conclusions

7.1 Conclusions

The Poisson and logistic models developed in this thesis and Martell's models fitted to B.C. data all provide better predictions than the model currently in use by the Protection Branch of the B.C. Ministry of Forests developed by Kourtz. However, the relative performance of these models varies from zone to zone within the Kamloops Region. The *Method CP*, common Poisson model, gives marginally the best predictions in the southern Kamloops Region where it is dryer and there are many fires. The *Method AP*, aggregated Poisson model, performs better than Martell's logistic models fitted to either aggregated or common data, in the southern Kamloops Region. It is not surprising that the Poisson regression models do better in areas with many fires, since logistic models assume only a binary response (0 fires or 1 fire). However, Martell's logistic models outperform the Poisson models in the northern Kamloops Region where fires are few. The Fourier terms in Martell's *AF* and *CF* models were only significant in the northern wet zone 21-C2X suggesting they may not be necessary in people-caused forest fire prediction in most of the Kamloops Region.

Overall, the models fitted to common data (*Method C*) give better

predictions than models fitted to aggregated data (*Method A*), including Martell's models. The common models are preferred over the aggregated models not only because they predict better over a zone, but also because, in the *CP* and *CL* models, the indicators for *road*, *lake*, *city*, or *park* can be included as surrogates for land use information.

Zoning the cells in the Kamloops Region to include fuel type and major climate type in the models seems valuable since different models and covariates were selected for different zones. Even in the zones 21-C2X and 21-C2Y where the fuel type and major climate type are the same, the selected logistic models differ from one another.

The initial goal of the thesis, to come up with a predictive model for the probability of people-caused forest fires in a small area shows that different models apply to different zones depending on weather and land use patterns. Poisson regression models appear to perform better for the dryer, southern parts of the region where many fires occur, while logistic regression seems more appropriate in the damper, northern region with fewer fires.

7.2 Recommendations

This section is subdivided into recommendations for the data used in analysing people-caused forest fires, for the implementation of predictive models in the province of B.C., and for further research.

7.2.1 Recommendations for People-caused Forest Fire Data

As discussed in Section 3.2, people-caused forest fire prediction may benefit from data on dates and locations of logging road and camp closures and campfire restrictions, and on locations of populated, recreational and industrial areas. The predictions from the models (Martell's included) revolve primarily around the fluctuation in the FFMC. The FFMC values in 1991 were not high signifying poor conditions for forest fire ignition; nonetheless several people-caused fires occurred. Both common sense and the inability of the FFMC (and other indices) to totally reflect people-caused fire occurrence suggests that not only weather, but people's use of the forests is required to predict people-caused forest fires. If the locations of roads, lakes, cities, parks, logging roads, and logging camps were digitalized so that these locations could be accessed by cell through the AFM database, then accurate indicators could be included in the common method models. Digitalizing this information would allow the percentage of the cell containing a road, lake, *etc.* to be included not only in the common models but also in the aggregated ones. As the aggregated models in this thesis do not contain any surrogate for land use patterns, such percentages may prove useful. Furthermore there is no way of assessing the efficacy of closures and campfire bans, unless a record is kept on when such closures occur. It is plausible that the presence or absence of a closure or campfire ban could be *the most important* factor in determining the probability of a people-caused fire, during the fire season. Once the data are digitized and available in the AFM database on the land use indicators roads, lakes, cities, and parks, it is recommended that the common models

be refitted with percentage of cell taken up by each of these indicators. It is also recommended that closures and campfire bans be incorporated into the models.

Major climate type also needs to be digitalized so that it could be accessed by cell through the AFM database. This is necessary to accurately construct zones based on major climate type. Fuel type is already available in the AFM database.

In Section 1.3 it was mentioned that the AFM database does not agree with the Datatrieve database with respect to the number of fires and the date of the occurrence of some fires. This discrepancy needs to be further examined and explained.

7.2.2 Recommendations for Implementing Predictive Models in B.C.

From the research in this thesis, it would appear that different models are fitted to different zones depending on climate, fuel type, and land use. In discussing the implementation stage of the predictive models from this thesis, the Protection Branch expressed an interest in having a single model for the entire province of B.C. To fit and test models on common data for even the Kamloops Region requires considerable amounts of computer memory and storage space. To fit and test models on aggregated data would be more manageable on regional data, but, again, memory and space requirements are considerably large. A tradeoff must occur in implementing predictive models. Either one can fit hundreds of models throughout the province, a very time consuming and costly procedure, or one can rely on general predictions over a fairly large area (such as an administrative

district) with fewer models being fitted. Choosing the appropriate balance is an administrative decision which will depend on cost factors, the mobility of fire-fighting equipment, *etc.* as well as computer memory available to the researcher.

Major climate type needs to be determined for all cells in B.C. Major climate type can then be paired with fuel type to set zone boundaries. Developing a model for a zone requires much labour-intensive work. One must decide on whether Poisson or logistic regression is to be used depending on dry or damp weather conditions, and then perform backwards elimination (and analysis-of-deviance) to select the model covariates. All models would then need to be updated with current data each year. For all of B.C. there would be approximately eight-hundred thousand data file records for every fire season that would have to be managed and processed as well.

The task could be simplified somewhat by selecting four zones within each of the remaining forest regions (outside Kamloops) then fitting Poisson and logistic models to each zone and by backwards elimination selecting a "best" model for each zone. The models fitted in the four preliminary zones may assist in the initial selection of model type and covariate selection for the remaining zones in each of the regions.

7.2.3 Recommendations for Future Research

It is recommended that the common models be refitted with indicator variables for roads, lakes, cities, parks, logging and campfire closures once these indicators are digitalized by cell.

There was overdispersion in the Poisson models in the wetter northern Kamloops Region where few fires occur. The logistic regression models gave better predictions in the two northern zones. One reason for this overdispersion may be that data is not always recorded beginning on April fifteenth even though this is officially the start of the fire season. By assuming a random length for the fire season, the negative binomial distribution may model the zones with few fires. See [19], chapter on Log-linear Modelling.

Forest fires are rare events and most data points in forest fire data are zeroes. It is recommended that the zero-inflated Poisson (ZIP) regression of Lambert [15] be used to fit forest fire data. The ZIP regression assumes that a Poisson random variable occurs with probability $1 - p$ and only zeroes occur with probability p. The ZIP regression models fit data with excess zeroes well.

While the performance of the models discussed in this thesis, in predicting people-caused fires, may not be spectacular, it should not imply that the effort to develop predictive models should be abandoned. At the very least the models proposed do far better than the current procedure (that of Kourtz) being used by the B.C. Forest Service. It should be remembered that forest fires are rare events, and predicting rare events is fraught with difficulties.

To improve the predictive performance of any model, we suggest that what is important is primarily *better data*, rather better models. Since predictions are made on the basis of weather forecasts, not actual weather, weather forecasts should be recorded for every cell and be included in the

AFM database. In addition, information on closures and campfire bans should be included in the AFM database. Likewise, information on land use such as presence or absence of roads, campgrounds, lakes, *etc.* may be very useful in prediction and should be recorded in the AFM database in a systematic way.

As far as developing the regression models is concerned, there may be value in investigating the problem of overdispersion in the Poisson aggregated models. The overdispersion may be due to the variability in the length of the fire season from year to year. The use of quasi-likelihood methods in this case should be investigated further. ZIP regression techniques should be explored.

Bibliography

1. Aitkin, M., Anderson, D., Francis, B., and Hinde, J., 1989. *Statistical Modelling in GLIM*, Oxford University Press, New York.

2. Baker, R.J., and Nelder, J.A., 1978. *The GLIM System: Generalized Linear Interactive Modelling*, Release 3, Royal Statistical Society, Oxford.

3. Bliss, C.I., 1970. *Statistics in Biology.* Vol 2. McGraw-Hill Co. Ltd., New York. pp. 219-287.

4. Bonneu, M., 1988. Model choice for prediction in generalized linear models. Statistics. **19**:369-382.

5. Canadian Forestry Service, 1984. Tables for Canadian Forest Fire Weather Index System, 4rth ed. Victoria, B.C.

6. Chatfield, C., 1988. *Problem Solving: a Statistician's Guide.* Chapman and Hall, London.

7. Cook, R.D., Weisberg, S., 1982. *Residuals and Influence in Regression.* Chapman and Hall, London.

8. Cox, D.R., Snell, E.J., 1989. *Analysis of Binary Data.* 2nd Ed. Chapman and Hall, London.

9. Cunningham, A.A., and Martell, D.L. 1972. A stochastic model for the occurrence of man-caused forest fires. Canadian Journal of Forest Research. **3**:282-287.

10. Cunningham, A.A., and Martell, D.L. 1976. The use of subjective probability assessments to predict forest fire occurrence. Canadian Journal of Forest Research. **6**:348-356.

11. Davison, A.C., Gigli, A., 1989. Deviance residuals and normal scores plots. Biometrika, **76**:211-221.

12. Dean, C., Lawless, J.F., 1989. Tests for detecting overdispersion in Poisson regression models. Journal of the American Statistical Association. **84**:467-471.

13. Dobson, A.J., 1983. *An Introduction to Statistical Modelling.* Chapman and Hall Ltd., New York.

14. Haines, D.A., Main, W.A., Frost, J.S., and Simard, A.J., 1983. Fire-danger rating and wildfire occurrence in the northeastern United States. Forest Science. **29**:679-696.

15. Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. **34**:1-14.

16. Lloyd, D., Angove, K., Hope, G., and Thompson, C., 1990. *A Guide to Site Identification and Interpretation for the Kamloops Forest Region.* Land Management Handbook Number 23, Province of British Columbia, Ministry of Forests, Victoria, B.C.

17. Martell, D.L., Bevilacqua, E., and Stocks, B.J. 1989. Modelling seasonal variation in daily people-caused forest fire occurrence. Canadian Journal of Forest Research. **19**:1555-1563.

18. Martell, D.L., Otukol, S., and Stocks, B.J. 1987. A logistic model for predicting daily people-caused forest fire occurrence in Ontario. Canadian Journal of Forest Research. **17**:394-401.

19. McCullagh, P., Nelder, J.A., 1989. *Generalized Linear Models*, 2nd Ed. Chapman and Hall, London.

20. Ministry of Forests, Protection Branch, 1983. *Fire Weather Indices: Decision Aids for Forest Operations in British Columbia*, Forest Protection Handbook 12, Victoria, B.C.

21. Ministry of Forests, Protection Information System, 1991. *Administrative Fire Systems Manual.* Ministry of Forests, Victoria, B.C.

22. Ministry of Forests, Protection Information System, 1991. *Operational Fire Systems Manual.* Ministry of Forests, Victoria, B.C.

23. Myers, R.H., 1986. *Classical and Modern Regression with Applications*, 2nd Ed. PWS Kent Publishing Co., Massechusets.

24. Pierce, D.A., Schafer, D.W., 1986. Residuals in generalized linear models. Journal of the American Statistical Association. **81**:977-986.

25. Protection Branch, British Columbia Ministry of Forests, 1988. *British Columbia Fire Suppression Program*, Ministry of Forests document, Victoria, B.C.

26. Tithecott, A.G., 1990. Evaluation of Martell's logistic regression models for people-caused fire occurrence prediction. Ontario Ministry of Natural Resources. AFMC Publication No. 273. 34 pp.

27. Todd, B., and Kourtz, P., 1990. *Predicting the Daily Occurrence of People-Caused Forest Fires.* Petawawa National Forestry Institute, Forestry Canada, Chalk River, Ontario.

28. Turner, J.A., 1961. *Report of the Working Group of the Commission for Agricultural Meteorology set up under resolution 8*, Forecasting for Forest Fire Services.

29. Turner, J.A., Lawson, B.D., 1978. *Weather in the Canadian Forest Fire Danger Rating System*, Pacific Forest Research Centre, Environment Canada Forestry Service, Victoria, B.C.

30. Van Wagner, C.E. 1974. Structure of the Canadian forest fire weather index. Canadian Forest Services Publication 1333, 44p. Petawawa Forest Exp. Stn. Chalk River, Ontario.

Appendix A
Aggregated and Common Data

The different models are referenced by letters throughout this thesis for convenience. The relationships are as follows.

A AGGREGATED: – *P* POISSON regression
– logistic regression: Martell
– *M* FFMC+BUI
– *F* FFMC+BUI+Fourier terms

C COMMON: – *P* POISSON regression
– logistic regression: – *L* LOGISTIC regression
– Martell:
– *M* FFMC+BUI
– *F* FFMC+BUI
+Fourier terms

So that, for example, the *aggregated Poisson* model is referenced as *AP*; and *Martell's logistic common model with Fourier terms* is referenced as *CF*.

Appendix B
PRESS Statistic

See [1], [19] and [23]. The PRESS statistic,

$$\frac{y_i - \hat{y}_{(i)}}{s_{(i)}\sqrt{1 + h_{(i)}}}$$

where,

$$s_{(i)} = \sqrt{\frac{(n-p)\hat{y}_i^2 - e_i^2/(1-h_{ii})}{n-p-1}}$$

can be calculated by fitting a model n times deleting the i^{th} observation, $i = 1, \ldots, n$, each time. Alternatively, the model can be fitted once and the PRESS statistic calculated as

$$\frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{1 - h_{ii}}} ,$$

as given in Equation 2.2.5. The following shows the derivation of the form of the PRESS used for calculation. The notation with subscript (i) indicates the model fitted without the i^{th} observation. The notation with subscript (-i) indicates a vector or matrix missing the i^{th} observation. The derivation uses the Sherman-Morrison-Woodbury Theorem (see [23]) which states that for an $n \times n$ matrix A and n-dimensional vector $\mathbf{z}$,

$$(A - \mathbf{z}\mathbf{z}^T)^{-1} = A^{-1} + \frac{A^{-1}\mathbf{z}\mathbf{z}^T A^{-1}}{1 - \mathbf{z}^T A^{-1}\mathbf{z}} .$$

The estimate of the parameters for the model missing the i^{th} observation is

$$\hat{\boldsymbol{\beta}}_{(i)} = (X_{(-i)}^T X_{(-i)})^{-1} X_{(-i)}^T \mathbf{y}_{(-i)}$$

and

$$X_{(-i)}^T X_{(-i)} = X^T X - \mathbf{x}_i \mathbf{x}_i^T ,$$

$$X_{(-i)}^T \mathbf{y}_{(-i)} = X^T \mathbf{y} - \mathbf{x}_i y_i .$$

The Sherman-Morrison-Woodbury Theorem is applied making the following assignments,

$$A = X^T X \text{ and } \mathbf{z} = \mathbf{x}_i$$

so that,

$$(X_{(-i)}^T X_{(-i)})^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} \mathbf{x}_i \mathbf{x}_i^T (X^T X)^{-1}}{1 - h_{ii}} .$$

Then,

$$\begin{aligned}
\hat{y}_{(i)} &= \mathbf{x}_i^T \boldsymbol{\beta}_{(i)} = \mathbf{x}_i^T (X_{(-i)}^T X_{(-i)})^{-1} X_{(-i)}^T \mathbf{y}_{(-i)} \\
&= \mathbf{x}_i^T (X^T X - \mathbf{x}_i \mathbf{x}_i^T)(X^T \mathbf{y} - \mathbf{x}_i y_i) \\
&= \mathbf{x}_i^T \left[(X^T X)^{-1} + (X^T X)^{-1} \mathbf{x}_i^{-1} (1 - h_{ii})^{-1} \mathbf{x}_i^T (X^T X)^{-1} \right] \left[X^T \mathbf{y} - \mathbf{x}_i y_i \right] \\
&= \left[\mathbf{x}_i^T (X^T X)^{-1} + \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i^{-1} (1 - h_{ii})^{-1} \mathbf{x}_i^T (X^T X)^{-1} \right] \left[X^T \mathbf{y} - \mathbf{x}_i y_i \right] \\
&= \mathbf{x}_i^T (X^T X)^{-1} X^T \mathbf{y} + \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i^{-1} (1 - h_{ii})^{-1} \mathbf{x}_i^T (X^T X)^{-1} X^T \mathbf{y} \\
&\quad - \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i y_i - \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i^{-1} (1 - h_{ii})^{-1} \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i y_i \\
&= \hat{y}_i (1 + \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i^{-1} (1 - h_{ii})^{-1}) \\
&\quad - h_{ii} y_i (1 + \mathbf{x}_i^T (X^T X)^{1} \mathbf{x}_i^{-1} (1 - h_{ii})^{-1}) \\
&= (\hat{y}_i - h_{ii} y_i)(1 + h_{ii} (1 - h_{ii})^{-1})
\end{aligned}$$

So,

$$\begin{aligned}
y_i - \hat{y}_{(i)} &= y_i - \frac{\hat{y}_i - h_{ii} y_i}{1 - h_{ii}} \\
&= \frac{y_i - \hat{y}_i}{1 - h_{ii}} .
\end{aligned}$$

Now for an expression for $h_{(i)}$ in terms of h_{ii}, using the Sherman-Morrison-Woodbury Theorem for $(X_{(-i)}^T X_{(-i)})^{-1}$,

$$\begin{aligned} h_{(i)} &= \mathbf{x}_i^T (X_{(-i)}^T X_{(-i)})^{-1} \mathbf{x}_i \\ &= \mathbf{x}_i^T \left[(X^T X)^{-1} + (X^T X)^{-1} \mathbf{x}_i^{-1} (1 - h_{ii})^{-1} \mathbf{x}_i^T (X^T X)^{-1} \right] \mathbf{x}_i \\ &= \frac{h_{ii}}{1 - h_{ii}} \end{aligned}$$

The PRESS residual becomes

$$\frac{\frac{y_i - \hat{y}_i}{1 - h_{ii}}}{s_{(i)} \sqrt{1 + \frac{h_{ii}}{1 - h_{ii}}}} = \frac{y_i - \hat{y}_i}{s_{(i)} \sqrt{1 - h_{ii}}} .$$

Appendix C
Glossary of Acronyms

This glossary contains explanations of the acronyms used in forestry and in model fitting.

63-C7 zone 63-C7: major climate type 63 or southern, very dry and fuel type C7 or Douglas Fir

6-O1 zone 6-O1: major climate type 6 or central-very dry and fuel type O1 or grass

21-C2X zone 21-C2X: major climate type 21 or northern-wet and fuel type C2 or Boreal Spruce

21-C2Y zone 21-C2Y, to the south of, but not contiguous with zone 21-C2X

AF Aggregated Fourier: Martell's model with Fourier terms fitted to aggregated data

AFM Advanced Fire Management database and software

AM Aggregated Martell: Martell's model fitted to aggregated data

ANODEV ANalysis Of DEViance

ANOVA ANalysis Of VAriance

AP Aggregated Poisson: Poisson regression model fitted to aggregated data

BUI Build Up Index

CF Common Fourier: Martell's model with Fourier terms fitted to common data

CFFWIS Canadian Forest Fire Weather Index System

CL Common Logistic: logistic regression model fitted to common data

CM Common Martell: Martell's model fitted to common data

CP Common Poisson: Poisson regression fitted to common data

DC Drought Code

DMC Duff Moisture Code

FFMC Fine Fuel Moisture Code

FFMC* a logistic transformation of the FFMC as given in Equation 3.1

FWI Fire Weather Index

GLM Generalized Linear Model

GLIM Generalized Linear Interactive Modelling software

ISI Initial Spread Index

NFDRS National Fire Danger Rating System

FFWI Fosberg Fire Weather Index

PRESS PREdicted Sum of Squares statistic

ZIP Zero Inflated Poisson regression

VITA

Surname: POULIN-COSTELLO Given Names: MELANIE RENÉE
Place of Birth: LONDON, ON Date of Birth: 66.01.14

Educationl Institute Attended:

University of Victoria	1990 to 1993
University Waterloo	1985 to 1990

Degrees Awarded:

B.Math. (Honours)	University of Waterloo	1990

Honours and Awards:

University of Victoria Dean's Scholarship	1990-1992
University of Victoria Graduate Teaching Award	1990-1991
JobTrac Supplement Award	1990-1991

PARTIAL COPYRIGHT LICENSE

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other university, or similar institute, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of Thesis: People-Caused Forest Fire Prediction
Using Poisson and Logistic Regression

Author [redacted]
(Signature)

MELANIE R. POULIN-COSTELLO
(Name in Block Letter)

93.07.22
(Date)