Modeling Wildfire Perimeter Formation for Strategic Containment Line Planning In the Boreal Forest Region of Alberta, Canada

by

Siqi Mo

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

The edge or perimeter of a wildland fire is an important characteristic that marks the extent of land area that has burned. It is not only an ecological boundary, but also an administrative way of documenting the direct impact of a wildfire. Weather has a well-established influence on fire cessation but changes rapidly and is difficult to predict over extended time periods. In contrast, more stable landscape factors such as land cover, road networks, water coverage, and topography are well-suited for informing strategic assessments over longer horizons. Previous studies have shown that the variations in these attributes influence where fires stop. However, these studies are mostly based on one or a few fire perimeter datasets, and the overall influences over a large spatial scale and a long temporal span (i.e., over a decade) are still unclear.

In this thesis, the hypothesis is that key landscape factors play an important role in fire cessation over large spatial and temporal scales. Over one hundred fire perimeters were selected from thousands of documented fires between 2008 and 2018 in the Boreal Forest Region of Alberta, covering various natural ecological sub-regions. Five categories of explanatory variables were chosen from multiple spatial datasets based on provincial statistical maps and remote sensing, to represent landscape factors, including water, topography, fuel, previous burns and human activity. Algorithms to automatically match case-control data sampling (i.e., inside and outside the fire perimeter) and perform data extraction were developed, including a custom Python toolbox and ArcGIS model, resulting in vastly improved processing time. A combined modeling framework connecting matched case-control conditional logistic regression and Random Forest classification was used to identify the influence of key landscape variables on fire cessation and to predict the probability of fire cessation at a given location.

Twenty key landscape factors were identified as having a clear influence on fire cessation, with an Area Under the Curve (AUC) of 0.7. Fire boundaries form in areas of aspen, grass, water, and lower topography. Mixedwood, with conifer contents up to 60%, also played a very strong role in fire cessation ($P < 0.001$). In contrast, lands where previous fires recently occurred showed strong fire-stopping behaviour, but the effect decreased with the passage of time to the point where land burned 70 years ago behaved equivalent to a hazardous fuel. Human activities also affected the location of wildfire perimeters. The closer a fire is to a roadway, the more likely it will stop. Four new fires were used to validate the model, and the highest AUC of the predictive model was 0.7.

This modelling framework could be used by wildfire management agencies to inform strategic planning. The tool supports decision-making by calculating the fire boundary forming probability for any area within the Alberta boreal zone, based on temporally stable landscape factors. It could potentially guide the allocation of firefighting resources, identify high and low-risk areas for proactive measures, and assist in more efficient fire line planning.

Key Words: Alberta, fire cessation modeling, fire perimeter prediction, fire cessation, matched case-control conditional logistic model, random forest model

Acknowledgements

I want to express my eternal gratitude to many people and organizations who have supported me in completing this thesis.

First, I would like to give my utmost thanks to my supervisor, Prof. Jen Beverly, for her invaluable guidance, support, and encouragement throughout my thesis research process. Her consistent assistance refining the research plan and providing insights shaped this work. I had little prior knowledge of wildfires before I started the thesis. Jen guided me into this fascinating world of fire through the graduate-level fire course she taught. She provided me with the academic freedom to work alone but always unfailingly welcomed my enthusiasm when I was eager to share my thought process and engage in discussions. I benefited so much from all the Wildfire Analytics group weekly meetings arranged by Jen, from which I gained a deeper understanding of wildfire investigation.

I am also grateful to Prof. Douglas Woolford from the University of Western Ontario for his help and guidance in statistical methods and thesis writing skills and for his patience in the committee meetings. Whenever I faced challenges with statistical problems, Doug was always guiding me, often pointing me towards useful resources and literature.

Many thanks to Prof. Scott Nielson for being external examiner and the Chair of my oral defence. The pertinent questions you raised made me think about logistical regression deeper. I am moved that you and Doug and Jen, all are so kind to put lot of time on reading my thesis with the careful revision as detailed as you could.

I appreciate the members of the Faculty of Agricultural, Life, and Environmental Science, Department of Renewable Resources. Thanks to the Faculty of Graduate Studies and Research, University of Alberta, who financially supported me to present an oral presentation in attending the IUGG2023 Assembly in Berlin to international audiences.

Special thanks to Department Chair Prof. Nadir Erbilgin for his trust in me and for giving me a unique opportunity to teach during my third year. I would also like to thank Ms. Christie Nohos for helping me with administrative tasks around campus.

Thanks to my dear mother and father, your unwavering financial and emotional support and encouragement sustained me during the challenges of this academic journey. Your belief in me has been my greatest motivation. Thanks also to all other family members for your strong support and care.

I acknowledge the financial support provided by NSERC/Canada Wildfire Strategic Network; this study would not have been possible without this tremendous support.

Thank you to all those who wrote me reference letters when I applied to this study, including Prof. Dehua Wang from the Chinese Academy of Sciences, Prof. Ken Butler, and Prof. Joanna Zigouris from the University of Toronto (Scarborough). Special thanks to Prof. Bruno Wichmann for providing me with a research-assistant position during my third year, which enhanced my research ability and provided a supplementary financial source to support me in finishing the thesis.

Special thanks to my peers and colleagues in the Wildfire Analytics group; they are the best people to hang around for studying and working. To Air, Jared, Carter, and Andrew: Best of luck in the coming years, and for finishing your studies. To Nima, thank you for all the seminars you

have hosted to help me with satellite data and data presentation. When I started the thesis work, Kiera Macauley led me to the gate of MCC data sampling.

This thesis would not have been possible without the collective contributions of these individuals and organizations. Thank you all for being an integral part of this academic journey.

Table of Contents

Contents

List of Tables

List of Figures

1. Introduction

1.1.Preamble

Wildfire naturally occurs in forests around the globe. In Alberta, Canada, wildfire is a major natural hazard within boreal forests, which, as part of the North American Boreal Zone, is one of the biggest intact forests left on earth. While many wildfires in Alberta tend to be small and frequent, the relatively rare large fires account for the vast majority of the area burned (Stocks et al., 2002). When a fire surpasses suppression efforts and escapes, it may cause extensive damage to timber resources and other forest values. When wildfire is paired with extreme weather conditions, it can result in devastating damage to properties and can endanger the lives of residents.

There is a long history of human interaction with wildfires in Alberta. Before European settlement, Indigenous people in Alberta used fire as an effective tool for ecosystem management (Nekrich, 2022). However, when Canada became a leading exporter of wood to consumers across the continent, there was a strong driver for the government to establish policies and agencies to maintain fire control within Alberta's provincial land (Tymstra et al., 2020). As more people move into Alberta, the increase of urban development in wildfire-dependent landscapes requires more attention from the government and fire management agencies.

Suppression of wildfires can lead to accumulations of fuel and fuel continuity, thus promoting conditions for future wildfires (Arno & Brown, 1991; Tymstra et al., 2020). Increases in the time since the last fire are associated with increased likelihood that a new fire will escape initial suppression in some forest types (Beverly, 2017).

Suppression of wildfires is costly. Throughout Canada, provinces and territories are responsible for the fire management costs within each region, and combined nationally, suppression expenditures have been $800 million to $1.5 billion dollars annually for the past decade (Natural Resources Canada, 2022). The province of Alberta invests heavily in fire suppression. Due to the devastating fire season of 2023, Alberta introduced funding of $151 million over the next three years to enhance the wildfire management and preparation. (Horner, 2024).

Provincial and territorial fire management agencies in Canada have strived to balance the disturbance and damage caused by wildfire with the beneficial effect of fire in a forest ecosystem. For example, in the province of Alberta, with limited resources at the disposal of fire management agencies, the policy dictates that all fires are responded to and suppressed before they grow larger than 2 hectares, and in the event these efforts are unsuccessful, an approved wildfire management plan is initiated (Tymstra et al., 2020). Throughout these fire suppression efforts, fire managers must decide how to allocate suppression resources to each fire.

Information about landscape factors in relation to possible, current, and expected fire perimeters is critical to several fire management decisions. A fire's perimeter is defined as the edge boundary of a wildfire, recorded after fire has ceased burning (i.e., the final perimeter) as well as during dormant periods throughout the life of the fire (i.e., daily fire progressions). Natural landscape factors like roads, water bodies, and natural topography interact differently with fire perimeters. The fire edge is a limited piece of administrative information; however, when combined with additional landscape data, it has research applications, providing a way of documenting the nature of fire disturbance on the land. Understanding the factors that dictate fire perimeters can also help an agency plan whether a fire needs to be suppressed (Hardy, 2005; Beverly et al., 2021). Agencies can reduce costs and support the ecological benefits of fires by allowing fires to burn when they

do not pose a threat to human lives, communities, property, or other valuable resources. However, to implement this approach effectively, fire managers must assess potential fire behaviour and predict fire spread. Strategic analysis derived from understanding the factors that influence the fire perimeter, may potentially assist fire managers in understanding the risks associated with a decision to let fires burn. Specific tools have been developed to help support fire managers' decisions with landscape information. In Canada, potential and actual fire behaviour is assessed with the Canadian Forest Danger Rating System (CFFDRS) and its two subsystems: the Canadian Forest Fire Weather Index (FWI) System (Van Wagner, 1987; Taylor & Alexander, 2006) and the Canadian Forest Fire Behaviour Prediction (FBP) System (Forestry Canada Fire Danger Group, 1992).

Emerging studies have investigated the interaction between fire perimeter formation and the surrounding spatial environment in the temperate forest (Narayanaraj & Wimberly, 2011; Holsinger et al., 2016; Macauley et al., 2022). Some studies have focused more on individual fires in single regions, such as the Canadian Rocky Mountains (Macauley et al., 2022). However, such studies lack a comprehensive understanding of the relationship within a larger geographical region over an extended period. This thesis aims to fill in the gap by expanding fire perimeter analysis to the scale of an entire province and over a much longer period. By combining fire perimeter and land cover information in a geographic information system (GIS), ArcGIS, a model was built to identify the key factors influencing the formation of a fire perimeter and predict expected perimeters within the boreal zone in Alberta. The tool developed by this study can potentially inform fire management decisions in Alberta‘s Boreal Zone, assisting with predictions of whether a fire is likely to be naturally contained. Implementing a potential provincial-wide fire perimeter

prediction system may help fire managers improve cost-efficiency in fire management planning and potentially enhance preparedness for dealing with escaped fires.

1.2. Literature Review

1.2.1. Influential Factors on Wildfires

A fire is a chemical reaction that requires three things: fuel (carbohydrate), oxygen, and heat. Combustion produces carbon dioxide, water vapour, heat, and light (Figure 1-1a). If any one of the three factors is missing, combustion would stop, and the fire would cease to burn. On a grander scale for wildfires, the cessation process extends from this triangle of combustion requirements to a 'fire behaviour triangle' (Countryman, 1972), consisting of three factors that influence fire behaviour (Figure 1-1b): topography, fuel, and weather.

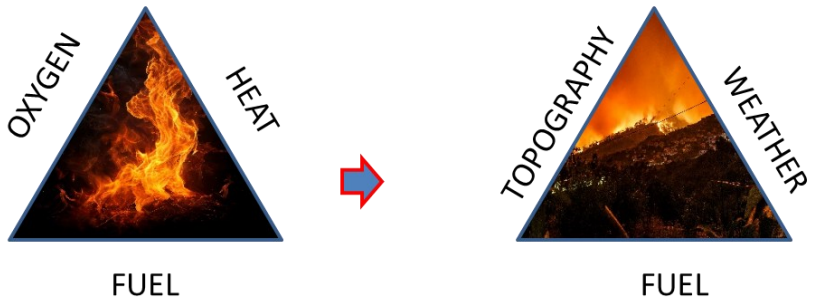


Figure 1-1 Illustration of the basic fire triangle for (Left) combustion, and (Right) a wildfire Event

Weather is directly related to wildfires, exhibiting a dynamic effect that varies from location to location and over short time-periods. Tymstra et al. (2021) found that extreme weather conditions is the main driver for the occurrence and spread of large spring wildfires in Alberta and hot, dry weather is known to make fire suppression in Alberta more difficult (i.e., Whitman et al. (2022). However, even though weather is a driving factor of fires, a recent study (Walker et al., 2020)

reported that fuel availability controls boreal wildfire severity and carbon emissions more than fire weather.

As a landscape factor, fuel influences fire as a bottom-up control (Parks et al., 2012; Fernandes et al., 2014). Like other landscape factors (e.g. topography, water bodies), fuel tends to remain unchanged for extended periods (i.e., 1-5 years). Changes in fuel availability have been found to affect the prediction of fire perimeter formation in various areas (Just et al., 2016; Rodrigues et al., 2020); however, these changes are limited over the short term.

Previously burned areas can provide a break in the continuity of the fuels. Areas burned by prior wildfires have exhibited a moderate to substantial influence on fire perimeter formation in the Western United States, and previous burns that happened long ago still impact fire dynamics, as demonstrated by Holsinger et al. (2016). Parks, Miller, et al. (2015) found that wildland fires exceeding 20 hectares have a regulating effect on the subsequent occurrence of fires in these regions. In the same study, they also found the regulating effect varies across different geographical locations. The effect lasts shorter (i.e., 9 years) in warm and dry area of the southwestern United States and longer (i.e., > 20 years) in the cool and wet areas of the northern Rocky Mountains (Parks, Miller, et al., 2015). Boreal spruce forests that burned 20-45 years prior were observed to have a protective effect such that new fires had a lower probability of escaping fire containment efforts (Beverly, 2017). Previous fire is rarely used as an input to fire spread pattern or perimeter prediction, and further research is needed.

Topography, which includes slope, aspect, elevation, configuration, and other related indices, plays a critical role in wildfire dynamics. Topography is a static and physical landscape feature, that has influence on vegetation (fuel), weather (wind direction and speed), and the way fire spreads. Ridges and slope increase wind speed as it pass through an up hill region, which allows

pre-heating of fuels and faster fire spread (Rothermel, 1983; Linn et al., 2007). Holsinger et al. (2016) reported that both valley bottoms and ridge tops showed moderate to high correlation with fire boundaries in the Northern Rockies and the Southwest of the United States. In the latter region, topography was a dominant factor, often serving as a fuel break. O'Connor et al. (2017) reported that with the exception of the most extreme fire weather conditions, topography and fuels are the most significant factors affecting potential fire spread and burn severity in the Northern Rocky Mountains of the United States. However, topography's ruggedness in different landscapes varies significantly across eco-regions. Topographic controls were most prominent in mountainous eco-regions and least influential in arid regions. Fire generally spread uphill, and ridge tops provided low-level control across all eco-regions of California (Povak et al., 2018).

Additionally, in the Mediterranean region, steep-relief mountainous areas show low levels of fire cessation likelihood, and plains show high fire perimeter presence (Rodrigues et al., 2020). Topography significantly impacts the cessation of wildfires, but regional differences and site-specific elements influences the relationship. The role of topography in fire cessation merits further study with additional research at multiple scales. Large water bodies, such as lakes, play a significant role in controlling wildfire spread. Nielsen et al. (2016) found that large lakes can effectively reduce the likelihood of wildfires in the boreal forest of Saskatchewan, Canada. Rodrigues et al. (2020) also found large rivers showing high fire perimeter presence in Catalonia, Spain.

Roads and trails are examples of human-made fire breaks that provide easy access for heavy equipment, support resource delivery, and are sometimes used to anchor the construction of a fire containment line. Narayanaraj and Wimberly (2011) are among the earliest to study the effects of roads on fires. They found that roads are most influential to fire boundary formation in lower-

elevation landscapes with high road network densities. Rodrigues et al. (2020) found that major roads show high fire perimeter presence. Povak et al. (2018) reported that roads were the dominant control across all ecoregions. However, they also found that removing roads from the analyses had no significant effect on the overall role of topography in wildfire extinguishment.

1.2.2. Sampling Methods

Various methods have been used to sample data in studies of fire edge formation. One approach is to use a vector layer with observed fire perimeters to establish the binary response variable where success is the presence of a fire perimeter (i.e., 1) and failure is the absence of a fire perimeter (i.e., 0). Observed perimeter locations represented successful control areas, whereas the rest of the burned area corresponds to the locations the fire spread through (i.e., failures). This method has been utilized in many studies, such as O'Connor et al. (2017) and Rodrigues et al. (2020). Just et al. (2016) also used a binary variable determined by whether an adjacent field had burned or not. .

Matched Case-Control (MCC) is a more complicated method inspired by prior applications in medical research (Breslow, 1996) and the ecology of animal movements (Compton et al., 2002; Whittington et al., 2005). Several studies (Narayanaraj & Wimberly, 2011; Holsinger et al., 2016; Macauley et al., 2022) have demonstrated its advantages in drawing meaningful conclusions between the fire cessation environment variables and the response variable (fire).

In MCC, sampling points are selected regularly along the fire perimeter. At each of these chosen points, a perpendicular transect intersects the fire perimeter, with two corresponding points selected at the end of the transect, inside and outside the fire's interior, to create pairs. Narayanaraj and Wimberly (2011) used this approach to sample points at 200-m intervals along the fire boundary and at 100-m intervals perpendicular to the fire boundary. Macauley et al. (2022) refined this method by distributing sample points along the transect segments in the burned and unburned

areas at 40 m, 100 m, 200 m, 300 m, 400 m, and 500 m. The 100-m interval perpendicular to the fire boundary was found to sufficiently capture the difference between burned and unburned states.

While the MCC method is expected to be highly efficient in capturing the key factors influencing the fire perimeter, previous research has shown that manual inspection was needed to remove points with mismatched burned/unburned status, which is labour-intensive. Holsinger et al. (2016) addressed this issue by choosing sampling points along the fire perimeter at 3-km intervals to reduce the data cleaning effort. Further studies are needed to evaluate the efficiency and labour costs involved in applying the MCC method over multiple fires in large regions.

1.2.3. Methods to Detect Key Factors in Perimeter Formation

In this thesis, the methods are divided into two categories: single factor and multiple factors. Grouping is determined by whether the influence of a given factor on fire perimeter formation is represented by one or multiple factors. For example, utilizing a single factor, Beverly et al. (2021) used stable physical fuel properties to develop a landscape metric of fire. By reclassifying conifer and mixedwood fuel types as hazardous fuels, any map cell that includes hazard fuels is treated as containing hazardous fuels. Parks, Miller, et al. (2015) investigated the factors influencing fire perimeter formation by assessing one representative factor: the time elapsed between the initial wildland fire and its subsequent ignition. This factor was used to explore further the role of previous fires on the fire cessation dynamics. Another example of the representative method was in Povak et al. (2018), which explored the interconnections between the topographic patch and fire size distributions. They implemented neutral landscape modelling to assess the significance of the topographic controls at fire boundaries across 16 ecoregions in California, USA. They used the topography feature as their single representative index. They assessed the differences in the topography between fire boundaries and fire interiors, which revealed distinct spatial control with

different fire sizes. Additionally, Nielsen et al. (2016) assessed a single factor: water(lakes) as a natural fire break influencing historical fire perimeters and found it had a strong bottom-up control on the local fire activities.

Instead of relying on one single factor, some studies comprehensively explore the influences of multiple factors on a fire perimeter, often through establishing a machine learning model. For example, based on prescribed fire experiments, Just et al. (2016) examined the influence of vegetation structure with microclimate condition on the extent of fire spread along savanna–wetland ecotonal gradients in Northern Carolina, USA, using linear mixed-effect models (GLMM). Rodrigues et al. (2020) effectively utilized a random forest model to predict fire cessation in Catalonia (northeastern Spain), achieving a high predictive accuracy with an Area Under the Curve (AUC) of 0.88. The model incorporated a variety of factors, including ground accessibility, fire breaks, and vegetation, among others. O'Connor et al. (2017) tested the spatial relationships between fire perimeter locations and physical landscape variables, potential fire behaviour, and access to suppression resources by extracting random samples over 238 US fires in the Northern Rocky Mountains.

Matched case-control (MCC) conditional logistic regression (Clogit) has been used widely to explore how different variables affect fire perimeters. In previous studies that applied MCC Clogit methods (Narayanaraj & Wimberly, 2011; Holsinger et al., 2016; Macauley et al., 2022), researchers modeled fire cessation using environment variables at each "case" sample point, which represent the hazard event, in this case fire cessation, that is outside the fire perimeter and assigned a value of 1. This case sample point is paired with a corresponding "control" sample point inside the fire perimeter, which is assigned a value of 0. The Clogit routine creates the dummy variable of times (all 1) and the strata.

In standard logistic regression, all the points are divided into two groups. However, in MCC Clogit, case and control points are paired based on stratifying variables (Hosmer, 2000). The pairing of the case points on fire boundary with control points inside the fire boundary is similar to a paired t-test (Whittington et al., 2005). Instead of modelling the overall distributions of cases versus controls, MCC quantifies the difference between each case and its matched control point. Clogit can identify the direction and the significance of variables influencing the fire perimeter. However, this also means that when utilizing an MCC Clogit model established from a training dataset for predicting outcomes using new data, the new test dataset must have the exact same strata as the training dataset. The mandatory data matching introduces a new challenge for any study that implements MCC Clogit for testing and prediction.

There are three types of methods to create test data for model evaluation (Xu & Goodacre, 2018). The first is *K*-fold cross-validation, which divides the data into *K* number of parts and selects one of the parts as the test data. The second is a deterministic split, which cuts the dataset into two determined parts by giving a test-to-train ratio. This is best for multiple runs of the same dataset (Xu & Goodacre, 2018). The third method is to split by randomly choosing a threshold ratio of data as the training dataset and using the remainder as test data. This method is best for avoiding potential biases (Xu & Goodacre, 2018). However, MCC Clogit could not use any of the three splitting methods, and the requirement for an exact match of strata between training and testing datasets became a limitation, making predicting new data using the MCC Clogit model problematic. The strength of the MCC Clogit (i.e., its exact matching) also turns out to be its shortcoming when processing predictions, especially for all new data points from a new data sample. Researchers who implemented the MCC Clogit method encountered this issue. Some, like Narayanaraj and Wimberly (2011), acknowledged this limitation and chose not to proceed with

prediction. Others, such as Macauley et al. (2022), explored alternative prediction methods that used the result of the MCC Clogit model but applied a different function for prediction. Addressing this issue remains an area in need of further research.

A promising alternative method for prediction is the Random Forest (RF) approach (Breiman, 2001). To establish a robust model to explore the influence of multiple predictor variables on response variables, namely, identifying fire-prone variables, an example of a decision tree model can be established, as described in Figure 1-2. The decision tree model is constructed using the entire dataset, using all the predictor variables.

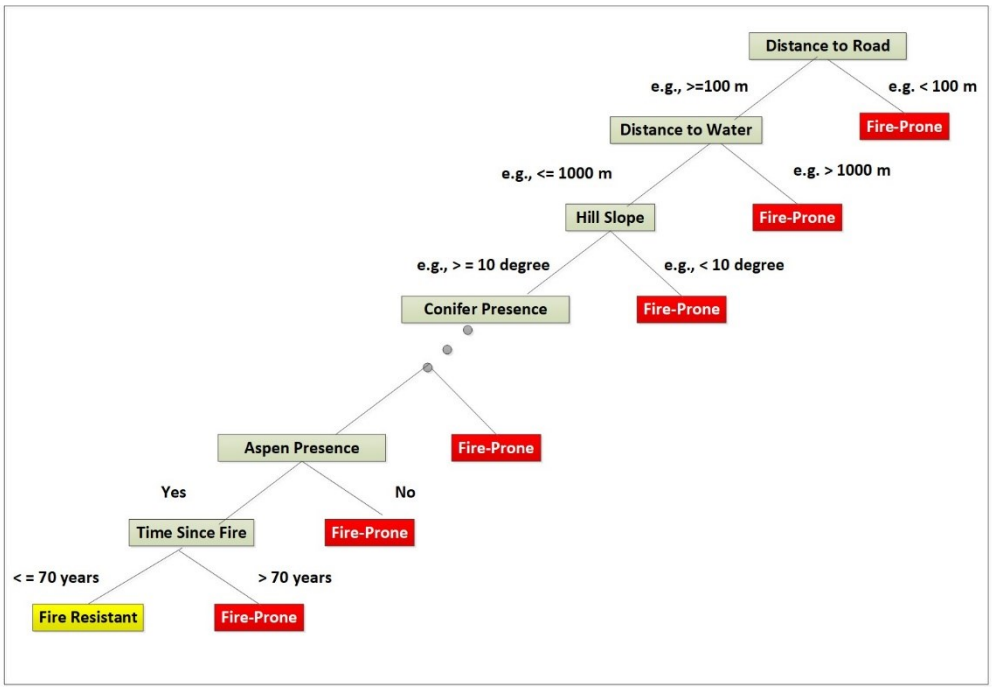


Figure 1-2 Example diagram of a decision tree model for the detection of a fire-resistant or fire-prone region.

The Random Forest model is an ensemble of decision trees that randomly selects a set of variables for each tree. Figure 1-3 illustrates an example of an RF model, where four decision trees are created, each taking two variables from the entire dataset. Each decision tree will predict the outcome based on the respective predictor variables used in that tree. The Random Forest model

aggregates the results by averaging predictions from all the decision trees. The advantage of the Random Forest model is that it increases predictive accuracy and robustness, making it a powerful tool for complex analyses involving multiple variables and potential interactions.

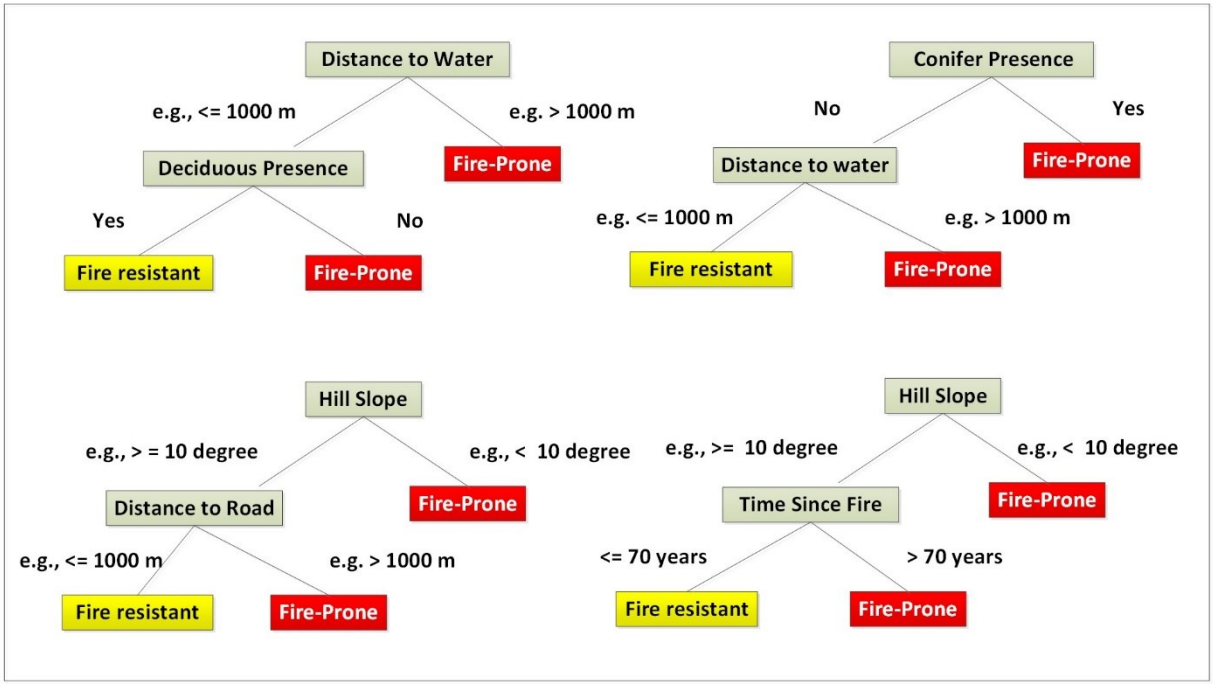


Figure 1-3 Example diagram of a Random Forest model consisting of four decision trees and two variables in each tree

Multiple decision trees must be created when implementing the Random Forest method. Each tree selects or votes the class (i.e., inside or outside the fire) based on the predictors' information at each specific site, and the predicted class is the one receiving the most votes by a simple majority.

Due to the random selection of variables in each decision tree, only random parts of the dataset are used. This randomness provides RF much more flexibility than a single decision tree. One particular advantage is avoiding overfitting issues in other algorithms. An overfitting issue happens when an algorithm reaches a very high model performance using training data but performs poorly predicting new data. In essence, this means that the disturbance can be well recorded and learned as concepts by the model in the training stage. Still, these concepts might not apply to the testing

data, which will negatively impact the model's ability to classify the new data, reducing the testing data's accuracy. The RF bagging technique can reduce the prediction variation by combining the result of multiple decision trees, where each tree was trained on a different random sampling of the dataset.

All the above characteristics demonstrate that RF has an advantage in handling high-dimensional (HD) issues. Couronné et al. (2018) showed that RF outperforms Clogit in large-scale data. Shomal Zadeh et al. (2020) proposed a high-dimensional matched case-control data method, introducing the Matched Forest (MF) algorithm. MF is based on the potential outcome model, which is flexible regarding the number of matching and exposure variables and can detect interaction effects. The method preserves each instance's case and control values but transforms the matched case-control data with added counterfactuals. A modified variable importance score from a supervised learner is used to detect important variables. Simulation studies show the effectiveness of MF in identifying important variables. MF modelling is also applied to data from the biomedical domain, and its performance is compared with that of alternative approaches.

Although RF and MF models excel in predicting outcomes with new data, they all struggle to detect the direction of the covariates (i.e., independent variables). In contrast, MCC Clogit excels in detecting the direction of the covariates but is limited for producing predictions. Combining the two algorithms could provide a possible solution to identify the key factors influencing fire cessation with prediction capability.

1.3. Research Opportunities

As summarized in Section 1.2, a relationship between landscape factors and fire cessation has been an emerging research topic, both scientifically and practically. Fire perimeter models for the boreal zone have yet to be developed. Prior studies have been limited to exploring specific factors. Studies

examining perimeters are more common; few studies have focused on the outside fire edge in the boreal zone, but often limited either spatially, by only focusing on single fires, or temporally, by only working with fires that occurred within a short time span. MCC techniques are effective in fire studies, but the data sampling method has largely remained manual and time consuming. Applying MCC for multiple fires over large-scale data is therefore challenging.

Fuel type is an explanatory variable used in fire perimeter modelling (Macauley et al., 2022). In statistical modelling, sometimes it is better to transform a categorical variable into a series of dummy variables (binary) to focus on each category individually. None of the documented studies to date have used dummy variables to explore the relationship between the fire perimeter and previously burned areas in the Alberta Boreal.

While the MCC Clogit regression model is effective for exploring the directional influence of various factors on fire cessation, it lacks capability in prediction. On the other hand, RF excels in its prediction capability but cannot identify the direction of influence. Therefore, a combined approach using both MCC Clogit and RF could provide more comprehensive insight about the relationship under investigation.

1.4.Research Objectives

The aim of this thesis is to identify key landscape factors that influence fire perimeter formation, based on numerous historical fires, to provide insight for decision-makers to enhance fire management strategies across the boreal forest zone in Alberta. Firstly, suitable fires and explanatory variables were selected. Secondly, an automated algorithm was established to efficiently sample and extract MCC data points for both fire and explanatory variables. Thirdly, a joint modeling framework was used to not only detect the influences but also make predictions of the fire perimeter.

A flowchart illustrating the overall research process is presented in Figure 1-4. Four working tasks (WT) were achieved as follows.

WT1: Fire and explanatory variables selection. Suitable fires and explanatory variables were selected. This involved choosing fires that align with the study area and have the least number of external conditions affecting fire cessation. Additionally, relevant explanatory variables related to fire environment landscape elements were identified. This was achieved by two sub-tasks:

WT1.1) Choose fires for better representativeness (rules and data quality control)

WT1.2) Choose relevant variables to represent fire environment landscape elements

WT2: MCC automation design and data cleaning. The workflow established automated algorithms for sampling and extracting MCC data points. Algorithms were developed to automate the MCC data sampling process and explanatory variable extraction process. In addition, data cleaning algorithms were developed to increase the efficiency and accuracy of the data collection process. Work Task 2 included three sub-tasks:

WT2.1) Develop algorithms to automate MCC data sampling

WT2.2) Develop algorithms to automate MCC data extraction

WT2.3) Develop algorithms to clean data

WT3: Identify the key variables by a joint modelling framework. In this task, MCC Clogit was used to identify key influential factors from the list of explanatory variables. This was achieved through a series of steps, starting with data preparation and pre-modeling processing, such as data cleaning and contingency testing. Various combinations of key influential factors called “schemes” were selected from MCC Clogit regression model runs. These schemes were refined using Random

Forest (RF) modelling. Additional schemes were determined based on importance scores from initial RF modelling. Subsequent RF modelling was conducted for each scheme resulting from the MCC Clogit and RF separately. The most effective scheme was selected based on performance measured by the RF model's Area Under the Curve (AUC). Three sub-tasks were included in Work Task 3:

WT3.1) Obtain key variable sets via MCC Clogit under Clogit schemes (cp0.1 stepwise)

WT3.2) Obtain key variable sets via MCC RF under RF schemes (importance score)

WT3.3) Obtain the best key variable set by rerunning the RF model based on the best AUC

WT4: Predict fire perimeter based on identified key variables. The key influential factors from the best-performing scheme were used to predict fire perimeter probability. This involved utilizing the trained model to predict brand new landscape environment data, and the fire perimeter maps predicted was verified with realistic fire maps. It included the following sub-tasks:

WT4.1) Sample the data for prediction

WT4.2) Predict fire perimeter based on identified key variables via RF

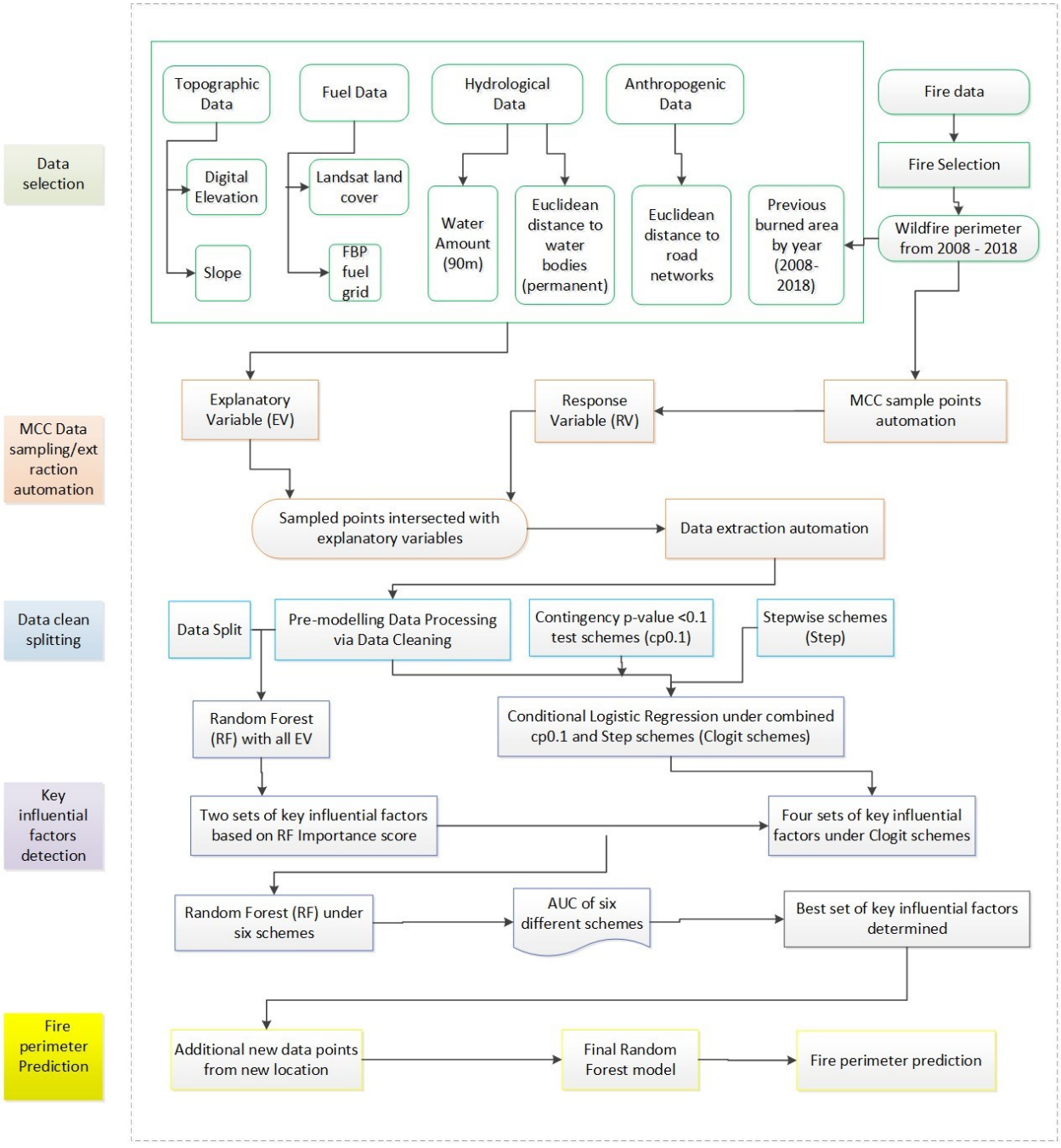


Figure 1-4 A flow chart diagram of the research carried out in this thesis.

2. Data Selection and Modelling Framework

2.1. Study Area

Canada has about 650×10^4 km^2 of forest ecosystems with a mosaic of trees, wetlands, and lakes (Wulder et al., 2008). Among them, the boreal forest occupies an area of 552×10^4 km^2 (with 270×10^4 km^2 of trees) (Brandt, 2009) across the country from east to west, forming one of the prominent features of Canada (Matasci et al., 2018).

Alberta is a western province in Canada located in the Northwestern Hemisphere. The total area of Alberta is 66.18×10^4 km^2 (Stamp, 2009; StatisticsCanada, 2021). The vast majority of Alberta is located within the Interior Plains region, featuring prairie grassland in the south, parkland in the center, and boreal forest in the north. The southwestern part of Alberta contains foothills that lead to the Rocky Mountains, and the northeastern corner transitions to the Canadian Shield. The prairie region of Alberta is relatively dry and consists mainly of grass, while the parkland region has a mixture of tall grasses and aspen trees. Within the Alberta boreal zone, the dominant tree species include white spruce (*Picea glauca*), black spruce (*Picea mariana*), jack pine (*Pinus banksiana*), balsam fir (*Abies balsamea*), paper birch (*Betula papyrifera*), and trembling aspen (*Populus tremuloides*) (Greene et al., 1999). The study area consists of the entire North American boreal range within the province of Alberta, which encompasses 48.33×10^4 km^2 in the north half of Alberta and 3 km^2 in the rocky mountain foothills (Figure 2-1).

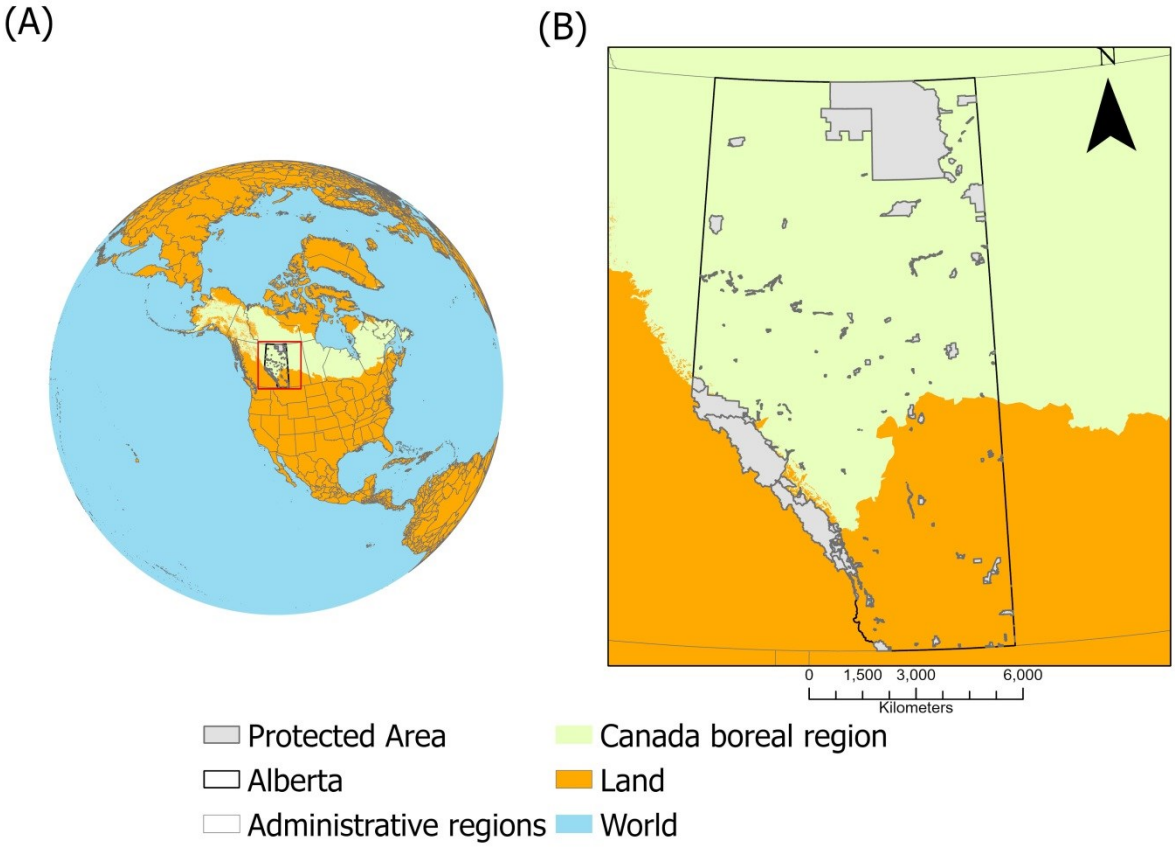


Figure 2-1 (A) The location of Alberta in the globe, (B) The map of Alberta within the North American boreal zone, with the protected areas coloured in grey. The green area within Alberta, minus the protected areas, is what the final study area looks like.

The Alberta boreal zone has five natural regions (Natural Regions Committee, 2006) (Figure 2-2): Rocky Mountain, Boreal Forest, Parkland, Canadian Shield and Foothills. Each natural region is further divided into sub-regions (Figure 2-2).

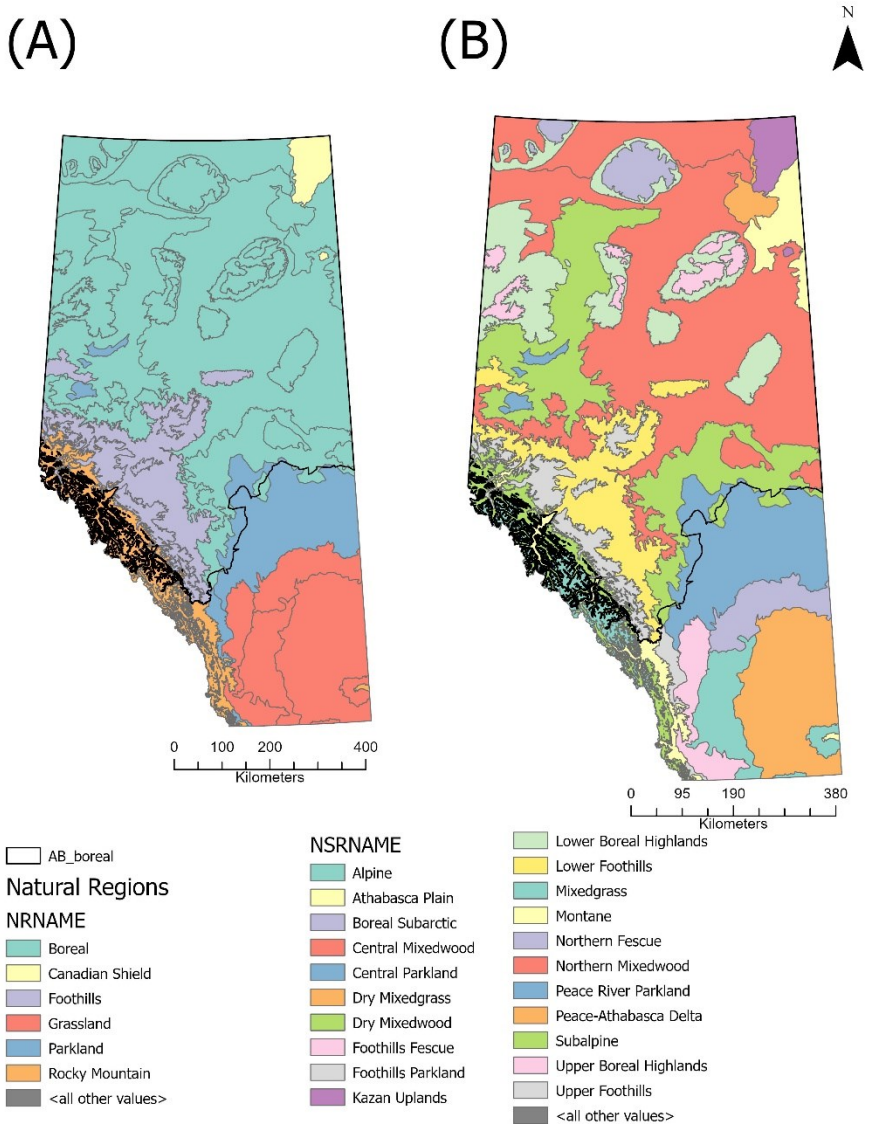


Figure 2-2 (A) The natural regions in Alberta (B) The natural sub-regions in Alberta, the different colours represent the different natural regions within Alberta, and the black outline represents the range of the North American Boreal Zone within Alberta. These data are from Natural Regions Committee (2006)

2.2. Selection of Fires

The fire (i.e., response variable) were selected from an historical fire polygon database (Alberta Agriculture and Forestry, 2016) delineating final perimeters of fires that occurred from 1931 to 2020. Each fire polygon record has multiple attributes including an identification number, the year of occurrence, fire name, burned area, timing of operational and administrative actions, and the method used to document the perimeter. As the data are imperfect, with the possibility of systemic or human errors, a set of six rules was used for selecting fires suitable for analysis. These are described as follows:

Rule 1. Time Span Matching: The year of the fire was set from 2008 to 2018, which represents the most recent decade of available data at the time of study initiation. The timing of fires also matched the temporal values of the explanatory data.

Rule 2. Spatial Restriction: All fires outside the Alberta boreal region or not entirely within the Alberta Boreal Region under provincial jurisdiction were removed using ArcGIS' "select by attribute" tool, including all fires that have majorly crossed the province border and fires within the federally protected parks. Those fires were removed with the geoprocessing tool “intersect.” Special consideration was made for stand-out fires that barely touched the border with a fire-by-fire case.

Rule 3. Only Lightning Caused: Only fires caused by lightning were selected. Fires that were not naturally started (i.e., caused by industry, railways, power lines, recreation, or other human-sources) could create bias due to different fire suppression efforts, so only lightning-caused fires were chosen.

Rule 4. Fire Class Filtering: Only fires with size class D and E were selected. Fire classes A, B, C, D, and E refer to fires with a fire area being (0 to 0.1], (0.1 to 4], (4 to 40], (40 to 200] and greater than 200 ha, respectively. It is impractical to mine the detailed spatial information of small fires. Fire perimeters with a fire class other than D or E were omitted.

Rule 5. Exclusion of unburned islands: The fire polygon database includes unburned islands and partially burned areas. Smaller fires with large partially burned areas were removed. Larger fires remained, but during the data cleaning process, a specific step was taken to remove all points sampled from areas considered "unburned islands."

Rule 6. Digitized Fires Only: The source of each fire polygon was classified as "0 - hand sketch of any type ", "1 - non-corrected ground GPS", "2 - non-corrected airborne GPS", "3 - corrected ground GPS", "4 - corrected airborne GPS", "5 - digitized from aerial photo", "12 - digital/satellite imagery (pixel 10 - 20m)" and "13 - digital/satellite imagery (pixel > 20m)". Only digitized fire data, i.e., with the data source being digitized from aerial photos, digital/satellite imagery (pixel 10m - 20m) and digital/satellite imagery (pixel > 20m) were analyzed, as the digitized fires often contained highly detailed polygons within the larger fire perimeter.

2.3. Selection of Explanatory Variables

In contrast to the conventional approach of categorizing the driving factors of fire perimeters into the fire triangle elements, as most literature did, the weather was excluded from the driving factors in this study, and only non-weather factors were investigated. While it is common knowledge that weather plays a significant role in fire behaviour, these factors are also challenging to quantify

across historical fires. Conversely, non-weather factors are relatively stable. The focus of this study is to examine which of those non-weather factors most affect the formation of fire perimeters.

Five types of variables were identified to explain fire perimeters using a literature review of factors affecting fire perimeter formation and fire-stopping and considering data availability on a provincial scale, in favour of data with less time-sensitive variables (i.e., remaining stable for at least a year). These variables were defined in relation to human activities, water, topography, fuel and previous fires.

2.3.1. Human Activities-related variables

Human footprint data (Alberta Biodiversity Monitoring Institute, 2010) provides information about human activities and artificial structures in Alberta and includes 21 spatial features that capture various human activities (e.g., rail lines, canals, roads, and railways… etc.). These data were originally in a polyline format in the shapefile layer.

The historical wildfire perimeters were overlayed with the human footprint. After manual inspection of all the possible human activities, the vast majority had no interactions with the wildfire perimeters. The only human feature that occasionally interacts with wildfire perimeters were roads. Thus, roads were chosen to represent human activity in the analysis. Like previous studies (Narayanaraj & Wimberly, 2011; O'Connor et al., 2017), a distance map to the nearest road using the "Euclidean Distance" Spatial Analysis tool in ArcGIS was created at a 30 m resolution.

2.3.2. Water-related variables

Water-related data were from the Boreal Surface Water Inventory, represented as polygons, and obtained from the ABMI (DeLancey et al., 2018). These data were in a shapefile comprising all of

Alberta's hydrography areas. The original dataset indicated the presence and location of water bodies, whether temporary or permanent.

Following Nielsen et al. (2016), the original polygon data for water was first converted into a binary format, then later transformed into two distinct variables:

(1) **Proportion of water:** The converted binary variable was used to calculate the amount of water in the surrounding area within a 90 m circular radius. The ArcGIS spatial analysis tool "Focal Statistics" was used to sum the occurrences of water within a circular moving window and output the total into a raster. The cell size was set as 30 m. Therefore, each cell in the output raster contains a number representing the total number of cells that contained water within a 90 m circular radius. A cell with a value of 0 indicates no water within the 90 m radius, while a value of 29 indicates it is fully surrounded by water (Figure 2-3).

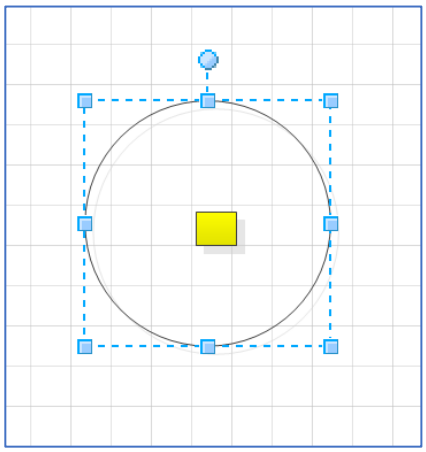


Figure 2-3 Example of the calculation of the water proportion around an example cell. Each square represents a 30×30 m cell pixel. The maximum of the spatial cells (30×30 m) outside of the yellow cell within a 90 m circular is $(\pi\ 90^2) \div 30^2 = 28.27$

(2) **Distance to nearest water bodies**. Using the spatial analysis "Euclidean Distance "tool in ArcGIS, the polygon layer is turned into a Euclidean distance layer. This layer depicts the closest distance in meters to the nearest water body, given a 30 m × 30 m resolution. The Euclidean

distance layer and the water proportion layer can better represent the effect of the water bodies in an area than simply the location or the status of whether the water body exists in a pixel. These two water variables are used to detect the influence of water, namely rivers, streams and lakes, on fire cessation.

2.3.3. Topography-related variables

Two variables are used to define the influence of topography on fire cessation, including a digital elevation model (DEM) and the slope. DEM data is in high resolution (10 m) at a datafile size of 40 gigabytes. These high-resolution data were clipped to cover the North American Boreal Zone in Alberta. Using the ArcGIS spatial analysis "slope" tool, the slope percentage was calculated with the formula (2-1)

$$\text{Slope percentage} = \text{rise/run} \times 100 \quad (2\text{-}1)$$

Where “rise” and “run” are the vertical rise and horizontal run of a slope angle, respectively. The 30 m elevation layer and the slope layer are created for representing topography.

2.3.4. Fuel-related Landscape Factors

To quantify the influence of land cover and fuel on fire cessation, the fuel type land cover classification of the FBP System were considered. The FBP System fuel data were in the form of individual raster data layers by year, with 100 m × 100 m resolution. In this case, for modelling consistency, the FBP fuel data were rescaled to 30 m x 30 m resolution using the ArcGIS spatial analysis tool "resize." This does not increase the resolution of the data; it only makes it convenient for the sampling process. Each cell contains a grid value that corresponds to an FBP System fuel type, including coniferous fuels: C1 – Spruce – Lichen Woodland, C2 – Boreal Spruce, C3 – Mature Jack or Lodgepole Pine, C4 – Immature Jack or Lodgepole Pine, deciduous fuels: D1/D2 – Aspen, mixedwood: M1/M2, 05PC to 95PC – Mixedwood with a percentage of coniferous fuel,

grass: O1 – Grass, non-fuels: NF – Non-Fuel, VNF – Vegetated Non Fuel, and W - Water. For modelling purposes, fuel variables were converted into dummy variables corresponding to each fuel type, each of which will be entered into the modelling. In the case of the mixedwood category, three additional dummy variables were created to represent three categories of conifer composition: 10-40%, 40-60% and 60- 90%. When sampling FBP System data, the year of the fuel variable is always set as the year prior to the fire.

2.3.5. Time-since-fire

To estimate the time-since-fire (TSF) of each fire, for each of the specific fire years, a sampled data point layer is created, and this layer is overlaid with the original fire polygon. This overlaid layer allows extraction of the year of past fire from the original polygon. The previously burned area was derived from the historical Spatial Wildfire database (Alberta Agriculture and Forestry, 2021; Government of Alberta, 2022b). For each fire year, additional layers in ArcGIS were created to analyze the effect of previous fires on the formation of the fire perimeter. A layer containing all fires from previous years dating back to 1931 was created for each year of fire perimeters.

There are two possibilities: NA (i.e., no fire data after 1931) or there is a prior fire present in a specific year. In the event of two overlapping past fires, the fire with the most recent year was used. The difference between that prior burn year and the year of the sampled fire from the geodatabase is TSF. TSF was converted into dummy variables representing decadal TSF categories (i.e., 0 to < 10 years, 10 to < 20 years, and continuing to the oldest TSF category of ≥ 70 years.

2.4. Summary of Data Sources

The data source associated with each variable is summarized in in Table 2-1.

Table 2-1 Variable data sources.

Description	Source	Spatial Extent/resolution	Time Frame/resolution	Format
Alberta Digital Elevation (DEM) data	AltaLIS (2021)	Alberta, 30×30 m (resized)	Not Applicable	Shapefile
Slope	Calculated from DEM data	North American Boreal Zone in Alberta, 30×30 m	North American Boreal Zone in Alberta	Raster data
Natural Regions and Subregions Polygon in Boreal Zone	Government of Alberta (2022a)	Alberta	Up to date	Shapefiles
Fire cause/Weather Info on the fire start date	Alberta Agriculture and Forestry (2016)	Alberta, fire by fire	2008-2018, fire by fire	Excel table
Historical Wildfire Polygons		Alberta	1931 - 2020	Shapefile
Fuel data		Alberta, 100×100 m, resampled into 30×30 m	2010-2020, yearly	Raster data
Wall-to-wall Human Footprint	Alberta Biodiversity Monitoring Institute (2010)	Alberta, 30×30 m	2010, 2014, 2015, 2016, 2017, 2018	Shapefile
Boreal Surface Water Inventory		Alberta, 30×30 m	2016, 2017	Shapefile
North American Boreal zone map	Natural Resources Canada (2009)	Canada	2010, five-yearly	Shapefile
Alberta Landsat Land Cover types	Government of Canada (2010)	Alberta	2010 & 2015	Shapefile

2.5.Process for Sampling Points

The key influential factors on fire cessation can be identified by finding the variation between factors inside and outside a fire perimeter. Following Narayanaraj and Wimberly (2011) and Macauley et al. (2022), a data sampling process was carefully designed to assign data as pairs.

In fire data, the fire perimeters are represented by polygons in ArcMap. However, actual wildfires never sharply stop at the edge of a polygon. Fire cessation is often influenced by gradual changes in the fire environment as much as abrupt changes, such as fuel breaks (Fernandes et al., 2016). Sampling directly at the edge of the digital fire perimeter may cause data pairs with no

variation if both points are within a band-shaped area created when the fire gradually stops. Therefore, a transect line is created at a 90-degree angle to the fire boundary at every sample location along the perimeter.

With reference to Narayanaraj and Wimberly (2011) and Macauley et al. (2022), the choice of sampling intervals along fire perimeters in this thesis is set at 200 m, with the transect lines reaching 100 m inside and 100 m outside of the fire perimeter. The 'control' point is inside the fire perimeter and is expected to have a 'burned' state. The 'case' point is outside of the fire perimeter and is expected to have an 'unburned' state. The paired points at each end of the transect line are considered a case-control pair of sample points. By sampling the points with the interval of 200 m along the perimeter line, enough variation in the information along the entire fire perimeter can be gathered while avoiding excessive sampling effort. At every 200 m on the fire perimeter, the difference in explanatory variables between the outside (unburn) and the inside (burn) of the perimeter can be explored (Figure 2-4).

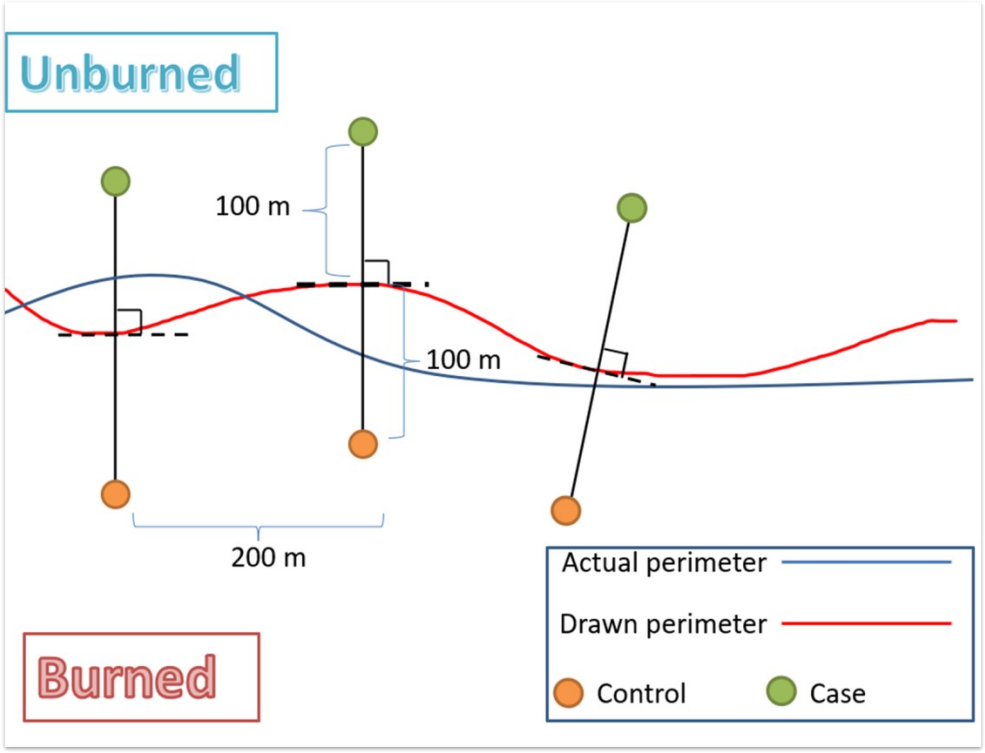


Figure 2-4 An illustration of the sampling methodology with 100 m transects perpendicular to the fire perimeter; the point on each end of the transect is a matched pair of case and control points

The detailed technical sampling process for a single fire polygon involves transforming the polygon into a line feature and then into a route feature under the ArcGIS spatial analysis tool. Through the ArcGIS calculation, the route feature file contains the total length of the route feature. By taking the value of the total length of the route, two different Excel tables called the "Event tables" are created, each with columns 'ROUTEID,' 'MEASURE_LOCATION,' and 'OFFSET.' The 'ROUTEID' column contains a universal 0, and the 'MEASURE_LOCATION' contains the sampling location along the perimeter, starting from 100 m, with an increment of 200 m to the total length of the route, and 'OFFSET' designates the sampled point being either control or case. Next, the ArcGIS Linear Referencing tool of "Make Route Event Layer" creates a feature layer of points as the sampled points. The Excel table created beforehand can control this process." ROUTEID" in the table can identify and match the route to the Event table,

"MEASURE_LOCATION" can pinpoint the exact locations along the route, and "OFFSET" creates points at exactly 100 m perpendicular to the original perimeter, the control point is inside the perimeter as it is represented by 100 (meaning sampled 100 m inward of the perimeter), and the case is represented by a -100 (meaning sampled 100 m outward of the perimeter) (Figure 2-5).

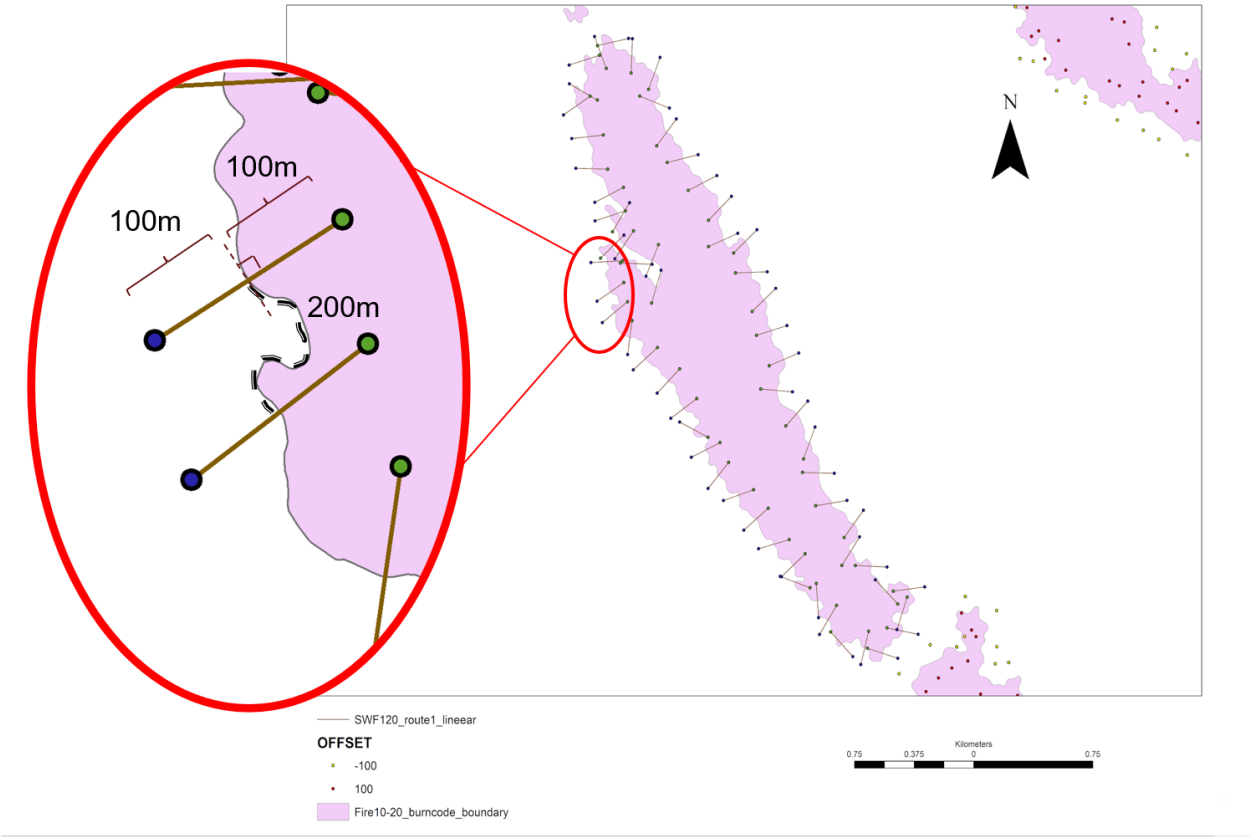


Figure 2-5 Examples of sampling points on a sample fire (SWF120-2010). Each pair of points is 200 m apart along the perimeter, and each point is 100 m away from the perimeter. The corresponding value of each explanatory variable was extracted for each pair.

This process generates two separate ArcGIS point layers for each pair of points, at which a list of explanatory variables can be extracted to describe the detailed fire environment at those point locations from all the different ArcMap spatial information layers of explanatory variables. This entire process was completed manually in previous studies (Macauley et al., 2022). If a parameter

needs to reset a value, all the work must be repeated individually using each spatial analysis tool. All the points must be visually inspected to ensure they are correctly positioned. When sampled points are along complex sections of the fire perimeter, a control point or a case point may not be in their respective location, and manual adjustments to the respective point positions would be required, which is laborious, especially in the case of processing multiple polygons.

2.6. Automating Data Sampling and Extraction for Multiple Fires

One of the main tasks of this study is to extend the sampling process from one or a few fires to more than one hundred fires, with vast spatial variability and temporal variability. It would be impractical and time-consuming to sample points manually. A new automated sampling technique is thus proposed, utilizing the ArcGIS Model Builder, custom Python script and Excel macro to streamline the process, reducing days of tedious, repetitive data processing into a simplified and fully automated process that takes just hours.

As detailed in the previous sections, the sampling process begins with initial filtering and selection of the fire perimeters. Since the fire polygons originate from different years, fires were first regrouped by year and stored separately in different geodatabases.

Individual models were built with the iterator tool in ArcGIS Model Builder. This tool loops through various elements, such as a location, a fire, a folder, or an ArcGIS polygon feature, and provides users with a flexible modeling process. In this case, the single polygon feature layer is the base layer of operation.

Before constructing the model in ArcGIS Model Builder, a Python-script-based tool to create event tables was developed and customized, as shown in Figure 2-6. This tool solves one of the Model Builder's key challenges for processing multiple fires, i.e., it automatically compiles two

individual event tables for every fire via the newly developed custom tool. This method eliminates manual work, making the process more efficient and practical.

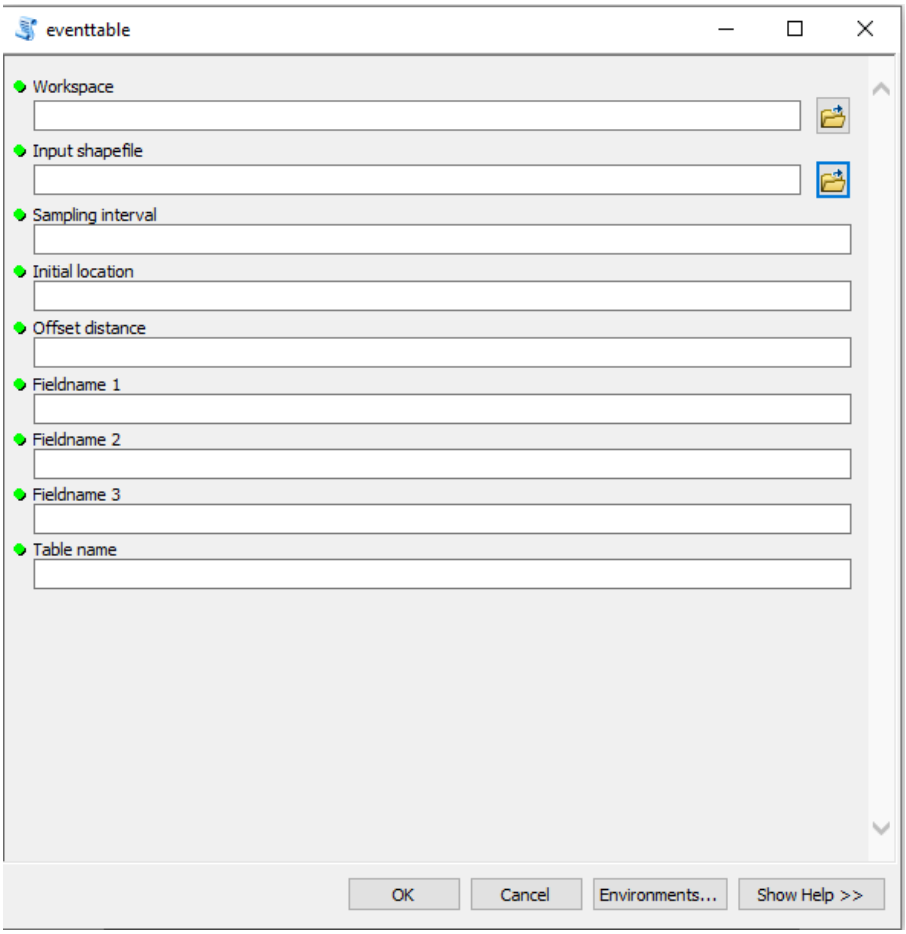


Figure 2-6 The interface of the custom-developed python-script tools for ArcGIS model builder. The tool takes in the 3 column headings of the route shape file, i.e., sampling interval, initial location, offset distance, as the input. The tool provides an output table with "routeid," and "measure_location."

The model builder starts with the creation of the route feature. It includes extracting each fire from the combined polygon feature layer, working with every single fire one at a time, converting it into a line feature, and converting each into a route feature, all stored into separate Geodatabases. The route feature contains an automated and calculated "Route length" attribute in its respective

attribute table. The details of these steps are shown in the first four rows of Figure 2-7.With the help of the custom tool, the model builder iterates through each fire, utilizing the individual fire perimeter's converted route length twice for each fire, and generating two tables, as shown in the fifth row of Figure 2-7.

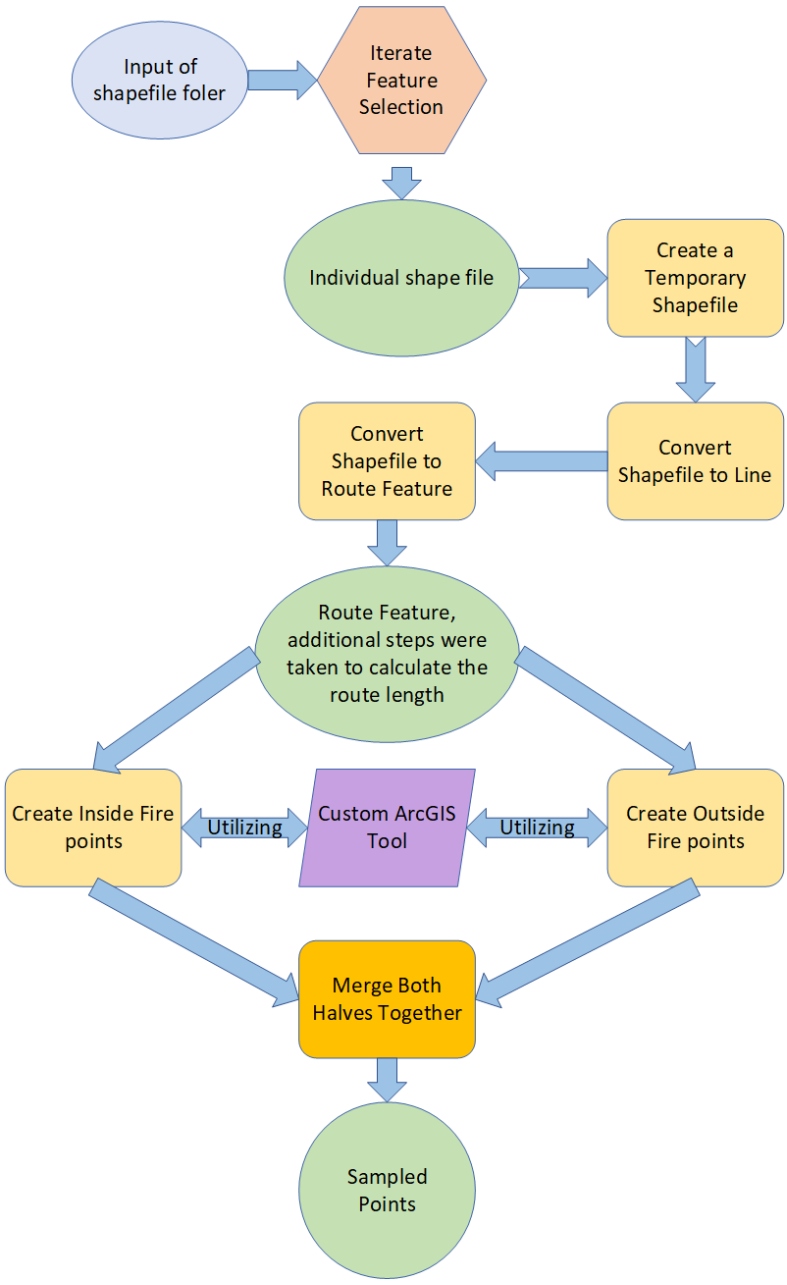


Figure 2-7 Data point sampling model for multiple fire polygons with automation

Storing all the generated tables in a Geodatabase, sample points can be generated using the ArcGIS Network Analysis tool "Make route event layer" twice for each fire. It uses the positive offset table (control) and the negative offset table (case) and creates two separate sets of points (case-control) that are uniquely stored. The two sets of corresponding points are then paired using the "merge" tool and output into a point shapefile by the "feature to point" tool, as shown in the sixth and seventh rows of Figure 2-7.

With all the sampled points from every single year stored in a separate Geodatabase, the last step is to extract associated spatial values and explanatory variables with minor changes such as changing the fuel data for different years, as shown in Figure 2-8. The sample points were intersected with all the processed explanatory variables: Euclidean distance to water, Euclidean distance to road, water proportion in a 90 m radius, slope (percentage rise), elevation, natural region, natural sub-region, fuel type, presence of previous fire within recent 70 years, and presence of previous fire in the entire fire history and the utility layers such as the cleaning layer (details described in the next section). Extraction was done using the ArcGIS Spatial Analysis tool "Extract Multi Values to Points," ensuring the extraction process was completed without bilinear interpolation. After the individual points had been stored with the required spatial information, they were exported to Excel tables through the ArcGIS conversion tool "Table to Excel."

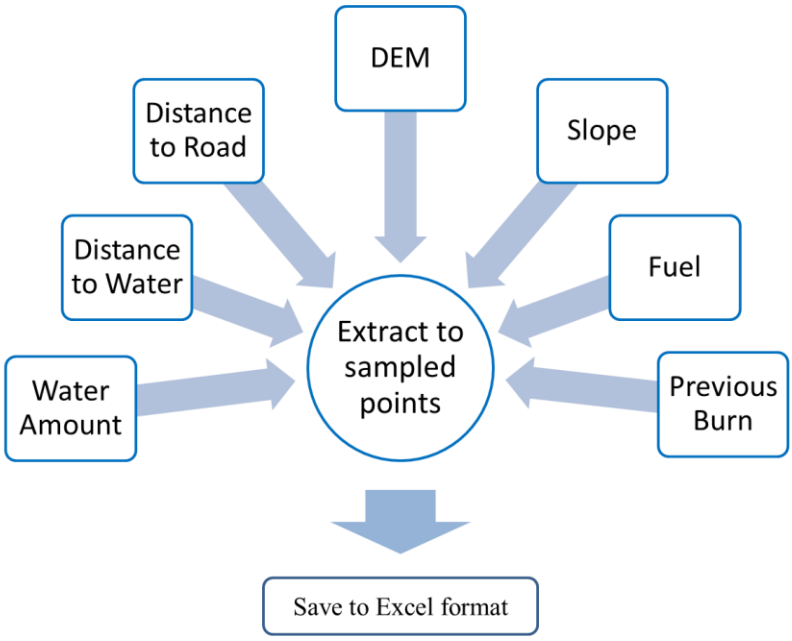


Figure 2-8 The illustration of automated extraction of all the explanatory variables to each sample point, and output to excel files

Before initiating modelling, an important issue had to be resolved. When ArcMap outputs tables from the ArcGIS model, the tables usually have the rows shifted in a random order. However, for modeling purposes, the format of the input file had to align the pairs of sampled points with each other (offset -100 pair with offset 100). While a simple re-sorting step in Excel could address this, the goal was to automate the process for hundreds of Excel files instead of using a manual command. A custom Microsoft Excel Macro was created to automatically loop through all the different Excel tables in the same folder and re-order them. This ensured that all the sampled points in each table were read into the regression model and paired without errors.

2.7. Algorithms for Data Cleaning

After the initial data sampling, a thorough cleaning process was applied to both the response variable and explanatory variables. After filtering the fires via the steps outlined in section 2.2, with further inspection, it was apparent that some fires encompass more than one distinct burned area, which necessitated a four-part cleaning process.

The first clean algorithm is an “inside-outside clean.” In the MCC data sampling, for a point along the fire perimeter, there is always a pair of data on the outside and inside of the fire. The points that are inside will be assigned a value for the flag variable OFFSET being 100, and those that are outside will be assigned a value for the flag variable OFFSET being -100. This is the normal case for most situations. However, in circumstances where the fire's perimeter is rigid and narrow, issues with the sample points might arise (Figure 2-9). The inside-outside cleaning algorithm retained normal points and removed abnormal points, as shown in Table 2-2. As the MCC sampling method is adopted in this thesis, all abnormal points were removed in pairs. Although it may reduce the sample size by excluding data, removal of matched pairs ensures that all the remaining data are in a strictly matched case-control status, which is necessary for Clogit modeling.

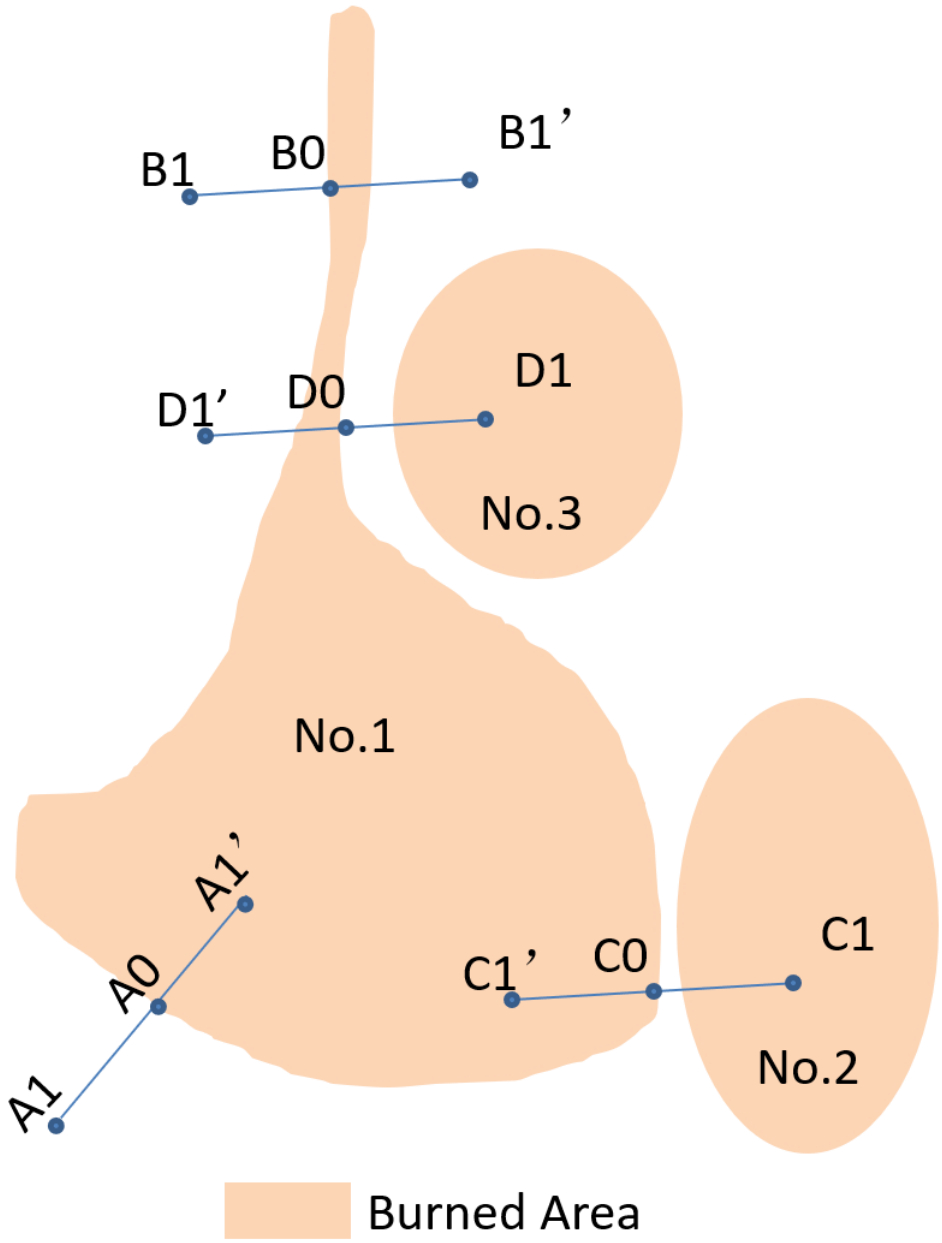


Figure 2-9 Illustration of the undesirable sample pairs that need cleaning. Brown areas represent the burned areas. In the text, 0 represents the sample point on the edge of the perimeter, and 1 and 1’ represent outside and inside fire points. Points A1 and A1’ show a normal pair. Point B1 is correctly at the outside of the fire, but its pair point B1’, is incorrectly outside the fire. Point C1, where its sample point is incorrectly inside fire perimeter polygon No.2. For the extreme case, point D1 is incorrectly inside another fire polygon No.3, and point D1’, which is supposed to be inside the perimeter, now is incorrectly outside the fire because of the irregular shape of the fire polygon No.1.

Table 2-2 Different conditions of sample data pairs, and corresponding action, and the symbols used are explained in Figure 2-9

Condition		outside fire	inside fire	Normal	Action
1	A1(OFFSET = -100)	Correct		Correct	Keep
	A1'(OFFSET = 100)		Correct	Correct	Keep
2	B1(OFFSET = -100)	Correct		Correct	Remove
	B1'(OFFSET = 100)		Incorrect	Incorrect	Remove
3	C1(OFFSET = -100)	Incorrect		Incorrect	Remove
	C1'(OFFSET = 100)		Correct	Correct	Remove
4	D1(OFFSET = -100)	Incorrect		Incorrect	Remove
	D1'(OFFSET = 100)		Incorrect	Incorrect	Remove

The second cleaning algorithm is called NVB-clean. The relatively coarse resolution of the provincial fuel grid introduced some obvious errors. For example, when the 100-m fuel grid was overlaid with a much finer detailed fire perimeter, it was possible for non-fuel areas to appear inside the fire boundary (Figure 2-10). An algorithm was created to clean those points with FBP-VNF (vegetated non-fuel) and FBP- NF (non-fuel) inside the fire to remove the bias created by this mismatched fineness.

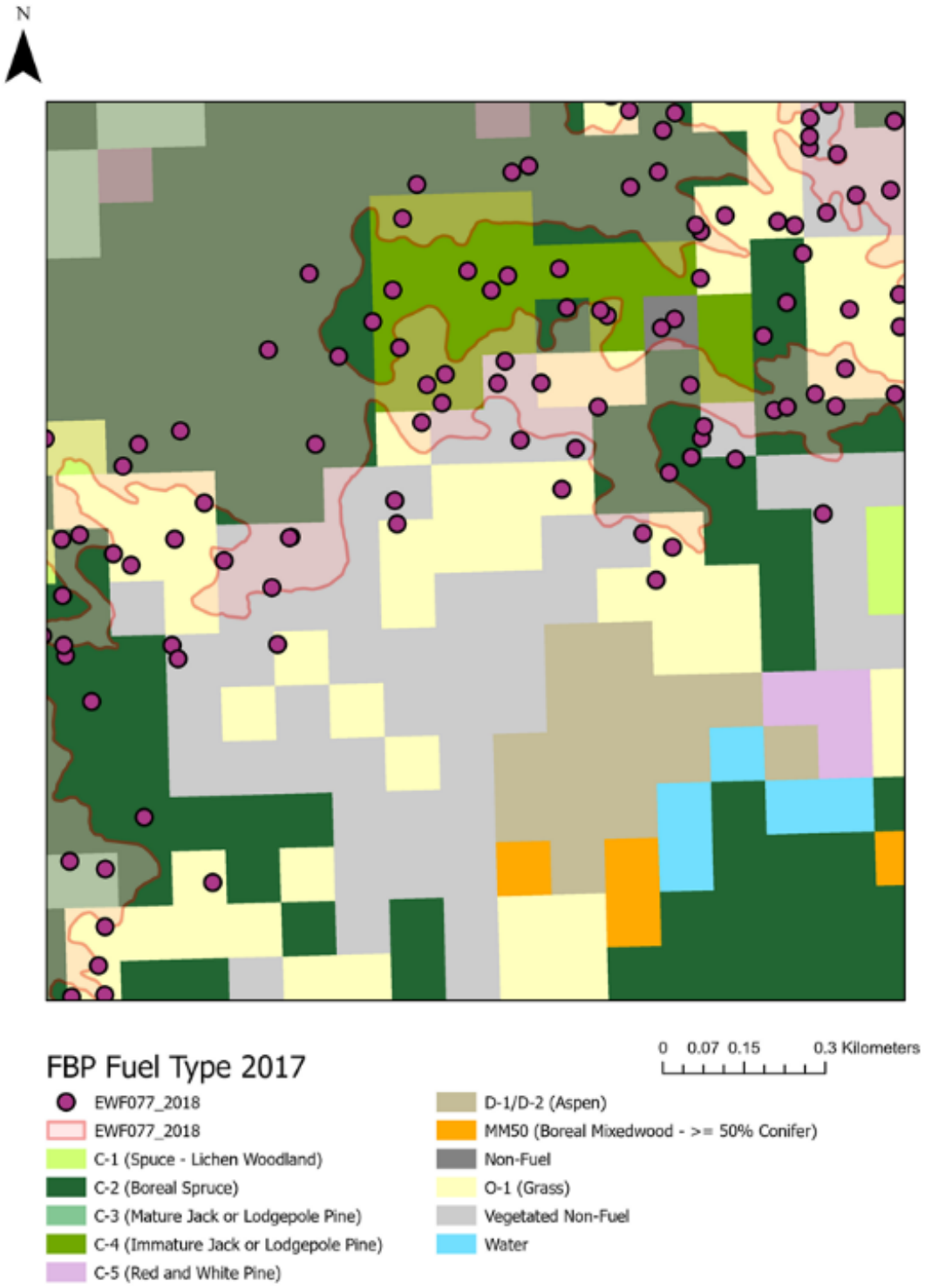


Figure 2-10 Example of the non-fuel uncertainty

The third cleaning algorithm is Input Data Cleaning. For points outside of the fire, the variable OFFSET must be -100; for points inside the fire, the value must be 100. This step of cleaning is mainly for the prevention of transcriptional errors. Due to the data needing to be transferred from ArcGIS to R to implement regression modelling through multiple transfers, occasionally, due to input errors, the data in the Excel file may not follow the standard order (first for data with Offset = -100, then for data with Offset = 100). A custom function was created using R-code to identify and correct those inputs.

The fourth algorithm is to handle missing values. The missing values at variables 'DataClean' and 'PreviousFireALL' each represent a specific meaning. For 'DataClean,' it represents a very clear physical meaning: no fire, and for 'PreviousFireALL,' it means no previous fires happened at these data points within the period of the record (i.e., since 1933). These missing values were not treated; they were assigned a specific, very large constant, 999999. Time-since-fire (TSF) can be easily estimated by calculating the difference between the sampled fire year and earliest prior fire years from all the sampled points. For all other missing values, if found, a custom function has been created to remove the data points and delete the corresponding data points in pairs.

2.8.Joint Modelling Framework

2.8.1. MCC Clogit Modelling

The Clogit model is established to explore the influence of water, anthropogenic factors, topography, fuel or land cover, and previous fires on the cessation of fire. From a famous quote, "The first law of geography: everything is related to everything else, but near things are more related than distant things." (Tobler, 1970), the importance of spatial autocorrelation when using and sampling spatial data was acknowledged by previous studies of fire boundaries (Narayanaraj & Wimberly, 2011; Holsinger et al., 2016; Negret et al., 2020; Macauley et al., 2022). They

universally recommended independent case matching, as case-control pairing has significantly reduced confounding issues and improved model effectiveness.

In the model, the binary variable, *CASE*, indicates whether (*CASE*=1) or not (*CASE*=0) a location can sustain a fire cessation event, which is the main interest. There are a series of explanatory variables $\boldsymbol{x}$ ($x_1, x_2, \ldots, x_K$), with coefficients $\boldsymbol{b}$ ($b_1, b_2, \ldots, b_K$). Like a linear regression, the relationship between CASE and x can be written as:

$$CASE = \mathrm{a} + \sum_{1}^{K} b_k x_k \qquad (2-2)$$

As CASE is a binary variable, a probability model is used instead to describe the probability of developing an event. The conditional probability formula for *CASE=1* given $\boldsymbol{x}$ can be written as a sigmoid function (Breslow et al., 1978):

$$p = pr\{CASE = 1|\boldsymbol{x}\} = \frac{1}{1 + e^{-a-\sum_1^K b_k x_k}} \qquad (2-3)$$

After substituting the equation (2-2) into Breslow's probability formula,

$$p = \frac{1}{1 + e^{-\mathrm{CASE}}} \qquad (2-4)$$

Based on the above probability *p* and the definition of the odds, where:

$$Odds = \frac{Probability\ of\ event}{Probability\ of\ non-event} \qquad (2-5)$$

There is:

$$Odds = \frac{p}{1-p} = \frac{\frac{1}{1+e^{-\mathrm{CASE}}}}{1-\frac{1}{1+e^{-\mathrm{CASE}}}} = \frac{1}{e^{-\mathrm{CASE}}} = e^{\mathrm{CASE}} \qquad (2-6)$$

Logging both sides of the Equation (2-6), there is:

$$\ln(\mathrm{odds}) = \ln\left(\frac{p}{1-p}\right) = \ln(e^{\mathrm{CASE}}) = \mathrm{CASE} = a + \sum_{1}^{K} b_k x_k \qquad (2-7)$$

By defining:

$$h(t) = \left(\frac{p}{1-p}\right), \qquad \ln\left(h_0(t)\right) = a \qquad (2-8)$$

Equation (2-6) is in the same form as the log function of Cox's hazard function, which is:

$$h(t) = h_0(t)\exp\left(b_K x_K\right) \qquad (2\text{-}9)$$

The Cox proportional hazards model is widely used in survival analysis in the medical field (Gail et al., 1981). It estimates hazard (Odds) for each covariate, which represents the relative likelihood of experiencing the event at any given time, comparing different levels of covariates. *t* is time of survival, *h(t)* is the hazard influenced by covariates $\boldsymbol{x}(x_1, x_2, \ldots, x_k)$ with coefficients $\boldsymbol{b}$ $(b_1, b_2, \ldots, b_K)$. The baseline hazard $h_0(t)$ represents the hazard when all covariates equal zero. $exp\ (b_k)$ is called Hazard of a specific covariate. If b_k is positive, this would make the value of HR $= exp\ (b_k)$ greater than 1, meaning as the value of the covariate x_i increases, contribution of b_i to the event hazard is positive. Thus, the probability of a hazard event increases. Since the fire cessation event is this study's main interest, here, the "hazard" means fire cessation, where fire boundary formed.

Due to the medical origin of the Cox proportional hazards model, the model often must deal with different human clinical clients as the input. The key assumption for the Cox model is that the survival curve of $h(t)\sim t$ for the two clients cannot be crossed over.

For example, for client *k*, its hazard function is:

$$h_K(t) = h_0(t)e^{\sum_1^K b_k x_k} \qquad (2-10)$$

For client *k'*, its hazard function is:

$$h_{K'}(t) = h_0(t)e^{\sum_1^{K'} b'_k x'_k} \qquad (2-11)$$

The ratio of the hazard (Hazard Ratio) between these two clients is

$$\frac{h_k(t)}{h_{k'}(t)} = \frac{h_0(t)e^{\sum_1^k b_k x_k}}{h_0(t)e^{\sum_1^{k'} b_{k'} x_{k'}}} = \frac{e^{\sum_1^k b_k x_k}}{e^{\sum_1^{k'} b_{k'} x_{k'}}} \qquad (2-12)$$

The hazard ratio is not related to time *t*. It means that for any set of clients (aka sample), the hazard of the events is only changing with the covariates. This is exactly the relative ratio in MCC Clogit (Breslow et al., 1978); by setting the matched case-control pair as the strata input for the model, this model can focus on the difference inside each case-control pair without external influence from other sample pairs (Harrel, 2015). This explains why, in R, the code for MCC Clogit is equivalent to the code for CoxPh, which is what will be used in this thesis under the R package "survival" (Therneau, 2023) to do MCC Clogit modelling.

2.8.2. Testing Multicollinearity (VIF) in the Clogit model

Multicollinearity among the explanatory variable for every model is tested using the Variance Inflation Factor (VIF). Because the variables contain both categorical variables and numeric

variables, a more generalized Variance Inflation Factor (GVIF) was used in the form of

$$Ajusted\ GVIF = GVIF^{\frac{1}{2 \times Degreeof\ Freedom}} \quad (2-13)$$

(Fox & Monette, 1992). Fox and Monette (1992) found that it provides a comprehensive measure of collinearity for each variable, and comparability among explanatory variables with different dimensions.

Attempting to use all the variables to run the MCC-Clogit models always produces very high VIF values. Many methods were explored to decrease VIF values (O'brien, 2007). In the earlier stage of this thesis, it was observed that the adjusted GVIF value is always very high for fuel dummy variables. By setting a threshold being 10, those variables with high adjusted GVIF were removed, same as (O'brien, 2007). However, it was found that some variables' GVIF will increase after such removal. It was found that dummy variables with too few data points can retain a high GVIF value and introduce unwanted bias to the model. Therefore, the final method to decrease GVIF is not to remove the highest GVIF variable but to remove dummy variables with too few sampled points. After this, all variables left in the established model have GVIF (< 10).

2.8.3. Contingency Evaluation

To raise model accuracy, an additional filtering option based on the contingency table was proposed, which means when applied, out of all the explanatory variables, only the ones with significant association were selected to build the final model. When this filtering option was applied, it is denoted as "cp0.1", meaning the variable with a p-value of the contingency table less than 0.1 was selected. The p-value in the contingency test detects each variable's association at the points inside and outside the fire perimeter before the modelling. It could be viewed as testing for a null hypothesis: "There is no association between this variable inside and outside the fire

perimeter in the sampled." With the P value smaller than 0.1, there is only a 10% chance of observing a result like this if there is no difference between inside and outside the fire perimeter. In this study, it is viewed that any variables with the Contingency Chi-square test's P value < 0.1 have an association and should be included in the model. The models that used and did not use cp0.1 filtering are compared.

2.8.4. Stepwise Selection and Four Schemes to Run Clogit

StepAIC (Choose a model by Akaike's Information Criterion (AIC) in a Stepwise Algorithm) function from the MASS package in R was used to complete an automatic stepwise backward model selection of the full model. The models with and without StepAIC were compared.

Four schemes are designed to run MCC Clogit, as shown in Table 2-3.

Table 2-3 The four schemes of Clogit modelling

Scheme	Abbreviation	Select variables with the contingency table with a P-value less than 0.1.	Select variables using stepwise reduction.
Scheme A	Nocp0.1Nostep	No	No
Scheme B	Nocp0.1Withstep	No	Yes
Scheme C	Withcp0.1Nostep	Yes	No
Scheme D	Withcp0.1Withstep	Yes	Yes

2.8.5. Combining Clogit Model and RF

The MCC Clogit modelling framework is efficient at quantifying the effect of explanatory variables. However, because of the strata feature in Clogit, making predictions using new data becomes impossible. On the other hand, the Random Forest (RF) model is good at prediction. Still,

its capability in detecting the direction and strength of each explanatory variable's effect is limited, especially when it would require high computation power, which is the case in this thesis.

Bootstrapping creates a dataset randomly selected from the original data points, with the possibility of a data point being sampled more than once (Henderson, 2005). This way, out-of-bag points are created to represent data points that were not sampled. The idea of RF is based on bagging, a technique widely used to reduce the variation in prediction by combining the result of multiple decision trees, where each tree was trained on a different random sampling of the dataset (Breiman, 2001).

The limited efficiency of existing tools in RF further complicates the analysis. Using RF alone is not sufficient to achieve the goal of this thesis. To overcome this limitation, a hybrid analysis approach was developed to combine RF and Clogit, leveraging the merits of both models to identify the key influential factors on fire cessation over a vast amount of sampling points, and predict fire cessation probability(Figure 1-4).

2.9.Model Performance and Variable Importance Quantification

2.9.1. Assessing model fit

The concordance coefficient (CC) was used to evaluate model performance of MCC Clogit as a standard output alongside the MCC Clogit results from R's Survival Package. When the response variable is binary, as in this study, CC equals the area under curve (AUC) from the Receiver Operating Characteristic (ROC) (Therneau, 2023), which was used to assess the fit of the RF model. Therefore, the assessment of the performance of models will focus on the AUC.

As mentioned in Chapter 2.8.1, in Clogit, only the probability of the response variable CASE (event = 1, non-event = 0) is estimated by Eq. (2-3), not the response variable itself. To assess the model fit, a probability cut-off point (threshold) in the range of [0,1] is needed to derive the binary

response variable from the estimated probability (Hosmer Jr et al., 2013). Take the cut point of 0.5, for example. If the estimated probability is higher than 0.5, the value of the derived binary response variable will be 1 (event); otherwise, it will be 0 (non-event).

To assess the performance of the whole model, and calculate the optimal cut point value, the most appropriate approach is to calculate the Receiver Operating Characteristics (ROC) curve, and the Area Under the Curve (AUC) (Fawcett, 2006). The methods originated from signal detection, where the test is about whether the receiver can correctly detect the true signal under the presence of noise (false signal) (Hosmer Jr et al., 2013).

The bases of the ROC curve come from understanding the two sets of terms. The first set is the sensitivity, which is the percentage of true positives (TP) within all the positive (P) samples. In this study, the ratio of areas that were truly unburned and predicted as unburned, versus all the areas unburned (Equation 2-13).

$$Sensitivity = \frac{TP}{P} \qquad (2-14)$$

The second set of terms, Specificity, is the percentage of True Negative (TN) out of all the negative (N) samples. In this study, this is calculated as the ratio of areas that were truly burned and predicted as burned, versus all areas burned (Equation 2-4).

$$Specificity = \frac{TN}{N} \qquad (2-15)$$

For a specific cut point within the range, there is a specific pair of Sensitivity and Specificity values. The optimal cut point is often identified where Sensitivity and Specificity are equalized, meaning the two values intersect. After using all values of cut point to calculate the corresponding Sensitivity and Specificity value pairs, they can be plotted on a graph, with Sensitivity as the Y-

axis and Specificity as the X-axis (in reverse order). The value pairs are joined by lines, and the curve that appears is the ROC curve, as shown in Figure 2-11. With the increasing value of the cut point, there is always a decrease in sensitivity and an increase in specificity. The AUC curve has a minimum value of 0.5, representing the probability of a binary decision is 50% at all thresholds, and a maximum value of 1. A generally acceptable AUC value for models of binary response variables is above 0.7 (Hosmer Jr et al., 2013).

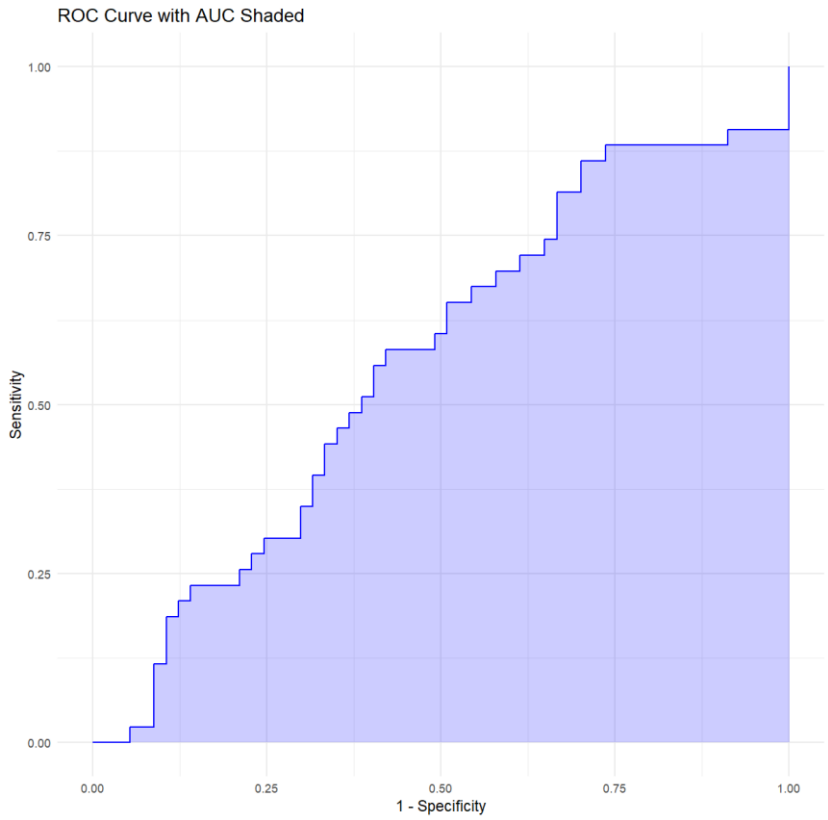


Figure 2-11 Example of AUC as shaded area under the ROC curve.

2.9.2. RF Model Importance

RF uses variable importance measures to rank the list of predictors (Breiman, 2001). There are two ways to measure the importance of covariates in RF modelling, namely Mean Decrease Gini

and Mean Decrease Accuracy (Bannerman-Thompson et al., 2013). Starting with Mean Decrease Accuracy, for example this thesis used 29 explanatory variables. The RF model would run using all the explanatory variables and calculate the baseline accuracy based on the 29 variables. Then, a single explanatory variable would be randomly shuffled across all data points (permuted). The RF model with the permuted explanatory variable would get a new and lower accuracy compared to the baseline. After many repetitions of this step, by measuring this drop in the prediction accuracy, Mean Decrease Accuracy is obtained for that explanatory variable. The higher the Mean Decrease Accuracy, the more important the variable is in the model (Bannerman-Thompson et al., 2013).

Mean Decrease Gini is based on the concept of the Gini criterion (Breiman, 2001). Still taking this thesis for example, when all the 29 explanatory variables are used to train an RF model, each variable is considered a parent node (root node). Each root node will need to be split into two daughter nodes (Bannerman-Thompson et al., 2013). Regarding each of the nodes, a Gini impurity index is used to measure how often a data point is incorrectly classified. The Gini impurity index of a parent node will always be larger than that of its daughter nodes, and this Gini difference is recorded. After the permutation of a single explanatory variable, a new Gini difference is recorded. The step is repeated and averaged, and the Mean Decrease Gini is obtained. A higher Mean Decrease Gini would suggest that the variable is more important in making accurate predictions. The importance value in the result is a relative measure, where the absolute numerical value does not represent anything, rather than serving the purpose of comparing each variable (Breiman, 2001; Hastie et al., 2001).

There were discussions about the stability of using the two indexes to rank the predictors. Calle and Urrea (2011) examined the stability of these measures and concluded that ranking based on

Mean Decrease Gini was more robust than those based on Mean Decrease Accuracy. However, Nicodemus (2011) later pointed out that ranking based on Mean Decrease Gini was sensitive to within-predictor correlation and differences in category frequencies, even when number of categories was held constant. Nicodemus (2011) argued that under a strong within-predictor correlation, the Mean Decrease Gini ranking was less stable than the Mean Decrease Accuracy. This thesis will use both measures to rank the predictors list and compare the results with those identified by Clogit. The significance of the importance measure for each variable in the RF model is calculated using the R package "rfpermute" (Archer & Archer, 2016). However, for higher dimensional data used in this thesis, the calculation is prolongated in time.

2.10. Fire-stop Prediction

Sampling new fires or regions is specially designed for fire perimeter prediction (Table 2-4).

Table 2-4 Process of sampling new areas

Step	Content
Step 1	Determine which fire will be evaluated for the prediction.
Step 2	Use the edit function in ArcGIS Pro to draw a rectangle around the specific fire, ensuring that each side of the rectangle is at least 100 meters away from the fire perimeter.
Step 3	Use the Create Fishnet geoprocessing tool in ArcGIS Pro to sample points at a 100 m resolution within the rectangle and tick the "Create Label points" option in the toolset, which would generate a points layer.
Step 4	Use the points layer as input and integrate it into the extract data model built previously to extract all the variables into the points layer.
Step 5	Use the Add XY coordinate geoprocessing tool to add the longitude and latitude of each point under the same projection.
Step 6	Export the table and then use the conversion tool to convert the ArcGIS table into Excel for RStudio to read.

Step 7	Read the Excel data and run through all the cleaning and classification of variable processes to prepare other sampled points.
Step 8	Predict the fire perimeter probability using the cleaned data model from the matched case-control sampled points.
Step 9	Output the results and plot in ArcGIS Pro.
Step 10	Rasterize the points and then use the Kriging interpolation method to produce a probability map of the rectangle area.

The result is the probability of forming a fire perimeter estimated at every one of the points on the map, with resolution only being determined by the computation power of the user. When viewed jointly with the other landscape features, such as waterways and road maps, there will be more confidence in explaining the map's high- and low-probability areas.

3. Quality Control and Descriptive Data Analysis

In this Chapter, the data points sampled and extracted through the proposed automatic algorithms will be evaluated from two aspects: first, by conducting a visual inspection of individual fires for quality control, and second, by performing a descriptive analysis of each explanatory variable.

3.1. Data Quality Control

3.1.1. Overview

After processing automated data sampling and extraction, data quality control was implemented for each pair of automated data sampling points. This was done by visual inspection of each fire one by one immediately after the sampling process. Additionally, after the extraction process, random inspections of the resulting input Excel files were performed to check for any issues in the data before importing it into modelling by R software. The automated data sampling and extraction process successfully samples matched case-control data points when the fire ID corresponds to a single unique fire polygon.

The automated algorithm processes each fire systematically, sampling one fire polygon per fire ID. However, if the fire ID is associated with multiple fire polygons, or if the fire polygons don't enclose into a complete polygon, or if the fire has particularly complicated fire perimeters, then large portions of the fire would remain unsampled or incorrectly sampled. Three types of quality control measures were implemented in response to each of these cases, as follows:

(1) Category 1: Fire with the same fire ID but with multiple fire polygons.

This happens when a single fire ID is associated with multiple fire polygons. Typically, the polygons of each fire were originally ordered by size under the same fire ID. Thus, the polygon to be sampled by the automation algorithm is usually the largest polygon of each fire, while other polygons are all relatively tiny. Even though this would mean that some

small parts of the fire perimeter were unsampled, the largest polygon's sampling was sufficient to cover most of the fire perimeter. This method effectively preserves the most effective sample points automatically without requiring special treatment. These fires are exactly those large fires with burn code=" B" but usually have unburned islands and partially burned areas scattered around, as described in the methods section.

However, through careful review, it was discovered that some polygons were incorrectly ordered by size due to unforeseen errors, resulting in sampling a tiny polygon instead of the largest one. For these fires, the polygons for each fire ID were reordered with fire size from high to low. This guarantees that the largest fire polygon will be sampled.

Additionally, there were cases where the two largest polygons of similar size existed within the same fire ID, leading to only one polygon being sampled, leaving nearly half of the fire area unsampled. For such fires, each of the equally large fire polygons was artificially separated into new fires with specific fire IDs starting from 900 before redoing the sampling. Fires that have those issues are identified as Category 1.

(2) Category 2, Same fire with different fire ID

In rare situations, two fire polygons with different fire IDs belong to one actual fire. Part of the fire perimeter polygons, a portion of the single fire, now appears on the map as two separate fires. Such fires will be combined into one new fire under a new name, counting down from 999, and then the sampling process is re-done.

(3) Category 3, complicated fires

For fires that are detailed and complicated, they must be removed. An example of this case is the Fort McMurry fire in 2015, the fire ID being MWF009_2016. This fire perimeter was mapped in detail, even for small elements, such as a single roadway. This exceeded the level

of detail of all other digitized fires, and the developed sampling algorithms cannot sample these fires without extensive manual adjustments; as such, these fires were removed. The detailed processing of all the above categories is listed in Table 3-1. An example of data quality control to process fires belonging to Category types 1 and 3 is shown in Figure 3-1.

Table 3-1 The procedure of data quality control

Category	Fires found with many missing values with explanatory variables	Processing procedure
1	HWF106_2012	
1	HWF107_2012	
1	HWF111_2015	
1	LWF175_2015	
1	HWF135_2012	Sampled the largest polygon by using the same fire ID
1	HWF142_2012	
1	LWF161_2015	
1	MWF091_2015	
1	MWF094_2015	
1	HWF120_2012	Separate into two new fires: HWF900_2012 & HWF901_2012
1	MWF106_2015	Separate into two new fires: MWF900_2015 & MWF901_2015
1	MWF110_2015	Separate into three new fires: MWF902_2015, MWF903_2015 & MWF904_2015

1	SWF163_2015	Separate into two new fires: SWF900_2015 & SWF901_2015
2	HWF212_2015	Combined into a new fire: HWF998_2015
2	HWF264_2015	
2	HWF217_2015	Combined into a new fire: HHWF999_2015
2	HWF220_2015	
2	MWF078_2015	Combined into a new fire: MWF999_2015
2	MWF101_2015	
3	HWF100_2016	Removed because too complicated
3	HWF193_2016	Removed because too complicated
3	MWF007_2011	Removed because containing too detailed information
3	MWF042_2013	Removed because the original fire is an empty file

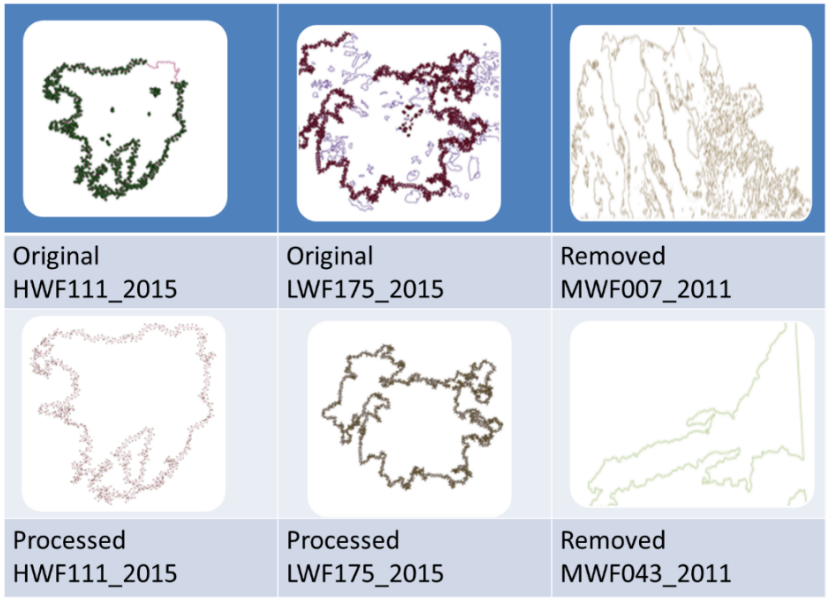


Figure 3-1 Example of quality control process of fires which belongs to Category 1 and 3

The example of data quality control to process fires belonging to Category 2 fires is shown in Figure 3-2. The fires are representative of ones belonging to Category 1 and 3

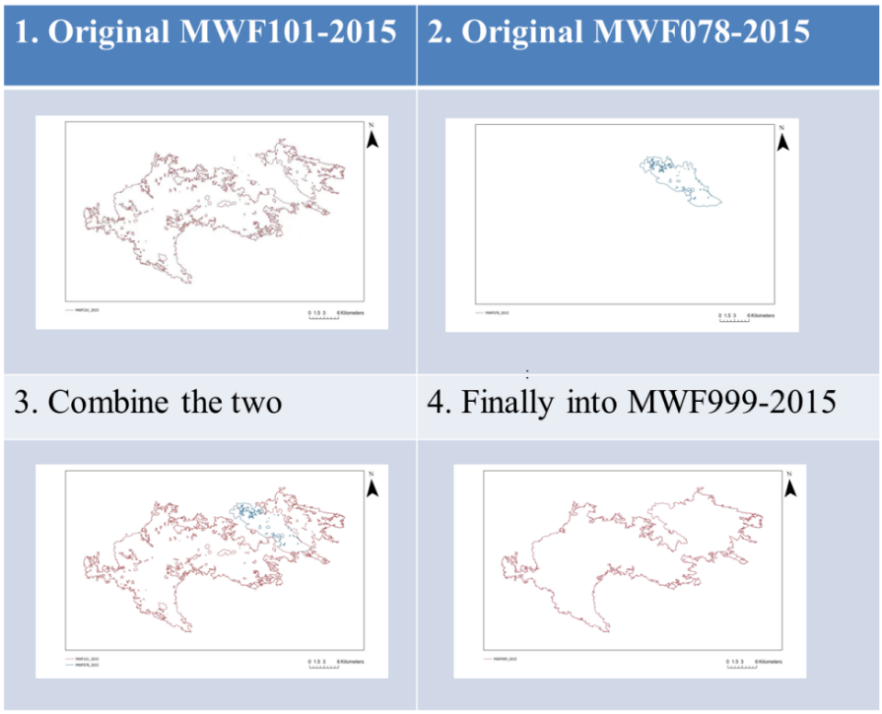


Figure 3-2 Example of data quality control process of fires which belongs to category 2

3.1.2. Final Fires Number and Data Pair Number

Following implementation of quality control measures, a total of 109 fires were selected. There were 52,037 data pairs (104,074 data points) sampled along the 109 fire perimeters. Following removal of missing data, 51,833 data pairs (103,666 data points) remained. After IO (inside-outside) cleaning, 71,030 data points remained. After the NVB clean, 65,078 data points remained. By applying a 70:30 percent random splitting for model training and testing , 45,336 data points remained for training data and 19,742 data points for testing as summarized in Table 3-3. The splitting approach was applied to all the data pairs instead of fire number, due to fires in Fire Class E vary greatly in sizes, and such that some large fires contain substantially more sample pairs than others. The Law of large numbers (Grimmett & Stirzaker, 2020) ensures that random sampling in a large sample size, would preserve the proportion of different categories (in this case, individual fires) for data points for training and testing respectively.

Table 3-2 Final fires number and data pair number sampled and extracted.

Total number of fires	109
Initial Data points	104074
Missing data clean applied	103666
After Inside-Outside clean applied	71030
After NVB clean applied	65078
Train and test data divided	45336: 19742

3.2. Descriptive Analysis of Explanatory Variables

The distribution and characteristics of the explanatory variables were explored with summary statistics and visualization, with the aim of selecting variables for later modelling that best capture the pattern of the fire perimeter formation. This analytical process was applied to both the training and testing dataset, ensuring the robustness of variable selection for subsequent modelling steps.

There were five categories of explanatory variables considered. These included one variable related to roads (EucDis_Water), two related to water (EucDis_Road and Water proportion), two associated with geomorphology (Slope_percentRise and Elevation), eight pertaining to previous fires (previous fire 0 – 10 years ago, previous fire 10 -20 years ago, previous fire 20 – 30 years ago, previous fire 30 – 40 years ago, previous fire 40 to 50 years ago, previous fire 50 to 60 years ago, previous fire 60 to 70 years ago, previous fire from more than 70 years ago) and sixteen generated FBP fuel dummy variables (C-1 (Spruce-Lichen Woodland), C-2 (Boreal spruce), C-3 (Mature Jack or Lodgepole Pine),C-4 (Immature Jack or Lodgepole Pine), C-5 (Red and White Pine), C-7 (Ponderosa Pine–Douglas-Fir), D-1/D-2 (Aspen), S-2 (White Spruce - Balsam Slash), O-1 (grassland), M1/2 (mixedwood),NF (non-fuel) W (Water), VNF (vegetated non-fuel), 10-40% confer mixedwood, 40-60% confer mixedwood and 60-90% confer mixedwood). Among them, the variables for roads, water, and geomorphology are continuous, and all other variables are categorical. In total, there are 29 explanatory variables. The responding variable (CASE) is also a binary variable, with 1 and 0 values representing outside fire and inside fire, respectively.

To reduce the number of variables used for modelling, descriptive analyses using summary statistics (minimum, first quartile, median, mean, third quartile, and maximum) and data visualization were used to evaluate variables between the case and control points. Binary variables

underwent additional analysis with contingency tables to explore the distribution shape and other data characteristics on each side of the perimeter.

3.2.1. Continuous Variables

The descriptive analysis results of all continuous variables are summarized in Table 3-2 and Table 3-3, and each individual continuous variable from Figure 3-3 to Figure 3-7. There was no distinct difference between Euclidean Distance to Roads, and the slope inside and outside the fire perimeter. It showed a clean and consistent pattern that the closest distance to water outside is less than the Euclidean Distance to water inside the fire perimeter, but only in the training data. There is a distinct and logical pattern indicating that the proportion of water outside is larger than the proportion inside the fire perimeter, for both training and testing data. Similarly, the elevation variable reveals a consistent, logically expected pattern, indicating that elevation outside the fire perimeter is lower than elevation inside, in both training and testing data.

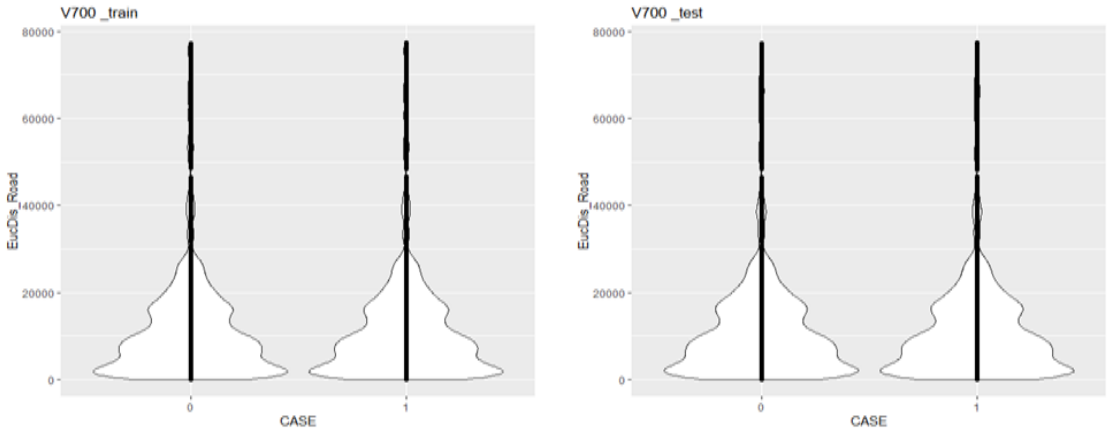


Figure 3-3 The distribution of the distance of the sampling points to a road (left panel: train dataset; right panel: test dataset)

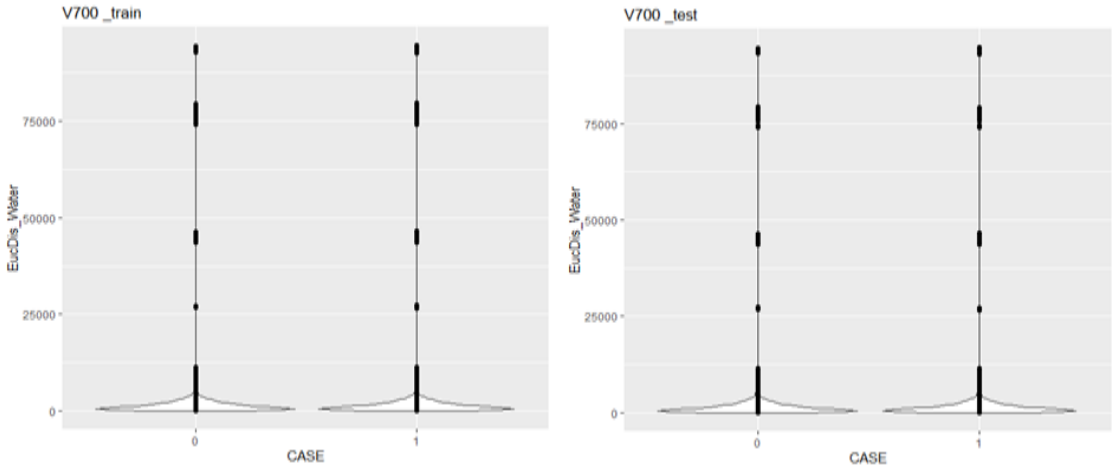


Figure 3-4 The distribution of the distance of the sampling points to water (left panel: train dataset; right panel: test dataset)

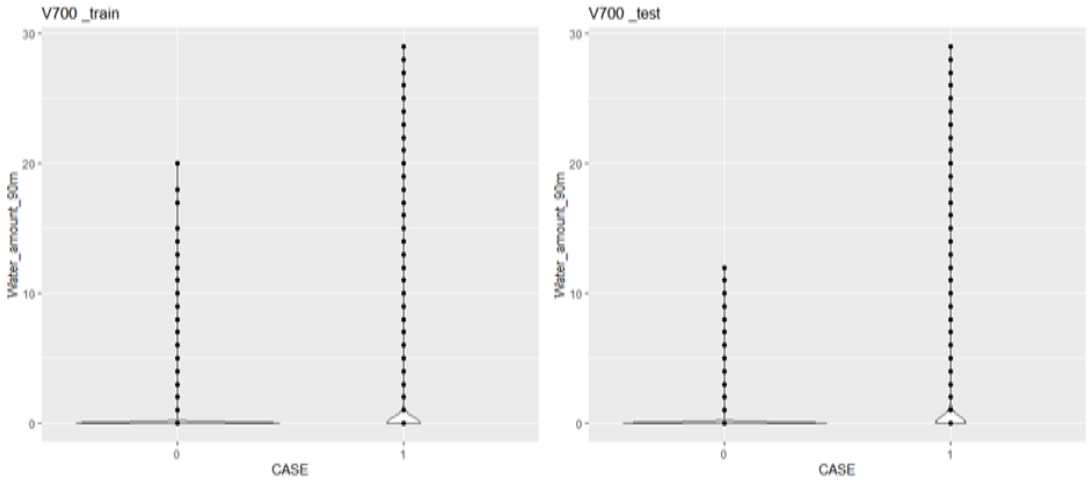


Figure 3-5 The distribution of the area of water proportion where the sampling point is located (left panel: train dataset; right panel: test dataset)

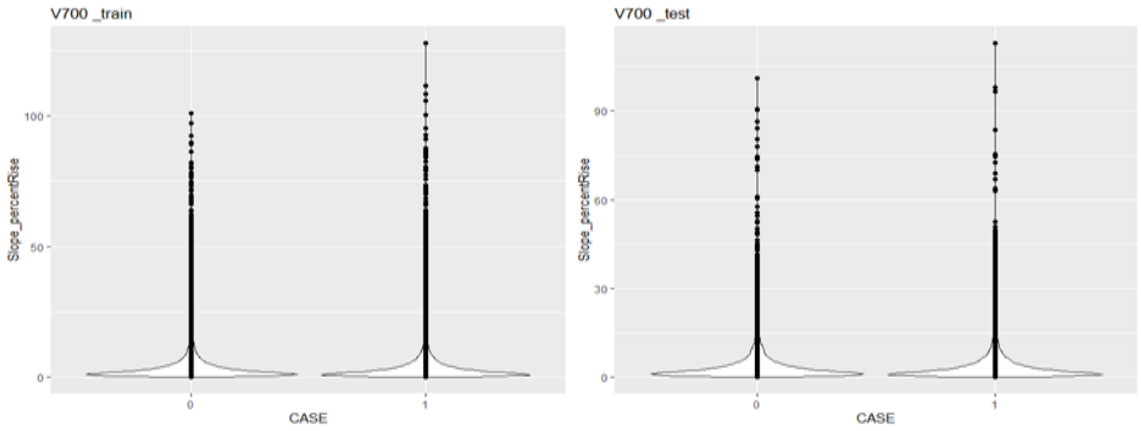


Figure 3-6 The distribution of slope where the sampling point is located (left panel: train dataset; right panel: test dataset)

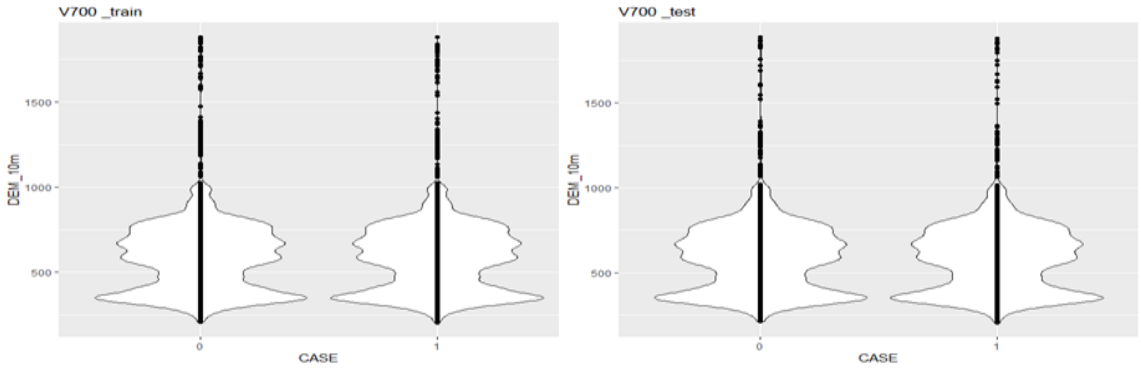


Figure 3-7 The distribution of the elevation where the sampling point is located (left panel: train dataset; right panel: test dataset)

Table 3-3 Descriptive analysis table for training data

Continuous variable	Inside Min	Inside First Quantile	Inside Median	Inside Mean	Third Quantile	Inside Max	Outside Min	Outside Frist Quantile	Outside Median	Outside Mean	Outside Third Quantile	Outside Max
Distance to road (m)	0	620	8859	12732	16798	77198	0	3623	9967	12730	16784	77392
Distance to water (m)	0	485	991	1794	1831	94549	0	484	983	1785	1829	94647
Water proportion	0	0	0	0.59	0	20	0	0	0	0.7	0	29
Slope (% rise)	0	1.00%	2.02%	3.68%	4.15%	101.01 %	0	0.92%	1.96%	3.93%	4.29%	127.80 %
Elevation (m)	211	4.3	590	580	714	1889	204	4.3	589	579	713	1883

Table 3-4 Descriptive analysis table for testing data

Continuous variable	Inside Min	Inside First Quantile	Inside Median	Inside Mean	Third Quantile	Inside Max	Outside Min	Outside Frist Quantile	Outside Median	Outside Mean	Outside Third Quantile	Outside Max
Distance to road(m)	0	3512	8754	12587	16738	77175	0	3504	8778	12588	16752	77385
Distance to water(m)	0	509	997	1738	1828	94597	0	489	999	1729	1830	94765
Water proportion	0	0	0	0.05	0	17	0	0	0	0.68	0	29
Slope	0	1.01%	2.01%	3.68%	4.13%	101.13 %	0	0.90%	1.91%	3.74%	4.10%	111.45 %
DEM(m)	209	401	586	574	709	1875	205	400	585	573	708	1848

3.2.2. Binary variables

3.2.2.1. Overview

The contingency relationship for the binary variable for both the V700 train data and test data are presented in Appendix 1 and 2, summarized in Table 3-4 and Table 3-5, respectively.

Table 3-5 Contingence table for training data

	Binary variable	Inside Non Existence	Inside Existence	Outside Non Existence	Outside Existence	p-value	Train	If it is cp0.1 Variable?
1	FBPC1	20345	2323	20648	2020	0.001	45336	N
2	FBPC2	10851	11817	14097	8571	0.001	45336	N
3	FBPC3	22089	579	22291	377	0.001	45336	N
4	FBPC4	22039	629	22227	441	0.001	45336	N
5	FBPC5	22623	45	22626	42	0.83	45336	Y
6	FBPC7	22666	2	22666	2	1	45336	Y
7	FBPD1/D2	20932	1736	19755	2913	0.001	45336	N
8	FBPS2	22638	30	22651	17	0.08	45336	N
9	FBPO1	19402	3266	18748	3920	0.001	45336	N
10	FBPM12	22668	0	22667	1	1	45336	Y
11	FBPNF	22668	0	22292	376	0.001	45336	N
12	FBPW	22525	143	21884	784	0.001	45336	N
13	FBPVNF	22668	0	22073	595	0.001	45336	N
14	FBP10_40	22423	245	22256	412	0.001	45336	N
15	FBP40_60	21651	1017	21277	1391	0.001	45336	N
16	FBP60_90	21834	834	21863	805	0.481	45336	N

17	PrevFire10	22632	36	22341	327	0.001	45336	N
18	PrevFire20	22616	52	22554	114	0.001	45336	N
19	PrevFire30	22483	185	22471	197	0.572	45336	Y
20	PrevFire40	20727	1941	20627	2041	0.1	45336	Y
21	PrevFire50	22293	375	22295	373	0.971	45336	Y
22	PrevFire60	22510	158	22519	149	0.647	45336	Y
23	PrevFire70	22057	611	22064	604	0.861	45336	Y
24	PrevFire70plus	3358	19310	3805	18863	0.001	45336	N

Table 3-6 Contingence table for testing data

	Binary variable	Inside Non Existence	Inside Existence	Outside Non Existence	Outside Existence	p-value	Train	If it is cp0.1 Variable?
1	FBPC1	8881	990	8989	882	0.009	19742	N
2	FBPC2	4720	5151	6132	3739	0.001	19742	N
3	FBPC3	9658	213	9714	157	0.004	19742	N
4	FBPC4	9627	244	9701	170	0.001	19742	N
5	FBPC5	9856	15	9854	17	0.86	19742	Y
6	FBPC7	9870	1	9870	1	1	19742	Y
7	FBPD1/D2	9098	773	8558	1313	0.001	19742	N
8	FBPS2	9843	28	9858	13	0.029	19742	N
9	FBPO1	8431	1440	8219	1652	0.001	19742	N

10	FBPM12	9871	0	9871	0	1	19742	Y
11	FBPNF	9871	0	9696	175	0.001	19742	N
12	FBPW	9818	53	9534	337	0.001	19742	N
13	FBPVNF	9871	0	9606	265	0.001	19742	N
14	FBP10_40	9763	108	9681	190	0.001	19742	N
15	FBP40_60	9415	456	9239	632	0.001	19742	N
16	FBP60_90	9473	398	9543	328	0.009	19742	N
17	PrevFire10	9861	10	9734	137	0.001	19742	N
18	PrevFire20	9853	18	9813	58	0.001	19742	N
19	PrevFire30	9796	75	9794	77	0.935	19742	Y
20	PrevFire40	9081	790	9056	815	0.532	19742	Y
21	PrevFire50	9707	164	9708	163	1	19742	Y
22	PrevFire60	9799	72	9805	66	0.669	19742	Y
23	PrevFire70	9597	274	9598	273	1	19742	Y
24	PrevFire70 plus	1403	8468	1589	8282	0.001	19742	N

Analyzing the Training data as shown in Table 3-5, the count of the occurrence for variables including C-1 (Spruce-Lichen Woodland), C-2 (Boreal spruce), C-3 (Mature Jack or Lodgepole Pine), C-4 (Immature Jack or Lodgepole Pine), FBPC5 (Red and White Pine), S-2 (White Spruce–Balsam Slash), 40-60% confer mixedwood, previous fire 40 to 50 years ago, previous fire 50 to 60 years ago, previous fire 60 to 70 years ago and previous fire from more than 70 years ago inside the fire perimeter exceeds the count outside the fire perimeter. In contrast, for variables including

D-1/D-2 (Aspen), O-1 (grassland), NF (non-fuel), W (Water), VNF (vegetated non-fuel), 10-40% confer mixedwood, 40-60% confer mixedwood, previous fire 0 to 10 years ago, previous fire 10 to 20 years ago, previous fire 20 to 30 years ago and previous fire 40 to 50 years ago, the count inside the fire perimeter is lower than the outside.

For FBPC7 (Ponderosa Pine–Douglas-Fir) and M1/2 (mixedwood), the count inside and outside the fire perimeter are the same and very small (0, 1 or 2).

Examining the testing data as shown in Table 3-6, the pattern remains consistent, with the count of occurrence for variables including C-1 (Spruce-Lichen Woodland), C-2 (Boreal spruce), FBPC3, C-4 (Immature Jack or Lodgepole Pine), S-2 (White Spruce–Balsam Slash), 60-90% confer mixedwood, previous fire 50 to 60, 60 to 70, and more than 70 years ago has more points inside the fire perimeter than outside. Conversely, variables D-1/D-2 (Aspen), O-1 (grassland), NF (non-fuel), W (Water), VNF (vegetated non-fuel), 10-40% confer mixedwood, 40-60% confer mixedwood, previous fire 0 to 10, 10 to 20, 20 to 30 and 30 to 40 years ago exhibit a lower count occurrence inside the fire perimeter compared with outside the fire perimeter.

For FBPC5 (Red and White Pine), FBPC7 (Ponderosa Pine–Douglas-Fir), M1/2 (mixedwood) and previous fire 40 to 50 years ago, the counts inside and outside the fire perimeter are almost the same or very small (0, 1, or 2). This comprehensive analysis of the binary variable distribution provides valuable insight into the prevalence of those variables within and outside the fire perimeter in both the training and testing datasets.

3.2.2.2. Variable selection based on contingency table

The descriptive analysis reveals that some variables, such as FBPC5 (Red and White Pine), FBPC7 (Ponderosa Pine–Douglas-Fir), M1/2 (mixedwood) previous fire 40 to 50 years ago,

exhibit nearly equal counter occurrence inside and outside the fire perimeter, often totalling only 0 or 1 or 2 instances among all the selected fires. Correspondingly, the p-values of the contingency table for these variables are 1. Additionally, several variables have p-values less than one but higher than 0.1. Specifically, 8 out of 29 variables, as observed from the contingency analysis, have a very weak association between occurrence inside and outside the fire perimeter. After removing weak variables, the next step was to run models with the reduced variables, denoted as cp0.1, either removed or retained.

3.3.Discussion

Matching pairs is a critical statistical step employed in MCC studies to ensure that the distribution of selected variables is similar across different groups. This method is instrumental in controlling for confounders variables that could potentially skew the analysis, thereby enhancing the statistical efficiency of the study. Matching on confounders can improve statistical efficiency (i.e., reduce the variance and improve power) for effect estimation. In this thesis, the implementation of pair matching is particularly challenging due to the vast scale and the intricate nature of the fire landscape. The complexity is compounded when dealing with large fires, where the perimeters can be highly irregular, presenting a significant challenge in ensuring accuracy and representation.

This thesis employs automated algorithms to streamline the data sampling process to mitigate these challenges. The time and labour required to collect such data have now been significantly reduced, transforming what used to be a thorough process spanning weeks or months into a task that can be completed in days or hours. However, despite the convenience, automated algorithms may not always navigate the nuances of complex fire perimeters with complete accuracy. For example, the fire perimeter, represented as MWF007_2011 as described in Section 3.1, is particularly complicated. Some fires present with unique perimeter characteristics that are

incomparable to any other fires. The algorithm may struggle with irregular shapes or fail to generate accurate data points, particularly when the fire perimeter is not fully enclosed. This can lead to inconsistencies or errors in the sampled data, necessitating a manual review to ensure accuracy. Perimeter irregularities can confuse the custom Python script written to determine the location of the sampled point, especially given it relies on the assumption of closed geometric shapes to function correctly. By combining the efficiency of automation with the discerning eye of human review, this thesis reached a balance between speed and accuracy in analyzing fire perimeters within the Alberta Boreal zone.

3.4. Summary

Through visual inspection and descriptive analysis, this chapter has shown that the automated data sampling process and extraction were successful. Through data quality control methods, solutions were found to deal with the algorithms' challenges. The largest polygon was guaranteed to be sampled for fire polygons under the same fire ID. For fires which have two or more equally large polygons, each of the large polygons was separately resampled. For single fires that belonged to one fire but were misclassified into two separate fires, each fire pair was combined into one new fire and then re-sampled. For fires that were extremely complicate, were removed.

Through the above processing, 109 fires were selected for analysis. After the cleaning process, 65,078 data points remained for modelling. By applying a 70:30 percent random splitting on all data points together for RF modelling, there are 45,336 data points for training data and 19,742 data points for testing. Through the descriptive analysis, most explanatory variables demonstrated clean association inside and outside the fire boundary, aligning with common knowledge in physics. This shows a good database for further modelling in the next chapter.

4. Key Factors Detection

In this Chapter, based on the sampled data pairs, six schemes (i.e., candidate lists of key factors), were produced, four from MCC Clogit and two from RF modelling. When modelling the variables in Clogit, the continuous variables of Euclidean distance to water, roads, water proportion, slope, and elevation were all standardized, with the aim of making the standardized coefficients for those continuous variables comparable.

Table 4-1 The six variable schemes for comparison

Scheme	Abbreviation	Description
Scheme A	S29Nocp0.1Nostep	Scheme with all 29 variables, no contingency test. No stepwise selection
Scheme B	S20Nocp0.1Withstep	Scheme with 20 variables, no contingency test, with stepwise selection
Scheme C	S20Withcp0.1Nostep	Scheme with 20 variables, include contingency test, no stepwise selection
Scheme D	S18Nocp0.1Withstep	Scheme with 18 variables, include contingency test, with stepwise selection
Scheme E	RF20	Scheme with 20 variables determined by Random Forest Importance (mean decrease accuracy)
Scheme F	RF18	Scheme with 18 variables determined by Random Forest Importance (mean decrease accuracy)

4.1.Clogit Modelling Schemes

By accounting for removing or retaining the cp0.1 variables as described in previous chapters, employing stepwise or not, four sets (i.e., schemes) of Clogit modelling results were obtained.

Each applied to V700 train data encompassing 109 fires. The specifics of each Clogit-scheme are described sequentially below.

4.1.1. Scheme A: S29Nocp0.1Nostep

This scheme utilizes Clogit modelling based on all 29 variables without stepwise selection and contingency table selection. The equation is:

$$\begin{aligned}\text{Clogit}(\text{CASE} \sim \text{EucDis\_Road} + \text{EucDis\_Water} + \text{Water\_amount\_90m} \\ + \text{Slope\_percentRise} + \text{DEM}_{10\text{m}} + \text{FBPC1} + \text{FBPC2} + \\ \text{FBPC3} + \text{FBPC4} + \text{FBPC5} + \text{FBPC7} + \text{FBPD12} + \text{FBPS2} + \text{FBPO1} \\ + \text{FBPM12} + \text{FBPNF} + \text{FBPW} + \text{FBPVNF} + \text{FBP10\_40} + \text{FBP40\_60} + \text{FBP60\_90} + \\ \text{PreviousFire10} + \text{PreviousFire20} + \text{PreviousFire30} + \text{PreviousFire40} + \\ \text{PreviousFire50} + \text{PreviousFire60} + \text{PreviousFire70} + \text{PreviousFire70Plus} \\ + \text{strata}(\text{Sample}), \text{data} = \text{mydata})\end{aligned}$$

In RStudio, it can also be written as:

$$\begin{aligned}&\text{coxph}(\text{formula} \\ &= \text{Surv}(\text{rep}(1, 45336\text{L}), \text{CASE}) \sim \text{EucDis\_Road} + \text{EucDis\_Water} \\ &+ \text{Water\_amount\_90m} + \text{Slope\_percentRise} + \text{DEM\_10m} + \text{FBPC1} \\ &+ \text{FBPC2} + \\ &\text{FBPC3} + \text{FBPC4} + \text{FBPC5} + \text{FBPC7} + \text{FBPD12} + \text{FBPS2} + \text{FBPO1} \\ &+ \text{FBPM12} + \text{FBPNF} + \text{FBPW} + \text{FBPVNF} + \text{FBP10\_40} + \text{FBP40\_60} + \text{FBP60\_90} + \\ &\text{PreviousFire10} + \text{PreviousFire20} + \text{PreviousFire30} + \text{PreviousFire40} +\end{aligned}$$

$$\text{PreviousFire50} + \text{PreviousFire60} + \text{PreviousFire70} + \text{PreviousFire70Plus} + \text{strata(Sample)}, \text{data} = \text{mydata}, \text{method} = \text{"exact"})$$

The modelling result is shown in Table 4-1. It is impossible to get VIF for all the variables as the model has aliased coefficients, i.e., with some variables having too few samples.

Table 4-2 The Clogit result for Scheme A

	Name	coef	exp(coef)	se(coef)	Pr(>|z|)	Sig.
1	EucDis_Road (standardized)	0.530	1.699	1.398	0.379	
2	EucDis_Water (standardized)	-2.110	0.121	0.633	-3.335	***
3	Water proportion (standardized)	0.458	1.580	0.040	11.463	***
4	Slope (standardized)	0.119	1.126	0.021	5.759	***
5	Elevation (standardized)	-12.520	0.000	0.639	-19.585	***
6	FBPC1	-0.329	0.720	1.434	-0.229	
7	FBPC2	-0.569	0.566	1.433	-0.397	
8	FBPC3	-0.768	0.464	1.436	-0.535	
9	FBPC4	-1.046	0.352	1.438	-0.727	
10	FBPC5	-0.077	0.926	1.462	-0.053	
11	FBPC7	-0.091	0.913	2.036	-0.045	
12	FBPD12	0.822	2.275	1.434	0.573	
13	FBPM12	-1.103	0.332	1.524	-0.724	
14	FBPS2	0.213	1.238	1.434	0.149	
15	FBPO1	18.660	1.269E+08	14790.000	0.001	
16	FBPNF	18.590	1.181E+08	730.300	0.025	
17	FBPW	0.612	1.844	1.440	0.425	
18	FBPVNF	18.560	1.155E+08	581.900	0.032	
19	FBP10_40	0.763	2.146	1.437	0.531	

20	FBP40_60	0.409	1.505	1.434	0.285	
21	FBP60_90	-0.032	0.968	1.433	-0.022	
22	PrevFire10	1.489	4.431	0.387	3.845	***
23	PrevFire20	0.711	2.036	0.368	1.935	.
24	PrevFire30	0.531	1.700	0.349	1.519	
25	PrevFire40	0.439	1.551	0.111	3.957	***
26	PrevFire50	0.111	1.117	0.247	0.449	
27	PrevFire60	-0.216	0.806	0.342	-0.631	
28	PrevFire70	-0.396	0.673	0.289	-1.368	
29	PrevFire70Plus	NA	NA	0.000	NA	

4.1.2. Scheme B: S20Nocp0.1Withstep

This scheme is the same as Scheme A (S29Nocp0.1Nostep) but with stepwise selection, based on the Akaike Information Criterion (AIC). The total number of variables identified is 20, the most statistically significant among the original 29 variables in the modelling process. The results of this scheme are presented in Table 4-3. The VIF for these variables are all less than 3, as shown in Table 4-4

Table 4-3 The Clogit result for Scheme B

	Name	**coef**	**exp(coef)**	**se(coef)**	**Pr(>|z|)**	**Sig.**
1	EucDis_Water (standardized)	-2.117	0.120	0.632	-3.347	***
2	Water proportion (standardized)	0.458	1.580	0.040	11.466	***
3	Slope (standardized)	0.118	1.126	0.021	5.726	***
4	Elevation (standardized)	-12.520	0.000	0.639	-19.595	***
5	FBPC1	-0.298	0.742	0.072	-4.137	***
6	FBPC2	-0.538	0.584	0.063	-8.55	***
7	FBPC3	-0.736	0.479	0.110	-6.685	***
8	FBPC4	-1.012	0.363	0.129	-7.861	***
9	FBPD12	0.853	2.347	0.073	11.718	***

10	FBPS2	-1.072	0.342	0.520	-2.06	*
11	FBPO1	0.244	1.277	0.068	3.604	***
12	FBPNF	18.620	1.216E+08	730.500	0.025	
13	FBPW	0.643	1.902	0.147	4.374	***
14	FBPVNF	18.590	1.189E+08	582.000	0.032	
15	FBP10_40	0.790	2.203	0.115	6.856	***
16	FBP40_60	0.441	1.554	0.077	5.752	***
17	PrevFire10	1.485	4.416	0.387	3.838	***
18	PrevFire20	0.732	2.080	0.367	1.998	*
19	PrevFire30	0.532	1.702	0.350	1.521	
20	PrevFire40	0.446	1.563	0.109	4.086	***

Table 4-4 The VIF for Scheme B

	Name	GVIF	df	GVIF^(1/(2*Df))
1	EucDis_Water	1.015	1	1.007
2	Water proportion	1.134	1	1.065
3	Slope	1.012	1	1.006
4	Elevation	1.023	1	1.011
5	FBPC1	2.423	1	1.557
6	FBPC2	5.118	1	2.262
7	FBPC3	1.386	1	1.177
8	FBPC4	1.244	1	1.115
9	FBPD12	2.537	1	1.593
10	FBPS2	1.011	1	1.006
11	FBPO1	3.714	1	1.927
12	FBPNF	1.000	1	1.000
13	FBPW	1.324	1	1.151
14	FBPVNF	1.000	1	1.000
15	FBP10_40	1.270	1	1.127
16	FBP40_60	2.004	1	1.416
17	PrevFire10	1.001	1	1.001
18	PrevFire20	1.005	1	1.003
19	PrevFire30	1.004	1	1.002
20	PrevFire40	1.004	1	1.002

4.1.3. Scheme C: S20Withcp0.1Nostep

This scheme is created by running the Clogit model based on the variables selected by the contingency table, but without stepwise selection, the results are shown in Table 4-5 , and the VIF is shown in Table 4-6.

Table 4-5 The Clogit result for Scheme C

	Name	coef	exp(coef)	se(coef)	Pr(>|z|)	Sig.
1	EucDis_Road (standardized)	0.507	1.660	1.398	0.71675	
2	EucDis_Water (standardized)	-2.102	0.122	0.632	0.000885	***
3	Water proportion (standardized)	0.458	1.580	0.040	2.00E-16	***
4	Slope (standardized)	0.117	1.124	0.021	1.41E-08	***
5	DEM_10 (standardized)	-12.540	0.000	0.639	2.00E-16	***
6	FBPC1	-0.298	0.742	0.072	3.42E-05	***
7	FBPC2	-0.539	0.583	0.063	2.00E-16	***
8	FBPC3	-0.735	0.480	0.110	2.51E-11	***
9	FBPC4	-1.005	0.366	0.129	5.76E-15	***
10	FBPD12	0.852	2.344	0.073	2.00E-16	***
11	FBPS2	-1.073	0.342	0.520	0.039288	*
12	FBPO1	0.244	1.276	0.068	0.00032	***
13	FBPNF	18.610	1.21E+08	730.600	0.979676	
14	FBPW	0.641	1.899	0.147	1.28E-05	***
15	FBPVNF	18.600	1.19E+08	581.900	0.974509	
16	FBP10_40	0.786	2.194	0.115	8.94E-12	***
17	FBP40_60	0.440	1.552	0.077	9.57E-09	***
18	PrevFire10	1.188	3.279	0.398	0.002835	**
19	PrevFire20	0.464	1.591	0.377	0.217981	
20	PrevFire70Plus	-0.298	0.743	0.094	0.001485	**

Table 4-6 The VIF for Scheme C

	Name	GVIF	df	GVIF^(1/(2*Df))
1	EucDis_Road	1.002	1	1.001

2	EucDis_Water	1.015	1	1.007
3	Water proportion	1.134	1	1.065
4	Slope	1.012	1	1.006
5	Elevation	1.023	1	1.011
6	FBPC1	2.425	1	1.557
7	FBPC2	5.122	1	2.263
8	FBPC3	1.386	1	1.177
9	FBPC4	1.244	1	1.115
10	FBPD12	2.539	1	1.593
11	FBPS2	1.011	1	1.006
12	FBPO1	3.716	1	1.928
13	FBPNF	1.000	1	1.000
14	FBPW	1.324	1	1.151
15	FBPVNF	1.000	1	1.000
16	FBP10_40	1.271	1	1.127
17	FBP40_60	2.005	1	1.416
18	PrevFire10	1.058	1	1.029
19	PrevFire20	1.057	1	1.028
20	PrevFire70Plus	1.114	1	1.056

4.1.4. Scheme D: S18Withcp0.1Withstep

This scheme proceeds with both AIC stepwise selection and contingency table selection. The total number of variables now becomes 18, the most statistically significant among 29 original variables participating in the modelling. The Clogit modelling results are shown in Table 4-7, and the VIF for the variables is less than 3, as shown in Table 4-8

Table 4-7 The Clogit results for the Scheme D

	Name	coef	exp(coef)	se(coef)	Pr(>|z|)	Sig
1	EucDis_Water (standardized)	-2.105	0.122	0.632	-3.329	***

2	Water proportion (standardized)	0.457	1.580	0.040	11.466	***
3	Slope (standardized)	0.118	1.125	0.021	5.719	***
4	Elevation (standardized)	-12.510	0.000	0.638	-19.593	***
5	FBPC1	-0.299	0.742	0.072	-4.150	***
6	FBPC2	-0.539	0.583	0.063	-8.573	***
7	FBPC3	-0.734	0.480	0.110	-6.667	***
8	FBPC4	-1.006	0.366	0.129	-7.817	***
9	FBPD12	0.852	2.344	0.073	11.701	***
10	FBPS2	-1.073	0.342	0.520	-2.061	*
11	FBPO1	0.245	1.278	0.068	3.614	***
12	FBPNF	18.630	1.23E+08	730.800	0.025	
13	FBPW	0.641	1.898	0.147	4.360	***
14	FBPVNF	18.620	1.23E+08	581.900	0.032	
15	FBP10_40	0.788	2.198	0.115	6.836	***
16	FBP40_60	0.439	1.552	0.077	5.735	***
17	PrevFire10	1.157	3.181	0.397	2.914	**
18	PrevFire70Plus	-0.325	0.723	0.091	-3.563	***

Table 4-8 The VIF for the Scheme D

	Name	GVIF	df	GVIF^(1/(2*Df))
1	EucDis_Water	1.015	1	1.007
2	Water proportion	1.134	1	1.065
3	Slope	1.010	1	1.005
4	Elevation	1.021	1	1.010
5	FBPC1	2.425	1	1.557
6	FBPC2	5.123	1	2.263
7	FBPC3	1.386	1	1.177
8	FBPC4	1.244	1	1.115
9	FBPD12	2.539	1	1.593
10	FBPS2	1.011	1	1.006

11	FBPO1	3.718	1	1.928
12	FBPNF	1.000	1	1.000
13	FBPW	1.324	1	1.151
14	FBPVNF	1.000	1	1.000
15	FBP10_40	1.270	1	1.127
16	FBP40_60	2.005	1	1.416
17	PrevFire10	1.054	1	1.027
18	PrevFire70Plus	1.059	1	1.029

4.1.5. Comparison summary of the four schemes for V700-109fire

Figure 4-1, Figure 4-2, Figure 4-3, Figure 4-4 provide a visualization of the results of the four schemes, highlighting differences and similarities.

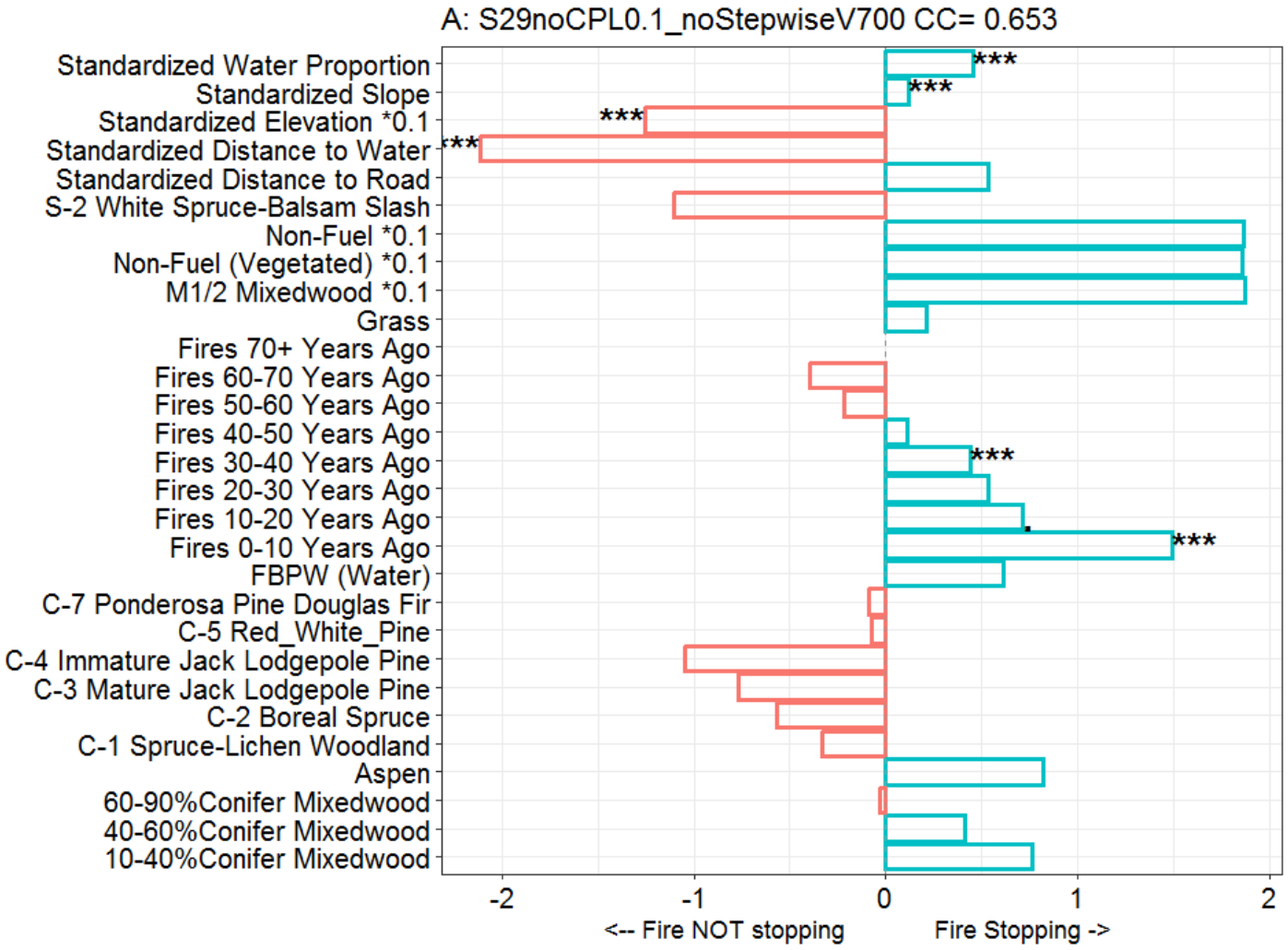


Figure 4-1 Clogit plot for Scheme A. The coefficient of non-fuel, vegetated non-fuel, elevation, and mixedwood has been reduced 10 times for better visuals.

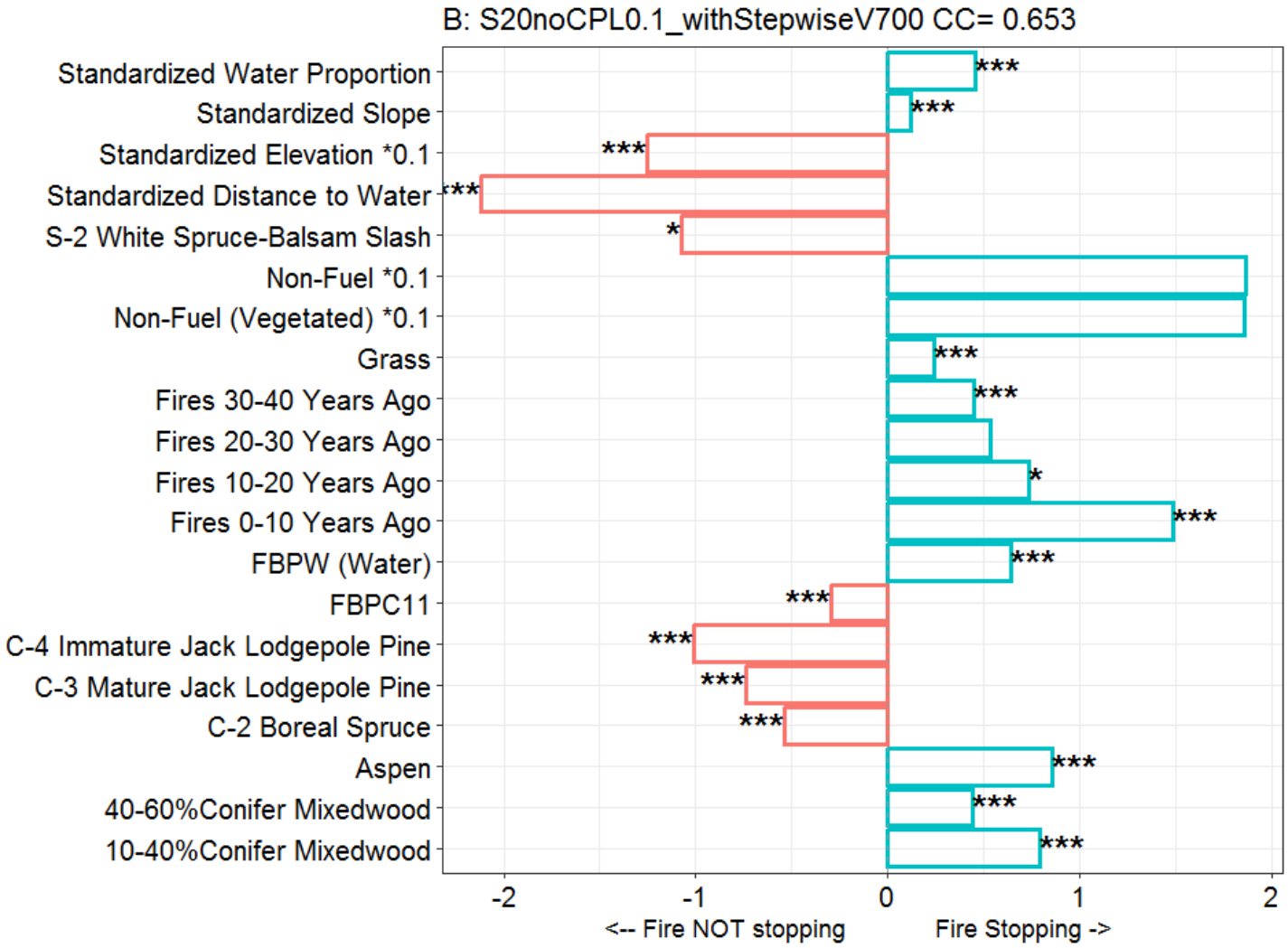


Figure 4-2 Clogit plot for Scheme B. The coefficient of non-fuel, vegetated non-fuel, and elevation has been reduced 10 times for better visuals.

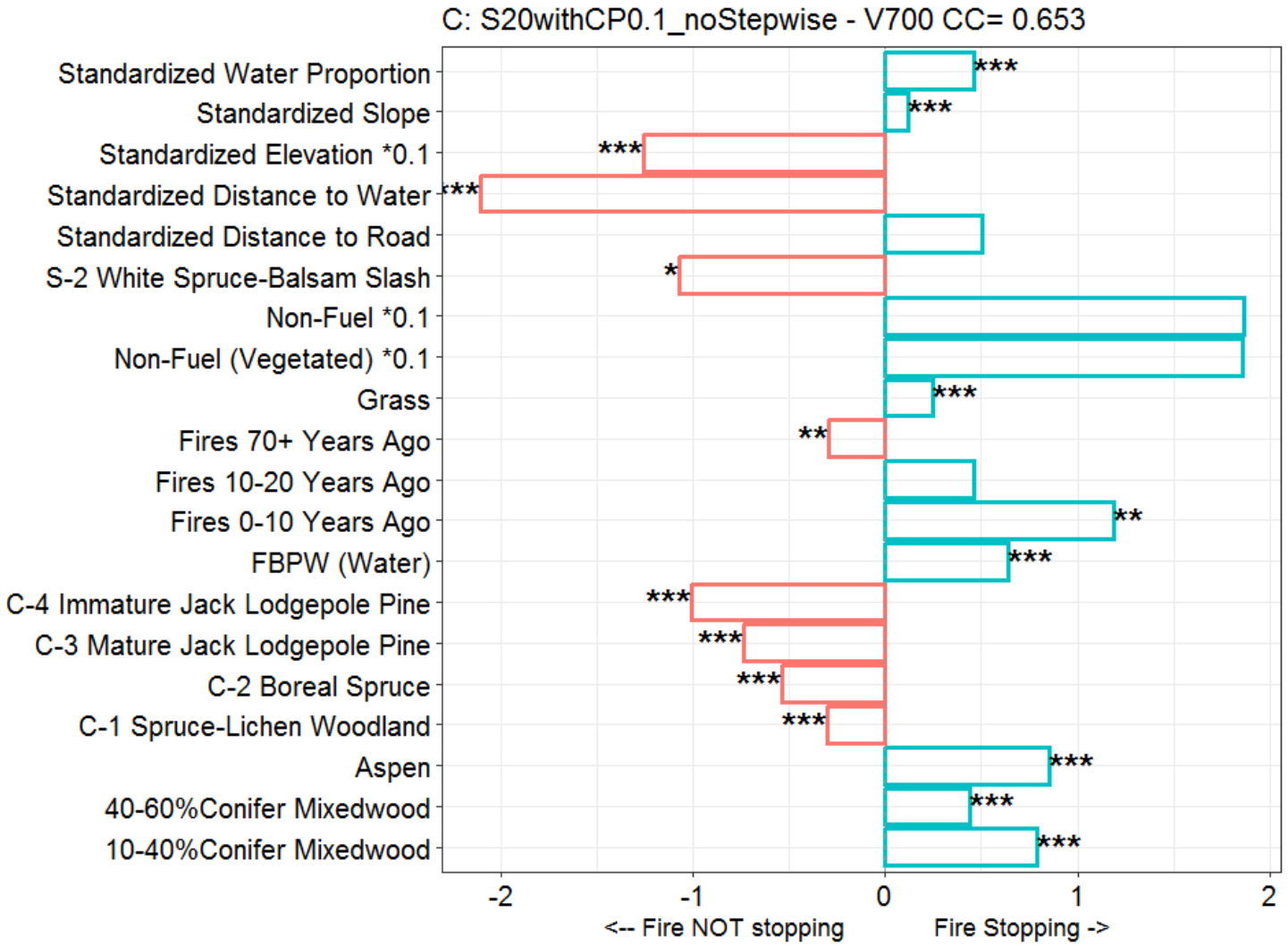


Figure 4-3 Clogit plot for Scheme C. The coefficient of non-fuel, vegetated non-fuel, and elevation has been reduced 10 times for better visuals.

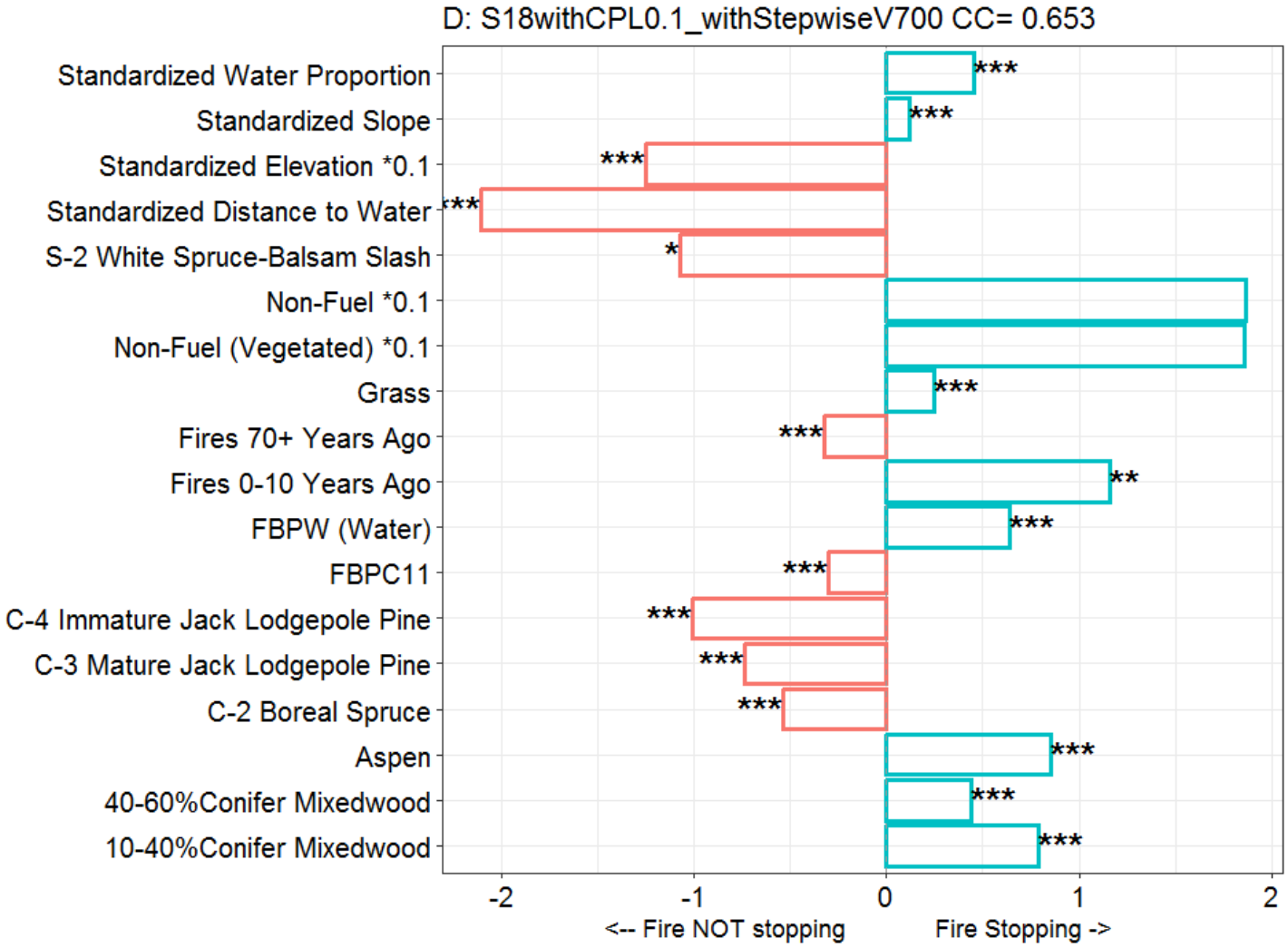


Figure 4-4 Clogit plot for Scheme D. The coefficient of non-fuel, vegetated non-fuel, and elevation has been reduced 10 times for better visuals.

The Clogit visualized result plot reveals that Clogit models can identify the coefficient direction of every variable in all four schemes. Stepwise selection reduces the number of variables used but decreases model performance. The contingency table is useful for selecting variables with consistent CC values across all four schemes. However, among the four schemes of A through D, only four out of five continuous variables were detected in the same significant class ($p<0.001$) in all four schemes. Only two out of the 24 binary variables (NF, non-fuel, and VNF, vegetated non-fuel) were found to be not significant ($p>0.1$) in all four schemes. This shows that further comparisons were needed.

Next, the Random Forest (RF) model was employed to prepare the key factor list for schemes based on RF importance, called the RF schemes. Comparison between each scheme was done to determine the best-performing model for prediction.

4.2.Identification of Two Schemes by RF Modelling

Running RF for all 29 variables obtains the variable importance for all variables as described in the methods section. To make a comparison regarding the number of variables in the four schemes by Clogit modelling, 20 and 18 were used as the threshold for preparing the RF schemes. Based on the importance measure by Mean Decrease Accuracy, variable importance (Figure 4-5) was ranked from high to low based on the two thresholds. The two RF schemes are identified, e.g., Scheme E (RF_comIMP_20) and Scheme F (RF_comIMP_18), consisting of 20 and 18 variables in descending order of importance, respectively.

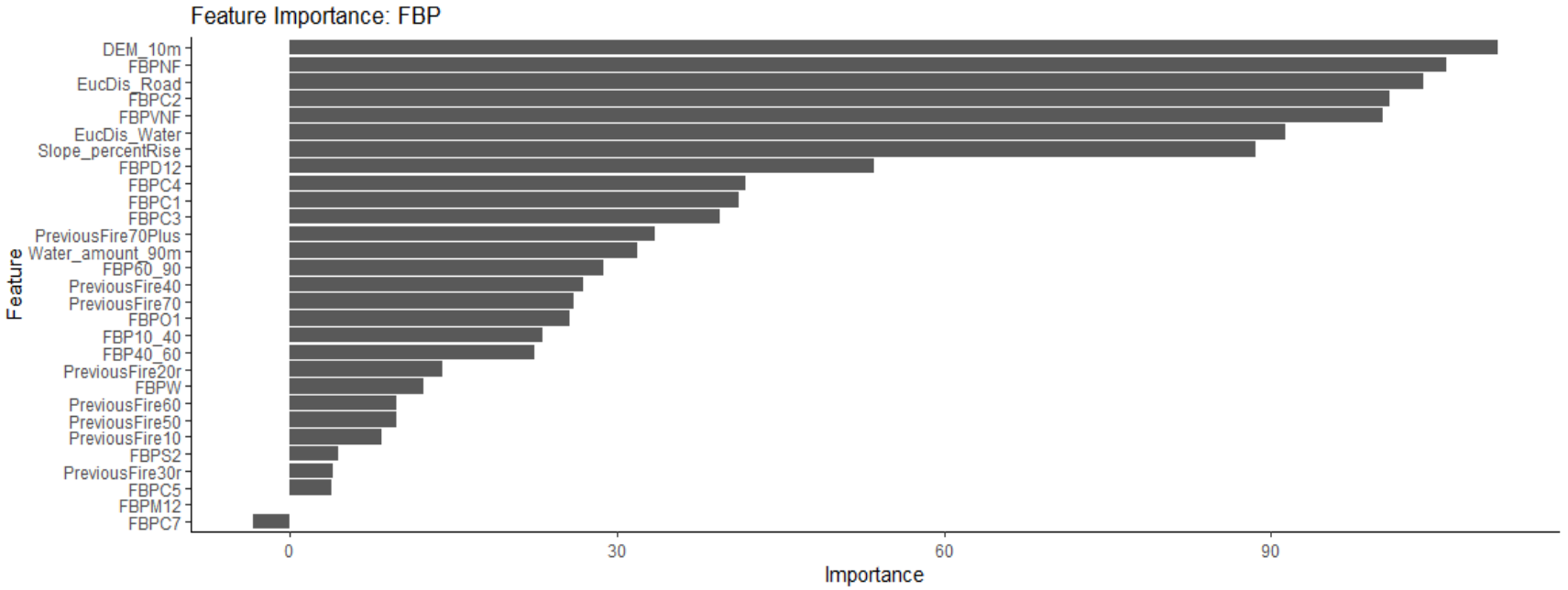


Figure 4-5 The ranked feature importance (Mean Decrease Accuracy) of 29 variables V700, train: test=0.7:0.3. Note: Water_amount_90m refers to water proportion, DEM refers to elevation, the same applies hereinafter.

4.3.Best Key Factor List from the Six Candidate Schemes

For each of the six schemes, their performance in RF is shown in Table 4-9 and Figure 4-6

Table 4-9 The comparison of the RF performance of 6 candidate schemes (among 109 fires, train: test=0.7:0.3, to 2 decimal place)

model name / Schemes	AUC for Train data of 109-V700	AUC for Test data of 109-V700	Accuracy for Test data of 109-V700
Scheme A	0.67	0.71	0.65
Scheme B	0.65	0.66	0.62
Scheme C	0.67	0.70	0.64
Scheme D	0.64	0.65	0.61
Scheme E	0.67	0.70	0.64
Scheme F	0.67	0.71	0.64

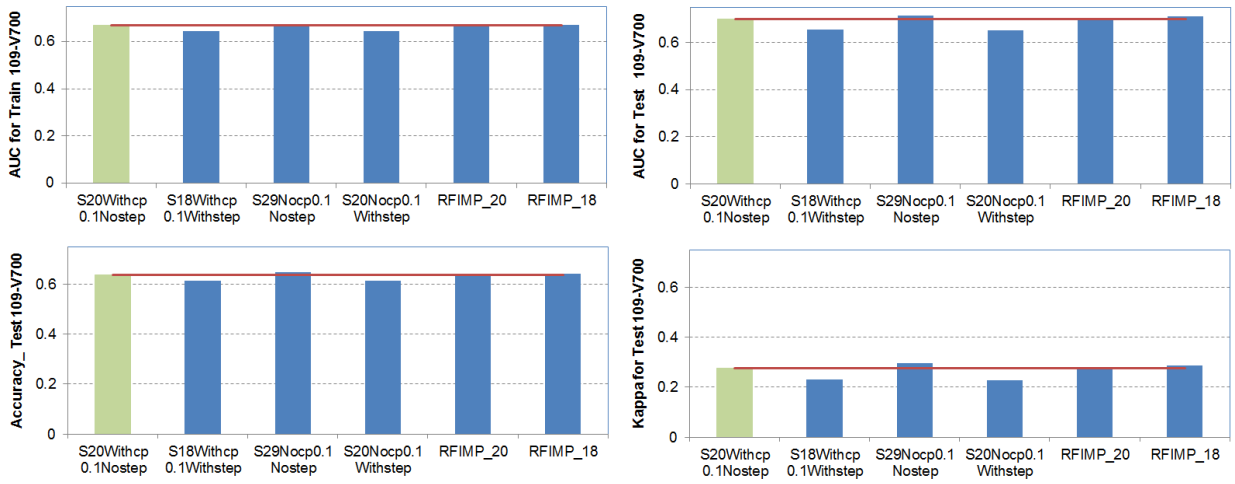


Figure 4-6 The comparison of RF modelling performance for each candidate scheme, green columns highlight the scheme chosen, red line indicates its value relative to others.

As presented in Table 4-9 and Figure 4-6, indices of model performance increased with the number of variables for each scheme. This was true for all the modelling performance indices, including AUC and Accuracy. The model performance of Scheme B (S18Withcp0.1Withstep) was lower than the other schemes with higher number of variables, aligning with studies showing that the model performance is worse with the decrease in the number of variables (Couronné et al., 2018).

However, from further comparison, it was obvious that Scheme C (S20withcp0.1Nostep) was the best scheme. Firstly, the model performance for this scheme was the same as the best-performing model, but the number of variables is much lower, as compared to Scheme A (S29NocpNostep). Secondly, compared to Scheme E (RF_comIMP_20), which had the same number of variables, after running the 20 variables selected by RF importance index in Clogit, from the results, it is clear that some variables among the 20 variables chosen by RF importance were insignificant (e.g. Euclidean distance to roads, W (Water), 40-60% confer mixedwood), as

shown in Figure 4-7. The result from the Scheme E when compared to Scheme C, covers less of the FBP fuel variables, and focuses more on the previous fire variables.

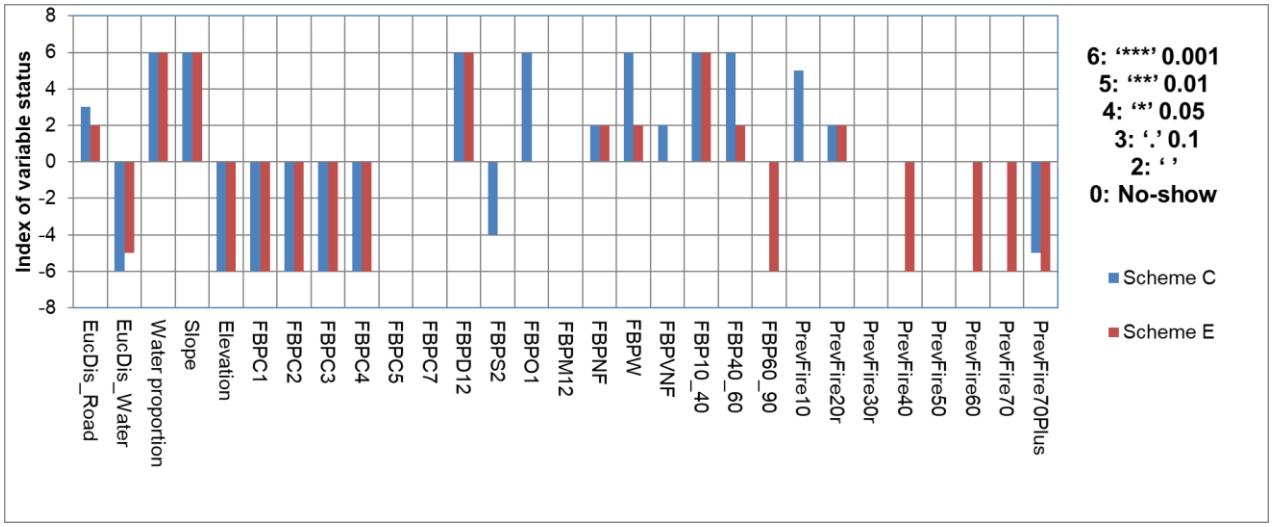


Figure 4-7 The comparison of the variable status between scheme C (S20Withcp0.1Nostep) and scheme E (RF_comIMP_20)

Although stepwise in Clogit allowed a reduction of variable number from 29 to 20, the model performance of Scheme B shows that if stepwise is applied, AUC was lower than that without stepwise as shown in Scheme C, shown in Table 4-9, indicating the pivotal role of the contingency table over stepwise in the modelling. Therefore, it was decided that Scheme C and its 20 corresponding variables are the most effective in predicting fire perimeters. The detailed interpretation for each key underlying influential factor was analyzed below, and then their utilization in fire perimeter prediction were explored.

4.4. Interpretation of the Identified Key Factors

4.4.1. Key Factor Interpretation Based on Clogit

The interpretation of the key underlying factors influencing fire cessation, based on Clogit, can be summarized in three parts. Firstly, the significance was determined by the p value in each Clogit table, representing the Wald statistic value. It corresponded to the ratio of each regression

coefficient (marked “coefficient” in Table and the y-axis of Clogit plot, or b_i, in the hazard function (Eq. 2-8)) to its standard error (se(coef)). The Wald statistic evaluates whether b_i the coefficient of a given variable is statistically significantly different from zero. As shown in Table 4-5, 13 out of 20 variables showed a high statistically significant difference between the coefficient of each of these three variables and zero. The null hypothesis is that the variable's coefficient is not different from zero and is rejected for these covariates, highlighting their significance in the analysis.

Secondly, the variable’s acting directional contribution to the event is determined by the sign of the regression coefficients (coef). A positive coefficient indicates an elevated hazard ratio (HR), represented as $\exp(b_i)$, suggesting a higher risk of event and a less favourable prognosis. In Table 4-5, variables such as Water proportion, Slope, D-1/D-2 (Aspen), O-1 (grassland), NF (non-fuel), W (Water), VNF (vegetated non-fuel), 10-40% confer mixedwood, 40-60% confer mixedwood, previous fire from 0 to 10 and 10 to 20 years ago exhibit positive coefficients, signifying a positive contribution to the event. In essence, for binary variables, an increase in the presence of these variables in a location and for continuous variables, an increase in their values would lead the Clogit hazard function to conclude that these variables enhance the likelihood of an area forming a fire boundary. Conversely, variables with a negative coefficient suggest a reduced likelihood of an area forming a fire boundary. In the case of binary variables, the presence of such variables, or an increase in their values for continuous variables, was associated with a lower probability of an area forming a fire boundary. The interpretation of the standardized continuous variables is different from the original, as the standardized coefficient of the standardized variable, means an increase in one standard deviation (SD) of the standardized variable, will result in an expected change in the likelihood of the area forming a fire boundary. For example, Euclidean distance to

water has a standardized coefficient of -2.102, and an exp(coef) of 0.122, meaning for each increase in one SD of Euclidean distance to water, the likelihood of forming a fire boundary would decrease by 88%.

The third component of the results involves interpreting the effect size of the covariant, which was shown by the HR (exp (b_i) coefficient). For example, disregarding the signs of the coefficient we mentioned in the previous paragraph, the variable C-3 (Mature Jack or Lodgepole Pine), encoding the presence or absence of C-3, demonstrated an HR of 0.48 (Table 4-5). This signifies that locations with C-3 have 0.48 (folds) times the risk of an event. It is the hazard ratio for the group taking a higher value relative to the group taking a lower value. In other words, because the coefficient sign for variable C-3 is negative (-7.35E-01), it can lead to the conclusion that the risk of an event (fire cessation) without C-3 is 0.48 times higher than the risk with C-3.

Taking water as another example, the variable water proportion is encoded as the quantity of water, with a higher value meaning a higher proportion of water. Its HR represents the ratio of hazards for the group with more water versus that with less water. The HR for Water proportion = exp(0.165)= 1.18, with a positive sign for coefficient, indicating that areas with more water have 1.18 times (fold) of risk of an event (fire cessation) than areas with less water. These HR interpretations provide insights into the impact of specific covariates on the likelihood of fire cessation. It is important to acknowledge that these conclusions are based on the assumption that all other covariates remained constant, as stipulated by the coxph model or Clogit model, which considers the interaction between the covariates; all the above conclusions are under the condition that all other covariates remain unchanged.

Figure 4-3 visually emphasize the significance of water proportion, Slope, D-1/D-2 (Aspen), O-1 (grassland), NF (non-fuel), W (Water), VNF (vegetated non-fuel), 10-40% confer mixedwood, 40-60% confer mixedwood, previous fire from 0 to 10, and 10 to 20 years ago as the significant fire-resistant factors. Notably, conifer content up to 60% in mixedwood remains a significant fire-stopping factor. With more conifer content beyond 60%, the probability of fire cessation decreases from 2.19 to 1.55 folds.

Furthermore, the analysis reveals some intriguing dynamics. Although previous fire from 0 to 10, and 10 to 20 years play a significant role in fire cessation, fires older than 70 years exhibit an opposite effect. Grass and aspen are fire-stopping elements, and all pines (C-1 (Spruce-Lichen Woodland) to C-4 (Immature Jack or Lodgepole Pine)) are identified as fire-prone.

NF (non-fuel) and VNF (vegetated non-fuel) showed highest exp(coefficient) values but with p value much higher than 0.1. This is expected as the NVB-clean step was applied which guarantees that these two fuel types only occur outside the fire perimeter without exceptions. For a comparison, Clogit for V700 for 109 fires without NVB-clean has been run, as shown in Table 4-10 and Table 4-11. It was seen that many Clogit- coefficients turn out to be physically unreasonable, with CC now reduced from 0.653 to 0.639 (when using NVB clean), which means NF and VNF play an important role. After applying stepwise, it was still physically unreasonable, with CC being 0.641. This is the reason for cleaning the NVB before the Clogit modelling.

Table 4-10 Clogit-contingency results of non-clean of VNF-NF-Barren (V700)-109 fires

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
EucDis_Road (standardized)	1.137	3.118	1.283	0.887	0.37529	
EucDis_Water (standardized)	-2.372	0.093	0.571	-4.152	3.30E-05	***

Water proportion (standardized)	0.494	1.638	0.040	12.497	2.00E-16	***
Slope (standardized)	0.080	1.083	0.019	4.273	1.93E-05	***
Elevation (standardized)	-12.640	0.000	0.615	-20.549	2.00E-16	***
FBPC1	0.055	1.057	0.213	0.260	0.794904	
FBPC2	-0.234	0.791	0.210	-1.116	0.264257	
FBPC3	-0.399	0.671	0.226	-1.769	0.076899	
FBPC4	-0.661	0.517	0.238	-2.776	0.005501	
FBPC5	1.155	3.175	0.213	5.425	5.79E-08	
FBPD12	0.541	1.718	0.211	2.566	0.010283	*
FBPS2	0.377	1.457	0.224	1.684	0.092118	
FBPO1	0.887	2.427	0.247	3.586	0.000335	
FBPNF	0.932	2.539	0.227	4.109	3.98E-05	
FBPW	1.172	3.228	0.230	5.106	3.30E-07	.
FBPVNF	0.777	2.175	0.215	3.622	0.000293	.
FBP10_40	0.229	1.258	0.217	1.058	0.289946	*
FBP40_60	1.385	3.993	0.207	6.702	2.05E-11	
FBP60_90	0.594	1.811	0.224	2.651	0.008026	
PrevFire10	0.595	1.812	0.166	3.582	0.000341	***
PrevFire20	0.086	1.089	0.137	0.624	0.532337	**
PrevFire30	1.137	3.118	1.283	0.887	0.37529	
PrevFire40	-2.372	0.093	0.571	-4.152	3.30E-05	***
PrevFire50	0.494	1.638	0.040	12.497	2.00E-16	
PrevFire60	0.080	1.083	0.019	4.273	1.93E-05	
PrevFire70	-12.640	0.000	0.615	-20.549	2.00E-16	
PrevFire70-Plus	0.055	1.057	0.213	0.260	0.794904	

Table 4-11 Clogit stepwise results of non-clean of VNF-NF-Barren (V700)-109 fires

	coef	exp(coef)	se(coef)	z	Pr(>|z|)

EucDis_Water (standardized)	-2.377	0.093	0.571	-4.161	3.17E-05	***
Water proportion (standardized)	0.494	1.638	0.039	12.498	2.00E-16	***
Slope (standardized)	0.080	1.083	0.019	4.265	2.00E-05	***
Elevation (standardized)	-12.640	0.000	0.615	-20.565	2.00E-16	***
FBPC1	-0.287	0.750	0.047	-6.119	9.41E-10	***
FBPC2	-0.450	0.638	0.097	-4.629	3.67E-06	.
FBPC4	-0.714	0.490	0.119	-5.983	2.19E-09	.
FBPC5	1.102	3.011	0.061	18.161	2.00E-16	*
FBPD12	0.488	1.629	0.052	9.453	2.00E-16	***
FBPO1	0.323	1.381	0.090	3.578	0.000346	***
FBPNF	0.833	2.301	0.138	6.061	1.35E-09	***
FBPW	0.878	2.406	0.096	9.105	2.00E-16	***
FBPVNF	1.117	3.055	0.108	10.359	2.00E-16	***
FBP10_40	0.725	2.064	0.067	10.810	2.00E-16	***
FBP40_60	0.176	1.193	0.072	2.454	0.014109	***
FBP60_90	1.298	3.662	0.155	8.359	2.00E-16	***
PrevFire10	0.512	1.668	0.179	2.852	0.004346	***
PrevFire20	0.517	1.677	0.110	4.719	2.37E-06	**
PrevFire40	-2.377	0.093	0.571	-4.161	3.17E-05	***

4.4.2. Influence of Key Factors by RF Importance

In this section, I explore the influence of each variable as detected by RF importance. Since the best scheme is Scheme C, the importance plot for this scheme is presented in Figure 4-8. In the figure, two measurements of importance were provided for each variable. The significance (p-value) is shown in Figure 4-9, Figure 4-10, Figure 4-11, and Figure 4-12.

The Mean Decrease Accuracy indicates how much the accuracy decreases when the variable was excluded, calculated separately for the two outcome classes (0 for burned and 1 for unburned perimeters). On the other hand, Mean Decrease Gini follows a different logic. As Random Forest models utilize a random selection of variables to create a decision tree, each variable acts as a notch on the decision tree, influencing the classification into one of the two classes. In this case, the variable would be used to determine burned and unburned. The Gini value is at its highest before any notch begins and decreases with each notch the decision tree progresses. The Mean Decreased Gini for a single variable is calculated as the mean of the decrease in Gini of this single variable across all the decision trees combined. The numeric value of Gini is irrelevant, and the relative value holds significance.

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
C-2 Boreal Spruce	37.4	39.0	105.2	391.5
Non-fuel	37.0	78.4	102.7	156.9
Vegetated non-fuel	42.3	64.3	96.9	248.9
Elevation	72.3	4.7	95.1	3661.1
EucDis_Road	58.0	30.3	90.4	3550.5
Slope	55.2	27.5	80.8	3567.4
EucDis_Water	56.0	37.3	74.2	3399.8
D-2 Aspen	3.8	42.0	67.0	91.4
C-4 Immature Jack/lodgepole pine	25.0	1.0	41.9	35.0
Fires 70+ years ago	35.0	1.9	41.7	163.3
C-1 Spruce-lichen woodland	26.7	-2.6	39.9	54.7
C-3 Mature Jack/Lodgepole pine	31.1	2.0	38.9	39.7
O-1 Grassland	2.6	24.1	37.8	74.3
10-40% conifer mixedwood	-8.0	35.8	34.8	30.0
Water proportion	16.4	22.7	31.3	220.1
40-60% conifer mixedwood	-10.8	27.6	31.3	51.6
FBPW (water)	9.7	10.9	14.3	42.9
Fires 10-20 years ago	10.9	5.2	13.5	13.6
Fires 0-10 years ago	14.7	-8.3	9.0	25.9
S-2 White Spruce-Balsam slash	9.2	-9.9	1.5	5.7

Figure 4-8 The comparison between the four importance indexes for each key factor, with the 2nd to 4th columns representing the index of mean decrease accuracy for burn, unburn and both, the 5th column being mean decrease gini

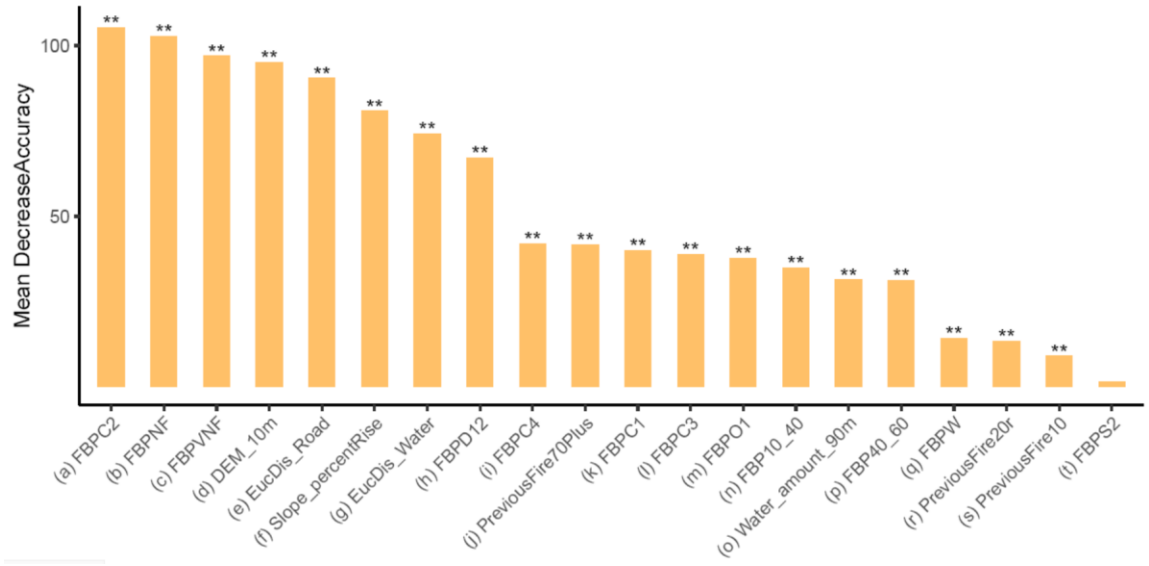


Figure 4-9 The p-value and importance in Mean Decrease Accuracy for scheme S20Withcp0.1Nostep

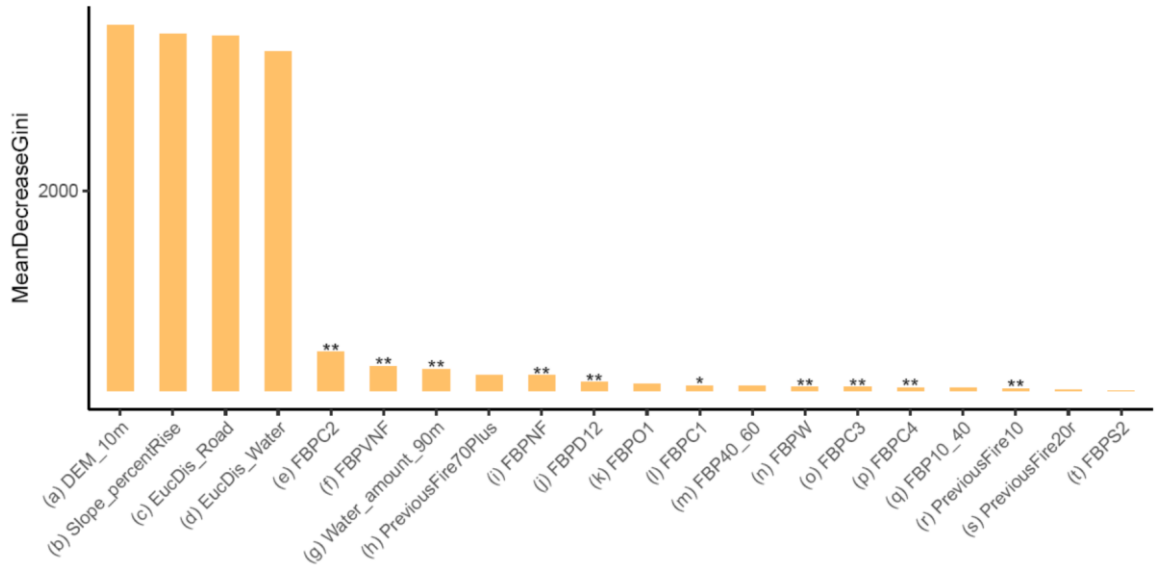


Figure 4-10 The p-value and importance in Mean Decrease Gini for scheme S20Withcp0.1Nostep

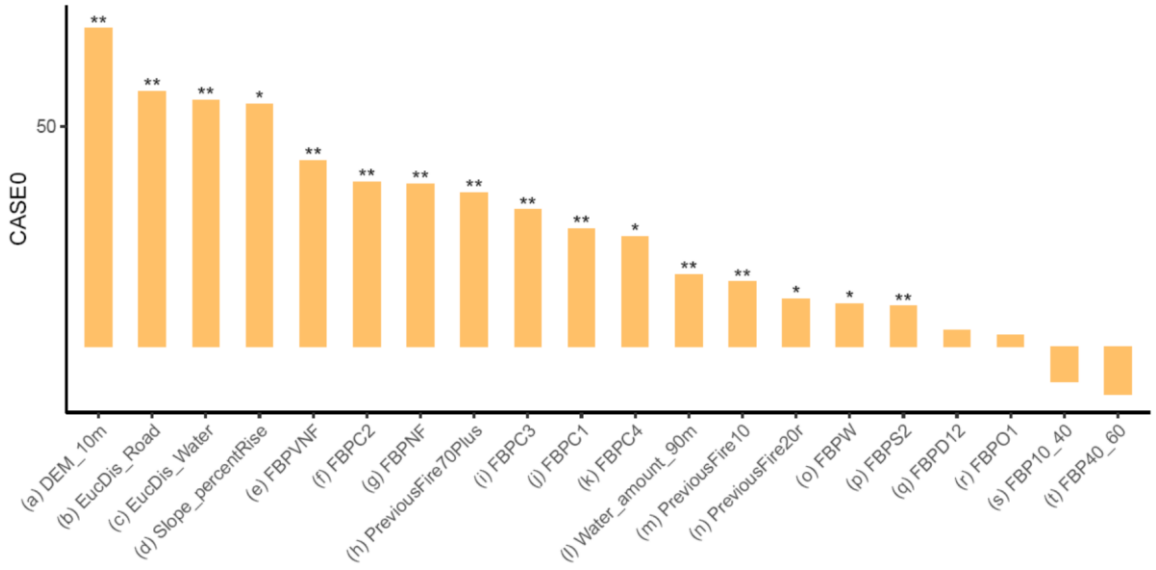


Figure 4-11 The p-value and importance in Mean Decrease Accuracy for burn for scheme S20Withcp0.1Nostep

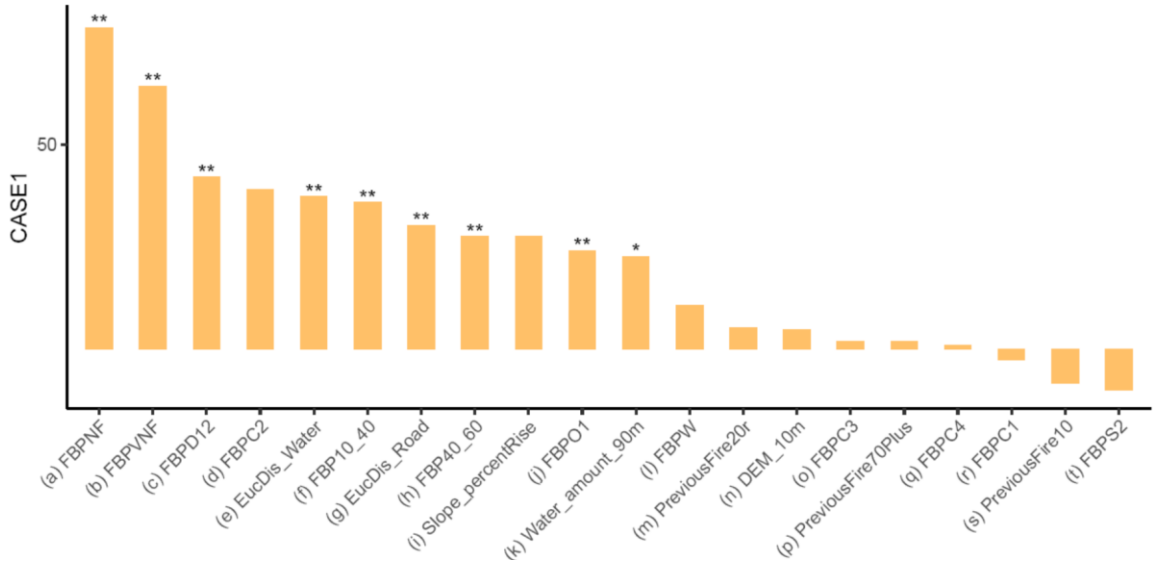


Figure 4-12 The p-value and importance in Mean Decrease Accuracy for unburn for scheme S20Withcp0.1Nostep

4.4.2.1. Influence of Euclidean Distances to Road and Water Respectively

The anthropogenic variable, namely the Euclidean distance to roads and water-related variables, such as Euclidean distance to water bodies, play an important role in predicting fire perimeters. The Euclidean distance to roads almost doubles the likelihood of predicting a burned area compared to an unburned one. Additionally, the predictive power of Euclidean distance to water bodies was lower, showing 1.5 times less the predictive power for unburned areas relative to Euclidean distance to roads. Both variables demonstrate statistical significance across all three categories of Mean Decrease Accuracy importance indexes (see Figure 4-9, Figure 4-10, Figure 4-11, Figure 4-12) with P -values less than 0.1.

From the Gini index, the variable Euclidean distance to roads was around 5% more important than Euclidean distance to water. However, both variables displayed no statistical significance in the Mean Decreased Gini. Despite this, the results across all four categories suggest the high importance of anthropogenic and water-related variables in forming fire boundaries in the boreal zone in Alberta. However, there is some discrepancy with Clogit's findings. The significance of Mean Decrease Accuracy agrees with Clogit, whereas Gini does not.

4.4.2.2. Influence of Topography and Amount of Surrounding Water

Among the topographic variables - slope (Percentage rise), digital elevation, and water proportion (90 m) – based on Mean Decrease Accuracy, digital elevation emerges as the absolute highest, followed by the slope (Percentage rise) with 70% of its predictive power. At the same time, water proportion has the lowest Mean Decrease Accuracy within the three variables, representing only 12% of the predictive power of digital elevation. All three variables demonstrate statistical significance ($P < 0.01$) under each variable's Mean Decrease Accuracy category.

Digital elevation is much more important when predicting the burned area than the unburned. In contrast, water proportion is 1/3 times more important for predicting unburned perimeters than burned ones, which is expected. Slope, in this case, is twice as important for predicting burned areas than for unburned. Digital elevation and water proportion are more statistically significant ($P < 0.01$) when predicting burned areas than slope ($P < 0.1$). On the other hand, all three variables were less statistically significant when predicting unburned areas ($P< 0.1$ or $P > 0.1$). When examining the categories of Mean Decrease Gini, "digital elevation" and "slope" ranked as the absolute highest among any variables in the entire model, though they are insignificant. Water proportion was only 3.9% as important as digital elevation but is more statistically significant ($P < 0.01$).

4.4.2.3. Influence of Fuels

The selected variables related to fuels included the FBP System fuel types C-1(Spruce–Lichen Woodland), C-2 (Boreal spruce), C-3 (Mature Jack or Lodgepole Pine), C-4 (Immature Jack or Lodgepole Pine), D-1/D-2 (Aspen), O-1 (grassland), S-2 (White Spruce - Balsam Slash), NF (non-fuel), VNF (Vegetated non-fuel), W (water), and additional specific mixedwood ratios: M10-40 (10 – 40% conifer mixedwood) and M40 – 60 (40 – 60% conifer mixedwood).

When considering Mean Decrease Accuracy for variables in both classes, C-2, NF and VNF had the highest Mean Decrease Accuracy, with D-1/D-2 showing about 60% of Mean Decrease Accuracy of C-2. In contrast, all other variables were far less impactful. When predicting the burned areas, all variables except O-1 (grassland) and D-1/D-2 ((Aspen)) were not statistically significant. C-1(Spruce–Lichen Woodland), C-3 (Mature Jack or Lodgepole Pine), C-4 (Immature Jack or Lodgepole Pine), and S-2 (White Spruce - Balsam Slash) showed a higher predictive power for burned areas than unburned, while others exhibited the opposite pattern. For predicting the

unburned fire perimeters, NF (non-fuel), VNF (vegetated non-fuel), D-1/D-2 (Aspen), 10-40% confer mixedwood, 40-60% confer mixedwood, and O-1 (grassland) were statistically significant ($P<0.01$), while others were insignificant. Looking at the Mean Decrease Gini for all variables, all the fuel variables were in the low-importance category, but all remained significant (at least $P<0.1$).

4.4.2.4. Influence of Previously Burned Areas

Variables representing previously burned areas ranged from within 10 to 20 years of the sampled point to previously burned areas from more than 70 years prior were focused on. Regarding the Mean Decrease Accuracy, Previous fire from 0 to 10 years and 10 to 20 years were less important than Previousfire70Plus for all four importance indices. All three previous fire variables were more powerful in predicting burned areas than predicting unburned, and they were all significant in Mean Decrease Accuracy ($P<0.01$). Only previousFire10 was significant for Gini, and the other three were not. They all were significant in predicting burned areas ($p<0.01$). However, they were not significant for predicting the unburned areas.

4.4.2.5. Summary of RF detection

Mean Decrease Accuracy was more in agreement with Clogit in variable selection. It was more powerful for predicting unburned than burned. However, more variables were less significant in predicting unburned states than in predicting burned states. Overall, the pattern of the significance of each variable in RF was not as clear as those in Clogit, but RF Exceled at prediction.

4.5. Fire Perimeter Prediction Using V1000

Having identified S20withcp0.1Nostep as the most effective scheme through the model comparison, the next step involved using its corresponding variables to run a prediction model

with RF. This final model, encompassing sample points from all 109 fires (noted as V1000), emerged as a potent tool for predicting fire perimeter in the Boreal region in Alberta. To assess the predictive capabilities of the model, the area immediately surrounding four new fires were selected from the Alberta Boreal Region – HWF120_2012, HWF137_2018, SWF107_2017 and SWF008_2011. These areas were chosen to represent a spectrum of landscapes with diverse landcover and historical fire incidents that could be compared to the predictive results.

The testing phase involved spatially sampling points at a resolution of 100 m. covering each of the four fires completely in the shapes of rectangles, and then having every point run in the RF model. Figure 4-13 to Figure 4-16 showed the fire perimeter prediction for the four selected regions. The grey scale colour gradient on these maps delineated varying levels of probabilities of fire stopping, providing a detailed understanding of how underlying factors contribute to the prediction outcomes.

The prediction was evaluated using the Area Under the Curve (AUC). Values in Table 4-12 served as a quantitative measure of the model's accuracy. Importantly, the AUC values for training (V1000 data) and testing data (newly sampled) affirmed the model's efficacy in capturing the key factors influencing fire perimeter.

Table 4-12 The AUC in four Regions predicted by the final model S20Withcp0.1Nostep

Fire name	WD	WS km/h	AUC-Tain	AUC-Test	Accuracy-Test
HWF120_2012	360	10	0.71	0.65	0.60
HWF137_2018	**0**	**0**	0.71	0.70	0.66
SWF107_2017	315	24	0.71	0.58	0.50
SWF088_2011	360	22	0.71	0.57	0.56

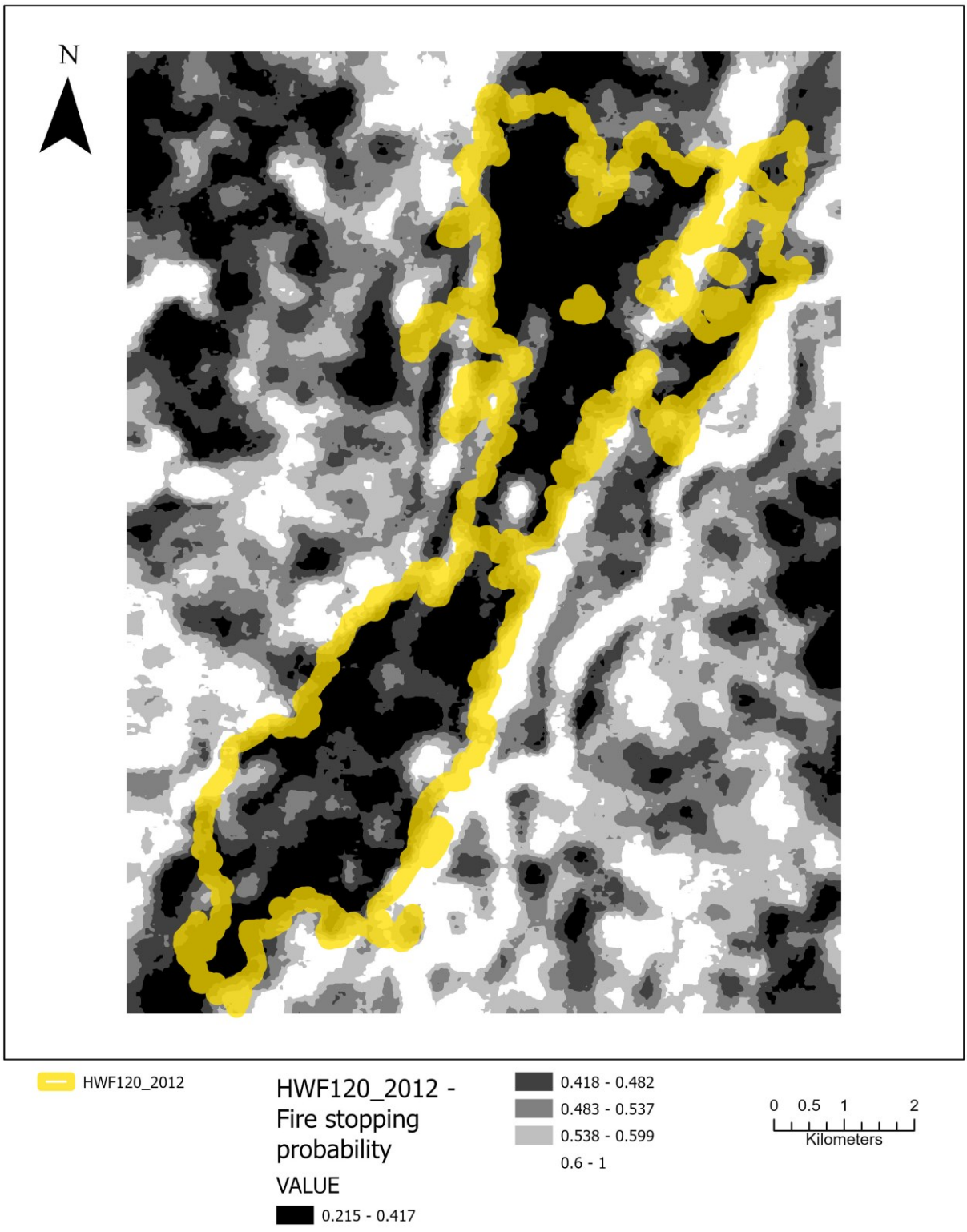


Figure 4-13 Fire perimeter prediction for fire HWF120-2012

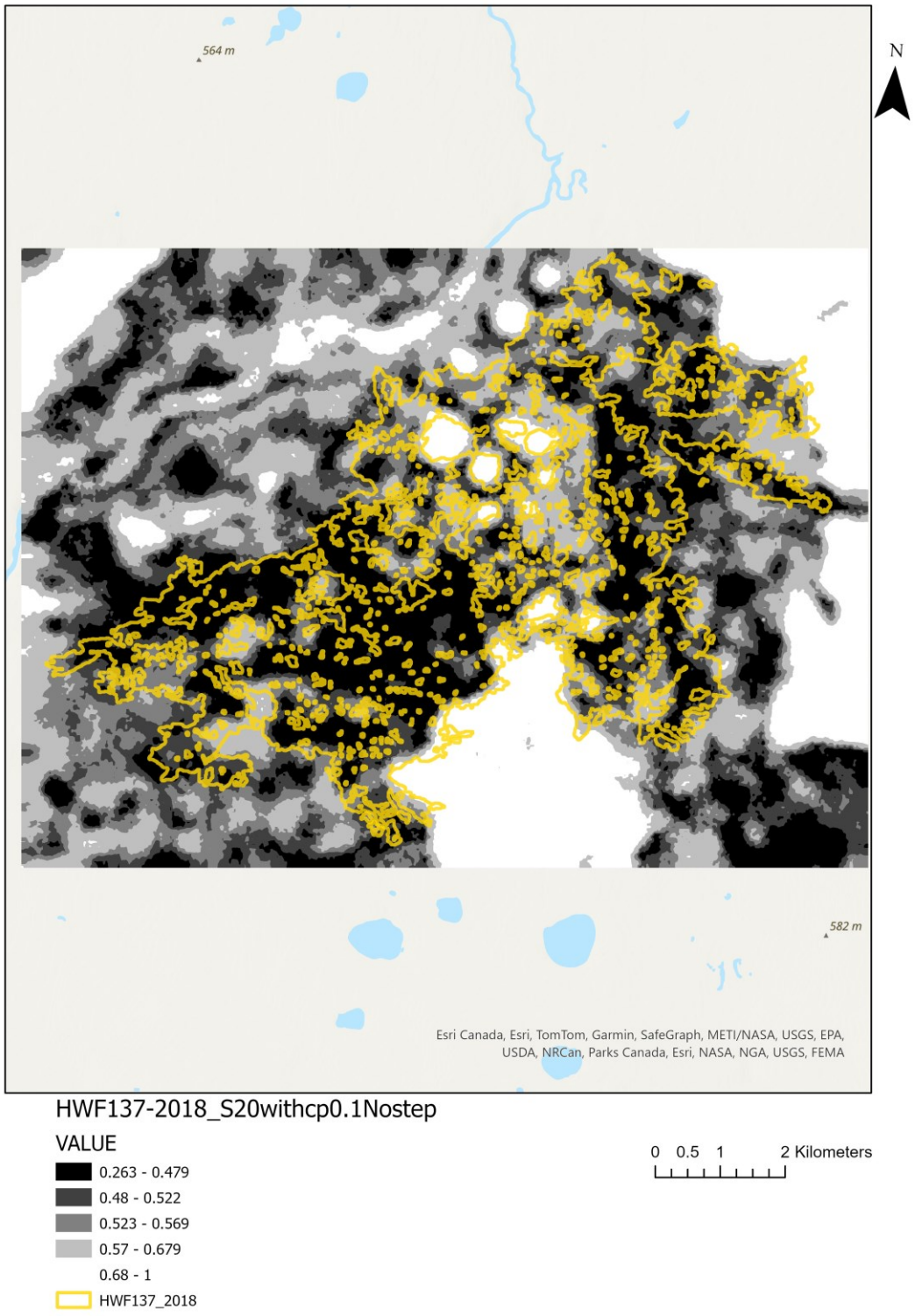


Figure 4-14 Fire perimeter prediction for fire HWF137-2018

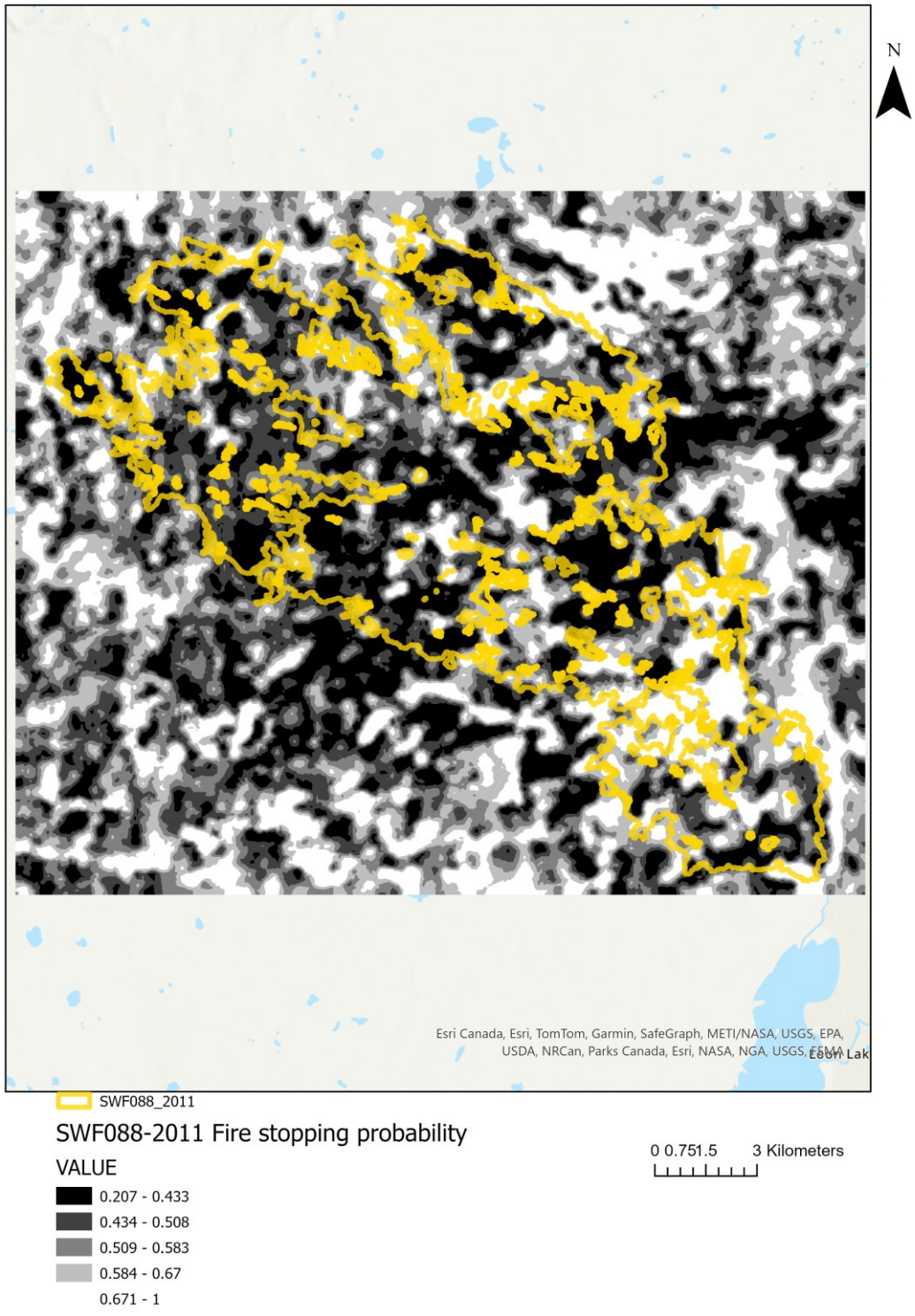


Figure 4-15 Fire perimeter prediction for fire SWF088-2011

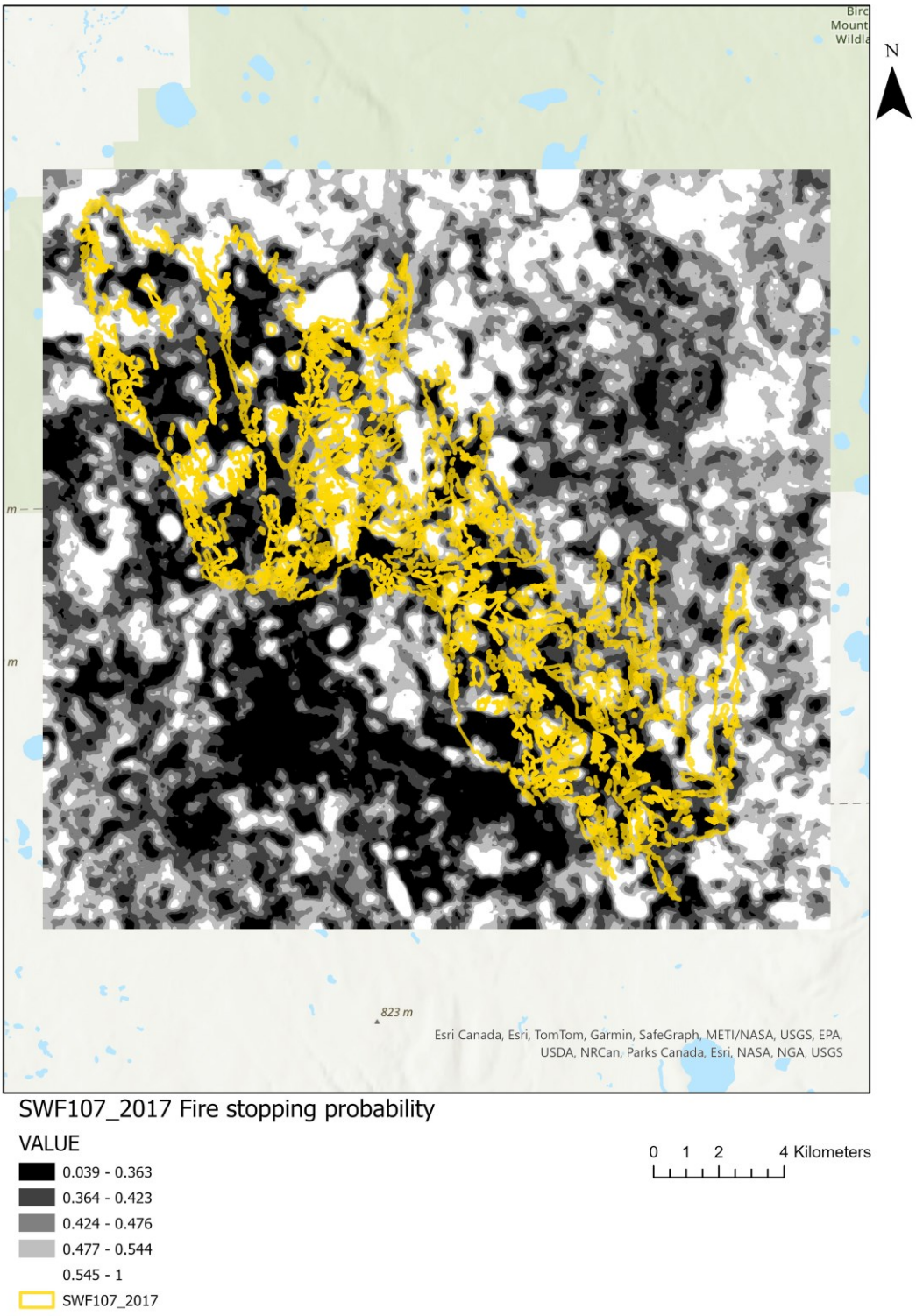


Figure 4-16 Fire perimeter prediction for fire SWF107-2017

For fire HWF120-2012, the prediction displayed two distinct fire-stopping lines associated with roads and rivers, with support from fuel distribution. This spatial distribution corresponded to HWF120-2012, showcasing the model's ability to capture the nuance relationship between variables and fire probabilities. For fire HWF137-2018, the south and the east boundaries of the fire had low fire probability; this was validated as the area contains rivers and lakes. Fire SWF088-2011 is characterised by area with a complex patchwork of both high and low fire-stopping probabilities. The distribution of fuels in this region plays an important fire-stopping role. However, the model's predictive ability was weak for the fire SWF088-2011, which suggests weather conditions may have been a dominant factor. A review of weather records revealed relatively high wind speeds, which could potentially explain the weak predictive results.

Lastly, fire SWF107-2017 demonstrated similar prediction results to SWF088-2011. The wind speed for SWF107-2017 data was also very high, at 24 km/h. The modelling process again had relatively lower modelling performance. However, the prediction was well-validated in many specific areas, such as the northern boundary, where the fire perimeter matched with the edges of D-1/D-2 (i.e., low flammability deciduous fuels) and water bodies, where the fire perimeter corresponded to the many different kinds of land cover, as shown in Figure 4-17.

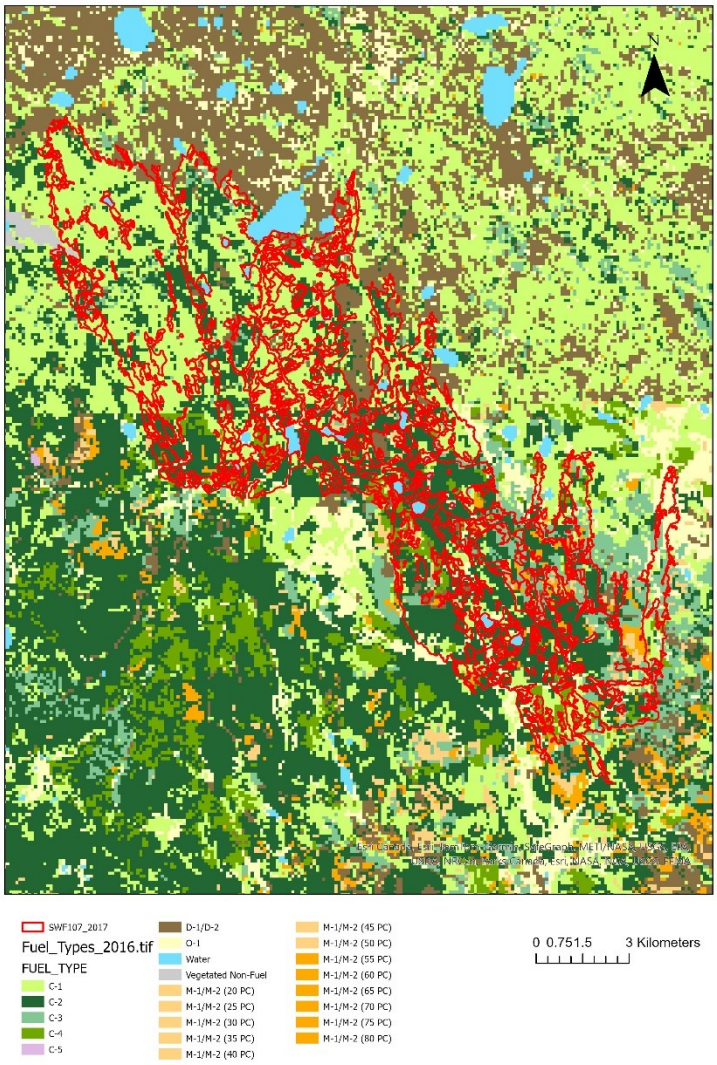


Figure 4-17 SWF107-2017 with FBP fuel map of 2016

This comprehensive evaluation shed light on the final model's practical application and limitations. The influence of weather conditions, especially high wind speeds, notably challenged modelling performance. Future refinements may involve integrating real-time weather data to enhance the model's accuracy in regions prone to weather-driven fire dynamics.

4.6. Discussions

4.6.1. Comparison of the Modelling Performance

After applying matched case-control (MCC) conditional logistic regression (Clogit) on the chosen scheme S20Withcp0.1Nostep, this Clogit model had a concordance coefficient index (CC) of 0.653, which was moderate at best. Prior studies that applied the MCC Clogit (Macauley et al., 2022) method reported a higher model CC of 0.79. Lower concordance reported in this thesis could be due to inclusion of over one hundred fires. A simulation experiment was conducted for each of the 109 fires individually, using the developed algorithm. The CC for an individual fire reached as high as 0.90 or even higher in a few fires (e.g. SWF175_2015). However, most of the fires has a CC between 0.6 - 0.8, with a few falling below 0.6 (e.g. MWF028_2009, PWF060_2012). For fires with exceptionally high CC, these cases were always correlated with extremely low sample size (<50). Due to the limited sample size, the statistical power was insufficient to detect the significant effects for any of the covariates. This also explained why the CC for most individual fires tended to be higher than the CC for our model with multiple fires. Therefore, for the model of 109 fires, the overall CC of 0.653 is reasonable on a high-dimensional dataset. It is worth noting that prior studies that reported higher concordance also considered weather conditions. Only landscape factors were considered in this thesis.

Couronné et al. (2018) showed that RF outperforms Clogit based on large-scale data. The study in this thesis of 109 fires also supported this conclusion, in which the highest AUC in the RF model was 0.67. In contrast, the highest CC from MCC Clogit was 0.65. However, RF could not identify the direction of explanatory variables. Thus, instead of choosing only Clogit or RF, this thesis used both by taking advantage of the merits of each to achieve the goal. A previous study by Shomal Zadeh et al. (2020), who proposed a method of Matched Forests, to a certain degree, supports this idea as well.

4.6.2. Fuels

Among the influence of various conifer fuel types on fire cessation, based on the result from the Clogit model (Table 4-5), the FBP C-4 and S-2 (White Spruce - Balsam Slash) fuel types were identified as the most fire-prone fuel, with C-4 showing a higher level of statistical significance ($p < 0.001$) compared with S-2 ($p < 0.05$), despite both having similar exp(coef.) values. Subsequently, the fuel types FBP C-3, FBP C-2, and FBP C-1 had decreasing susceptibility to fire. However, all these fuel types shared the same high level of statistical significance as FBP C-4 ($p < 0.001$). C-2 fuel had the highest importance measures compared with all other fuel variables (Figure 4-8). This result was unsurprising, given that boreal conifers were considered highly fire-prone (Forestry Canada Fire Danger Group, 1992). C-3, C-1, and C-4 all had relatively similar Mean Decrease Accuracy, but C-2 alone had more than double the Mean Decrease Accuracy importance measure. This meant C-2 was much more influential and important for the RF model's accuracy in prediction. VNF and NF had the highest variable coefficients in the result. Still, the two variables were plagued with data inconsistencies and did not have a significant P value. This will be further discussed in Section 4.6.7. D-1/D-2 remained the most influential fuel variable (Coef. = 8.52E-01, $P < 0.001$) in stopping fires, which aligns with the common knowledge that deciduous forests are generally less flammable than coniferous forests (Cumming, 2001; Bernier et al., 2016).

Previous studies used mixedwood as a simple fuel category; in contrast, this thesis explored the influence of differing conifer levels within the mixedwood fuel type. As would be expected, mixedwood forests with 10 to 40% of coniferous fuel were better at stopping fire than mixedwood forests with 40 to 60% of coniferous fuels. From the results of a different scheme RF20, as shown in Figure 4-7, in which all three mixedwood fuel variables were included, we can see that mixedwood forests with 10 to 40% of coniferous fuel are significantly fire-stopping, mixedwood

forest with 40 -60% showed no significance, and that mixedwood forest with 60-90% coniferous fuel were significantly fire-prone.

This result provides insights into how different fuel types affect fire cessation. Fire management agencies could use this information to make informed decisions when planning strategic fire containment lines. Mixedwood forests can be viewed differently depending on the coniferous proportion. The findings indicate that a lower proportion of coniferous trees in mixedwood forest can help stop fires, whereas higher proportions increase fire susceptibility. Forest management agencies could use this understanding of conifer proportion inside the mixedwood forest as a guideline for fuel management to reduce coniferous proportion within a mixedwood forest to a threshold of 40 percent to maintain the mixedwood forest's fire-stopping properties.

4.6.3. Topography

This study indicated that topography, such as elevation and slope difference, had a significant influence on halting and promoting wildfires. As shown in Figure 4-4, the standardized coefficient of slope and elevation is 0.117 and -12.540. The exp(coef) was 1.124 and < 0.001 respectively. This means the increase in one standard deviation of slope difference would increase probability of fire perimeter formation by 12%, and one standard deviation increase in elevation would increase the probability of fire stopping by more than 10 folds. The influence of those two variables should be viewed together. Normally, it is common to attribute elevation differences to the existence of hills or clips, and the negative coefficient of the continuous elevation coefficient matched this description perfectly. As mentioned in previous chapters, the negative coefficient means with increasing elevation as you move from the inside of a fire to the outside, there is a decreasing chance of forming a fire cessation event. Although the landscape in between is not accounted for, there is an uphill slope between the two points. It is in line with the common

knowledge that fire likely travels uphill much easier due to the preheating of fuels (Whelan, 1995; Linn et al., 2002; Dupuy et al., 2011).

Slope also had a high level of statistical significance ($P<0.001$), but it had a positive coefficient. In this context, fire cessation is more likely to happen if the average slope (i.e., a unit in percentage rise) of the sampled point outside is higher than the inside point. A difference in slope between two sampled pairs typically indicates changing topographic features and often signifies a break in fuel continuity; this finding aligns with previous research showing that steeper slope changes usually resulted in such discontinuities (Beaty & Taylor, 2001; Heyerdahl et al., 2001; Knapp & Keeley, 2006). It is also noted that the slope variable had a relatively tiny coefficient, representing the influence of this variable on fire stopping, which was rather small compared to others.

4.6.4. Water-related Variables

By using both the proportion and the Euclidean distance to water, the effect of water to fire cessation had a high statistical significance ($p<0.001$). The respective coefficients suggest the presence of more water, and the closer to a water body, can increase the probability of fire cessation. This is to be expected, given the physical presence of waterways interrupts fuel continuity, encouraging fire cessation and providing strategic points for firefighting efforts. Water bodies also create cool, moist, and shady environments that support vegetation growth in adjacent areas that will have higher moisture content, thus acting as natural firebreaks. Moreover, these findings aligned with those from a previous study in Saskatchewan, Canada, where fire was more likely to occur in areas distant from circular-shaped lakes, in areas with less surrounding water (Nielsen et al., 2016).

4.6.5. Previous Burned Areas

Areas that experienced a fire 0 to 10 years prior, demonstrated a fire-stopping capability even when compared with low flammability fuels like the FBP D-1/D-2 deciduous fuel type. This result suggests that recent fire history in a region contributes positively to reducing the spread of new fires, possibly due to the absence of fuel. The statistical significance ($P < 0.01$) underlined the robustness of this relationship, emphasizing the importance of considering recent fire history in fire management and modelling efforts.

Conversely, the impact of fires occurring more than 70 years ago presented a stark contrast, presenting as fire prone with a high statistical significance ($P < 0.001$). The result suggests that the fire-stopping influence of historical fires diminishes over time, aligning with the characteristics of the spruce fuel types known for their highly flammable nature. After an extended period, the ecological and fuel conditions appeared to revert to a state that is indistinguishable from areas without a known fire history, indicating that the legacy of the past fire has been entirely erased. The result confirms the phenomenon of ecosystem memory, as the fuel limitation created by past fire would only last until the stands recovery and vegetation filled in (Peterson, 2002; Parks, Holsinger, et al., 2015; Harris & Taylor, 2017). In the result from scheme RF20 (Figure 4-7), this scheme included multiple variables such as the previous burned area from 30 to 40, 50 -60, 60 – 70, and 70 plus years. Each of those previous burned variables in this scheme was significantly fire prone ($P<0.001$), reiterating the finding that the fire-stopping of the historical fires diminishes over time. Beverly (2017) reported the protective effect of previous fire that decreased the probability of a new fire escaping containment diminishes after 20 to 45 years. Parks, Miller, et al. (2015) reported the self-limiting nature of fire would last up to 20 years. Results of both of these prior studies were further confirmed with the results of this thesis.

These findings highlight the dynamic nature of fire-ecosystem interactions and the need to consider the temporal dimension of historical fire impacts on current fire perimeters. Understanding these temporal effects is necessary for improving predictive models and developing more effective fire containment line planning and land management strategies. These findings open new directions for further explorations of how fuel types recover post-fire and influence the behaviour of subsequent fires.

4.6.6. Automated Inside-outside Fire Points Data Clean Technique

To ensure an unbiased MCC Clogit result, modellers must identify paired sample points exactly in the expected locations. Previous studies by Macauley et al. (2022) had specifically addressed the influence of the distance between matched pairs and discovered the best distance for model performance is 100 m, i.e., 100 m sampling distance both inside and outside the fire. The 100 m was used as a default in this thesis without regard to the size of the fires. While the 100-meter distance ensured capture of the complete fire-environment interactions leading to fire cessation, working with digitized fire perimeters by using this 100 m sampling distance increased the chance of incorrect point sampling for the ridged fire edges. Macauley et al. (2022) addressed this issue visually by manually reviewing the data. This type of manual visual cleaning was not feasible for any study with a large-scale data set.

In this thesis, creating a custom function in R by utilizing the "DataClean" variable for cleaning was used to automate the cleaning process to accommodate any number of fires. The custom function increased the efficiency of the cleaning process. It helped clarify the data and significance of these sampling issues in fire cessation modelling. An additional notable achievement of the function was that it cleaned the sampled data of the points accidentally sampled in the unburned islands or partially burned areas. As described previously, the fire polygons were transformed into a single line to ensure

the sampling model could correctly sample along the fire perimeter; some fires even required additional treatment for splitting and merging. The transformation had unintentionally created edges, resulting in sampled points occurring within unburned islands inside the fire, which are invalid sample points. Using the "DataClean" variables, areas of unburned islands were also considered outside the fire. Hence, the points sampled within were cleaned as well. In total, 30 percent of the total points were deemed unfit and removed.

4.6.7. Automated NVB-clean Technique

Out of all the dummy variables converted from the FBP System fuel grid, NF (non-fuel) and VNF (vegetated non-fuel) emerged as the most influential in the model (largest value of Clogit coefficient) but non-significant ($p>0.1$). An automated NVB-clean guaranteed that these two fuel types only occur outside the fire perimeter without exceptions. This cleaning step is crucial for maintaining the integrity and accuracy of the model, as it helps to eliminate potential biases that could arise from the misclassification of fuel types, as shown the comparison results in Section 4.4.1.

The main reason for using this clean is that there are possible biases in the fuel grid data regarding VNF and NF. This was substantiated with evidence when reviewing some of the selected fires in the study. For example, Figure 4-18 shows the MWF052_2015 fire perimeter overlaid on the 2014 fuel grid. Most of the fire perimeter burned into areas considered VNF. Figure 4-19 shows the perimeter of the latest previously burned fire near MWF052_2015, a fire called E01025_1998, which occurred 17 years prior. A further check using Google Earth imagery (Figure 4-20) showed that the land had undergone significant revegetation post-fire, contradicting its classification as VNF. It showed that other than the charred black areas, everywhere else was revegetated and contained possible burnable fuels. It is expected that the fuel grid is imperfect, and the assigned

classification may not represent conditions on the ground. Ecosystems are dynamic and the fuel grid is updated regularly as conditions change; however, due to the nature of large remotely sensed landcover data, misclassification can be expected, the fire perimeter from 1998 in this area (not shown in the figure) was also marked as VNF when burnable fuels had accumulated and sustained a burn already.

As a result, extra care was taken when running models with the VNF or NF variable because fire perimeters in Boreal Alberta were very sensitive to these two fuel types. Understanding the limitations and potential inaccuracies in existing fuel classification methods is vital for improving fire prediction and management strategies, emphasizing the necessity of integrating field data and empirical observations into the modelling process.

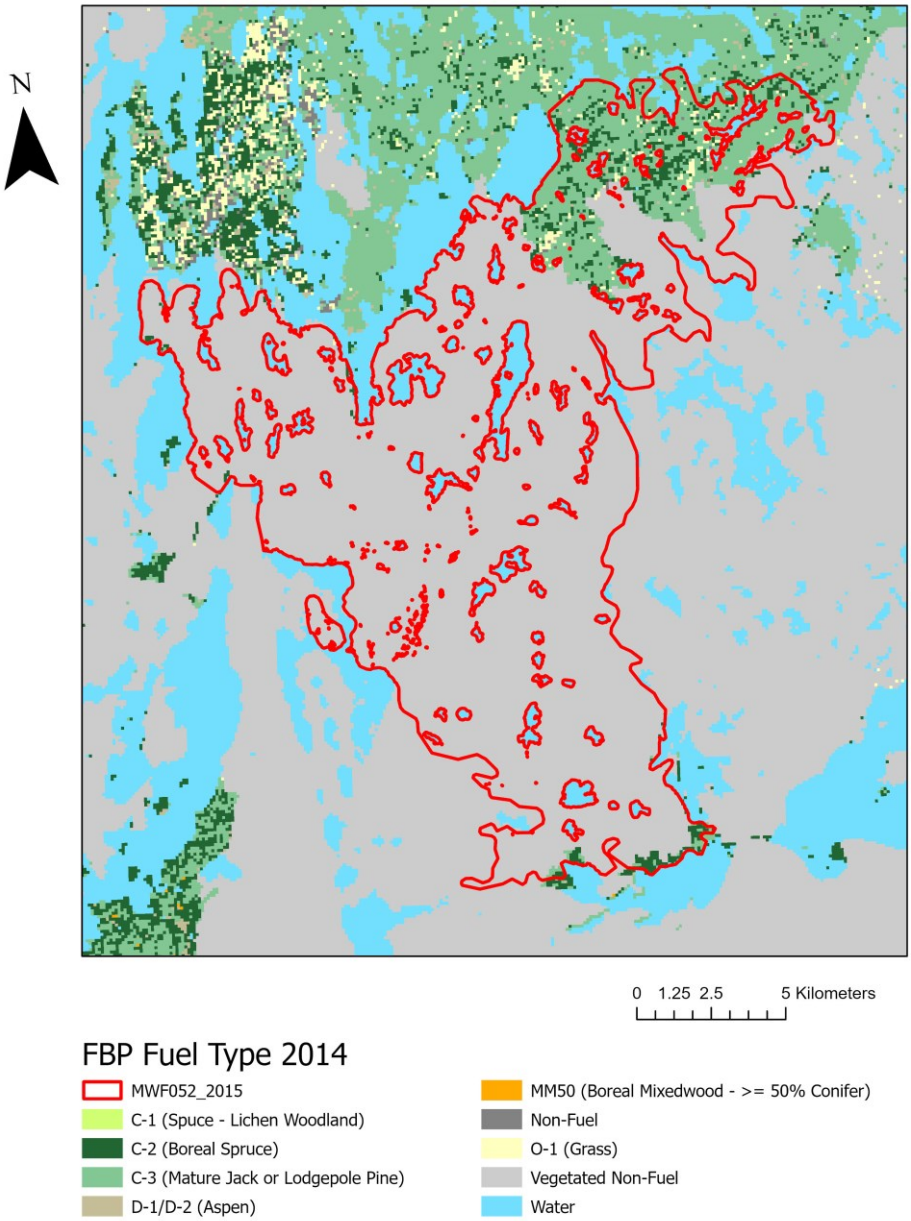


Figure 4-18 The perimeter of MWF052_2015 and the 2014 Fuel map

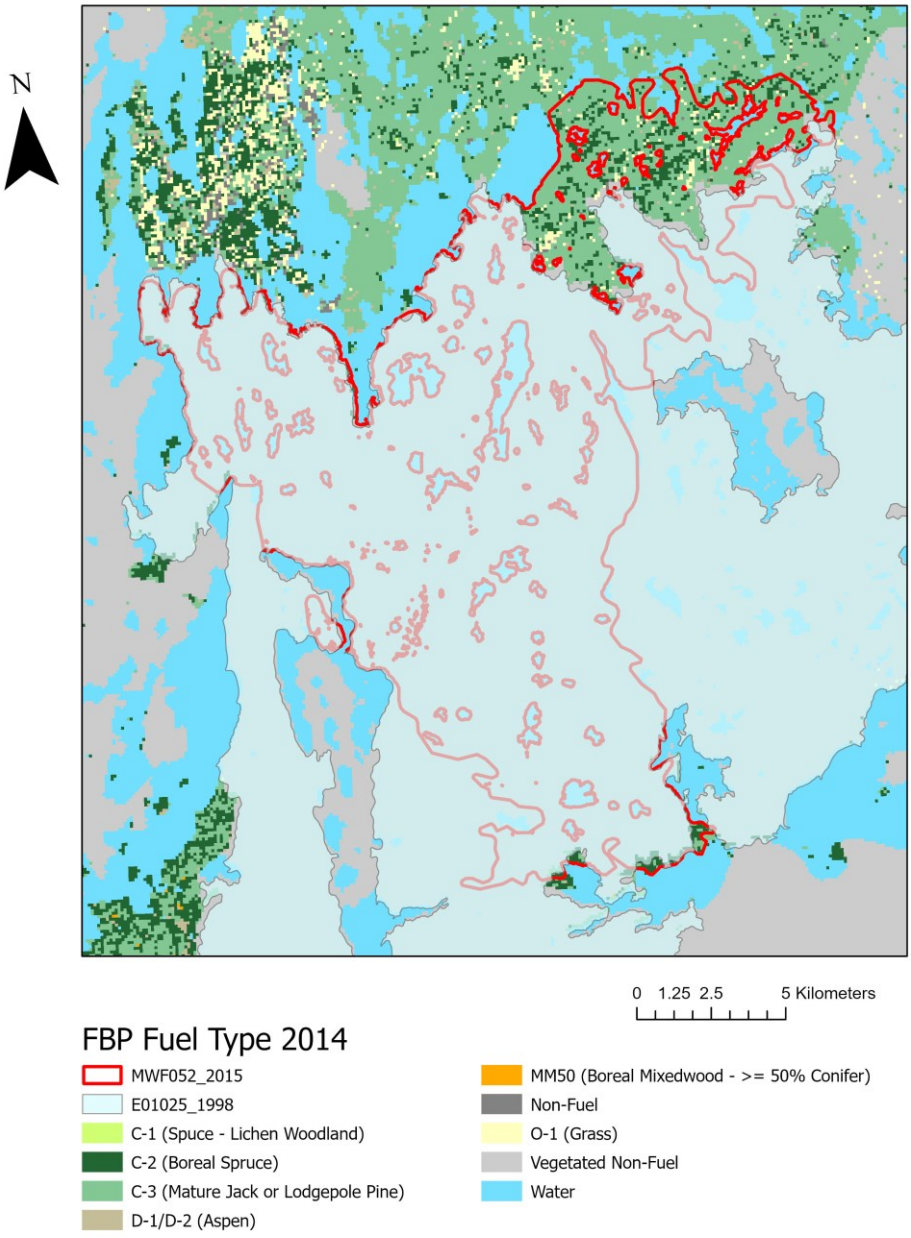


Figure 4-19 The perimeter of E01025_1998 on top of MWF052_2015 and the 2014 Fuel map

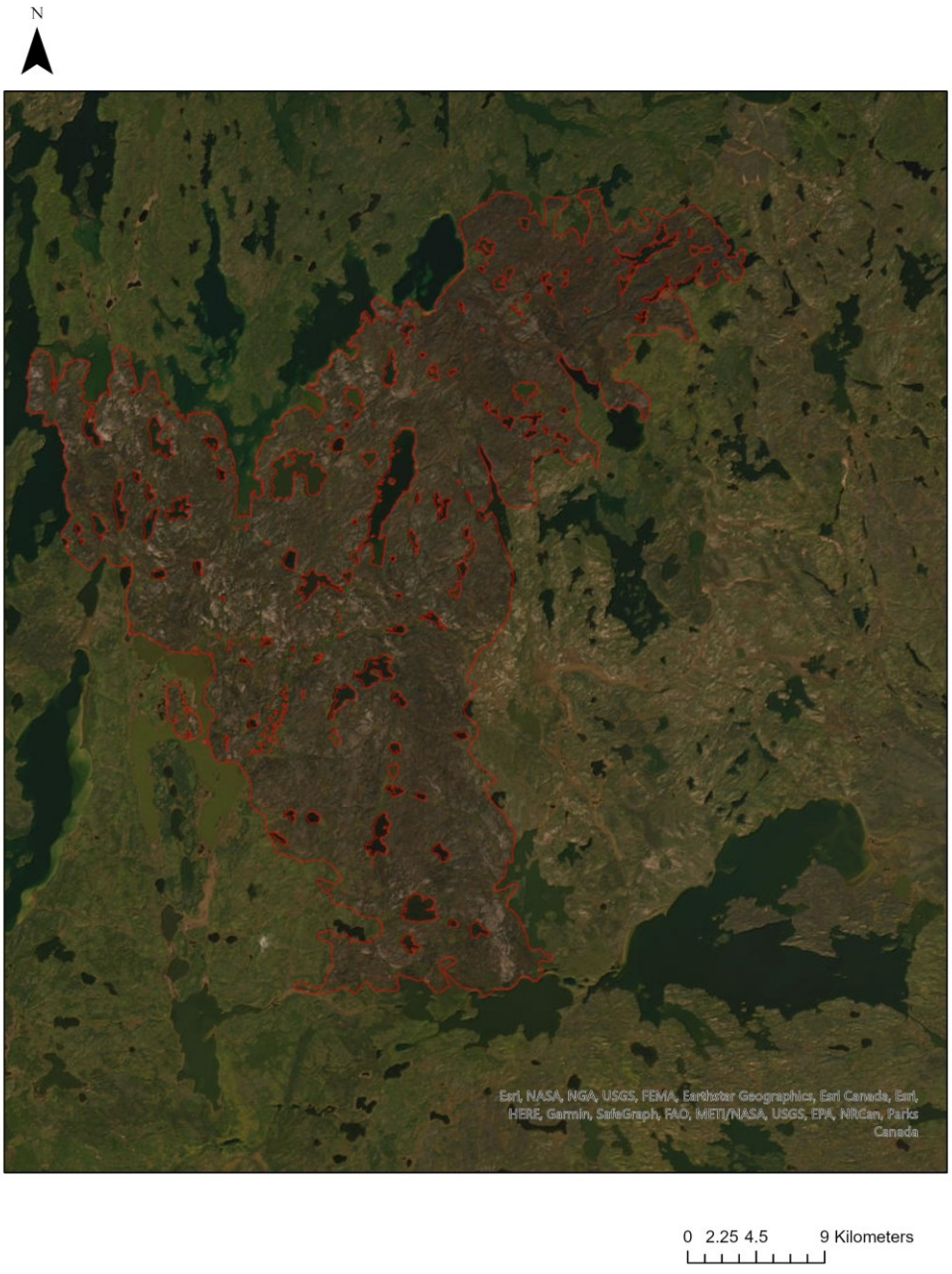


Figure 4-20 The satellite image of MWF052_2015 from Google Earth

4.6.8. Fully Clean Technique

As described in the data clean methodology, data cleaning was used to remove pairs of data points if at leas one of the points in the pair was erroneous. For convenience, this clean was defined as “full clean.” The alternative cleaning process was a partial clean, in which only the wrong data points were removed. Full clean was necessary for Clogit modelling as it needs to retain the feature of match of data pairs for the modelling. If only the Random Forest model was run, doing a full clean may not be necessary. Doing partial cleaning retains more sample points. Couronné et al. (2018) showed that RF model performance increased with the increase of data samples based on their large-scale benchmark datasets.

A simulation experiment was done to compare the full and partial clean. The result was shown in Table 4-13. The second column of the table was from the first row of Table 4-9, i.e., full clean. Partial clean includes IO (inside-outside) partial clean and NVB partial clean. There are 86,227 data points left after only using IO partial clean. After both partial IOclean and partial NVB clean, the model performance was raised significantly from AUC_train being 0.67 to 0.70. However, the AUC for Test data did not change very much.

Recalling the methodology section about NVB clean, considering the non-Fuel classification was too coarse, this thesis created an algorithm to fully clean those points with FBP-VNF and FBP-NF inside the fire. To support this algorithm, the comparing simulation experiment was designed to simulate for partial IO clean but without NVB-clean, as shown in the final column of Table 4-13. After this, the model performance was not raised. This confirmed the same conclusions from the previous section in that VNF and NF were very sensitive. The model performance decreased without cleaning its inaccurate points, and the influence were not explored clearly. Partial cleaning can indeed raise model performance. However, as partial data cannot use Clogit,

Clogit is required to detect the variable direction. Ultimately, the MCC feature was kept to run RF with full clean, not using partial clean.

Table 4-13 Comparison between partial clean and full clean for Scheme C for V700-109 fires

	Full IO clean - Full NVB clean	Partial IO clean - Partial NVB clean	Partial IO clean – No NVB clean
Full or Partial Inside-Outside clean	Full	Partial	Partial
Data points After IO clean	71030	86227	86227
NVB clean	Full	partial	No
Data points After NVB clean	65078	68054	71030
AUC_Train	0.67	0.70	0.67
AUC_Test	0.70	0.70	0.66
Accuracy	0.64	0.64	0.61

4.6.9. Mixed effect in modelling

As mentioned in the previous sections, over 100 fires were used in the automated sampling process. This produced over 60,000 sampled points that were later used in modelling. It would raise the concern that each individual fire produced many sample points, meaning multiples observations are on the same "individual" (i.e., the same fire). The question of whether the independent variables are truly independent arises (Zuur et al., 2009).

In a normal logistic regression model, there are possibilities of including a random effect to deal with this issue. However, in this thesis, a specific conditional logistic regression model was used (conditional (fixed effect) logistic regression), where the focus was primarily on the matched pairs, and the intercept was implicitly accounted for by the conditioning process. Therefore, as shown in the different schemes in Section 4.1, the model did not estimate any overall intercept across the dataset, but rather focusing on the matched case-control pairs (strata). It is possible incorporate

mixed effect into conditional logistical regression model, such as by using the “mclogit” package from R. However, the current algorithms in R do not have the capability to compute a mixed effect with paired data yet. In addition to that, the Clogit model from the “survival” package, that was used in the study does not support the additional of a mixed effect, therefore it is currently impossible to account for a mixed effect in this current study. Tackling this will be a future work for logistic regression.

4.7. Summary

The Matched Case-Control conditional logistic regression (MCC Clogit) model was used to identify the key influential factors on the fire perimeter under four schemes by mixing and matching the two conditions, with and without the help of stepwise regression (Step), with and without the consideration to the association of the variable inside and outside the fire perimeter based on the p-value of the contingency table being less than 0.1 (cp0.1). Due to the lack of prediction capability of MCC Clogit, the Random Forest (RF) model was used to evaluate the four MCC Clogit schemes, along with two additional RF schemes obtained using the variable importance measure of RF. Upon rerunning RF over the six schemes, it was found the scheme obtained by MCC Clogit while considering cp0.1 and without using Stepwise regression (S20withcp0.1Nostep), showed the best model performance, achieving the highest Area Under the Curve (AUC) and using the minimum number of variables. Selecting this scheme follows the principle of using the least number of variables to achieve the best model performance.

Among the identified 20 key factors, water, slope, aspen, grass, non-fuel, vegetated non-fuel, mixedwood with 10-40% conifer, mixedwood with 40-60% conifer, previous fire 0-10 years ago were the most significant fire-stopping factors. Notably, conifer content up to 60% in mixedwood remained a significant fire-stopping factor. However, as the conifer percentage increased, the

effectiveness of mixedwood forming fire boundary decreased from 2.19 to 1.55 folds. Although previous fire from 0 to 10 years ago played an important role in fire cessation. Previous fires older than 70 years exhibited the opposite effect. Grass and aspen were fire-stopping variables, whereas conifer species (C-1 (Spruce-Lichen Woodland) to C-4 (Immature Jack or Lodgepole Pine)) were identified as fire-prone.

Overall, RF's variable importance was not as clear as that identified by MCC Clogit, highlighting the role of MCC Clogit in identifying the key factors contributing to fire cessation. Nevertheless, RF remained helpful in selecting the best scheme among the options comparing train and test datasets. RF helped the prediction of fire perimeters for four example fires that were not used in model building. Based on the best-performing scheme (S20withcp0.1Nostep), the identified variables were used to predict fire perimeters using RF. The established final model, encompassing all 109 fire data points (train and test combined, noted as V1000), provided a potent tool for predicting fire perimeters in the Boreal region in Alberta. The model's predictive capabilities were assessed with randomly selected diverse areas within the Alberta Boreal Region, each related to a fire record. The AUC of the predictive model was 0.70, indicating a prediction capability of 70%. The higher prediction capability was noted with regions less susceptible to fire weather conditions.

5. Conclusions and Prospective

5.1. Main Takeaways of this Thesis

(1) An automated algorithm for matched case-control data pair sampling and data extraction over 109 selected fires within the boreal region of Alberta during recent years was established. An event-table toolbox was developed to help sample matched pairs of data automatically on each side of the fire perimeters to represent burned and unburned states. Python models in ArcGIS Model Builder, custom Python tools, and custom Excel Macro were used to automatically extract spatial information representing the explanatory variables associated with each data point. Compared with a manual process, this automation reduced workload from several weeks to hours.

(2) Visual review of each analyzed fire was conducted individually to verify that the automated algorithm's performance was consistent with expectations. Various procedures were adopted for dealing with complicated fires with multiple fire polygons corresponding to the same fire ID. These included sampling the largest polygon with a small polygon neglected, combining two very closely related fires into a new fire, and separating large fires with two or more equal large polygons into two or more new ones. Several cleaning algorithms were employed in conjunction with those procedures, including the inside-outside clean, NVB-clean, and others. This approach produced an accurate dataset, from which the basic association of the binary variable inside and outside the fire boundary was tested by statistical tests with p-values. Comparative descriptive features inside and outside the fire boundary for continuous variables were analyzed.

(3) An advanced R-based modelling framework was established by combining matched case-control (MCC) conditional logistic regression (Clogit) and random forest (RF) to identify the key factors and predict fire perimeters. Four sets of factors were generated using MCC Clogit under different conditions: with and without stepwise regression (Step) and with and without considering the association of variables inside and outside the fire perimeter, based on the contingency table ($P < 0.1$). Additionally, two sets of factors were produced using RF according to the variable importance measure, with the same number of variables as the MCC Clogit schemes. RF was rerun over the six schemes, and the most optimal scheme was identified as the one considering cp0.1 without stepwise regression, achieving the highest Area Under the Curve (AUC) and utilizing the minimum number of variables.

(4) Among the 20 variables in the best scheme, water proportion, slope, aspen, grass, non-fuel, water, vegetated non-fuel, mixedwood with 10-40% conifer, mixedwood with 40-60% conifer, previous fire from 0-10 years ago and previous fire from 10-20 years ago were the most significant fire-resistant variables. Notably, conifer content up to 60% in mixedwood remained a significant fire-stopping factor. With more conifer content, the probability of fire cessation decreased from 2.19 to 1.55 folds. Although previous fire from 0 to 10 and 10 to 20 years ago played important roles in fire cessation, fires older than 70 years exhibited the opposite. Grass and aspen were fire-stopping elements, and all conifers (C-1 (Spruce-Lichen Woodland) to C-4 (Immature Jack or Lodgepole Pine)) were identified as fire-prone, especially C-2, as it was the most important variable in RF models. The identified results in terms of Mean Decrease Accuracy in RF were more in agreement with results by Clogit detection than Mean Decrease Gini. Overall, the pattern of each variable identified only by RF's variable importance was not as clear as that identified by MCC

Clogit, showing the value of MCC Clogit in identifying key factors that influence fire cessation. However, RF helped select the best factor list among the choices.

(5) Based on the best factor list (S20withcp0.1Nostep), the identified variables were used to predict fire perimeters using RF. The established final model, encompassing all 109 fire data points (noted as V1000), provided a potent tool for predicting fire perimeter in the Boreal region in Alberta. The AUC of the predictive model was 0.70, indicating a prediction capability of 70%. It showed a higher prediction capability for a fire with milder fire weather conditions than those with high wind speed.

(6) The findings of this thesis have implications for strategic containment line planning in wildfire management. By identifying the key factors influencing fire perimeter formation, the automated data collection methods and modelling framework of this study provide wildfire managers with valuable information on areas with high potential for containment and natural fire containment barriers. This information could aid in tailoring containment strategies and managing fire containment lines more effectively. Furthermore, this study offers wildfire managers a deeper understanding of the landscape variables influencing fire containment, enabling more informed strategic decision-making.

5.2.Future Perspectives

5.2.1. Seasonality

The fire season in Alberta is from March 1 to October 31 (Government of Alberta, 2022b), beginning in early spring and extending to the end of autumn. The fire behaviours vary, influenced by the changing conditions of each season.

In the early months of the fire season, particularly in spring, a notable phenomenon known as the “spring dip” occurs. This term refers to the period when deciduous trees and grasses, out of

dormancy, exhibit exceptionally low moisture content (Alexander & Cruz, 2013; De Jong et al., 2016). This condition arises because these plants have not yet begun their new growth cycle after winter. Consequently, they have not yet replenished moisture through uptake from soil and initialized photosynthesis. As a result, these fuels become highly flammable during this period. However, as spring progresses and plants begin their growth, the moisture content increases, and their flammability gradually decreases.

Despite the susceptibility to fire in the spring due to the spring dip, out of the 109 fires of my choice, only 32 out of the 109 fires occurred during the spring season, all later in spring, after the spring dip period. On the other hand, summer fires are influenced by different factors, including high temperatures and lower relative humidity, which collectively contribute to a peak in fire occurrence during summer. Future research considering seasonality's role in fire perimeter formation will provide deeper insights into temporal fire patterns, ultimately aiding in developing more effective fire management and mitigation strategies.

5.2.2. Future Research: FBP and LSAT

Future research could explore potential improvements in model performance by using Landsat landcover data (Landsat) rather than the fuel grids used here, which identify the fuel types of the Canadian Forest Fire Behaviour Prediction (FBP) System. Landsat provides comprehensive, satellite-based landcover data that offers more precise data (i.e., 30 m resolution) than the FBP System fuel grid. Landsat data could be particularly beneficial in regions or scenarios where the FBP system might have limitations due to the unavailability of specific ground-based data or the need for more granular spatial resolution.

A hybrid model incorporating the FBP System fuel grid, and Landsat factors could offer more robust predictive capabilities. By comparing these two data sources, researchers can identify the

strengths and weaknesses inherent to each and potentially uncover synergies between them. Future research that involves periodic ground-truthing of non-fuel fuel types through on-site investigations is also a promising direction for improving the accuracy of fuel maps.

References

Alberta Agriculture and Forestry. (2016). *Historical wildfire database*. Retrieved (Date Accessed): 1 May 2021 from https://open.alberta.ca/dataset/wildfire-data#summary

Alberta Agriculture and Forestry. (2021). *Spatial wildfire database*. Retrieved May 2, 2021 from https://wildfire.alberta.ca/resources/historical-data/spatial-wildfire-data.aspx

Alberta Biodiversity Monitoring Institute. (2010). *Wall-to-Wall Human Footprint Inventory*. https://abmi.ca/home/data-analytics/da-top/da-product-overview/Human-Footprint-Products/HF-inventory.htm

Alexander, M. E., & Cruz, M. G. (2013). Corrigendum to: Assessing the effect of foliar moisture on the spread rate of crown fires. *International Journal of Wildland Fire*, *22*(6), 869-870.

AltaLIS. (2021). *Alberta Digital Elevation Model*. https://www.altalis.com/

Archer, E., & Archer, M. E. (2016). Package 'rfPermute'. *R Project: Indianapolis, IN, USA*.

Arno, S. F., & Brown, J. K. (1991). Overcoming the paradox in managing wildland fire.

Bannerman-Thompson, H., Rao, M. B., & Kasala, S. (2013). Bagging, boosting, and random forests using R. In *Handbook of Statistics* (Vol. 31, pp. 101-149). Elsevier.

Beaty, R. M., & Taylor, A. H. (2001). Spatial and temporal variation of fire regimes in a mixed conifer forest landscape, Southern Cascades, California, USA. *Journal of Biogeography*, *28*(8), 955-966.

Bernier, P., Gauthier, S., Jean, P., Manka, F., Boulanger, Y., Beaudoin, A., & Guindon, L. (2016). Mapping local effects of forest properties on fire risk across Canada. Forests 7: 157. In.

Beverly, J. L. (2017). Time since prior wildfire affects subsequent fire containment in black spruce *International Journal of Wildland Fire*, *26*(11), 919-929. https://doi.org/https://doi.org/10.1071/WF17051

Beverly, J. L., McLoughlin, N., & Chapman, E. (2021). A simple metric of landscape fire exposure. *Landscape Ecology*, *36*(3), 785-801.

Brandt, J. P. (2009). The extent of the North American boreal zone [Review]. *Environmental Reviews*, *17*, 101-161. https://doi.org/10.1139/A09-004

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Breslow, N., Day, N., Halvorsen, K., Prentice, R., & Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology*, *108*(4), 299-307.

Breslow, N. E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, *91*(433), 14-28.

Calle, M. L., & Urrea, V. (2011). Stability of Random Forest importance measures. *Briefings in bioinformatics*, *12*(1), 86-89.

Compton, B. W., Rhymer, J. M., & McCollough, M. (2002). Habitat selection by wood turtles (Clemmys insculpta): an application of paired logistic regression. *Ecology*, *83*(3), 833-843.

Countryman, C. M. (1972). *The fire environment concept*. Pacific Southwest Forest and Range Experiment Station.

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, *19*, 1-14.

Cumming, S. (2001). Forest type and wildfire in the Alberta boreal mixedwood: what do fires burn? *Ecological applications*, *11*(1), 97-110.

De Jong, M. C., Wooster, M. J., Kitchen, K., Manley, C., Gazzard, R., & McCall, F. F. (2016). Calibration and evaluation of the Canadian Forest Fire Weather Index (FWI) System for improved wildland fire danger rating in the United Kingdom. *Natural Hazards and Earth System Sciences*, *16*(5), 1217-1237.

DeLancey, E. R., Kariyeva, J., Cranston, J., & Brisco, B. (2018). Monitoring hydro temporal variability in Alberta, Canada with multi-temporal Sentinel-1 SAR data. *Canadian Journal of Remote Sensing*, *44*(1), 1-10.

Dupuy, J.-L., Maréchal, J., Portier, D., & Valette, J.-C. (2011). The effects of slope and fuel bed width on laboratory fire behaviour. *International Journal of Wildland Fire*, *20*(2), 272-288.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Fernandes, P. M., Loureiro, C., Guiomar, N., Pezzatti, G. B., Manso, F. T., & Lopes, L. (2014). The dynamics and drivers of fuel and fire in the Portuguese public forest. *Journal of environmental management*, *146*, 373-382.

Fernandes, P. M., Pacheco, A. P., Almeida, R., & Claro, J. J. E. J. o. F. R. (2016). The role of fire-suppression force in limiting the spread of extremely large forest fires in Portugal. *135*(2), 253-262.

Forestry Canada Fire Danger Group. (1992). *Development and structure of the Canadian forest fire behavior prediction system* (Vol. 3). Forestry Canada, Science and Sustainable Development Directorate.

Fox, J., & Monette, G. J. J. o. t. A. S. A. (1992). Generalized collinearity diagnostics. *87*(417), 178-183.

Gail, M. H., Lubin, J. H., & Rubinstein, L. V. (1981). Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika*, 703-707.

Government of Alberta. (2022a). *Natural Regions and Subregions of Alberta*. Retrieved August 31

from https://open.alberta.ca/opendata/gda-2f36921e-41e3-4cd8-813e-3333ea3c5983

Government of Alberta. (2022b). *Wildfires*. https://open.alberta.ca/dataset/a221e7a0-4f46-4be7-9c5a-e29de9a3447e/resource/80480824-0c50-456c-9723-f9d4fc136141/download/fp-historical-wildfire-data-2006-2023.xlsx

Government of Canada. (2010). *2010 Land Cover of Canada*. https://open.canada.ca/data/en/dataset/c688b87f-e85f-4842-b0e1-a8f79ebf1133

Greene, D. F., Zasada, J. C., Sirois, L., Kneeshaw, D., Morin, H., Charron, I., & Simard, M.-J. (1999). A review of the regeneration dynamics of North American boreal forest tree species. *Canadian Journal of Forest Research*, *29*(6), 824-839.

Grimmett, G., & Stirzaker, D. (2020). *Probability and random processes*. Oxford university press.

Hardy, C. C. (2005). Wildland fire hazard and risk: Problems, definitions, and context. *Forest ecology management*, *211*(1-2), 73-82.

Harrel, F. E. (2015). In *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (pp. P283-286). Springer International Publishing. . https://doi.org/10.1007/978-3-319-19425-7

Harris, L., & Taylor, A. H. (2017). Previous burns and topography limit and reinforce fire severity in a large wildfire. *Ecosphere*, *8*(11), e02019.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). In *The Elements of Statistical Learning- Data Mining, Inference, and Prediction* (pp. 270 - 272). Springer International Publishing Switzerland. https://doi.org/I 10.1007/978-0-387-21

Henderson, A. R. (2005). The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. *Clinica chimica acta*, *359*(1-2), 1-26.

Heyerdahl, E. K., Brubaker, L. B., & Agee, J. K. (2001). Spatial controls of historical fire regimes: a multiscale example from the interior west, USA. *Ecology*, *82*(3), 660-678.

Holsinger, L., Parks, S. A., & Miller, C. (2016). Weather, fuels, and topography impede wildland fire spread in western US landscapes [Article]. *Forest Ecology and Management*, *380*, 59-69. https://doi.org/10.1016/j.foreco.2016.08.035

Horner, N. (2024). *Fiscal Plan - A Responsible Plan for a Growing Province*. February 29, 2024: GOVERNMENT OF ALBERTA Retrieved from https://www.alberta.ca/expense

Hosmer, D. (2000). Lemeshow S. Applied logistic regression. In: USA: John Wiley and Sons.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Just, M. G., Hohmann, M. G., & Hoffmann, W. A. (2016). Where fire stops: vegetation structure and microclimate influence fire spread along an ecotonal gradient. *Plant ecology*, *217*(6), 631-644.

Knapp, E. E., & Keeley, J. E. (2006). Heterogeneity in fire severity within early season and late season prescribed burns in a mixed-conifer forest. *International Journal of Wildland Fire*, *15*(1), 37-45.

Latifovic, R., Pouliot, D., & Olthof, I. (2017). Circa 2010 Land Cover of Canada: Local Optimization Methodology and Product Development [Article]. *Remote Sensing*, *9*(11). https://doi.org/10.3390/rs9111098

Linn, R., Reisner, J., Colman, J. J., & Winterkamp, J. (2002). Studying wildfire behavior using FIRETEC. *International Journal of Wildland Fire*, *11*(4), 233-246.

Linn, R., Winterkamp, J., Edminster, C., Colman, J. J., & Smith, W. S. (2007). Coupled influences of topography and wind on wildland fire behaviour. *International Journal of Wildland Fire*, *16*(2), 183-195.

Macauley, K. A., McLoughlin, N., & Beverly, J. L. (2022). Modelling fire perimeter formation in the Canadian Rocky Mountains. *Forest ecology management*, *506*, 119958.

Matasci, G., Hermosilla, T., Wulder, M. A., White, J. C., Coops, N. C., Hobart, G. W., & Zald, H. S. (2018). Large-area mapping of Canadian boreal forest cover, height, biomass and other structural attributes using Landsat composites and lidar plots. *Remote sensing of environment*, *209*, 90-106.

Narayanaraj, G., & Wimberly, M. C. (2011). Influences of forest roads on the spatial pattern of wildfire boundaries. *International Journal of Wildland Fire*, *20*(6), 792-803. https://doi.org/10.1071/WF10032

Natural Regions Committee. (2006). *Natural Regions and Subregions of Alberta*. Government of Alberta Retrieved from http://www.albertaparks.ca/media/2942026/nrsrcomplete_may_06.pdf

Natural Resources Canada. (2022). *The State of Canada's Forests Annual Report*. Retrieved from https://www.nrcan.gc.ca/our-natural-resources/forests/state-canadas-forests-report/16496

Natural Resources Canada, B., J.P. (2009). *North American boreal zone map shapefiles*. https://natural-resources.canada.ca/our-natural-resources/forests/sustainable-forest-management/boreal-forest/north-american-boreal-zone-map-shapefiles/14252

Negret, P. J., Marco, M. D., Sonter, L. J., Rhodes, J., Possingham, H. P., & Maron, M. J. C. B. (2020). Effects of spatial autocorrelation and sampling design on estimates of protected area effectiveness. *34*(6), 1452-1462.

Nekrich, A. (2022). Key factors determining scales of burned areas in state Victoria (Australia) and province Alberta (Canada) during 1980-2019. *Journal of Wildlife and Biodiversity*, *6*(2), 87-99.

Nicodemus, K. K. (2011). On the stability and ranking of predictors from random forest variable importance measures. *Briefings in bioinformatics*, *12*(4), 369-373.

Nielsen, S. E., DeLancey, E. R., Reinhardt, K., & Parisien, M. A. (2016). Effects of Lakes on Wildfire Activity in the Boreal Forests of Saskatchewan, Canada [Article]. *Forests*, *7*(11). https://doi.org/10.3390/f7110265

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673-690.

O'Connor, C. D., Calkin, D. E., & Thompson, M. P. J. I. j. o. w. f. (2017). An empirical machine learning method for predicting potential fire control locations for pre-fire planning and operational fire management. *26*(7), 587-597.

Parks, S. A., Holsinger, L. M., Miller, C., & Nelson, C. R. (2015). Wildland fire as a self-regulating mechanism: the role of previous burns and weather in limiting fire progression. *Ecological applications*, *25*(6), 1478-1492.

Parks, S. A., Miller, C., Holsinger, L. M., Baggett, L. S., & Bird, B. (2015). Wildland fire limits subsequent fire occurrence. *International Journal of Wildland Fire*, *25*(2), 182-190.

Parks, S. A., Parisien, M.-A., & Miller, C. (2012). Spatial bottom-up controls on fire likelihood vary across western North America. *Ecosphere*, *3*(1), 1-20.

Peterson, G. D. (2002). Contagious disturbance, ecological memory, and the emergence of landscape pattern. *Ecosystems*, *5*, 329-338.

Povak, N. A., Hessburg, P. F., & Salter, R. B. (2018). Evidence for scale-dependent topographic controls on wildfire spread [Article]. *Ecosphere*, *9*(10). https://doi.org/10.1002/ecs2.2443

Rodrigues, M., Alcasena, F., Gelabert, P., & Vega-García, C. (2020). Geospatial modeling of containment probability for escaped wildfires in a Mediterranean region. *Risk analysis*, *40*(9), 1762-1779.

Rothermel, R. C. (1983). *How to predict the spread and intensity of forest and range fires* (Vol. 143). US Department of Agriculture, Forest Service, Intermountain Forest and Range ….

Shomal Zadeh, N., Lin, S., & Runger, G. C. (2020). Matched Forest: supervised learning for high-dimensional matched case–control studies. *Bioinformatics*, *36*(5), 1570-1576.

Stamp, R. M. (2009, November 30, 2021). *Alberta*. https://www.thecanadianencyclopedia.ca/en/article/alberta

StatisticsCanada. (2021). *Focus on Geography Series, 2021 Census of Population*. https://www12.statcan.gc.ca/census-recensement/2021/as-sa/fogs-spg/page.cfm?topic=1&lang=E&dguid=2021A000248

Stocks, B., Mason, J., Todd, J., Bosch, E., Wotton, B., Amiro, B., Flannigan, M., Hirsch, K., Logan, K., & Martell, D. J. J. o. G. R. A. (2002). Large forest fires in Canada, 1959–1997. *107*(D1), FFR 5-1-FFR 5-12.

Taylor, S. W., & Alexander, M. E. J. I. J. o. W. F. (2006). Science, technology, and human factors in fire danger rating: the Canadian experience. *15*(1), 121-135.

Therneau, T. M. (2023). *Survival: Survival Analysis*. In https://github.com/therneau/survival

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, *46*(sup1), 234-240.

Tymstra, C., Jain, P., & Flannigan, M. D. (2021). Characterisation of initial fire weather conditions for large spring wildfires in Alberta, Canada. *International Journal of Wildland Fire*, *30*(11), 823-835.

Tymstra, C., Stocks, B. J., Cai, X., & Flannigan, M. D. J. P. i. D. S. (2020). Wildfire management in Canada: Review, challenges and opportunities. *5*, 100045.

Van Wagner, C. (1987). Development and structure of the canadian forest fireweather index system. Can. For. Serv., Forestry Tech. Rep,

Walker, X. J., Rogers, B. M., Veraverbeke, S., Johnstone, J. F., Baltzer, J. L., Barrett, K., Bourgeau-Chavez, L., Day, N. J., de Groot, W. J., Dieleman, C. M., Goetz, S., Hoy, E., Jenkins, L. K., Kane, E. S., Parisien, M. A., Potter, S., Schuur, E. A. G., Turetsky, M., Whitman, E., & Mack, M. C. (2020). Fuel availability not fire weather controls boreal wildfire severity and carbon emissions [Article]. *Nature Climate Change*, *10*(12), 1130-U1100. https://doi.org/10.1038/s41558-020-00920-8

Whelan, R. J. (1995). *The ecology of fire*. Cambridge university press.

Whitman, E., Parks, S. A., Holsinger, L. M., & Parisien, M.-A. (2022). Climate-induced fire regime amplification in Alberta, Canada. *Environmental Research Letters*, *17*(5), 055003.

Whittington, J., St. Clair, C. C., & Mercer, G. (2005). Spatial responses of wolves to roads and trails in mountain valleys. *Ecological applications*, *15*(2), 543-553.

Wulder, M. A., White, J. C., Cranny, M., Hall, R. J., Luther, J. E., Beaudoin, A., Goodenough, D. G., & Dechka, J. A. (2008). Monitoring Canada's forests. Part 1: Completion of the EOSD land cover project. *Canadian Journal of Remote Sensing*, *34*(6), 549-562.

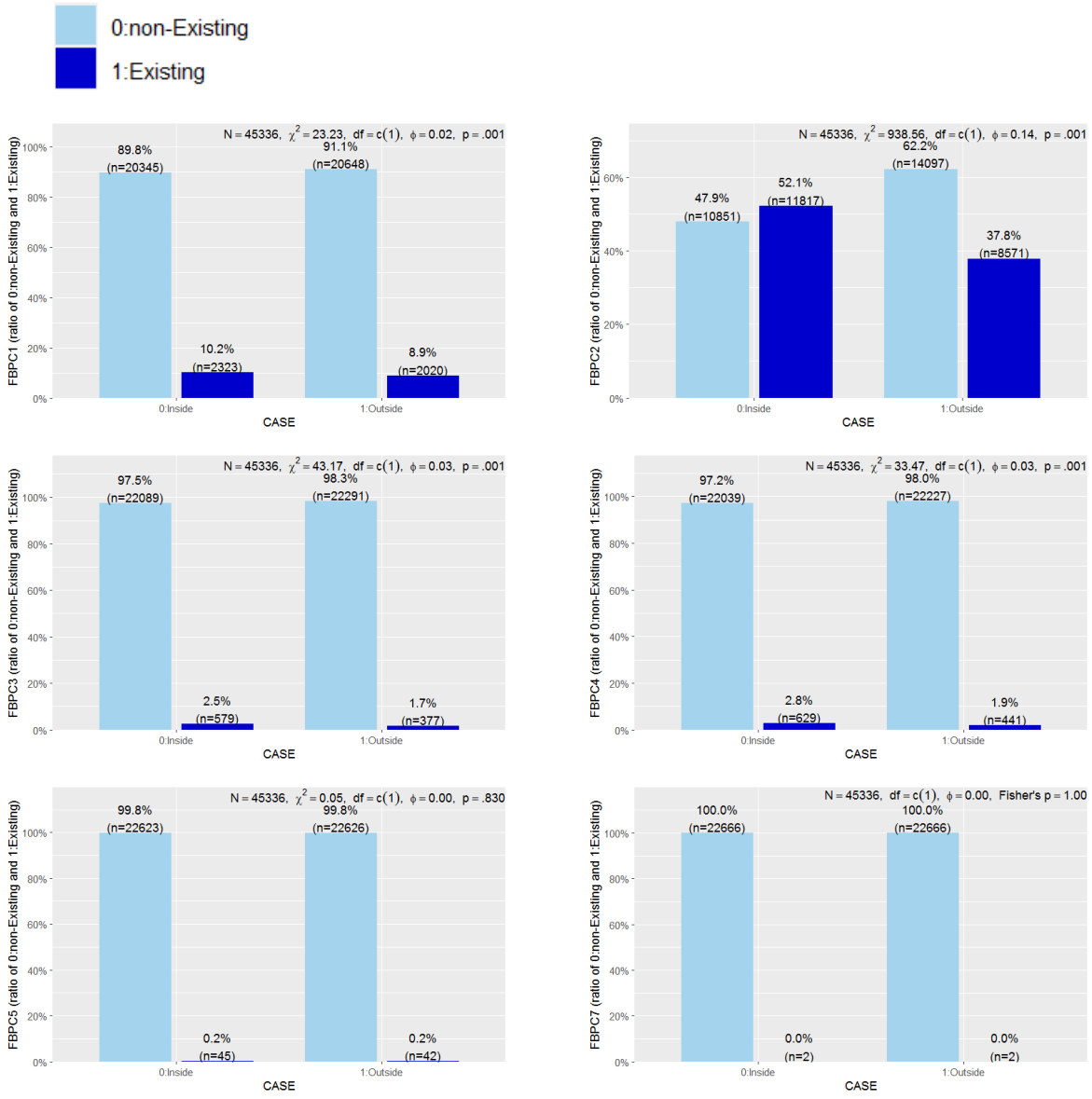
Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of analysis and testing*, *2*(3), 249-262.

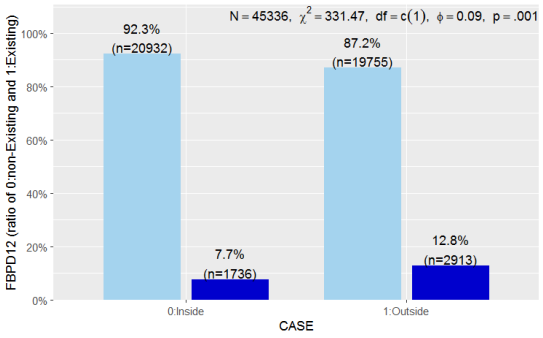
Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R* (Vol. 574). Springer.

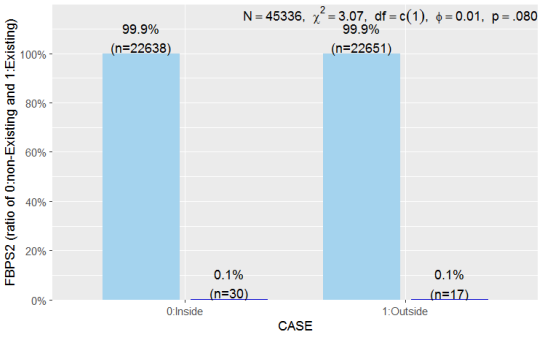
Appendix

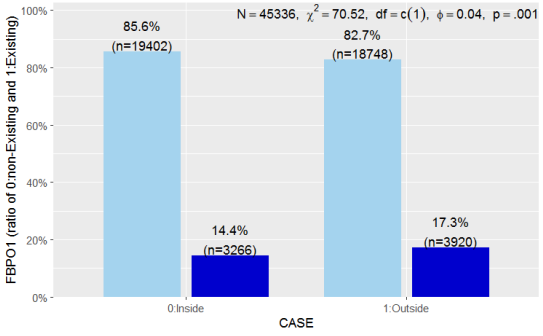
Appendix A1: Contingency Plot for Binary Variables in V700-Train Data

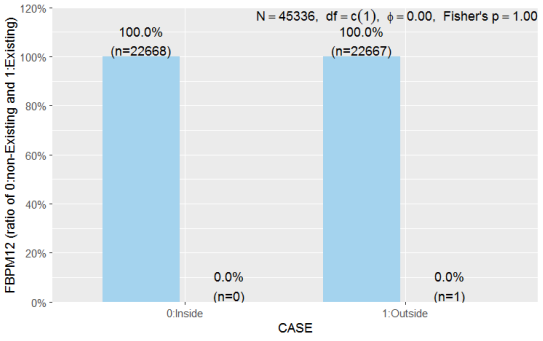
The number (n) and the percentage of non-existing (lightskyblue2) and existing (medium blue) data points for each of 29 variables inside (CASE=0) and outside (CASE=1) the fire perimeter for train data

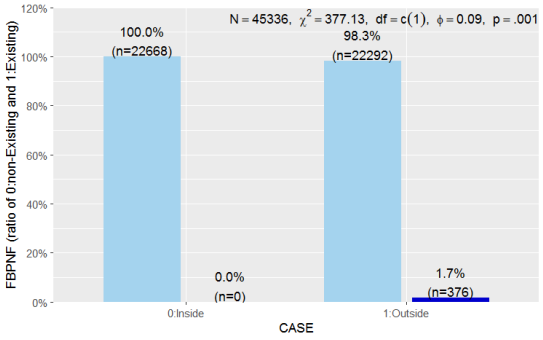


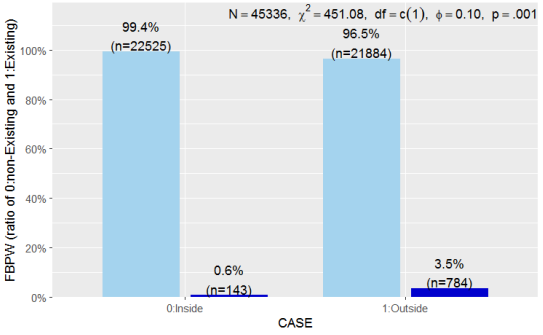

N = 45336, χ^2 = 331.47, df = c(1), ϕ = 0.09, p = .001
FBPD12 (ratio of 0:non-Existing and 1:Existing)
92.3% (n=20932)
7.7% (n=1736)
87.2% (n=19755)
12.8% (n=2913)
0:Inside
1:Outside
CASE

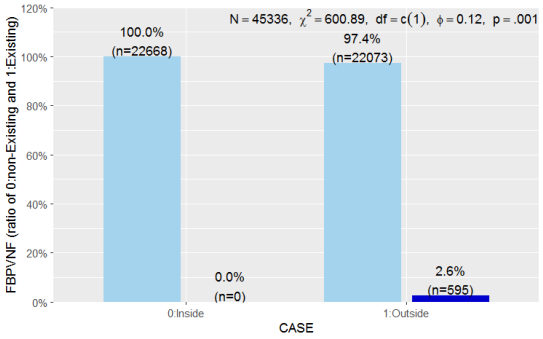

N = 45336, χ^2 = 3.07, df = c(1), ϕ = 0.01, p = .080
FBPS2 (ratio of 0:non-Existing and 1:Existing)
99.9% (n=22638)
0.1% (n=30)
99.9% (n=22651)
0.1% (n=17)
0:Inside
1:Outside
CASE

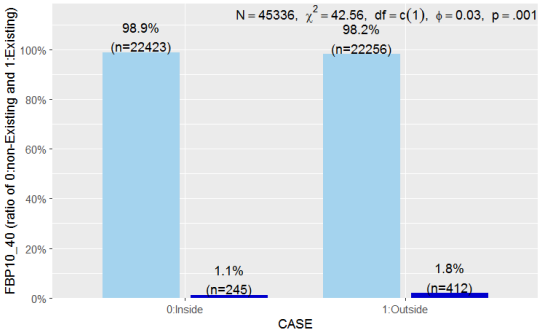

N = 45336, χ^2 = 70.52, df = c(1), ϕ = 0.04, p = .001
FBPO1 (ratio of 0:non-Existing and 1:Existing)
85.6% (n=19402)
14.4% (n=3266)
82.7% (n=18748)
17.3% (n=3920)
0:Inside
1:Outside
CASE

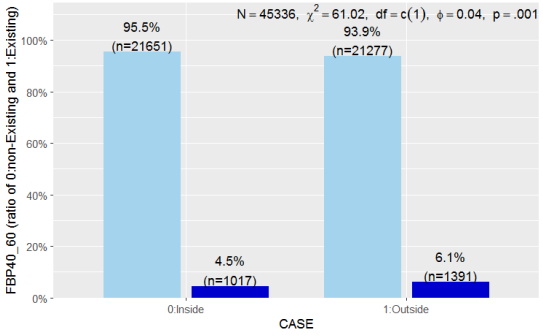

N = 45336, df = c(1), ϕ = 0.00, Fisher's p = 1.00
FBPM12 (ratio of 0:non-Existing and 1:Existing)
100.0% (n=22668)
0.0% (n=0)
100.0% (n=22667)
0.0% (n=1)
0:Inside
1:Outside
CASE

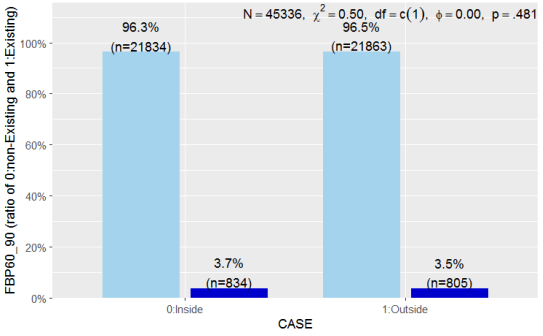

N = 45336, χ^2 = 377.13, df = c(1), ϕ = 0.09, p = .001
FBPNF (ratio of 0:non-Existing and 1:Existing)
100.0% (n=22668)
0.0% (n=0)
98.3% (n=22292)
1.7% (n=376)
0:Inside
1:Outside
CASE

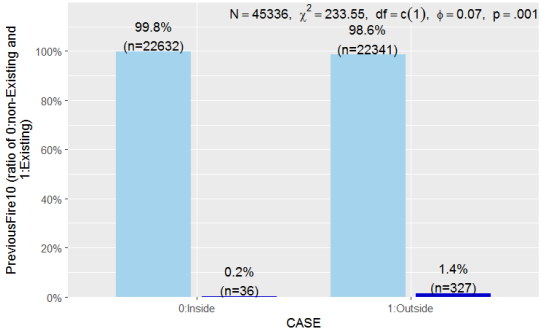

N = 45336, χ^2 = 451.08, df = c(1), ϕ = 0.10, p = .001
FBPW (ratio of 0:non-Existing and 1:Existing)
99.4% (n=22525)
0.6% (n=143)
96.5% (n=21884)
3.5% (n=784)
0:Inside
1:Outside
CASE

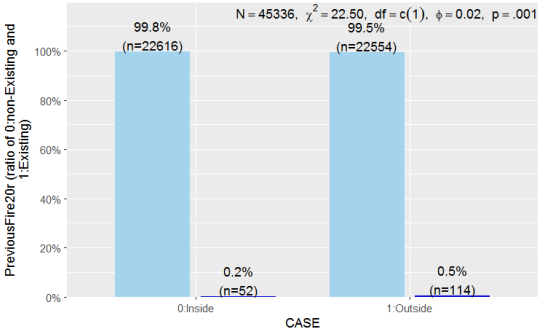

N = 45336, χ^2 = 600.89, df = c(1), ϕ = 0.12, p = .001
FBPVNF (ratio of 0:non-Existing and 1:Existing)
100.0% (n=22668)
0.0% (n=0)
97.4% (n=22073)
2.6% (n=595)
0:Inside
1:Outside
CASE

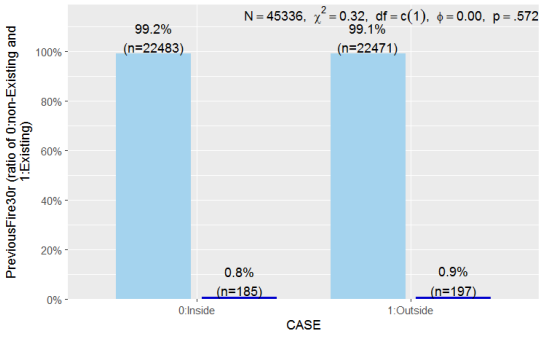

N = 45336, χ^2 = 42.56, df = c(1), ϕ = 0.03, p = .001
FBP10_40 (ratio of 0:non-Existing and 1:Existing)
98.9% (n=22423)
1.1% (n=245)
98.2% (n=22256)
1.8% (n=412)
0:Inside
1:Outside
CASE

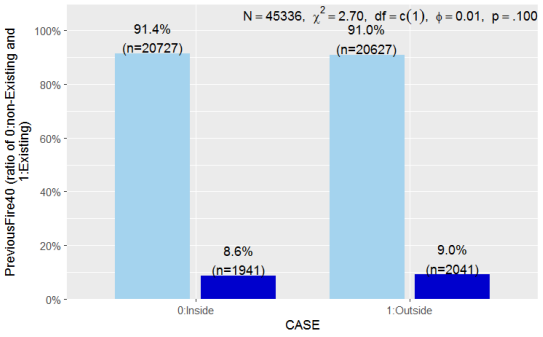

N = 45336, χ^2 = 61.02, df = c(1), ϕ = 0.04, p = .001
FBP40_60 (ratio of 0:non-Existing and 1:Existing)
95.5% (n=21651)
4.5% (n=1017)
93.9% (n=21277)
6.1% (n=1391)
0:Inside
1:Outside
CASE

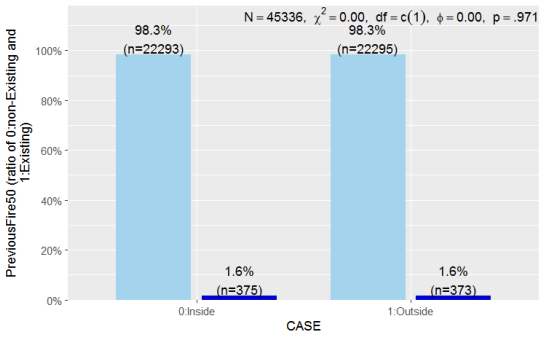

N = 45336, χ^2 = 0.50, df = c(1), ϕ = 0.00, p = .481
FBP60_90 (ratio of 0:non-Existing and 1:Existing)
96.3% (n=21834)
3.7% (n=834)
96.5% (n=21863)
3.5% (n=805)
0:Inside
1:Outside
CASE

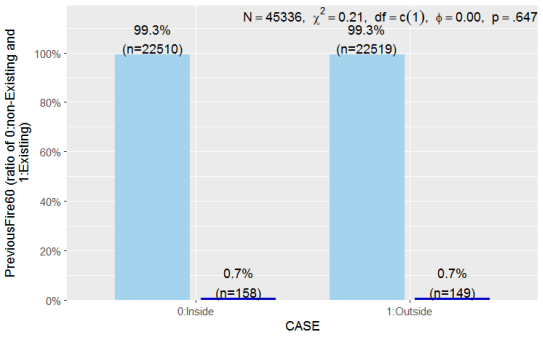

N = 45336, χ^2 = 233.55, df = c(1), ϕ = 0.07, p = .001
PreviousFire10 (ratio of 0:non-Existing and 1:Existing)
99.8% (n=22632)
0.2% (n=36)
98.6% (n=22341)
1.4% (n=327)
0:Inside
1:Outside
CASE

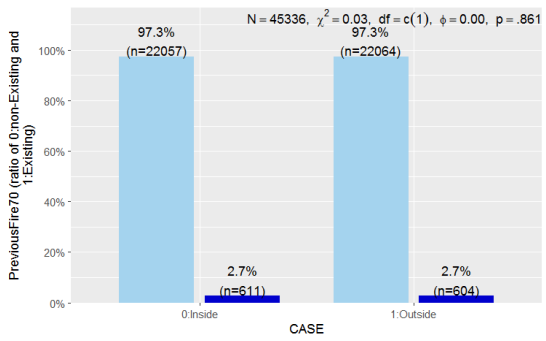

N = 45336, χ^2 = 22.50, df = c(1), ϕ = 0.02, p = .001
PreviousFire20r (ratio of 0:non-Existing and 1:Existing)
99.8% (n=22616)
0.2% (n=52)
99.5% (n=22554)
0.5% (n=114)
0:Inside
1:Outside
CASE

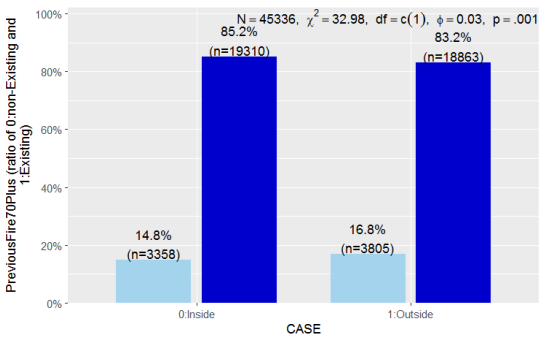

N = 45336, χ^2 = 0.32, df = c(1), ϕ = 0.00, p = .572
PreviousFire30r (ratio of 0:non-Existing and 1:Existing)
99.2% (n=22483)
0.8% (n=185)
99.1% (n=22471)
0.9% (n=197)
0:Inside
1:Outside
CASE


N = 45336, χ^2 = 2.70, df = c(1), ϕ = 0.01, p = .100
PreviousFire40 (ratio of 0:non-Existing and 1:Existing)
91.4% (n=20727)
8.6% (n=1941)
91.0% (n=20627)
9.0% (n=2041)
0:Inside
1:Outside
CASE


N = 45336, χ^2 = 0.00, df = c(1), ϕ = 0.00, p = .971
PreviousFire50 (ratio of 0:non-Existing and 1:Existing)
98.3% (n=22293)
1.6% (n=375)
98.3% (n=22295)
1.6% (n=373)
0:Inside
1:Outside
CASE

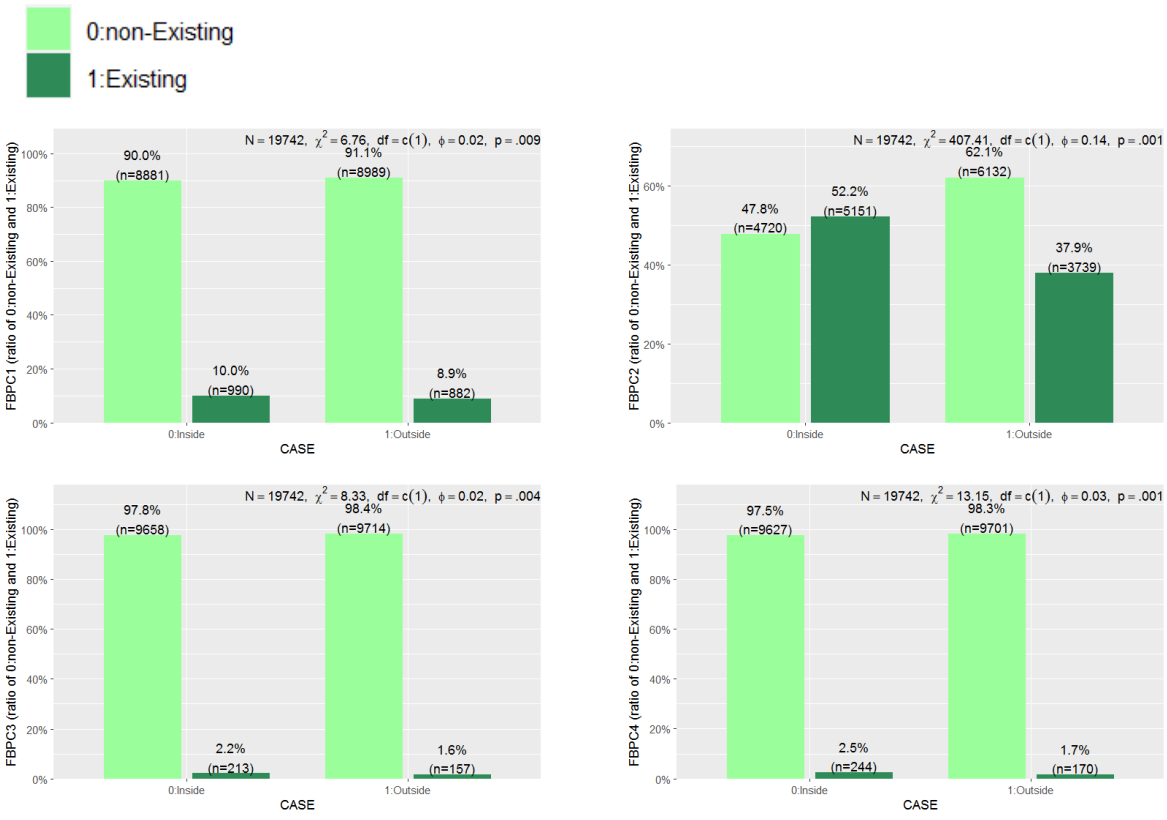

N = 45336, χ^2 = 0.21, df = c(1), ϕ = 0.00, p = .647
PreviousFire60 (ratio of 0:non-Existing and 1:Existing)
99.3% (n=22510)
0.7% (n=158)
99.3% (n=22519)
0.7% (n=149)
0:Inside
1:Outside
CASE

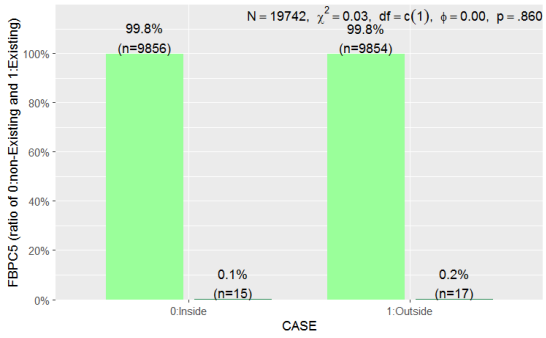


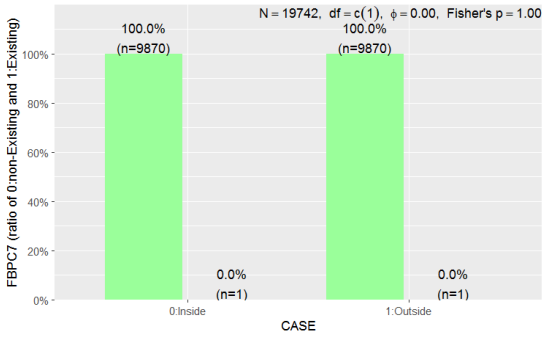


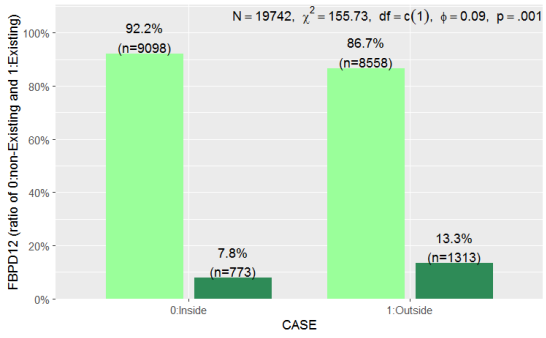
Appendix A2: Contingency Plot for Binary Variables in V700 Test Data

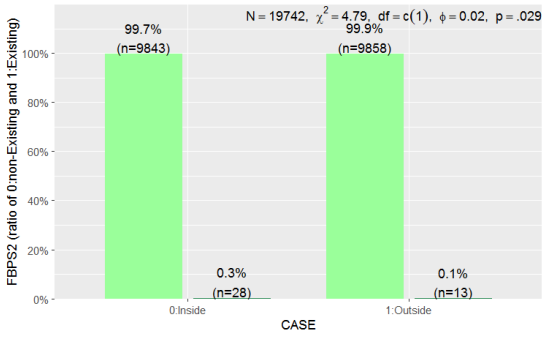
The number (n) and the percentage of non-existing (palegreen1) and existing (seagreen) data points for each of 29 variables inside (CASE=0) and outside (CASE=1) the fire perimeter for test data

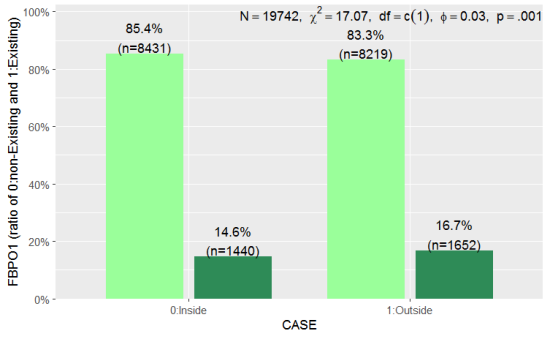


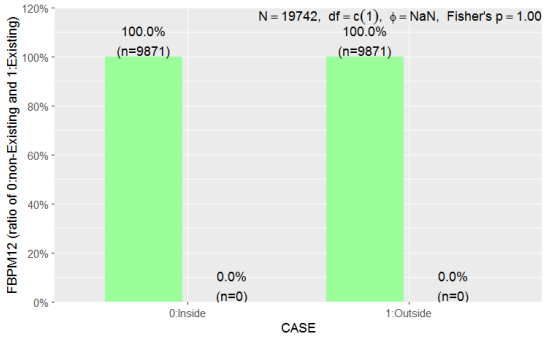

N = 19742, χ^2 = 0.03, df = c(1), ϕ = 0.00, p = .860
FBPC5 (ratio of 0:non-Existing and 1:Existing)
99.8% (n=9856)
0.1% (n=15)
99.8% (n=9854)
0.2% (n=17)
0:Inside
1:Outside
CASE

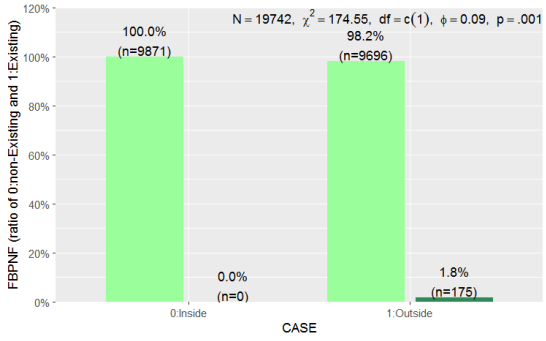

N = 19742, df = c(1), ϕ = 0.00, Fisher's p = 1.00
FBPC7 (ratio of 0:non-Existing and 1:Existing)
100.0% (n=9870)
0.0% (n=1)
100.0% (n=9870)
0.0% (n=1)
0:Inside
1:Outside
CASE

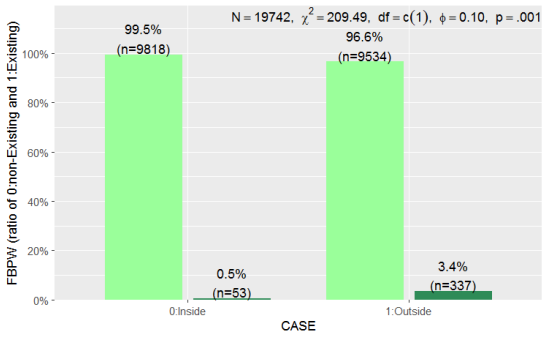

N = 19742, χ^2 = 155.73, df = c(1), ϕ = 0.09, p = .001
FBPD12 (ratio of 0:non-Existing and 1:Existing)
92.2% (n=9098)
7.8% (n=773)
86.7% (n=8558)
13.3% (n=1313)
0:Inside
1:Outside
CASE

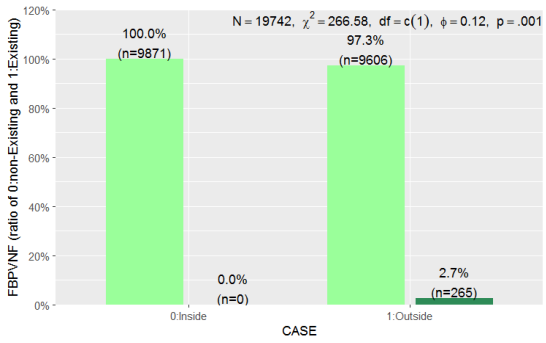

N = 19742, χ^2 = 4.79, df = c(1), ϕ = 0.02, p = .029
FBPS2 (ratio of 0:non-Existing and 1:Existing)
99.7% (n=9843)
0.3% (n=28)
99.9% (n=9858)
0.1% (n=13)
0:Inside
1:Outside
CASE

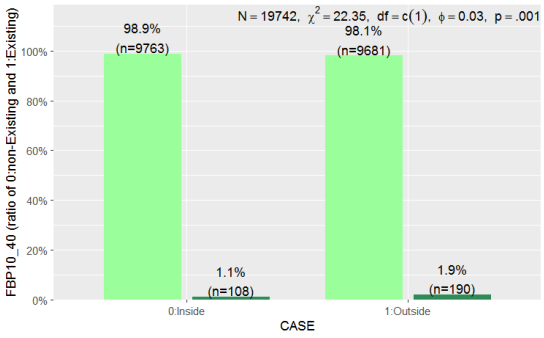

N = 19742, χ^2 = 17.07, df = c(1), ϕ = 0.03, p = .001
FBPO1 (ratio of 0:non-Existing and 1:Existing)
85.4% (n=8431)
14.6% (n=1440)
83.3% (n=8219)
16.7% (n=1652)
0:Inside
1:Outside
CASE

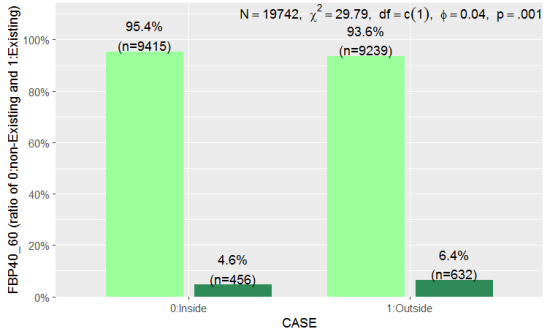

N = 19742, df = c(1), ϕ = NaN, Fisher's p = 1.00
FBPM12 (ratio of 0:non-Existing and 1:Existing)
100.0% (n=9871)
0.0% (n=0)
100.0% (n=9871)
0.0% (n=0)
0:Inside
1:Outside
CASE

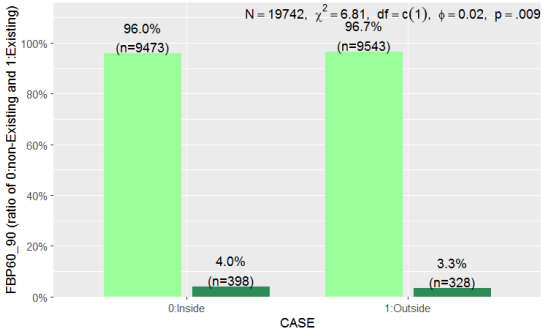

N = 19742, χ^2 = 174.55, df = c(1), ϕ = 0.09, p = .001
FBPNF (ratio of 0:non-Existing and 1:Existing)
100.0% (n=9871)
0.0% (n=0)
98.2% (n=9696)
1.8% (n=175)
0:Inside
1:Outside
CASE

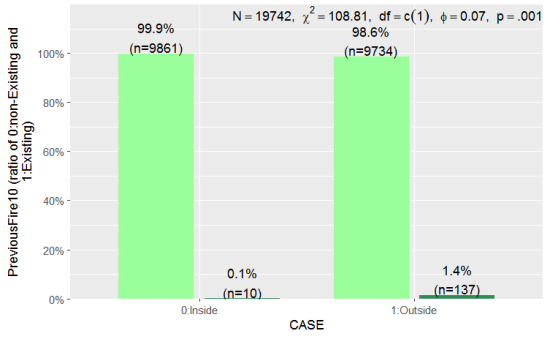

N = 19742, χ^2 = 209.49, df = c(1), ϕ = 0.10, p = .001
FBPW (ratio of 0:non-Existing and 1:Existing)
99.5% (n=9818)
0.5% (n=53)
96.6% (n=9534)
3.4% (n=337)
0:Inside
1:Outside
CASE

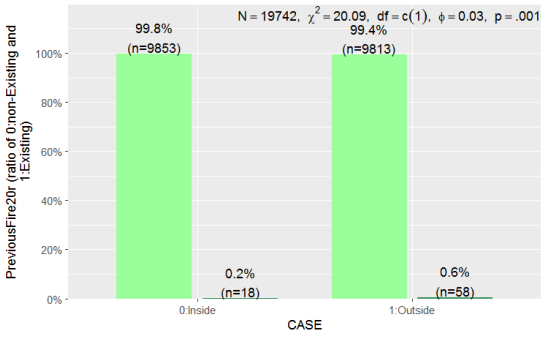

N = 19742, χ^2 = 266.58, df = c(1), ϕ = 0.12, p = .001
FBPVNF (ratio of 0:non-Existing and 1:Existing)
100.0% (n=9871)
0.0% (n=0)
97.3% (n=9606)
2.7% (n=265)
0:Inside
1:Outside
CASE

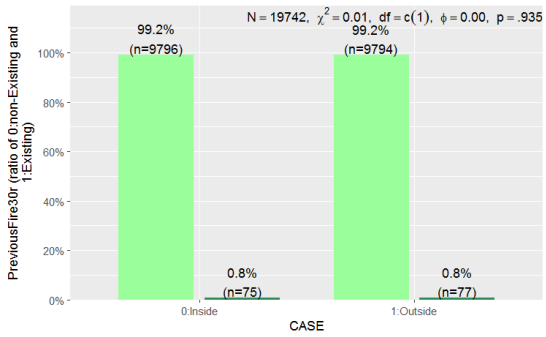

N = 19742, χ^2 = 22.35, df = c(1), ϕ = 0.03, p = .001
FBP10_40 (ratio of 0:non-Existing and 1:Existing)
98.9% (n=9763)
1.1% (n=108)
98.1% (n=9681)
1.9% (n=190)
0:Inside
1:Outside
CASE

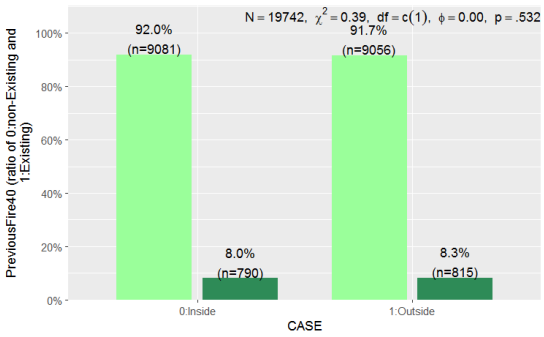

N = 19742, χ^2 = 29.79, df = c(1), ϕ = 0.04, p = .001
FBP40_60 (ratio of 0:non-Existing and 1:Existing)
95.4% (n=9415)
4.6% (n=456)
93.6% (n=9239)
6.4% (n=632)
0:Inside
1:Outside
CASE

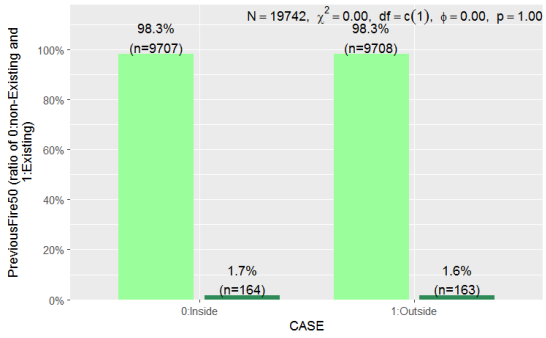

N = 19742, χ^2 = 6.81, df = c(1), ϕ = 0.02, p = .009
FBP60_90 (ratio of 0:non-Existing and 1:Existing)
96.0% (n=9473)
4.0% (n=398)
96.7% (n=9543)
3.3% (n=328)
0:Inside
1:Outside
CASE

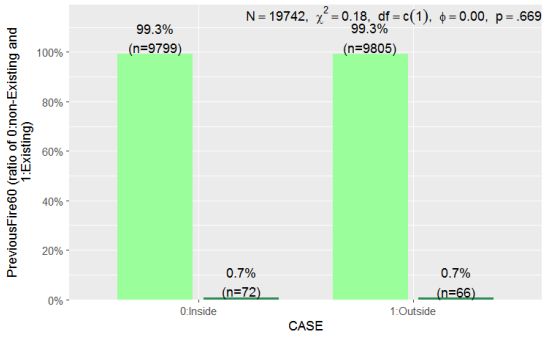

N = 19742, χ^2 = 108.81, df = c(1), ϕ = 0.07, p = .001
PreviousFire10 (ratio of 0:non-Existing and 1:Existing)
99.9% (n=9861)
0.1% (n=10)
98.6% (n=9734)
1.4% (n=137)
0:Inside
1:Outside
CASE

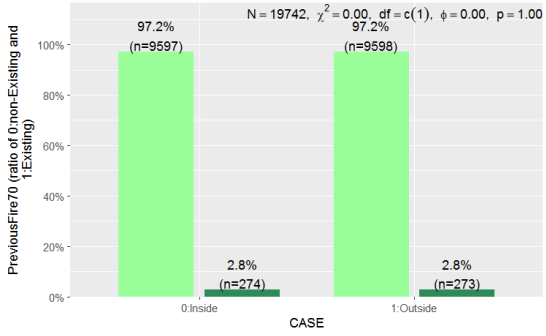

N = 19742, χ^2 = 20.09, df = c(1), ϕ = 0.03, p = .001
PreviousFire20r (ratio of 0:non-Existing and 1:Existing)
99.8% (n=9853)
0.2% (n=18)
99.4% (n=9813)
0.6% (n=58)
0:Inside
1:Outside
CASE

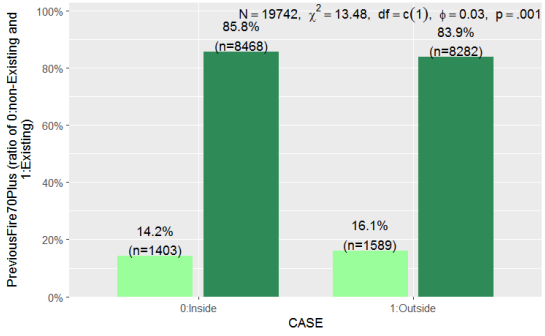

N = 19742, χ^2 = 0.01, df = c(1), ϕ = 0.00, p = .935
PreviousFire30r (ratio of 0:non-Existing and 1:Existing)
99.2% (n=9796)
0.8% (n=75)
99.2% (n=9794)
0.8% (n=77)
0:Inside
1:Outside
CASE


N = 19742, χ^2 = 0.39, df = c(1), ϕ = 0.00, p = .532
PreviousFire40 (ratio of 0:non-Existing and 1:Existing)
92.0% (n=9081)
8.0% (n=790)
91.7% (n=9056)
8.3% (n=815)
0:Inside
1:Outside
CASE


N = 19742, χ^2 = 0.00, df = c(1), ϕ = 0.00, p = 1.00
PreviousFire50 (ratio of 0:non-Existing and 1:Existing)
98.3% (n=9707)
1.7% (n=164)
98.3% (n=9708)
1.6% (n=163)
0:Inside
1:Outside
CASE


N = 19742, χ^2 = 0.18, df = c(1), ϕ = 0.00, p = .669
PreviousFire60 (ratio of 0:non-Existing and 1:Existing)
99.3% (n=9799)
0.7% (n=72)
99.3% (n=9805)
0.7% (n=66)
0:Inside
1:Outside
CASE


N = 19742, χ^2 = 0.00, df = c(1), ϕ = 0.00, p = 1.00
PreviousFire70 (ratio of 0:non-Existing and 1:Existing)
97.2% (n=9597)
2.8% (n=274)
97.2% (n=9598)
2.8% (n=273)
0:Inside
1:Outside
CASE


N = 19742, χ^2 = 13.48, df = c(1), ϕ = 0.03, p = .001
PreviousFire70Plus (ratio of 0:non-Existing and 1:Existing)
14.2% (n=1403)
85.8% (n=8468)
16.1% (n=1589)
83.9% (n=8282)
0:Inside
1:Outside
CASE