Factors that Influence Daily Human-Caused Forest Fires in Alberta

By

Kimberly Morrison

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

In

Forest Biology and Management

Department of Renewable Resources

University of Alberta

Abstract

Humans are the major cause of forest fires in the spring in Alberta, and have resulted in major property damage in both the Flat Top Complex fires in 2011 and the Fort McMurray fire in 2016. Fire occurrence prediction (FOP) models can help predict when and where fires can be expected in order to help fire managers manage resources. In Alberta, these FOP models need to be improved, especially in regards to spring fire starts and the timing of the end of the spring fire season as most human-started forest fires in Alberta occur in the spring. Candidate models were created to explore which independent variables best predict human-caused fire starts in Alberta. The independent variables are separated into four groups: spatial distribution, Fire Weather Index System codes and indices, human influences, and seasonality. Finally, several model forms were explored to determine the best model, as determined by best fit to the data and/or best predicted fire occurrence. These were: Generalized Linear Model (GLM), Hurdle Model and Zero-Inflated Model, each with a Poisson and negative binomial link. A GLM with a negative binomial link and the following variables predicted the number of human-caused fire starts the best: FFMC, FWI, ECOREGION, FFMC X ECOREGON, and SEASON3 (a three level variable with a transition season between spring and summer). This model had a RMSE of 0.697 when tested on a bootstrapped set of test data. This model could be used as the basis of future FOP model research in Alberta, and management of wildfire fighting resources in conjunction with other fire activity prediction methods.

Acknowledgements

There are many people I would like to acknowledge, without whom this thesis would not exist. I would like to thank my supervisor Dr. Mike Flannigan for his guidance and support throughout this process. I appreciate the feedback and technical guidance from my second committee member Dr. Mike Wotton. Also, a big thanks to everyone who helped me with my statistics, but in particular Dr. Xianli Wang and Dr. Piyush Jian, and to those that helped me with my editing, in particular Karen Blouin and Keith Wollenberg. I am also thankful to all my labs mates for their support and help or just for being a sounding board for my thoughts. I would like to thanks all my friends and family for supporting me in the endeavor. I would like to thank the University of Alberta and the Western Partnership of Wildland Fire Science for giving me this opportunity. Finally I would like to thank the TRANSFOR-M program for giving me the opportunity to spend the first year of my Masters in Freiburg, Germany.

The funding for this study was provided by the government of Alberta.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AAF	Alberta Agriculture and Forestry (also formally part of the ESRD)
AEP	Alberta Environment and Parks ministry (formally known as ESRD)
AIC	Akaike Information Criterion
AP	Absolute Value of predicted number of fires starts minus actual number of fire starts
BUI	Build-Up Index
CFFDRS	Canadian Forest Fire Danger Rating System
DC	Drought Code
DMC	Duff Moisture Code
ESRD (now	Alberta Ministry of Environment and Sustainable Resource Development known as AEP and AAP)
FFMC	Fine Fuel Moisture Code
FOP	Fire Occurrence Prediction
FWI	Fire Weather Index
GDD	Growing Degree Day
GLM	Generalized Linear Model
ISI	Initial Spread Index
NB	Negative Binomial
PC	Pearson Correlation
RMSE	Root-Mean-Square Error
SD	Sample Data
TD	Test Data
WMA	Wildfire Management Area

Introduction

Although forest fires are a natural part of the boreal forest, some forest fires can threaten infrastructure, resources and even human life. Accordingly, Canadian provinces with substantial forest stands devote considerable resources to forest fire management. In order to allocate fire control resources effectively, it would be desirable to have a system that could accurately predict the number of forest fires at a given time, in a given area and under a given set of circumstances. This is the goal of Fire Occurrence Prediction (FOP) modelling. As the processes involved in human and lightning-caused fires are different they are modeled separately. In this thesis, I seek to develop an FOP model for human caused-fires specific to Alberta, which builds on past research. In this process, I will consider both what independent variables (such as fuel moisture/weather, human activity and seasonality) best predict day-to-day changes in fire occurrence and what model forms are most useful under these conditions.

General

Forest fires are a natural part of the boreal forest ecosystem. Fire controls many aspects of the boreal forest including species composition, stand structure, seral stage, age class distribution and spatial heterogeneity. Many boreal forest species are adapted to, and rely upon, fires as part of their life histories. For example, tree species such as jack pine (*Pinus banksiana*) have serotinous cones which require extreme heat to open and release the seeds. Jack pine also prefer to germinate on mineral soil in full sun and are consequently well adapted to growing in ecosystems with stand replacing fires. The same can be said to some extent for lodgepole pine (*Pinus contorta*) and black spruce (*Picea mariana*). Deciduous tree species such as trembling aspen (*Populus tremuloides*) and paper birch (*Betula papyrifera*) that reproduce through suckering prefer full sunlight and are

therefore also able to grow well after fire. All of these tree species benefit from the reduction in competition resulting from the removal of vegetation during a fire (Edmonds et al., 2000).

Fire processes in a given location and time period can vary in severity, intensity, seasonality, size and return interval, which are collectively called the fire regime. Fire severity is defined as the depth of the fuels on the forest floor consumed by a fire (Stocks et al., 2002), or the effect of the fire on vegetation. Fire intensity is a measure of forest fire behavior defined as the energy output rate per unit length of fire front or as a physical attribute of the fire itself, directly related to flame height (Byram, 1959).

Fires can vary in intensity and severity depending on the season in which they occur. In general the deep forest floor fuels tend still to be wet in the spring, likely preventing fires from consuming as much of the forest floor. However, the low foliar moisture levels of coniferous trees before bud burst (Chrosciewicz, 1986), and the absence of leaves on deciduous trees and vegetation allow for more intense fires in the spring. These variations in intensity and severity have a distinct effect on the landscape, including the recovery after the fire, creating heterogeneity of the landscape. There can also be variation of severity and intensity within the same fire created by smaller scale variations in topography, fuel s (e.g. structure and moisture) and the atmosphere (Countryman, 1972). Some areas may be almost completely burned, while others burn partially, leaving patches completely untouched. This creates further heterogeneity on the landscape and habitat for a variety of flora and fauna.

There are three main types of fires: ground fires, surface fires and crown fires (and a mix of crown and surface) (Countryman, 1972; Kasischke et al., 1995). Ground fires burn deep into the forest floor and often occur in bogs, consuming large amount of dead fuels and sphagnum moss, and releasing

large amount of carbon into the atmosphere. These fires can smolder for a long time. Surface fires burn the surface of the forest floor. Although surface fires can leave thick-barked tree species alive, in the boreal surface fires will most likely kill the more thin-barked trees, but still leave them standing with their foliage (Kasischke et al., 1995). Crown fires burn the crowns of trees. These fires can spread from crown to crown, or along the surface, candling the trees as they spread. Either way, these are high intensity fires, resulting in the consumption of live foliage, and partial consumption of the tree boles (Van Wagner, 1977). A fire can change between all three types within the same fire, or burn as only one type, creating even more heterogeneity on the landscape.

Historically, fires in the boreal were often started by lightning, and to a lesser extent by aboriginal peoples (Larsen, 1997). The number of human-caused fires has increased since the arrival of colonists and subsequent population growth. Lightning fires caused fires and a natural occurrence, and therefore helpful to maintaining a healthy boreal forest, although changes in climate and fuels may cause such fires to burn in 'unnatural' ways. Human-caused fires are not considered natural, and while these fires may also help maintain a healthy boreal forest, they also change the fire regime of an area in regards to seasonality, and possibly frequency, severity, and intensity, as forest fuels have different properties in the spring than the summer. Seasonal variations in forest fuels may include degrees of curing, leaf flush, and deep fuel drying. Also, human-caused fires occur more often near important values (e.g. towns, property, and human life (Vega-Garcia et al., 1993a).

Between 1959 and 1997, only 3% of fires in Canada were over 200 ha in size, but these fires accounted for about 97% of the area burnt. Of these large fires, about 28% were caused by humans. On average there are approximately 8500 fires a year in Canada burning approximately 1.8 million

hectares, although this number varies greatly with over 7 million hectares burning some years (Stocks et al., 2002).

In contrast between 2005 and 2014, about 65% of fires starts in Alberta were human-caused, resulting in 47% of the area burnt, a much higher percentage than the country as a whole. These human-caused fire starts averaged 989 per year, burning an average of 90,095 ha per year (Alberta Agriculture and Forestry, 2014a). Although human-caused fires are responsible for less area burnt than lightning caused fires, the damage can be severe due to their proximity to human values. The 2011 Flat Top Complex fires near the Town of Slave Lake, Alberta and the Fort McMurray fire in May 2016 are examples of spring human-ignited fires that threatened human life and caused extensive property damage. These fires have several things in common; they are human-caused, in the spring and there were multiple fires in the same area at the same time causing firefighting resources to be spread thin.

Furthermore, Alberta had a very active springe fire season in 2011, and most spring fires in Alberta are caused by people. This resulted in about 84% of fires in that year being caused by humans (1,139 human-caused fires), burning 91% (731,634 hectares) of the total area burnt (Alberta Agriculture and Forestry, 2014a). This year therefore represents a significant deviation from even the more recent human-influenced fire regimes. These extreme years tax forest managers' resources.

The Fire Weather Index System

The Fire Weather Index System (FWI System) was created to give an indication of the moisture content of forest fuels and a relative rating of potential fire behaviour through the use of its six codes and indices. The FWI System provides an indication of the receptivity of fuels to ignition, potential rate of spread, and potential fire front intensity. This System determines values for each of its

indices based on past and present weather conditions, rather than measurement of actual fuel moistures (Van Wagner, 1987).

The Fine Fuel Moisture Code (FFMC) represents the moisture content of fine fuels and was developed by relating weather to fine fuel moisture contents in a pine stand in Petawawa, Ontario. These fuels are at the surface of the ground (originally pine needles) and are the smallest fuel class. Due to the high surface area to volume ratio of these fuels, this index represents the drying of the forest fuels over a short period of time (2/3 of a day). The FFMC is calculated using rainfall, relative humidity (RH), wind speed and temperature (Van Wagner, 1987).

The Duff Moisture Code (DMC) represents the moisture content of the duff layer. This measure starts below the fine surface fuels, covering medium-sized fuels that continue to an average of 7 cm in depth, to a maximum of about 10 cm in depth (Alberta Agriculture and Forestry, 2010). This index represents drying over an intermediate time frame (weeks). The DMC is calculated using rainfall, RH and temperature (Van Wagner, 1987).

The Drought Code (DC) represents the deeper soil organic matter, greater than approximately 10 cm in depth (Alberta Agriculture and Forestry, 2010), averaging up to 18 cm in depth. This measure also includes large dead woody debris such as downed logs. As such materials take a long time to dry, this index represents the drying of forest fuels over the spring and summer (i.e., periods of months). That said, these fuels are not always fully rewetted over the winter, in which case an overwintering DC can be calculated to account for this. The DC is calculated using only rainfall and temperature, as wind speed and RH do not significantly affect the drying of large fuels (Van Wagner, 1987).

The final three indices of the FWI System are calculated using the FFMC, DMC and DC and provide a relative rating of potential fire behaviour. The Initial Spread Index (ISI) is calculated using the FFMC

and wind speed and is an indication of how quickly a fire will spread on the ground (without taking fuel quantities into account). The Buildup Index (BUI) represents the total fuel available for consumption by a fire and is a weighted average of DC and DMC. The Fire Weather Index (FWI) takes all of the other indices into account and provides an indication of fire front intensity (Van Wagner, 1987)

The FWI System indices, although used across Canada, do not have consistent meaning across the country. Different index values can indicate different fire danger ratings in different parts of the country (Van Wagner, 1987) and an index value can represent different fuel moisture content depending on fuel type (Van Wagner, 1987; Wotton & Beverly, 2007).

Forest Fire Occurrence Prediction Modeling

Forest fire occurrence prediction (FOP) models can help wildfire managers better allocate resources. Since forest fires are best fought while small (Edmonds et al., 2000; Todd & Kourtz, 1991; Wotton et al., 2003), having enough resources close by, and crews on standby when large numbers of fires are predicted, should allow fires to be detected and responded to more quickly and insure the number of fires does not overwhelm the available resources.

It is common practice to model the different sources of ignition – lightning and humans – separately, due to differences in the ignition process arising from such factors as seasonality, spatial distribution, and FWI System code/index most closely related to fire occurrence, as well as the ignition source. In Alberta, lightning-caused fires are more common in the summer, whereas human-caused fires are more common in the spring (Figures 1 and 2). Human-caused fire starts are more closely related to the FFMC (Martell et al., 1987; Martell et al., 1989; Wotton et al., 2003; Wotton et al., 2010), while lightning-started fires are better predicted by the DMC as the ignition source often gets into the duff

layer and smolders before actively spreading (Wotton & Martell, 2005). Additionally predicting where and when ignition sources will be available is different for lightning and human-caused fires as predicting where and when lightning and humans will be are two completely different processes (Vega-Garcia et al., 1993b; Wotton & Martell, 2005). This thesis will focus on human-caused forest fires in Alberta.

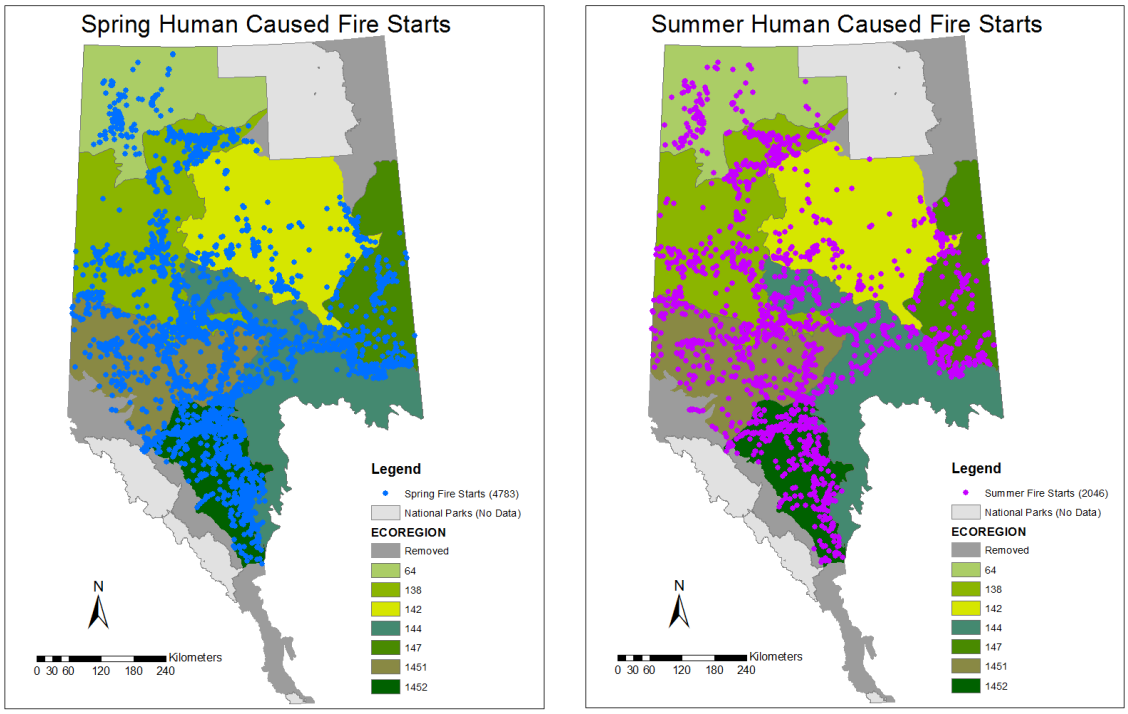

Figure 1: Human-caused fire starts from 1983-2014 in Alberta before June 1st (spring) and June 1st or later (summer). National Parks were removed, as these fires weren't park of the provincial record.

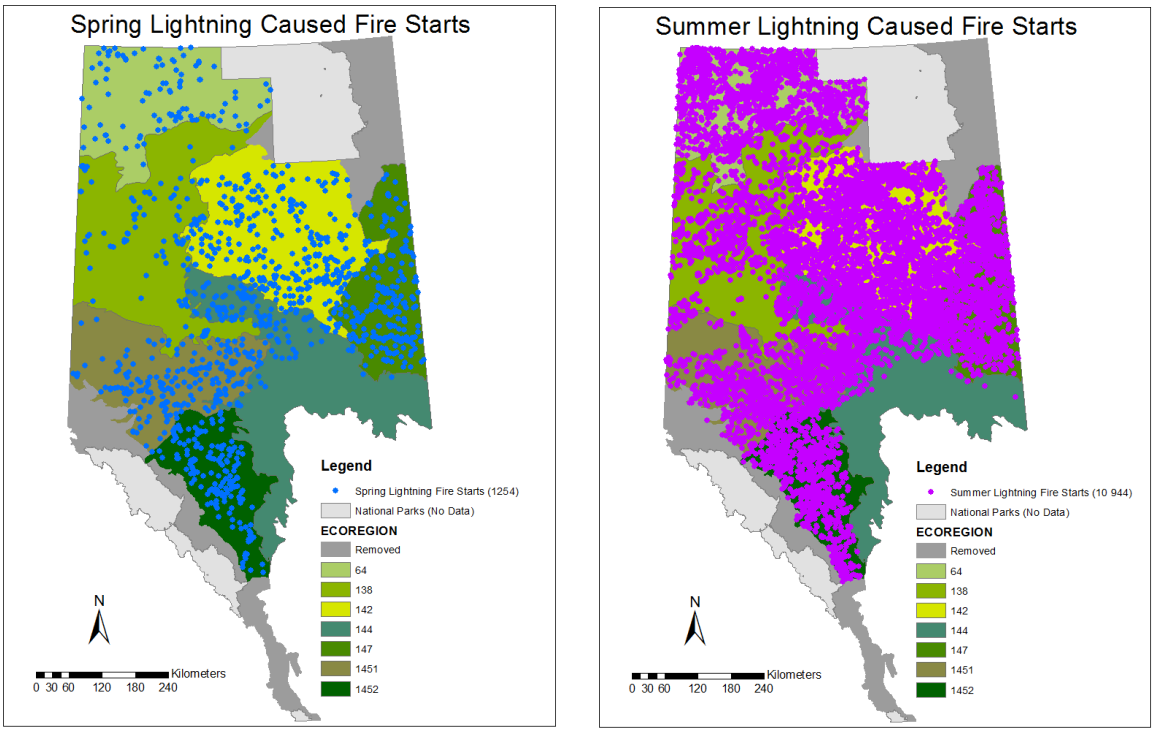

Figure 2: Lightning-caused fire starts from 1983-2014 in Alberta before June 1st (spring) and June 1st or later (summer). National Parks were removed, as these fires weren't park of the provincial record.

Fire Occurrence Prediction models can be used to predict the likelihood (or number) of fires for the upcoming day (or in the near future) using weather forecasts (Todd & Kourtz, 1991) or to help predict the effects of climate change on the number of fires in the future, by inputting the results of general circulation models (Wotton et al., 2010).

Currently FOP models have limited use operationally in Canada. Instead, intelligence officers use experience combined with the FWI System codes and indices to predict fire numbers and locations. Although it has been shown these predictions by fire managers are reasonable (Cunningham & Martell, 1976; Todd & Kourtz, 1991), supplementing this approach with more objective computer based prediction models would be expected to increase prediction reliability. In BC and Ontario,

models are used to assist in these predictions already. British Columbia (BC) uses lightning and human-caused fire predictions models named SPARKY and PEOPLE respectively. These models were created by Bernie Todd in 1991 (Todd & Kourtz, 1991) and modified in later years for use in BC. Ontario uses a lightning model created by Wotton and Martell (2005), mostly to determine potential locations of high fire activity. Ontario also currently uses a human-caused FOP model prototype based on Woolford el al. (2011), which was originally made for only a small area of Ontario (the Romeo Malette Forest). There are FOP models being built for other provinces, but they are not currently in use (Mike Wotton, personal communication, September 17, 2015).

Variables used to predict fires

Fires occur when there is oxygen, an ignition source, and fuel present. Of these, fire occurrence modellers usually just consider ignition source and fuel, as oxygen is relatively constant in forests. In regards to human-caused FOP models, the number of ignition sources (e.g. number of people in the forest or where people are in the forest (Vega-Garcia et al., 1993b)) or type of ignition (e.g. the activity the people are doing) may be included (Martell et al., 1987; Martell et al., 1989). Human-ignited fires typically occur near areas humans frequent — with the notable exception of powerlines — such as cities, towns, parks, roads, railways, campsites and industrial sites. For example, Vega-Garcia found 90% of human-ignited fires started within 4.8 km of a road in the Whitecourt Forest, Alberta (Vega-Garcia et al., 1993b). However, human-caused fires are more likely to be detected quickly, as there are more people around that are likely to see and report them. These variables can be separated into static variables that don't change from day-to-day such as proximity to road or town and non-static variables that change from day to day such as the day of the week or fire bans. This thesis will focus on the day-today changes in human activity.

The Ontario Ministry of Natural Resources (OMNR) classifies human-caused forest fires into seven categories: industry (forestry), industry (other), incendiary (arson), recreation, resident, railway, and miscellaneous. Unknown fire causes are also often included with human-caused fires. Sometimes these ignition types are grouped into two classes: those with a peak of fire occurrence in the spring and those with a peak of fire occurrence in the summer. Class 1, with its peak in the summer, includes: recreation, industrial (forestry), industrial (other) and incendiary. Class 2, with its peak in the spring, includes: resident, railway, miscellaneous and unknown (Martell et al., 1989; Wotton et al., 2003; Wotton et al., 2010). Additionally, Martell et al. (1989) theorized that fires such as recreational fires (e.g. from campfires) would more likely start in the duff and be related to the DMC, whereas fires from class 2, such as railway fires were more likely to start in the grass and be related to the FFMC.

Fuel is the final factor necessary for fires to occur. The drier the fuel, the less energy is required to ignite said fuel. Fuel moisture is often represented in models by the FWI System indices, rather than gravimetric moisture contents. For spring fires, the FFMC was found to be positively related to the likelihood of a human-ignited fire, as many human sources of ignition depend on the dry fine fuels (Martell et al., 1987; Martell et al., 1989; Todd & Kourtz, 1991; Wotton et al., 2003). While the relationship between human-caused fire starts and FFMC is consistently supported by literature for spring fires, Martell et al. (1987) found that in the summer in Northern Ontario, different FWI System codes and indices can be more closely related to human-ignited fire starts. Depending on the ignition type, BUI, DMC, FWI, as well as FFMC, can be the more important code/index in these conditions. For example, railway fires were most likely to ignite fine, top fuels in the spring and summer, and were therefore modeled using the FFMC in both seasons. However, forest industry fires were most closely related to FFMC in the spring, but to BUI in the summer.

However, there are also seasonal trends in fuel moisture not accounted for using the FWI System. The same FFMC value will indicate drier fuels in the spring than the summer and fall (Wotton & Beverly, 2007). In the early spring, deciduous trees and understory vegetation have not yet leafed out (also known as flushing), allowing larger amounts of solar radiation, and possibly wind, to reach (and dry) the forest floor(Kiil & Grigel, 1969). It is unclear why this relationship holds true for all forest types, not just deciduous and mixedwood stands (Wotton & Beverly, 2007). However, the presence of dry, dead, cured fine fuels from the previous summer in all forest types may account for this (Kiil & Grigel, 1969). These dry fuels combined with the spring physiological drop in live coniferous tree crown needle moisture makes for very flammable forests in the spring (Chrosciewicz, 1986; Little, 1970; Van Wagner, 1967). Additionally, without deciduous vegetation to shade fuels or block wind, fuels dry quicker (Kiil & Grigel, 1969). Once the vegetation has flushed, the likelihood of fire decreases until the vegetation dries later in the summer. Every plant greens-up at a different time, and the likelihood of fire does not decrease immediately at leaf flush. For example Kiil and Grigel (1969) noted that during the fires in 1968 in Alberta, the aspen and poplar leaves were 25% developed.

Current models use various techniques to account for this seasonal importance. Wotton et al. (2010) uses a binary indication of season (i.e. spring/ summer). Others use ordinal day (Morin, 2015), three sub-seasons based on weekly average number of fires (Martell, et al., 1987), or complicated periodic functions based on ordinal day and historical fire occurrence (Martell, et al., 1989) to account for the effects of seasonality. With the exception of Wotton et al. (2010), much of this work was done in Ontario, Canada, which has different seasonal trends in fire occurrence and different vegetation than Alberta.

Morin (2015) used generalized additive models to compare models separating season by spring/summer, ordinal date and first bloom of the saskatoon (*Amelanchier alnifolia*) plus a lag time (resulting in a different division of spring and summer for each year) in Alberta. She found that including variable season divisions based on saskatoon first bloom resulted in the model with the lowest deviance, with the ordinal date model having the second lowest deviance.

When considering fuel, modellers should also consider ecoregion, as different vegetation behaves differently in regard to fire. Additionally, the actual fuel moisture content and the likelihood of fire at a given FFMC varies among vegetation types (Wotton & Beverly, 2007). Therefore for accurate modeling, different vegetation types should be modeled separately. Wotton et al. (2010) took this into account by modeling each forested ecoregion in Canada separately, as defined by the Ecological Stratification Working Group (1996). Finally, as different provinces have different fire management policies, and different record keeping histories, each province was modeled separately in Wotton et al. (2010).

The goal of this thesis is to develop FOP models specific to Alberta. Additionally, this thesis will explore the following questions:

1. Which FWI System indices best predict human-caused fires in Alberta?
2. Which seasonality variables best predict human-caused fires in Alberta?
3. What other variables are important in human-caused fire prediction in Alberta?
4. Does using more complex models such as a zero-inflated and hurdle model improve the predictions?

Data

Study Area

Alberta is located in the North-Western Hemisphere between 49 and 60 degrees latitude and 110 and 120 degrees longitude. Alberta is a province in Western Canada, of about 661,000 square kilometres (Government of Alberta, 2015) and composed of prairie in the south, parkland in the centre and boreal forest in the north. The west of the province is bordered by the Rocky Mountains and the far northeast of the province transitions to Canadian Shield (Stamp & Warnell, 2016).

The province has a continental climate, with four distinct seasons. The summers are generally mild, with daytime highs of 20 to 25 degrees Celsius, and winters cold. Although Alberta has the most sunny days in Canada (2,300 hours) (Government of Alberta, 2015) in the summer, isolated thundershowers are common. On average, Alberta receives about 300 mm of precipitation per year in the south, 450 mm in the north and 550 mm to 600 mm in the foot hills of the Rocky Mountains. The growing season in Alberta ranges from 120 days in the south to 60 days in the north, with the northern part of the province having much longer days in the summer than the south (Stamp & Warnell, 2016).

This study uses data spanning 30 years (1983-2014). In 1981 (just before the start of the study period), there were 2,237,724 people living in the province, with 77% of the population living in urban areas and 23% in rural areas. As of 2011, there were 3,645,257 people living in Alberta, with 83% living in urban areas and 17% living in rural areas. Although there was a higher percentage of the population living in the city in 2011 compared to 1981, there was still an increase of 100,000 people living in rural areas (Statistics Canada, 2011). While not all of these people are living in the forested part of Alberta, they man still travel and recreate in forested areas. This means that there

was an increase in the number potential ignition sources of human-caused fires as more people lived, worked and recreated in rural areas.

In Alberta, due to oil and gas and other resource industries, as well as the geography of the province, there are people working and living throughout the province. By way of comparison, the mountainous geography of BC makes building access roads to remote locations difficult; this is not a problem for much of Alberta. Therefore, unlike other provinces in Canada which have both full suppression zones and observation zones, Alberta only has full suppression. In full suppression zones all fires are, in theory, actively supressed. In observation zones, fires are observed and only supressed if they threaten human values. For example, fire in regions with little or no people or property may be allowed to let burn their natural course while just being monitored by the firefighting agency (Wotton et al., 2010). This distribution of people across the landscape allows for human-caused fires distributed throughout the province, although still concentrated near major roadways and towns/cites (Figure 1).

Data Description

This project makes use of six data sets (Table 1). These data sets were all modified to suit the projects needs and aggregated into one file in the comma-separated values format (csv).

Table 1: Description of data used in this thesis.

Description	Obtained From	Spatial Extent	Time Frame	Format
Noon LST weather data	AAF (ESRD)	Alberta	1983-2014	csv file
Fire occurrence data	AAF (ESRD)	Alberta	1983-2014	3 csv files
Ecoregion polygons	Ecological Stratification Working Group	Alberta	Not Applicable	shapefile
Fire bans	AAF	Alberta	2006-2015	pdfs

Data was included from 1983 through 2014 when available. This time frame was deemed appropriate as it balances the need for a complete fire record and consistency in people's behaviours with having enough data. Various advances in technology and changes in policy have made for a more complete fire record. Starting in the 1950's, aircraft were used for fire detection, because they allowed for a much greater area to be viewed. Around this same time, a more concerted effort was put into fire detection and fighting. By the mid 1970's Alberta also had a network of lookout towers with observers assigned to monitor the landscape for fires. In 2016 there were 127 lookout towers in Alberta (Alberta Agriculture and Forestry, 2016a). This was particularly true for northern Alberta, where the fire record was far from complete before 1951 (Murphy et al., 2000). Furthermore, the advent of lighting detection systems (i.e. the Canadian Lightning Detection Network started in 1998 (Burrows & Kochtubajda, 2010), and the provincial lighting detection system for Alberta started in 1983 (Alberta Agriculture and Forestry, 2014b)) made it easier to differentiate between lightning and human-caused fire starts, allowing for more accurate records of human-caused fire starts. Finally, climate changes over time, as does the distribution and habits of people, in turn causing changes in fire regimes (Stocks et al., 2002; Wotton et al., 2010). All of these changes make prediction of fires less reliable as the past is no longer a good representation of the present or the future. Although there has been some change in fire regime since 1983, this date was a good compromise between data quantity, quality and relevance.

Weather Data

The weather data set contains noon Local Standard Time (LST) weather records of relative humidity, 24 hour precipitation, temperature, and wind speed as well as longitude, latitude and elevation. These data are from 1983-2014 inclusive and cover the province of Alberta. The data set also

contained over wintering calculations for start-up DC values done using Turner and Lawson equations (Turner & Lawson, 1978) (Paul Kruger, personal communication, August 17, 2014).

Fire Data

The fire start data were obtained from the Alberta Department of Agriculture and Forestry Historical Wildfire Database (AAF) (Alberta Agriculture and Forestry, 2015) in three separate shape files and compiled into one file. The data set contains latitude and longitude, fire start date and cause for Alberta from 1983-2014 inclusive. These data are provincial government data, and are therefore missing the national parks.

Ecoregion Data

The ecoregions are a combination of ecoprovince and ecoregion and were obtained as two shapefiles from the Ecological Stratification Working Group (Ecological Stratification Working Group, 1996).

Fire Ban Data

Fire ban data was obtained by compiling the ministerial order files in Portable Document Format (pdf) from the ESRD webpage, but only includes the dates 2006 through 2014 (Alberta Agriculture and Forestry, 2016b). One fire ban was missing an end date and could not be included.

Data Processing

Weather Data

Weather Station Selection

Data processing was required to get a complete weather data set for the chosen time frame, with FWI System indices and codes calculated for each day in the fire season.

Three weather stations were chosen to represent the weather of each ecoregion. The selection criteria are as follows, in order of importance:

1. Continuous weather station for the time period of the study (i.e. no missing years from 1983 to 2014).
2. Stations needing as little temporal interpolation as possible but no more than 4 days in a row requiring all weather data to be interpolated.
3. Stations that include as much of the fire season as possible (i.e. it is preferable that the station starts recording before the middle of May and stops recording after the end of August). The start date is particularly important as my study focuses on spring fires (before June 1st), therefore data are needed before this month.
4. Located as close to the centre of the ecoregion as possible or as evenly distributed in the ecoregion as possible (i.e. insure weather stations represent ecoregion as best as possible. For example weather stations should not be clumped in one corner of an ecoregion whenever possible).

In some instances, up to 8 days in a row of only relative humidity had to be interpolated early in the season due to problems associated with negative temperatures. Although up to 4 days in a row of missing data was allowed, there was only one such instance, the rest of the data had at most 3 days missing in a row.

Not all the chosen stations had the desired spatial distribution, and/or started recording by mid-May, and/or stopped recording after the end of August. These data were still considered usable. However, the ecoregions in the far northeast and the Rocky Mountains were dropped from the study as there was too much missing weather data (six years and nine years of data respectively).

Additionally, once the National Parks were removed the ecoregions were quite small. Figure 3 shows the selected weather stations, with the dark grey areas representing the ecoregions that were dropped from the study.

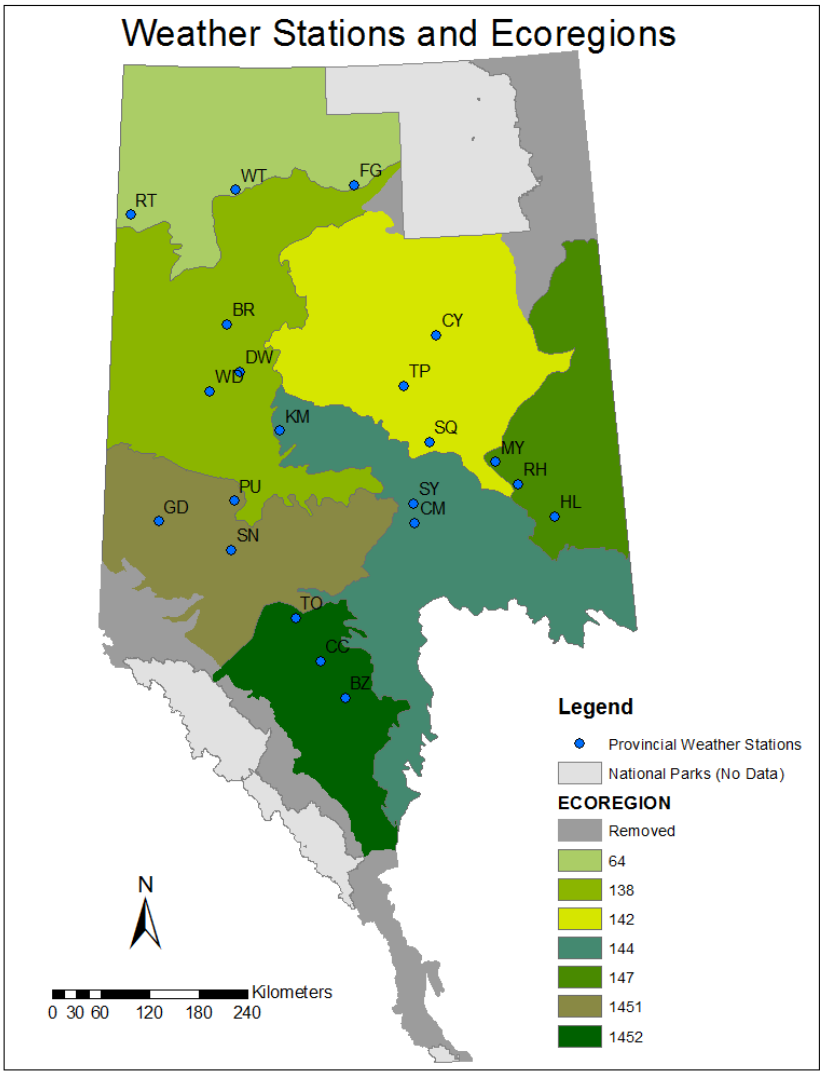

Figure 3: Study area showing selected ecoregion and selected weather stations.

Weather Interpolation

Missing weather data was interpolated for the selected weather stations. These values were interpolated in two ways. If only one value was missing in a row, then a linear interpolation was performed; that is, the daily values before and after the missing value were averaged.

If two or more values were missing in a row, a thin plate spline interpolation was performed using all weather stations, from the original AAF weather data set. At least ten stations were required with weather data recorded, preferably within three degrees of latitude and longitude of the station with missing data. If this was not possible, ten or more stations within five degrees of latitude and longitude were used. If this was not possible, no interpolation was done.

If no interpolation could be done, and no FWI System indices for that date were already calculated, the station was dropped and replaced. CL was the only station dropped for this reason, it was replaced with KM in ecoregion 144. If no interpolation was done, but only the RH values were missing (as was common early in some seasons where there was a cold spell early in the fire season), the FWI System indices and codes calculated by AAF were kept for the fire season up to that date. The missing RH values were always only at the beginning of the fire season. The FWI System indices and codes were likely calculated using the RH as measured at another, representative station, although this was not recorded in the data set (Personal communication, Brett Moore, 2015). All other FWI System indices and codes were recalculated using the CFFDRS package in R (Wang et al., 2015), using the initial (start-up) values used in the original AF weather file, as they included overwintering for DC. If the RH was missing, all values were calculated after the final missing RH value, using the FWI System indices and codes for the last day with a missing RH for the initial values.

After the FWI System indices and codes were calculated for each weather station, they were averaged by each ecoregion. The result is a table with one value for each FWI System index and code for each ecoregion and day.

All data were used from the start of the weather station's record through ordinal date 243 (August 31st, August 30th for leap years). This is referred to as the study fire season from this point forward. If the weather stations started at different times, the record starts with the first station and includes the other stations in the average as they come online (FWI System codes and indices were averaged).

Fire Data

The fire data were downloaded from the AF webpage in multiple files and formats (Alberta Agriculture and Forestry, 2015). They were then combined to all be the same format. Fields that were not needed were removed.

The fire occurrence data set was processed as follows:

1. All fires that were not of human or unknown cause type were removed including restarts, prescribed fires or lightning fires. (It is assumed that unknown cause type fires were caused by humans (Martell et al., 1987; Martell et al., 1989; Wotton & Martell, 2005; Wotton et al., 2010)).
2. All records 0.01 ha or less were removed as this was a new size class in 1995, and in 2002 there was another change in fire record changing procedure to include all fires no matter how small. This resulted in an increase in the numbers of very small human-caused fires (Quince, 2009), some of which may have been no more than an illegal campfire. In order to maintain consistency, all fires under 0.01 ha in size were removed.
3. All records without a complete start date were removed.

4. Only fires started within the study area were included.

Prior to 2004, industry fires were only grouped into forest industry and other industry cause types. From 2004 onwards information was collected on oil and gas fires and power line fires and from 2012 onwards agricultural fires. To keep records consistent, all industry fires other than forest industry were grouped into the category of other industry. Additionally, the few fires considered to be under investigation were grouped with the undetermined fires.

Ecoregion Data

The Ecoregion data are a combination of ecoprovince and ecoregion from the Ecological Stratification Working Group (Ecological Stratification Working Group, 1996), and modified to approximately match the Ecoregions from Wotton et al. (2010). Wotton et al. (2010) divided the ecoregions in Alberta by provincial boundaries and significant vegetation changes. They also divided ecoregions by fire management zones if policies regarding firefighting changed between zones (e.g. a modified response zone versus a full suppression zone). This is not relevant to Alberta, as the entire province is a full suppression zone. A spatial intersect was run in ESRI's ArcMap (Version 10.2.2) on the Ecoregions and the fire data to have an ecoregion variable for each record.

Fire Ban Data

Fire ban data was extracted from its original pdf format, and recorded in an Excel spreadsheet. As the fire ban data were separated into Fire Control Zones, rather than ecoregions, these data were further processed so the Fire Control Zones under a fire ban were visually overlaid with ecoregions. The Fire Control Zones and ecoregions did not align well, and so the portion of the ecoregion under a fire ban was approximated. The ecoregion was considered under a fire ban if more than approximately 25% of the area was under a fire ban. This somewhat small fraction was chosen

because it is assumed that the portion of the ecoregion under the fire ban is the portion with the highest likelihood of fire occurrence, and therefore the area of interest of the model. The data on fire bans were limited to just 9 years (2006 to 2014) and therefore had limited value, as the FOP models in this thesis were developed for a much longer time frame (1983 to 2014).

Variable Calculation (Holidays and Week of the Year)

All the data were combined into one file, with one record per ecoregion per day. Each record included by date the following variables; ECOREGION, FFMC, DMC, DC, ISI, BUI, FWI, HOLIDAYS, DAY OF WEEK, ORDINAL DATE, WEEK OF YEAR, SEASON2, SEASON3, FIRE BANS and ACTUAL FIRES. ACTUAL FIRES is the dependent variable, and is the total number of human-caused forest fires in each ecoregion for each day. All variable calculations were done in R (R Core Team, 2015).

The HOLIDAYs variable, which includes provincial and federal holidays, was determined using the R library of timeDate (Rmetrics Core Team et al., 2015). The following holidays were only included if they fell within the study fire season:

1. Good Friday
2. Easter Monday
3. Victoria Day
4. Canada Day
5. Heritage Day

If the holiday fell on a Monday or Friday, the adjacent weekend was also counted as a holiday. For holidays occurring on a Saturday or Sunday, the entire weekend was considered to be a holiday. The HOLIDAY variable is binary with 1 representing a holiday and 0 representing no holiday.

SEASON2 was a binary variable with 1 being before June 1st and 0 being June 1st or later. SEASON3 was a factor variable with 3 levels represented by the numbers 0, 1 and 2. With 2 being days before ordinal date 147 (May 27th, May 26th for leap years), 1 being ordinal dates 147 through 160 (June 10th, June 9th for leap years) and 0 being ordinal dates 161 or later.

WEEK OF YEAR is a factor variable, with the week of the year the fire occurred, from 1 to 52 and was calculated using the format function as part of the timeDate package in R. The PREVIOUS FIRES variable is the number of fire starts that occurred in a respective ecoregion the previous day. DAY OF WEEK is the day of the week the fire started calculated using the R function weekdays. The ORDINAL DATE is a continuous numeric date with 1 being January 1st of each year, and 365 being December 31 except for leap years. It was calculated in R using the function yday as part of the lubridate package.

Sample and Test Data

Finally, the data were divided into two data sets, a sample data (SD) and a test (TD). The TD was composed of 10 years of data randomly selected using the sample function in R, without replacement. The years included in the TD are: 1986, 1988, 1990, 1992, 1994, 1998, 2000, 2010, 2011, and 2012. The remaining 22 years makes up the SD. The SD was used to fit the model, the TD was used to see how the model performed on a data set that was not used to create it.

As the TD and SD need to have the same levels of the factor variables of WEEK OF YEAR and ORDINAL DATE (or at least there should be no new factor levels introduced in the TD, as the model will not have the information to predict a new factor level), several records needed to be removed from the TD since this data set had a few records start earlier than the SD. Only 5 records were removed and none of them had any fire occurrence, as they were right after the start of spring. These data sets were used for the rest of this thesis.

Methods

There are many different approaches to creating and selecting models. However, some general guidelines seem to be common in the ecological sciences. Model variables should be selected based on sound scientific theory (e.g. does it make sense ecologically that the independent variable can be used to predict the dependent variable). Common sense and logic also need to be used when selecting model variables and candidate models. In the case of models with similar goodness of fit (e.g. Akaike Information Criterion) other aspects of the models should be compared such as model predictions, cost of variable measurement (or perhaps reliability/ease of obtaining variable) (Carruthers et al., 2008; Faraway, 2002), *"…parsimony, and coherence of the underlying assumptions* [and] *the consistency with known behavioral phenomena…"*(Vandekerckhove et al., 2014)

In that vein, the data were explored including independent variable correlation, dependent and independent variable relationships and general data summaries. Then, the different model types and data distributions used in this thesis were discussed. Thirdly, the candidate models were formed using information gained from this exploration, previous research and ecological knowledge. Finally, the methods for evaluating and comparing the candidate models were presented.

Data Exploration

The data were thoroughly explored to ensure quality, to summarize its contents, and to explore relationships between the independent variables and the dependent variable. The data were first checked for completeness and correctness, and for outliers.

The first step in creating the candidate models was to start with a list of logical, scientifically plausible independent variables. The second and third steps were to explore how well these variables explained the dependent variable and how correlated said variables were with each other. From this

exploration, candidate models were formed, from which the best model was selected. Variables were selected based on the results of previous FOP research (with a priority on Canadian research), data exploration, logic and experience/observations of when fires occur.

Much work has been done on this topic already for other areas of Canada, or the country as a whole. This research was used as a basis in order to investigate appropriate predictor variables for the model. From this research, four subgroups of independent variables were investigated: spatial variation, seasonality, FWI System (weather), and human variables. The following four sections explore these subgroups.

Spatial Variation

Since vegetation, climate and human activity differ across the province of Alberta, it is expected that fire occurrence rates, fuel type and seasonality will also differ across the province. Additionally, it is expected that there will be a different relationship between FFMC and number of fire starts by ecoregion, as one FFMC value can indicate different fuel moisture levels in different vegetation (Wotton & Beverly, 2007)

The fire start data was summarized by ecoregion. Table 2 shows that area burned and number of human-caused fires varies greatly by ecoregion. Additionally, the average area burned does not necessarily correlate with the average number of fire starts. For example, ecoregion 136 has only 14 fires per year on average, but has on average over 10 times the annual area burnt than ecoregion 138, which has 75 fires. This difference in fire regimes between ecoregions further justifies the inclusion of an ecoregion variable in the model, in addition to changes in vegetation by ecoregion.

Table 2: Summary of data used for study, showing total number of human-caused fires and area burnt by ecoregion for entire study period (1983 through 2014), from the seasonal start of FWI calculations to Ordinal date 243. Fires outside the study area were not included in this summary, nor were fires 0.01 hectares in size or smaller.

Ecoregion	Area Burnt (ha)	# of Fires	Average Area Burnt a Year	Average # Fires Year
64	28921	379	904	11.8
138	42759	1558	1336	48.7
142	23128	588	723	18.4
144	217204	1389	6788	43.4
147	255850	745	7995	23.3
1451	86415	855	2700	26.7
1452	26870	1322	840	41.3
Total	681148	6836	21286	214

Since all evidence points towards ecoregion being an important independent variable, it was included as an independent variable in all candidate models. Therefore, all exploratory analysis henceforth is separated by ecoregion, further highlighting the difference between ecoregions.

Seasonality

Alberta is prone to having most of its human-caused fires in the spring, with a sharp drop off around green-up and then a slight (if any) increase in the late summer (Figure 4). Therefore some indication of the time of year was deemed necessary to include in the models. Four options were considered:

1. A binary season of spring before June 1st and summer starting on June 1st (called SEASON2)
2. A tertiary season with spring, summer and a transition season between spring and summer (called SEASON3)
3. ORDINAL DATE (a count of the day of the year with January 1st being one and December 31st being 365/366 depending on whether it is a leap year or not)

4. WEEK OF YEAR (a count of the week of the year with the first week in January starting with Sunday being week 1)

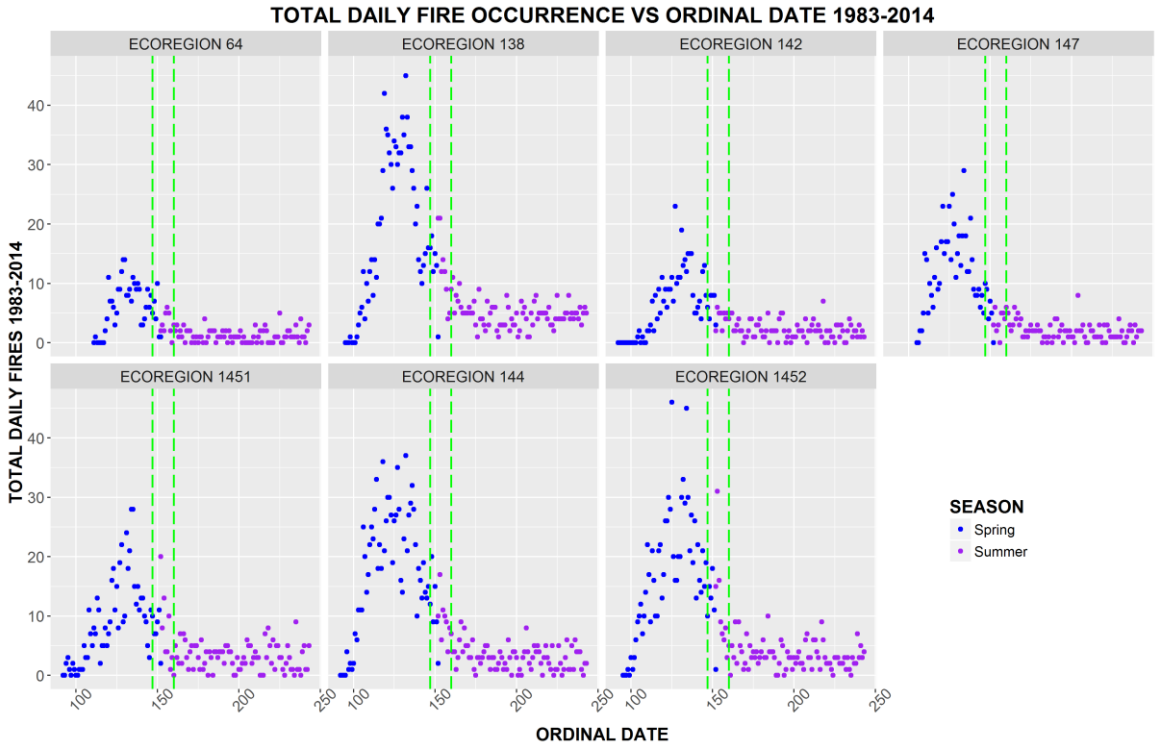

Figure 4: Cumulative number of fires by ordinal date, separated by ecoregion and season with spring being before June 1st and summer June 1st or later. The green dotted line represents the start and finish of the transition period of the variable SEASON3 (inclusive).

Using a simple two season variable (SEASON2), has been done before (Wotton et al., 2003; Wotton et al., 2010), and June 1st appears to roughly approximate the end of the spring peak in fire occurrence (Figure 5). However, in some years, June 1st is too early and in other years too late. Hence the need for the somewhat more complex variables from options 2-4.

SEASON3 attempts to encompass some of the year to year variation not accounted for in SEASON2, by including a transition season. The break points of SEASON3 were determined by visually inspecting the data (Figure 4). The transitional season was determined to be the ordinal date from

the point the number of fire starts begins to slowly decrease (after the sudden decrease after the spring peak) to the ordinal date where the number of daily fires has decreased almost to the average daily rate of fire occurrence in the summer. Spring occurs right before this transition period and summer right after. Although this transitional season works well in general for all ecoregions, the graph shows some indication that slight variations of the ordinal dates delineating the transition season by ecoregion could be beneficial. For example, ecoregion 138's could perhaps be extended by a day or two, but that is beyond the scope of this thesis.

ORDINAL DATE, adds many factor levels to the model (each day was 1 level, and therefore needed its own coefficient). This level of detail may be too great as there is much day to day variation. WEEK OF YEAR greatly reduced the number of factor levels compared to ORDINAL DATE, and may smooth out the day to day variation, yet be more detailed than SEASON2 and SEASON3. All four options were compared in four different candidate models to explore which variable produced the best model.

All further data exploration was separated by season in order to see the importance of seasonal variation. In order to simplify matters, the simplest option was chosen (SEASON2).

FWI System

As dryness of fuels (especially the fine fuels) has been shown to be highly correlated with human-caused fire occurrence, some measure of fuel moisture must be included in the model. The FWI System codes and indices are used to indicate fuel moisture and potential fire behaviour in Canada. Wotton et al. (2010) showed that FFMC, DMC and DC were useful to include in a Canada wide model. BUI and FWI could also be relevant (Martell et al., 1987). This thesis looks at FFMC, DMC, DC, BU, ISI and FWI and their abilities to predict daily fire starts.

In order to visually compare the FWI System codes and indices to the actual number of fires, the FWI System codes indices were divided into bins, and then plotted against the average number of daily fires starts that occurred for that bin, for each ecoregion and two-level season. These graphs were fitted with a negative binomial GLM fit line created using a model that employed only the variables graphed. These bins were only used as part of the data exploration, and not in the candidate models, where non-binned data was used.

FFMC is an indication of the moisture content of the fine surface fuels, and is well supported in the literature as being highly correlated with human-caused fire occurrence, especially in the spring (Martell et al., 1987; Wotton et al., 2003; Wotton et al., 2010). The relationship between rounded FFMC (i.e. FFMC expressed as an integer) and average number of daily fire starts for the data used in this study is shown in Figure 5. The daily actual number of fires were binned by FFMC integer value and then divided by the number of days in that bin. Also shown is a fit line for the relationship, which was produced using a GLM with a negative binomial link. Although some noise is evident, overall the average number of daily fires fits quite well with the FFMC, and accordingly FFMC is included in all candidate models. Potential sources of the noise in the relationship will be addressed in the Discussion.

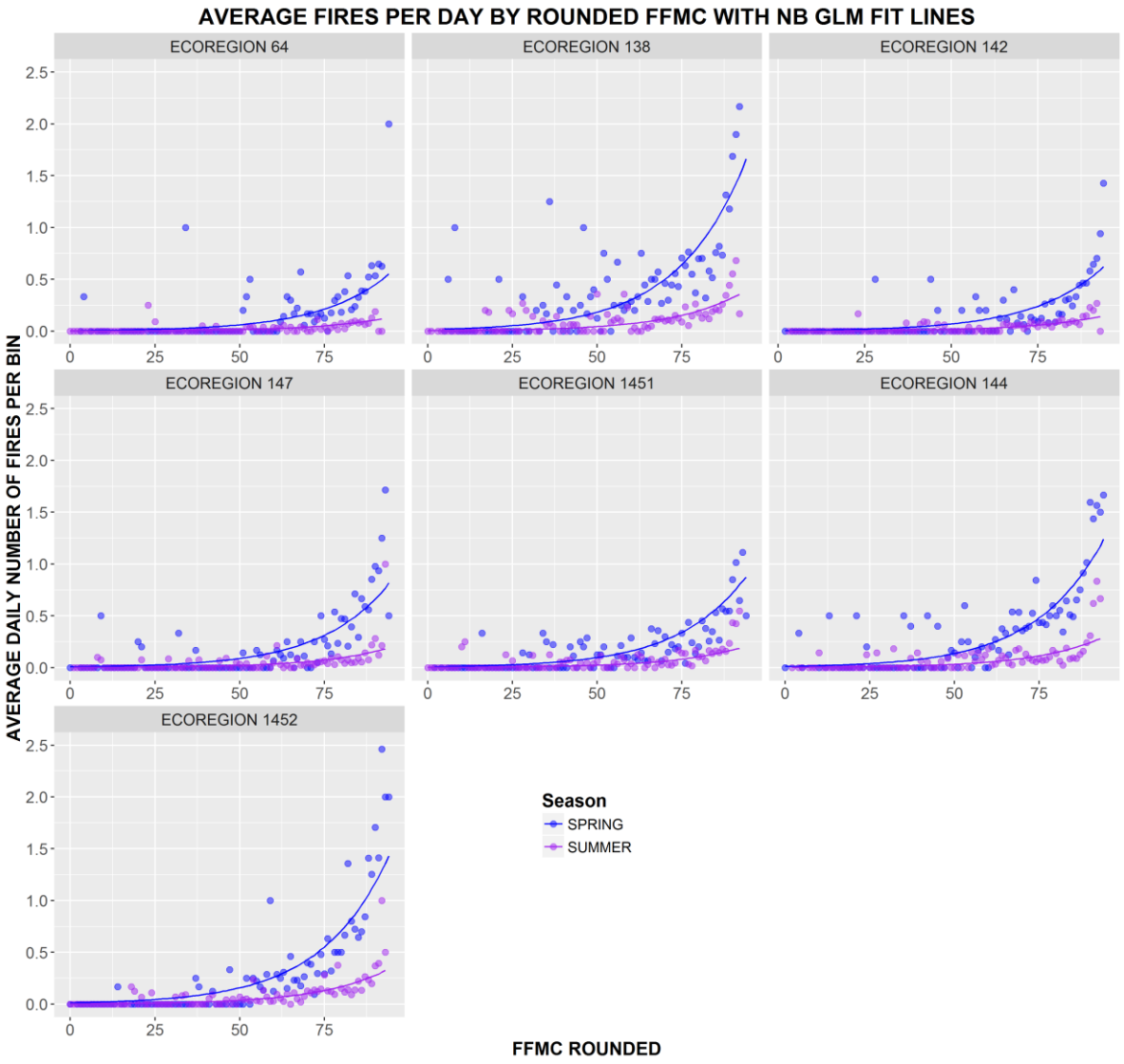

Figure 5: Average number of fires a day versus rounded FFMC (i.e. a bin 20 FFMC would be FFMCs of 19.5 to 20.49), separated by ecoregion and season. Lines were fit using a GLM with a negative binomial distribution. Two points are not shown to improve visualization (ecoregion 138 x=2.67, y=93 and x=5.00, y=94).

Martell et al. (1987) found that although human-caused fire occurrence was best predicted using FFMC in the spring, other variables in addition to FFMC, including BUI, DMC and FWI, could be the best predictors in the summer in Northern Ontario, depending on cause. For example, in the summer FWI was the best predictor of resident-started fires, whereas FFMC was the best predictor of railway-started fires in the summer. In the spring, indiscriminate of cause, human-ignited fire

starts were best predicted by FFMC. Therefore it is prudent to consider BUI, DMC and FWI as potential predictors of human-caused fires as well. ISI is not commonly indicated as a predictor of human-caused fires (Martell et al., 1987; Wotton et al., 2003; Wotton et al., 2010). Additional exploratory analysis of ISI also did not indicate a good fit at high ISI values (Figure 6). ISI appeared to be a good fit until an ISI of about 30 is reached (Figure 7). At this point, the fit line over-predicts fire starts greatly. This would still be workable, as there are very few ISIs over 30, except the over predictions are huge (predicting about 239,000 fires for one day in the most extreme case). ISI is not included in the candidate models.

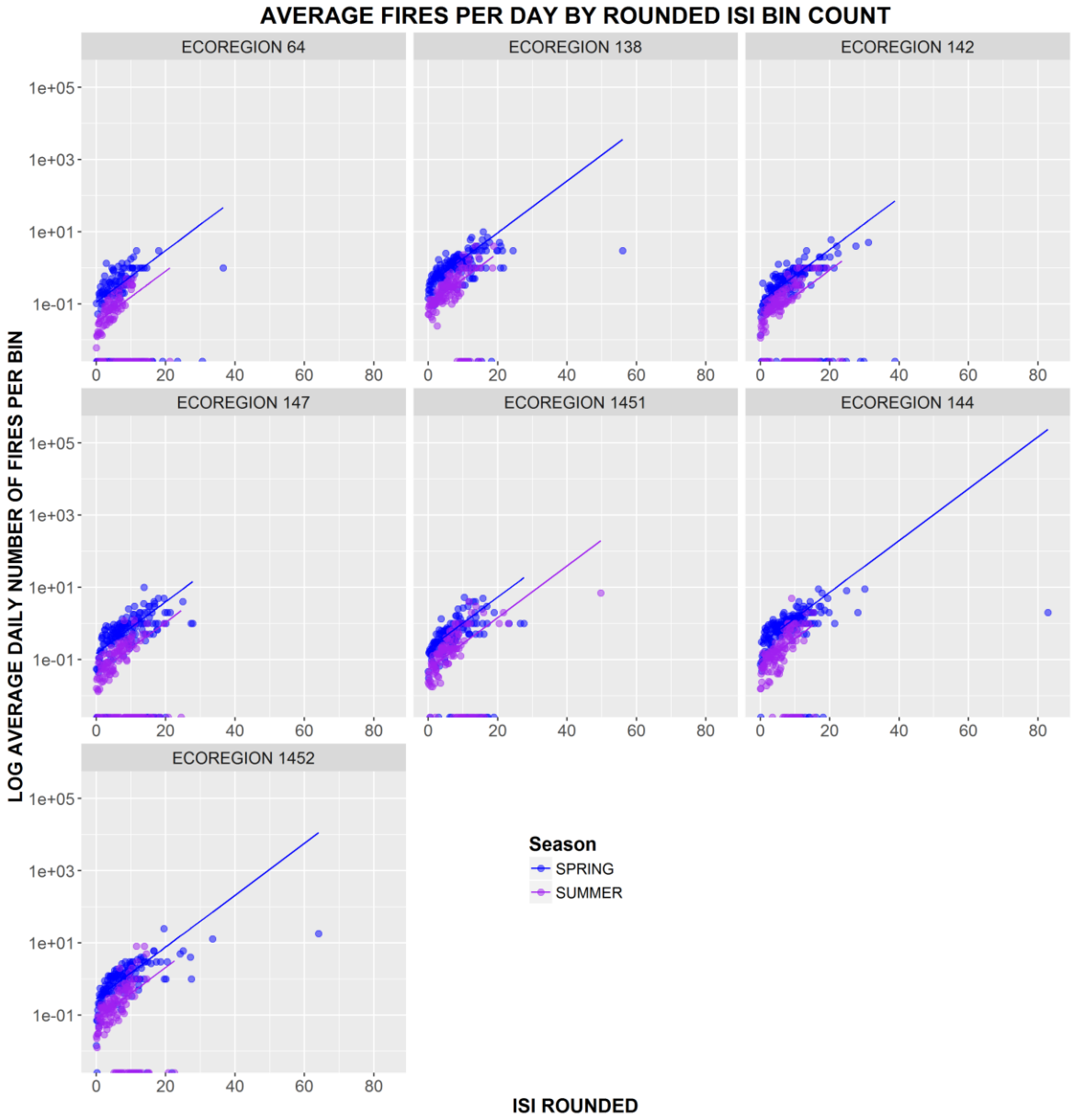

Figure 6: Average number of fires a day versus rounded ISI (i.e. a bin 20 ISI would include ISIs of 19.5 to 20.49), separated by ecoregion and season. Lines were fit using a GLM with a negative binomial distribution. If there were no days in a bin, no dot was shown.

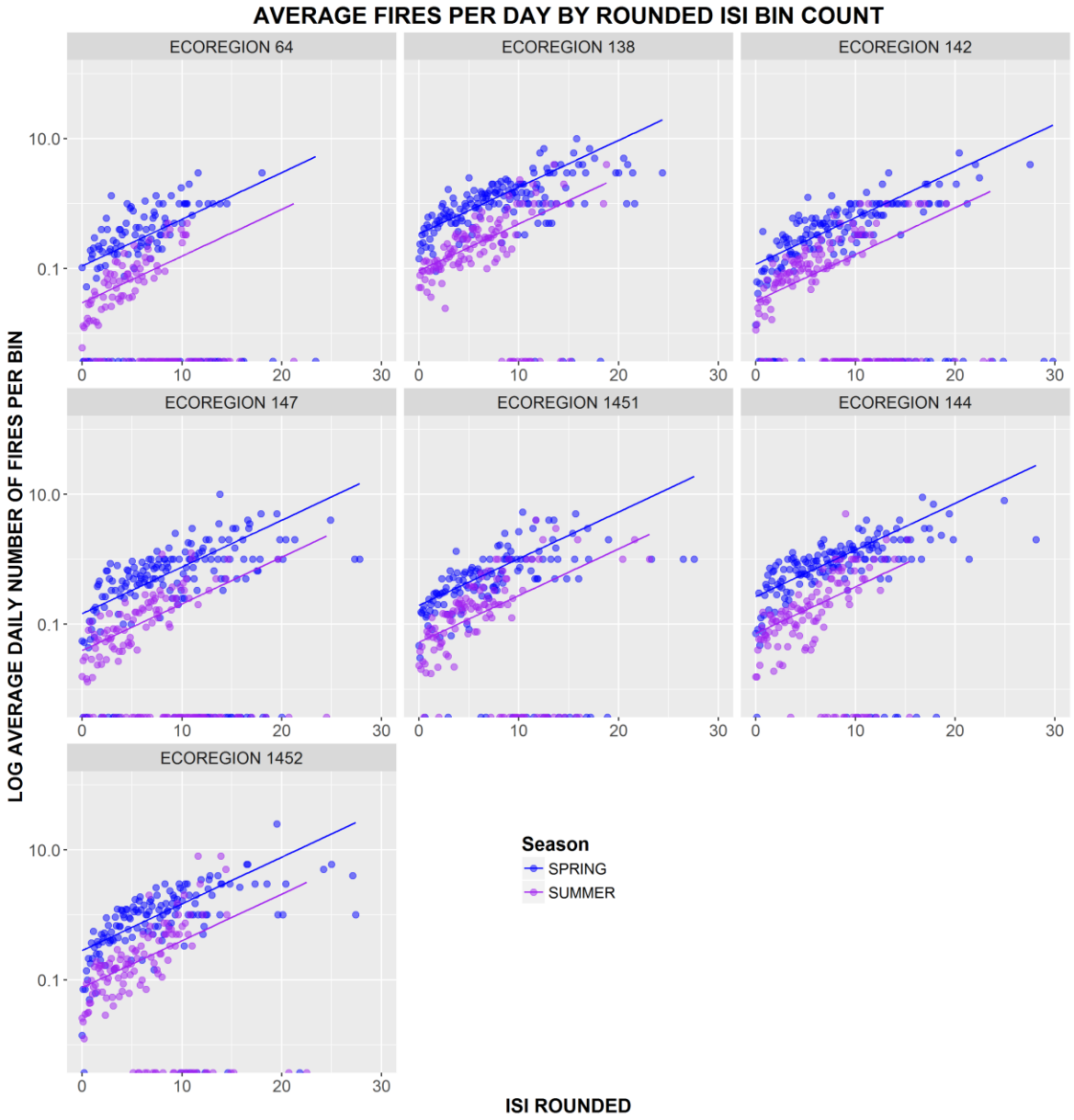

Figure 7: Average number of fires a day versus rounded ISI (i.e. a bin 20 ISI would include ISIs of 19.5 to 20.49), separated by ecoregion and season. Lines were fit using a GLM with a negative binomial distribution. This graph eliminates the extreme values for better visualization.

FWI and actual fire starts were compared in a similar manner to FFMC and ISI as shown above.

Figure 8 shows a good fit between FWI and actual fire starts up until an approximate FWI of 30

(varying by ecoregion and season), at which point FWI generally over-predicts actual fire starts.

However, as this over-prediction is not large, there are very few FWI values over 30, and the fit up to 30 is very good, FWI was include in the candidate models.

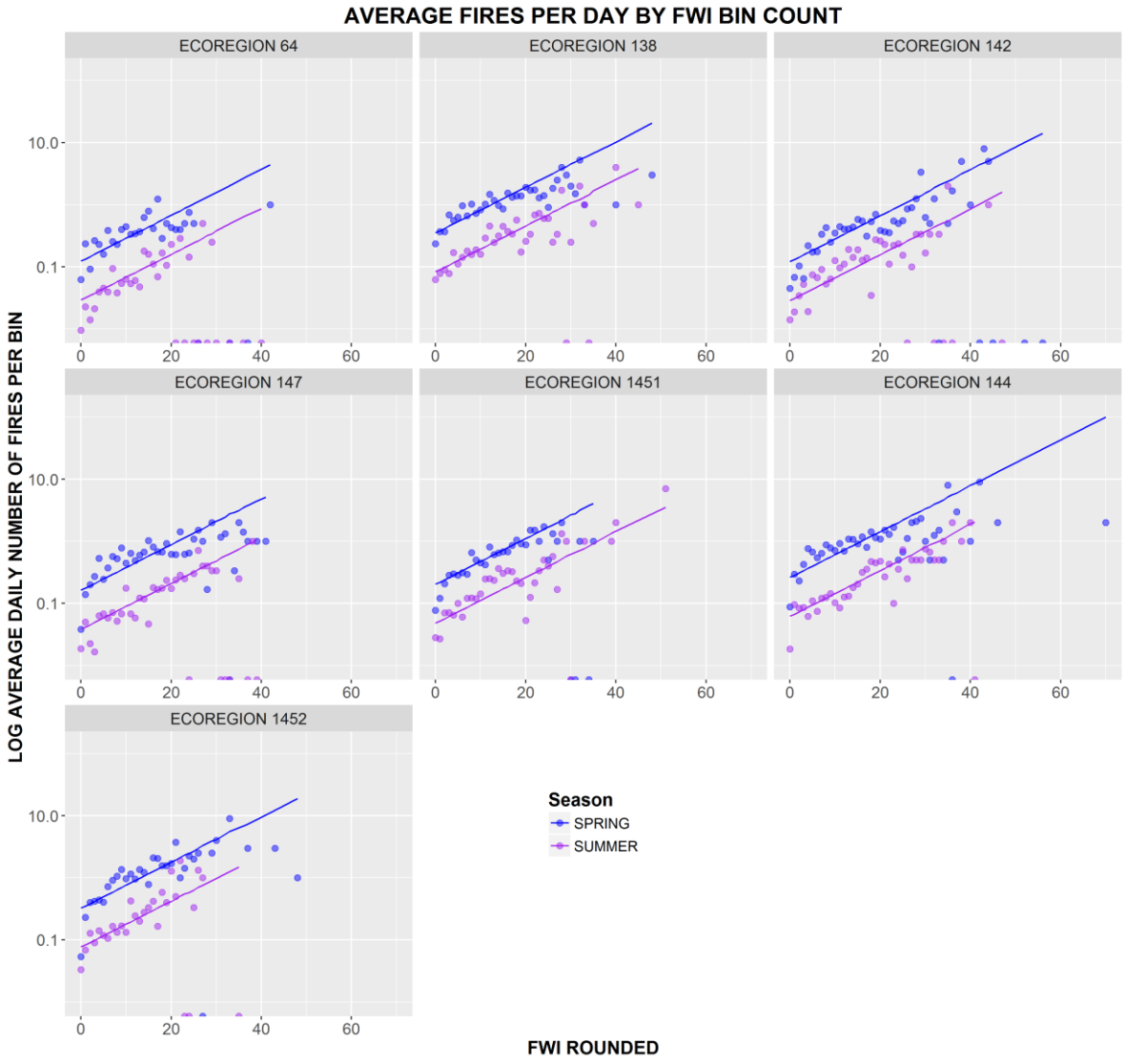

Figure 8: Average number of fires a day versus rounded FWI (i.e. a bin 20 FWI would include FWIs of 19.5 to 20.49), separated by ecoregion and season. Lines were fit using a GLM with a negative binomial distribution.

Similar graphs were produced plotting DMC, DC and BUI against actual number of fires. There were positive relationships between each of DMC, DC and BUI with actual number of fires, but not as strong as for FFMC and FWI. They are not shown for reasons of brevity.

Included in Wotton et al.'s (2010) model was an interaction term between ecoregion and FFMC based on the evidence that FFMC values should represent different levels of fuel moistures in different vegetation types, such as species composition and stand density (Wotton & Beverly, 2007). This was considered applicable to Alberta and included in all candidate models. Wotton and Beverly (2007) also found that actual moisture contents related to FFMCs varied based on the time of year, therefore an interaction term between season and FFMC was included in candidate models with two or three seasons. An interaction term between FFMC and WEEK OF YEAR or ORDINAL DATE seemed impractical.

A Pearson correlation was run to check for correlation between the FWI System codes and indices. Some were found to be highly correlated with each other (Table 3). As it would be redundant to use highly correlated variables in the same candidate model, a threshold of 0.7 was used to determine which variables were too correlated to include together (Dormann et al., 2013).

Table 3: Pearson product-moment correlation of FWI System codes and indices. An * denotes correlation coefficients of 0.700 or higher. All correlations have a p-value of less than 0.001.

	FFMC	DMC	DC	ISI	BUI	FWI
FFMC	1.000	0.482	0.175	0.660	0.490	0.668
DMC	0.482	1.000	0.447	0.462	0.978*	0.710*
DC	0.175	0.447	1.000	0.095	0.583	0.271
ISI	0.660	0.462	0.095	1.000	0.445	0.924*
BUI	0.490	0.978*	0.583	0.445	1.000	0.698
FWI	0.668	0.710*	0.271	0.924*	0.698	1.000

FWI is highly correlated with DMC and BUI (0.710 and 0.0.698 respectively) and therefore are not included in the same candidate model. Although FFMC and FWI were fairly highly correlated (0.668), the correlation was under the threshold of 0.7, therefore the variables were included in the same candidate model. BUI is very highly correlated with DMC (0.978) and is not included with DMC (Table 3) in the same model.

Human Variables

There are several commonly used ways to include how variations in human activities affect human-caused fire occurrence. Some modelers are interested in static variables that influence human-caused fires, such as distance to nearest road, distance to nearest town or land ownership (Vega-Garcia et al., 1993b). This paper looks at variables that influence day-to-day fire occurrence. These include grouping fire causes by the time of year their peak number of starts occur (CAUSE GROUP) (Martell et al., 1987; Wotton et al., 2003; Wotton et al., 2010), or when people are more likely to be in the wilderness (HOLIDAYS and DAY OF WEEK). This thesis will look at cause, day of week, holidays, fire bans and previous day's fires.

In previous studies in Ontario, human-caused fire starts were placed into two groups. Group 1 (summer peak) included Other Industry, Forest Industry, Recreation and Incendiary. Group 2 (spring peak) included Railroad, Resident, Undetermined and Miscellaneous Known (Martell et al., 1989; Wotton et al., 2003; Wotton et al., 2010). However, as this work was done in Ontario and on viewing the data for Alberta (Figure 9), it was determined that Alberta did not display the same seasonal trends by cause as Ontario. Most significantly was the lack of fire occurrence peaks in the summer (recreation-started fires were the only one with a large summer peak). CAUSE GROUP was not included as only one cause type had a summer peak in occurrence, and omitting CAUSE GROUP made

for a simpler model. Additionally, comparing models with CAUSE TYPE included to those without CAUSE TYPE would not allow for the use of Akaike Information Criterion (AIC) to compare the models.

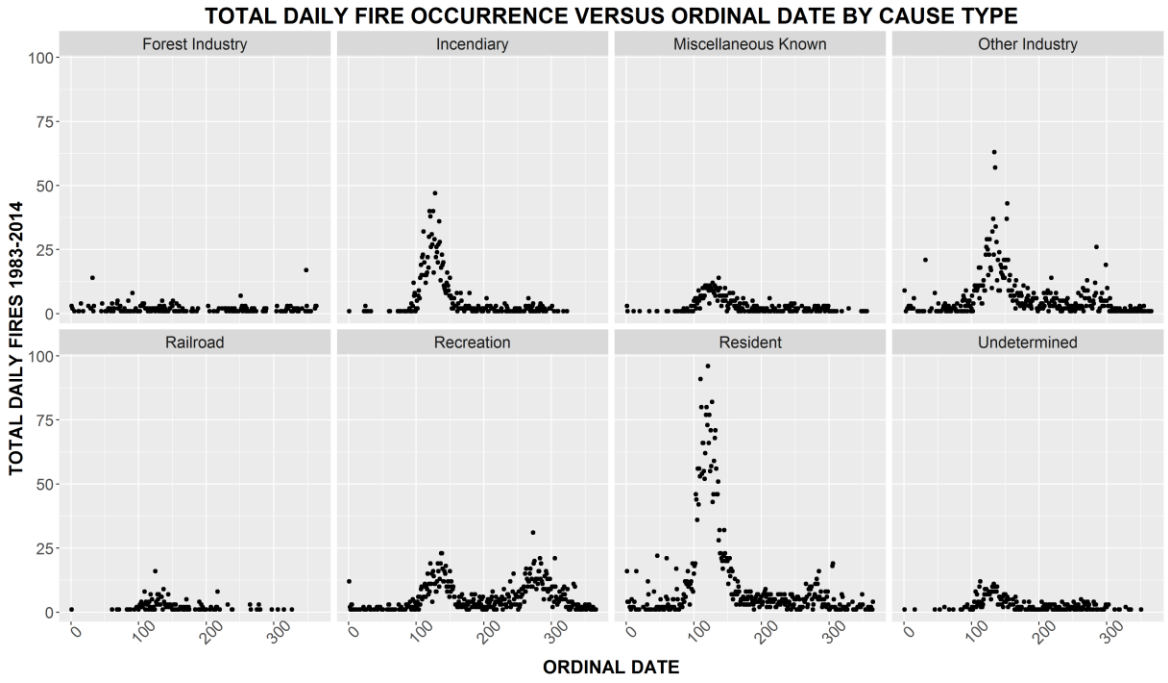

Figure 9: Summed fire starts for any given ordinal date for study period by each cause type.

Other variables considered were HOLDIAYS, DAY OF WEEK and FIRE BANS. People's activities vary based on the day of the week and whether it is a holiday or not. It was speculated that people are more likely to go on short trips in forested areas on weekends and holidays (especially if the weather is nice), and more likely to be doing industrial activity on the weekdays. Upon visual inspection of the data, no obvious trend between holidays and non-holidays was observed, although there is a notable difference in the spring in ecoregions 138 and 1452 (Figures 10 and 11). Ecoregion 138 has more fires on non-holidays in the spring and ecoregion 1452 has more fires on holidays in the spring. The empirical CDF graphs show a very slight difference between holidays and holidays with non-holidays

having about two percent more zeros than the holidays. It was decided there was not enough of a difference between holidays and non-holidays to include in the candidate models.

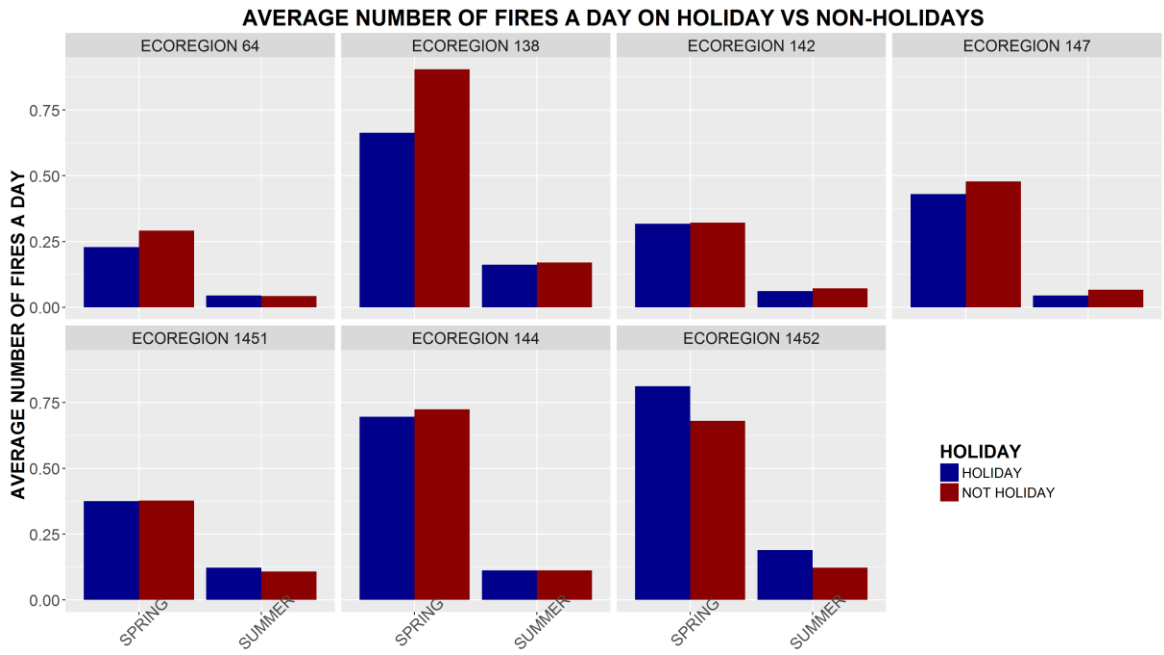

Figure 10: Average number of fires a day on holidays versus non-holidays by ecoregion and season.

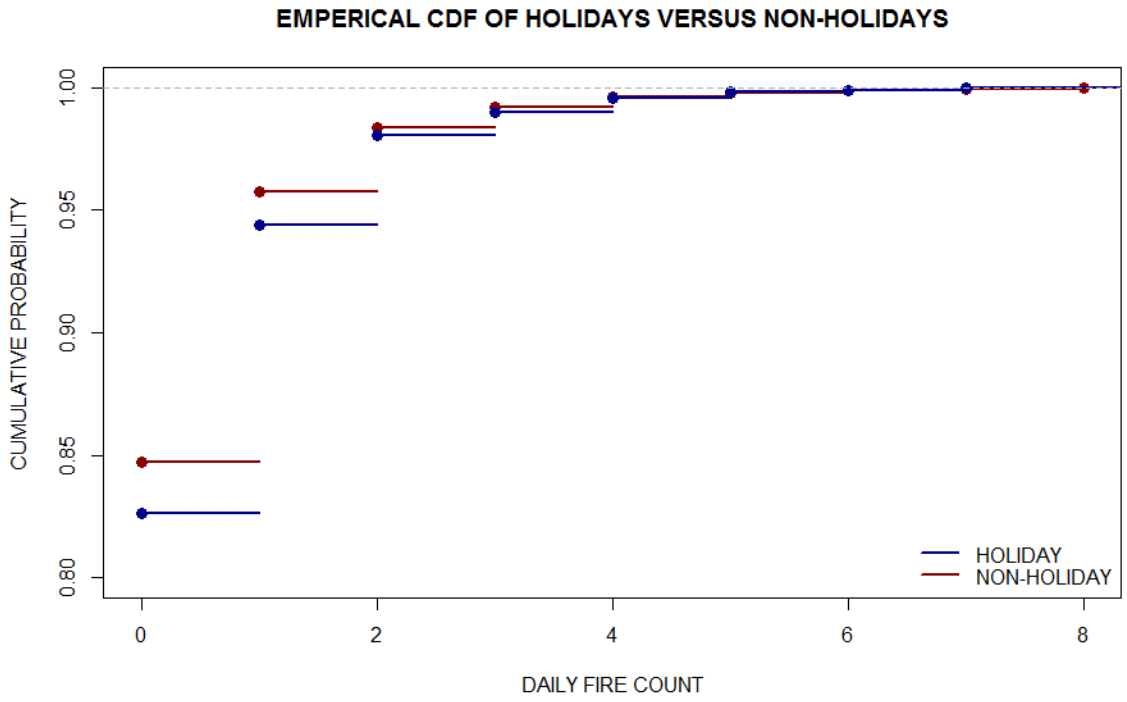

Figure 11: Empirical CDF of holidays versus non-holidays for all ecoregions and seasons together. The x-axis was cut off at eight as there are very few values greater than eight. Both data sets have between 80 and 85 percent zeros, so the y-axis is truncated as well.

Visual inspection of number of fires versus day of the week showed more weekend fires in most ecoregions to varying degrees, but only in the spring. Ecoregions 1451, 144 and 1452 (Figure 12) showed the most increase in fires on the weekends. There was very little difference by day of the week in the summer in any ecoregion.

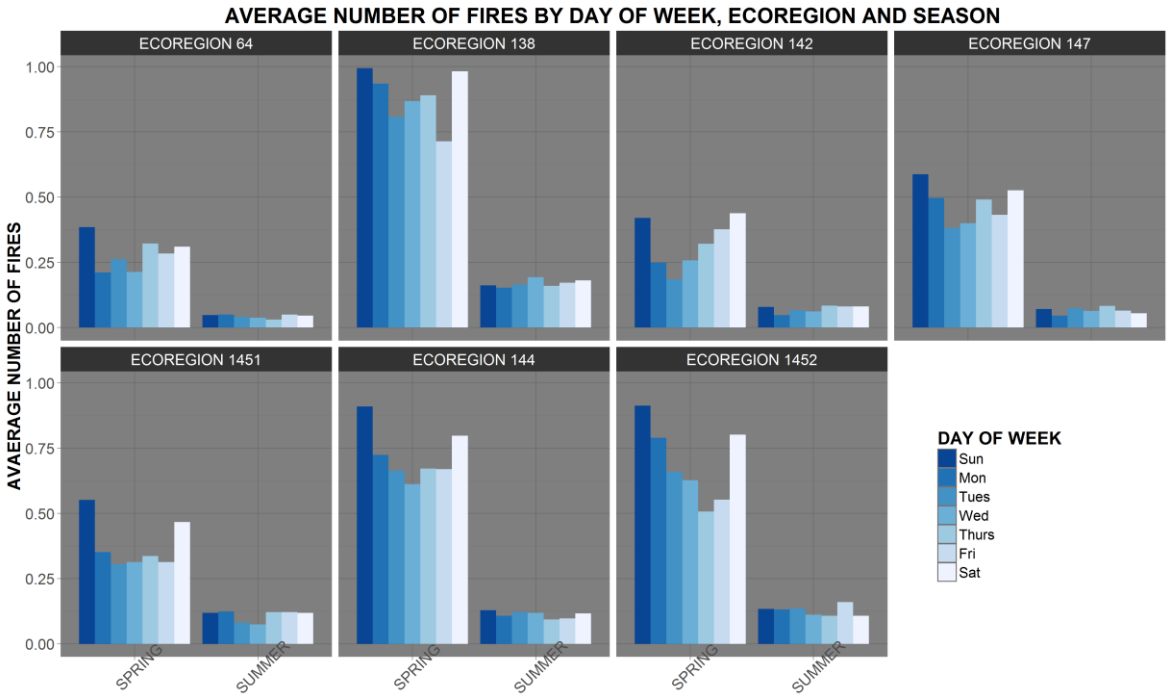

Figure 12: Total number of fire starts by day of week, season and ecoregion for study period.

Logically, fire bans should decrease the number of human-caused fires, and viability of including this data in the model was explored. Fire bans occur on days that have a very high likelihood of fire occurring if ignitions are present, and therefore should only be compared to other days of very high likelihood of fire. In order to have consistent comparisons between the number of fires starts on fire ban days and non-fire ban days, it was decided to only look at days with FFMCs of greater than 70. As a rule of thumb fires do not occur at FFMCs of lower than 75. An FFMC of 70 was chosen as the cut off to take into account the heterogeneity of FFMCs throughout the ecoregions and to include as much of the limited fire band data as possible (Table 4). This also removed the anomalously low FFMC fire ban days from the study, and resulted in comparisons of relatively equivalent FFMCs (fire

ban days then had a mean FFMC of 84.8, and non-fire ban days a mean FFMC of 83.1). On average, there are more fires on ban days (Figures 13 and 14).

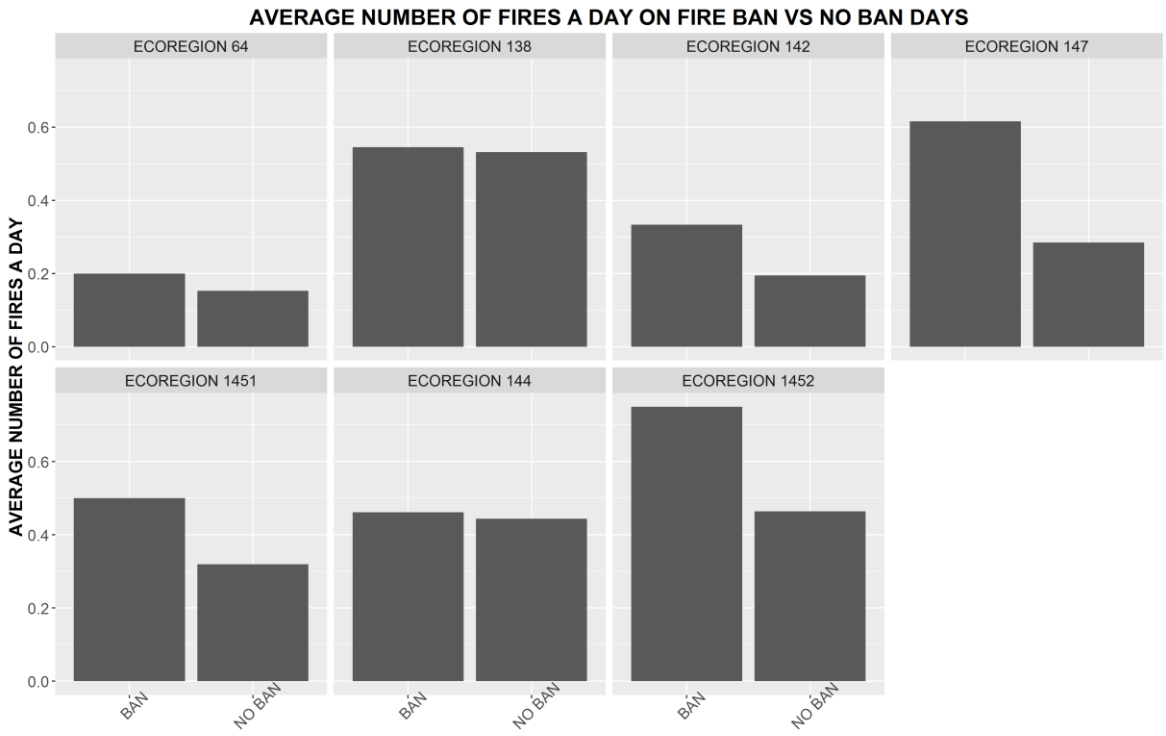

Figure 13: Average number of fires on days with fire bans and days without fire bans from 2006-2014 for days with FFMCs of 70 or greater.

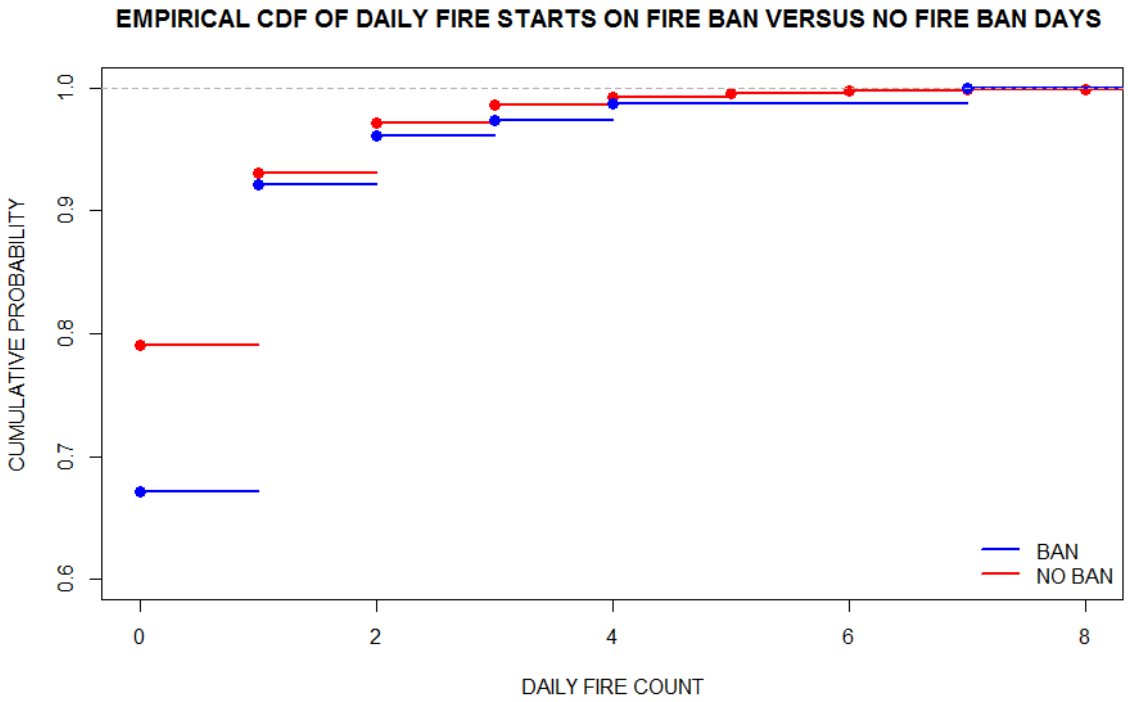

Figure 14: Empirical CDF graph of daily fire starts on ban s versus non-ban days from 2006-2014 for days with FFMCs of 70 or greater. The y-axis is truncated for plot clarity.

Table 4: Total number of days with fire bans in each ecoregion for days with FFMC 70 or greater.

Ecoregion	Fire ban days
64	5
138	11
142	12
147	13
1451	10
144	13
1452	12

It was decided not to keep the fire ban data in the model because of the limited time frame of the data (2006 onwards), low sample size, and only partial spatial overlap with the ecoregions.

Ultimately, the only human factor considered in the candidate models was DAY OF WEEK; all the other factors did not indicate that they would be good predictors of fire starts.

Model Type

Many different types of models exist, with varying levels of complexity and each with its own strengths and weaknesses. Previous studies on wildfire occurrence prediction have used Generalized Linear Models (GLMs)(Wotton & Martell, 2005; Wotton et al., 2010), zero-inflated models, hurdle models, Generalized Linear Mixed Models (Xiao et al., 2015) and logistic generalized additive models (Morin, 2015)

Generalized Linear Model

GLMs require a link function which relates the random component (distribution of the dependent variable) and the systematic component (the independent variables). The link function helps deal with non-normally distributed data. Since the fire occurrence data in this study is not normally distributed, this made using a GLM the obvious choice. Additionally, this model is relatively simple to understand and implement. However, when modeling forest fire occurrence, there tends to be a disproportionately high number of zero fire days, which may not be accurately modeled with a GLM alone. There are a several methods to deal with this problem including zero-inflated models and hurdle models (Mullahy, 1986; Xiao et al., 2015).

Distribution and Link Function

The distribution of daily number of fire starts has a few important properties. Firstly, the data is not normally distributed. Secondly, the data is count data, rather than continuous, further limiting the distributions that are appropriate.

Common count data distributions are Poisson and Negative Binomial (NB). The Poisson distribution represents the number of events that could happen in a given period of time (Hu et al., 2011). The following is the probability mass function for the Poisson distribution:

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

Where y represents a count variable, and λ represents both the variance and the mean, and is greater than zero. In the Poisson distribution the variance is equal to the mean, and over-dispersion is when the variance exceeds the mean (Mullahy, 1986; Xiao et al., 2015).

The NB distribution looks very similar to the Poisson distribution, but with a longer and flatter tail, resulting in over-dispersion. The NB distribution deals with over-dispersion by adding dispersion parameter (Hu et al., 2011). Therefore instead of the variance equaling the mean (λ), the variance in the NB is $\lambda + \lambda\theta^2$, where θ is the dispersion parameter that accounts for the unobserved heterogeneity in the data (MacNeil et al., 2009; Xiao et al., 2015). Both the Poisson and NB distributions use a log link. The following the probability mass function of the NB distribution (Xiao et al., 2015):

$$P(Y = y) = \frac{\Gamma(y + \theta^{-1})}{\Gamma(y + 1)\Gamma(\theta^{-1})}\left(\frac{\theta^{-1}}{\theta^{-1} + \lambda}\right)^{\theta^{-1}}\left(\frac{\lambda}{\theta^{-1} + \lambda}\right)^{y}$$

Much of the previous FOP work done in Canada was done using a Poisson distribution, as this distribution was considered to adequately represent the distribution of daily fire starts (Martell et al., 1987; Wotton et al., 2003; Wotton et al., 2010) and is simpler than the NB. However, more recent work done in China showed using a NB distribution resulted in better fit models based on AIC and BIC (Xiao et al., 2015). With the increased computing power over the past decade and pre-made

functions in R, computational simplicity is no longer as great a virtue in a model as it once was. This does not mean the model should be needlessly complex, as simple models are easier for practitioners to use. Also more complex models do not necessarily make for better a better representation of reality or for better predictions.

The study data did not meet the assumptions of the Poisson distribution as the variance of daily fire start count is 0.522 and the mean is 0.233, clearly not equal, as required by the Poisson distribution. Since the NB distribution has a term to deal with this over dispersion, it was decided to use the NB distribution in this study. Additionally, the data set has a long tail (Table 5), another indication that the NB distribution is more appropriate than the Poisson distribution. After about 10 fires per ecoregion per day, the number of fire starts becomes very irregular, with anywhere from 0 to 2 fires starts per day up to 25 fires per day. This demonstrates that the days with extremely high fire starts are very rare (making them rather difficult to model).

Table 5: Counts of days with each number of fire starts in the study data versus a Poisson distribution.

Number of Fires	0	1	2	3	4	5	6	7	8	9	10	11+
Poisson	23145	5411	632	49	3	0	0	0	0	0	0	0
NB	24760	3033	908	325	126	51	21	9	4	2	1	1
Actual Fires	24728	3238	783	254	118	53	26	19	8	4	4	6

Table 5 shows the distribution of the number of days with each number of fire starts in the study data versus a theoretical Poisson and NB distributions. To create the theoretical NB distribution, the fitdistr function was used to find the mean and dispersion parameter (0.234 and 0.257 respectively) of the best fit NB distribution to the study data. These values were entered into the dnbinom

function which produces a NB probability distribution. This distribution was multiplied by 29, 241 (the number of observations). The same process was repeated for the Poisson distribution but with the dpois function instead of the dnbinom function and no dispersion parameter (R Core Team, 2015).

In the theoretical Poisson data set, there are very few days with four fires (three days) and there are no days with five or more fires (Table 5). In comparison, the study data has 118 days with four fires, 53 days with six fires and so on up to ten fire starts a day. After that there a total of six days with fire starts ranging from 11 to 25. This demonstrates a much flatter tail than expected in a Poisson distribution.

There are a disproportionately large number of days with zero fire starts in the study data (Table 5). In the study data there are 24,728 zero fire days, and in the theoretical Poisson data set there are around 1,600 fewer zeros at 23,145. This prompted the exploration of hurdle models and zero-inflated models to see if these fit the data better.

The NB theoretical distribution had a much better fit with the study data. While the NB theoretical distribution has a much flatter tail than the Poisson distribution, it is not as flat the study data. It also does not have as many total days in the tail. However, the fit is still very good, justifying the use of the negative binomial distribution.

Zero-Inflated Model

Zero-inflated models (ZIM) use two regressions. The first is a binary prediction of fire or no fire. If this first equation predicts at least one fire, a second regression is used to predict how many fires, which can also predict a zero. This method can be used with a GLM to fit the second regression. Zeros can be from 'structural' sources or 'sampling' sources. In other words, the inflated number of

zeroes can be predicted unrelated to the distribution of the rest of the data, and as part of the data distribution, meaning zeros have two opportunities to be predicted. Sample sources are zeros that occur from random chance as a regular part of the distribution. For example, for a given FFMC value, a certain percent of fire brands would be expected to fizzle out and result in a fire (think of how many matches it takes to light a camp fire). The structural sources are sources that occur as part of the structure of the data (Hu et al., 2011). For example, these would be from situations with an FFMC that is receptive to ignition, but there are no ignition sources, perhaps in situations such as fires bans.

$$P(Y = y) = \begin{cases} p + (1-p)f(0) & y = 0 \\ (1-p)f(y) & y > 0 \end{cases}$$

Where p is the probability of a 'structural' 0, and (1-p) is the probability of belonging to the 'sample' distribution. In this thesis, Y is the number of daily fire starts per ecoregion (Cameron & Trivedi, 2013; Ridout et al., 1998; Xiao et al., 2015).

Hurdle Model

Hurdle models (HM) are very similar to ZIM, except that in hurdle models, all zero data are from 'structural' sources. In other words all the zeros are unrelated to the distribution of the non-zero days, therefore if the value is not zero, a truncated distribution (with no zeros) is used to predict the non-zero days (meaning a zero cannot be predicted at this stage) (Hu et al., 2011; Mullahy, 1986)

$$P(Y = y) = p \qquad if\ y = 0$$

$$P(Y = y) = (1-p)w\frac{f(y)}{1-f(0)} \qquad if\ y > 0$$

Where p is the probability of a zero and (1-p) is a probability of being in the truncated 'sample' distribution (Cameron & Trivedi, 2013; Xiao et al., 2015).

The Candidate Models

Candidate models provide a way to compare a predefined set of models. Ideally variables and models are selected *a priori*, based on current ecological and experiential knowledge, information gathered through data exploration, and logic (Burnham & Anderson, 2003). In the case of this thesis, four groups of models were created, with the first set created *a priori*. The best model was selected from this group and the second set of models was created based on this model, and so on. This method was chosen because it allows for a manageable number of candidate models while still exploring FWI System codes and indices, human variables, seasonality separately.

Part 1 consists of the first group of candidate models, created *a priori,* and explores the FWI System codes and indices as independent variables. Part 2 uses the best candidate model from part 1 with the addition of a human factor for an independent variable. Part 3 uses the best model from part 2, with the addition of seasonality independent variables. Part 4 uses the best model from part 3 and compares models with different models forms and dependent variable distributions.

The following section describes the candidate models. All models in parts 1 through 3 use a GLM with a negative binomial distribution. Models in part 4 have model form and distribution specified.

Part 1: FWI System

Candidate models M1 through M4 explore how effective various FWI System codes and indices are at predicting daily fire occurrence.

All models in part 1 included ECOREGION, SEASON2 and an ECOREGION/FFMC interaction as these variables were indicated to be of importance in previous research and in the preliminary data exploration.

MODEL 1 (M1):

$$\ln(N_{HUM}) = \alpha_0 + \alpha_1 * \text{ECOREGION} + \alpha_2 * \text{FFMC} + \alpha_3 * \text{FFMC} * \text{ECOREGION} + \alpha_4 * \text{SEASON2} + \alpha_5 * \text{DMC} + \alpha_6 * DC$$

Where N_{HUM} is the daily number of human-caused fire starts.

MODEL 2 (M2):

In M2 BUI was used to replace both DMC and DC as BUI is a weighted average of the two (heavily weighted towards DMC).

$$\ln(N_{HUM}) = \alpha_0 + \alpha_1 * \text{ECOREGION} + \alpha_2 * \text{FFMC} + \alpha_3 * \text{FFMC} * \text{ECOREGION} + \alpha_4 * \text{SEASON2} + \alpha_7 * \text{BUI}$$

MODEL 3 (M3):

In M3 FWI replaces BUI since FWI is highly correlated with BUI.

$$\ln(N_{HUM}) = \alpha_0 + \alpha_1 * \text{ECOREGION} + \alpha_2 * \text{FFMC} + \alpha_3 * \text{FFMC} * \text{ECOREGION} + \alpha_4 * \text{SEASON2} + \alpha_8 * \text{FWI}$$

MODEL 4 (M4):

M4 is the same as M3 with the addition of DC. DMC is highly correlated with FWI, but DC is not and so it seemed prudent to see the effect of adding a measurement of the longer term drying of the larger or deeper fuels.

$$\ln(N_{HUM}) = \alpha_0 + \alpha_1 * \text{ECOREGION} + \alpha_2 * \text{FFMC} + \alpha_3 * \text{FFMC} * \text{ECOREGION} + \alpha_4 * \text{SEASON2} + \alpha_8 * \text{FWI} + \alpha_6 * \text{DC}$$

Part 2: Human Variables

MODEL 5 (M5):

The selected model from part 1 (M3) was used to test the one human factor considered in this thesis. M5 is the same as M3 but with the addition of the variable DAY OF WEEK.

$$\ln(N_{HUM}) = \alpha_0 + \alpha_1 * \text{ECOREGION} + \alpha_2 * \text{FFMC} + \alpha_3 * \text{FFMC} * \text{ECOREGION} + \alpha_4 * \text{SEASON2} + \alpha_8 * \text{FWI} + \alpha_9 * \text{DAY OF WEEK}$$

Part 3: Seasonality

The next 5 candidate models (M6-M10) explored seasonality, and how to best include it in the model. Additionally, an interaction between FFMC and seasonality variable was included for variables SEASON2 and SEASON3. ORDINAL DATE and WEEK OF YEAR had so many levels that an interaction term would have been impractical. While a particular FFMC value may mean something different in terms of fire likelihood in the spring versus the summer, it is unlikely to mean something different from one day to the next.

MODEL 6 (M6):

M6 is the same as M5 but instead of SEASON2, SEASON3 was used.

$$\ln(N_{HUM}) = \alpha_0 + \alpha_1 * \text{ECOREGION} + \alpha_2 * \text{FFMC} + \alpha_3 * \text{FFMC} * \text{ECOREGION} + \alpha_8 * \text{FWI} + \alpha_{10} * \text{SEASON3}$$

MODEL 7 (M7):

M7 is the same as M6, with the addition of an interaction term between FFMC and SEASON3.

$$\ln(N_{HUM}) = \alpha_0 + \alpha_1 * \text{ECOREGION} + \alpha_2 * \text{FFMC} + \alpha_3 * \text{FFMC} * \text{ECOREGION} + \alpha_8 * \text{FWI} + \alpha_{10} * \text{SEASON3} + \alpha_{11} * \text{FFMC} * \text{SEASON3}$$

MODEL 8(M8):

M8 is the same as M6, except with WEEK OF YEAR instead of SEASON3.

$$\ln(N_{HUM}) = \alpha_0 + \alpha_1 * \text{ECOREGION} + \alpha_2 * \text{FFMC} + \alpha_3 * \text{FFMC} * \text{ECOREGION} + \alpha_8 * \text{FWI} + \alpha_{12} * \text{WEEK OF YEAR}$$

Model 9 (M9):

M9 is the same as M8 except with ORDINAL DATE instead of SEASON3.

$$\ln(N_{HUM}) = \alpha_0 + \alpha_1 * \text{ECOREGION} + \alpha_2 * \text{FFMC} + \alpha_3 * \text{FFMC} * \text{ECOREGION} + \alpha_8 * \text{FWI} + \alpha_{13} * \text{ORDINAL DATE}$$

Model 10 (M10):

M10 is the same as M5 except with the addition of an interaction term between FFMC and SEASON2.

$$\ln(N_{HUM}) = \alpha_0 + \alpha_1 * \text{ECOREGION} + \alpha_2 * \text{FFMC} + \alpha_3 * \text{FFMC} * \text{ECOREGION} + \alpha_4 * \text{SEASON2} + \alpha_8 * \text{FWI} + \alpha_{14} * \text{FFMC} * \text{SEASON2}$$

Part 4: Model Types

The results from part 3 were used to investigate different model formats, and dependent variable distributions, to see if other model formats can improve the results. The independent variables from M6 were using to run a model with Poisson and NB links for each of standard GLM, hurdle and zero-inflated models. All of these models have been used to predict fire occurrence in past research, and

the hurdle and zero-inflated models would, in theory, be expected to help with the issue of the large number of zeros in fire occurrence data.

Model 11 (PM6)

PM6 (and the rest of the models in this section) use the same equation (independent variables) as M6, the only difference being PM6 uses a GLM assuming a Poisson distribution of the dependent variable rather than a NB distribution.

Model 12 (HNM6)

HNM6 uses a hurdle model assuming a NB distribution of the dependent variable.

Model 13 (ZNM6)

ZNM6 uses a zero-inflated model assuming a NB distribution of the dependent variable.

Model 14 (HPM6)

HPM6 uses a hurdle model assuming a Poisson distribution of the dependent variable.

Model 15 (ZPM6)

ZPM6 uses a zero-inflated model assuming a Poisson distribution of the dependent variable.

Model Selection

Once the candidate models were created, they needed to be compared and evaluated. Several evaluation methods were used. Some methods evaluate the fit of the model to the data used to create it (SD), some methods evaluate the predictive ability of the model (its ability to predict the TD). Two commonly used methods of evaluating model fit are the AIC and deviance.

Akaike Information Criterion

AIC is used to give a relative indication of the goodness of fit of a model (Burnham et al., 2010; Maindonald & Braun, 2010). It cannot give an indication of the absolute model fit, but can be used to compare models that use the same data set. By penalizing for number of model parameters, the AIC balances goodness of fit with model complexity (Fortin & DeBlois, 2007; Xiao et al., 2015). We want the most parsimonious model that represents the data, in order to avoid fitting model to data noise rather than the data signal. Additionally, parsimonious models tend to be easier to understand and require the collection of less data (Vandekerckhove et al., 2014). AIC is calculated as follows:

$$AIC = -2logML + 2K$$

Where ML is the maximum likelihood and K the model's total number of parameters (Akaike, 1974; Burnham et al., 2010). However, the model with the smallest AIC is not guaranteed to be the only model with useful outputs, or the best model. It is just most likely to be the best model (Burnham et al., 2010).

Deviance

The residual deviance (referred to as only deviance) can also be used to compare model goodness of fit and was calculated for each model except those in part 4. Deviance is very similar to AIC except with no correction for the number of model parameters.

$$Deviance = -2(logML\ saturated\ model - logML\ fitted\ model)$$

Deviance compares the candidate model (fitted model) to the saturated model, where the saturated model has one data point for each observation and perfectly fit the data set. A smaller deviance represents a model with a better fit, as it is closer to the saturated model (Clewer, 1999).

Model Predictions

The following tests are used to compare the predictive ability of the candidate models. The tests were run on both the SD and TD, but the results from the TD were used to decide which model had the best predictive capacity.

Bootstrapping

All predictive model results were bootstrapped using a random selection of 30% of the TD, ran 200 times. The mean of the bootstrapped results were recorded for RMSE, Pearson product-moment correlation coefficient, Root-mean-square error and the absolute difference between actual and predicted number of fires for each model.

Correlation

A Pearson product-moment correlation coefficient (PC) was run on the number of fires predicted by each model versus the actual recorded number of fire starts for each ecoregion per day on both the SD and TD sets. A PC was chosen rather than a Spearman's rank correlation coefficient because the results are expected to be linearly related. Spearman's correlation could also be used as this relationship is also monotonic, however Spearman's correlation allows the relationship to be non-linear, which is not want is wanted here (Zar, 2010). Ideally, the regression should have a slope of 1 with an intercept of zero. The correlation coefficient can be between -1 and 1 (Clewer, 1999). In this thesis, a coefficient of 1 would be ideal.

Root-Mean-Square Error

Root-mean-square error (RMSE) of the predicted values versus the actual values for each of the candidate models using the TD was calculated. RMSE gives an indication of the spread of the predicted values around the actual values. These RMSE was calculated 200 times for each candidate

model, using 30% of the data each time. These 200 values were checked for normality, and equal variance.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

Mean Absolute Error

While correlation gives a good indication of how well the predicted number of fire starts linearly relates to the actual number of fire starts, it does not give an indication of how close the predictions are to the actual number of fire starts. To evaluate this, predicted number of fire starts were subtracted from the actual number of fires for each ecoregion and date. The absolute value was taken of the difference, then summed and averaged. This gives the average daily difference between the actual number of fires and the predicted number of fires. From here on this result will be call AP. Ideally, the best model would have the largest correlation coefficient and the smallest AP.

ANOVA

An ANOVA test was run on the RMSE from all the candidate models to compare the means of the APs from candidate models to see if they differed significantly. A Tukey's test was then performed to see which models had significantly different predictions. Tukey's test is like a t-test but for multiple comparisons (Zar, 2010).

Results

The following section displays the results and comparisons of the candidate models divided into four parts; FWI System, human variables, seasonality and model type.

Part 1: FWI System Variables

Candidate models M1 through M4 were compared and the best model selected. All evaluation criteria were considered to gain an understanding of the fit, the predictive ability of the model, and to watch for over-fitting. Extra weight was given to the predictive ability of the model when considering which model to choose. However, if one model was not clearly the best based on all criteria, the simplest model was chosen (the model with the fewest independent variables). If the models were equal in simplicity RMSE was used to make the final decision. The PC SD and AP SD were only included to look for anomalies between the test and sample data sets.

Table 6: Goodness of fit and predictive capability of models M1-M4 for all ecoregions together.

	M1	M2	M3	M4
AIC	19160	19158	18986	18988
DEV	9603.9	9604	9612.7	9610.4
PC SD	0.447	0.447	0.444	0.444
AP SD	0.301	0.301	0.300	0.300
PC TD	0.394	0.399	0.488	0.489
AP TD	0.327	0.326	0.317	0.317
RMSE TD	0.744	0.742	0.706	0.705
RMSE TD 18	0.726	0.724	0.702	0.702
PC TD 18	0.395	0.400	0.459	0.459

Model 4 was chosen as it has the lowest AIC and RMSE, the fewest variables (adding DC to create M4 added an independent variable with very little improvement to the predictions (Table 6)). However, caution should be used as M1 and M2 did have better deviances and better PCs for the sample data than M3. Additionally, M3 had a higher PC TD than PC SD.

M1 and M3 were compared in more detail, as it was unexpected that FWI (M3) would be the better predictor of fire starts than DMC and DC. May 3rd, 1998 in Ecoregion 1452, is a high fire day (18 fires), with very different fire occurrence predictions numbers depending on the model. Because of the unusually high number of fires, this one day has the potential to greatly influence the RMSE. To test this influence, it was removed and the RMSE TD and PC TD were recalculated (RMSE TD 18 and PC TD 18 respectively). Although M3 still scores better on the RMSE TD 18 and PC TD 18 than M2, the difference is much smaller than with this day included. The 20 days with the biggest differences in predictions were also compared. Of these, M3 performed better on 14, demonstrating that it is not just one day that the model is predicting much better, but many days. This finding is in contrast to later examples, wherein one day with a greatly improved prediction accounted for most of the increase in the model's predictive ability.

M1 and M3 were also compared visually (Figures 15-17). The graphs were selected from the ecoregions and years with the largest difference between models. In general, M3 did a better job of catching the peaks in fire starts. On one day in 1998 in ecoregion 1452 there were 18 fires. M3 over-predicts this peak (predicting about 25 fires) but M1 does not catch this extremely large number of fires at all (predicting about 1.8 fires) (Figure 17).

Overall M3 was considered the best and used in part 2.

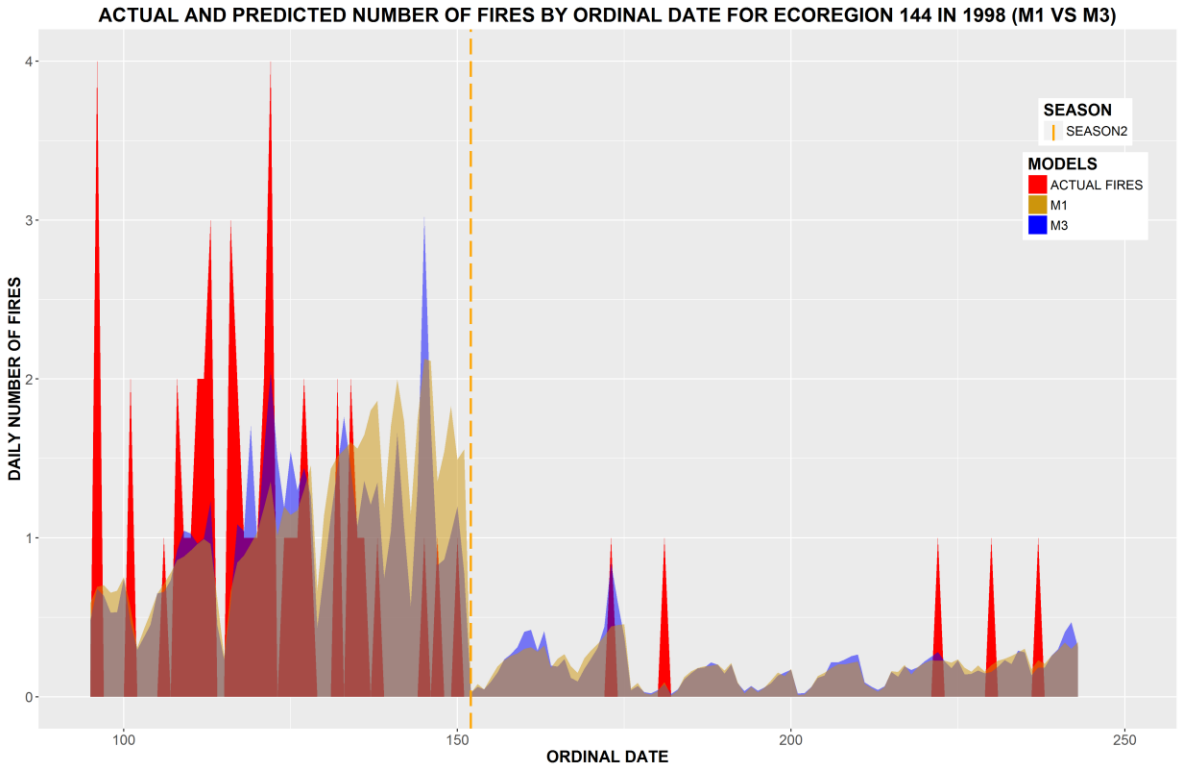

Figure 15: Actual daily fire starts compared to the predicted number of fires starts for M1 and M3 for ecoregion 144 in 1998. As the colours for the models are semi-transparent, the brown grey colour is where the two models overlap. The vertical dashed orange line represents the border between spring and summer, as described by the variable SEASON2 (June 1st).

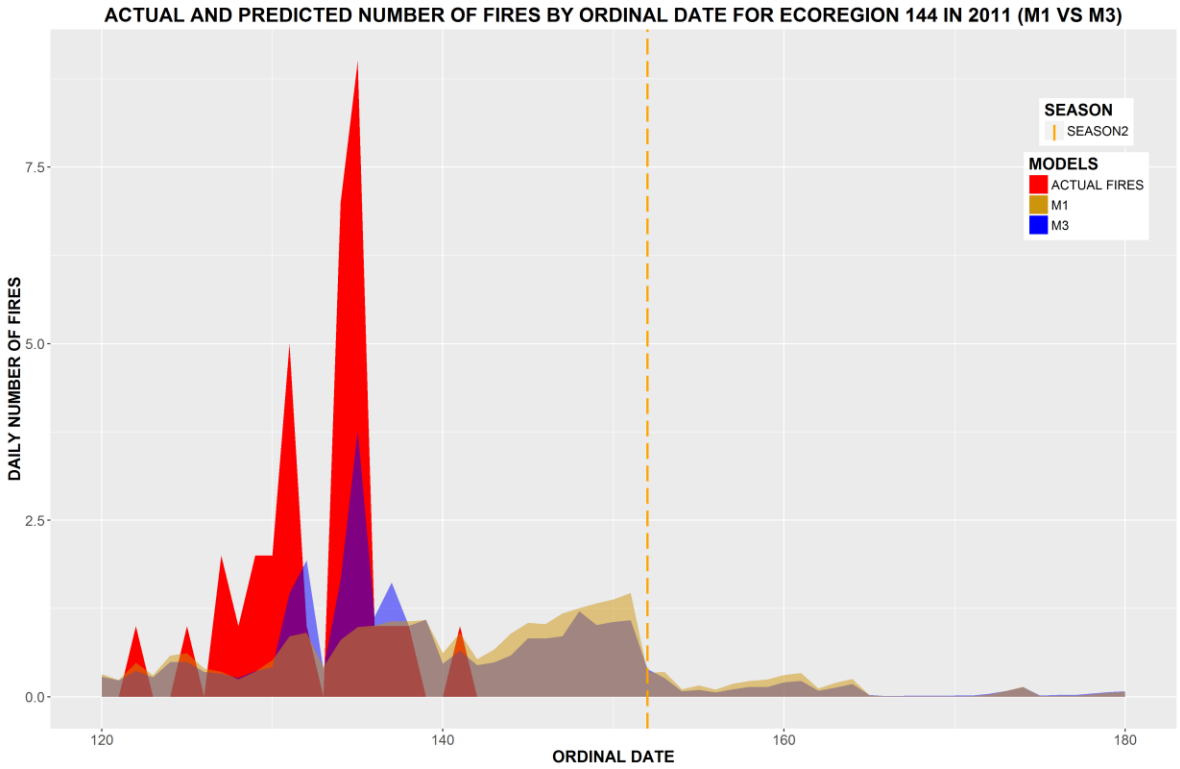

Figure 16: Actual daily fire starts compared to the predicted number of fires starts for M1 and M3 for ecoregion 144 in 2011. As the colours for the models are semi-transparent, the brown grey colour is where the two models overlap. The vertical dashed orange line represents the border between spring and summer, as described by the variable SEASON2 (June 1st).

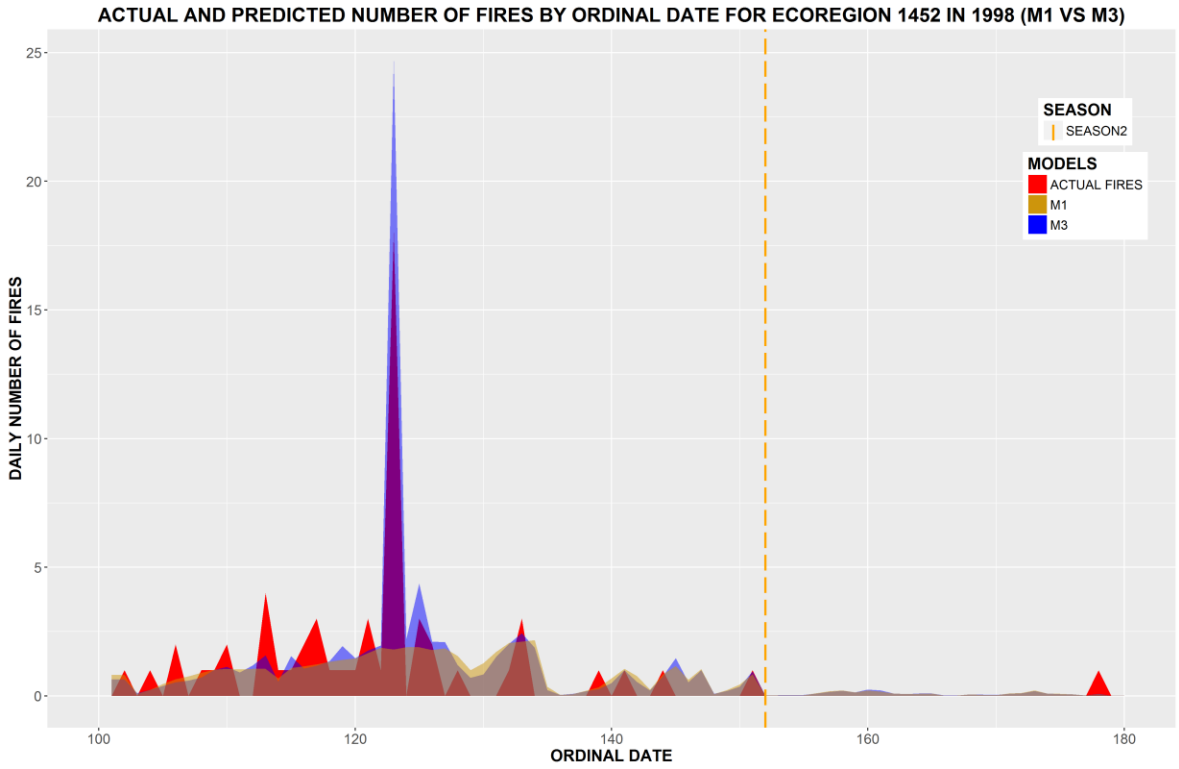

Figure 17: Actual daily fire starts compared to the predicted number of fires starts for M1 and M3 for ecoregion 1452 in 1998. As the colours for the models are semi-transparent, the brown grey colour is where the two models overlap. The vertical dashed orange line represents the border between spring and summer, as described by the variable SEASON2 (June 1st).

Part 2: Human Variables

Adding a variable for day of the week decreases the AIC, increases the deviance, decreases the PC of the TD, and increases the TD RMSE (Table 7). Since the RMSE in M3 was lower, and since M3 also has fewer independent variables, it was selected as the best model.

Table 7: Goodness of fit and predictive capability of models M3 and M5.

	M3	M5
AIC	18986	18971
DEV	9612.7	9626
PC SD	0.444	0.446
AP SD	0.300	0.299
PC TD	0.488	0.493
AP TD	0.317	0.317
RMSE TD	0.705	0.708
RMSE TD 18	0.702	0.699
PC TD 18	0.459	0.467

These models were explored in further detail as the PC TD and RMSE TD were in conflict with each other. The top 20 days with the biggest differences in predictions between models were compared. Of these, M5 predicted 10 days better and M3 predicted 10 better. However, the day with the biggest difference between models (May 3rd, 1998 in Ecoregion 1452) contributed greatly to the difference in RMSE between the two models, as the RMSE penalizes larger differences more heavily. On this day there were 18 fires in Ecoregion 1452. M3 predicted 24.67 fires and M5 predicted 29.32 fires -- a much larger over prediction (Figure 18). If this ecoregion day was removed from the analysis, M5 performed slightly better on all measurements (PC TD, AP TD and RMSE TD). This shows that it is one day that is largely driving the difference between RMSE of M3 and M5.

Furthermore, Figures 18 and 19 show the ecoregions and years with the largest difference between models, and they show very little other difference between the models besides the day with 18 fires. Since there is very little to choose from between the models, M3 was used going forward as it was the simplest.

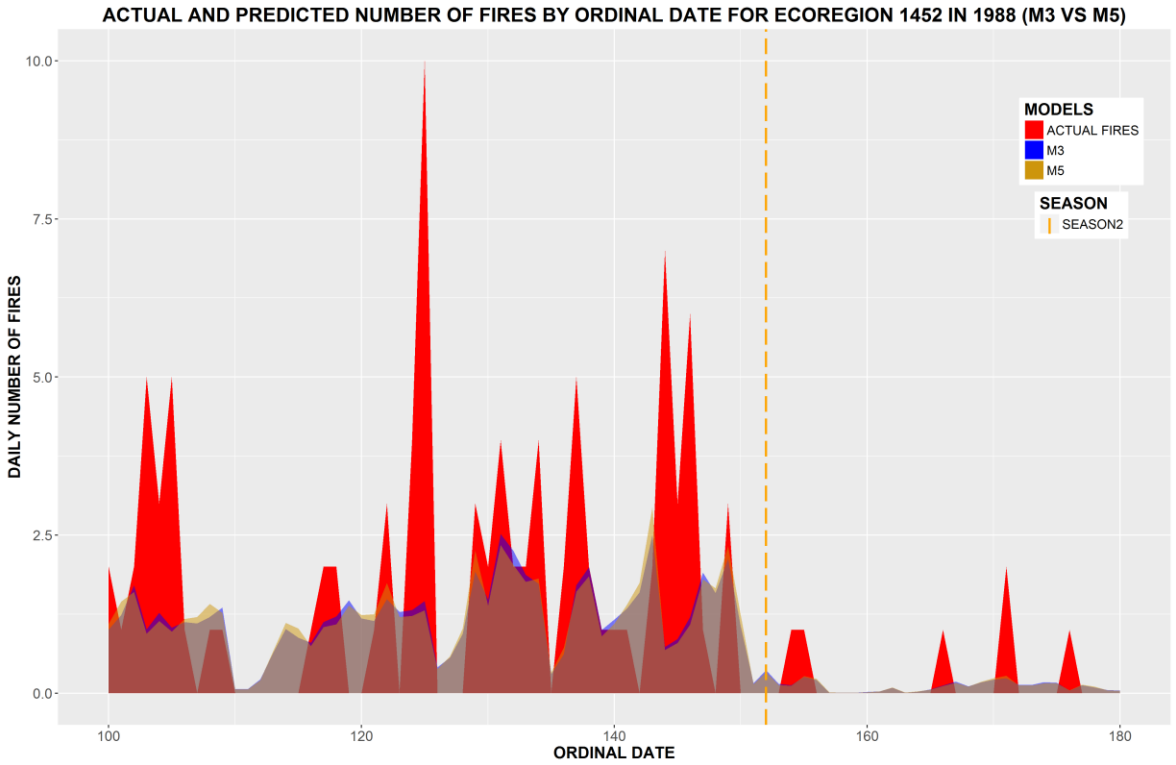

Figure 18: Actual daily fire starts compared to the predicted number of fires starts for M3 and M5 for ecoregion 1452 in 1988. As the colours for the models are semi-transparent, the brown grey colour is where the two models overlap. The vertical dashed orange line represents the border between spring and summer, as described by the variable SEASON2 (June 1st).

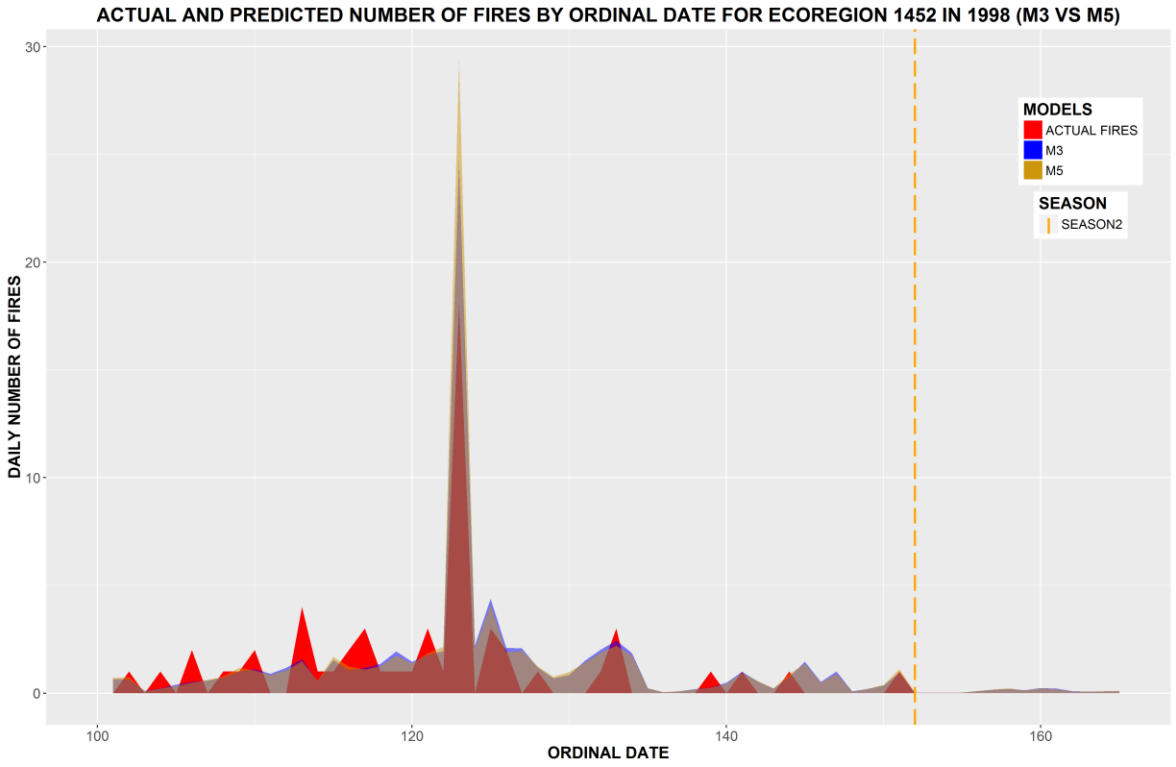

Figure 19: Actual daily fire starts compared to the predicted number of fires starts for M3 and M5 for ecoregion 1452 in 1998. As the colours for the models are semi-transparent, the brown grey colour is where the two models overlap. The vertical dashed orange line represents the border between spring and summer, as described by the variable SEASON2 (June 1st).

Part 3: Seasonality

In general, increasing the detail covered by the seasonality variable decreased the AIC (Table8) (in order of least detailed to most detailed, SEASON2, SEASON3, WEEK OF YEAR, ORDINAL DATE), with the exception of ORDINAL DATE (M9), which produced a model with a higher AIC than WEEK OF YEAR (M8).

However, this trend did not hold true when looking at the predictive capability of the models. The PC TD and the RMSE TD improved when moving from SEASON2 (M3) to SEASON3 (M6), but worsened when using WEEK OF YEAR (M8) and ORDINAL DATE (M9). Therefor SEASON3 was chosen as the variable best able to account for the seasonal trend in human-caused forest fire starts in Alberta.

Table 8: Goodness of fit and predictive capability of models M3, M5-M10.

	M3	M6	M7	M8	M9	M10
AIC	18986	18757	18739	18663	18731	18974
DEV	9612.7	9590.4	9581.2	9608.4	9556.3	9600.9
PC SD	0.444	0.469	0.470	0.489	0.503	0.445
AP SD	0.300	0.294	0.294	0.290	0.287	0.300
PC TD	0.488	0.504	0.504	0.488	0.479	0.488
AP TD	0.317	0.312	0.311	0.310	0.312	0.317
RMSE TD	0.705	0.697	0.698	0.707	0.713	0.706
RMSE TD 18	0.702	0.696	0.696	0.702	0.707	0.703
PC TD 18	0.469	0.475	0.475	0.460	0.449	0.458

This leaves M6 and M7 to compare, with M7 adding and interaction term between SEASON3 and FFMC. These models had very similar results with M7 having the lowest AIC and M6 being the simpler and having the better PC TD and RMSE TD. Therefore, moving forward M6 was chosen.

The day to day predictions were visually inspected, to gain an understanding of how selected models predict fire starts, with particular regard to seasonality. M3 was compared to M6 and M6 was compared to M7 on the ecoregions and years that had some of the largest differences between models on the predictions using the TD (Figures 20-22). In general the day-to-day predictions of all three models were very similar.

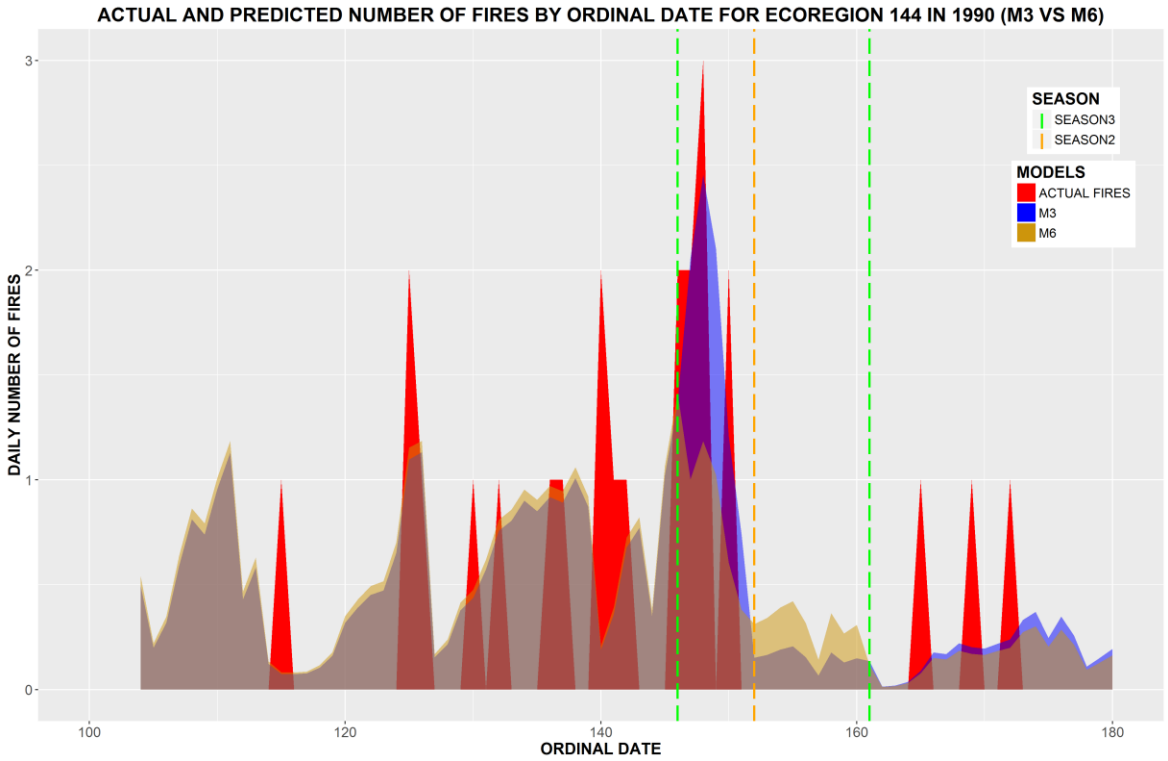

Figure 20: Actual daily fire starts compared to the predicted number of fires starts for M3 and M6 for ecoregion 144 in 1990. As the colours for the models are semi-transparent, the brown grey colour is where the two models overlap. The vertical green dashed line represents the start and end of the SEASON3 variable transition season, with the first line being the last day of the spring season and the second line being the first day of the summer season. The vertical dashed orange line represents the border between spring and summer, as described by the variable SEASON2 (June 1st).

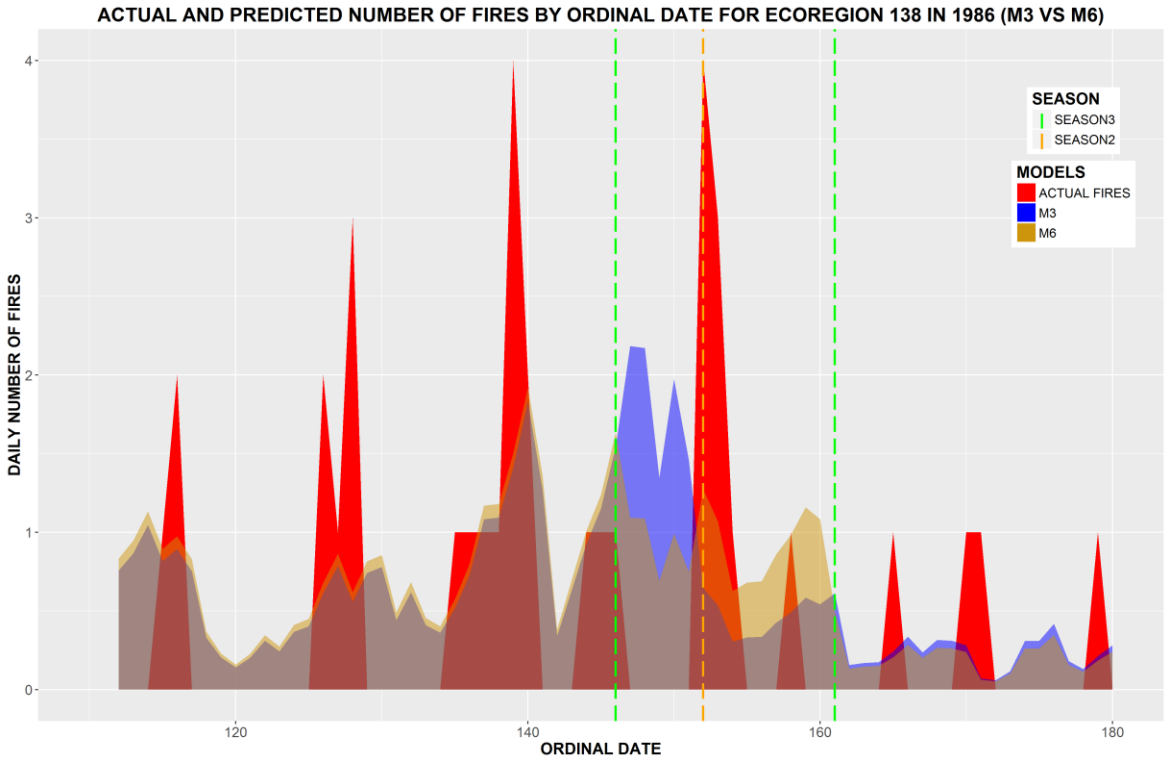

Figure 21: Actual daily fire starts compared to the predicted number of fires starts for M3 and M3 for ecoregion 138 in 1986. As the colours for the models are semi-transparent, the brown grey colour is where the two models overlap. The vertical green dashed line represents the start and end of the SEASON3 variable transition season, with the first line being the last day of the spring season and the second line being the first day of the summer season. The vertical dashed orange line represents the border between spring and summer, as described by the variable SEASON2 (June 1st).

In regard to capturing the seasonal trend, some years M6 captured this better, other years M3. M6 often over-predicted the number of fires after June 1st, unlike M3. However, some years had more daily fire starts in the latter half of the transition season than the summer season, and M6 was more able to predict this than M3. Overall it appears that M6 captures the variability of the transition time of year better than M3.

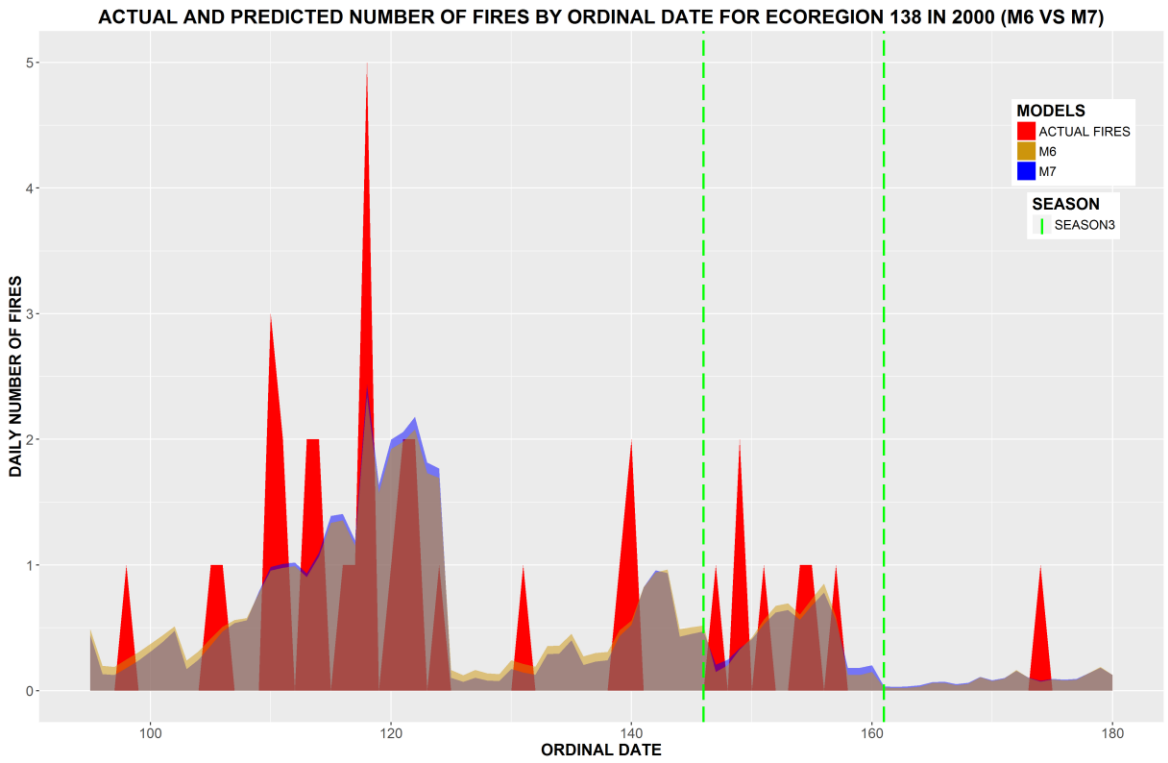

Figure 22: Actual daily fire starts compared to the predicted number of fires starts for M6 and M7 for ecoregion 138 in 2000. As the colours for the models are semi-transparent, the brown grey colour is where the two models overlap. The vertical green dashed line represents the start and end of the SEASON3 variable transition season, with the first line being the last day of the spring season and the second line being the first day of the summer season.

As suspected, M6 and M7 had very similar results (Figure 22). Since this figure showed the year with the most variation between models and still has very little difference between models, M6 was chosen over M7 as it was simpler (fewer independent variables).

Part 4: Model Types

Although all of the zero-inflated and hurdle models have better AICs than M6, none of them had better PC TDs or RMSE TD's (Table 9). PM6 had by far the worst AIC, but had only slightly lower PC TD and RMSE TD than M6. Since M6 had the lowest PC TD and RMSE TD, it was selected as the best model. There is no indication that the added complexity of hurdle and zero-inflated models improves the predictive ability of fire occurrence models that were tested.

Table 9: Goodness of fit and predictive capability of models in part 4.

	M6	PM6	HNM6	ZNM6	HPM6	ZPM6
AIC	18757	19245	18701	18680	18899	18918
PC SD	0.469	0.480	0.484	0.487	0.486	0.486
AP SD	0.294	0.294	0.293	0.291	0.293	0.292
PC TD	0.504	0.503	0.495	0.493	0.492	0.486
AP TD	0.312	0.312	0.313	0.312	0.312	0.313
RMSE TD	0.697	0.698	0.705	0.705	0.706	0.708
RMSE TD 18	0.696	0.697	0.698	0.700	0.698	0.699
PC TD 18	0.475	0.472	0.472	0.466	0.471	0.467

As with many of the previous models, much of the difference between RMSE TDs of the models is from the one day in ecoregion 1452 with 18 fires. If predictions are run on the TD without this day (May 3rd 1998 in ecoregion 1452), then the RMSE TD and PC TDs of all the models in part 4 are still very similar, with M6 still being slightly better. These results are called RMSE TD 18 and PC TD 18 in Table 9. Figure 23 shows an example of this, where M6 greatly over-predicts that day (about 22 fires) and HNM6 greatly under-predicts this day (about 8 fires). M6 is off by about 4 fires, HNM6 by about 8 fires. In a data set with very few days with fires starts above 8, a difference of 8 on one day can make a huge difference in the RMSE, even though many of the other predictions are very similar.

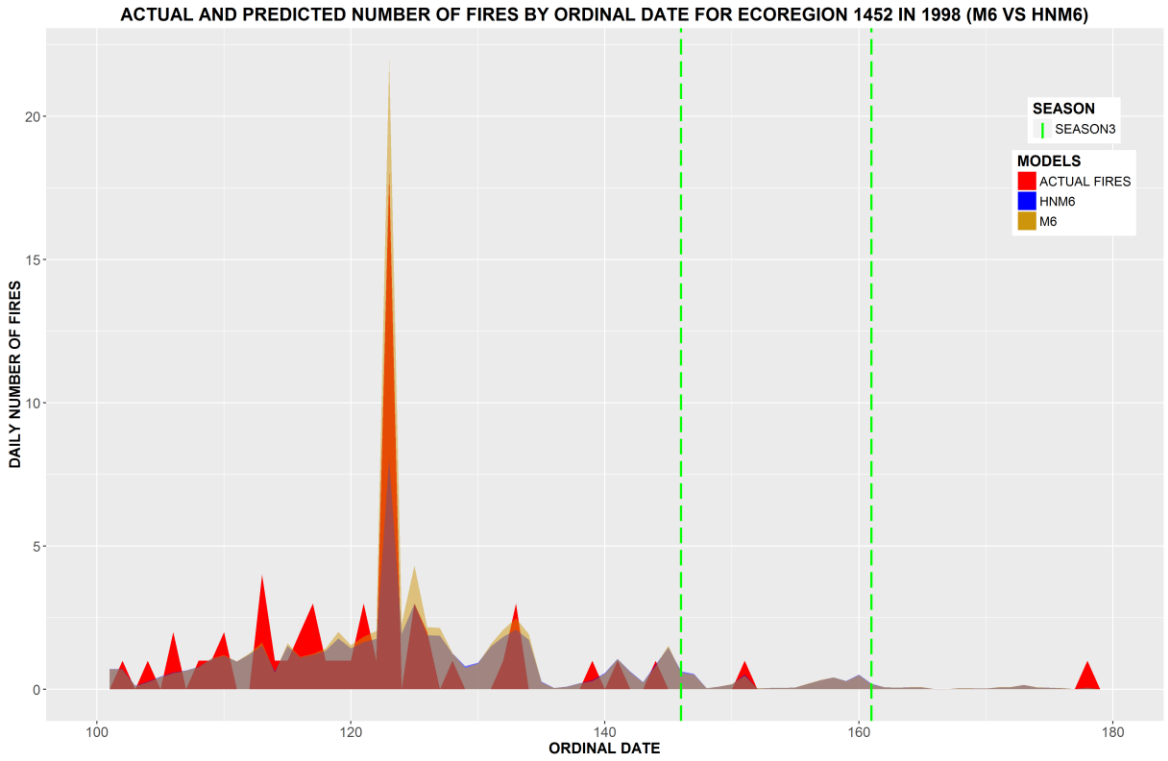

Figure 23: Actual daily fire starts compared to the predicted number of fires starts for M6 and HNM6 for ecoregion 1452 in 1998. As the colours for the models are semi-transparent, the brown grey colour is where the two models overlap. The vertical green dashed line represents the start and end of the SEASON3 variable transition season, with the first line being the last day of the spring season and the second line being the first day of the summer season.

Figures 25 and 26 both show the years with the second highest difference between the two models (the highest difference excluding 1998 in ecoregion 1452). There is very little difference between the models.

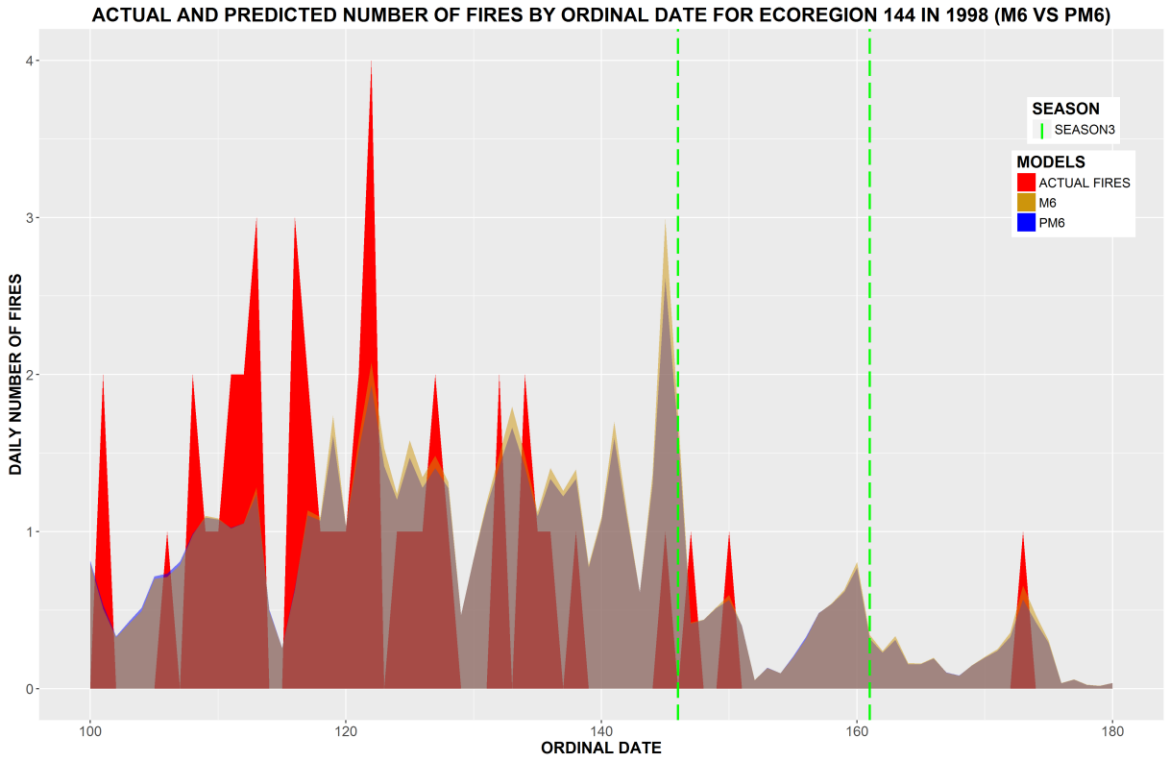

Figure 24: Actual daily fire starts compared to the predicted number of fires starts for M6 and PM6 for ecoregion 144 in 1998. As the colours for the models are semi-transparent, the brown grey colour is where the two models overlap. The vertical green dashed line represents the start and end of the SEASON3 variable transition season, with the first line being the last day of the spring season and the second line being the first day of the summer season.

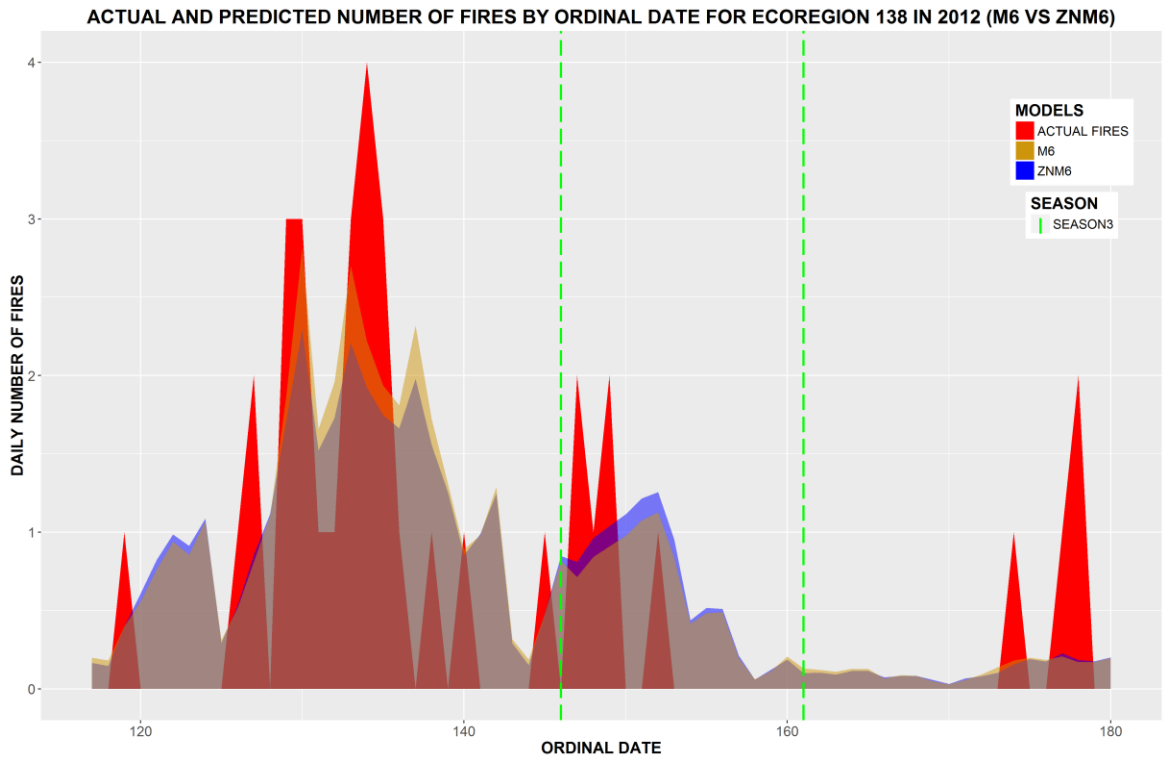

Figure 25: Actual daily fire starts compared to the predicted number of fires starts for M6 and ZNM6 for ecoregion 138 in 2012. As the colours for the models are semi-transparent, the brown grey colour is where the two models overlap. The vertical green dashed line represents the start and end of the SEASON3 variable transition season, with the first line being the last day of the spring season and the second line being the first day of the summer season.

ANOVA RMSE

Finally, all models' predictive abilities were assessed by comparing the bootstrapped RMSE TD using an ANOVA table (Table 10). The RMSE errors were approximately normally distributed for each model, and the variances of the RMSE were approximately the same for each of the models.

Table 10: ANOVA comparing RMSE of the TD for all models.

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Model	14	0.561	0.04007	6.899	3.83xe-14
Residuals	2985	17.336	0.00581		

The ANOVA shows that there is a significant difference between the bootstrapped RMSE TD of one or more of the models. Therefor the TukeyHSD test was run to find out which one(s).

The bootstrapped RMSE TDs only came out significantly different when comparing M1 or M2 to any of the other models. Otherwise, the results were not significantly different (including a comparison between M1 and M2).

Discussion

Overall Model Performance

The results of the TukeyHSD show that there is very little difference between the models, with the exception of M1 and M2, which are significantly worse than the rest. M3 and all later models improve significantly on M1 and M2, however no other model has predictions which are significantly different than M3's.

Since the maximum likelihood process of model fitting used by the GLM tries to limit the variance of the residuals, the model results end up giving the average of the number of fire starts in which one particular day's conditions would result. Days with conditions unlikely to have fires will still have the odd fire, meaning all days with these conditions will have a prediction of just of over zero fire starts, even when the mostly likely outcome is zero fires. This is of minor concern, as there can't be a partial fire start on any given day and so the number can be rounded down to zero.

The opposite is true for extremely high fire days. Conditions conducive to a large number of fire starts do not always results in a large number of fire starts and in these conditions the models will under-predict days with a particularly large number of fires and over-predict days with the same

conditions and few fires. Of more concern to fire managers are the days with very high fire starts. In this study there are only ten days with ten or more fire starts (TD and SD)—the model greatly under predicts nine of these. This problem may have been exacerbated by the fact that 4/5th of these high fire days occur in the TD, which is only about 1/3 of the data. To address this problem, future research may want to separate the TD and SD based on fire starts rather than randomly.

There are a lot of zero fire days at all FFMC values, but the proportion of fire days to zero fire days increases as FFMC increases. This results in most of the predictions from the model being between 0 and 1. However, there are of course no fire days with fire starts between zero and 1 in reality. Therefore, the results of the model should be taken as a relative indication of the number of fires rather than an absolute indication of the number of fires. This means the predictions for particularly high fire days (relatively rare events), will often be low. The variability in fire starts on days with the same predictor values, combined with the large number of zero fire days, make predicting fire occurrence difficult.

The day-to-day predictions (using M6) show a PC of 0.504 using the TD (Figure 26). When a PC is done comparing the TD predicted versus actual fires per year the correlation is much better (0.783) (Figure 27). In general the over-predictions and under-predictions average out over the entire fire season. This may be useful when predicting the effects of climate change on fire occurrence and determining resources that should be hired before a fire season starts, provided accurate weather predictions can be made in advance. However, the best model for predictions of daily fire occurrence was not the same as the model with the best PC over the course of a year. If predicting the number of fires per year were the main point of the model, a different selection process would be more appropriate.

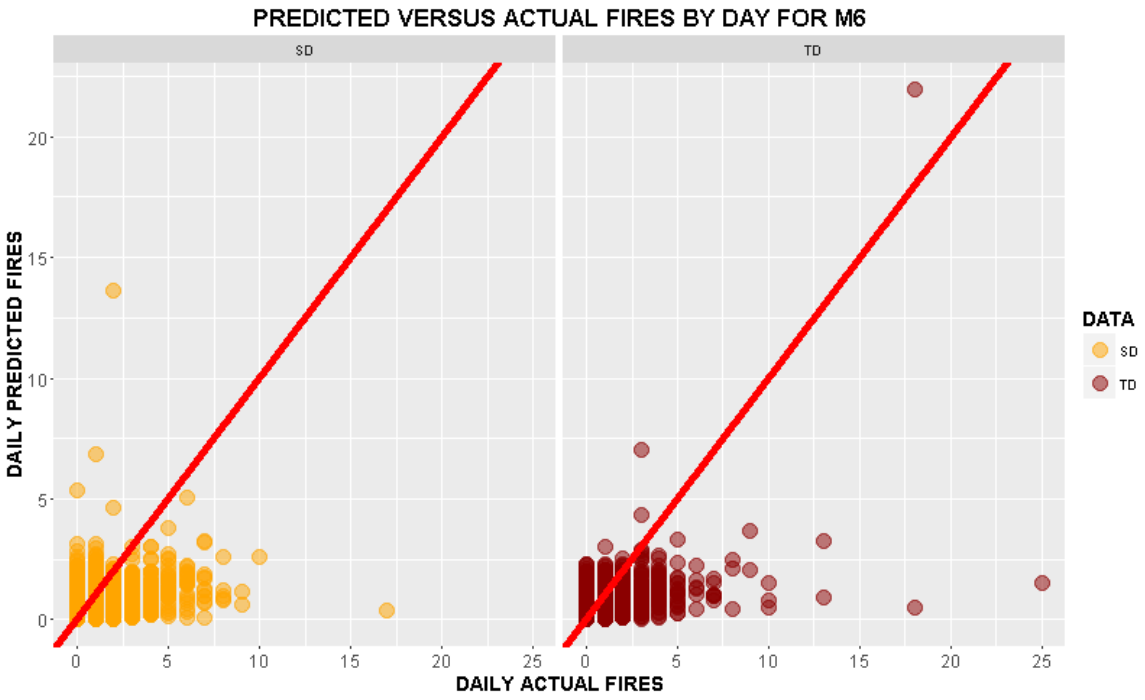

Figure 26: Predicted versus actual fire starts by day and ecoregion.

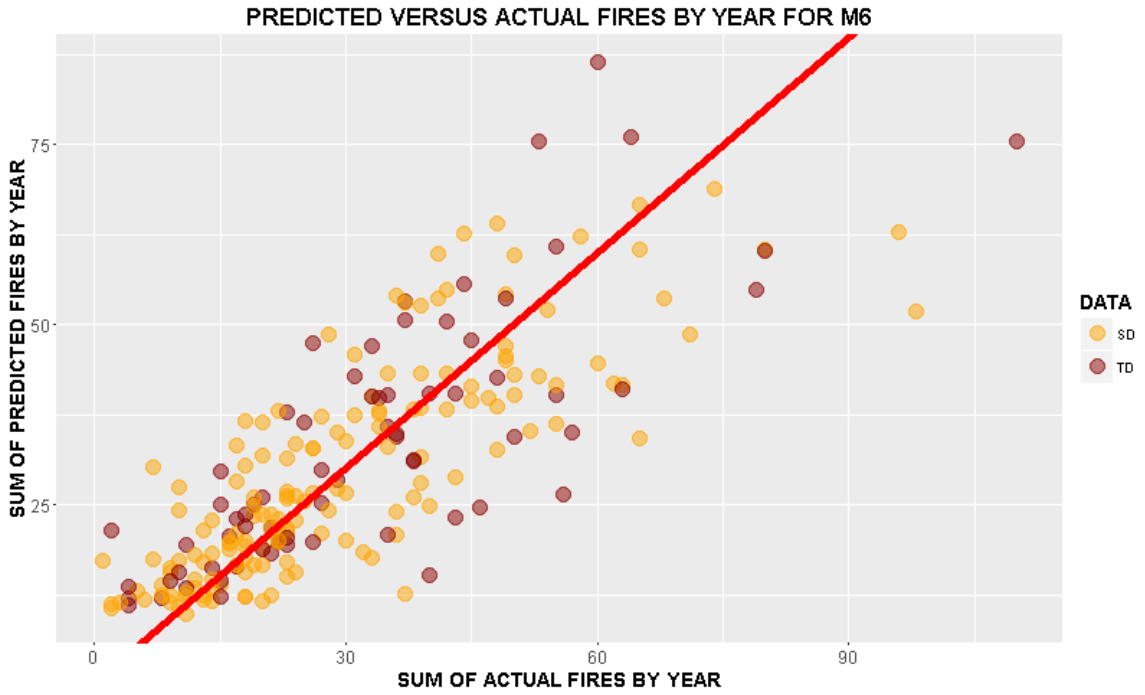

Figure 27: Sum of predicted versus actual fire starts per year and ecoregion.

Part 1: FWI System

As fuel moisture is a very important indication of the potential for a fire to start, various FWI System codes and indices were tested in the candidate models. As expected, FFMC predicted human-caused fire starts well, DMC and BUI contributed significantly, and DC had very little impact. However, unexpectedly, the FWI Index greatly improved the overall results of the models, more so than DMC and BUI.

For all the models, if a very high fire day falls on a very high FFMC day, the model generally performs well. However, if a high fire day occurs on a day without an extremely high FFMC then the model does not do a good job at predicting this extremely high fire day. This shows that although FFMC is a good predictor of the number of fire starts, there are other, unknown factors involved —factors not included in the model. Possibly this has to do with the spatial heterogeneity of FFMC values and the unpredictability of human actions. Additionally, this result highlights the difficulty of generalizing FFMC over a large area, as it can be quite heterogeneous at both the ecoregion and stand level.

The TukeyHSD test shows there is not a significant difference between M1 and M2, therefore there is not much difference in model predictive capacity when using DMC and DC together, or BUI alone. DC and DMC are correlated (0.447), therefore DMC may already provide some of the same information as DC.

DC is indicative of season long drying trends. This may make DC less useful than DMC for predicting human-caused fire starts, as these often happen in the spring before much deep soil drying has occurred. While a dry winter may cause the DC to be very high in the spring, and therefore could in theory contribute to an increase in spring human-caused fire starts, there is no indication of this

being the case from the data. This is likely because human-caused fires are less likely to start deep in the forest floor as fire brands usually land on the top of the forest floor.

Since the start-up DC is an estimate, and not physically measured, early season DCs tend to be less accurate than late season DCs. This is because incorrect start-up values will either be reset after a significant, multiday rainfall, or the fuels will eventually dry out to equilibrium moisture contentment, providing very similar DC values later in the fire season regardless of their starting value. It takes the DC 52 days to lose about 2/3 of the free moisture above equilibrium (Van Wagner, 1987), which is longer than the spring fire season in most cases. Therefore more exact measurements of the DC may yield different results, especially since most of the fires in this study occur in the spring when the DC is least accurate. In order to get more accurate DCs, field measurements would have to be taken directly after snowmelt in multiple locations throughout the province of Alberta. This may not be feasible.

Since the models all have some indication of fire season trends through the inclusion of seasonality variables, some of the information provided by the DC may have already been covered by the seasonality variable. These seasonality variables do not actually measure fuel moisture, but account for trends that happen from year to year. Since most years follow a similar pattern of overall climate, they also follow a similar pattern in DC values throughout the year (Figure 28). DC values are not prone to sudden changes from short term weather events. This is very different from FFMC which may show some seasonal trends, but can change greatly with just one rainfall.

Interestingly, the change in season from spring to summer (orange vertical line) corresponds to a change in slope of the DC, indicating a change in the weather at the change in the season, with the

exception of ecoregion 64, where the slope change happens much later. This may indicate the ecoregion 64 should have its spring season end much later.

In Figure 28, the really high values at the beginning of the spring are often a result of only 1 or 2 year's measurements and not an indication of how dry all years start as most years the weather station did not start recording that early. However it may be interesting to see if these early starting years, where the fuels weren't rewetted during winter, often lead to a high number of spring fire starts.

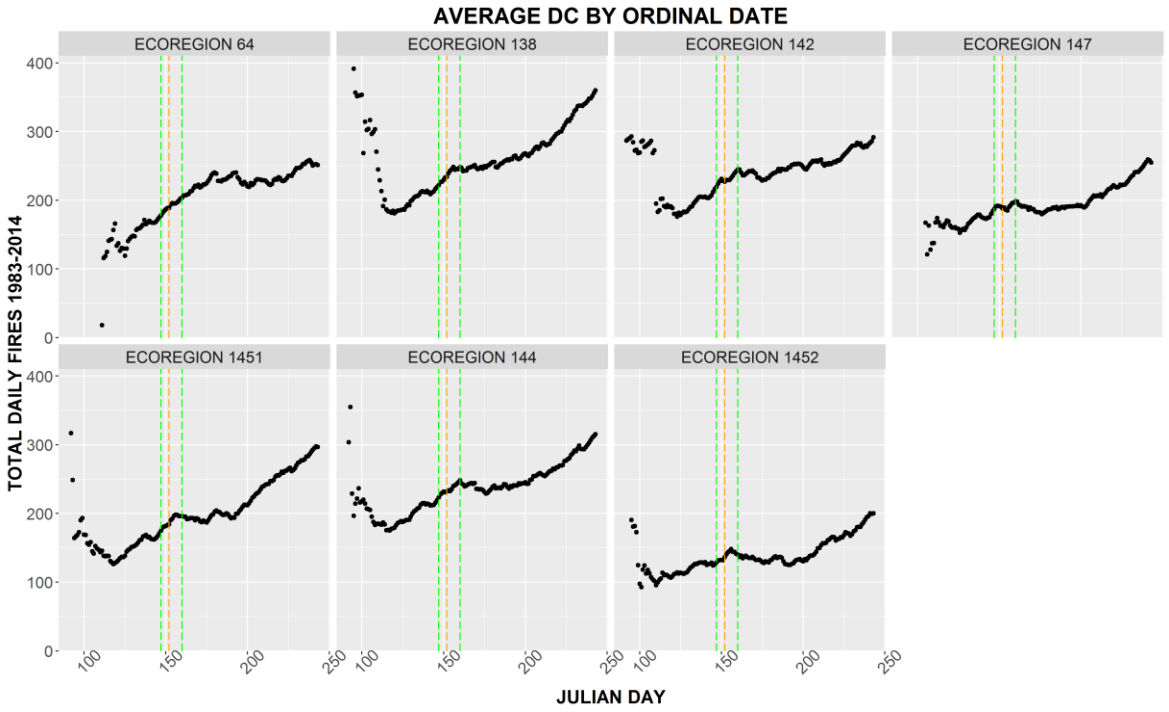

Figure 28: Average DC by ordinal date, separated by ecoregion. The vertical green dashed lines represent the changes of season for SEASON3, the orange for SEASON2.

DC values tend to become higher as the fire season progresses, whereas the number of human-caused fire starts tends to decrease. This may confuse the model as it may appear that higher DCs lead to fewer fires, when this is not the case as other confounding factors are involved. An unusually

low DC in the spring may mean something quite different than the same DC value in the summer. Perhaps in a model of the spring fire season only, DC would be more relevant. Finally, DC is an indicator of deep organic fuels and large fuels, which are unlikely to be what is initially ignited by a fire brand. The DC may be a better indicator of how likely these fires are to grow, or how hard these fires would be to extinguish once established, both of which are not part of this study.

As expected, FFMC does the best of the FWI codes and indices at predicting human-caused fire occurrence. It was not anticipated that FWI would predict human-caused fire starts better than DMC and DC together or BUI alone. FWI is a combination of BUI and ISI (and therefore a combination of FFMC, DMC, DC and wind speed) and is meant to represent the potential fire intensity.

It is still unclear why FWI (M3) fit better than DMC or BUI (M1 and M2) as an independent variable (as far as improving the AIC and test model predictions), but did not have the lowest deviance or sample data predictions. It is possible that by chance the test data had a couple of points that were predicted well by FWI, making the test stats better overall. The day with by far the highest FWI was in the TD. This is also the day where 18 fires occurred, which could contribute to this result (as M3 predicted this day much better than M1 and M2).

Unfortunately, as will be the case with all independent variables in this study, the zero fire days are influencing the fit far more than the days where fires occur. There are 24,728 zero fire days out of 29,241 days total (TD and SD combined). By trying to minimize the variance of the residuals, the maximum likelihood procedure is much more influenced by the large number of zero fire days than the small number of high fire days which are of higher concern to fire managers. That said, as discussed later in this thesis, using hurdle and zero-inflated models which specifically deal with this problem did not improve the model much, or at all. This was likely due to the fact that zero fires can

occur at any FWI value (or any other FWI System code and index) if there is no ignition source. This will skew the results when trying to fit fire starts to FWI System values whether basic GLM, zero-inflated or hurdle.

Additionally, a PC of 0.7 or greater was chosen as the cut-off of excessive correlation between independent variables in the same model (variables with a PC of 0.7 or higher were not included in the same model) and FWI was correlated with FFMC by 0.669, which is still rather high, which indicates the inclusion of FWI with FFMC requires further investigation.

It may be beneficial in the future to produce a model with two fit lines: one for an FWI index of 0 to roughly 30 (this value varies by ecoregion and season) and one for higher FWI index values. This is because a GLM model (with only FWI, ECOREGION and SEASON) fits very well up until this point, after which it starts greatly over predicting fire occurrence (Figure 8). This is likely because there are relatively few days with very high FWI Index values, so they don't influence the model fit much. However, when combined with FFMC (which under-predicts high fire days), the fit line improves greatly (Figure 29). Since the FFMC tends to under-predict human-caused fire starts on high fire occurrence days, and FWI tends to over-predict on high FWI index day, the two indices balance each other out. It is unclear if this result has a physical cause, or if it is just the mathematical result of adding the over-predictions indicated by the FWI Index on very dry days. Therefore it is unclear if M3 really predicts human-caused fire occurrence better than M1 or M2.

In Alberta, an FWI Index of 29.5 indicates the start of the extreme fire hazard rating class (Alberta Agriculture and Forestry, 2010). These classes are made by compiling FWI values over several fire seasons, deciding how many extreme fire days there should be each season, then composing a logarithmic scale so the number of fire days in each class increases at the same rate between classes.

This means that there will be very few days in the extreme category, and most days will be in the low hazard category. The FWI Index is a good indicator of fire intensity and of overall fire danger (Alberta Agriculture and Forestry, 2010; Van Wagner, 1987). In this study about 0.65% of days have an extreme FWI, resulting in about 3.57% of fire starts and 40.69% of the area burnt. It is worth noting that the Richardson fire, which at 577,646.8 ha almost equalled the entire area burnt for this study, was not included because it started in ecoregion 136. Therefore these high FWI Index days, while few in number, are an important indicator of fire danger, and future investigation into fitting a better model to these values is indicated.

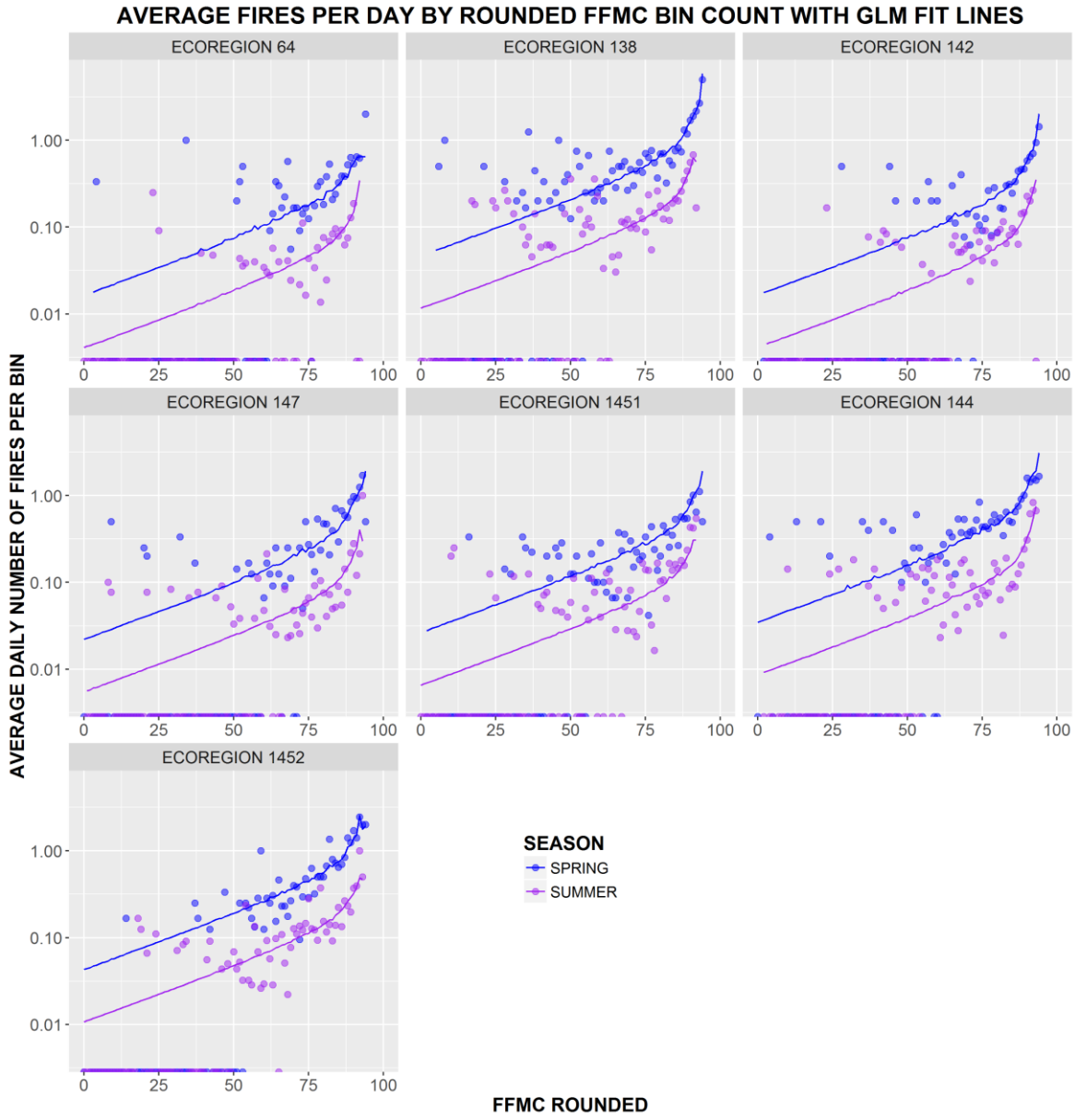

Figure 29: Average number of daily fires versus rounded FFMC bins separated by ecoregion and season. For example an FFMC bin of 50 would contain all FFMCs from 49.5 to 50.49. Fit lines are a negative binomial GLM model with FFMC, FWI, ECOREGION and SEASON2 as independent variables.

Additionally, the FWI System variables are a generalization of the fuel conditions of large areas, which may not be representative of the whole area or the areas where the fires are occurring. For example, in ecoregion 138 the weather stations chosen to represent the ecoregion were near the center of the ecoregion, however most of the fires occur in the fingers along the fringes of the

ecoregion (Figure 30). Since weather conditions can vary quite considerably over such a large distance, the FWI System values calculated from this weather data may not represent these fingers well. Furthermore, FFMCs can vary within the same stand. If the forest floor is under a canopy it can be shaded from the sun, but also sheltered from the rain, unlike forest floors under gaps in the canopy. Areas with depression, hills, different aspects or thicker/thinner understory vegetation can all make small scale FFMC variations within a stand and make fire prediction difficult.

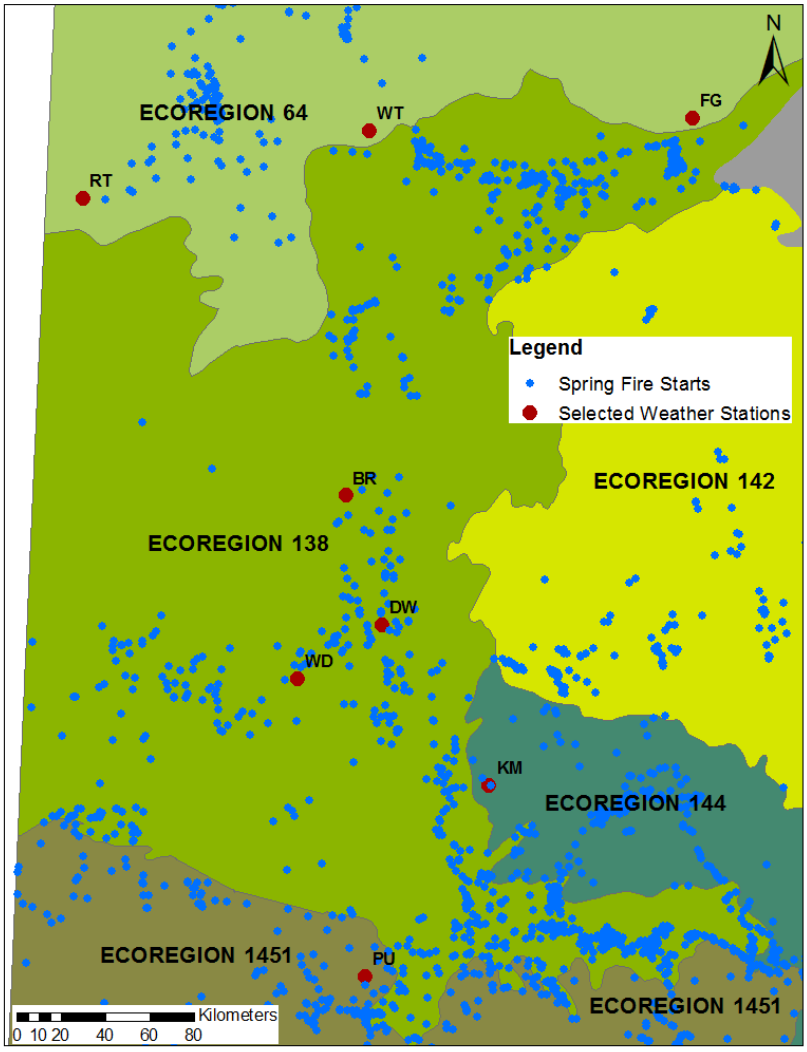

Figure 30: Map showing the distribution of spring human-caused fire starts in ecoregion 138 in relation to the selected weather stations. The summer human-caused fire starts have a similar spatial distribution.

There is no easy solution to this issue. Since there are two fingers, far apart, picking weather stations in these fingers (if even possible), may not improve predictive ability as the average conditions between the two fingers may represent neither. The other option would be to include these fingers in different ecoregions to make ecoregions without fingers. However, past research indicates vegetation type affects the meaning of FFMC values in regards to fuel moisture content, and therefore combining in this way may reduce the predictive benefit of separating by ecoregion.

If vegetation were deemed to not be very important, changing the ecoregions to Wildfire Management Areas (WMA; Figure 31) may be more practical for use. It would give fire managers an idea of how many fires would be in their particular area of interest, rather than an ecoregion. A WMA can contain several ecoregions and an ecoregion can overlap with several WMAs, so there is no way to determine which WMA the fire predicted in the current model would occur in, thus making resource allocation based on these predictions difficult.

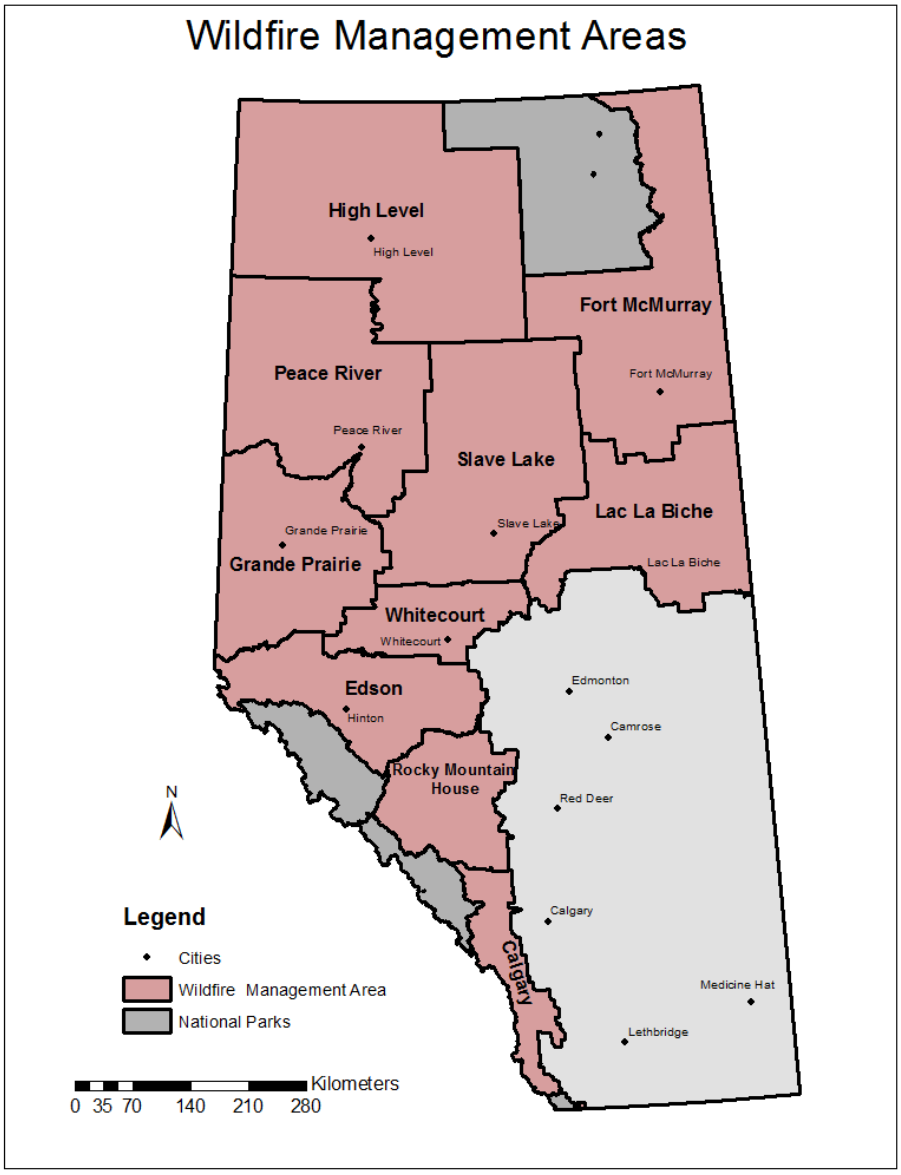

Figure 31: Wildfire Management Areas of Alberta.

Theoretically, FFMC should be closely related to the number of fires occurring, with FFMCs of 75 or less indicating the fuel is too wet for ignition. This holds true for the most part in this study (Figure 5). Interestingly, when comparing FFMC to number of fire starts, ecoregion 138 shows the most variation around the GLM fit line—this may be due to the factors just discussed. It is also interesting

to note that although there is very little difference between M6 and M7, M7 preforms slightly better on AIC, while M6 performs slightly better on mean RMSE TD and PC TD. This indicates that an interaction term between SEASON2/SEASON3 and FFMC makes very little difference to the model. This is counterintuitive in light of past research that found that FFMC values in different season represented different actual fuel moistures (Wotton & Beverly, 2007)

Part 2: Human Variables

The behaviour of people is particularly hard to predict, which may contribute to the result that no human activity variables were found to be sufficiently predictive to merit inclusion in the final model. DAY OF WEEK had potential for inclusion in a human-caused fire starts model. Since the improvement to the model by adding DAY OF THE WEEK was slight, the simpler model was chosen. However, as the original exploratory analysis indicated, in the spring some ecoregions had greater variability of fire starts by day of the week than other ecoregions, indicating future research, possibly separated by ecoregion or season, is warranted. Additionally, grouping day of the week into a binary variable by weekend and weekday may be useful. This grouping should be done for each ecoregion separately, as whether Monday and/or Friday should be grouped with weekend or weekday varies by ecoregion. While all the ecoregions had the highest fire occurrence rates on Saturday and Sunday, in some ecoregions Monday/and or Friday had fire occurrence rates closer to the weekend than the weekdays.

Although CAUSE GROUP was not included in this study, it should be considered for future research. Future research could consider grouping by how distinct/extreme the spring peak of fire occurrence is, with those cause groups with very high spring peaks grouped together, and those with less distinct or non-existent spring peak grouped together. Alternatively, recreation fires could make up one

cause group, and all the other human-caused fires the other cause group, as recreation fires consist of a large portion of human-caused fires, and are the only cause type with a large summer peak. When originally making these cause groups, Martell et al made separate models for each fire occurrence cause type, and used which variables predicted them best to group the causes into cause groups (1989). A similar procedure could be reproduced for Alberta.

The variable HOLILDAYS did not show enough significance to include in the candidate models. This is possibly because outside factors, such as the weather, confounded this variable. For example, people are far less likely to go out if the weather is cold and wet, and the forest is also not as likely to be flammable when the weather is uninviting. Therefore, on many holidays there are two factors reducing the incidence of fires, fewer people and wetter conditions, with a single cause, namely the weather. While the same reasoning applies to non-holidays, this behaviour pattern does mean that only holidays with nice weather would have the possibility of having more fires than on non-holidays. Since the number of days meeting these criteria is relatively small, it is likely any increase in number of fires on holidays would be lost in the aggregate. Additionally, it makes sense that people are more likely to go to places close to where they reside when they only have three days off (Wotton et al., 2003). Therefore, there is a possibility that HOLIDAYS would be more significant on sunny days in ecoregions farther to the south, where there are more people.

If a model were created for each ecoregion separately, HOLIDAYS should be reconsidered, as the visual inspection of the data showed a possible trend in ecoregion 1452. However, future research should also look into weekends versus weekdays, as the fact that most holidays are on weekends may be confounding things. Additionally, there may be holiday trends by cause type (e.g. there may be more recreation fires and fewer industry fires on holidays), which wasn't researched in this thesis.

If there is a difference by cause type, it could have confounded the results in this thesis and the industry fires may have balances out the holiday fires, indicating that there is no difference, when there actually is.

There are many more days without holidays versus days with holidays (2,069 versus 27,172 in this study's data set). This may influence the results, as some very rare, extremely high fire occurrence days are in the non-holiday category. Since there are over 10 times as many data points in the non-holiday data set, these rare events are far more likely to fall in this data set, regardless of the effect holidays have on fire starts.

Although fire bans were not included in the final model, the exploratory analysis did bring up some points worth discussion. There was a huge variety of FFMC's on fire ban days. While the majority of fire ban days had high FFMCs, the minimum FFMC on a fire ban day was 9.32. This is not an FFMC where fires should be starting and therefore a surprising day to have a ban; however, some portions of the ecoregion could be dry enough for fires to occur. It is hard to compare number of fires per day fire on ban days to number of fires per day on non-fire ban days with the same conditions, especially when viewing large regions with generalized FFMC values. This highlights the difficulty of comparing fire ban days to non-fire ban days for such large areas.

Future research could compare one Fire Control Zone with a fire ban, with the same Fire Control Zone with similar conditions (e.g. FWI System codes and indices, and dates) from previous years, to see if there is a drop in fires with a fire ban.

Perhaps there are other factors other than dry fuels that cause managers to initiate a fire ban, such as the predicted activity of people. It is possible that this increased activity would make for more fires on fire ban day than a non-ban day, so although fewer fires occurred than would have that day

without a ban, there were still more fires than a regular day. It is also unclear on how rigorously people adhere to fire bans.

An additional difficulty was encountered in that the ban often didn't cover the whole ecoregion. For purposes of analysing the potential utility of this variable, a minimum of approximately 25% of the ecoregion had to be covered by the fire ban for the region to be counted has having a fire ban. It was assumed that the areas in the ecoregion with the high likelihood of fire were the ones under the fire ban, and the other areas were unlikely to have fires starts. However, this is not necessarily the case, and was not verified in this analysis. Using WMAs to compare fire ban areas may be better as the administrative boundaries of the Fire Control Zones line up better with the WMA's than the ecoregion's.

Fires may have started outside the banned area in a particular ecoregion, thereby counting as a fire on a ban day, but not have been affected by the ban. In combination, the factors of messy spatial overlap, limited timeframe, and low sample size were deemed to place too great a limitation on the utility of fire ban data to warrant inclusion in this analysis. However, this is by no means a conclusive result, and fire ban data could warrant further study.

Finally, human activities (as well as the population of Alberta) change over time and the long time frame of the study could have obscured more recent trends in people's activities. Future research could be done on a much small time frame, such as from 1995 onward, which would still provide 20 years' worth of data.

Part 3: Seasonality

Adding SEASON3 instead of SEASON2 (M6 and M3 respectively) created very little difference in the model's predictive abilities. This is because most of the predictions for the summer and early spring were very similar. This makes sense because the majority of the data for these two seasons remains the same between the two models. Since the majority of fires start are in the spring (including the high fire days), there is not a large difference in the RMSEs. However, the transition season created a new, different combination of data which often predicted different values than the original two season variable (e.g. Figures 20 and 21). Therefore if only the predictions during the transition season were compared, these models would likely come out as different.

All of the seasonality models presented show the difficulty in modeling human-caused fire occurrence, due to the variability in fire starts by year, ecoregion, and day of the year. They also highlight the difficulty of modeling seasonality of fire occurrence, as every year has a different transition between the high daily number of fires in the spring and the low number of fires in the summer. Some years have a distinct peak in human-caused fire starts in the spring (Figure 16), other years have more uniform spring human-caused fire occurrence (Figure 15). Some years have an abrupt end to the spring fire season (Figure 20), some more gradual (Figure 24). None of the seasonal variables incorporate the timing of this change particularly well. This emphasizes the need to monitor plant phenology or other indicators of seasonal stage of the forest, and perhaps vary the cut-off between spring and summer, or the transition season by each year and ecoregion, in order to more accurately predict fire starts.

Adding the variables WEEK OF YEAR or ORDINAL DATE, although more detailed than SEASON3, did not account for more of the seasonal variability. WEEK OF YEAR and ORDINAL DATE added a large

number of factors to the model without adding much useful information as there is much yearly variation, especially by day but even in the week of year, resulting in overfitting of the model. This even further emphasized the need to find a way to enter the yearly variation in seasonality into the model.

In the future, it may also be interesting to see if using ordinal date in the spring, and week of the year or just one variable for all of the summer, works better, as most of the variation is in the spring, and this would reduce the factor levels.

All variables of time and seasonality are, in some measure, proxies for the physical and ecological realities on the ground. Although over time fire starts follow a similar average pattern year to year, therefore allowing for the use of variables such as SEASON2 and SEASON3 to approximate the seasonal phenological changes (and the resulting changes in fire likelihood), no variables in this thesis explicitly address the ecologically significant factor of phenology. It seems reasonable that this variable would have significant predictive value in fire models, particularly in the Alberta system, and previous modelling efforts would appear to support this assessment (Morin, 2015). Since different plants green-up at different times, and fire starts don't reduce with the first signs of green-up, additional thought would have to be put into which plants to observe and how long of a lag time would be required after the phenological observation before the end of the spring fire season. For example, does the spring fire season end five days, or twenty days, after aspen leaf out, or does this lag vary from year to year? Is aspen even the best plant to use? Province-wide phenological observation data was not available with the same time frame as this study. Another option may to use satellite imagery to record the state of green-up of the forest.

Part 4: Model Types

Since the majority of days in this study have zero fires, the models' fits are influenced greatly by these days. However, of more interest to fire managers are the high fire days. Taking this into account, it was expected that the zero-inflated and hurdle models would perform better that the ordinary GLM. While the hurdle and zero-inflated models had slightly better AICs, the predictive ability of these models was slightly worse. Overall there really wasn't a lot of difference between the models. It is unclear why this is the case.

It was also predicted that the negative binomial distribution would be the better fit for the data than the Poisson distribution. Although the AICs of the models using a Poisson link were much higher than the models with a negative binomial link, the predictive abilities of these models were not much different (Table 9). When looking at how well the negative binomial and Poisson models fit an FFMC, ECOREGION and SEASON2 only model, there was very little difference in the fit lines, with the larger differences being at the low FFMCs and therefore low fire occurrences.

Although PM6 had by far the worst AIC of any of the models, it is tied for the second best RMSE TD and PC TD of any of the models. It is unclear why the AIC is so high, but the predictive ability was comparable to the best model (M6).

Operational Use

To make this model useful for managers, several steps need to occur. Firstly, having a good hourly weather prediction for at least the following day is required for the model to function at all. In addition, the predictions would need to be grouped into classes, and a user friendly computer interface created to make getting and interpreting results easier for managers. Changing the model to give results by WMA would also be a useful step before attempting implementation, whether

through refitting the model with the WMAs instead of ecoregion, or by redistributing the predictions as explained below. A trial period would be necessary to test results before operational use. Finally, potential users would require training in order to gain acceptance and understanding of the model.

A Graphical User Interface (GUI) should be programmed to allow for easy use. This program would just require the user to enter the model variables and it would give them the desired predictions. These predictions could be grouped into classes to help managers interpret the results. For example, five classes could be made; very low, low, moderate, high and very high. While separate classes should be done for each ecoregion when implementing, when done for ecoregion 1452 it could look something like (Table 11).

Table 11: Example of how fire prediction classes could be delineated, using M6 on ecoregion 1452.

Class	Very Low	Low	Moderate	High	Very High
% of Days	80	10	5	3.5	1.5
Prediction Range	0-0.476	0.477-0.915	0.916-1.229	1.230-1.789	1.790+
Average Actual Fires Starts	0.12	0.65	1.12	1.64	2.32

Training should be done for all users of the model. This should be done in such a way so that the users understand the limitations of the model. Operational fire managers are often hesitant to use new research in their day-to-day decision making processes (Martell, 1982). Training and understanding may encourage these managers to use the model.

The biggest limitation in the model is the fact that the predictions are averages of fire occurrence rates for a given set of conditions. There are still a wide range of potential outcomes. For example in Table 11, the very high occurrence class has a potential to have anywhere from 0 to 25 fires, with an

average of 2.32 fires. The predictions are also limited by the accuracy of the weather forecast, and will not tell the users where in an ecoregion fires will occur.

Another option would be to follow the method of Todd and Kourtz (1991), where they predict a range of fire possibilities. They use a Poisson process to make this range. For example, if the predicted number of fires was 3, there is a 90% chance that the actual number of fires will be between 1 and 6. Both of these two classification are meant to be used as a tool for fire managers to help them better predict fire occurrence, not as a final prediction used to make operational decisions.

In the trial period, the managers would employ their regular prediction methods, and then make a prediction using their regular prediction methods in conjunction with the model. These two predictions would need to be analysed to test the usefulness of the model. A weighted scoring procedure similar to that in Todd and Kourtz (1991) could be used. They compared the fire occurrence predictions to actual number of fires, giving a large reward to high fire days that were predicted close to accurately and a large penalty to high fire days that were not predicted well. If predictions with the model prove better than predictions without the model, it would be recommended for operational use.

Making the FOP model useful for fire managers while basing the model on ecoregions is difficult, as in Alberta management decisions are made for WMAs (Figure 32). Basing the model on ecoregions allowed variations in vegetation throughout the province, and related differences in FFMC value meanings, to be included in the model. However, redoing this model with WMAs is possible, and should be considered due to the potential for operational utility— even if some of the model's

predictive ability may be lost. That said, modeling by WMA may actually improve predictive ability as fire management decisions vary by WMA, especially in regards to effort put into detection.

Alternatively, the ecoregions could be spatially overlapped with the WMAs, and the fires predicted by ecoregions could be evenly distributed to the WMAs by percent area. In other words, the percentage of the ecoregion in the WMA would determine the percentage of fires predicted that belong to the WMA. If multiple ecoregions are in one WMA, these predicted fire values would be added together. This could be done in a more complex manner with a weighted average, giving more weight to the ecoregion sections with the higher FFMC, or higher historical number of fires.

Finally, work should be done to improve the model, such as the suggestions mentioned in the Discussion and Future Research sections of this thesis. This can be done before implementation and/or done as the model is tested and as new research/technology arises. Additionally, feedback from the users could be incorporated to improve the model.

Future Research

While composing this thesis, many interesting questions and ideas came up that were outside the scope of this thesis and are recommended for future research.

Of the FWI System codes and indices, FFMC has been shown to be the best predictor of human-caused fire occurrence in Alberta. However, using FFMC alone causes poor predictions on high fire days, not because high FFMC values don't correspond with high fire days, but because the relationship between FFMC and human-caused fire starts appears to be non-linear. Quadratic or other non-linear functions should be considered in future research to accommodate this problem.

Although much work has already been done on relating FWI System codes and indices to fire occurrence, weather variables alone, such as relative humidity and temperature, have also been shown to be good predictors of fire occurrence (Xiao et al., 2015), and could be considered for future FOP models for Alberta.

Another problem in fitting these models was the large number of data points. This large number of data points (and therefore a very large n value), allowed for many variables that barely changed the model to come out with significant p values (less than or equal to 0.05). There are several ways to reduce this n value, but still use all of the data in creating the model(s). One would be to bootstrap the data used to fit the models, the other would be to create separate models for each ecoregion and/or season. These methods could also be combined. The first approach would use a similar method to that used to test the models' predictive ability in this study. In order for the models to work, each set of randomly selected data points would need to have sufficient data from each season and ecoregion, something that was not important in testing the data.

This method would also allow more models to be fitted on the data, theoretically allowing for a better fit. However, it would make using variables such as WEEK OF YEAR and ORDINAL DATE difficult as it would be hard to ensure each had enough data (e.g.in order for a model to make a prediction for a particular ordinal day, the fitted model would need information for that day. If a fire season started particularly early in a year that was being predicted, not all the days may have been included in the model fitting, therefore no predictions for those days would be possible). However, as WEEK of YEAR and ORDINAL DATE were not deemed the best method of delineating season in these models, the point is likely moot.

The second option, to make a model for each ecoregion and/or season separately, would allow the model to be fit to the specific properties of ecoregion and/or season. In some cases this could improve the model, as relationships between independent variables (FWI codes and indices) and the number of fires starts varies with ecoregion and season, and separate models may be better able to accommodate this. Additionally, running a spring only model would greatly decrease the proportion of zero fire days, which may help the model fit.

Future research could also look into high fire days only, as these are of the most interest to fire managers. Another option would be to look for temporal autocorrelation within the data. It is unclear if the data is temporally auto correlated, but if it is, taking that into account could improve the models.

Conclusions

Wildfires ignited by people account for the majority of fires starts in Alberta in the spring, and a significant number of wildfire start in the summer. These wildfires tend to occur near populated areas and infrastructure as this is where people tend to be (Vega-Garcia et al., 1993a). Human-caused fires, including the Flat Top Complex fires around the Town of Slave Lake in 2011 and the Fort McMurray fire in 2016, have resulted in much damage to infrastructure. Human-caused fires have also changed the seasonality of the fire regime in Alberta, as most lightning caused fires occur in the summer.

In order to fight wildfire effectively, fire managers must have the needed resources on hand to attack fires early, while they are still relatively easy to extinguish. Fire Occurrence Prediction (FOP) Models can be used as a tool to help fire managers make this judgement. This thesis explored various

models to see which predicted daily number of fires starts the best. A series of candidate models was created for May through August, and much of the forested areas of Alberta, using weather and fire occurrence data from 1983 through 2014. These models explored Fire Weather Index (FWI System) codes and indices (FFMC, DMC, DC, BUI and FWI), human variables (day of the week), seasonality (SEASON2, SEASON3) and model form (Zero-Inflated, Hurdle and Generalized Liner Models each with negative binomial and Poisson links).

It was found that a GLM model with a negative binomial link predicted daily fire occurrence the best while still being parsimonious. The model had the following form:

$$\ln(N_{HUM}) = \alpha_0 + \alpha_1 * \text{ECOREGION} + \alpha_2 * \text{FFMC} + \alpha_3 * \text{FFMC} * \text{ECOREGION} + \alpha_4 * \text{SEASON3} + \alpha_8 * \text{FWI}$$

SEASON2 is a binary variable representing spring and summer. SEASON3 is a three variable season variable representing spring, summer and a transition season between the two. This model had, for the daily predictions, a RMSE of 0.697 and a Pearson correlation of 0.504 when tested against a bootstrapped set of test data independent from that used to make the model. While these were the best (or tied for the best) statistics of all the candidate models, there was no significant difference in the predictive ability between this model and all but two of the other candidate models, indicating more research into some independent variables would be useful. This model could be used as is as an additional tool to help fire managers, or as a basis for future research.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions On, 19*(6), 716-723.

Alberta Agriculture and Forestry. (2010). Understanding fire weather. Retrieved from http://wildfire.alberta.ca/fire-weather/understanding-fire-weather.aspx

Alberta Agriculture and Forestry. (2014a). 10-year wildfire statistics. Retrieved from http://wildfire.alberta.ca/wildfire-maps/historical-wildfire-information/10-year-statistical-summary.aspx

Alberta Agriculture and Forestry. (2014b). Lightning detection. Retrieved from http://wildfire.alberta.ca/fire-weather/lightning-detection/default.aspx

Alberta Agriculture and Forestry. (2015). Historical wildfire database. Retrieved from http://wildfire.alberta.ca/wildfire-maps/historical-wildfire-information/historical-wildfire-database.aspx

Alberta Agriculture and Forestry. (2016a). Fire detection. Retrieved from http://wildfire.alberta.ca/wildfire-operations/wildfire-detection.aspx

Alberta Agriculture and Forestry. (2016b). Ministerial orders. Retrieved from http://wildfire.alberta.ca/wildfire-maps/historical-wildfire-information/ministerial-orders.aspx

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2010). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology, 65*(1), 23-35.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach* Springer Science & Business Media.

Burrows, W. R., & Kochtubajda, B. (2010). A decade of cloud-to-ground lightning in Canada: 1999–2008. Part 1: Flash density and occurrence. *Atmosphere-Ocean, 48*(3), 177-194.

Byram, G. M. (1959). Combustion of forest fuels. In K. P. Davis (Ed.), *Forest fire: Control and use* (1st ed., pp. 61-89) McGraw-Hill, New York.

Cameron, C. A., & Trivedi, P. K. (2013). *Regression analysis of count data* Cambridge university press.

Carruthers, E., Lewis, K., McCue, T., & Westley, P. (2008). Generalized linear models: Model selection, diagnostics, and overdispersion. *Memorial University of Newfoundland.*

Chrosciewicz, Z. (1986). Foliar moisture content variations in four coniferous tree species of central Alberta. *Canadian Journal of Forest Research, 16*(1), 157-162.

Clewer, A. (1999). Cambridge dictionary of statistics. *Journal of Applied Ecology, 36*(5), 842-842.

Countryman, C. M. (1972). *The fire environment concept.* ().U.S. Forest Service.

Cunningham, A. A., & Martell, D. L. (1976). The use of subjective probability assessments to predict forest fire occurrence. *Canadian Journal of Forest Research, 6*(3), 348-356.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., . . . Leitão, P. J. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27-46.

Ecological Stratification Working Group. (1996). *A national ecological framework for Canada* Centre for Land and Biological Resources Research.

Edmonds, R. L., Agee, J. K., & Gara, R. I. (2000). *Forest health and protection*. Long Grove, IL: Waveland Press, Inc.

Faraway, J. J. (2002). *Practical Regression and ANOVA using R.*

Fortin, M., & DeBlois, J. (2007). Modeling tree recruitment with zero-inflated models: The example of hardwood stands in southern Québec, Canada. *Forest Science, 53*(4), 529-539.

Government of Alberta. (2015). Climate and geography. Retrieved from http://www.albertacanada.com/opportunity/choosing/province-climate-geography.aspx

Hu, M., Pavlicova, M., & Nunes, E. V. (2011). Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial. *The American Journal of Drug and Alcohol Abuse, 37*(5), 367-375.

Kasischke, E. S., Christensen, N. L., & Stocks, B. J. (1995). Fire, global warming, and the carbon balance of boreal forests. *Ecological Applications, 5*(2), 437-451.

Kiil, A. D., & Grigel, J. E. (1969). The may 1968 forest conflagrations in central Alberta. *Forest Research Laboratory, Edmonton Alberta, Information Report.*

Larsen, C. P. S. (1997). Spatial and temporal variations in boreal forest fire frequency in northern Alberta. *Journal of Biogeography, 24*(5), 663-673.

Little, C. H. A. (1970). Seasonal changes in carbohydrate and moisture content in needles of balsam fir (*abies balsamea*). *Canadian Journal of Botany, 48*(11), 2021-2028.

MacNeil, M. A., Carlson, J. K., & Beerkircher, L. R. (2009). Shark depredation rates in pelagic longline fisheries: A case study from the northwest Atlantic. *ICES Journal of Marine Science: Journal Du Conseil.*

Maindonald, J., & Braun, J. W. (2010). In Ghahramani. Z., Gill, R., Kelly, F. P., Ripley, B. D., Silverman, B. W. and Stein, M. (Eds.), *Data analysis and graphics using R* (Third ed.). New York: Cambridge University Press.

Martell, D. L. (1982). A review of operational research studies in forest fire management. *Canadian Journal of Forest Research, 12*(2), 119-140.

Martell, D. L., Otukol, S., & Stocks, B. J. (1987). A logistic model for predicting daily people-caused forest fire occurrence in Ontario. *Canadian Journal of Forest Research, 17*(5), 394-401.

Martell, D. L., Bevilacqua, E., & Stocks, B. J. (1989). Modelling seasonal variation in daily people-caused forest fire occurrence. *Canadian Journal of Forest Research, 19*(12), 1555-1563.

Morin, A. (2015). *Alberta innovates project: Summary* (Tech. rep.).

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics, 33*(3), 341-365.

Murphy, P. J., Mudd, J. P., Stocks, B. J., Kasischke, E. S., Barry, D., Alexander, M. E., & French, N. H. F. (2000). Historical fire records in the North American boreal forest. In E. S. Kasischke, & B. J. Stocks (Eds.), (pp. 274-288). New York: Springer-Verlag.

Quince, A. F. (2009). *Performance measures for forest fire management organizations: Evaluating and enhancing initial attack operations in the province of Alberta's boreal natural region.* (Unpublished Master). University of Toronto.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria}: R Foundation for Statistical Computing.

Ridout, M., Demétrio, C. G. B., & Hinde, J. (1998). Models for count data with many zeros. Paper presented at the *Proceedings of the XIXth International Biometric Conference, 19* 179-192.

Rmetrics Core Team, Wuertz, D., Setz, T., Chalabi, Y., Maechler, M., & Byers, J. W. (2015). *timeDate: Rmetrics - chronological and calendar objects* (R package version 3012.100 ed.)

Stamp, R. M., & Warnell, D. (2016). Alberta. Retrieved from http://www.thecanadianencyclopedia.com/en/article/alberta/

Statistics Canada. (2011). Population, urban and rural, by province and territory. Retrieved from http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo62j-eng.htm

Stocks, B. J., Mason, J. A., Todd, J. B., Bosch, E. M., Wotton, B. M., Amiro, B. D., . . . Skinner, W. R. (2002). Large forest fires in Canada, 1959--1997. *Journal of Geophysical Research: Atmospheres (1984--2012), 107*(D1), FFR--5.

Todd, B., & Kourtz, P. H. (1991). *Predicting the daily occurrence of people-caused forest fires*

Turner, J. A., & Lawson, B. D. (1978). Weather in the Canadian forest fire danger rating system. A user guide to national standards and practices.

Van Wagner, C. E. (1967). Seasonal variation in moisture content of eastern Canadian tree foliage and the possible effect on crown fires.

Van Wagner, C. E. (1977). Conditions for the start and spread of crown fire. *Canadian Journal of Forest Research, 7*(1), 23-34.

Van Wagner, C. E. (1987). *Development and structure of the Canadian forest fire weather index system.*

Vandekerckhove, J., Matzke, D., & Wagenmakers, E. (2014). Model comparison and the principle of parsimony.

Vega-Garcia, C., Woodard, P. M., & Lee, B. S. (1993a). Geographic and temporal factors that seem to explain human-caused fire occurrence in Whitecourt forest, Alberta. Paper presented at the *GIS'93 Symposium.*

Vega-Garcia, C., Woodard, P. M., & Lee, B. S. (1993b). Mapping risk of wildfires from human sources of ignition with a GIS.

Wang, X., Cantin, A., Parisien, M., Wotton, M., Anderson, K., Moore, B., & Flannigan, M. (2015). *CFFDRS: Canadian forest fire danger rating system* (R package version 1.1/r13 ed.) Retrieved from http://R-Forge.R-project.org/projects/cffdrs/

Woolford, D. G., Bellhouse, D. R., Braun, W. J., Dean, C. B., Martell, D. L., & Sun, J. (2011). A spatio-temporal model for people-caused forest fire occurrence in the Romeo Malette Forest. *Journal of Environmental Statistics, 2*, 2-16.

Wotton, B. M., & Beverly, J. L. (2007). Stand-specific litter moisture content calibrations for the Canadian fine fuel moisture code. *International Journal of Wildland Fire, 16*(4), 463-472.

Wotton, B. M., Martell, D. L., & Logan, K. A. (2003). Climate change and people-caused forest fire occurrence in Ontario. *Climatic Change, 60*(3), 275-295.

Wotton, B. M., & Martell, D. L. (2005). A lightning fire occurrence model for Ontario. *Canadian Journal of Forest Research, 35*(6), 1389-1401.

Wotton, B. M., Nock, C. A., & Flannigan, M. D. (2010). Forest fire occurrence and climate change in Canada. *International Journal of Wildland Fire, 19*(3), 253-271.

Xiao, Y., Zhang, X., & Ji, P. (2015). Modeling forest fire occurrences using count-data mixed models in Giannan autonomous prefecture of Guizhou province in China. *PloS One, 10*(3).

Zar, J. H. (2010). In Lynch, D., Cumming, C., Lepre, C., Mendoza de Leon, B. and Behrens, L. (Eds.), *Biostatistical analysis* (5th Edition ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.

Appendix 1: Regression Coefficients, Standard Errors and Dispersion Parameters

Table 12: Regression coefficients and the dispersion parameter (theta) with the standard error in brackets. Formulas as presented in the methods section. Model coefficients were not included for M9.

Model	α_0	α_1	α_2	α_3	$\alpha_4/\alpha_{10}/\alpha_{12}/\alpha_{13}$	α_5/α_7	α_6	α_8	$\alpha_9/\alpha_{11}/\alpha_{14}$	Theta
M1	-6.57 (0.48)	Table 13	0.046 (0.0058)	Table 14	1.48 (0.041)	0.012 (0.0015)	0.00039 (0.00024)	-	-	1.118
M2	-6.52 (0.47)	Table 13	0.045 (0.0058)	Table 14	1.49	-0.01 (0.0011)	-	-	-	1.118
M3	-5.64 (0.44)	Table 13	0.033 (0.0054)	Table 14	1.38 (0.037)	-	-	0.049 (0.0029)	-	1.229
M4	-5.67 (0.44)	Table 13	0.033 (0.0055)	Table 14	1.39 (0.041)	-	0.00012 (0.00023)	0.049 (0.0031)	-	1.227
M5	-5.59 (0.44)	Table 13	0.033 (0.0054)	Table 14	1.38 (0.037)	-	-	0.049 (0.0029)	Table 15-	1.256
M6	-5.78 (0.44)	Table 13	0.034 (0.0055)	Table 14	Table 16	-	-	0.046 (0.0029)	-	1.381
M7	-5.46 (0.45)	Table 13	0.030 (0.0056)	Table 14	Table 16	-	-	0.046 (0.0029)	Table 16	1.391
M8	-4.56 (0.53)	Table 13	0.032 (0.0054)	Table 14	Table 16	-	-	0.45 (0.0030)	-	1.509
M10	-5.24 (0.45)	Table 13	0.028 (0.0055)	Table 14	0.49 (0.24)	-	-	0.048 (0.0030)	0.011 (0.0029)	1.234
PM6	-5.92 (0.43)	Table 13	0.036 (0.0053)	Table 14	Table 16	-	-	0.040 (0.0022)	-	-
Count HNM6	-3.16 (1.13)	Table 13	0.0058 (0.013)	Table 14	Table 16	-	-	0.033 (0.0059)		0.566
Zero HNM6	-5.81 (0.49)	Table 13	0.031 (0.0061)	Table 14	Table 16	-	-	0.063 (0.0041)	-	-

Model	α_0	α_1	α_2	α_3	$\alpha_4/\alpha_{10}/\alpha_{12}/\alpha_{13}$	α_5/α_7	α_6	α_8	$\alpha_9/\alpha_{11}/\alpha_{14}$	Theta
Count ZNM6	5.36 (0.58)	Table 13	0.036 (0.0070)	Table 14	Table 16	-	-	0.033 (0.0033)	-	1.629
Zero HNM6	-3.20 (2.73)	Table 13	0.058 (0.035)	Table 14	Table 16			-0.19 (0.038)	-	1.629
Count HPM6	-2.05 (0.87)	Table 13	0.0055 (0.01)	Table 14	Table 16	-	-	0.024 (0.0034)	-	-
Zero HPM6	-5.81 (0.49)	Table 13	0.031 (0.0061)	Table 14	Table 16	-	-	0.063 (0.0041)	-	-
Count ZPM6	-1.93 (0.88)	Table 13	0.0053 (0.010)	Table 14	Table 16	-	-	0.021 (0.0035)	-	-
Zero ZPM6	3.8 (1.00)	Table 13	-0.024 (0.012)	Table 14	Table 16	-	-	-0.082 (0.013)	-	-

Table 13: Regression coefficients for α_1 (ECOREGION) with ecoregion 64 as the base factor (coefficient of 0).

Model	64	138	142	144	147	1451	1452
M1	0	1.83	-0.72	0.85	-0.70	1.22	0.50
M2	0	1.83	-0.73	0.83	-0.72	1.18	0.0045
M3	0	1.93	-0.12	1.10	-0.24	1.16	0.46
M4	0	1.92	-0.13	1.10	-0.23	1.16	0.47
M5	0	1.92	-0.12	1.09	-0.25	1.16	0.44
M6	0	1.95	-0.13	1.11	-0.27	1.13	0.46
M7	0	1.98	-0.053	1.20	-0.19	1.25	0.57
M8	0	1.93	-0.15	1.09	-0.30	1.04	0.37
M10	0	1.95	-0.077	1.16	-0.19	1.23	0.54
PM6	0	1.87	-0.21	1.09	-0.43	1.16	0.45
Count HNM6	0	-1.91	-5.91	-1.37	-2.75	-1.38	-0.96
Zero HNM6	0	2.02	0.19	1.13	0.096	1.20	0.31
Count ZNM6	0	1.56	-0.56	0.49	-0.80	0.83	-0.13
Zero ZNM6	0	-1.89	-2.13	-6.90	-6.63	0.11	-7.41
Count HPM6	0	-2.45	-7.06	-1.77	-2.71	-1.35	-1.20
Zero HPM6	0	2.02	0.19	1.13	0.096	1.20	0.31
Count ZPM6	0	-1.91	-4.20	-2.38	-3.09	-3.22	-3.19
Zero ZPM7	0	-5.76	-6.80	-4.97	-4.06	-4.64	-5.67

Table 14: Regression coefficients for α_3 (SEASON2 and ECOREGION interaction) with ecoregion 64 as the base factor (coefficient of 0).

Model	64	138	142	144	147	1451	1452
M1	0	-0.012	0.0067	-0.0037	0.010	-0.012	0.0041
M2	0	-0.011	0.0068	-0.0035	0.011	-0.012	0.0047
M3	0	-0.013	-0.00099	-0.0065	0.0039	-0.011	0.0047
M4	0	-0.013	-0.00094	-0.0066	0.0039	-0.011	0.0047
M5	0	-0.012	-0.00097	-0.0064	0.0041	-0.011	0.0050
M6	0	-0.013	-0.00096	-0.0070	0.0041	-0.011	0.0043
M7	0	-0.013	-0.0020	-0.0082	0.0032	-0.012	0.0030
M8	0	-0.013	-0.00097	-0.0070	0.0043	-0.0096	0.0053
M10	0	-0.013	-0.0016	-0.0073	0.0033	-0.012	0.0038
PM6	0	-0.012	0.000071	-0.0068	0.0063	-0.011	0.0042
Count HNM6	0	0.026	0.060	0.018	0.032	0.014	0.017
Zero HNM6	0	-0.011	-0.0035	-0.0054	-0.000085	-0.0099	0.0081
Count ZNM6	0	-0.0096	0.0030	-0.00039	0.011	-0.0094	0.49
Zero ZNM6	0	0.013	0.016	0.079	0.086	-0.022	0.075
Count HPM6	0	0.032	0.074	0.022	0.031	0.014	0.019
Zero HPM6	0	-0.011	-0.0035	-0.0054)	-0.000085	-0.0099	0.0081
Count ZPM6	0	0.027	0.042	0.030	0.036	0.024	0.042
Zero ZPM6	0	0.054	0.071	0.050	0.047	0.045	0.056

Table 15: Regression coefficients for α_9 (DAY_OF_WEEK) with Friday as the base factor (coefficient of 0).

Model	Mon	Tues	Wed	Thurs	Fri	Sat	Sun
M5	-0.11	-0.12	-0.15	-0.11	0	0.044	0.11

Table 16: Regression coefficients for α_{10} and α_{11} (SEASON3 and SEASON3 FFMC interaction) with Friday as the base factor (coefficient of 0).

Model	α_{10}(Spring)	α_{10}(Transition)	α_{11}(Spring)	α_{11}(Transition)
M6	1.62	0.89	-	-
M7	1.30	0.66	-0.0048	0.012
PM6	1.62	0.91	-	-
Count HNM6	1.45	0.91	-	-
Zero HNM6	1.74	0.86		
Count ZNM6	1.24	0.81	-	-
Zero ZNM6	-18.37	-0.09	-	-
Count HPM6	1.29	0.83	-	-
Zero HPM6	1.74	0.86	-	-
Count ZPM6	1.15	0.88	-	-
Zero ZPM6	-1.17	-0.18	-	-

Table 17: Regression coefficients for α_{12} (WEEK_OF_YEAR) with week 14 as the base factor (coefficient of 0).

WEEK	**15**	**16**	**17**	**18**	**19**	**20**	**21**
M8	0.069	0.48	0.60	0.65	0.45	0.12	-0.11
WEEK	**22**	**23**	**24**	**25**	**26**	**27**	**28**
M8	-0.28	-0.74	-0.87	-1.16	-1.14	-1.28	-1.30
WEEK	**29**	**30**	**31**	**32**	**33**	**34**	**35**
M8	-1.25	-1.13	-1.17	-1.76	-1.23	-1.45	-1.48