11-27-2020 2:00 PM

Statistical Methods with a Focus on Joint Outcome Modeling and on Methods for Fire Science

Da Zhong Xi, *The University of Western Ontario*

Supervisor: Dean, Charmaine B., *University of Waterloo*
Co-Supervisor: Woolford, Douglas G., *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Statistics and Actuarial Sciences

Abstract

Understanding the dynamics of wildland fires contributes significantly to the development of fire science. Challenges in the analysis of historical fire data include defining fire dynamics within existing statistical frameworks, modeling the duration and size of fires as joint outcomes, identifying how fires are grouped into clusters of subpopulations, and assessing the effect of environmental variables in different modeling frameworks. We develop novel statistical methods to consider outcomes related to fire science jointly. These methods address these challenges by linking univariate models for separate outcomes through shared random effects, an approach referred to as *joint modeling*. Comparisons with existing approaches demonstrate the flexibilities of the joint models developed and the advantages of their interpretations. Models used to quantify fire behaviour may also be useful in other applications, and here we consider modeling disease spread. The methodologies for fire modeling can be used, for example, for understanding the progression of Covid-19 in Ontario, Canada.

The key contributions presented in this thesis are the following: 1) Developing frameworks for modelling fire duration and fire size in British Columbia, Canada, jointly, both through modelling using shared random effects and also through copulas. 2) Illustrating the robustness of joint models when the true models are copulas. 3) Extending the framework into a finite joint mixture to classify fires into components and to identify the subpopulation to which the fires belong. 4) Incorporating the longitudinal environmental variables into the models. 5) Extending the method into the analysis of public health data by linking the daily number of Covid-19 hospitalizations and deaths as time series processes using a shared random effect. A key aspect of the research presented here is the focus on extensions of the joint modeling framework.

Keywords

Joint modeling, finite mixture model, time series, fire duration, fire size, Covid-19 data

Summary for Lay Audience

This thesis develops novel statistical techniques for analyzing data associated with fire science and disease modeling. In general terms, a mathematical model can be used to describe relationships observed in the real world. We create modeling frameworks in which different types of data (e.g. time to event occurrence and repeated environmental observations) can be incorporated into a single model.

Understanding how wildland fires grow contributes to the development of fire science. Some research areas we study include analyzing historical fire data to learn how they behave, and studying predictive variables such as the time to supress the fire and the area burned. We also consider variables such as seasonality, location, and weather, and the impact of these variables on fire behaviour.

We use a technique called joint modeling that allows the incorporation of multiple types of data into one model simultaneously, and we build on this approach to describe fire behavior. Using this approach, we show the effect of predictive variables on two outcomes, duration and size of fires. Models used to quantify fire behaviour may also be useful in other applications, such as modeling disease spread. The methodologies for fire modeling can be used, for example, for understanding the progression of an infectious disease. We apply our techniques developed for studying fire science to the study of Covid-19 in Ontario, Canada.

Co-Authorship Statement

Paper 1: Chapter 2

Paper title: Statistical models of key components of wildfire risk

List of authors: Dexen D. Z. Xi, Steve W. Taylor, Douglas G. Woolford, C. B. Dean

Publication:

Xi, D. D. Z., Taylor, S. W., Woolford, D. G., & Dean, C. B. (2019). Statistical Models of Key Components of Wildfire Risk. *Annual Review of Statistics and Its Application*, 6(1), 197–222. https://doi.org/10.1146/annurev-statistics-031017-100450

Author contributions:

Dr. Dean was invited to write a paper with the purpose of providing a review for the role that statistics and its application has played in the development of fire science. Dr. Woolford formulated the outline of the paper and wrote section 2. Fire scientist Steve Taylor wrote section 1, 3 and 7. Dr. Dean wrote section 8 and took lead in the editing. I wrote section 2, 5, and 6. I organized the development of the paper and took lead in its revision. All authors contributed to the preparation of the manuscript with regard to the content and relevance of the work.

Paper 2: Chapter 3

Paper title: Modeling the duration and size of extended attack wildfires as dependent outcomes

List of authors: Dexen D. Z. Xi, C. B. Dean, Steve W. Taylor

Publication:

Xi D. D. Z., Dean, C. B., & Taylor, S.W. (2020). Modeling the duration and size of extended attack wildfires as dependent outcomes. *Environmetrics*. 31(e2619). https://doi.org/10.1002/env.2619

Author contributions:

The topic of jointly modeling fire duration and fire size using shared random effects was initiated by Dr. Dean, Steve Taylor and the visiting scholar Dr. Giovani da Silva. Mr. Taylor provided the data and Dr. Silva carried out exploratory work. I proposed and formulated the model frameworks, then carried out the analysis. Dr. Dean and Mr. Taylor provided statistical guidance and scientific interpretation during the analysis. All authors contributed to the preparation of the manuscript regarding the content and relevance of the work.

Paper 3: Chapter 4

Paper title: Modeling the Duration and Size of Wildfires Using Joint Mixture Models

List of authors: Dexen D. Z. Xi, C. B. Dean, Steve W. Taylor

Publication:

Xi, D. D. Z., Dean, C. B., & Taylor, S. W. *Modeling the Duration and Size of Wildfires Using Joint Mixture Models*. Submitted for publication.

Author contributions:

After discussions with Dr. Reg Kulperger at my thesis proposal defense, I initiated the topic of developing a multivariate mixture model for fire duration and fire size through joint modeling. Dr. Dean proposed examining the estimated component probabilities while I proposed using a Dirichlet regression to assess the effect of covariates on the probabilities. Mr. Taylor and Dr. Khurram Nadeem prepared the fire data and I performed the analysis of the data. Dr. Dean and Mr. Taylor provided statistical guidance and scientific interpretation during the analysis. All authors contributed to the preparation of the manuscript with regard to the content and relevance of the work.

Paper 4: Chapter 5

Paper title: Joint Modeling of Hospitalization and Mortality of Ontario Covid-19 cases

List of authors: Dexen Xi, C. B. Dean, Elizabeth M. Renouf

Publication: In preparation

Author contributions:

The seriousness of the pandemic led to many discussions about how statistical science and my work on joint modeling could contribute to understanding aspects of the impact of Covid-19. The topic of analyzing the observed and underlying relationship between daily number of hospitalizations and deaths was initiated by Dr. Dean. I raised the idea of viewing the outcome processes as time series and conducting a cointegration analysis. Cointegration analysis was studied and applied in my M.Sc. summer project. Dr. Elizabeth Renouf and Dr. Georges Bucyibaruta collected data and provided suggestions in model building. I formulated the joint model and carried out the analysis. All authors contributed to the preparation of the manuscript with regard to the content and relevance of the work.

To Qi Lin and Yuan

“Where tree leaves dance... one shall find flames... the fire's shadow will illuminate the village... and once again tree leaves shall bud anew.”

– Hiruzen Sarutobi, the Third Hokage

Kishimoto, M. (2007). Naruto (Shonen jump manga ed.). San Francisco, CA: Viz.

Acknowledgments

Words cannot express my gratitude and appreciation to Charmaine Dean, my senior supervisor and academic mother; I could not have the events of today without having you in my life. Thank you to Charmaine Dean and Douglas Woolford for establishing the Wildland Fire Science Laboratory and for many helpful research discussions and activities. Thanks to Steve Taylor for bringing us interesting research projects. Our collaborations gave strong support to the firm steps statistics has made in the evolution of wildland fire studies. Thank you also to Giovani da Silva who guided me at the early stage of my thesis.

Thanks to all my lab members and colleagues through my past six years at Western, as well as people from the Department of Statistical and Actuarial Sciences. The sense of belonging and the recognition from you have been my motivation through this journey. Special thanks to my virtual lab members during this special time: Elizabeth Renouf, Georges Bucyibaruta, Wenyu Shen, and Catherine Tian. We have achieved more than ever expected from collaborating remotely. The True North strong, we stand on guard for thee.

I gratefully acknowledge the support of the Pacific Forestry Centre for providing the fire data for my research, the Natural Sciences and Engineering Research Council of Canada through its Discovery Grants program and its Research Support Fund.

I take this opportunity for expressing my deep love and appreciation to my parents and parents-in-law. Saving the best to the last, Qi Lin and Yuan, you are the amazing grace from God to my life.

Table of Contents

Abstract i

Summary for Lay Audience i

Co-Authorship Statement i

Acknowledgments ii

Table of Contents iii

List of Tables vi

List of Figures vii

List of Appendices xi

Chapter 1 1

1 Introduction 1

Chapter 2 4

2 Statistical Models of Key Components of Wildfire Risk 4

2.1 Introduction 4

2.2 Occurrence 9

2.2.1 A Point Process Viewpoint 9

2.2.2 Logistic Models as a Discretized Approach to Estimate a Point Process 10

2.2.3 Changes in Fire Occurrence 12

2.2.4 The Two Cultures of Fire Occurrence Prediction Modeling Statistical Modeling Versus Algorithmic Methods 15

2.3 Fire Spread, Intensity and Growth 15

2.4 Duration 17

2.5 Size 24

2.6 Modeling Duration and Size as Joint Outcomes 27

2.7 Event Sets and Burn Probability 30

2.8 Discussion 33

References *37*

Chapter 3 47

3 Modeling the Duration and Size of Extended Attack Wildfires as Dependent Outcomes 47

3.1 Introduction *47*

3.2 Modeling and Estimation of Joint Outcomes *49*

3.2.1 The Copula Model Framework ***52***

3.2.2 The Joint Model Framework ***55***

3.3 British Columbia Fire Data *58*

3.3.1 Data Description ***58***

3.3.2 Construction of Derived Covariates ***63***

3.4 Analysis and Results *68*

3.5 Robustness under Joint Modeling *72*

3.6 Discussion *78*

References *80*

Chapter 4 85

4 Joint Mixture Models for the Duration and Size of Wildfires 85

4.1 Introduction *85*

4.2 Modelling Frameworks *88*

4.2.1 Finite Mixture Joint Models ***88***

4.2.2 Finite Mixture Bivariate Model ***90***

4.2.3 The Four-Component Mixture Models ***90***

4.2.4 Dirichlet Model for the Effect of Covariates on Component Membership ***94***

4.3 British Columbia Fire Study *95*

4.3.1 Data Description ***95***

4.3.2 Parameter Estimates ***97***

4.3.3 Effect of Covariates in the Dirichlet Model ***101***

4.4 Discussion *110*

References *114*

Chapter 5 117

5 Joint Modeling of Hospitalization and Mortality of Ontario Covid-19 Cases 117

5.1 Introduction 117

5.2 Models and Methods 118

5.2.1 Cointegration Analysis 118

5.2.2 Joint Modeling 120

5.3 Results and Analysis 122

5.3.1 Ontario Data 122

5.3.2 Cointegration Analysis of Ontario Data 124

5.3.3 Joint Modeling of Ontario Data 126

5.4 Discussion 130

References 131

Chapter 6 133

6 Future Work 133

6.1 Future Work as Identified in the Articles Integrated to Form the Thesis 133

6.2 A Framework for Predicting Daily Fire Load 134

References 137

Appendices 138

Appendix 3A: Full Conditional Posterior Distributions of The Models Used in The Analysis 138

Appendix 3B: Comparison of Fit of Candidate Models Based on DIC and WAIC for the Fire Data 141

Appendix 4C: Full Conditional Posterior Distributions of the Models in the Analysis .. 142

Appendix 4D: Sensitivity to Priors in the FMJM 144

Appendix 4E: Comparison of Estimated Component Membership Probabilities from FMJM and FMBM 145

Appendix 4F: Additional Covariates Not Discussed in Detail in Section 3 146

List of Tables

Table 3.1: Distributions for the outcomes under the accelerated failure time (AFT) model 51

Table 3.2: Parameterization of the copulas. 53

Table 3.3: Parameterization the joint models.....56

Table 3.4: Data used in the study.....60

Table 3.5: Descriptive statistics of the covariates.....65

Table 3.6: Posterior estimates of model parameters and statistics accessing model fits of the three dependent models.....69

Table 3.7: Posterior estimates of the covariate effects for model 2a. 70

Table 3.8: The lower limit $Q.025$, median $Q.500$ and the upper limit $Q.975$ of the distribution of the joint model estimates for data generated from the normal copula with both margins as lognormal.....75

Table 4.1: Parameterization of the models considered in the fire science application 91

Table 4.2: Posterior estimates of model parameters 99

Table 4.3: Posterior estimates (exponentiated) of the covariate effects obtained from FMJM.....109

Table 5.1: Parameterization of the joint models 123

Table 5.2: Statistics assessing model fits for the candidate models.....127

Table 5.3: Posterior estimates of the model parameters 128

List of Figures

Figure 2.1: The risk triangle concept from the insurance and wildland fire perspectives. (a) The general risk triangle framework in insurance (adapted from Crichton, 1999). (b) The risk triangle concept as it applies to assessing wildland fire risk (modified from Scott, 2006). Risk in general, as well as in the context of wildland fires, can be viewed as having three connected components, as highlighted by the sides of the risk triangle. 6

Figure 2.2: Some factors contributing to wildfire hazard and risk are estimated with various qualitative, deterministic, and stochastic models. 7

Figure 2.3: Panels a and b plot the estimated component-specific fire occurrence probability curves (red lines) for the regular and extreme components, respectively. A fire day refers to a day when one or more wildland fires are reported. Overlaid on each of these curves are the observed empirical weighted proportion of the number of fire days per week over all years (red circles), where the observed data were weighted by the posterior probabilities of membership for the corresponding regular or extreme component. Panel c compares observed and expected frequencies of excess zeros. The light blue line is the expected number of zeros from the zero-heavy component plotted versus year. The blue circles are the empirical number of excess zeros: the number of observed zeros minus the number of zeros expected to arise from both the regular and extreme seasonal components. Adapted with permission from Woolford et al. (2014). ...14

Figure 2.4: Choropleth map of the frailty terms for lightning-caused fires in the former intensive fire management zone of Ontario, Canada, which was partitioned into a set of fire management compartments (FMCs). Each FMC polygon was assigned a heat map color based on the estimate of the latent effect of the FMC (i.e., the frailty). FMCs that are outside of the study region are white. Exponentiated values of posterior frailty estimates can be viewed as multiplicative factors on the hazard function of fire lifetimes. Negative estimates imply an increase in survival probability via a reduction in hazard rate. Adapted with permission from Morin (2014). 22

Figure 2.5: (a) Wildfire event sets can be generated with stochastic point process models and fire growth simulations of specified duration (ellipses used for illustration only; figure courtesy of Carol Miller, US Department of Agriculture Forest Service). Elements within a cell are homogenous respect to weather fuels and topography. The ellipses are individual fires, and the blue square represents a sampling point. (b) Monte Carlo methods have been used to map burn probability by simulating large event sets representing many thousands of potential outcomes in modeled landscapes, such as the Thompson-Okanagan region of southern British Columbia (Wang et al., 2016). 32

Figure 3.1: The locations of the fires (left) and a scatter plot of duration versus size, with a log base 10 scale on both axes (right). Fires are clustered around the Rocky Mountain Trench. Duration and size have a moderate positive dependence. 62

Figure 3.2: The estimated parametric and nonparametric survivor functions of the outcome. The lognormal distribution seems to fit both outcomes well. 62

Figure 3.3: The trajectories of the environmental variables for 30 randomly chosen fires. The threshold values are plotted in dashed lines. BUI = Buildup Index; DC = Drought Code; DMC = Duff Moisture Code; FFMC = Fine Fuel Moisture Code; FWI = Fire Weather Index; ISI = Initial Spread Index; PCP = precipitation; RH = relative humidity; TEMP = temperature; WIND = windspeed. 67

Figure 3.4: Residual diagnostics for the final models. For both duration and size, the standardized residuals are roughly normal with no significant outliers. The straight edges along the bottom of the points arise from the truncation at duration >2 days and size > 4 hectares. These features can be identified among all three final models (1n, 2a, 2m). 73

Figure 3.5: The lower limit, median, and the upper limit of the distribution of the joint model estimates of $\sigma 2$, γ, and σb(left panel) and $\mu 1$, $\mu 2$, $\beta 1$, $\beta 1$ (right panel) for data generated from the normal copula with both margins as lognormal. As the association parameter τ increases, the shared variability increases. The distribution of the estimates of γ and $\sigma 2$ becomes narrower, whereas that for σb remains about the same. Joint models also provide robust location parameters and coefficient estimates. 77

Figure 4.1: The data and the estimated distributions of fire duration and fire size. The top row contains the estimated marginal distributions of the outcomes, overlaid on their histograms, with duration on the left panel and size on the right. The marginal distributions of duration and size are both captured by a narrowly spread normal component and a widely spread extreme component and seem to provide reasonable fits. Fires that are normal or extreme in both outcomes tend to have outcomes correlated. 98

Figure 4.2: The probability estimates by fire centre are presented in violin plots. The plots show the posterior estimate of the probability of each fire belonging to component 1 to 4 for each of the fire centres. The medians of the probabilities displayed in the violin plots demonstrate clear variation over fire centres for the extreme duration components (component 3, displayed in blue and component 4, displayed in red). 104

Figure 4.3: Posterior estimates of the probability of each fire belonging to component 1 to 4 by month. October data are combined into September because of its small number of observations. The seasonality of the fire behavior displays different patterns depending on component. The months of August through October are associated with a much higher risk of fires being extreme in duration and size. On the other hand, fires of extreme size and normal duration tend to occur in May. Fires of extreme duration and normal size are more likely to present at the end of the season than at the beginning.....105

Figure 4.4: The posterior probability estimates (y-axis) by the ADFT of average wind speed (km/h, x-axis) measured over a 10-minute period on the two left panels and the ADFT of the amount of rain (mm, x-axis) accumulated in the 24-hour period from noon to noon on the two right panels. As wind speed increases, the probability of being identified in the component corresponding to normal duration and extreme size increases, while as precipitation increases, the probability of being identified in the component corresponding to extreme duration and extreme size decreases. 106

Figure 4.5: The posterior probability estimates (y-axis) by the ADFT of DC (x-axis) on the two left panels and by DMC (x-axis) on the two right panels. As the ADFT of DC increases, the probability of fires with short duration (components 1 and 2) decreases and the probability of fires having long duration tends to increase (components 3 and 4). This

suggests that exceedance in temperature and shortage of precipitation will increase the containment time of the fire. As the ADFT of DMC increases, the probability of being identified as normal size components (component 1 and 3) decreases while the probability of being identified as extreme size components (component 2 and 4) increases. 107

Figure 5.1: The left panel illustrates the logarithm of the cumulative number of hospitalizations 6 days prior (black) and the cumulative number of deaths (red). Hospitalizations and deaths grow with a decreasing rate over time. The right panel plots the daily number of these quantities and their residuals (blue) against time. The processes are identified as having a long-term correlation through the cointegration analysis described in the text. 125

Figure 5.2: Posterior estimates of the shared random effect (left panel) and the estimated joint distribution (right panel) of the outcomes, $\boldsymbol{y1}$ and $\boldsymbol{y2}$ being the first order difference of hospitalizations six days prior and deaths, respectively. The posterior estimates of the shared random effect have a peak on May 02, reflecting the peak in daily hospitalizations. The estimated joint distribution of the outcomes demonstrates a weak dependence between the outcomes. 129

List of Appendices

Appendix 3A: Full Conditional Posterior Distributions of The Models Used in The Analysis.....138

Appendix 3B: Comparison of Fit of Candidate Models Based on DIC and WAIC for the Fire Data.....141

Appendix 4C: Full Conditional Posterior Distributions of the Models in the Analysis ..142

Appendix 4D: Sensitivity to Priors in the FMJM 144

Appendix 4E: Estimated Component Membership Probabilities of FMJM and FMBM 145

Appendix 4F: Additional Covariates Not Discussed in Detail in Section 3 146

Chapter 1

1 Introduction

Scientific studies often consider different types of outcomes obtained from the same individual. These outcomes of interest are usually modeled by current understanding of the scientific principles from which the outcomes arise, and as well, their distributions can be modeled statistically and empirically when historical records are available. These kinds of studies are commonly seen in biometrics, environmetrics, econometrics, and other fields of science, with the main purpose of understanding, if any, the relationship between the outcomes. For example, in biostatistics, the progression of CD4 (i.e. cluster of differentiation 4 counts, a longitudinal biomarker measuring white blood cells of a patient in AIDS research) and lifetime are outcomes of different types that are often studied together using statistical models. The purposes of such a study are (i) to understand the within-subject pattern of CD4 and (ii) to characterize the relationship between CD4 and the lifetime.

Statistical models also play a similar role in the development of wildland fire science. Fire danger systems have evolved from qualitative indices, to process-driven deterministic models of fire behavior and growth, to data-driven stochastic models of fire occurrence and simulation systems. However, there has often been little overlap or connectivity in these frameworks, and validation has not been common in deterministic models. Examples of validation approaches for such deterministic models are the use of expert intuition, the contrast with real system measurements, and comparisons with theoretical analysis.

Yet, marked increases in annual fire costs, losses, and fatality costs over the past decade draw attention to the need for better understanding of fire risk to support fire management decision making through the use of science-backed, data-driven tools. Contemporary risk modeling systems provide a useful integrative framework. Chapter 2 discusses a variety of important contributions for modeling fire risk components over recent decades, certain key

fire characteristics that have been overlooked, and areas of recent research that may enhance risk models.

Understanding the complex relationship between the duration and size of forest fires is important in order to better predict these key characteristics of fires for fire management purposes in a changing climate. Describing this relationship is also important for our fundamental understanding of fire science. In Chapter 3, we develop and utilize novel techniques for characterizing the distribution of multiple outcomes related to a specific event, placed in the fire science context. In this framework, we jointly model time spent (duration), in days, and area burned (size), in hectares, from ground attack to final control of a fire as a bivariate survival outcome using two broad methodologies: a copula model that connects the two outcomes functionally, and a joint modeling framework that connects the two outcomes with a shared random effect. We compare these two methodologies in terms of their utility and predictive power. We also consider how longitudinal environmental variables (e.g. precipitation, drought indices) are best incorporated in this context, and challenges related to the complexity of computation associated with the analysis of two outcomes considered jointly.

As well, fire behaviour, linked to hidden effects, tends to yield that fires arise from different subpopulations. Indeed, it is not unusual for fire behaviour to be identified as arising from either normal or extreme subpopulations, for example. In Chapter 4, we embed these two concepts into a new framework for jointly modeling fire duration and fire size. We develop a bivariate finite mixture framework that can be used to model duration and size with four subpopulations of the outcomes whereby duration and size are either normal or extreme. We utilize a shared random effect model as well as a bivariate Gaussian mixture model for such mixture modeling. We also incorporate the effect of explanatory variables associated with each fire event, on the posterior probability of the component that the fire belongs to, through a Dirichlet model. In an analysis of fire outcomes from British Columbia, Canada, we find that the majority of the fires are of normal or extreme magnitude in both outcomes, with strong evidence indicating correlation between duration and size. The effect of fire centre, month, and several environmental covariates are identified as key predictors and we

are able to determine through these approaches how these covariates differentially affect the four subpopulations.

The concepts of joint modeling developed in the previous chapters can be applied to a wide variety of settings. Given the current focus on pandemic modeling, we also consider its utility in modeling public health data related to Covid-19. Daily number of hospitalizations and deaths are key outcomes in quantifying the outbreak of infectious diseases. For the purposes of understanding the trend of the processes and the effect of observations from previous days, it may be useful to consider time series approaches for modeling the outcomes. Using such an approach, cointegration analysis may be employed to identify the long-run relationship between those multiple processes that are key to understanding trends in infectious disease such as hospitalization and death. As an alternative perspective, relationships between outcomes can be modeled through a shared latent stochastic error term; in Chapter 5, we propose a novel framework to study the underlying correlation between two time series processes through this method called joint modeling. In our Ontario Covid-19 study, a cointegration analysis utilizes statistical tests to identify the long-run relationship between the daily number of new hospitalizations six days prior and the daily number of new deaths in Ontario. Additionally, a joint autoregressive model provides a framework to model the underlying correlation between the processes.

The remainder of the thesis is organized as follows: Chapter 2 provides the background on statistical models developed in fire science as they relate to the work contained in the thesis. The joint modeling frameworks for fire duration and fire size, as well as the analysis of historical fires in British Columbia, Canada are presented in Chapter 3 and Chapter 4. Chapter 5 provides the framework to jointly model daily number of hospitalizations and deaths for Covid-19 studies. Chapter 6 concludes the thesis with a discussion of future work related to the research presented here. The thesis is organized in the integrated-article format with the chapters treating discrete but related problems (see https://grad.uwo.ca/administration/regulations/8.html#8321, Section 8.3 The Thesis Preparation and Format).

Chapter 2

2 Statistical Models of Key Components of Wildfire Risk

2.1 Introduction

The global average annual area burned due to wildfires was recently estimated (Giglio et al., 2013) to be approximately 3.48 million km^2 for the 1997–2011 period, about the area of India. Wildfire characteristics such as the number of fires, their size, their severity, the season during which they occur, and the annual area burned in a region vary considerably with climate, vegetation, topographic controls, and human influence at local (Heyerdahl et al., 2007), regional (Parks et al., 2012), and global (Krawchuk et al., 2009) scales. Fires are moderately rare events at a daily scale. For example, while about 5,000 fires occur in Canada every year, this translates to a background rate of approximately one new fire per ten million hectares per day during an approximately 5-month fire season. However, this may be punctuated by surges in the number of fire ignitions associated with high pressure systems or lightning storms at local or regional scales, resulting in many dozens to hundreds of fire starts being discovered within a few hours to days. Consumption of biomass, smoke emissions, and changes in land cover associated with vegetation fires have an important influence on global atmospheric chemistry, the global carbon budget, and energy balance, as well as the structure and function of affected ecosystems (Ryan, 1991; GLOBAL, 2013). As well, unwanted fires may also cause loss of life and property, impacts on air quality and human health, and loss of business revenue.

The field of fire science evolved over approximately 100 years from early descriptive studies (e.g., Plummer, 1912) to the development of complex models of fire spread and other physical processes (e.g., Linn et al., 2007). Research has followed two streams: basic research to enhance understanding of wildfire as an ecological process, and applied research to inform fire management decision making. Contemporary fire management organizations follow the four pillars of emergency management: prevention and mitigation, planning and preparedness, response, and recovery. Thus, to inform preparedness and

response actions, wildfire managers would like know, at a daily to weekly scale, how many fires will likely occur, whether and how fast they will spread, how intense and how large they will grow, and how long they will last. To inform prevention and mitigation activities, they would also like to know the long-term likelihood of a vegetated area burning. Because of the close connection between weather and fire, much early effort was devoted to the development of fire danger rating systems to predict fuel flammability, fire occurrence, and fire behavior in different vegetation types with changing weather conditions to support preparedness and suppression response decision making (Taylor & Alexander, 2006; Hardy & Hardy, 2007; Fujioka et al., 2008). An important historical development is that independent systems have been developed to portray fire danger in different countries; no single global fire danger system has emerged. Examples include the Canadian Fire Weather Index (FWI) System, a subsystem of the Canadian Forest Fire Danger Rating System; the National Fire Danger Rating System in the United States; and the McArthur Forest Fire Danger Index in Australia.

Because wildfire is a natural process that cannot be completely eliminated from some environments, even where unwanted, fire management is increasingly being recognized as a form of risk management. Natural hazard risk, the expected loss or impact arising from a natural event, is considered to have three components: hazard, vulnerability, and exposure (Cardona et al., 2012), which can be visualized as a risk triangle (Crichton, 1999). It is important to note that this is not a statistical representation of risk but one that has been developed and utilized by the natural hazards community. We present it here because of its common usage in environmental science. Scott (2006) adapted this concept, defining the wildland fire risk triangle as including the three components: fire probability (i.e., the hazard or the risk of fire occurrence), fire behavior (i.e., the severity or potential behavior of a fire if it occurs), and fire effects (i.e., the exposure or potential impact of the fire). Statistical science has made many contributions to modeling some of the components of this risk triangle, as we discuss later. Figure 2.1 illustrates the concept of the general risk triangle and its adaptation to wildland fire risk.

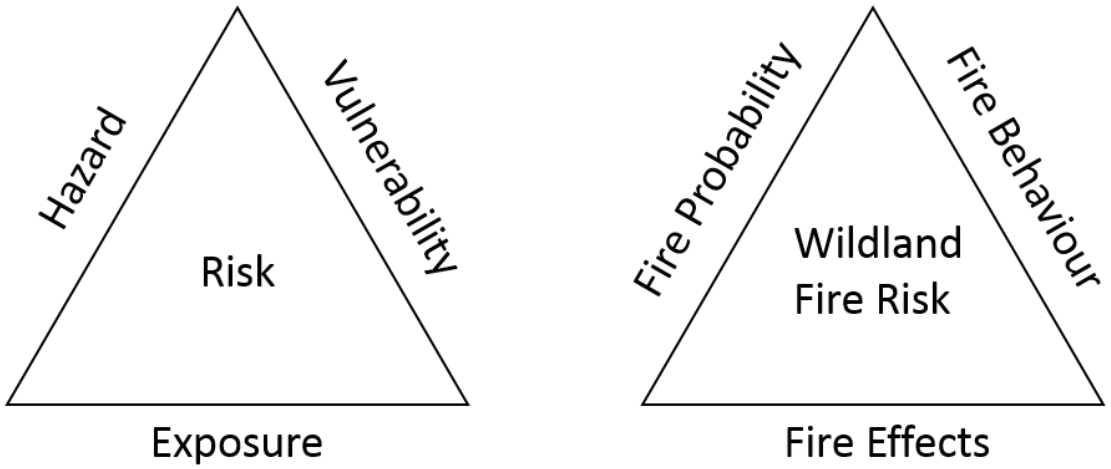


Figure 2.1: The risk triangle concept from the insurance and wildland fire perspectives. (a) The general risk triangle framework in insurance (adapted from Crichton, 1999). (b) The risk triangle concept as it applies to assessing wildland fire risk (modified from Scott, 2006). Risk in general, as well as in the context of wildland fires, can be viewed as having three connected components, as highlighted by the sides of the risk triangle.

Consequently, there has been increasing focus on the development of quantitative risk analysis methods (Miller & Ager, 2013). Quantitative risk assessment to inform management decision making has its foundations in decision theory and utility theory (Morgan et al., 1992). Figure 2.2 illustrates a number of factors that contribute to wildfire risk, including the likelihood and severity of fires in a region and the exposure, vulnerability, and value of valued assets. Finney (2005) defined wildfire risk as the expected change in net present value obtained from the aggregate losses and benefits in n values or assets over all N possible fire behaviors (under all weather conditions from all ignition locations):

$$\sum_{i=1}^{N}\sum_{j=1}^{n} p(F_i)\left[B_{ij} - L_{ij}\right],$$

where $p(F_i)$ is the probability of the i th fire behavior, and B_{ij} and L_{ij} are the benefits and losses resulting from the effects of the i th fire behavior on the j th asset type, respectively.

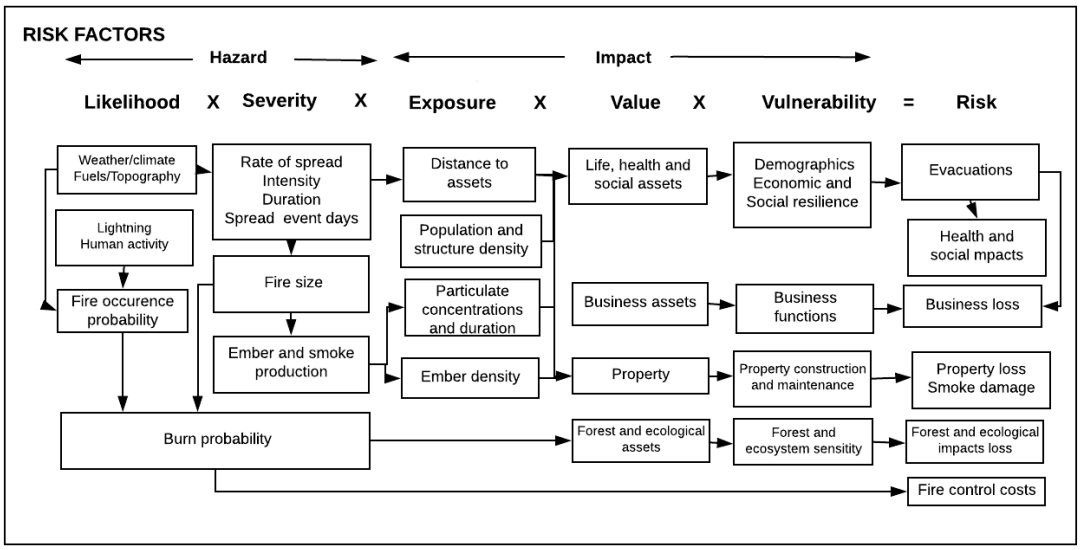


Figure 2.2: Some factors contributing to wildfire hazard and risk are estimated with various qualitative, deterministic, and stochastic models.

More recently, Papakosta et al. (2017) defined wildfire risk to the j th asset as

$$\int_{\text{Hazard scenarios h}} f_H(h) \int_{\text{Damage scenario d}} f_{D_j|H}(d|h)C_j(d,h)\mathrm{d}d\mathrm{d}h,$$

where $f_{D_j|H}(d|h)$ is the conditional density of damage d given a wildfire event h (vulnerability), $C_j(d,h)$ is the cost associated with the damage by the wildfire, and $f_H(h)$ is the probability density of a wildfire event. In a further refinement of the model, the density $f_H(h)$ may be related to a certain fire severity characteristic, for example, the likelihood of a particular type of fire, fire intensity, duration, size, or incident complexity. These key characteristics represent positive continuous outcomes (or marks in the context of a point process; e.g., Daley & Vere-Jones, 2003).

The precise definition of each fire severity characteristic may vary in different contexts. As an example, the characteristic may be the size being larger than some specific value, for instance, a fire being class classified as a large fire as defined by Stocks et al. (2002). In this case, the risk on assets relates to risks especially for large fire events. Alternatively,

a certain epoch of the fire's duration may be of interest (e.g., Morin et al., 2015). In that case, the risk on assets refers to the risk during this epoch of the fires. Note that this framework can be further extended by decomposing the fire characteristic into components, such as decomposing fire size into the probability of a large fire given fire occurrence and then modeling the size distribution for large fires. This approach was utilized in Preisler et al. (2011), where size was decomposed even further by coupling to a model for cost per acre in order to forecast future suppression costs.

Statistical science has an important role in modeling and quantifying uncertainties in the various components that make up wildland fire risk, through such conditional and marginal models, in order to better understand wildland fire science and inform wildland fire management (Preisler & Ager, 2013; Taylor et al., 2013). The latter review also includes a thorough commentary on the history of statistical modeling of fire occurrence, starting with the pioneering work of Bruce (1963) and Cunningham & Martell (1973), followed by the seminal work of Brillinger et al. (2003) and Preisler et al. (2004) that led to substantial developments over the next decade.

This chapter reviews key statistical models that have been used to predict wildfire risk components, including some very recently developed novel modeling strategies. Section 2 reviews models for fire occurrence prediction, while Section 3 discusses deterministic models for fire spread, intensity, and growth. Section 4 presents models for fire duration, and Section 5 examines models for estimating fire size. Sections 2 through 5 discuss past research investigating each of those characteristics of fire regimes separately or through the conditional framework as discussed above. In Section 6, we introduce the use of joint modeling methods using duration and size as an illustration, which we believe has the potential for gaining further insight into fire behavior. In Section 7, we turn to an alternative conditional framework for modeling fire hazard and its utilizations through computer simulations. We conclude with a discussion.

2.2 Occurrence

2.2.1 A Point Process Viewpoint

It is important to note that not all wildland fire ignitions may appear in fire management agency records (Taylor et al., 2013). Ignitions that lead to sustained fire spread may be detected by wildland fire management agencies (e.g., aerial detection or stationary towers), by the public, or by satellite-borne sensors. Detected fires that are subsequently reported and then recorded by a fire management agency are referred to as fire occurrences. Observed patterns in fire occurrences can be viewed as realizations of a spatio-temporal point process. Examples of applying methodology from the point process literature include Podur et al. (2003), Wang & Anderson (2011), and Turner (2009). The spatio-temporal point process underlying the generation of fire occurrence is denoted $N(s_1, s_2, t)$, where x and y are location variables and t is time. Since the rate of wildfire occurrences depends on environmental conditions favorable for ignition, the presence of an external ignition source, and detection capability, the point process can be assumed to have an inhomogeneous conditional intensity function λ that depends on parameter $\theta = \theta(z)$, where z is a vector of such predictors. The log-likelihood of the spatio-temporal point process is

$$L(\theta) = \int_0^T \int_{s_1} \int_{s_2} log[\lambda(s_1, s_2, t|\theta)] \, \mathrm{d}\, N(x, y, t) - \int_0^T \int_{s_1} \int_{s_2} log[\lambda(s_1, s_2, t|\theta)] \, \mathrm{d}\, x \mathrm{d}y \mathrm{d}t.$$

A discretized approach to approximating this likelihood has been the preferred framework for modeling fire occurrences with the underlying conditional intensity function approximated by a Bernoulli probability of a fire occurrence; the response and covariates are recorded on a set of discrete space-time voxels, chosen to be at a fine enough scale so that the counts of fire occurrence are reduced to presence/absence of a fire occurrence in any given voxel. A common scale for dividing space-time is 1 km × 1 km by daily cells (voxels). Dynamic covariates, such as weather and fire weather indices, or lightning counts

for lightning-caused fires, are interpolated to the centroid of that voxel. Static covariates, including measures of key predictors related to human-caused fires, such as measures related to roads, railways, or population density, are integrated over each voxel. For more details on the connection between the underlying spatio-temporal point process likelihood function and discretized approximations in the context of modeling fire occurrence and an example of such a model, readers are directed to Brillinger et al. (2003) or the review discussion in Taylor et al. (2013).

2.2.2 Logistic Models as a Discretized Approach to Estimate a Point Process

Using the discretized approach, the most widely employed method for modeling fire occurrence appears to be logistic regression or related extensions such as logistic generalized additive models (GAMs); sometimes models with random effects are also considered. Separate models, stratified by the cause of the fire, are commonly developed due to differences between the underlying processes generating the different types of ignitions. For example, different types of ignition sources can lead to different lag periods between the ignition of a fire and its eventual arrival to a fire management agency as a reported wildfire. This was reflected in the set of models characterizing the probabilities of ignition and eventual arrival (i.e., occurrence) of lightning fires in Ontario as developed by Wotton & Martell (2005). Lightning ignitions and their subsequent arrivals as reported forest fires are modeled separately. Then, the probability of a lightning strike igniting a fire at time s and that fire being reported at time $s + t$ is calculated by

$$\mathrm{P}(\text{lightning strike at time } s \text{ leads to a lightning fire occurrence at time } s+t) = \mathrm{P}(\text{occurrence at time } s+t \mid \text{ignition at time } s)\mathrm{P}(\text{ignition at time } s)\ .$$

There are also commonly highly nonlinear relationships between the probability of fire occurrence and other predictors, such as for seasonality or spatial effects. These nonlinear relationships on the log-odds scale are commonly modeled by spline-based smoothers using logistic GAMs. Wood (2006), for example, provides a general discussion of GAMs, and

Preisler & Ager (2013) provide a high-level overview of GAMs in the fire occurrence context. Prior to the introduction of GAMs, nonlinear seasonality components were modeled using periodic functions (e.g., Martell et al. 1989).

Let $Y_i, i = 1, \ldots, n$ be a set of random variables representing an indicator for fire occurrence (Yes = 1, No = 0) in voxel i, assumed to be independently distributed as $\text{Bernoulli}(p_i)$, conditional on observed covariates. Here, $p_i = P(Y_i = 1|x_i)$ where x_i denotes a vector of covariates for the i th voxel. We may model p_i through

$$\text{logit}\, p_i = \beta_0 + \sum_{p=1}^{P} g_p\left(x_{ip}\right),$$

where $\text{logit}\, p_i = \log\left(\frac{p_i}{1-p_i}\right)$, β_0 is an intercept, $x_{ip}, p = 1, \ldots, P$ are covariates, and g_p are corresponding zero-mean smoothers of these covariates. The terms in the model may include multidimensional smoothers [e.g., $g(s,t)$, where $s_i = (s_{i1}, s_{i2})$ represents the location and $t_i = (t_{i1}, t_{i2})$ represents day of year and year] to model spatial and temporal effects, where the latter allows for trends both within (e.g., seasonality) and across (e.g., climate change or other trends) years, as well as other nonlinear and/or linear effects of other key predictors, such as measures of fuel moisture and human-land use characteristics.

As noted by Woolford et al. (2011), the volume of data can present computational difficulties when modeling on a fine spatio-temporal scale such as the discretized space-time voxel approach as outlined previously. For example, in their case study of the Romeo Malette Forest in Ontario, Canada, Woolford et al. (2011) noted that discretizing the data to a set of 1 km × 1 km × daily voxels led to nearly 90 million records. For larger-scale studies, such as developing provincial or national modeling frameworks, this problem compounds immensely. Interestingly, however, the solution to this problem lies at the heart of one of the underlying dogmas of statistical inference: Rather than trying to fit a model to all data, a representative sample is used for model fitting. Since fires are an moderately rare event on any fine space-time scale, a response-dependent sampling scheme is

commonly employed, where all the voxels with fire occurrences are kept, along with a simple random sample of the nonfire voxels.

From a decision-support point of view, a key contribution of spatio-temporal fire occurrence modeling is that it produces a relative occurrence probability map where cells with higher probability of fire occurrence are identified, which can aid decision support, such as aerial detection routing. The expected number of fire occurrences in a given region on a given day can then be estimated. We also note that it is common to achieve greater specificity (correct identification of cells without fire occurrences) than sensitivity (correct prediction of cells with fire occurrences) because of the stochastic nature of the ignition process and because an overwhelming number of voxels refer to nonfire day and areas. The mathematical details of this framework are discussed in depth by Brillinger et al. (2003) and Taylor et al. (2013). The latter also summarizes the connections between this technique and logistic retrospective case-control studies.

2.2.3 Changes in Fire Occurrence

Whether and where fire occurrence is changing with climate is of considerable interest to fire managers. For example, fire occurrence has been shown to be increasing under a warming climate in the western United States (e.g., Westerling et al. 2006). Observed increases in fire occurrence have been found to be associated with anomalies in fire weather indices (Woolford et al. 2014). The fire season as measured by fire occurrence probability has been getting longer (Woolford et al. 2010; Albert-Green et al. 2013) and, based on results from studies using global climate model data under various scenarios, fire occurrence probability is predicted to increase under a warming climate, (e.g., Krawchuk et al. 2009; Wotton et al. 2003, 2010).

As commented on by Taylor et al. (2013), difficulties with historical analyses (e.g., Woolford et al., 2010; Albert-Green et al., 2013) arise because fire detection system effectiveness can change over time, leading to potential confounding with any climate change effect. Woolford et al. (2010) found that the median size at detection for lightning-

caused fires had been decreasing over the 42 years of their study, suggesting that the detection system may have become more effective over time, under the assumption that a fire would continue to grow after ignition.

Woolford et al. (2010) noticed three dominant characteristics in the lightning-caused fire occurrence records they analyzed, namely, regular seasonal patterns (as commonly quantified in other fire occurrence work such as Martell et al. (1989); Brillinger et al. (2003); Preisler et al. (2004); Woolford et al. (2009, 2011)); large deviations from these patterns, including zero-heavy behavior where no fires were observed even though fires were typically observed around such a period; and extreme behavior where many more fires were observed than what is typical.

The mixture framework of Woolford et al. (2014) identified significant increases to the lightning-caused fire occurrence probability that were associated with temperature and fire-weather index anomalies. Their study monitored long-term trends in a set of historical fire records for the period 1963–2009 for a region of northwestern Ontario, Canada, using a three-component mixture of logistic GAMs with the component densities representing seasonal, zero-heavy, and extreme behavior as discussed above. They noted that potential confounders, such as improved wildland fire detection systems, make it difficult to tease out climate change trends and that longer than a half-century of historical records is required to have strong confidence in correctly concluding significance in such a study monitoring temporal changes to fire occurrence. They also noted that determining power to detect trends in fire occurrence probability as a function of the number of years in the historical records was a "key, yet commonly overlooked, point in many quantitative scientific investigations of trends that may be related to climate change." (Woolford et al., 2014, p. 407) For their study, 47 years of fire records yield a power of 20% to detect trend changing. This power evaluation offers an opportunity to consider how many years of records are required to detect changes in environmental effects with high confidence.

Regardless of the underlying model and study, goodness-of-fit checking is a key step in any model-building framework. Typically, the goodness-of-fit logistic-based models can

be assessed by comparing observed counts versus those expected under the model where such counts are aggregated on various scales. Ideally, cross-validation (Wood, 2006; James et al., 2013) is also used to assess predictive accuracy. Vilar et al. (2010) provide simple examples of this in the context of fine-scale spatio-temporal fire occurrence prediction. However, assessing goodness of fit in the mixture modeling framework is more complicated. Woolford et al. (2014) examined goodness of fit by comparing observed versus expected counts for each subcomponent of their mixture model. This is illustrated in Figure 2.3. Such an assessment approach offers the opportunity to determine which of the subcomponents are not appropriate and could find wide application in other mixture modeling frameworks.

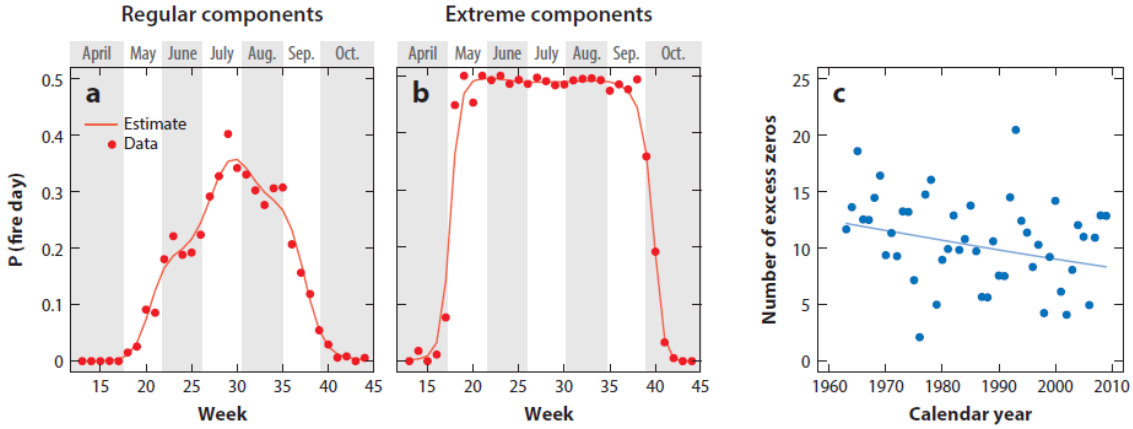


Figure 2.3: Panels a and b plot the estimated component-specific fire occurrence probability curves (red lines) for the regular and extreme components, respectively. A fire day refers to a day when one or more wildland fires are reported. Overlaid on each of these curves are the observed empirical weighted proportion of the number of fire days per week over all years (red circles), where the observed data were weighted by the posterior probabilities of membership for the corresponding regular or extreme component. Panel c compares observed and expected frequencies of excess zeros. The light blue line is the expected number of zeros from the zero-heavy component plotted versus year. The blue circles are the empirical number of excess zeros: the number of observed zeros minus the number of zeros expected to arise from both the regular and extreme seasonal components. Adapted with permission from Woolford et al. (2014).

2.2.4 The Two Cultures of Fire Occurrence Prediction Modeling Statistical Modeling Versus Algorithmic Methods

Breiman (2001) noted that the objective of a statistical analysis is to use data to make inferences, observing that the two dominant cultures for doing so are "statistical" and "algorithmic," where the latter focuses on finding a function to predict the response as a function of other variables without assuming a specific stochastic model. The preceding subsections have focused on summarizing key developments from a statistical modeling standpoint. However, algorithmic modeling approaches have also been used in the context of fire occurrence prediction.

For example, Garcia et al. (1996) developed artificial neural networks for wildfire occurrence and compared them to those of Garcia et al. (1995), who had utilized logistic regression models to analyze the data. This early study of fire occurrence using algorithm methods found reasonable predictive accuracy with neural nets; it correctly predicted 85% of the nonfire days and 78% of the fire days. They also commented that the improvement in predictions over traditional logistic regression modeling results were "not as dramatic as it has been in other applications" **(**Garcia et al., 1996, p. 14**)** that compared neural nets to logistic regression methods. The total percentage correctly predicted by the neural net model was found to improve by only 2% when compared with the logistic regression model for the same independent validation data set. They postulated that this could be due to the limited amount of data used in these studies (only five fire seasons). Ongoing research for large fire prediction in Canada (e.g., Nadeem et al., 2016) is exploring extensions to that work using lasso logistic regression, as well as algorithmic approaches, such as random forests. Hastie et al. (2011) and James et al. (2013) provide details on lasso and random forests.

2.3 Fire Spread, Intensity and Growth

Once ignited, a fire will continue to spread from fuel particle to particle as a self-sustaining process as long as the heat produced by combustion is sufficient to heat the adjacent

particles to the ignition temperature, or the fuel is exhausted. The rate of fire spread is influenced by fuel properties, particularly the moisture content and temperature, and the ambient atmospheric conditions, particularly wind speed. Over the past decades, several dozen mathematical fire spread models have been developed using approaches varying from simple empirically based nonlinear regressions to detailed computational fluid dynamics models (Sullivan, 2009). Fire intensity, the amount of energy released per unit length of fire front, is usually modeled as a function of fire spread rate, fuel consumption, and heat content of the fuel. A number of empirical models have been developed (assuming local homogeneity) to model the two-dimensional spread of the fire perimeter through heterogeneous fuel and topographic conditions at landscape scales (< 1–100 km) (Sullivan, 2009). The study area and time period of interest represent a set of voxels, with vegetation, topographic and other geographic covariates varying between points in a two-dimensional grid, and weather and other covariates varying spatially across the grid and temporally for each time step in the period. Early models represented fire spread as cellular automata, where a cell along a fire perimeter composed of grid cells could ignite adjacent grid cells in a time step, depending on vegetation, topographic, and weather covariates in the adjacent cell. Higher-resolution models simulate fire spread as a wave process, projecting the angular velocity of a number of discrete points around the fire perimeter as a vector over a discrete time step, depending on vegetation, topographic and weather covariates, but where the spread distance and directions are unconstrained by the grid resolution. The fire perimeter after each time step is remapped as the convex hull of the new points. However, although fire prediction is inherently probabilistic (because of the difficulty in accurately representing fuel properties and assessing and predicting atmospheric conditions; Taylor et al. 2013), most fire spread models are deterministic. Recently several authors have used ensemble methods to introduce stochasticity to fire spread (Cruz, 2010) and fire growth models (Braun & Woolford, 2013; McLoughlin & Gibos, 2016; Pinto et al., 2016) to better represent uncertainty.

Statistical models may provide important alternative risk measures or adjuncts to deterministic models. Noting that large, intense fires are rare events, Hernandez et al.

(2015) fitted generalized extreme value (GEV) distributions to remote-sensing based observations of fire intensity (fire radiative power) and size for fires in Portugal and used a nearest neighbor procedure to estimate the parameters of the distribution from meteorological covariates. They suggested that this approach provides an important estimate of uncertainty beyond qualitative fire danger indices. Price et al. (2015) fitted binomial regression models of large fire spread distances through cells with varying fuel conditions and weather conditions for 677 large fires in the Sydney region of Australia. They used the models to estimate the likely spread distance and the probability that a fire starting from the 667 ignition points would reach one of 26,000 3.4-hectare receiver points in the study area. The heuristic of modeling potential fire spread from an ignition to a receiver point of interest provides a simpler alternative to explicitly modeling fire spread from all points on the fire perimeter, which is computationally demanding.

2.4 Duration

Although the probability of fire occurrence has been well studied using statistical models for over half of a century, quantifying the survival distribution of fires during its containment for management purposes has not received much attention until more recently. Finney et al. (2009) categorized stages of containment of a fire into spreading intervals based on fire occurrences during 2001 to 2005 in the United States. The number of days in each interval were modeled using generalized linear mixed models using the same framework as for a repeated measures problem. Other quantitative studies also have been carried out for studying fire containment elsewhere with various foci in their statistical methods, such as for Italy (Marchi et al., 2014), Spain (Costafreda-Aumedes et al., 2015), Canada (Xiong, 2015), Mediterranean Europe (DaCamara et al., 2014), and Portugal (Fernandes et al., 2016). The latter three modeled duration directly, while the latter two considered the survival probabilities of duration. The former two studies used analysis of variance and regression trees, respectively. However, integrating these models into a unified framework incorporating fire occurrence models to describe and predict the complete dynamics of wildfires remains a challenge.

The duration of a fire may also be modeled directly as a survival outcome using statistical survival models. These models exist commonly in industrial and medical research; they work in similar ways as logistic regression under fire occurrence modeling but differ in that the quantity of interest is the survivor function or the hazard function of an outcome (typically a time quantity, obtained by measuring from an origin to an event). Since the survivor and hazard functions represent, respectively, the probability that the individual can survive more than a certain time and the instantaneous rate of death at a certain time given survival up to that time, they seem well-suited to describing the dynamics of wildfires.

Let $t_i, i = 1, \dots, n$ be the duration of fire i, assumed to be independently distributed. We may model t_i through a log-location-scale model, sometimes referred to as the accelerated failure time (AFT) model:

$$\log(t_i) = \mu + \boldsymbol{\beta^T x_i} + \sigma\epsilon_i,$$

where μ and σ are the location and scale parameters, $x_i = (x_{i1}, \dots, x_{iP})^T$ is a vector of P covariates and $\beta^T = (\beta_1, \dots, \beta_P)$ are the corresponding coefficients, and ϵ_i are random errors. The survivor function of the outcome is

$$S(t_i|x_i) = S_0\left(\frac{\log t_i - (\mu + x_i^T\beta)}{\sigma}\right),$$

where S_0 is the survivor function of the random error, termed the baseline survivor function. Since fire duration tends to be heavily right-skewed, the survivor functions may be modeled through parametric right-skewed distributions. Three common such baseline survivor functions are standard Gumbel, standard normal, and standard logistic, which correspond respectively to Weibull, log-normal, and log-logistic distributions for the outcome. As the covariate effects are multiplicative on time, the model assumes that different covariate values will scale the time axis of the survivor function. Covariates that are exogenous environmental variables (e.g., wind speed and temperature) then serve as stress factors that accelerate/decelerate the time to containment of a fire but keep the shape of the survivor function the same.

The hazard function may also be modeled directly using a Cox proportional hazards (PH) model:

$$h(t_i|\boldsymbol{x}_i) = h_0(t_i)\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i),$$

where h is the hazard function of the outcome; h_0 is the baseline hazard function corresponding to the hazard when $x_i = 0$, which can be either parametrically specified or unspecified to capture potential irregular features. In the PH framework, the covariate effects act multiplicatively on the baseline hazard rate. The model assumes that fires with different covariate values will result in hazard functions that are proportional to each other. This modeling strategy particularly lends itself to covariates such as endogenous fire characteristic variables (e.g., initial size and drought indices), where the intrinsic tendency of burning is different for fires with different values of these covariates. In particular, fires with large initial size occurring during drought conditions will have less steep hazard curves.

Recently, Morin et al. (2015) used survival techniques to model the duration of forest fires in Ontario's intensive fire management zone using data on more than 18,000 fires recorded during 1989 through 2004. Fire management zones are partitions of a study region that are assumed to be approximately internally homogeneous with respect to ecological characteristics such as fuel, weather, topography, and fire management strategy, and so may have a similar range or pattern of fire characteristics including size, duration, intensity, frequency, and season (Morin et al., 2015). They restricted their analysis to a period up to 2004 due to a change to Ontario's fire management strategy that led to a change in the number and location of fire management zones in the province after 2004. Response time, initial size, and several other FWI System indices were considered as covariates. The duration of each fire was defined to be the time interval from the start of initial attack to the time that a fire was declared as being under control, measured in hours. To capture changes in shapes of the hazard that were observed in nonparametric estimates of the survival function, and to ensure that the requirement for proportional covariate effects was not violated, a nonparametric stratified PH model was used to model survival times of

lightning-caused fires. Their work appears to be the first of its kind to model duration on a fine timescale using a stratified PH model, demonstrating that survival models that include covariate effects, such as the PH model, can be used as building blocks for more complicated structures in wildfire modeling.

Within fire management zones, the durations of fires within the same zone are dependent. An important extension of univariate regression type models to account for such dependence is the inclusion of a shared random effect z_i to explain the variation in homogenous space polygon $i = 1, \dots, n$ for fires $k = 1, \dots, n_i$ occurring in that polygon. A typical modeling framework in survival models is

$$S(t_{ik}|x_{ik}, z_i) = S(t_{ik}|x_{ik})^{z_i},$$

or equivalently,

$$h(t_{ik}|\boldsymbol{x}_{ik}, z_i) = z_i h(t_{ik}|\boldsymbol{x}_{ik}),$$

where $\mathrm{S}(\mathrm{t_{ik}}|\mathrm{x}_{ik}, \mathrm{z_i})$ is the conditional survivor function for fire k of polygon i with covariate vector $x_{ik} = (x_{ik1}, \dots, x_{ikP})^T$ and $h(t_{ik}|x_{ik}, z_i)$ is the conditional hazard function. The term z_i is commonly referred to as a shared frailty, and the framework is referred as a shared frailty model. The frailty extension of the PH model has been discussed in many texts because of the popularity of PH frailty models in medical studies (Hougaard, 2000; Therneau & Grambsch, 2000; Duchateau & Janssen, 2008; Wienke, 2010). Using the fact that $h(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\log S(t)$ and the formulation of the PH model, the above expression leads to

$$h(t_{ik}|x_{ik}, z_i) = z_i h_0(t_{ik})\exp(\beta_k^T x_{ik}) = h_0(t_{ik})\exp(\beta_k^T x_{ik} + b_i),$$

where the term $b_i = \log z_i$ can be interpreted as a latent covariate.

In her study of the lifetimes of forest fires in Ontario, Morin (2014) developed a set of PH frailty models to explore and quantify spatial differences in duration across a set of fire management compartments (FMC). The FMC partition was developed by Martell & Sun

(2008). Morin (2014) found that a Gaussian frailty term b_i, representing an FMC effect, had an estimated variance significantly different from zero for lightning-caused fires, which is evidence in favor of a positive dependence between the durations of fires in the same FMC. Mapping posterior estimates of the frailties showed that the western region of Ontario experiences lightning-caused fires with shorter survival times (Figure 2.4).

It is worth noticing that the AFT model can be extended to a shared frailty model of the form described earlier. For example, if the outcomes follow a Weibull distribution, with location parameters $\lambda_k = \exp\left(-\frac{\mu_k + x_{ik}^T \beta_k}{\sigma_k}\right)$ and scale parameters $\upsilon_k = \frac{1}{\sigma_k}$, including the term b_i as an additive latent covariate yields

$$h(t_{ik}|x_{ik}, b_i) = \lambda_k \upsilon_k t^{\upsilon_k - 1} = \exp\left(-\frac{\mu_k + x_{ik}^T \beta_k + b_i}{\sigma_k}\right) \frac{1}{\sigma_k} t^{\frac{1}{\sigma_k} - 1}$$

$$= \exp\left(-\frac{b_i}{\sigma_k}\right) h(t_{ik}|x_{ik}).$$

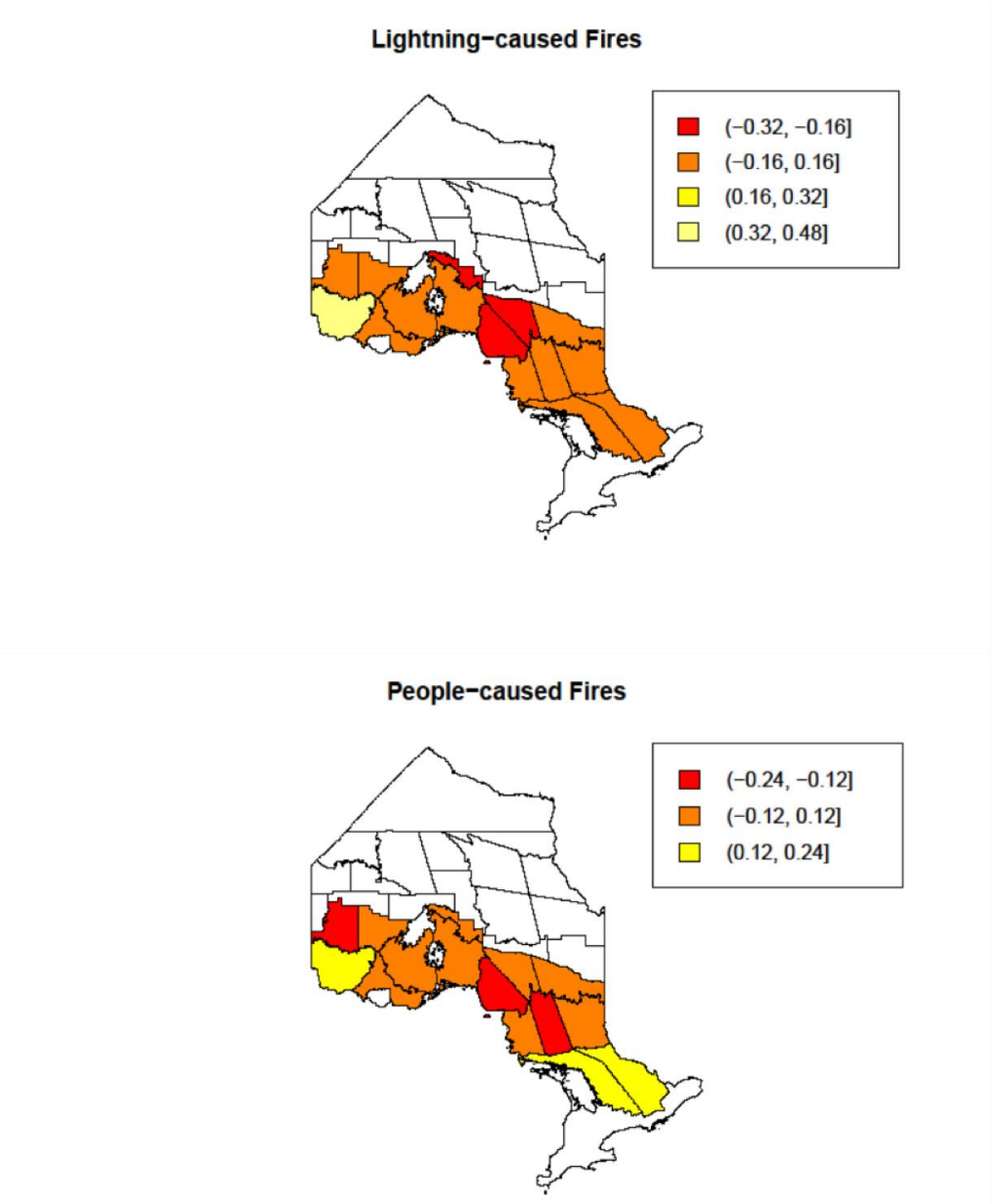


Figure 2.4: Choropleth map of the frailty terms for lightning-caused fires in the former intensive fire management zone of Ontario, Canada, which was partitioned into a set of fire management compartments (FMCs). Each FMC polygon was assigned a heat map color based on the estimate of the latent effect of the FMC (i.e., the frailty). FMCs that are outside of the study region are white. Exponentiated values of posterior frailty estimates can be viewed as multiplicative factors on the hazard function of fire lifetimes. Negative estimates imply an increase in survival probability via a reduction in hazard rate. Adapted with permission from Morin (2014).

The AFT model may be an appropriate alternative when modeling long fires with environmental variables (i.e., having smooth survivor functions and exogenous covariates). We come back to this formulation in section 6.

The frailty model and the copula model serve as important pieces of the foundation of modeling multivariate survival outcomes. Here, we briefly note the connection between shared frailty models and Archimedean copulas. For simplicity, we model without covariates. Shared frailty models assume that, conditional on z_i , t_{ik} are independent across all fires within the same polygon; thus, the joint survivor function conditioning on z_i is

$$S(t_{i1}, \dots, t_{in_i}|z_i) = S(t_{i1}|z_i) \dots S(t_{in_i}|z_i)$$

$$= \exp\{-H(t_{i1}|z_i)\} \dots \exp\{-H(t_{in_i}|z_i)\}$$

$$= \exp\{-z_i H(t_{i1})\} \dots \exp\{-z_i H(t_{in_i})\}$$

$$= \exp\{-z_i[H(t_{i1}) \dots H(t_{in_i})]\},$$

where $H(t) = -\log S(t)$ is the cumulative hazard function. Taking the expectation of the right-hand side of the expression above over z_i, the joint survivor function, yields

$$S(t_{i1}, \dots, t_{in_i}) = E\, S(t_{i1}, \dots, t_{in_i})$$

$$= E \exp\{-z_i[H(t_{i1}) + \cdots + H(t_{i2})]\}$$

$$= \boldsymbol{L}[H(t_{i1}) + \cdots + H(t_{i2})],$$

where $\mathrm{L}(a) = Ee^{-xa}$ is the Laplace transformation of the random variable x. Using the fact that $S(t) = e^{-H(t)} = Ee^{zH(t)} = \mathrm{L}H(t)$, we have

$$S(t_{i1}, \dots, t_{i2}) = \mathrm{L}[H(t_{i1}) + \cdots + H(t_{i2})] = \mathrm{L}[\mathrm{L}^{-1}\mathrm{S}(\mathrm{t_{i1}}) + \dots + \mathrm{L}^{-1}\mathrm{S}(\mathrm{t_{i2}})],$$

which yields the so-called Archimedean copula family (Nelsen, 2006; Liu, 2012; Joe, 2014). Embrechts & Hofert (2014) provide a detailed overview of the connections between

the two frameworks, as well as the development of copulas in a quantitative risk management perspective.

2.5 Size

The increase in fire size during the life of a fire and its ultimate size at extinction are important to the difficulty of control, and fire size also has many other impacts of scientific and economic concern. Qualitative publications based on physical/process models derived in the natural sciences historically dominated the study of this phenomenon; however, quantitative studies based on empirical/statistical models have been appearing in the literature at an increasing rate since approximately the late 1990s (Cui & Perera, 2008). Early empirical models assumed a power law (i.e., Pareto) distribution for wildfire sizes:

$$f_X(x; b) \propto x^{-b},$$

where X is the random variable representing fire size, and its density function $f_X(x; b)$ depends on a parameter b. An example using this approach was presented by Schoenberg et al. (2003), who considered several parametric models for the distribution of wildfire sizes in Los Angeles County, California. Using visual diagnostics and nonparametric tests for comparing distributions, they advocated for the use of a tapered Pareto distribution for modeling size distributions in that area. Cumming (2001) modeled the survivor function of the size of fires in the province of Alberta, Canada, using a right-truncated exponential distribution under the assumption that there was a maximum size a fire could grow to, based on characteristics of the study area. Recent models for fire size include environmental variables into models. Butry et al. (2008) incorporated environmental variables using linear regression for modeling the logarithm of the size of large fires in northeast Florida from 1981–2001. Chen et al. (2014) used quantile regression to study the effect of precipitation on fires in southwestern China. A comprehensive review of fire size models appears in Cui & Perera (2008).

Here, we review two key threads of research in the development of statistical methods for modeling fire size. Power-law behaviors are commonly observed in nature. If fire growth follows a preferential attachment or Yule process (Gibrat's Law), the distribution of randomly killed states (or states observed once) under stochastic processes follows a power law in one or both tails (Reed, 2001; Reed & Hughes, 2002). Using percolation theory, Reed (1999) observed that a piecewise probability distribution, partitioned at the percolation threshold, fits the distribution of forest fire size reasonably well. Reed & McKelvey (2002) derived the density function and survival function of the killed state. Let the fire size at time t be $X(t) = \exp(\mu t)$ and the growth rate at size X be $\mu(X) = \mu X$, proportional to size by a constant, μ. We further assume that the killing rate $k(t)$ takes the form of

$$k(t) = \lim_{dt \to 0} \frac{P(T < t + dt | T > t)}{dt} = \mathrm{v}\big(X(t)\big),$$

where $v(x)$ is a nonincreasing function referred to as the extinguishment rate. Let $\bar{X}$ denote the killed state, and then it can be shown that the density function of $\bar{X}$ is

$$f_{\bar{X}}(x) = \rho(x)\exp(-\int_{x_0}^{x} \rho(x')dx'),$$

where $\rho(x) = v(x)/\mu(x)$ is the hazard rate function. The survival function of $\bar{X}$ is

$$S_{\bar{X}}(x) = \exp(-\int_{x_0}^{x} \rho(x')dx').$$

The plot of empirical $\log S_{\bar{X}}(x)$ against $\log x$ demonstrates a linear trend if the data exhibit power-law behavior. The authors suggest plotting the extinguishment growth-rate ratio (EGRR) against $\log x$, which is expressed as

$$\mathrm{EGRR} = R(x) = \frac{xv(x)}{\mu(X)} = \frac{xf_{\bar{X}}(x)}{S_{\bar{X}}(x)}.$$

Power-law behavior will occur on an interval that EGRR is constant. Conditions when ECGR is not constant that may lead to thin- or thick-tailed distributions include the following: (*a*) In regions where the fire season is limited to a portion of the year (e.g., by winter), the distribution of killing times for fires starting later in the season may be right truncated; this also applies for regions where fire size may be limited by available fuel. (*b*) In a managed environment where all fires are suppressed but where occasional extreme weather such as Santa Ana winds favors large fire growth (Moritz, 1997), distributions may be thick tailed; this may also occur when climate over a long sampling period is nonstationary. When power-law behavior does not occur, the authors recommended using non-power-law distributions such as a 3-parameter Weibull distribution for certain cases. The theoretical foundation of the work above connects well with methods developed to model the distribution of size in other fields of science (Reed & Hughes 2002, 2004; Reed & Jorgensen 2004; Reed 2011, 2012).

Another thread of research in modeling fire size has developed in engineering. To examine extreme fire size, Holmes et al. (2008) utilized GEV methods for analyzing fire sizes with heavy-tailed distributions. GEV methods play important roles in engineering and actuarial science because of their focus on rare but extreme events (Castillo, 2012; Longin, 2016). Foss et al. (2011) provide a probabilistic perspective of heavy-tailed distributions. The distribution of a random variable Y is said to be heavy-tailed if, for any $u > 0, v > 0$,

$$\lim_{u\to\infty} \mathrm{P}\,(Y > u + v | Y > u) = \lim_{u\to\infty} \frac{S(u+v)}{S(u)} = 1.$$

That is, if the observation already exceeds a large value u , then it will likely exceed a larger value $u + v$. The maximum value of a sample of observations is traditionally used for constructing the models under GEV methods. To overcome the limitation of information loss, the authors used all the observations beyond a threshold (e.g., size > 200 ha) instead. Thus, the survivor function of the observations beyond a threshold is

$$S(u + y_i|u) = P(Y_i > u + y_i | Y_i > u) = \frac{S(u+y_i)}{S(u)},$$

where Y_i is the size for fire i and u is the threshold. It follows that the resulting distribution function follows a generalized Pareto distribution (Davison & Huser, 2015):

$$f(y_i|\mu,\xi,\sigma_i) = \frac{1}{\sigma_i}\left(1+\xi\frac{y_i-\mu}{\sigma_i}\right)^{-\left(1+\frac{1}{\xi}\right)},$$

where μ, σ, and ξ are the location, scale, and shape parameters. Covariates z_i can be included through $\sigma_i = \sigma(z_i) = \mu + \beta^T z_i$ to model and simulate fire sizes given environmental variables. The model leads to a set of integrated frameworks (e.g., Preisler & Westerling, 2007; Westerling & Bryant, 2008; Preisler et al., 2011; Westerling et al., 2011; Bryant & Westerling, 2014) that can be used to understand, for example, the impact of climate change and human development on fire-related losses in different regions.

2.6 Modeling Duration and Size as Joint Outcomes

As mentioned in the section on fire occurrence modeling, it is possible to combine models for fire occurrence with other models, such as those for duration, to model fire load (e.g., Morin, 2014), or such as those for fire size and cost distributions, to develop spatially explicit forecasts for suppression costs (e.g., Preisler et al., 2011). These frameworks commonly decompose the problem through a multi-stage approach, developing separate, independent models for each component as building blocks for the overall model, such as an occurrence model coupled to an independent survival model (Morin, 2014), or coupling occurrence models to independent models for fire size and cost distributions (Preisler et al., 2011). However, components may be linked.

For example, marked point process models have been proposed for wildfire modeling: The point process identifies the occurrence of the fire, with size as the mark. However, the marks may not be separable from the points. This was illustrated by Schoenberg (2004), who found evidence of a lack of separability between fire occurrences and sizes in Los Angeles County, California, due to small-scale clustering. Moreover, even outside of the context of developing marked point processes models, key wildfire characteristics are

likely linked. An obvious example of this is fire duration and size, under the principle that the longer a fire lasts, the larger it grows. Such situations motivate the need to consider alternative modeling frameworks where outcome characteristics are modeled jointly. In this section, we give an overview of joint modeling of two random variables, using fire duration and size as an illustration.

Jointly modeling the duration and size of fires with environmental variables as covariates offers a potential novel direction for effectively quantifying these outcomes. Since smaller-sized fires' (<2 hectares) lifetimes are usually short (<2 days), while larger ones are usually long (days to months), such modeling accounts for the dependence between duration and size. In managed regions, more than 90% of fires are contained during an initial attack, and for fires that escape extended attack, there is a clear connection between the time to containment and fire size (Fried & Gilless, 1989). The two-dimensional framework for bivariate extreme value models (e.g., Weibull, log-normal, logistic) has been recently adopted in some pioneering work because duration and size are often weakly correlated with heavy tails. (Yoder & Gebert, 2012; Sun, 2013). Bayham (2013), in his dissertation, modeled the duration, size, and cost of containment on 3,829 US fires using a tri-outcome PH frailty model with environmental and geographical variables as covariates. Endogenous time-varying covariates were lagged for one period, and the median value was used instead of the complete covariate trajectories.

Past work modeling duration and size as a function of environmental variables shares four common features. First, although the survivor or hazard functions of the outcomes are often of interest, they can be obtained easily from estimates of the distribution of these outcomes and hence do not need to be modeled directly. Nevertheless, they need to be measured from the same origin to the same event (e.g., from the start of initial attack to the time of final control). Second, heavy-tailed distributions may be used to model both outcomes. Although power-law or extreme value distributions have received much attention in the context of wildfire science, basic location-scale distributions also fit well, and such empirical approaches have been overlooked. Third, although AFT frameworks have been commonly used, they do not necessarily lead to a model where the frailty can be interpreted as a latent

covariate acting multiplicatively on the hazard. Finally, we note that covariate coefficients in the PH frailty model may not be estimated well when the PH assumption does not hold (He & Lawless, 2005; He, 2014), which may be of concern in the use of these frailty models in wildfire science. As a result, placing the random effect additively as a latent covariate in an AFT model would provide a compromise to both frameworks (Lambert et al., 2004; Komárek & Lesaffre, 2008) and a foundation to model duration and size jointly.

To illustrate a joint modeling framework that addresses the issues above, we consider a simple model that has been discussed extensively in the literature. Assuming that both the duration and size of fires follow a location-scale distribution, AFT models can be linked to model the two outcomes jointly:

$$\log(t_{ik}) = \mu_k + \beta_k^T x_{ik} + b_{ik} + \sigma_k \epsilon_{ik},$$

where $b_i = (b_{i1}, b_{i2})^T$ is a random effect with components that are dependent. Here, k equals 1,2, for duration and size respectively; t_{ik} is the outcome; x_{ik} are covariates with associated coefficients β_k; μ_k is the intercept term (the mean of the logarithm of t_{ik} when $x_{ik} = 0$); ϵ_{ik} is the outcome-specific error with unit variance, associated with outcome k for fire i; and σ_k are variance parameters associated with outcome k.

Various forms of b_i have been discussed in the literature. He & Lawless (2005) and Duchateau & Janssen (2008), among many others, note that $\boldsymbol{b}_i$ may be parametrized with $b_{i1} = b_{i2} = b_i$, as a shared frailty acting additively on the logarithm of the outcomes. To account for the scale difference between the two outcomes, an additional parameter, γ, often called the factor loading parameter, can be introduced by letting $b_i = (b_i, \gamma b_i)^T$ with $b = (b_1, \ldots b_n)^T \sim MVN(0, \Sigma_b)$, with b_i and ϵ_{ik} independent. The term b_i can be viewed as an individual-specific error that is shared across the two outcomes. With the assumption that individual fire lifetimes and sizes are independent, a simple form for Σ_b is $\sigma_b^2 I$ (Renouf et al., 2016; Juarez-Colunga et al., 2017). Having σ_b significantly different from 0 suggests that there is dependence between the two outcomes. When $\gamma = 1$ the shared random effect influences the two outcomes identically. This is not likely in situations where the two

outcomes, such as duration and size, have very different scales. In general, having γ significantly different from 1 suggests that the terms b_i have different scales by which they act on the outcomes. If prior knowledge suggests that the frailty is correlated, for example, spatially, then Σ_b may take more complicated forms (Feng & Dean, 2012). Additional constraints are required (i.e., removing ϵ_{i1} from the model) to ensure that the model is identifiable. Alternative forms such as assuming $b_i = (b_{i1}, b_{i2})^T$ with $b_i \sim N_2(0, D)$ have also being considered in Komárek & Lesaffre (2008) and Bogaerts et al. (2018). For other methods that also use random effects or latent variables to model multiple outcomes jointly, readers are directed to, for example, Molenberghs & Verbeke (2017).

2.7 Event Sets and Burn Probability

Many fire characteristics, such as ignition probability, spread rate and duration, and fire size contribute to the fire hazard. Reed (2006) defined the local hazard of burning at a point x in a study area, at time t, as

$$\lambda(t; x) = \lim_{\mathrm{d}t \to 0} \{\mathrm{P}(\text{fire at location } x \text{ in} [t, t + \mathrm{d}t])/\mathrm{d}t\},$$

the area-wide hazard of burning as

$$\Lambda(t) = \lim_{\mathrm{d}t \to 0} \{\mathrm{P}\ (\text{fire ignited somewhere in the area in } [t, t + \mathrm{d}t])/\mathrm{d}t\},$$

and the relationship between local and area wide hazard of burning as

$$\lambda(t{:}\, x) = \Lambda(t) \int_A h(x, y{:}\, t) f(y; t)\, dy,$$

where $h(x, y{:}\, t)$ is the conditional probability of a fire ignited at point y spreading to x at time t and $f(y; t)$ is the probability density function of where ignitions will occur over the area A given that a fire occurs in time t. The integral above can be simplified, where $p(t, x)$ is the conditional probability of a fire occurring at x given that a fire starts somewhere in the area, as

$$\lambda(t{:}\,x) = \Lambda(t)p(t,x).$$

It is noteworthy that most of the local hazard of burning at a point *x* obtains from incursions of fire from adjacent locations. This model is analogous to population system epidemiology (Koopman & Lynch, 1999), where infection connections between individuals and joint effects of possible multiple exposures are incorporated into infectious disease spread analysis. The local hazard of burning in a region has been estimated from empirical data from fire scars and forest stand age data (see summary in Taylor et al., 2013) in a region, assuming spatial homogeneity, while the area-wide hazard can be estimated from administrative fire records or remote sensing data. However, a number of authors have found that the local hazard of burning may vary within a landscape at scales important to fire and land management due to topographic and vegetation conditions. For example, forest stands on warm slopes in the Rocky Mountains have a greater likelihood of burning (Rogeau & Armstrong, 2017), while those adjacent to nonvegetated areas such large lakes have a lower likelihood of burning (Bergeron, 1991).

Over the past decade, a body of work termed burn probability modeling (Miller et al., 2008) has developed to estimate the local hazard of burning while incorporating the influences of varying vegetation types and topographic positions within regional landscapes. Briefly, burn probability is estimated geometrically by modeling fire event sets, where an event is a spatially referenced fire perimeter map, the final fire perimeter obtained from the cumulative spread over a fire's lifetime. Monte Carlo methods are used to simulate a large number of fire events in a study area. As with the spread modeling described earlier, the study area and simulation period represent a set of voxels, with vegetation, topographic, and other geographic covariates varying spatially and weather covariates varying temporally for each day in the simulation period. In one approach, a fire is ignited at a point using a conditional spatial point process model of fire occurrence (Woo et al., 2017) including covariates at the grid points; the fire spreads between points using a deterministic fire growth model, depending on weather, vegetation, and topographic covariates at the grid points on a particular day; and individual fire events are modeled for a number of days

informed by duration models. The burn probability of a given cell is estimated as the empirical proportion of fire events in that cell over the number of simulation iterations (usually years), as shown in Figure 2.5. The size distribution of fires in resulting event sets can be compared against fire size models to assess goodness of fit (e.g., appendix S.6 in Wang et al., 2016). Several systems have developed around different fire spread models. Parisien et al. (2013) provide a flowchart of the modeling process for an application of the BurnP3 system that uses the Prometheus spread model; similar procedures are used for simulations with the FSim system, which uses the FarSite fire growth simulator (Finney et al., 2011). Further challenges may include closer integration of joint models of size and duration and covariance of numbers of fires and fire size, as well as explicit representation of fire suppression.

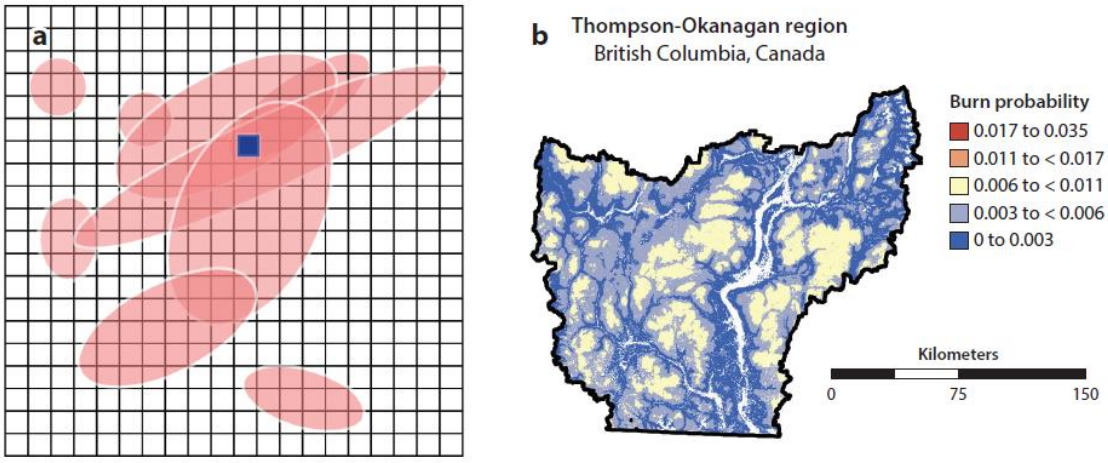


Figure 2.5: (a) Wildfire event sets can be generated with stochastic point process models and fire growth simulations of specified duration (ellipses used for illustration only; figure courtesy of Carol Miller, US Department of Agriculture Forest Service). Elements within a cell are homogenous respect to weather fuels and topography. The ellipses are individual fires, and the blue square represents a sampling point. (b) Monte Carlo methods have been used to map burn probability by simulating large event sets representing many thousands of potential outcomes in modeled landscapes, such as the Thompson-Okanagan region of southern British Columbia (Wang et al., 2016).

2.8 Discussion

Fire danger and risk research has evolved from the development of qualitative indices, to deterministic models of fire characteristics, to stochastic models of fire characteristics. It is an ongoing challenge to integrate these models and approaches in probabilistic, quantitative hazard and risk models. Improving prediction of daily wildfire dynamics is critical for proactive rather than reactive fire management decision making. Effective prediction for decision making requires (*a*) an understanding of the physical and management processes influencing ignition, growth, and survival; (*b*) the acquisition and assembly of historical, longitudinal data on daily fire starts, size, management actions, extinguishment, and covariates for building empirical models; (*c*) the development of appropriate statistical models; and (*d*) the implementation of predictive models in fire management decision support systems, preferably with an ability to use new data on fires for continuous improvement of predictive models.

Much work has focused on developing models to understand ignition and growth processes, and far fewer studies have considered containment and extinction. Even so, there are few models that consider changes in fire size by day while also accounting for resources allocated to suppression; this is due in part to limited availability of such longitudinal data on a daily scale. More work is also needed in the development of appropriate models that take into account fire-pest interaction effects and the health of trees in the path of a fire. New remote-sensing products may assist here, but there will be considerable work involved in building historical archives. Importantly, there are substantial challenges associated with mounting investigations and developing predictive models because of the massive effort involved in the assembly of historical fire databases, validation of these databases, linkage and data fusion across regions and across governmental agencies that record environmental and management variables associated with fire, and management of differences in spatial and temporal resolution that are associated with each database.

Verification of historical records can be very difficult, as can homogenization of long-term series of environmental data, when monitoring stations change location over time. There

are also challenges associated with appropriately accommodating the differences in fire suppression management protocols over time, changes in detection efficiency over time, and the differences in tools and techniques for fire suppression that have evolved over recent decades. These large data issues are not inconsequential, especially when developing provincial/national models at high spatial/temporal resolution. Finally, accurate prediction models require incorporating variability associated with the differential use of fire suppression resources between fires and variability associated with future weather conditions. Importantly, we note that few statisticians are willing to expend the effort necessary to take models and methods to an implementation or knowledge translation stage as identified in item *d* above, but this is a key critical process step for impact. Further work is also needed to better represent uncertainty in models of spread, growth, and intensity and also for visualization of these characteristics.

Comprehensive, fine-scale fire occurrence modeling over a large study area introduces specific challenges. For example, the province of Ontario uses a suite of person-caused and lightning-caused fire occurrence prediction models operationally on a daily basis (Woolford et al., 2016). For this decision support tool, human-caused and lightning-caused fire occurrence predictions based on models need to be integrated into a single probability scale. This can be challenging because occurrence probabilities for lightning-caused fires in a given cell can be much larger than the probabilities for human-caused fires. This is because lightning-caused fire occurrence models incorporate lightning strike observations, which have high daily variability (e.g., Wotton & Martell, 2005), whereas indicators of human presence or activity in human-caused fire models don't have strong daily variation (e.g., Woolford et al., 2011). This difference in scale may occur because lightning-caused fire occurrence models incorporate information about the observed strikes that are recorded by a network of sensors (e.g., Wotton & Martell, 2005), whereas human-caused fire occurrence prediction models summarize historical patterns in fire ignitions without incorporating information about potential ignition sources (e.g., Woolford et al., 2011). In addition, outputs from fitting complex models in statistical software, such as a logistic GAM model object fit in R software, need to be summarized (as, e.g., a set of lookup tables

for each partial effect in the model) for easy implementation into a non-R-based fire management operations decision support tool.

Simulation systems that are currently used to estimate the annual local burn probability use statistical models to represent stochastic components in that complex system. However, there are many components that are modeled as separate subprocesses. In order to enhance quantitative risk assessment models, a joint modeling framework should be considered when key characteristics may not necessarily be independent. Further development of quantitative risk assessment methods across all temporal scales will require, as in statistical physics, hybrid approaches that combine mathematical and statistical models with simulation methods to estimate very complex processes. For example, key stakeholders such as fire management agencies and property insurers are interested not only in annual burn probability maps but also in burn probabilities at other temporal scales, such as the probability that a fire may be ignited and spread into a nearby town on a given day.

We comment that it would be very interesting to compare simulation methods with other means of estimating these complex processes, which may be effective at some spatial scales. It would be also useful to compare modern statistical learning algorithmic techniques to the well-established logistic-based modeling techniques. Developing methodology for combining these and other models together in an ensemble framework, thereby building on the benefits of each of these approaches, would be particularly helpful.

It is challenging to incorporate estimates of uncertainty in fire management strategies, in part because it is a highly dynamic and multiscalar decision environment. Although advances have been made in developing stochastic models of characteristics such as fire occurrence, medium-term fire spread, and burn probability, few studies have connected hazard measures, including uncertainty, with damage functions and impacts (e.g., Preisler et al., 2011). Implementation of new models within a fire management decision environment presents special challenges at the interface between data analytics and human factors (that are not unique to the fire community). These include:

1. The time available for decision making decreases in the series of activities: mitigation/prevention, planning/preparedness, and response. At the sharp end of fire response, the time for decision making may be reduced to a few minutes or less (e.g., Alexander et al., 2016). Models have to be simple to use and easy to interpret; visualization techniques should be used whenever possible.

2. Decision makers within an operations background tend to be "men (or women) of action, rather than men of letters" (Macleod, 1964, p. 8) coming from an institutional culture that values fast, intuitive decision making over slower, rational decision processes (e.g., Kahneman, 2011) and have a healthy skepticism of models. It is important to validate models and provide case studies showing the value of information. In counterpoint, fire managers with long experience with weather dependent fire phenomena may have an intuitive appreciation for the stochastic nature of fire characteristics. Current machine learning algorithms based on historical data cannot adequately replicate such experience.

Collaborative approaches have proved successful in developing and implement the models currently used in fire management. Whereas commonly the statistician's goal is finding a useful application, it is important at a project's outset to set common goals, find champions who are influential members of the user community, create relationships, and seek to understand the decision maker's way of doing business and constraints.

References

Albert-Green, A., Dean, C., Martell, D. L., & Woolford, D. G. (2012). A methodology for investigating trends in changes in the timing of the fire season with applications to lightning-caused forest fires in Alberta and Ontario, Canada. *Canadian Journal of Forest Research, 43*(1), 39-45.

Alexander, M., Taylor, S., & Page, W. (2015). *Wildland firefighter safety and fire behavior prediction on the fireline.* Paper presented at the Proceedings of the 13th international wildland fire safety summit & 4th human dimensions wildland fire conference.

Bayham, J. (2013). *Characterizing incentives: an investigation of wildfire response and environmental entry policy.* PhD Thesis, Washington State University, Pullman,WA

Bergeron, Y. (1991). The influence of island and mainland lakeshore landscapes on boreal forest fire regimes. *Ecology, 72*(6), 1980-1992.

Bogaerts, K., Komarek, A., & Lesaffre, E. (2017). *Survival analysis with interval-censored data: A practical approach with examples in R, SAS, and BUGS*: CRC Press.

Braun, W. J., & Woolford, D. G. (2013). Assessing a Stochastic Fire Spread Simulator. *Journal of Environmental Informatics, 22*(1).

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science, 16*(3), 199-231.

Brillinger, D. R., Preisler, H. K., & Benoit, J. W. (2003). Risk assessment: a forest fire example. *Lecture Notes-Monograph Series*, 177-196.

Bruce, D. (1963). How many fires. *Fire Control Notes, 24*(2), 45-50.

Bryant, B. P., & Westerling, A. L. (2014). Scenarios for future wildfire risk in California: links between changing demography, land use, climate, and wildfire. *Environmetrics, 25*(6), 454-471.

Butry, D. T., Gumpertz, M., & Genton, M. G. (2008). The production of large and small wildfires. *The Economics of Forest Disturbances: Wildfires, Storms, and Invasive Species, 79*, 79.

Cardona, O. D., Van Aalst, M. K., Birkmann, J., Fordham, M., Mc Gregor, G., Rosa, P., . . . Décamps, H. (2012). Determinants of risk: exposure and vulnerability *Managing the Risks of Extreme Events and Disasters to Advance Climate Change*

Adaptation: Special Report of the Intergovernmental Panel on Climate Change (pp. 65-108): Cambridge University Press.

Castillo, E. (2012). *Extreme value theory in engineering*: Elsevier.

Chen, F., Fan, Z., Niu, S., & Zheng, J. (2014). The influence of precipitation and consecutive dry days on burned areas in Yunnan Province, Southwestern China. *Advances in Meteorology, 2014*.

Costafreda-Aumedes, S., Cardil, A., Molina, D. M., Daniel, S. N., Mavsar, R., & Vega-Garcia, C. (2015). Analysis of factors influencing deployment of fire suppression resources in Spain using artificial neural networks. *iForest-Biogeosciences and Forestry, 9*(1), 138.

Cruz, M. G. (2010). Monte Carlo-based ensemble method for prediction of grassland fire spread. *International Journal of Wildland Fire, 19*(4), 521-530.

Cui, W., & Perera, A. H. (2008). What do we know about forest fire size distribution, and why is this knowledge useful for forest management? *International Journal of Wildland Fire, 17*(2), 234-244.

Cumming, S. (2001). A parametric model of the fire-size distribution. *Canadian Journal of Forest Research, 31*(8), 1297-1303.

Cunningham, A. A., & Martell, D. L. (1973). A stochastic model for the occurrence of man-caused forest fires. *Canadian Journal of Forest Research, 3*(2), 282-287.

DaCamara, C. C., Calado, T. J., Ermida, S. L., Trigo, I. F., Amraoui, M., & Turkman, K. F. (2014). Calibration of the Fire Weather Index over Mediterranean Europe based on fire activity retrieved from MSG satellite imagery. *International Journal of Wildland Fire, 23*(7), 945-958.

Daley, D. J., & Vere-Jones, D. (2003). An introduction to the theory of point processes, volume 1: Elementary theory and methods. *Verlag New York Berlin Heidelberg: Springer*.

Duchateau, L., & Janssen, P. (2008). The Frailty Model New York. *Inc.: Springer-Verlag*.

Embrechts, P., & Hofert, M. (2014). Statistics and quantitative risk management for banking and insurance. *Annual Review of Statistics and Its Application, 1*, 493-514.

Feng, C., & Dean, C. (2012). Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models. *Environmetrics, 23*(6), 493-508.

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

Finney, M., Grenfell, I. C., & McHugh, C. W. (2009). Modeling containment of large wildfires using generalized linear mixed-model analysis. *Forest Science, 55*(3), 249-255.

Finney, M. A. (2005). The challenge of quantitative risk analysis for wildland fire. *Forest Ecology and Management, 211*(1-2), 97-108.

Finney, M. A., McHugh, C. W., Grenfell, I. C., Riley, K. L., & Short, K. C. (2011). A simulation of probabilistic wildfire risk components for the continental United States. *Stochastic Environmental Research and Risk Assessment, 25*(7), 973-1000.

Foss, S., Korshunov, D., & Zachary, S. (2011). *An introduction to heavy-tailed and subexponential distributions* (Vol. 6): Springer.

Fried, J. S., & Gilless, J. K. (1989). Notes: Expert opinion estimation of fireline production rates. *Forest Science, 35*(3), 870-877.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1): Springer series in statistics Springer, Berlin.

Fujioka, F. M., Gill, A. M., Viegas, D. X., & Wotton, B. M. (2008). Fire danger and fire behavior modeling systems in Australia, Europe, and North America. *Developments in Environmental Science, 8*, 471-497.

Garcia, C. V., Woodard, P., Titus, S., Adamowicz, W., & Lee, B. (1995). A logit model for predicting the daily occurrence of human caused forest-fires. *International Journal of Wildland Fire, 5*(2), 101-111.

Giglio, L., Randerson, J. T., & Werf, G. R. (2013). Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4). *Journal of Geophysical Research: Biogeosciences, 118*(1), 317-328.

GLOBAL, F. M. C. (2013). Vegetation fires and global change. *Challenges for concerted international action. A white paper directed to the United Nations and International Organizations. Germânia*.

Hardy, C. C., & Hardy, C. E. (2007). Fire danger rating in the United States of America: an evolution since 1916< a class=. *International Journal of Wildland Fire, 16*(2), 217-231.

He, W. (2014). Analysis of multivariate survival data with Clayton regression models under conditional and marginal formulations. *Computational Statistics & Data*

Analysis, 74, 52-63.

He, W., & Lawless, J. F. (2005). Bivariate location–scale models for regression analysis, with applications to lifetime data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(1), 63-78.

Hernandez, C., Keribin, C., Drobinski, P., & Turquety, S. (2015). *Statistical modelling of wildfire size and intensity: a step toward meteorological forecasting of summer extreme fire risk.* Paper presented at the Annales Geophysicae.

Heyerdahl, E. K., Lertzman, K., & Karpuk, S. (2007). Local-scale controls of a low-severity fire regime (1750–1950), southern British Columbia, Canada. *Ecoscience, 14*(1), 40-47.

Holmes, T. P., Huggett Jr, R. J., & Westerling, A. L. (2008). Statistical analysis of large wildfires *The Economics of Forest Disturbances* (pp. 59-77): Springer.

Hougaard, P. (2000). Shared frailty models. *Analysis of Multivariate Survival Data*, 215-262.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6): Springer.

Joe, H. (2014). *Dependence modeling with copulas*: CRC Press.

Juarez-Colunga, E., Silva, G., & Dean, C. (2017). Joint modeling of zero-inflated panel count and severity outcomes. *Biometrics, 73*(4), 1413-1423.

Kahneman, D. (2011). *Thinking, fast and slow*: Macmillan.

Komárek, A., & Lesaffre, E. (2008). Bayesian accelerated failure time model with multivariate doubly interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association, 103*(482), 523-533.

Koopman, J. S., & Lynch, J. W. (1999). Individual causal models and population system models in epidemiology. *American Journal of Public Health, 89*(8), 1170-1174.

Krawchuk, M. A., Cumming, S. G., & Flannigan, M. D. (2009). Predicted changes in fire weather suggest increases in lightning fire initiation and future area burned in the mixedwood boreal forest. *Climatic change, 92*(1), 83-97.

Krawchuk, M. A., Moritz, M. A., Parisien, M.-A., Van Dorn, J., & Hayhoe, K. (2009). Global pyrogeography: the current and future distribution of wildfire. *PloS one, 4*(4), e5102.

Lambert, P., Collett, D., Kimber, A., & Johnson, R. (2004). Parametric accelerated failure

time models with random effects and an application to kidney transplant survival. *Statistics in medicine, 23*(20), 3177-3192.

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.

Liu, J. (2012). Modeling dependence induced by a common random effect and risk measures with insurance applications.

Longin, F. (2016). *Extreme Events in Finance: A Handbook of Extreme Value Theory and Its Applications*: John Wiley & Sons.

Macleod, J. (1964). *Planning for forest fire control*: ottawa: queen's printer.

Marchi, E., Neri, F., Tesi, E., Fabiano, F., & Brachetti Montorselli, N. (2014). Analysis of helicopter activities in forest fire-fighting. *Croatian Journal of Forest Engineering, 35*(2), 233-243.

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

Martell, D. L., & Sun, H. (2008). The impact of fire suppression, vegetation, and weather on the area burned by lightning-caused forest fires in Ontario. *Canadian Journal of Forest Research, 38*(6), 1547-1563.

McLoughlin, N., & Gibos, K. A 72-day Probabilistic Fire Growth Simulation used for Decision Support on a Large Mountain Fire in Alberta, Canada.

Miller, C., & Ager, A. A. (2013). A review of recent advances in risk analysis for wildfire management. *International Journal of Wildland Fire, 22*(1), 1-14.

Miller, C., Parisien, M.-A., Ager, A., & Finney, M. (2008). Evaluating spatially-explicit burn probabilities for strategic fire management planning. *WIT Transactions on Ecology and the Environment, 119*, 245-252.

Molenberghs, G., & Verbeke, G. (2017). Modeling Through Latent Variables. *Annual Review of Statistics and Its Application, 4*(1).

Morgan, M. G., Henrion, M., & Small, M. (1990). *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*: Cambridge university press.

Morin, A. A. (2014). *A Spatial Analysis of Forest Fire Survival and a Marked Cluster Process for Simulating Fire Load.* The University of Western Ontario.

Morin, A. A., Albert-Green, A., Woolford, D. G., & Martell, D. L. (2015). The use of survival analysis methods to model the control time of forest fires in Ontario, Canada. *International Journal of Wildland Fire, 24*(7), 964-973.

Moritz, M. A. (1997). Analyzing extreme disturbance events: fire in Los Padres National Forest. *Ecological Applications, 7*(4), 1252-1262.

Nadeem, K., Taylor, S. W., Dean, C. B., Woolford, D. G., Magnussen, S., & Wotton, B. M. (2016 October). *Severe wildland fire risk prediction in Canada.* Poster session presented at Wildland Fire Canada 2016: Building Resilience, Kelowna, BC

Nelsen, R. (2006). An introduction to copulas, ser. *Lecture Notes in Statistics. New York: Springer*.

Papakosta, P., Xanthopoulos, G., & Straub, D. (2017). Probabilistic prediction of wildfire economic losses to housing in Cyprus using Bayesian network analysis. *International Journal of Wildland Fire, 26*(1), 10-23.

Parisien, M.-A., Walker, G. R., Little, J. M., Simpson, B. N., Wang, X., & Perrakis, D. D. (2013). Considerations for modeling burn probability across landscapes with steep environmental gradients: an example from the Columbia Mountains, Canada. *Natural hazards, 66*(2), 439-462.

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.

Pinto, R. M., Benali, A., Sá, A. C., Fernandes, P. M., Soares, P. M., Cardoso, R. M., . . . Pereira, J. M. (2016). Probabilistic fire spread forecast as a management tool in an operational setting. *SpringerPlus, 5*(1), 1205.

Plummer, F. G. (1912). *Forest fires: their causes, extent, and effects, with a summary of recorded destruction and loss* (Vol. 117): US Dept. of Agriculture, Forest Service.

Podur, J., Martell, D. L., & Csillag, F. (2003). Spatial patterns of lightning-caused forest fires in Ontario, 1976–1998. *Ecological Modelling, 164*(1), 1-20.

Preisler, H. K., & Ager, A. A. (2013). Forest-Fire Models. *Encyclopedia of Environmetrics*.

Preisler, H. K., Brillinger, D. R., Burgan, R. E., & Benoit, J. (2004). Probability based models for estimation of wildfire risk. *International Journal of Wildland Fire, 13*(2), 133-142.

Preisler, H. K., & Westerling, A. L. (2007). Statistical model for forecasting monthly large wildfire events in western United States. *Journal of Applied Meteorology and Climatology, 46*(7), 1020-1030.

Preisler, H. K., Westerling, A. L., Gebert, K. M., Munoz-Arriola, F., & Holmes, T. P. (2011). Spatially explicit forecasts of large wildland fire probability and suppression costs for California. *International Journal of Wildland Fire, 20*(4), 508-517.

Price, O., Borah, R., Bradstock, R., & Penman, T. (2015). An empirical wildfire risk analysis: the probability of a fire spreading to the urban interface in Sydney, Australia. *International Journal of Wildland Fire, 24*(5), 597-606.

Reed, W. J. (1999). Forest fires and oilfields as percolation phenomena. *http://www. math. uvic. ca/faculty/reed/*.

Reed, W. J. (2001). The Pareto, Zipf and other power laws. *Economics letters, 74*(1), 15-19.

Reed, W. J. (2006). A note on fire frequency concepts and definitions. *Canadian Journal of Forest Research, 36*(7), 1884-1888.

Reed, W. J. (2011). A flexible parametric survival model which allows a bathtub-shaped hazard rate function. *Journal of Applied Statistics, 38*(8), 1665-1680.

Reed, W. J. (2012). Power-Law Adjusted Survival Models. *Communications in Statistics-Theory and Methods, 41*(20), 3692-3703.

Reed, W. J., & Hughes, B. D. (2002). From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. *Physical Review E, 66*(6), 067103.

Reed, W. J., & Hughes, B. D. (2004). A model explaining the size distribution of gene and protein families. *Mathematical biosciences, 189*(1), 97-102.

Reed, W. J., & Jorgensen, M. (2004). The double Pareto-lognormal distribution—a new parametric model for size distributions. *Communications in Statistics-Theory and Methods, 33*(8), 1733-1753.

Reed, W. J., & McKelvey, K. S. (2002). Power-law behaviour and parametric models for the size-distribution of forest fires. *Ecological Modelling, 150*(3), 239-254.

Renouf, E., Dean, C. B., Bellhouse, D. R., & McAlister, V. C. (2016). Joint Survival Analysis of Time to Drug Change and a Terminal Event with Application to Drug Failure Analysis using Transplant Registry Data. *International Journal of Statistics in Medical Research, 5*(3), 198-213.

Rogeau, M.-P., & Armstrong, G. W. (2017). Quantifying the effect of elevation and aspect on fire return intervals in the Canadian Rocky Mountains. *Forest Ecology and Management, 384*, 248-261.

Ryan, K. C. (1991). Vegetation and wildland fire: implications of global climate change. *Environment International*, *17*(2-3), 169-178.

Schoenberg, F. P. (2004). Testing separability in spatial-temporal marked point processes. *Biometrics*, 471-481.

Schoenberg, F. P., Peng, R., & Woods, J. (2003). On the distribution of wildfire sizes. *Environmetrics, 14*(6), 583-592.

Scott, J. H. (2006). *An analytical framework for quantifying wildland fire risk and fuel treatment benefit.* Paper presented at the Andrews, PL, Butler, BW (Comps), Fuels Management-How to Measure Success: Conference Proceedings, March.

Stocks, B., Mason, J., Todd, J., Bosch, E., Wotton, B., Amiro, B., . . . Martell, D. (2002). Large forest fires in Canada, 1959–1997. *Journal of Geophysical Research: Atmospheres, 107*(D1).

Sullivan, A. L. (2009). Wildland surface fire spread modelling, 1990–2007. 1: Physical and quasi-physical models. *International Journal of Wildland Fire, 18*(4), 349-368.

Sullivan, A. L. (2009). Wildland surface fire spread modelling, 1990–2007. 2: Empirical and quasi-empirical models. *International Journal of Wildland Fire, 18*(4), 369-386.

Sullivan, A. L. (2009). Wildland surface fire spread modelling, 1990–2007. 3: Simulation and mathematical analogue models. *International Journal of Wildland Fire, 18*(4), 387-403.

Sun, C. (2013). Bivariate Extreme Value Modeling of Wildland Fire Area and Duration. *Forest Science*, *59*(6), 649-660.

Taylor, S. W., & Alexander, M. E. (2006). Science, technology, and human factors in fire danger rating: the Canadian experience. *International Journal of Wildland Fire, 15*(1), 121-135.

Taylor, S. W., Woolford, D. G., Dean, C., & Martell, D. L. (2013). Wildfire Prediction to Inform Management: Statistical Science Challenges. *Statistical science*, 586-615.

Therneau, T. M. T. M., & Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model* (0387987843). Retrieved from

Turner, R. (2009). Point patterns of forest fire locations. *Environmental and ecological statistics, 16*(2), 197-223.

Vega-Garcia, C., Lee, B., Woodart, P., & Titus, S. (1996). Applying neural network

technology to human-caused wildfire occurrence prediction. *AI applications, 10*(3), 9-18.

Vilar, L., Woolford, D. G., Martell, D. L., & Martín, M. P. (2010). A model for predicting human-caused wildfire occurrence in the region of Madrid, Spain. *International Journal of Wildland Fire, 19*(3), 325-337.

Wang, X., Parisien, M.-A., Taylor, S. W., Perrakis, D. D., Little, J., & Flannigan, M. D. (2016). Future burn probability in south-central British Columbia. *International Journal of Wildland Fire, 25*(2), 200-212.

Wang, Y., & Anderson, K. R. (2011). An evaluation of spatial and temporal patterns of lightning-and human-caused forest fires in Alberta, Canada, 1980–2007. *International Journal of Wildland Fire, 19*(8), 1059-1072.

Westerling, A., & Bryant, B. (2008). Climate change and wildfire in California. *Climatic Change, 87*(1), 231-249.

Westerling, A., Bryant, B., Preisler, H., Holmes, T., Hidalgo, H., Das, T., & Shrestha, S. (2011). Climate change and growth scenarios for California wildfire. *Climatic Change, 109*(1), 445-463.

Westerling, A. L., Hidalgo, H. G., Cayan, D. R., & Swetnam, T. W. (2006). Warming and earlier spring increase western US forest wildfire activity. *science, 313*(5789), 940-943.

Westerling, A. L., Turner, M. G., Smithwick, E. A., Romme, W. H., & Ryan, M. G. (2011). Continued warming could transform Greater Yellowstone fire regimes by mid-21st century. *Proceedings of the National Academy of Sciences, 108*(32), 13165-13170.

Wienke, A. (2010). *Frailty models in survival analysis*: CRC Press.

Woo, H., Chung, W., Graham, J. M., & Lee, B. (2017). Forest fire risk assessment using point process modelling of fire occurrence and Monte Carlo fire simulation. *International Journal of Wildland Fire, 26*(9), 789-805.

Wood, S. (2006). *Generalized additive models: an introduction with R*: CRC press.

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

Woolford, D., Braun, W., Dean, C., & Martell, D. (2009). Site-specific seasonal baselines for fire risk in Ontario. *Geomatica, 63*(4), 355-363.

Woolford, D. G., Cao, J., Dean, C. B., & Martell, D. L. (2010). Characterizing temporal changes in forest fire ignitions: looking for climate change signals in a region of the Canadian boreal forest. *Environmetrics, 21*(7-8), 789-800.

Woolford, D. G., Dean, C., Martell, D. L., Cao, J., & Wotton, B. (2014). Lightning-caused forest fire risk in Northwestern Ontario, Canada, is increasing and associated with anomalies in fire weather. *Environmetrics, 25*(6), 406-416.

Woolford, D. G, Wotton, B. M., Martell, D.L., McFayden, C., & Stacey A., et al. (2016). *Daily lightning- and person-caused fire prediction models used in Ontario.* Poster presented at Wildland Fire Canada Conf. 2016. http://www.wildlandfire2016.ca/wp-content/uploads/2017/03/McFayden-Fire-Occurence-Prediction-Poster-Ontario-2016-10-17V2Final.pdf

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

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

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

Xiong, Y. (2015). Analysis of Spatio-Temporal Data for Forest Fire Control.

Yoder, J., & Gebert, K. (2012). An econometric model for ex ante prediction of wildfire suppression costs. *Journal of Forest Economics, 18*(1), 76-89.

Chapter 3

3 Modeling the Duration and Size of Extended Attack Wildfires as Dependent Outcomes

3.1 Introduction

Two outcomes that have been studied extensively to quantify fire survivorship are the containment time and the area burned, commonly referred as duration and size. Although fire size and duration have been studied separately (e.g. Morin et al., 2015 for duration and Tremblay et al., 2018 for size), it is important to note that these are likely dependent outcomes and hence are prime candidates for *so-called* joint outcome analysis that allow for such potential for dependence. Indeed, very early work studying fire outcomes considered this concept of dependency; Beall (1949) showed graphically the relationship between time to control and fire size at control. There has also been early work on this relation by considering area burned as a function of time (Mcarthur, 1968; Van Wagner, 1969). Additionally, some authors have recently considered using multivariate distributions (Yoder and Gebert, 2012; Sun, 2013) or shared frailty Cox proportional hazards models (Bayham, 2013; Morin et al., 2019) to capture dependence in a variety of fire related outcomes. However, these models usually assume that outcomes are measured on the same scale, which is not suitable for jointly modeling time to containment and area burned when considering fire survivorship. As well, the inclusion of environmental variables and other information as covariates should be incorporated in any analysis related to these outcomes because of the substantial environmental influence on these outcomes. For detailed reviews of related work in fire science, see, for example, Taylor et al. (2013) and Xi et al. (2019).

Here we focus on providing novel approaches and insight for fire science through the adoption of two modern statistical frameworks that may be used to model dependence among multiple outcomes that are measured on different scales, while also accounting for the effects of covariates. One is the copula modeling framework, which has been used, for

instance, in Wu (2014), for linking the age and the mileage of automobiles in order to assess changes in warranty plans. Whereas alternative frameworks represent mileage as a function of age (e.g. Lawless et al., 1995), or use standard bivariate distributions (e.g. Pal and Murthy, 2003) for both outcomes, copula models have gained advantages in reliability analysis by offering flexibility in the types of tail dependence that can be accommodated in the joint distribution of the outcomes, as well as allowing the outcomes to take distinct marginal distributions (Genest and Favre, 2007). For a comprehensive overview of copulas, see for example, Nelsen (2006) and Joe (2014).

Since both containment time and area burned can be considered as survival outcomes, another suitable framework is an additive frailty modeling framework, which uses cluster-specific random effects (i.e. frailties) to incorporate variation that is common to the outcomes. Compared to the traditional frailty models (e.g. Hougaard, 2000; Therneau and Grambsch, 2000; Duchateau and Janssen, 2008; Wienke, 2010), which constrain frailties to have multiplicative effects on the hazard, the additive frailty model framework offers more flexible specification and interpretation for the frailties when the outcomes are measured with different scales or the hazard is not directly of interest. Using random effects to construct models in this way is commonly referred to as joint modeling. For recent publications of joint modeling in biostatistics and ecology, see, for example, Feng and Dean (2012), Renouf et al. (2016) and Juarez-Colunga et al. (2017).

In this chapter, we model the duration of a fire, in days, and its area burned, or size, in hectares, from two critical points in the life history of a fire: (1) ground attack, to (2) final control, for lightning-caused, extended attack fires in British Columbia (BC), Canada. A typical life history of fires in BC is characterized by critical points, including for example, the date and time of fire discovery, ground attack, final control, and mop-up. Here, ground attack and final control are the common origin and event, related to both survival outcomes regarding fire containment.

We extend univariate accelerated failure time (AFT) models for each of these outcomes using both a copula model framework and a joint model framework. AFT models assume

location-scale distributions for the outcomes, which align well with our knowledge of the distributions of duration and area burned during fire containment (DaCamara et al., 2014; Fernandes et al., 2016; Schoenberg et al., 2003; Reed and McElvey, 2002; Cumming, 2001; Butry et al., 2008; Holmes et al., 2008).

For the joint analysis of duration and size, we will consider the *Normal copula* and three *Archimedean copulas— Clayton*, *Gumbel*, and *Frank*. These forms of copulas have often been used in applications, and, as well, the three Archimedean copulas can be constructed from the traditional frailty models by assuming different distributions of the frailties, hence linking the two frameworks we consider here. For modeling the distribution of the frailties in the joint models, we will consider a *factor loading* (e.g. Feng and Dean, 2012) form and a *multivariate form* (e.g. Komárek and Lesaffre, 2008). Static location variables are included to incorporate topographical and temporal effects. Dynamic environmental variables are included by summarizing their trajectories. We consider whether these frameworks offer an improvement over utilizing univariate approaches for modeling the outcomes in terms of their model fits and predictabilities.

Section 2 outlines our proposed modeling frameworks. The data that motivate this research are discussed in Section 3. Section 4 discusses the analysis of the data, contrasting interpretation of the results in the context of the two modeling frameworks. Section 5 considers the effect of model misspecification. A direct comparison of the robustness of the two broad statistical frameworks under model misspecification is investigated by simulation. Section 6 closes with a discussion and recommendations.

3.2 Modeling and Estimation of Joint Outcomes

We develop two frameworks for modeling bivariate survival outcomes, one based on copulas, the other based on frailties, and we note immediately that generalization of the frameworks to multivariate or censored outcomes is straightforward. For a continuous outcome $t > 0$, we use $f(t)$ to denote its density function, $F(t) = \int_0^t f(x)dx$ to denote the probability that the outcome is less than or equal to t, termed the distribution function, and

$S(t) = 1 - F(t)$ to denote the probability that the outcome is greater than or equal to t, termed the survival function.

Let $t_{ik},\ k = 1,\ 2$ be the two outcomes considered in the fire science context, respectively, the duration and the size of fire $i = 1, \dots, n$, where the logarithm of the outcome, $y_{ik} = \log t_{ik}$, follows a location-scale distribution, and n is the number of observations. A univariate AFT model where outcomes are not linked can be written as

$$\log t_{ik} = \mu_k + \boldsymbol{\beta}_k^T \boldsymbol{x}_{ik} + \sigma_k \varepsilon_{ik},$$

where $\boldsymbol{x}_{ik} = \left(x_{ik1}, \dots, x_{ikR_k}\right)^T$ is a vector of R_k covariates associated with outcome k for fire i, $\boldsymbol{\beta}_k^T = \left(\beta_{k1}, \dots, \beta_{kR_k}\right)$ are the corresponding coefficients, ε_{ik} represents the random error, and μ_k and σ_k are the location and scale parameters associated with outcome k. We refer to model parameters as $\boldsymbol{\theta}_k = (\mu_k, \boldsymbol{\beta}_k, \sigma_k)^T$. Hence the survival function of outcome k is

$$S_k(t_{ik}|\boldsymbol{x}_{ik}, \boldsymbol{\theta}_k) = S_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k),$$

where $S_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k)$ is the survival function of the random error, termed the baseline survival function. Three common baseline survival functions discussed in the survival analysis literature in biostatistics (e.g. Lawless, 2011) and considered here are — the standard Gumbel, standard normal, and standard logistic, which correspond respectively to the Weibull, lognormal, and loglogistic distributions on the scale of the outcomes. Analogously, let $F_k(t_{ik}|\boldsymbol{x}_{ik}, \boldsymbol{\theta}_k) = F_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k) = 1 - S_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k)$ be the distribution function of outcome k, and let $f_k(t_{ik}|\boldsymbol{x}_{ik}, \boldsymbol{\theta}_k) = f_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k) = dF_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k)/\varepsilon_{ik}$ be the density function of outcome k, where $F_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k)$ and $f_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k)$ are termed respectively as the baseline distribution function and the baseline density function. Table 3.1 provides the parameterizations of the three distributions used here for modeling each of t_{ik} and ε_{ik}, noting that the model provides flexibility that the forms need not be identical for the two outcomes $k = 1, 2$.

Table 3.1: Distributions for the outcomes under the accelerated failure time (AFT) model

	Weibull	lognormal	loglogistic
$\lambda(\mu, \boldsymbol{\beta}, \sigma)$	$\exp\left[-\frac{\mu + \boldsymbol{\beta}^T \boldsymbol{x}_i}{\sigma}\right]$	$\mu + \boldsymbol{\beta}^T \boldsymbol{x}_i$	$-\frac{\mu + \boldsymbol{\beta}^T \boldsymbol{x}_i}{\sigma}$
$\nu(\sigma)$	$\frac{1}{\sigma}$	σ^2	$\frac{1}{\sigma}$
$f(t \mid \lambda, \nu)$	$\nu\lambda t^{\nu-1}\exp(-\lambda t^{\nu})$	$\frac{1}{t\sqrt{2\pi\nu}}\exp\left[-\frac{1}{2\nu}(\log t - \lambda)^2\right]$	$\frac{\exp(\lambda)\nu t^{\nu-1}}{[1+\exp(\lambda)t^{\nu}]^2}$
$S(t \mid \lambda, \nu)$	$\exp(-\lambda t^{\nu})$	$1 - \Phi\left(\frac{\log t - \lambda}{\sqrt{\nu}}\right)$	$\frac{1}{1+\exp(\lambda)t^{\nu}}$
$f_0(\varepsilon)$	$\exp[-(\varepsilon + e^{-\varepsilon})]$	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\varepsilon^2\right)$	$\frac{\exp(\varepsilon)}{[1+\exp(\varepsilon)]^2}$
$S_0(\varepsilon)$	$1 - \exp(-e^{-\varepsilon})$	$1 - \Phi(\varepsilon)$	$\frac{1}{1+\exp(\varepsilon)}$

To represent the two multivariate frameworks, we further define $\boldsymbol{t}_i = (t_{i1}, t_{i2})$, $\boldsymbol{t} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n)$, $\boldsymbol{y}_i = (y_{i1}, y_{i2})$, $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2})$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)$, $\boldsymbol{\mu} = (\mu_1, \mu_2)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$, and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\sigma})$. Framework-specific parameters will be defined later in the corresponding subsections. For both multivariate frameworks, model parameters are estimated by maximizing their posterior distribution through a Bayesian MCMC approach. We assume vague priors commonly used in the literature (see, for example, Feng and Dean (2012)), independent and identically distributed as: $\mu_k \sim N(0,1)$ $, k = 1, 2$, $\beta_{kr} \sim N(0, 100), k = 1, 2, r = 1, \ldots, R_k$, and $\sigma_k \sim U(0, 100)$ $iid, k = 1, 2$. The joint prior distribution, required for estimation of the model parameters, is

$$p(\boldsymbol{\mu})p(\boldsymbol{\beta})p(\boldsymbol{\sigma}) = \prod_{k=1}^{2} p(\mu_k)p(\beta_{k1}) \ldots p\left(\beta_{kR_1}\right)p(\sigma_k).$$

This product will be referred to in constructing the posterior distributions under both multivariate frameworks. Other priors will be required for specific models and these are identified in the subsections below.

The two frameworks and the models developed under each will be discussed below, using the following nomenclature: a digit with 1 representing copula models and 2 representing joint models, and a letter for the form these models take under each framework. See Tables 3.2 and 3.3 for details on the nomenclature.

3.2.1 The Copula Model Framework

The definition of copulas is stated in the following theorem taken from Sklar (1959):

Theorem 2.1 (Sklar's theorem): Let F be a continuous joint distribution function of the outcomes t_1 and t_2 with margins F_1 and F_2. The copula associated with F is a distribution function

$C(u_1, u_2|\delta)$: $[0,1]^2 \rightarrow [0,1]$ that satisfies

$$F(t_1, t_2) = C(F_1(t_1), F_2(t_2)|\delta).$$

Such C exists for all t_1 and t_2 and is termed a copula function or copula, with association parameter δ. Conversely, if C is a copula and $F_1(t_1)$ and $F_2(t_2)$ are distribution functions, then $F(t_1, t_2)$ is a joint distribution function.

Essentially, copulas link two univariate models as the marginal models of a multivariate framework by plugging their distribution functions into a copula function. Table 3.2 summarises the four different forms of copula function we consider.

Table 3.2: Parameterization of the copulas.

Form	Copula Function	Range of δ	Kendall's τ
Normal (1n)	$C(u_1, u_2\|\delta) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\|\delta)$ $c(u_1, u_2\|\delta) = (1-\delta^2)^{-1/2} \exp\left\{-\frac{x_1^2 + x_2^2 - 2\delta x_1 x_2}{2(1-\delta^2)}\right\} \exp\left\{\frac{x_1^2 + x_2^2}{2}\right\}$ where $x_k = \Phi^{-1}(u_k), k = 1,2$	$[-1, 1]$	$2\pi^{-1} \arcsin(\delta)$
Clayton (1c)	$C(u_1, u_2\|\delta) = \left(u_1^{-\delta} + u_2^{-\delta} - 1\right)^{-1/\delta}$ $c(u_1, u_2\|\delta) = (1+\delta)(u_1 u_2)^{-\delta-1}\left(u_1^{-\delta} + u_2^{-\delta} - 1\right)^{-2-\frac{1}{\delta}}$	$[0, \infty)$	$\delta/(\delta+2)$
Gumbel (1g)	$C(u_1, u_2\|\delta) = \exp\left\{-\left([-\log u_1]^\delta + [-\log u_2]^\delta\right)^{1/\delta}\right\}$ $c(u_1, u_2\|\delta) = C(u_1, u_2\|\delta)(u_1 u_2)^{-1}\left(\tilde{u}_1^\delta + \tilde{u}_2^\delta\right)^{-2+\frac{2}{\delta}}(\tilde{u}_1 \tilde{u}_2)^{\delta-1}\left[1 + (\delta-1)\left(\tilde{u}_1^\delta + \tilde{u}_2^\delta\right)^{-1/\delta}\right]$ where $\tilde{u}_k^\delta = -\log u_k, k = 1,2$	$[1, \infty)$	$(\delta-1)/\delta$
Frank (1f)	$C(u_1, u_2\|\delta) = \frac{-1}{\delta}\log\left(\frac{1 - e^{-\delta} - \left(1-e^{-\delta u_1}\right)\left(1-e^{-\delta u_2}\right)}{1-e^{-\delta}}\right)$ $c(u_1, u_2\|\delta) = \frac{\delta\left(1-e^{-\delta}\right)e^{-\delta(\mu_1+\mu_2)}}{[1 - e^{-\delta} - (1-e^{-\delta u_1})(1-e^{-\delta u_2})]^2}$	$(-\infty, \infty)$	$1 + \frac{4}{\pi}[D_1(\delta) - 1]$

Here $\boldsymbol{D_1(x) = x^{-1}\int_0^x t^1(e^t - 1)^{-1}}d\boldsymbol{t}$ f. See Nelson (1986) and Genest (1987).

To model dependence between two outcomes of AFT models using copulas, we make use of the multivariate extension of the survival function, discussed by He and Lawless (2005), in terms of distribution functions:

$$F(\boldsymbol{t}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = F_0(\boldsymbol{\varepsilon}_i|\boldsymbol{\theta}) = F_0(\varepsilon_{i1}, \varepsilon_{i2}|\boldsymbol{\theta}),$$

where $F(\boldsymbol{t}_i|\boldsymbol{x}_i, \boldsymbol{\theta})$ is the joint distribution of the outcomes, and $F_0(\boldsymbol{\varepsilon}_i|\boldsymbol{\theta})$ is the joint distribution function of the corresponding random errors. By the equation of copula, we have

$$F_0(\varepsilon_{i1}, \varepsilon_{i2}|\boldsymbol{\theta}) = C(F_{01}(\varepsilon_{i1}|\boldsymbol{\theta}_1),\ F_{02}(\varepsilon_{i2}|\boldsymbol{\theta}_2)|\delta).$$

Hence, the joint posterior distribution is expressed as:

$$p(\boldsymbol{\theta}, \delta|\boldsymbol{t}) \propto p(\boldsymbol{t}|\boldsymbol{\theta}, \delta)p(\delta)p(\boldsymbol{\mu})p(\boldsymbol{\beta})p(\boldsymbol{\sigma}).$$

The first term on the right-hand side of the above is the likelihood:

$$p(\boldsymbol{t}|\boldsymbol{\theta}, \delta) = \prod_{i=1}^{n}\left\{c(F_{01}(\varepsilon_{i1}|\boldsymbol{\theta}_1),\ F_{02}(\varepsilon_{i2}|\boldsymbol{\theta}_2)|\delta)\prod_{k=1}^{2} f_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k)\right\},$$

where $c(u,\ v|\delta) = d^2C(u,\ v|\delta)/du\ dv$, and $f_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k)$ is the density function of the random error. Note that the two error random variables for given i are not independent. However, their joint density can be shown to be written in this way.

We estimate parameters utilizing the approach described in Kelly (2007). The association parameter δ is assumed to follow a prior distribution of $U(0,1)$ for the Normal copula, $U(0,50)$ for the Clayton copula, $U(1,50)$ for the Gumbel copula and $U(0,50)$ for the Frank copula. Since different copulas have different ranges for δ, for comparing copulas, we will instead report their Kendall's τ (Kendall, 1938), a standardized value between -1 and 1 for measuring the ordinal association between two random variables. Table 3.2 provides the relationship between δ and τ for the copulas considered.

3.2.2 The Joint Model Framework

Another way to model the dependence between two outcomes of AFT models is by utilizing an additive frailty framework through what has been termed joint outcome modeling. For fire $i = 1, \dots, n$, outcome $k = 1$ for duration and $k = 2$ for size, the framework takes the general form by extending the univariate AFT model:

$$\log t_{ik} = \mu_k + \boldsymbol{\beta}_k^T \boldsymbol{x}_{ik} + b_{ik} + \sigma_k \varepsilon_{ik},$$

where $\boldsymbol{b}_i = (b_{i1}, b_{i2})^T$ is a random effect, independent of $\boldsymbol{\varepsilon}_i$, and assumed to be independent and identically distributed as $Q(\boldsymbol{b}_i|\boldsymbol{D}) = N_2(\boldsymbol{0}, \boldsymbol{D})$. The distribution of $\boldsymbol{b}_i$, often called the mixing distribution, is bivariate normal with a zero-mean 2×1 vector, $\boldsymbol{0}$, and a 2×2 symmetric and positive definite variance-covariance, $\boldsymbol{D}$. The form of $\boldsymbol{D}$ defines the dependence in the outcomes t_{i1} and t_{i2}. Before we consider the form of $\boldsymbol{D}$ in depth, we note that regardless of its form, the joint posterior distribution is expressed as

$$p(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{D}|\boldsymbol{t}) \propto p(\boldsymbol{t}|\boldsymbol{\theta}, \boldsymbol{b}) Q(\boldsymbol{b}|\boldsymbol{D}) p(\boldsymbol{\mu}) p(\boldsymbol{\beta}) p(\boldsymbol{\sigma}) p(\boldsymbol{D}),$$

where $\boldsymbol{b} = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_n)$, and $Q(\boldsymbol{b}|\boldsymbol{D})$ is the product of $Q(\boldsymbol{b}_i|\boldsymbol{D})$ over $i = 1, \dots, n$ by independence of $Q(\boldsymbol{b}_i|\boldsymbol{D})$ over i; $p(\boldsymbol{D})$ is the prior of $\boldsymbol{D}$. The first term on the right-hand side is the likelihood:

$$p(\boldsymbol{t}|\boldsymbol{\theta}, \boldsymbol{b}) \propto \prod_{i=1}^{n} f(\boldsymbol{t}_i|\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{b}_i),$$

where $f(\boldsymbol{t}_i|\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{b}_i)$ is the conditional joint density function of the outcomes given $\boldsymbol{b}_i$.

Various forms of $\boldsymbol{D}$ under the above framework have been discussed in the literature (e.g. He and Lawless, 2005; Duchateau and Janssen, 2007; Verbeke and Molenberghs, 2017). Table 3.3 summarises the three different forms, along with $f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_k, \boldsymbol{b}_i)$, the joint distribution of the outcomes on the logarithm scale, that align with the application considered as discussed in Section 3 and developed here. The factor

Table 3.3: Parameterization the joint models

Form	$\boldsymbol{b}_i$	$\boldsymbol{D}$	Model Constraints	$f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_k, \boldsymbol{b}_i)$
Factor Loading form (2a)	$\begin{bmatrix} b_{\mathrm{i}} \\ \gamma b_{\mathrm{i}} \end{bmatrix}$	$\begin{bmatrix} \sigma_b^2 & \gamma\sigma_b^2 \\ \gamma\sigma_b^2 & \gamma^2\sigma_b^2 \end{bmatrix}$	$\log t_{i1} = \mu_1 + \boldsymbol{\beta}_1^T \boldsymbol{x}_{i1} + b_i$ $\log t_{i2} = \mu_2 + \boldsymbol{\beta}_2^T \boldsymbol{x}_{i2} + \gamma b_i + \sigma_2 \varepsilon_{i2}$	$N_2\left(\begin{bmatrix} \mu_1 + \boldsymbol{\beta}_1^T \boldsymbol{x}_{i1} \\ \mu_2 + \boldsymbol{\beta}_2^T \boldsymbol{x}_{i2} \end{bmatrix}, \begin{bmatrix} \sigma_b^2 & \gamma\sigma_b^2 \\ \gamma\sigma_b^2 & \gamma^2\sigma_b^2 + \sigma_2^2 \end{bmatrix}\right)$
Separate form (2s)	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	not applicable	$\log t_{i1} = \mu_1 + \boldsymbol{\beta}_1^T \boldsymbol{x}_{i1} + \sigma_1 \varepsilon_{i1}$ $\log t_{i2} = \mu_2 + \boldsymbol{\beta}_2^T \boldsymbol{x}_{i2} + \sigma_2 \varepsilon_{i2}$	$N_2\left(\begin{bmatrix} \mu_1 + \boldsymbol{\beta}_1^T \boldsymbol{x}_{i1} \\ \mu_2 + \boldsymbol{\beta}_2^T \boldsymbol{x}_{i2} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right)$
Multivariate form (2m)	$\begin{bmatrix} b_{\mathrm{i}1} \\ b_{\mathrm{i}2} \end{bmatrix}$	$\begin{bmatrix} \sigma_{b1}^2 & \sigma_{b12}^2 \\ \sigma_{b12}^2 & \sigma_{b2}^2 \end{bmatrix}$	$\log t_{i1} = \mu_1 + \boldsymbol{\beta}_1^T \boldsymbol{x}_{i1} + b_{i1}$ $\log t_{i2} = \mu_2 + \boldsymbol{\beta}_2^T \boldsymbol{x}_{i2} + b_{i2}$	$N_2\left(\begin{bmatrix} \mu_1 + \boldsymbol{\beta}_1^T \boldsymbol{x}_{i1} \\ \mu_2 + \boldsymbol{\beta}_2^T \boldsymbol{x}_{i2} \end{bmatrix}, \begin{bmatrix} \sigma_{b11}^2 & \sigma_{b12}^2 \\ \sigma_{b12}^2 & \sigma_{b22}^2 \end{bmatrix}\right)$

loading form whereby $b_{i1} = b_i$, and $b_{i2} = \gamma b_i$ is a modification of the traditional shared frailty model and has been applied in the joint modeling studies discussed earlier. In Table 3.3, the distributional form labelled 2a uses a factor loading framework where the parameter γ accounts for the different scale of the effect of the frailty term b_i on the two outcomes. The term b_i can be viewed as a fire-specific error, shared commonly and additively on the logarithm of both outcomes. Note that one of the outcome specific errors, ε_{ik}, is set to zero (variance set to zero) to avoid over-parameterization. Under this form, having σ_b significantly different from zero suggests that there is dependence between the two outcomes, and having γ significantly different from 1 suggests that such b_i are acting on the outcomes with different scales. Hence the dependence between the outcomes can be measured by, $\gamma^2\sigma_b^2/(\gamma^2\sigma_b^2 + \sigma_k^2) \times 100\%$, the percentage of heterogeneity explained by the shared component, similar to the intraclass correlation coefficient for random effect models (Faraway, 2006). We assume that the outcomes are independent given their shared frailties, thus the right-hand side of equation above can be simplified as:

$$\prod_{i=1}^{n} f(\boldsymbol{t}_i|\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{b}_i) = \prod_{i=1}^{n}\prod_{k=1}^{2} f_k(t_{ik}|\boldsymbol{x}_{ik}, \boldsymbol{\theta}_k, b_{ik}) = \prod_{i=1}^{n}\prod_{k=1}^{2} f_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k, b_{ik}),$$

where $f_k(t_{ik}|\boldsymbol{x}_{ik}, \boldsymbol{\theta}_k, b_{ik})$ is the conditional density function of outcome k given b_{ik} with $f_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k, b_{ik})$ as the conditional baseline density. We utilize vague priors, with $\sigma_b \sim U(0,100)$, $\gamma \sim N(0,100)$, and $p(\boldsymbol{D}) = p(\sigma_b)p(\gamma)$.

We also consider a model where the outcomes are not dependent, called the separate form (Table 3.3, distribution form labelled 2s), where no linking frailty is introduced, and each outcome has its own outcome-specific error σ_1 and σ_2. Hence b_{i1} and b_{i2} are identically zero and the conditional likelihood (11) is proportional to:

$$\prod_{i=1}^{n} f(\boldsymbol{t}_i|\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{b}_i) = \prod_{i=1}^{n}\prod_{k=1}^{2} f_k(t_{ik}|\boldsymbol{x}_{ik}, \boldsymbol{\theta}_k) = \prod_{i=1}^{n}\prod_{k=1}^{2} f_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k),$$

where $f_k(t_{ik}|\boldsymbol{x}_{ik}, \boldsymbol{\theta}_k)$ is the density function of outcome k with $f_{0k}(\varepsilon_{ik}|\boldsymbol{\theta}_k)$ as the baseline density. Models constructed under this form will be used to contrast the fits of, and assess the benefits of, joint models.

Alternatively, the multivariate form (Table 3.3, distribution form labelled 2m), a modification of the correlated frailty model (Wienke, 2011) that has been considered, for example, in the cluster-specific AFT model by Komárek and Lesaffre (2008), may be used here. Under this form, the frailties b_{i1} and b_{i2}, follow a multivariate normal distribution with covariance taking a non-zero value. The correlation of the frailties is defined by $\rho = D_{12}/\sqrt{D_{11}D_{22}}$, where D_{11} and D_{22} each represents the variance of the b_{i1} and b_{i2}, and D_{12} represents their covariance. Having ρ significantly different from zero suggests that there is dependence between the two outcomes. For priors, we further assume that $\boldsymbol{D} \sim \text{Wishart}(2, \boldsymbol{R})$, where $\boldsymbol{R}$ is a 2×2 matrix such that $R_{11} = 0.01, R_{22} = 0.1, R_{12} = R_{21} = 0$.

3.3 British Columbia Fire Data

3.3.1 Data Description

Our work is motivated by an interest in understanding the relationship, if any, between fire size and fire duration as well as the effect of environmental variables on theses outcomes. Records of wildland fires that occurred in British Columbia from 1953-2000 were obtained from the BC Wildfire Service. Weather variables for the same period were obtained from an unpublished Canadian Forest Service study (Flannigan et al., 2002). The weather observations (noon temperature, relative humidity, wind speed, 24-hour precipitation) in this analysis were obtained from the Meteorological Service of Canada (MSC) and BC Wildfire Service weather stations for the 1953-1970 and 1971-2000 periods, respectively, and were interpolated to a 5 km grid using inverse distance weighting (temperature and relative humidity were corrected for elevation); the interpolated weather observations were subsequently used to calculate the six standard indices of the Canadian Forest Fire Weather Index System (Van Wagner, 1987). We used the location of the centroid of the fires

(latitude, longitude) to extract the observations for each day of each fire's life history from the appropriate grid cell.

The data contain the following information on historical fire activity from 1953-2000 and are described in Table 3.4: duration in days and size in hectares, six static location variables, and ten interpolated dynamic environmental variables recorded daily through the complete life history of the fires. The ten environmental variables include four weather observations and six indices calculated from the weather observations. The four weather observations include temperature (TEMP), wind (WIND), relative humidity (RH), and Precipitation (PCP). The six indices include three fuel moisture codes and three fire behavior indices. The former contains Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), and Drought Code (DC), which increase as the dryness of the corresponding layer of the forest floor increases. The latter contains Initial Spread Index (ISI), Buildup Index (BUI), and Fire Weather Index (FWI), which increase as the fire spread rate, the available fuel, and the intensity of the fire-line increases correspondingly. The ten environmental variables are also functionally related in a hierarchical structure (Natural Resources Canada, 2017). There will be need for care in employing these variables because of potential multicollinearity. De Groot (1998), Lawson and Armitage (2008) and Wotton (2008) provide an excellent review of the scientific interpretation of these variables. We validate the estimates of the linear effect of the by comparing with scientific knowledge.

For this project, we are interested in the 911 lightning-caused, extended attack fires. Extended attack fires are those that have escaped initial attack and for which duration exceeds 2 days and size exceeds 4 hectares, and therefore require additional resources to contain. These fires account for around 93% total area burned by lightning fires, and a large percentage of suppression costs and damage. The left panel of Figure 3.1 is a plot of the locations of these fires. Fire occurrence is more severe along a ridge from the north-east to the south-west of the province. This occurs because there is a high density of lightning strikes and lightning caused fires in south east BC (Magnussen and Taylor, 2013). The right panel of Figure 3.1 is a scatter plot of duration versus size, with a log base 10 scale on both

Table 3.4: Data used in the study

Outcomes and Variables	Descriptions
Outcomes:	
Duration (days)	Time spent from ground attack to final control
Size (ha)	Area burned from ground attack to final control
Location Variables:	
Slope (degree)	Steepness of the landscape
Elevation (m)	Height above sea level
Ground attack size (ha)	Burned area at the ground attack stage
Fire centre	Administrative regions of the province, coded as: Coastal, Northwest, Prince George, Kamloops, Southeast, Cariboo
Decade	Decades that the fires occur
Month	Months that the fires occur
Weather Observations:	
Temperature (TEMP; °C)	The noon temperature recorded in Celsius
Wind (WIND; km/h)	The average wind speed measured over a 10-minute period
Relative Humidity (RH; %)	The fraction of moisture present in the atmosphere
Precipitation (PCP; mm)	The amount of rain accumulated in the 24-hour period from noon to noon

Outcomes and Variables	Descriptions
Fuel Moisture Codes:	
Fine Fuel Moisture Code (FFMC)	An index of the moisture content of litter and other cured fine fuels
Duff Moisture Code (DMC)	An index of the moisture content of loosely compacted organic (duff) layers of moderate depth
Drought Code (DC)	An index of the moisture content of deep, compact organic layers
Fire Behavior Indices:	
Initial Spread Index (ISI)	A relative measure of the expected rate of fire spread, which combines the effects of wind and Fine Fuel Moisture Code
Buildup Index (BUI)	A weighted combination of Duff Moisture Code and Drought Code, a relative measure of the total amount of fuel available for combustion
Fire Weather Index (FWI)	A combination of the Initial Spread Index and Buildup Index, a relative measure of the potential intensity of a spreading fire as energy output rate per unit length of fire front

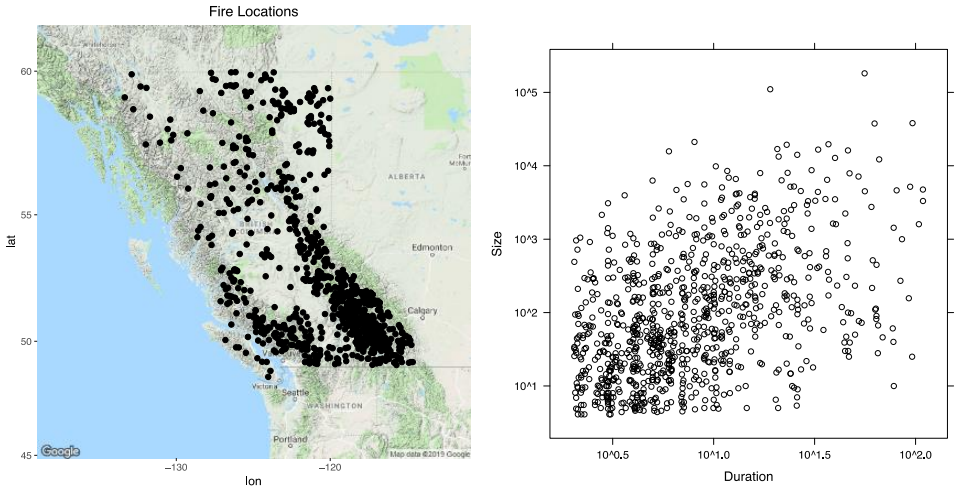


Figure 3.1: The locations of the fires (left) and a scatter plot of duration versus size, with a log base 10 scale on both axes (right). Fires are clustered around the Rocky Mountain Trench. Duration and size have a moderate positive dependence.

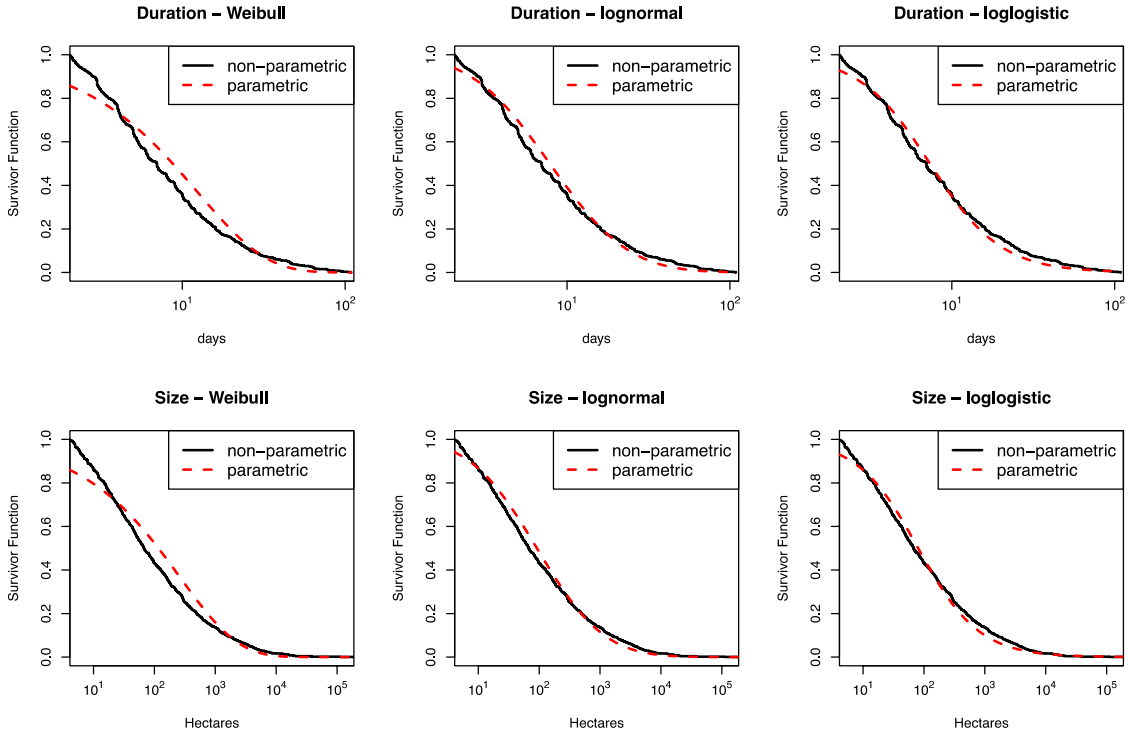


Figure 3.2: The estimated parametric and nonparametric survivor functions of the outcome. The lognormal distribution seems to fit both outcomes well.

axes. There are more short-and-small extended attack fires than long-and-large such fires, and a moderate positive correlation seems to exist between the outcomes. Figure 3.2 has six plots; each plot contains the estimated parametric and nonparametric survival functions of the outcome based on the parametric forms mentioned in the previous section and not accounting for covariates. The top row corresponds to the outcome of duration and the bottom row correspond to the outcome of size. The columns, from left to right, are the plots for estimated Weibull, lognormal, and loglogistic distributions. The parametric forms are estimated using maximum likelihood (Lawless, 2011) using the R package `survival` (Therneau and Lumley, 2014) as a simple exploratory analysis. For $t \geq 0$, the nonparametric estimate, the empirical survival function (Lawless, 2011), is calculated as $\hat{S}(t) = [\text{Number of observations } \geq t]/n$. The lognormal distribution seems to fit both outcomes reasonably well and is used for each outcome in the two multivariate frameworks. To illustrate the dynamic environmental variables, Figure 3.1 plots the trajectories of the ten environmental variables from 30 randomly chosen fires. These trajectories illustrate the high variability in these variables as well as the sharp changes that may be exhibited in some variables showing how susceptible they are to changes in moisture and wind speed.

3.3.2 Construction of Derived Covariates

For fire i outcome k, we partition its associated R_k covariates $\boldsymbol{x}_{ik}$ as $\boldsymbol{x}_{ik} = \left(\boldsymbol{x}_{ik}^P, \boldsymbol{x}_{ik}^Q\right)^T$, in which $\boldsymbol{x}_{ik}^P = \left(x_{ik1}^P, \dots, x_{ikP_k}^P\right)^T$ is the vector of P_k static covariates representing the six location variables, and $\boldsymbol{x}_{ik}^Q = \left(x_{ik1}^Q, \dots, x_{ikQ_k}^Q\right)^T$ is a set of Q_k derived covariates constructed by summarizing the trajectories of the corresponding environmental variables into relevant indices. Precisely, a derived covariate $x_{ikq}^Q, q = 1, \dots, Q_k$, is defined as $g\left(X_{ikq}^Q\right)$, where $X_{ikq}^Q = \left\{x_{ijkq}^Q, j = 1, \dots, m_i\right\}$ denotes the complete history of the observed dynamic variables, and g is a function that summarizes the history over its m_i observations. For all Q derived covariates of fire i, outcome k, we define

$$x_{ikq}^{Q} = g\left(X_{ikq}^{Q}\right) = \frac{\sum_{j=1}^{m_i}\left(x_{ijkq}^{Q} - x_{i0kq}^{Q}\right)}{m_i},$$

where x_{i0kq}^{Q} is a threshold value for the associated variable, based on scientific knowledge. These thresholds are either weather conditions in a normal day of July (Van Wagner, 1987) or critical values of the fuel moisture codes (Stocks et al., 1989) and fire behavior indices (Podur and Wotton, 2011) which the intensity of fire activity increases. For example, days in which FWI>19 are considered "spread event days" that have the potential to yield large fire size. The quantity, x_{ikq}^{Q}, can be interpreted as the average deviation from threshold (ADFT) over the complete history of the fire. In our nomenclature, the threshold is appended to the variable name; for example, DC400 refers to the effect of the average deviation from the threshold of 400 for the variable DC. Table 3.5 summarizes the covariates used in the study and identifies the thresholds through the ADFT variable names. The threshold value for each environmental variable is also plotted in red dashed lines on Figure 3.3. Additionally, for those environmental variables that demonstrate clear linear trend in their trajectories (DC, DMC, BUI), we centre them, fit a simple linear regression model by maximum likelihood from the first day (day of ground attack) to the last day (day of final control) of the fire, and use the estimated intercept and slope as additional covariates. Table 3.5 also summarizes the estimated slopes and intercepts. Note that around 86%, 60%, and 62% of the estimated slopes were positive for the regression analysis for DC, DMC, and BUI. The derived covariates used in this study are then: summaries of the linear trends, ADFT of weather observations, ADFT of fuel codes, and ADFT of fire behaviour indices.

Exploratory analyses indicate that summaries of the linear trends are generally not correlated with the other covariates, while the ADFT of the fuel codes are strongly correlated with the ADFT of the fire behavior indices. Fuel code covariates are generally positively correlated; the same is true for fire behaviour indices. These correlations have been noted by other authors for other fire data.

Table 3.5: Descriptive statistics of the covariates

Continuous Location Variables					
	Mean	SD	Min	Median	Max
Slope	57.8	30.6	0.0	60.0	99.0
Elevation	33.3	170.4	0.0	12.0	1900.0
Ground attack size	65.4	253.6	0.0	5.0	2952.5
ADFT of the Environmental Variables					
	Mean	SD	Min	Median	Max
TEMP21	-1.4	5.2	-21.5	-1.5	16.5
RH45	9.9	18.5	-26.7	8.3	55.0
WIND13	-3.3	3.9	-13.0	-3.7	15.2
PCP12	-11.0	1.3	-12.0	-11.4	-1.3
FFMC74	4.4	14.9	-74.0	9.3	22.8
DMC20	41.1	42.9	-20.0	30.8	246.3
DC400	21.1	134.6	-329.8	10.7	466.3
ISI7.5	-2.2	4.1	-7.5	-3.2	19.9
BUI50	33.8	49.3	-50.0	26.4	233.8
FWI19	-1.8	13.4	-19.0	-5.3	49.7

Categorical Location Variables		
Fire Centre	Duration	Size
Coastal	8.3	36.3
Northwest	9.6	281.9
Prince George	6.0	130.5
Kamloops	6.0	52.5
Southeast	8.4	58.5
Cariboo	5.0	131.5
Decade	Duration	Size
1950s	5.3	68.8
1960s	7.0	105.7
1970s	5.9	52.8
1980s	8.0	97.2
1990s	7.3	44.9
Month	Duration	Size
May	5.0	408.0
June	5.6	236.1
July	8.0	81.8
August	6.2	53.2
September	8.0	34.4

Linear Model Summaries		
	Median	Percent of Positive
DC intercept	-22.5	-
DC slope	6.0	0.85
DMC intercept	-2.7	-
DMC slope	0.6	0.60
BUI intercept	-4.8	-
BUI slope	1.0	0.62

The continunous location and the average deviation from threshold (ADFT) for the envionmental variables are summarized in terms of mean, standard deviation (SD), minimum (Min), median, and maximum (Max). The median duration and size of the

categorical location variables are summarized for each of their categories. The linear model estimates (i.e. estimated intercepts and slopes) are summarized by their medians and the percentage of estimated slope parameters that are positive is provided. With regard to duration, fires at Prince George, Kamloops, and Cariboo tend to be shorter than those in other regions. Fires from the 60's, 80's, and 90's, as well as fires in July tend to be long. With regard to size, fires at Northwest, Prince George, Cariboo, from the 60's and 80's, as well as in May and June tend to be larger.

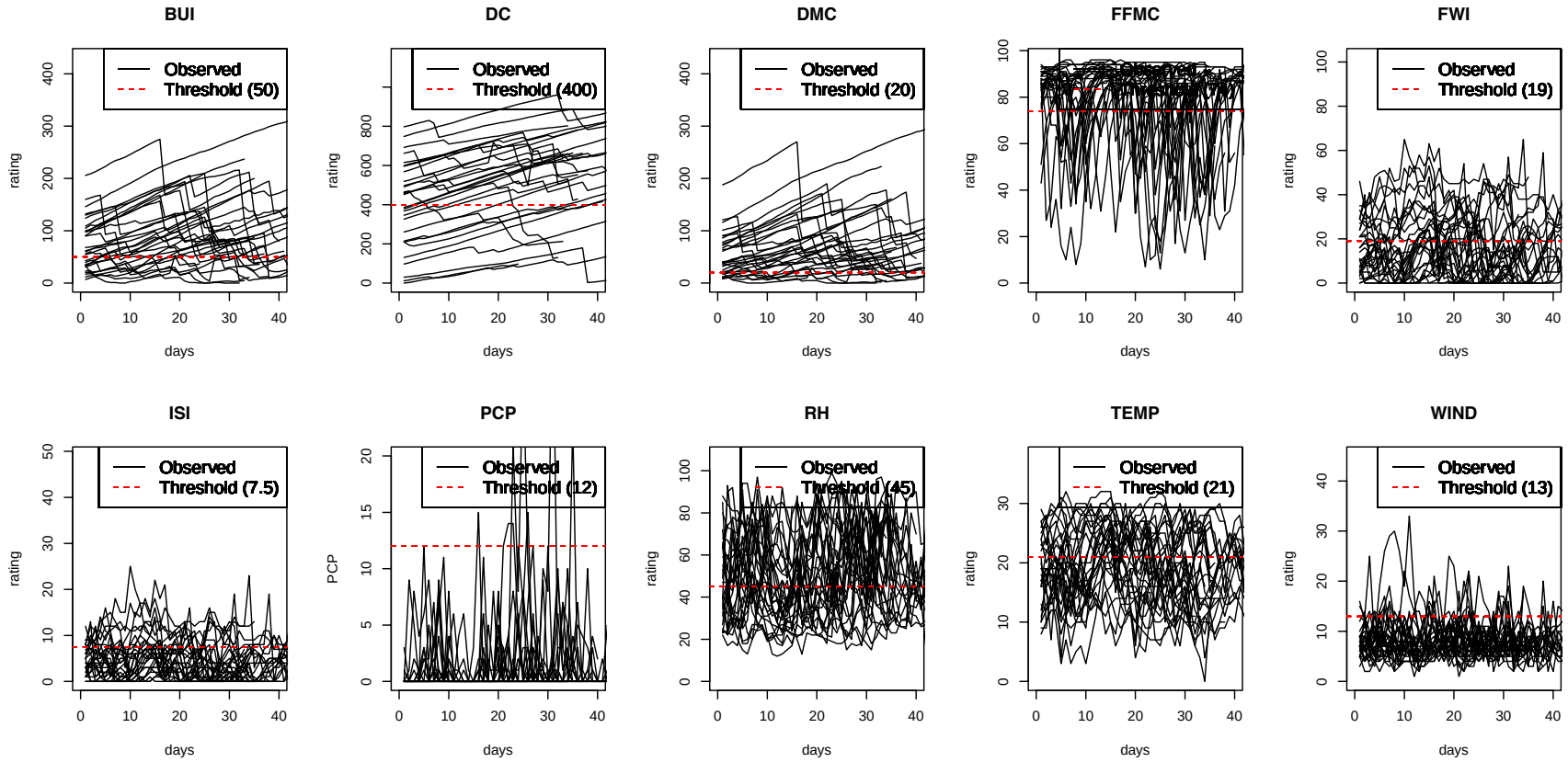


Figure 3.3: The trajectories of the environmental variables for 30 randomly chosen fires. The threshold values are plotted in dashed lines. BUI = Buildup Index; DC = Drought Code; DMC = Duff Moisture Code; FFMC = Fine Fuel Moisture Code; FWI = Fire Weather Index; ISI = Initial Spread Index; PCP = precipitation; RH = relative humidity; TEMP = temperature; WIND = windspeed.

3.4 Analysis and Results

Model fitting is carried out by adaptive MCMC using the R package runjags (Denwood, 2016) with 3 chains. Each chain has 10000 adapting steps, 5000 burn-in steps, and 30000 steps thinned at 4. The parameter estimate is its posterior median. Convergence is assessed by visually examining chain trajectories and density plots of the sampled parameter values, as well as by calculating the Gelman-Rubin statistic (Gelman and Rubin, 1992). Autocorrelations of the values of the chains are plotted to assess if the chains are of sufficient length. Chains are run on parallel hardware to improve computational efficiency. Credible intervals are obtained as the lower/upper 2.5% quantiles of the posterior density. Covariate identification proceeded by forward selection. Model fits are assessed using the Deviance information criteria (DIC) by Spiegelhalter et al. (2002) and the Watanabe–Akaike information criterion (WAIC) by Watanabe (2010). The WAIC uses the computed log pointwise posterior predictive density and adds a correction for effective number of parameters to account for overfitting.

The final models, discussed in depth in this section, utilize the normal form with the copula model (1n), and the factor loading and multivariate forms of the joint model (2a and 2m, respectively). With the copula model, the normal form outperforms the other forms in terms of DIC and WAIC. Additionally, the factor loading form of the joint model outperforms the separate form. The full posterior distributions of the final models are provided in Appendix 3A. A summary of the fit of all models considered is provided in Appendix 3B.

Tables 3.6 and 3.7 present parameter estimates (95% credible intervals) obtained from fitting the three models and resulting from the selection procedure for the normal copula model, the factor loading model, and the multivariate model. We include the static covariates, and employ a forward selection procedure for each of the four categories of derived covariates as defined earlier: summaries of the linear trends, ADFT of weather observations, ADFT of fuel codes, and ADFT of fire behaviour indices. Covariate effects Table 3.6 identifies that for model 1n, the dependence between the outcomes is captured

Table 3.6: Posterior estimates of model parameters and statistics accessing model fits of the three dependent models

	Model 1n					
	Model Parameters			**Model Fits**		
	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$			
μ_1	1.049	1.525	1.977		DIC	18238791
μ_2	0.607	1.857	3.058		WAIC	168
σ_1	1.295	1.363	1.437			
σ_2	2.910	3.065	3.234			
τ	0.573	0.602	0.630			
	Model 2a					
	Model Parameters			**Model Fits**		
	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$			
μ_1	0.839	1.437	1.998		DIC	14728
μ_2	0.426	1.261	2.340		WAIC	17130
σ_2	0.017	0.125	0.442			
γ	3.741	4.386	4.918			
σ_b	0.377	0.429	0.501			
	Model 2m					
	Model Parameters			**Model Fits**		
	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$			
μ_1	1.089	1.563	2.015		DIC	5688
μ_2	0.764	1.951	3.068		WAIC	5596
σ_{11}	0.636	0.697	0.767			
σ_{12}	0.722	0.833	0.954			
σ_{22}	3.209	3.519	3.862			

For model 1n, the dependence between the outcomes is captured by a Normal copula with a moderate τ estimated as 0.602 with 95% credible interval $(0.573,\ 0.630)$. For model 2a, the shared error, σ_b , is estimated as 0.429 with 95% credible interval $(0.377, 0.501)$and attached to a factor loading of 4.386 with 95% credible interval $(3.741, 4.918)$on size. For model 2m, the correlation between the frailties, ρ, is estimated as 0.283 with 95% credible interval $(0.255,\ 0.308)$. These three models yield the best fit among all the other model candidates (see Appendix 3B).

Table 3.7: Posterior estimates of the covariate effects for model 2a.

	Duration Coefficients			Size Coefficients		
	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$
Slope	-0.001	0.001	0.003	-0.008	-0.004	0.001
Elevation	0.000	0.000	0.000	0.000	0.001	0.001
Ground attack size	0.000	0.000	0.000	0.001	0.001	0.002
Northwest	-0.098	0.160	0.405	0.887	1.479	2.076
Prince George	-0.458	-0.272	-0.064	0.621	1.059	1.507
Kamloops	-0.511	-0.312	-0.088	-0.471	0.019	0.522
Southeast	-0.221	-0.053	0.143	-0.283	0.144	0.568
Cariboo	-0.701	-0.440	-0.160	-0.164	0.463	1.096
1960s	0.097	0.319	0.531	-0.109	0.390	0.888
1970s	-0.042	0.197	0.444	-0.081	0.435	0.966
1980s	0.249	0.474	0.697	0.053	0.546	1.053
1990s	0.318	0.571	0.804	-0.292	0.253	0.796
June	-0.013	0.351	0.712	-0.499	0.392	1.305
July	0.222	0.517	0.848	-0.659	0.149	0.977
August	0.030	0.331	0.676	-0.920	-0.103	0.749
September	-0.101	0.441	0.945	-1.168	0.018	1.249
BUI intercept	0.006	0.010	0.014			
BUI slope	0.029	0.047	0.066			
WIND13				-0.006	0.025	0.057
PCP12				-0.252	-0.168	-0.088
DMC20				0.004	0.007	0.010
DC400	0.000	0.000	0.001			

Parameter estimates ($Q_{.500}$) are reported as well as the lower limit ($Q_{.025}$) and upper limit ($Q_{.975}$)of their 95% credible intervals. Dominant covariate effects are highlighted in green. Effects of fire centre, decade, and month roughly agree with our findings in Table 3.5, except that Cariboo, the decade of the 60's and month are not significant for modeling size. BUI intercept and BUI slope are positively related to duration. Ground attack size and the average deviation from threshold (ADFT) for DMC are positively related to size, while the ADFT for PCP is negatively related. Results are reported in comparison to the reference group (i.e. at Coastal, from the 50's, and in May). Though not shown here, results related to these covariates are about the same for the three dependent models (i.e. 1n, 2a, 2m).are nearly identical across the three dependent models, thus only the effects corresponding to model 2a are presented in Table 3.7 to avoid redundancy.

by a Normal copula with a moderate τ estimated as $0.602\ (0.573, 0.630)$. For model 2a, the error shared across both outcomes is significant with an estimate of σ_b as $0.429\ (0.377,0.501)$. The factor loading parameter γ is estimated as $4.386\ (3.741,4.918)$, suggesting that the effect of the shared error on the logarithm of size is about four times as large as its effect on the logarithm of duration. Furthermore, the size-specific error, σ_2, estimated as $0.125\ (0.017,0.442)$, is quite small compared to the shared outcome error. As a result, about 99.6% of heterogeneity in size is explained by the shared variability. For model 2m, the correlation between the frailties, ρ, is estimated as $0.283\ (0.255, 0.308)$. The dependence across the two outcomes is significant for all three models as described by these parameter estimates.

Dominant covariate effects are highlighted in green in Table 3.7 and summarized as follows:

- *Ground attack size*: Size (the difference in fire size between ground attack and final control) tends to be larger as the size at ground attack increases.
- *Fire centre*: Compared to the fires from the Coastal fire centre, fires from Prince George, Kamloops, and Cariboo fire centres tend to have shorter duration while fires from Northwest and Prince George fire centres tend to have larger sizes.
- *Decade*: Fires in recent decades tend to have considerably longer duration and larger sizes compared the ones in the 1950's, except for the 70's for duration and the 90's for size.
- *BUI intercept and BUI slope*: In our fire data, one standard deviation in the distribution of the estimated BUI intercept is about 20 units. In the Canadian Fire Behaviour Prediction System (CFS Fire Danger Group 1992) the effect of BUI on fuel consumption varies non-linearly by fuel (forest) type. For example, increasing BUI by 20 units in the C-3 jackpine-lodgepole pine fuel type represents an increase in surface fuel consumption of 0.78 kg/m^2 at BUI 40 and 0.11 kg/m^2 at BUI 200 when most of the surface fuel will have been consumed. However, this does not account for the effect of BUI and fuel consumption on crowning. Such a change in the fuel for combustion will multiply duration by $\exp(20 \times 0.01) = 1.22$. A fire

with an initial BUI of 150 would have a duration 2.77 times that of a fire at 50 BUI. Likewise, one standard deviation in the distribution of the estimated BUI slopes is about 5 units, which represents a day-over-day increase in surface fuel consumption of 0.16 kg/m^2 at BUI 80 and will multiply duration by $\exp(5 \times 0.047) = 1.26$. That is, such changes in initial and day-over-day fuel combustion will result in a 22% and 26% increase in duration, respectively.

- *The average deviation from threshold (ADFT) for precipitation and DMC*: Increasing the ADFT of precipitation by 1mm will multiply size by $\exp(-0.168) = 0.85$ (See Table 3.5 for a summary of ADFT values for precipitation in our data). Increase in the ADFT of DMC by 10 units will multiply size by $\exp(10 \times 0.007) = 1.07$. These results quantify, through the lifetime of fires, how precipitation leads to smaller fire sizes, and how dryer organic layers at moderate depth will lead to larger fire sizes. Importantly, the small change in the ADFT of precipitation substantially affects fire size.

Figure 3.4 presents the histogram of the standardized residuals and a plot of the standardized residuals vs. fitted value for each of the two outcomes based on model 2a. For both outcomes, residuals are distributed around zero with no extreme outliers. The right skewness of the histograms and the heteroscedasticity observed in the plot of residuals versus fitted values suggest that the variability of residuals is increasing as the outcomes become large, an effect that will be discussed later.

3.5 Robustness under Joint Modeling

Joint modeling offers a helpful framework for interpreting the relationship between the two outcomes considered here in the fire science context. Even so, it is of interest to determine how robust joint models are when the true model is a copula, and to assess whether the factor loading form captures the variability for each of the outcomes. We consider a simulation study to investigate if the joint outcome models can effectively describe data

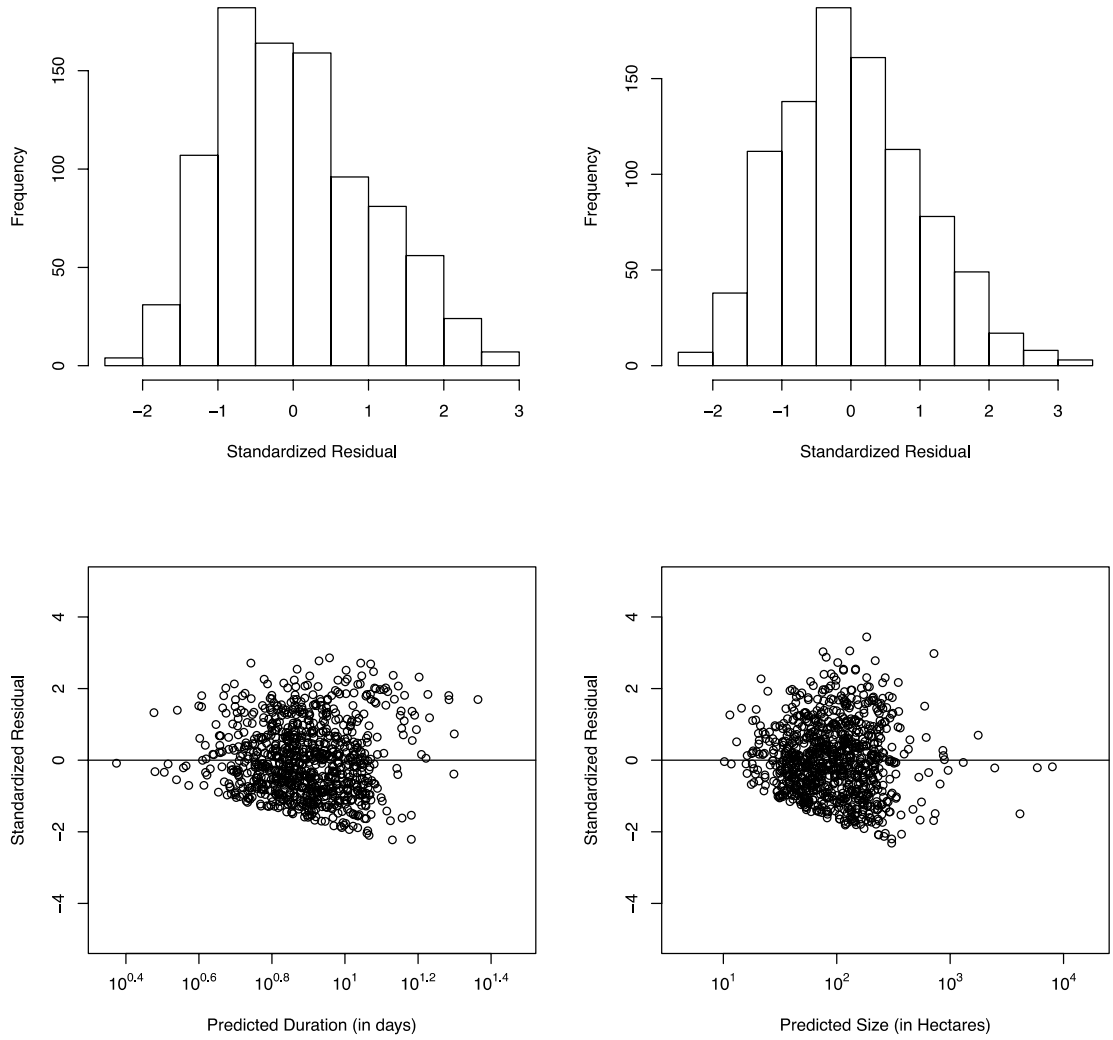


Figure 3.4: Residual diagnostics for the final models. For both duration and size, the standardized residuals are roughly normal with no significant outliers. The straight edges along the bottom of the points arise from the truncation at duration >2 days and size > 4 hectares. These features can be identified among all three final models (1n, 2a, 2m).

generated from the four copulas discussed earlier.

We generate n observations from each of the four copulas with the marginal for each outcome distributed as either Weibull, lognormal, or loglogistic. Data are generated from copula models using the conditional approach (Frees and Valdez, 2014; Hofert et al., 2014). This approach randomly generates t_{i1} from $F_1(t_{i1}|\boldsymbol{x}_{i1}, \boldsymbol{\theta}_1), i = 1, \dots, n$ using the inverse method, and generates t_{i2} from the conditional distribution of t_{i2} given t_{i1}. Parameters are set as $\mu_1 = 2.0$, $\mu_2 = 4.5$, $\sigma_1 = 1.0$, $\sigma_2 = 2.0$, and we incorporate a single covariate x for both outcomes with $x_{i1} = x_{i2} = 0$ for $i = 1, \dots, n/2$ and $x_{i1} = x_{i2} = 1$ for $i = n/2, \dots, n$. The true covariate effects are $\beta_1 = 0.100$, $\beta_2 = 0.075$; n is set at 200. These parameter values generate outcomes that are about the same scale and variability as the fire data. A range of values for the association parameter for the copula was considered with $\tau = 0.1, 0.2, \dots, 0.9$. One hundred data sets were generated at each of the 216 combinations of the parameter values.

Here we focus on understanding the decomposition of the variability in each outcome under the joint model and how the decomposition is affected by the changes in the value of the association parameter for the copula model. Table 3.8 and the left panel of Figure 3.5 identify the 2.5% and 97.5% quantile of the distribution of the estimates of σ_2, γ, and σ_b when data are generated from the normal copula with both margins as lognormal. The medians of these estimates are also identified in Table 3.8. When the dependence between the outcomes approaches zero, the copula function will converge to the *independence copula*, which is the same model as the joint model with no dependence (the separate form of the model). In this case, the median of the estimates of σ_2 is about 2 while that for σ_b is 1, accurately capturing the variability of the marginal distributions of the outcomes. As the association parameter τ increases, the shared variability increases. The distribution of the estimates of γ and σ_2 become narrower, while that for σ_b remains about the same. As τ increases to one, the median of estimates of σ_b remain approximately 1. The estimates of σ_2 decrease substantially , while the estimates of γ increases to about 2, again capturing the

Table 3.8: The lower limit $(Q_{.025})$, median $(Q_{.500})$ and the upper limit $(Q_{.975})$ of the distribution of the joint model estimates for data generated from the normal copula with both margins as lognormal. As the association parameter τ increases, the shared variability increases, the distribution of the estimates of γ and σ_2 become narrower, while that for σ_b remains about the same. Joint models also provide robust location parameters and coefficient estimates (see next page).

		Kendall's τ								
Parameter	Summary Statistic	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\mu_1 = 2.0$	$Q_{.025}$	1.779	1.763	1.737	1.711	1.706	1.698	1.704	1.717	1.715
	$Q_{.500}$	1.967	1.952	1.939	1.927	1.932	1.889	1.902	1.903	1.913
	$Q_{.975}$	2.154	2.141	2.140	2.143	2.159	2.080	2.101	2.089	2.112
$\mu_2 = 4.5$	$Q_{.025}$	3.947	3.998	3.903	3.895	3.899	3.838	3.891	3.936	3.935
	$Q_{.500}$	4.326	4.328	4.308	4.301	4.324	4.248	4.281	4.307	4.325
	$Q_{.975}$	4.704	4.658	4.713	4.707	4.748	4.658	4.671	4.677	4.714
$\beta_1 = 0.1$	$Q_{.025}$	-0.155	-0.108	-0.101	-0.094	-0.135	-0.070	-0.091	-0.096	-0.096
	$Q_{.500}$	0.123	0.149	0.177	0.182	0.160	0.227	0.196	0.187	0.183
	$Q_{.975}$	0.401	0.406	0.454	0.458	0.454	0.524	0.484	0.470	0.462
$\beta_2 = 0.075$	$Q_{.025}$	-0.389	-0.363	-0.424	-0.357	-0.523	-0.419	-0.443	-0.443	-0.447
	$Q_{.500}$	0.119	0.116	0.131	0.179	0.087	0.192	0.145	0.095	0.093
	$Q_{.975}$	0.627	0.595	0.687	0.716	0.696	0.803	0.733	0.633	0.633
σ_2	$Q_{.025}$	1.793	1.727	1.609	1.442	1.271	1.058	0.799	0.525	0.208
	$Q_{.500}$	1.980	1.912	1.790	1.609	1.418	1.164	0.899	0.588	0.247
	$Q_{.975}$	2.167	2.098	1.971	1.776	1.565	1.269	1.000	0.650	0.285
γ	$Q_{.025}$	0.098	0.331	0.638	0.960	1.241	1.472	1.682	1.827	1.952
	$Q_{.500}$	0.352	0.613	0.926	1.197	1.458	1.638	1.798	1.922	1.994
	$Q_{.975}$	0.607	0.895	1.213	1.434	1.674	1.804	1.915	2.017	2.037
σ_b	$Q_{.025}$	0.893	0.893	0.908	0.901	0.910	0.899	0.882	0.902	0.920
	$Q_{.500}$	0.997	0.998	0.999	1.003	0.999	0.990	0.994	1.004	1.013
	$Q_{.975}$	1.101	1.104	1.091	1.106	1.088	1.081	1.106	1.106	1.106

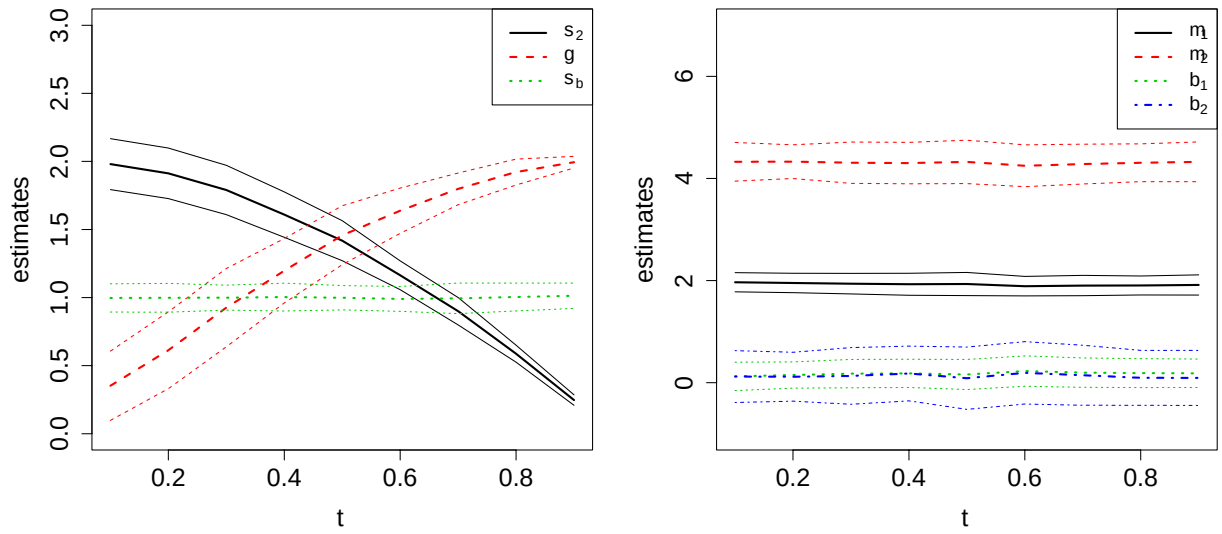


Figure 3.5: The lower limit, median, and the upper limit of the distribution of the joint model estimates of σ_2, γ, and σ_b (left panel) and μ_1, μ_2, β_1, β_1 (right panel) for data generated from the normal copula with both margins as lognormal. As the association parameter τ increases, the shared variability increases. The distribution of the estimates of γ and σ_2 becomes narrower, whereas that for σb remains about the same. Joint models also provide robust location parameters and coefficient estimates.

variability of the marginals. These features suggest that as the dependence between outcomes becomes stronger, the joint model captures such dependence through increased shared variability.

Though not presented here, these features hold for all six combinations of the three margins under all four types of copulas, which suggest that the joint model framework can describe copulas and the increase in outcome dependence is captured as an increase in the amount of shared outcome variability and a reduction in the total variability over both outcomes. The right panel of Figure 3.5 provides summary of the estimates of the μ's and β's. We see that joint models also provide robust location parameters and coefficient estimates. Note that the study only considers a fixed sample size and did not incorporate a covariate effect. The theoretical properties of the parameters estimate under model misspecification also deserve a further investigation.

3.6 Discussion

This chapter has developed a copula model framework and a joint model framework that can be utilized to model and predict the survivorship of an extended attack fire in terms of its containment time and area burned, given its environmental information as covariates. As a joint outcome analysis, duration and size are defined by a common origin and event, while two flexible frameworks (i.e. copula and joint modeling) are used to model their dependence. The factor loading form of the joint model reflects the scale difference between the shared error of the two outcomes.

We focused on understanding the relationship between and utility of the copula model and the factor loading form, as well as developing novel techniques to construct covariates and providing estimates of their effects. Our results suggest that duration and size are significantly dependent, and joint modeling outperforms modeling the outcomes separately. Our simulation studies show that as the outcome dependence in a copula increases, the shared variability in a joint model increases and the outcome-specific

variability decreases, while estimates of associated parameters become more precise. Some striking covariate effects were observed. Fire center and decade affect both duration and size. Increases in initial and day-over-day organic layer dryness have positive effects on duration, while increase in the ADFT of precipitation has negative effect on size. The findings provide a comprehensive perspective for understanding the statistical uncertainty quantified in modeling fire duration and fire size through copulas and joint models. The findings are also significant in a climate change context as BUI and DMC are expected to increase, and precipitation decrease, in parts of the fire season in central to southern BC in future decades. They also help to explain the large fire sizes in BC in the 2017 and 2018 fire seasons.

With regard to the moderate heteroscedasticity and skewness observed in the residuals, developing methods to handle clustering effects in the data may provide an effective mechanism to reduce these effects. Different containment strategies may result in more than one population of fires (i.e. mild and severe) and, hence, clustered outcome distributions. Under the framework of joint model, such clustering can be accounted for by introducing another latent variable as an unobserved label of the clusters. For instance, the nesting of joint modeling and mixture model utilizes one method as a foundation model and applies the other method in one of its sub-models (e.g. Dean et al., 2007; Huang et al., 2016). The bivariate normal mixture is also a comparable alternative often used in medical studies (e.g. Vink et al., 2016). Developing a finite mixture of the joint model for fire duration and fire size may be of both scientific and statistical interest to extend methods for this dependent modeling framework.

References

Bayham, J. (2013). *Characterizing incentives: an investigation of wildfire response and environmental entry policy.* PhD Thesis, Washington State University, Pullman,WA

Beall, H.W. (1949). An outline of forest fire protection standards. Forestry Chronicle 25:82-106.

Butry, D. T., Gumpertz, M., & Genton, M. G. (2008). The production of large and small wildfires. *The Economics of Forest Disturbances: Wildfires, Storms, and Invasive Species, 79*, 79.

Canada. Forestry Canada. Fire Danger Group. (1992). *Development and structure of the Canadian forest fire behavior prediction system*. Ottawa: Forestry Canada, Fire Danger Group.

Cumming, S. (2001). A parametric model of the fire-size distribution. *Canadian Journal of Forest Research, 31*(8), 1297-1303.

DaCamara, C. C., Calado, T. J., Ermida, S. L., Trigo, I. F., Amraoui, M., & Turkman, K. F. (2014). Calibration of the Fire Weather Index over Mediterranean Europe based on fire activity retrieved from MSG satellite imagery. *International Journal of Wildland Fire, 23*(7), 945-958.

De Groot, W. J. (1998). *Interpreting the Canadian Forest Fire Weather Index (FWI) System*. Paper presented at the Proc. of the Fourth Central Region Fire Weather Committee Scientific and Technical Seminar.

Dean, C., Nathoo, F., & Nielsen, J. (2007). Spatial and mixture models for recurrent event processes. *Environmetrics*. The official journal of the International Environmetrics Society, *18*(7), 713-725.

Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software, 71*(9), 1-25.

Duchateau, L., & Janssen, P. (2008). The Frailty Model New York. *Inc.: Springer-Verlag*.

Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Boca Raton, FL:CRC Press.

Feng, C., & Dean, C. (2012). Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models. *Environmetrics, 23*(6), 493-

508.

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

Flannigan, M.D, B.M. Wotton, B.J. Stocks, B. Todd, H. Cameron, K. Logan. 2002. Assessing Past, Current and Future Fire Occurrence and Fire Severity in BC. *Collaborative Research Agreement Report for the BC Forest Service Protection Program*. Canadian Forest Service.

Frees, E. W., & Valdez, E. A. (1998). Understanding relationships using copulas. *North American actuarial journal, 2*(1), 1-25.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2): CRC press Boca Raton, FL.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457-472.

Genest, C., & Favre, A.C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering, 12*(4), 347-368.

Hofert, M., Kojadinovic, I., Maechler, M., & Yan, J. (2014). copula: Multivariate dependence with copulas. *R package version 0.999-9, URL http://CRAN. R-project. org/package= copula*, C225.

He, W., & Lawless, J. F. (2005). Bivariate location–scale models for regression analysis, with applications to lifetime data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(1), 63-78.

Holmes, T. P., Huggett Jr, R. J., & Westerling, A. L. (2008). Statistical analysis of large wildfires. *The Economics of Forest Disturbances* (pp. 59-77): Springer.

Hougaard, P. (2000). Shared frailty models. *Analysis of Multivariate Survival Data*, 215-262.

Huang, Y., Yan, C., Yin, P., & Lu, M. (2016). A mixture of hierarchical joint models for longitudinal data with heterogeneity, non-normality, missingness, and covariate measurement error. *Journal of biopharmaceutical statistics, 26*(2), 299-322.

Joe, H. (2014). *Dependence modeling with copulas*: CRC Press.

Juarez-Colunga, E., Silva, G., & Dean, C. (2017). Joint modeling of zero-inflated panel count and severity outcomes. *Biometrics, 73*(4), 1413-1423.

Kelly, D. L. (2007). *Using copulas to model dependence in simulation risk assessment.* Paper presented at the ASME 2007 International Mechanical Engineering Congress and Exposition.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1/2), 81-93.

Komárek, A., & Lesaffre, E. (2008). Bayesian accelerated failure time model with multivariate doubly interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association, 103*(482), 523-533.

Lawless, J., Hu, J., & Cao, J. (1995). Methods for the estimation of failure distributions and rates from automobile warranty data. *Lifetime Data Analysis, 1*(3), 227-240.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data* (Vol. 362): John Wiley & Sons.

Lawson, B. D., & Armitage, O. (2008). Weather guide for the Canadian forest fire danger rating system.

Magnussen, S., & Taylor, S. W. (2012). Prediction of daily lightning-and human-caused fires in British Columbia. *International Journal of Wildland Fire*, 21(4), 342-356.

McArthur, A.G. (1968). The effect of time on fire behaviour and fire suppression problems. S.A.Emergency Fire Services, Keswick, South Australia. *Emergency Fire Services (EFS) manual*: 3-6, 8, 10-13.

Morin, A. A., Albert-Green, A., Woolford, D. G., & Martell, D. L. (2019). Frailty Models for the Control Time of Wildland Fires in the Former Intensive Fire Management Zone of Ontario, Canada. *Journal of Environmentl Statistics,* 9(5), 1-16.

Morin, A. A., Albert-Green, A., Woolford, D. G., & Martell, D. L. (2015). The use of survival analysis methods to model the control time of forest fires in Ontario, Canada. *International Journal of Wildland Fire, 24*(7), 964-973.

Natural Resources Canada (2017). *Canadian Forest Fire Weather Index (FWI) System*. Retrieved from http://cwfis.cfs.nrcan.gc.ca/images/fwi_structure.gif

Nelsen, R. (2006). An introduction to copulas, ser. *Lecture Notes in Statistics. New York: Springer*.

Pal, S., & Murthy, G. (2003). An application of Gumbel's bivariate exponential distribution in estimation of warranty cost of motor cycles. *International Journal of Quality & Reliability Management, 20*(4), 488-502.

Podur, J., & Wotton, B. M. (2011). Defining fire spread event days for fire-growth modelling. *International Journal of Wildland Fire, 20*(4), 497-507.

Reed, W. J., & McKelvey, K. S. (2002). Power-law behaviour and parametric models for the size-distribution of forest fires. *Ecological Modelling, 150*(3), 239-254.

Renouf, E., Dean, C. B., Bellhouse, D. R., & McAlister, V. C. (2016). Joint Survival Analysis of Time to Drug Change and a Terminal Event with Application to Drug Failure Analysis using Transplant Registry Data. *International Journal of Statistics in Medical Research, 5*(3), 198-213.

Schoenberg, F. P., Peng, R., & Woods, J. (2003). On the distribution of wildfire sizes. *Environmetrics, 14*(6), 583-592.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris, 8*, 229-231.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583-639.

Stocks, B. J., Lynham, T., Lawson, B., Alexander, M., Wagner, C. V., McAlpine, R., & Dube, D. (1989). Canadian forest fire danger rating system: an overview. *The Forestry Chronicle, 65*(4), 258-265.

Sun, C. (2013). Bivariate Extreme Value Modeling of Wildland Fire Area and Duration. *Forest Science, 59*(6), 649-660.

Taylor, S. W., Woolford, D. G., Dean, C., & Martell, D. L. (2013). Wildfire Prediction to Inform Management: Statistical Science Challenges. *Statistical Science*, 28(4), 586-615.

Therneau, T. M., & Lumley, T. (2014). Package 'survival'. *Survival analysis. Published on CRAN*. https://tbrieder.org/epidata/course_reading/e_therneau.pdf

Therneau, T. M. T. M., & Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. New York, NY: Springer.

Tremblay, P.-O., Duchesne, T., & Cumming, S. G. (2018). Survival analysis and classification methods for forest fire size. *PloS one, 13*(1), e0189860.

Van Wagner, C. E. (1969). A simple fire-growth model. *The Forestry Chronicle*. 145(2):103-104.

Van Wagner, C. E. (1987). *Development and Structure of the Canadian Forest Fire Weather Index System*. Can. For. Serv., Forestry Tech. Rep. 35, Petawawa National Forestry Institute. Chalk River, Ont.

Verbeke, G., & Molenberghs, G. (2017). Modeling Through Latent Variables. *Annual*

Review of Statistics and Its Application, 4, 267-282.

Vink, M. A., Berkhof, J., van de Kassteele, J., van Boven, M., & Bogaards, J. A. (2016). A bivariate mixture model for natural antibody levels to human papillomavirus types 16 and 18: baseline estimates for monitoring the herd effects of immunization. *PloS one, 11*(8), e0161109.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*(Dec), 3571-3594.

Wienke, A. (2010). *Frailty models in survival analysis*: CRC Press.

Wotton, B. M. (2009). Interpreting and using outputs from the Canadian Forest Fire Danger Rating System in research applications. *Environmental and Ecological Statistics, 16*(2), 107-131.

Wu, S. (2014). Construction of asymmetric copulas and its application in two-dimensional reliability modelling. *European Journal of Operational Research, 238*(2), 476-485.

Xi, D. D. Z., Taylor, S. W., Woolford, D. G., & Dean, C. B. (2019). Statistical Models of Key Components of Wildfire Risk. *Annual Review of Statistics and Its Application, 6*(1), 197-222. doi:10.1146/annurev-statistics-031017-100450

Yoder, J., & Gebert, K. (2012). An econometric model for ex ante prediction of wildfire suppression costs. *Journal of Forest Economics, 18*(1), 76-89.

Chapter 4

4 Joint Mixture Models for the Duration and Size of Wildfires

4.1 Introduction

Fire duration and fire size, representing how long a fire lasts and how much area is burned, respectively, have been studied as key outcomes of wildland fire risk in fire science (Fried and Gilless 1989; Taylor et al. 2013; Xi et al., 2019). Early studies have been motivated by an assessment of their relationship with environmental variables for ecological and managerial purposes (Cumming, 2001; Finney et al., 2009). Both outcomes are non-negative, right skewed, quantifying the survivorship of the fires from an origin to extinguishment.

Fires with long durations tend to be large in size, hence fire duration and fire size are often correlated outcomes (Yoder and Gebert 2012; Sun, 2013; Bayham, 2013; Xi et al., 2020). These authors note indications of *multimodality*, namely, that there are distinct peaks in the density functions of the outcomes. From the fire management perspective, this is not unexpected as some fires are contained quickly on initiation of fire suppression activities, while others escape or indeed are left to burn (Filmon, 2004; Xi et al., 2019). Such fire suppression strategies are widely adopted in fire management, hence aside from being correlated, both fire duration and fire size can also be regarded as being generated from multiple management strategies yielding distinct subpopulations of fires.

Usual parametric distributions that rely on location and scale parameters are often not suitable for modeling data with multimodality in their distribution. Some authors choose to avoid modeling irregular shapes by analyzing only the subset of fires that exceed a threshold of duration (DaCamara et al., 2014) or size (Holmes et al., 2008), while others handle such irregularities by utilizing non-parametric survival models (Morin et al., 2015;

Tremblay et al., 2018). These studies have considered only a single outcome (i.e. either duration or size). While correlation in the outcomes has been noted, there have been no models developed to address both multimodality and correlation in the outcomes simultaneously. It is of both scientific and statistical interest to develop a comprehensive model to account for the correlation of fire duration and fire size, while also modeling the potential multimodality observed in their marginal and joint distributions.

Two types of statistical methods more commonly used in biostatistics may initiate such development: joint modeling, and mixture models. For the application of joint modeling in environmental and other circumstances, see for example, Feng and Dean (2012), Renouf et al. (2016), Juarez-Colunga et al. (2017) and Lundy and Dean (2018). Joint modeling provides an approach where correlation of the outcomes may be addressed (Dunson, 2000; Henderson et al., 2000). By assuming that the distributions of the outcomes are independent, conditional on a shared latent variable, the joint distributions of the outcomes may be obtained by integrating their product over the support of the latent variable. The latent variable included in each of the outcomes induces correlation. As well, multimodality can be accommodated through finite mixture models, in which the outcome distribution arises from a mixture of components reflecting subpopulations, and a categorical latent variable identifies the subpopulation to which a fire is associated. (McLachlan and Peel, 2000).

A mechanism by which both joint models and mixture models may be employed, reflecting the scientific context of fire science, uses each of these as building blocks in constructing an overarching model. Vink et al. (2016) use a bivariate Gaussian mixture model for estimating vaccine-type seroprevalence from correlated antibody responses, hence incorporating mixtures in correlated outcomes. In a forestry study, Dean et al. (2007) developed a multi-state model for tree disease status using a two-component mixture. In the component of affected trees, the forward and backward transition probabilities of the disease status are linked with a tree-specific spatial random effect. In AIDS research, Huang et al. (2016) developed a three-component skewed-t mixture model for longitudinal viral load. The underlying trajectories of a covariate are linked with the viral load model through

a latent covariate process. To date, there has been little research on the development of models that correlate outcomes through shared latent variables to form multivariate joint mixture distributions.

Additionally, covariate effects could be incorporated in such mixtures by formulating a logistic model linking covariate effects to probabilities of the underlying component membership. Such an approach is computationally unattractive for a variety of key reasons. Importantly, model building becomes more computationally intensive in determining covariate selection (Asparouhov and Muthén, 2014; Murphy and Murphy, 2019). As well, such techniques sometimes require estimates of component membership to be approximated as the component which has the largest posterior. Instead, a two-stage approach is adopted here whereby the estimated probabilities of component membership from a mixture model are considered as a function of covariates in a Dirichlet regression. This chapter therefore aims to address several gaps in crucial research regarding joint outcome models in a mixture context. Importantly, this is a critical statistical advancement that seems particularly applicable in the fire science context we are considering.

In this chapter, we propose and develop a finite mixture framework for the joint modeling of fire duration and fire size. Duration and size are modelled simultaneously using univariate lognormal distributions, which are linked through shared errors to form a four-component bivariate mixture. The posterior estimates of the probabilities of component membership for each fire are modeled as a function of explanatory variables using Dirichlet regression. Our framework provides a novel perspective to study the underlying mechanism linking fire duration and fire size, while being flexible and having the advantage of a straightforward interpretation when the number of outcomes is large or the marginal distributions are complex.

We present the models for fire duration and fire size in section 2 and provide methods of estimation of the joint mixture model in section 3. In section 4, we describe the fire data from British Columbia, Canada, that motivated this research. Section 5 discusses the

analysis and the interpretation of the results from models fitted. The chapter closes with a discussion in section 6.

4.2 Modelling Frameworks

We describe two hierarchical frameworks for joint modeling of fire duration and fire size, a finite mixture joint model (FMJM) and a finite mixture bivariate model (FMBM). The distributions of the two models are provided in detail later. Individual fires are indexed by $i = 1, \dots, n$, with unobserved component labels $j = 1, \dots, J$, specifying the unique mixture component from which the joint distribution of duration and size arises. Outcomes are indexed by k, with $k = 1$ for duration and $k = 2$ for size. The bivariate random variable, $\boldsymbol{t}_i = (t_{i1}, t_{i2})^T$ is a 2×1 vector of the duration and size outcomes, where t_{i1} is the duration of the fire in days and t_{i2} is the size in hectares, with $\boldsymbol{t}_1, \dots, \boldsymbol{t}_n$ independent. We conduct a 2-stage analysis. In the first stage we estimate the parameters of the mixture models. In the second stage, the estimated probabilities that $\boldsymbol{t}_i$ belongs to each component are regressed against explanatory variables in a Dirichlet model to assess the effect of covariates.

4.2.1 Finite Mixture Joint Models

Let $z_i = 1, \dots, J$ be the unobserved component label of $\boldsymbol{t}_i$. The distribution of z_i is defined as i.i.d. Multinomial$(1, \boldsymbol{\pi})$, where $\boldsymbol{\pi} = \left(\pi_1, \dots, \pi_J\right)^T$ is a vector of mixture probabilities such that $\sum_{j=1}^{J} \pi_j = 1$, with π_j denoting the probability that $z_i = j$. We represent z_i by a $J \times 1$ latent vector $\boldsymbol{z}_i = \left(z_{i1}, \dots, z_{iJ}\right)^T$, where $z_{ij} = 1$ if $z_i = j$, and $z_{ij} = 0$ otherwise. Let $\boldsymbol{b}_{ij} = \left(b_{ij1}, b_{ij2}\right)^T$ be a 2×1 vector of random effects that accounts for potential correlation between t_{i1} and t_{i2} given membership in component j, with the correlation depending on the component to which they belong. The distribution of $\boldsymbol{b}_{ij}$ is defined as i.i.d. $Q_j\left(\boldsymbol{b}_{ij} | \boldsymbol{D}_j\right) = N_2\left(\boldsymbol{0}, \boldsymbol{D}_j\right)$, with a zero-mean 2×1 vector, $\boldsymbol{0}$, and a 2×2 symmetric and positive definite variance-covariance, $\boldsymbol{D}_j$. Given membership in component j, and given $\boldsymbol{b}_{ij}$, the outcomes t_{i1} and t_{i2} are independent. We define the $2 \times n$ matrix $\boldsymbol{t} = (\boldsymbol{t}_1, \dots, \boldsymbol{t}_n)$, the $J \times n$ matrix $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)$, $2 \times J$ matrices $\boldsymbol{\mu} = \left(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_J\right)$ where $\boldsymbol{\mu}_j =$

$(\mu_{j1}, \mu_{j2})^T$ and $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_J)$ where $\boldsymbol{\sigma}_j = (\sigma_{j1}, \sigma_{j2})^T$, as well as $\boldsymbol{b} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$, the collection of $2 \times J$ matrices such that $\boldsymbol{b}_i = (\boldsymbol{b}_{i1}, \ldots, \boldsymbol{b}_{iJ})$. The joint distribution of the data $\boldsymbol{t}$ and the latent variable $\boldsymbol{z}$, given all model parameters and random effects, is:

$$p(\boldsymbol{t}, \boldsymbol{z} | \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{b}, \boldsymbol{\pi}) = \prod_i \prod_j \left[\pi_j f_j(\boldsymbol{t}_i | \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \boldsymbol{b}_{ij})\right]^{z_{ij}},$$

where $f_j(\boldsymbol{t}_i | \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \boldsymbol{b}_{ij})$ is the conditional joint density function of $\boldsymbol{t}_i$ given $\boldsymbol{\mu}_j$, $\boldsymbol{\sigma}_j$, and the random effect $\boldsymbol{b}_{ij}$.

To model the correlation between the outcomes, we represent the relationship in a loglinear model (Duchateau and Janssen, 2008). Given that the outcomes belong to component j, we assume that $\boldsymbol{b}_{ij}$ has an additive effect on the logarithm of t_{ik}:

$$\log t_{ik} = \mu_{jk} + b_{ijk} + \sigma_{jk} \varepsilon_{ik},$$

where ε_{ik} follows an i.i.d. $N(0,1)$ and is the outcome-k-specific random error associated with fire i. The subscript j can be replaced by z_i for a more coherent notation. We assume that $\boldsymbol{b}_{ij}$ and ε_{ik} are independent for all i. When duration and size in component j are dependent, that is, when the covariance entries of $\boldsymbol{D}_j$ are not zero, we assume that $b_{ij1} = b_{ij}$, $b_{ij2} = \gamma_j b_{ij}$, where b_{ij} follows i.i.d. $q_j(b_{ij} | \sigma_{bj}) = N(0, \sigma_{bj}^2)$. In this case, b_{ij} is a shared frailty that produces the correlation between duration and size, while γ_j is the factor loading on b_{ij} that accounts for the scale difference between the outcomes. When duration and size are independent, b_{ij1} and b_{ij2} are freely varying with independent distributions. Then $f_j(\boldsymbol{t}_i | \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \boldsymbol{b}_{ij})$ becomes

$$f_j(\boldsymbol{t}_i | \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \boldsymbol{b}_{ij}) = \prod_k f_{jk}(t_{ik} | \mu_{jk}, \sigma_{jk}, b_{ijk}),$$

where $f_{jk}(t_{ik} | \mu_{jk}, \sigma_{jk}, b_{ijk})$ is the conditional density function of outcome k given membership in component j and associated random effect b_{ijk}.

4.2.2 Finite Mixture Bivariate Model

We also consider a finite mixture of bivariate distributions (FMBM) for modeling correlation in mixture models. Such a framework is used for comparison with the latent model framework developed in the previous section. We assume that $\boldsymbol{y}_i = \log(\boldsymbol{t}_i)$ follows i.i.d. $N_2(\mathbf{0}, \boldsymbol{\Sigma}_j)$ with a zero-mean 2×1 vector, $\mathbf{0}$, and a 2×2 symmetric and positive definite variance-covariance matrix, $\boldsymbol{\Sigma}_j$. In $\boldsymbol{\Sigma}_j$, the marginal variability of duration, marginal variability of size, and covariance of duration and size for component j are each specified directly by its elements: $\boldsymbol{\Sigma}_j^{11} = \sigma_{j1}^2$, $\boldsymbol{\Sigma}_j^{22} = \sigma_{j2}^2$, and $\boldsymbol{\Sigma}_j^{12} = \boldsymbol{\Sigma}_j^{21} = \rho_j \sigma_{j1} \sigma_{j2}$, where $\boldsymbol{\rho} = (\rho_1 \ldots \rho_J)$ is the vector of correlation parameters between y_{i1} and y_{i2} in component j. The terms $\boldsymbol{t}$, $\boldsymbol{z}$, $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, follow from their definition in section 2.1.

The joint distribution of $\boldsymbol{t}$ and $\boldsymbol{z}$ given all model parameters is:

$$p(\boldsymbol{t}, \boldsymbol{z} | \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\pi}) = \prod_i \prod_j \left[\pi_j f_j(\boldsymbol{t}_i | \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \rho_j) \right]^{z_{ij}},$$

where $f_j(\boldsymbol{t}_i | \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \rho_j)$ is the joint density function of $\boldsymbol{t}_i$ given $\boldsymbol{\mu}_j$, $\boldsymbol{\sigma}_j$ and ρ_j.

4.2.3 The Four-Component Mixture Models

We consider a special case in the fire science context for the two frameworks discussed above for modeling fire duration and size, with the parameterization of $f_j(\boldsymbol{y}_i | \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \boldsymbol{b}_{ij})$ and $f_j(\boldsymbol{y}_i | \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \rho_j)$ provided in Table 4.1 under the columns FMJM and FMBM, where $\boldsymbol{y}_i = \log(\boldsymbol{t}_i)$.

Fires tend to occur in two main clusters: of typical size and duration, given the time of the fire season in which they occur; or, of extreme fire size and duration, contrasted with typical fires at that time of the year. This results in four groups of fires according to the

Table 4.1: Parameterization of the models considered in the fire science application

Component Label	$f_j(\boldsymbol{y}_{ij}\|\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \boldsymbol{b}_{ij})$	
	FMJM	FMBM
$z_i = 1$	$\boldsymbol{y}_{i1} \sim N_2\left(\begin{bmatrix}\mu_{N1}\\ \mu_{N2}\end{bmatrix}, \begin{bmatrix}\sigma_{bN}^2 + \sigma_{N1}^2 & \gamma_N \sigma_{bN}^2\\ \gamma_N \sigma_{bN}^2 & \gamma_N^2 \sigma_{bN}^2 + \sigma_2^2\end{bmatrix}\right)$	$\boldsymbol{y}_{i1} \sim N_2\left(\begin{bmatrix}\mu_{N1}\\ \mu_{N2}\end{bmatrix}, \begin{bmatrix}\sigma_{N1}^2 & \rho_1 \sigma_{N1}\sigma_{N2}\\ \rho_1 \sigma_{N1}\sigma_{N2} & \sigma_{N2}^2\end{bmatrix}\right)$
$z_i = 2$	$\boldsymbol{y}_{i2} \sim N_2\left(\begin{bmatrix}\mu_{N1}\\ \mu_{E2}\end{bmatrix}, \begin{bmatrix}\sigma_{bN}^2 + \sigma_{N1}^2 & 0\\ 0 & \gamma_E^2 \sigma_{bE}^2 + \sigma_2^2\end{bmatrix}\right)$	$\boldsymbol{y}_{i2} \sim N_2\left(\begin{bmatrix}\mu_{N1}\\ \mu_{E2}\end{bmatrix}, \begin{bmatrix}\sigma_{N1}^2 & \rho_2 \sigma_{N1}\sigma_{E2}\\ \rho_2 \sigma_{N1}\sigma_{E2} & \sigma_{E2}^2\end{bmatrix}\right)$
$z_i = 3$	$\boldsymbol{y}_{i3} \sim N_2\left(\begin{bmatrix}\mu_{E1}\\ \mu_{N2}\end{bmatrix}, \begin{bmatrix}\sigma_{bE}^2 + \sigma_{E1}^2 & 0\\ 0 & \gamma_N^2 \sigma_{bN}^2 + \sigma_2^2\end{bmatrix}\right)$	$\boldsymbol{y}_{i3} \sim N_2\left(\begin{bmatrix}\mu_{E1}\\ \mu_{N2}\end{bmatrix}, \begin{bmatrix}\sigma_{E1}^2 & \rho_3 \sigma_{E1}\sigma_{N2}\\ \rho_3 \sigma_{E1}\sigma_{N2} & \sigma_{N2}^2\end{bmatrix}\right)$
$z_i = 4$	$\boldsymbol{y}_{i4} \sim N_2\left(\begin{bmatrix}\mu_{E1}\\ \mu_{E2}\end{bmatrix}, \begin{bmatrix}\sigma_{bE}^2 + \sigma_{E1}^2 & \gamma_E \sigma_{bE}^2\\ \gamma_E \sigma_{bE}^2 & \gamma_E^2 \sigma_{bE}^2 + \sigma_2^2\end{bmatrix}\right)$	$\boldsymbol{y}_{i4} \sim N_2\left(\begin{bmatrix}\mu_{E1}\\ \mu_{E2}\end{bmatrix}, \begin{bmatrix}\sigma_{E1}^2 & \rho_4 \sigma_{E1}\sigma_{E2}\\ \rho_4 \sigma_{E1}\sigma_{E2} & \sigma_{E2}^2\end{bmatrix}\right)$

magnitude of their duration and size–normal (N) or extreme (E), suggesting a four-component bivariate mixture joint model to reflect components:

$$\begin{cases} \text{normal duration} - \text{ normal size } (j=1) \\ \text{normal duration} - \text{extreme size } (j=2) \\ \text{extreme duration} - \text{normal size } (j=3) \\ \text{extreme duration} - \text{extreme size } (j=4) \end{cases}.$$

We put constraints on certain univariate terms, namely, μ_{jk}, b_{ijk}, γ_j, σ_{jk} and σ_{bjk}, if the associated term is describing the distribution of the corresponding outcome in a normal cluster or an extreme cluster. For both FMJM and FMBM, we assume that the centres of the related components are the same for model parsimony and identifiability: $\boldsymbol{\mu}_1 = (\mu_{N1}, \mu_{N2},)$, $\boldsymbol{\mu}_2 = (\mu_{N1}, \mu_{E2},)$, $\boldsymbol{\mu}_3 = (\mu_{E1}, \mu_{N2},)$, $\boldsymbol{\mu}_4 = (\mu_{E1}, \mu_{E2},)$, where $\mu_{E1} = \mu_{N1} + \Delta_{\mu 1}$, $\mu_{E2} = \mu_{N2} + \Delta_{\mu 2}$.

For FMJM, we further assume that only the outcomes in component 1 and 4 are linked through a latent variable. Since the factor loading parameter defines the scale difference of the random effect on the outcome of fire size, this parameter would be the same for components representing normal size (components 1 and 3), and also the same for components related to extreme size (components 2 and 4). Hence $\gamma_1 = \gamma_N$, $\gamma_2 = \gamma_E$, $\gamma_3 = \gamma_N$, $\gamma_4 = \gamma_E$, where we parameterize $\gamma_E = \gamma_N + \Delta_\gamma$, and $\boldsymbol{b}_{i1} = (b_{iN}, \gamma_N b_{iN})^T$, $\boldsymbol{b}_{i2} = (b_{iN}, \gamma_E b_{iE})^T$, $\boldsymbol{b}_{i3} = (b_{iE}, \gamma_N b_{iN})^T$, $\boldsymbol{b}_{i4} = (b_{iE}, \gamma_E b_{iE})^T$. We allow duration-specific variabilities to be distinct, but set size-specific variabilities equal across all components to avoid over-parameterization: $\boldsymbol{\sigma}_1 = (\sigma_{N1}, \sigma_2)$, $\boldsymbol{\sigma}_2 = (\sigma_{N1}, \sigma_2)$, $\boldsymbol{\sigma}_3 = (\sigma_{E1}, \sigma_2)$, $\boldsymbol{\sigma}_4 = (\sigma_{E1}, \sigma_2)$, where $\sigma_{E1} = \sigma_{N1} + \Delta_{\sigma 1}$. In other words,

$$y_{ik} \sim N\left(\boldsymbol{\mu}_{z_i}^{(k)} + \boldsymbol{b}_{z_i}^{(k)}, \boldsymbol{\sigma}_{z_i}^{(k)}\right),$$

where $\boldsymbol{\mu}_{z_i}^{(k)}$, $\boldsymbol{b}_{z_i}^{(k)}$, and $\boldsymbol{\sigma}_{z_i}^{(k)}$ are the k-th elements of $\boldsymbol{\mu}_{z_i}$, $\boldsymbol{b}_{z_i}$, and $\boldsymbol{\sigma}_{z_i}$, respectively.

For FMBM, we further assume that the marginal variabilities in components that reflect normal or extreme duration and normal or extreme size are the same, but the correlation parameters among the components are distinct:

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} \sigma_{N1}^2 & \rho_1\sigma_{N1}\sigma_{N2} \\ \rho_1\sigma_{N1}\sigma_{N2} & \sigma_{N2}^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} \sigma_{N1}^2 & \rho_2\sigma_{N1}\sigma_{E2} \\ \rho_2\sigma_{N1}\sigma_{E2} & \sigma_{E2}^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_3 = \begin{bmatrix} \sigma_{E1}^2 & \rho_3\sigma_{E1}\sigma_{N2} \\ \rho_3\sigma_{E1}\sigma_{N2} & \sigma_{N2}^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_4 = \begin{bmatrix} \sigma_{E1}^2 & \rho_4\sigma_{E1}\sigma_{E2} \\ \rho_4\sigma_{E1}\sigma_{E2} & \sigma_{E2}^2 \end{bmatrix}.$$

Hence, the joint posterior distributions of a FMJM and a FMBM become:

$$p(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{z}, \boldsymbol{b}, \boldsymbol{\pi}, \boldsymbol{D}|\boldsymbol{t}) \propto p(\boldsymbol{t}, \boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{b}, \boldsymbol{\pi})p(\boldsymbol{b}|\boldsymbol{D})p(\boldsymbol{\mu})p(\boldsymbol{\sigma})p(\boldsymbol{D}),$$

and

$$p(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{z}, \boldsymbol{\rho}, \boldsymbol{\pi}|\boldsymbol{t}) \propto p(\boldsymbol{t}, \boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\sigma})p(\boldsymbol{\rho}),$$

respectively, where the joint prior distributions, required for estimation of the model parameters, are

$$\begin{aligned} p(\boldsymbol{b}|\boldsymbol{D})p(\boldsymbol{\mu})p(\boldsymbol{\sigma})p(\boldsymbol{D}) &= \prod_i \prod_j \left[Q(\boldsymbol{b}_{ij}|\boldsymbol{D}_j)p(\boldsymbol{\mu}_j)p(\boldsymbol{\sigma}_j)p(\boldsymbol{D}_j)\right]^{z_{ij}} \\ &= \prod_i \prod_j \left[p(\gamma_j)\prod_k q(b_{ijk}|\sigma_{bjk})p(\sigma_{bjk})p(\mu_{jk})p(\sigma_{jk})\right]^{z_{ij}}, \end{aligned}$$

and

$$p(\boldsymbol{\mu})p(\boldsymbol{\sigma})p(\boldsymbol{\rho}) = \prod_i \prod_j \left[p(\boldsymbol{\mu}_j)p(\boldsymbol{\sigma}_j)p(\rho_j)\right]^{z_{ij}} = \prod_i \prod_j \left[p(\rho_j)\prod_k p(\mu_{jk})p(\sigma_{jk})\right]^{z_{ij}}.$$

Model fitting is carried out by the adaptive MCMC method. We assume vague priors commonly used in the literature (e.g. Feng and Dean, 2012; Vink et al., 2016): for $j = N, E$ and $k = 1, 2$, $p(u_{jk})$ is distributed as $N(0, 10000)$; $p(\sigma_{jk})$, $p(\sigma_{bj})$, $p(\gamma_j)$, $p(\Delta_{uk})$, $p(\Delta_\gamma)$ and $p(\Delta_{\sigma 1})$ are distributed as half. $N(0, 10000)$; $p(\rho_j)$ is distributed as $U(-1, 1)$, where $U(a, b)$ is the uniform distribution over (a, b); $p(\boldsymbol{\pi})$ is distributed as Dirichlet$(\mathbf{1})$ where Dirichlet$(\boldsymbol{\alpha})$ has density

$$p(\boldsymbol{\pi}) = \Gamma\left(\sum_{j=1}^{J} \alpha_j\right) \prod_{j}^{J} \frac{\pi_j^{\alpha_j - 1}}{\Gamma(\alpha_j)},$$

with the shape parameter vector, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$ and $\mathbf{1}$ is a vector of 1's with dimension J. The posterior estimates of parameters and of latent variables are obtained as their posterior medians. The full posterior distributions of the final models are provided in Appendix 4C.

4.2.4 Dirichlet Model for the Effect of Covariates on Component Membership

Let $p_{i1}, \ldots, p_{iJ}$ be the estimated probabilities that $z_i = j$, $j = 1, \ldots, J$, given $\boldsymbol{t}_i$, the estimated probabilities of component membership. We model these as a function of the covariates, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iR})^T$, in a Dirichlet regression (Douma and Weedon, 2019). The term p_{ij} is obtained through Bayes' Rule as

$$p_{ij} = P(z_i = j|\boldsymbol{t}_i) = \frac{P(z_i = j)P(\boldsymbol{t}_i|z_i = j)}{\sum_j P(z_i = j)P(\boldsymbol{t}_i|z_i = j)} = \frac{\pi_j p(\boldsymbol{t}_i|z_i = j)}{\sum_j \pi_j p(\boldsymbol{t}_i|z_i = j)},$$

where $p(\boldsymbol{t}_i|z_i = j)$ is the posterior density function of $\boldsymbol{t}_i$ given z_i, which is obtained using the estimated model parameters in the first stage analysis. The membership probabilities

are rescaled in the manner $p_{ij}^* = \left[p_{ij}(N-1) + 0.5\right]/N$ to avoid values very close to zero or one (Smithson and Verkuilen, 2006).

In the second stage of the analysis, we model the $J \times 1$ vector $\boldsymbol{p}_i^* = \left(p_{i1}^*, \dots, p_{iJ}^*\right)^T$ is distributed as Dirichlet$(\boldsymbol{\alpha}_i)$, where each element of the shape parameter vector, $\boldsymbol{\alpha}_i = \left(\alpha_{i1}, \dots, \alpha_{ij}\right)$, is parametrized as

$$\text{logit}\left(\alpha_{ij}\right) = \alpha_0 + \beta_{0j} + \sum_{r=1}^{R} \beta_{1jr} x_{ir},$$

and α_0 is the global mean; β_{0j} is the component-j specific mean, and the $1 \times R$ vector $\boldsymbol{\beta}_{1j} = \left(\beta_{1j1}, \dots, \beta_{1jR}\right)$ is the component-j-specific vector of covariate coefficients. For identifiability, we set β_{04} and all the elements in $\boldsymbol{\beta}_{14}$ as zero. The sum $\alpha_0 + \beta_{0j}$ is interpreted as the log-odds of a fire belonging in the jth component, relative to the fourth component, with all covariates constant. As discussed by Maier (2014), in the development of Dirichlet regression models, the variable $\exp\left(\beta_{1jr}\right)$ is interpreted as the odds ratio corresponding to the increase of x_{ir} by one unit, given that the observation is in the jth component. Then

$$\hat{p}_{ij}^* = \frac{\hat{\alpha}_{ij}}{\sum_j \hat{\alpha}_{ij}},$$

is the estimate of the transformed probability that fire i belongs to component j, conditional on its covariates. We assume vague priors $p(\alpha_0)$, $p\left(\beta_{0j}\right)$, and $p\left(\beta_{1jr}\right), j = 1, \dots, 3, r = 1, \dots, R$ distributed as $N(0, 10000)$.

4.3 British Columbia Fire Study

4.3.1 Data Description

Our study is motivated by an interest in understanding the correlation between fire duration and fire size, as well as the effect of environmental variables. We consider an approach that

is based on the mixture model context discussed earlier. Duration and size are defined as the days and the hectares burned from two critical points in the life history of a fire: (1) start of ground attack, to (2) time of final control. Here we focus on only lightning-caused, extended attack fires (i.e. fires for which duration exceeds 2 days and size exceeds 4 hectares).

The data, assembled by fire scientists at the Pacific Forestry Centre, Natural Resources Canada, include historical fire records and the associated environmental records obtained from the British Columbia Wildfire Service and weather stations. The data comprise information about 1285 fires. There are six regional location variables identifying the fire centres in which the fire occurred: fire centres are geographic areas varying in size from about 73,000 to 319,000 km^2, with varying forest and topographic conditions and fire weather conditions that influence fire growth and difficulty of control, as well as values at risk that may influence fire management strategies and allocation of suppression resources. The fire management offices in each fire centre is responsible for wildland fire management within its regional boundaries. Additional variables are: temporal variables—decade and month in which the fire occurred; slope; elevation; size of the fire at the time of attack; and ten environmental variables recorded at weather stations for which daily records are available. The environmental variables include four weather observations and six standard fire indices of the Canadian Forest Fire Weather Index System (Van Wagner, 1987), derived from the weather observations. The four weather observations, temperature, wind, relative humidity, and precipitation are interpolated to a 20 km by 20 km grid using smoothing splines after adjusting for elevation and snowmelt/snow onset effects (Nadeem et al., 2020). The interpolated values are then used to calculate the six standard indices. The indices include three fuel moisture codes, Fine Fuel Moisture Code, Duff Moisture Code, Drought Code, that describe the dryness of the corresponding layer of the forest floor, and three fire behavior indices, Initial Spread Index, Buildup Index, and Fire Weather Index, that describe the fire spread rate, the available fuel, and the intensity of the fire-line respectively.

We adopt simple, meaningful ways to summarize environmental variables through their lifetime. As in Xi et al. (2020), variables demonstrating a clear trend through their trajectories are centered and regressed against time. The estimated intercept and the slope are used to summarize the trajectory. The remaining environmental variables are summarized into an index, referred to as the *average deviation from threshold (ADFT)*, describing the amount of exceedance, averaged across the lifetime of the fire, from a threshold value determined by scientific input. The terminology referring to the covariates identifies the names of the variables, either intercept or slope, or the values of the threshold (see Table 4.3).

4.3.2 Parameter Estimates

The four panel plots in Figure 4.1 present the data and the estimated distributions of fire duration and fire size. The top row contains the estimated marginal distributions of the outcomes, overlaid on their histograms, with duration on the left panel and size on the right. The estimated FMJM and FMBM distributions are provided in red dashed lines and green dotted lines respectively. The marginal distributions of duration and size are both captured by a narrowly spread normal component and a widely spread extreme component, and seem to provide reasonable fits. The bottom row identifies the component with the highest posterior probability of membership for each of the fires. The plots on the bottom row contain estimated contours based on the estimated normal joint distributions of the outcomes for each of the model components, with the panel on the left based on the fitted FMJM while that on the right is based on the fitted FMBM. Estimated components are identified with different colours and symbols.

Table 4.2 presents parameter estimates of the two models. The posterior median and the 95% credible interval of the parameters are reported. Under FMJM, the probability that a fire belongs to components 1 to 4, $\pi_1, \ldots, \pi_4$, are estimated as 0.339(0.278, 0.407), 0.052(0.016, 0.095), 0.109(0.054,0.170) and 0.497(0.420, 0.571). For convenience for the following discussion, recall that the specification of the means and the variabilities of the outcomes in each component are given in Table 4.1. For the means, μ_{N1} and μ_{E1} are

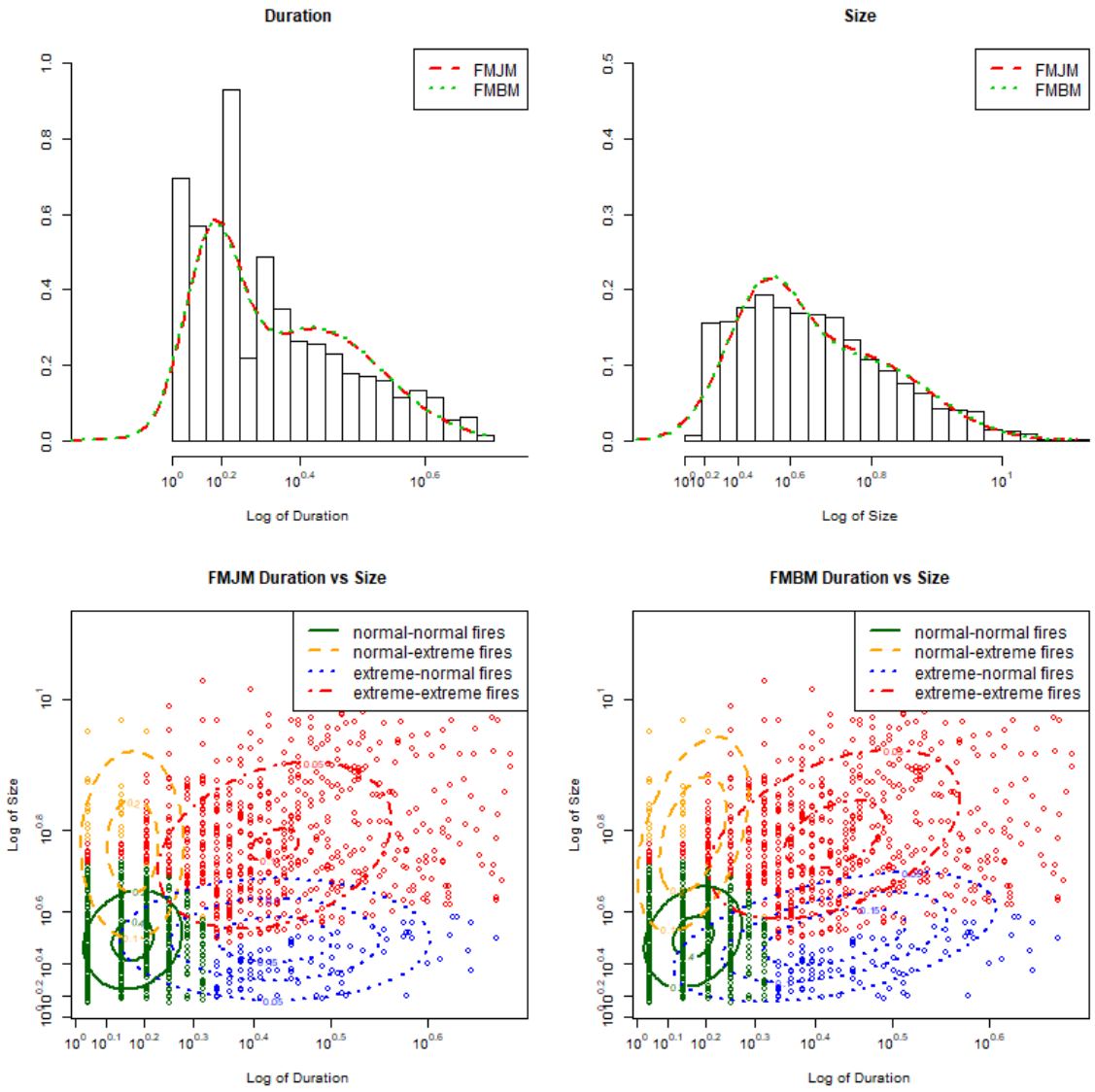


Figure 4.1: The data and the estimated distributions of fire duration and fire size. The top row contains the estimated marginal distributions of the outcomes, overlaid on their histograms, with duration on the left panel and size on the right. The marginal distributions of duration and size are both captured by a narrowly spread normal component and a widely spread extreme component and seem to provide reasonable fits. Fires that are normal or extreme in both outcomes tend to have outcomes correlated.

Table 4.2: Posterior estimates of model parameters

FMJM					FMBM			
	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$			$Q_{.025}$	$Q_{.500}$	$Q_{.975}$
π_1	0.278	0.339	0.407		π_1	0.267	0.323	0.388
π_2	0.016	0.052	0.095		π_2	0.020	0.056	0.098
π_3	0.054	0.109	0.170		π_3	0.124	0.192	0.270
π_4	0.420	0.497	0.571		π_4	0.341	0.426	0.505
μ_{N1}	1.425	1.483	1.551		μ_{N1}	1.419	1.473	1.540
μ_{E1}	2.592	2.690	2.804		μ_{E1}	2.577	2.674	2.782
μ_{N2}	3.006	3.205	3.414		μ_{N2}	3.070	3.272	3.497
μ_{E2}	5.562	5.872	6.222		μ_{E2}	5.813	6.184	6.622
γ_N	0.516	2.951	6.410		σ_{N1}	0.278	0.318	0.367
γ_E	4.247	5.933	8.681		σ_{E1}	0.782	0.829	0.879
σ_{bN}	0.077	0.136	0.247		σ_{N2}	0.978	1.110	1.252
σ_{bE}	0.200	0.285	0.385		σ_{E2}	1.696	1.866	2.032
σ_{N1}	0.196	0.291	0.354		ρ_1	0.084	0.241	0.397
σ_{E1}	0.718	0.772	0.827		ρ_2	-0.073	0.493	0.775
σ_2	0.782	1.001	1.158		ρ_3	0.312	0.463	0.581
					ρ_4	0.306	0.433	0.557

.

estimated as 1.483 and 2.690, while μ_{N2} and μ_{E2} are estimated as 3.205 and 5.872. For the factor loading parameters, γ_N^2 and γ_E^2 are estimated as 2.951 (0.516, 6.410) and 5.933(4.247, 8.681). For the variabilities, σ_{bN}^2 and σ_{bE}^2 are estimated as 0.136(0.077, 0.247) and 0.285(0.200, 0.385), while σ_{N1}^2, σ_{E1}^2, and σ_2^2 are estimated as 0.291(0.196, 0.354), 0.772(0.718, 0.827) and 0.772(0.782, 1.158). Corresponding values from fitting the FMBM are very close and omitted here (see Table 4.2). Note that Appendix 4D provides a sensitivity analysis to the choice of other priors, indicating robustness to the choice of priors.

A focus here is to understand the correlation between duration and size. Under FMJM, the estimates of the standard error of the shared error distribution, σ_{bN} and σ_{bE}, are 0.136(0.077, 0.247) and 0.285(0.200, 0.385), while the factor loading parameters, γ_N and γ_E, are estimated as 2.951(0.516, 6.410) and 5.933(4.247, 8.681). The effect of the shared error on the logarithm of size is about three times as large as its effect on the logarithm of duration in component 1 and is about six times in component 4. Furthermore, the size-specific error, σ_2, estimated as $0.125\ (0.017, 0.442)$, is quite small compared to the shared outcome error, suggesting that much of the variabilities in component 1 and 4 is shared. Under FMBM, the correlation of component 1, 3, and 4, ρ_1, ρ_2, and ρ_3 are significant and estimated respectively as 0.241(0.084, 0.397), 0.463(0.312, 0.581) and 0.433(0.306, 0.557).

As both models constrain the means of the components similarly, the estimated means are similar across the models. The estimates of the component labels, z_{ij}, from both models are very close, with high positive correlation (see Appendix 4E).

The covariance entries, $Cov(y_{ij1}, y_{ij2})$, are 0.054, 0, 0 and 0.482 for component $j = 1,..,4$ for FMJM, while the corresponding entries are 0.085, 0.295, 0.426 and 0.670 for FMBM (See Table 4.1). Essentially, when the shared variability is normally distributed, both models are Gaussian mixtures, while FMJM forces the outcome covariance in two of the components to be zero but FMBM does not, which is shown by the difference of the

directions of the estimated contours of component 2 and 3 in Figure 4.1. On the other hand, joint models allow that the shared variability may have different distributions than the normal, which offers one component of flexibility that is not reflected in FMBM.

4.3.3 Effect of Covariates in the Dirichlet Model

In this section we discuss covariates which are seen to have dominant effects on the response. Results are presented here for the FMJM given the similarity in results for the two mixture models and the benefits offered based on this model. Appendix 4F provides supplemental material related to other covariates considered. Figures 4.2 to 4.5 and Figures B.6 to B.8 in the supplemental materials display the estimated transformed membership probabilities, $\hat{p}_{ij}^{*}$, plotted against each of the covariates in the model. These scatterplots include a smoothing loess for numerical covariates, providing the overall trend of the plots using weighted linear least squares regressions over the span of the value of the covariates; for categorical covariates, the plots are side-by-side violin plots. The exponentiated estimated covariate effects, relative to component 4, $\exp(\hat{\beta}_{1jr}), r = 1, \dots, R$, for components $j = 1,2,3$ are summarized in Table 4.3.

Fire Centre: The estimated transformed membership probability by fire centre are presented in violin plots in the left panel of Figure 4.2. The plots show the posterior estimate of the probability of each fire belonging to component 1 to 4 for each of the fire centres. Fire management strategies vary by fire centre, and differences in such strategies may be exacerbated for fires with extreme duration which tend to receive more containment resources. Hence, we expect to see some variation by fire centre. As evidenced in Figure 4.2, the medians of the probabilities displayed in the violin plots demonstrate clear variation over fire centres for the extreme duration components. Within component 3 (displayed in blue), which identifies fires with extreme duration and normal size, the Coastal Region (Co) and the Cariboo Region (Ca) have the highest and the lowest probabilities respectively. The Southeast Region (So) also has a high corresponding probability. Within component 4 (displayed in red), which identifies fires with extreme duration and extreme size, the Coastal Region and the Cariboo Region have the lowest and the highest

probabilities respectively. This suggests that, for fires with extreme duration, fires in the Coastal Region of the province tend to have high probability of being small in size and the fires in the interior Cariboo Region tend to have high probabilities of being large in size.

Month: Figure 4.3 provides estimates of the probability of each fire belonging to component 1 to 4 by month and by year. The seasonality of the fire behavior displays different pattems depending on component. In component 1 (green), identifying fires with normal duration and normal size, the probabilities tend to a minimum in the middle of the fire season, whereas in component 4 (red), identifying fires with extreme duration and extreme size, probabilities tend to a maximum in the middle of the fire season. The months of August through October are associated with a much higher risk of fires being extreme in duration and size (in component 4, displayed in red). These components include 85% of the fires in the study. Fires of extreme size and normal duration (in component 2, displayed in yellow) tend to occur in May. Fires of extreme duration and normal size (in component 3, displayed in blue) are more likely to present at the end of the season than at the beginning.

Wind and Precipitation: Figure 4.4 presents the posterior probability estimates by the average wind speed (km/h) measured over a 10-minute period on the left panel and the amount of rain (mm) accumulated in the 24-hour period from noon to noon on the right. As wind speed increases, the probability of being identified in the component corresponding to normal duration and extreme size increases, while as precipitation increases, the probability of being identified in the component corresponding to extreme duration and extreme size decreases.

Drought Code and Duff Moisture Code: Figure 4.5 demonstrates posterior probability estimates by the ADFT of DC on the left panel and by DMC the right. DMC and DC are correlated with the moisture content of forest floor organic layers approximately 5-10 cm, and 10-20 cm thick, respectively, and indicate the average amount of available fuel in mid to deeper organic layers throughout the lifetime of the fire. The DMC is modeled from cumulative observations of relative humidity, temperature, and precipitation, and the DC from temperature and precipitation observations over the fire season. In our analysis, DMC

influenced fire size, and DC influenced fire duration. As the ADFT of DMC increases, the probability of being identified as normal size components (component 1 and 3) decreases while the probability of being identified as extreme size components (component 2 and 4) increases. As the ADFT of DC increases, the probability of fires with short duration (components 1 and 2) decreases and the probability of fires having long duration tends to increase (components 3 and 4).

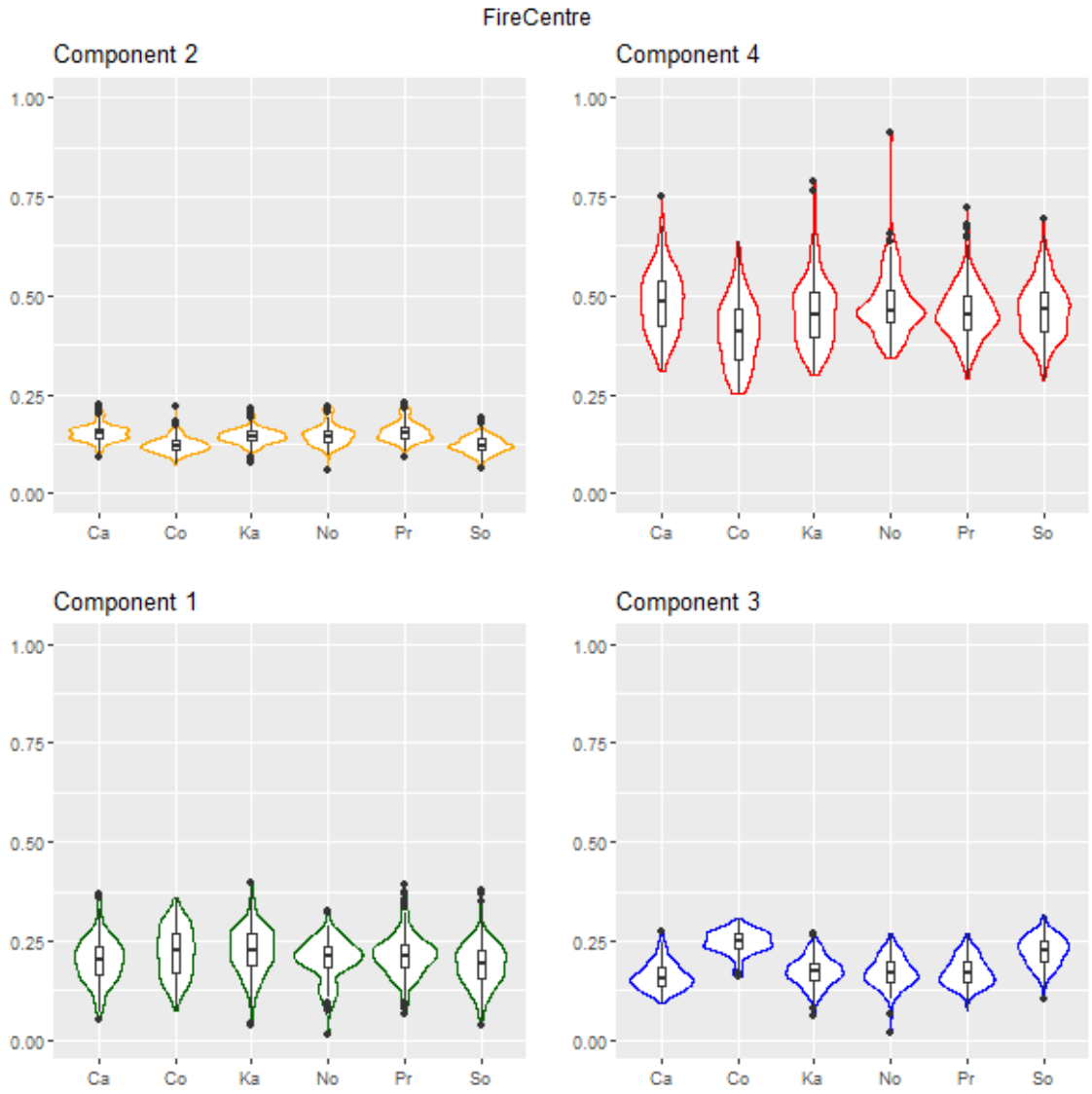


Figure 4.2: The probability estimates by fire centre are presented in violin plots. The plots show the posterior estimate of the probability of each fire belonging to component 1 to 4 for each of the fire centres. The medians of the probabilities displayed in the violin plots demonstrate clear variation over fire centres for the extreme duration components (component 3, displayed in blue and component 4, displayed in red).

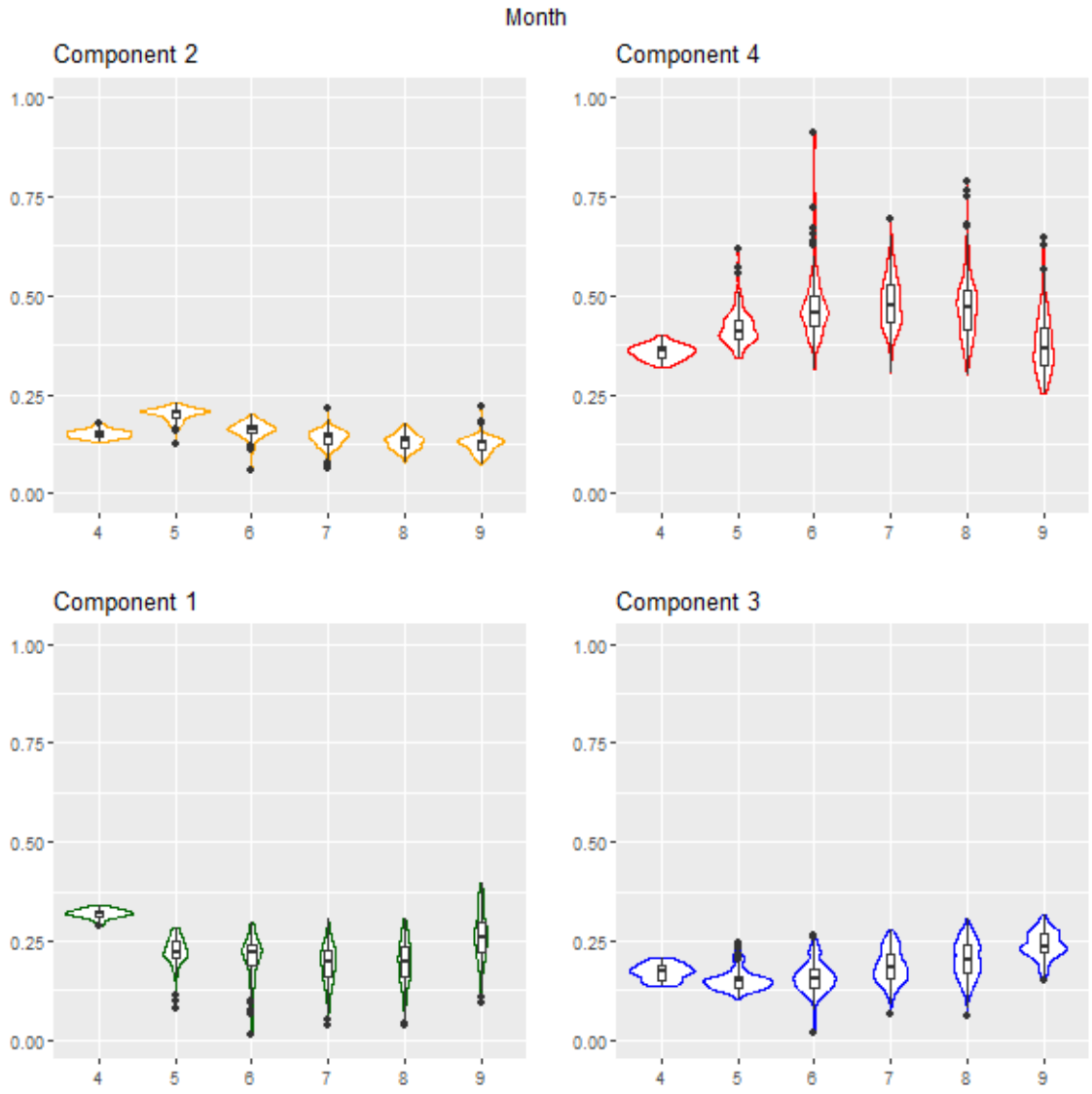


Figure 4.3: Posterior estimates of the probability of each fire belonging to component 1 to 4 by month. October data are combined into September because of its small number of observations. The seasonality of the fire behavior displays different patterns depending on component. The months of August through October are associated with a much higher risk of fires being extreme in duration and size. On the other hand, fires of extreme size and normal duration tend to occur in May. Fires of extreme duration and normal size are more likely to present at the end of the season than at the beginning.

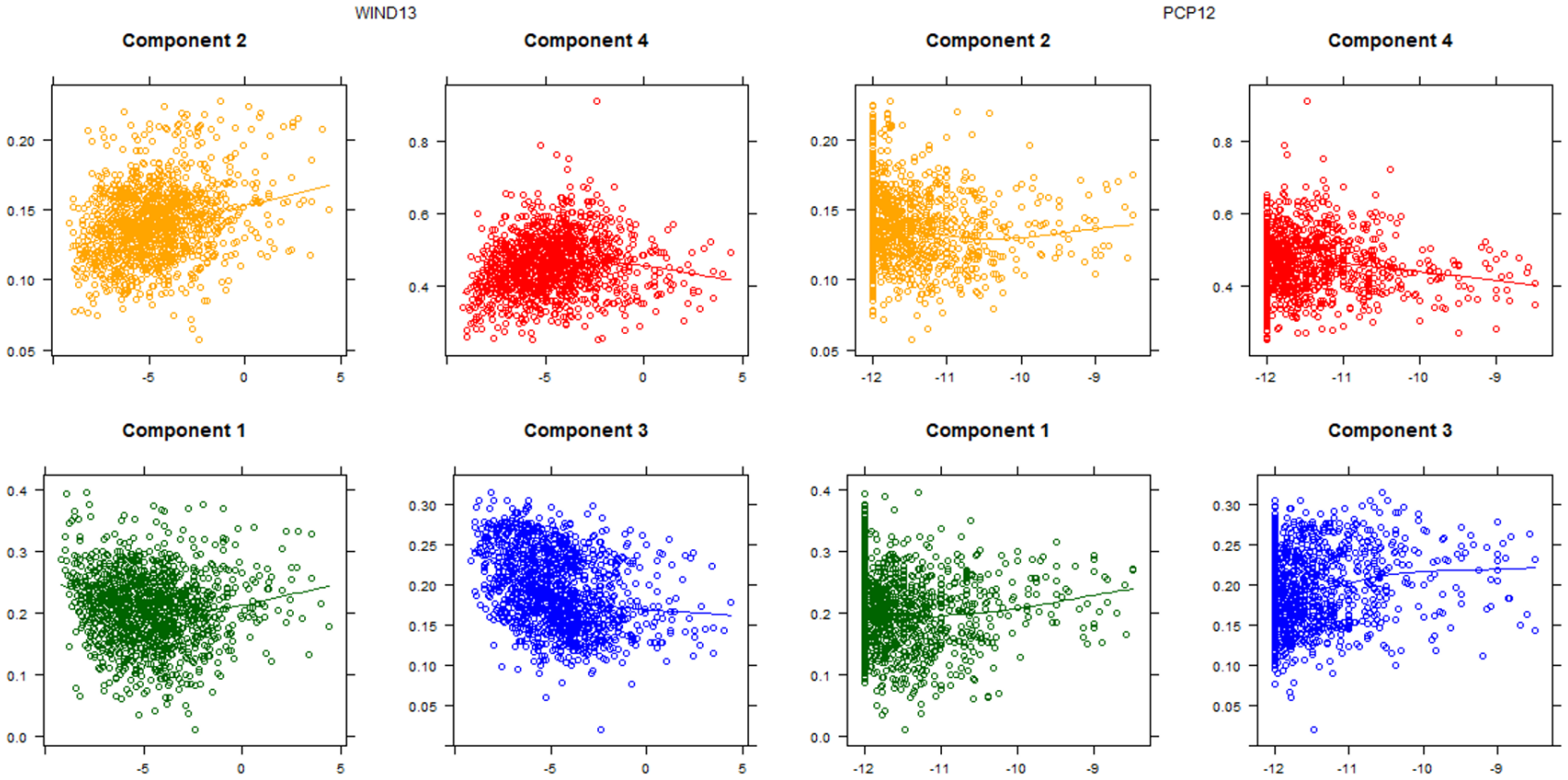


Figure 4.4: The posterior probability estimates (y-axis) by the ADFT of average wind speed (km/h, x-axis) measured over a 10-minute period on the two left panels and the ADFT of the amount of rain (mm, x-axis) accumulated in the 24-hour period from noon to noon on the two right panels. As wind speed increases, the probability of being identified in the component corresponding to normal duration and extreme size increases, while as precipitation increases, the probability of being identified in the component corresponding to extreme duration and extreme size decreases.

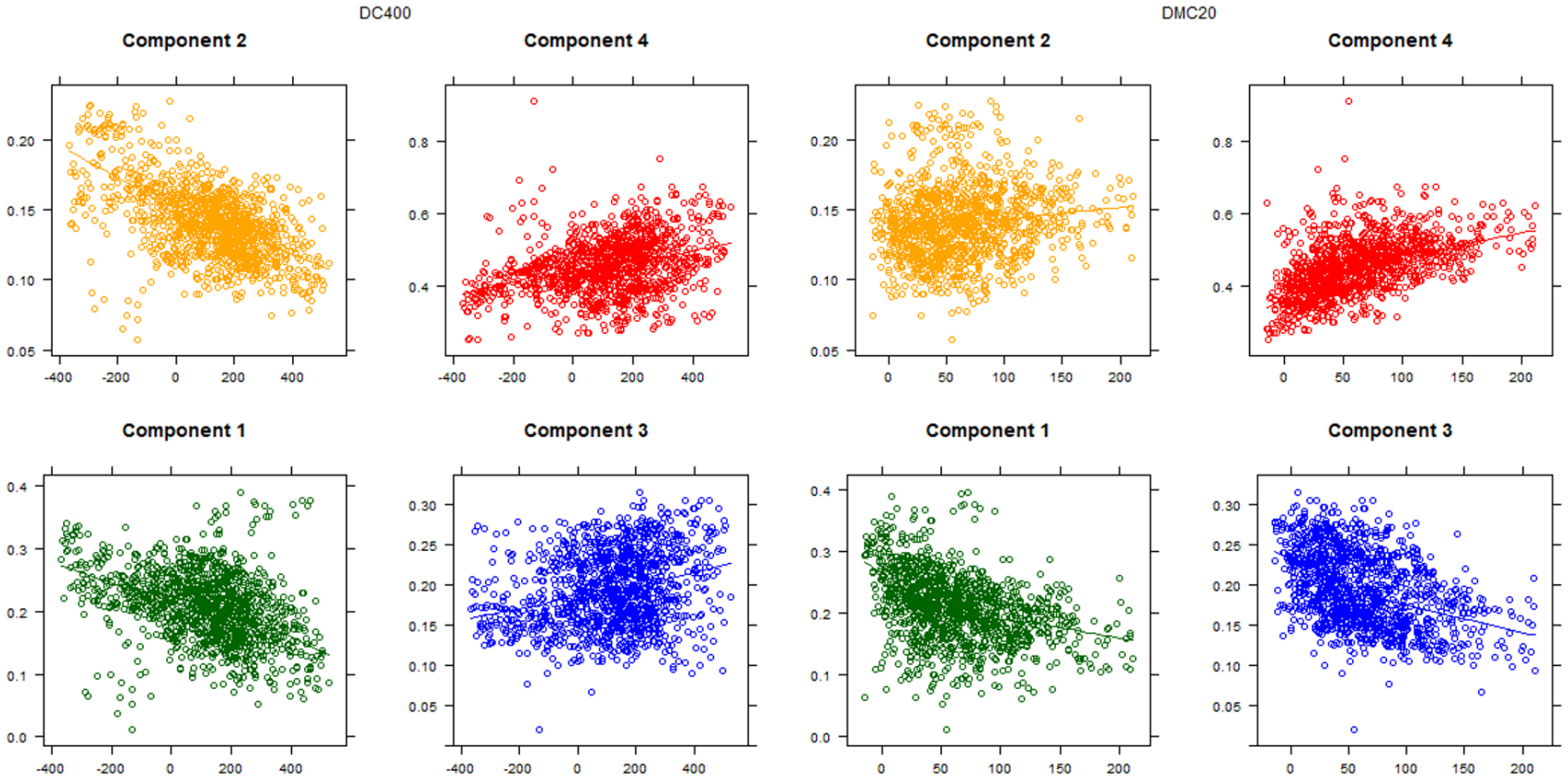


Figure 4.5: The posterior probability estimates (y-axis) by the ADFT of DC (x-axis) on the two left panels and by DMC (x-axis) on the two right panels. As the ADFT of DC increases, the probability of fires with short duration (components 1 and 2) decreases and the probability of fires having long duration tends to increase (components 3 and 4). This suggests that exceedance in temperature and shortage of precipitation will increase the containment time of the fire. As the ADFT of DMC increases, the probability of being identified

as normal size components (component 1 and 3) decreases while the probability of being identified as extreme size components (component 2 and 4) increases.

Table 4.3: Posterior estimates (exponentiated) of the covariate effects obtained from FMJM

	Component 1			Component 2			Component 3		
	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$
intercept	0.032	0.110	1.091	0.419	4.388	47.564	1.000	1.000	1.000
Slope	0.995	0.998	1.001	0.996	0.998	1.001	0.998	1.002	1.005
Elevation	0.999	0.999	0.999	1.000	1.000	1.000	0.999	1.000	1.000
G. Attack Size	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
NorthWest	0.297	0.456	0.702	0.644	0.909	1.283	0.259	0.411	0.651
PrinceGeorge	0.264	0.383	0.551	0.662	0.875	1.153	0.273	0.410	0.609
Kamloops	0.547	0.799	1.153	0.757	1.010	1.350	0.382	0.574	0.852
Southeast	0.410	0.590	0.844	0.608	0.802	1.057	0.561	0.840	1.260
Cariboo	0.297	0.451	0.687	0.673	0.944	1.319	0.297	0.466	0.726
Decade90	0.902	1.176	1.536	0.806	0.992	1.222	0.911	1.192	1.561
Decade00	0.880	1.188	1.617	0.732	0.947	1.226	0.846	1.142	1.553
Decade10	1.165	1.730	2.594	0.705	0.974	1.348	0.907	1.324	1.959
May	1.277	2.890	6.632	1.172	2.212	4.299	0.371	0.885	1.921
Jun	2.435	5.368	12.476	1.071	1.951	3.703	0.440	1.047	2.274
Jul	4.506	10.088	25.108	1.191	2.221	4.326	0.657	1.598	3.685
Aug	8.571	20.028	53.542	1.424	2.752	5.546	0.944	2.389	5.882
Sep and Oct	18.278	47.386	142.783	1.681	3.337	6.965	1.834	5.100	13.957
BUI.intercept	0.894	0.927	0.960	0.951	0.978	1.007	0.948	0.986	1.023
BUI.slope	0.941	0.981	1.022	0.971	1.003	1.038	0.921	0.958	0.996
WIND13	0.668	0.766	0.882	0.827	0.944	1.040	0.868	1.053	1.336
PCP12	0.995	0.998	1.000	0.999	1.001	1.003	0.992	0.995	0.997
DMC20	0.994	0.995	0.996	0.997	0.998	0.999	0.997	0.998	0.999
DC400	0.995	0.998	1.001	0.996	0.998	1.001	0.998	1.002	1.005

4.4 Discussion

In this chapter, we developed a finite mixture model for the joint modeling of fire duration and fire size. The model can be viewed as an extension of the model by Xi et al. (2020) where the joint distribution of the outcomes is separated into components for capturing multimodality in the first stage of the analysis, and effect of covariates are assessed in the second stage. Compared to existing multivariate frameworks such as the Gaussian mixture model, joint modeling has the flexibility to link outcomes to enable a better understanding of how each outcome is related to the other, in this case, whether and how fire duration and size are connected. A factor loading parameter is utilized to account for the scale difference between duration and size in developing the shared variability model. For fires that are classified as having extreme duration and size, as discussed here, the shared variability across these outcomes is identified as the dominant variability term. Compared to the joint model, the Gaussian mixture model is more suitable if the marginals are known to follow univariate Gaussian distributions. Joint modeling, on the other hand, offers an intuitive approach to link the two distributions that provides a natural interpretation of how the two outcomes are connected.

In this analysis, the research objective from the fire science context leads to the development of the four-component mixture model, and how covariates differentially affect each of the components. For instance, the effects of environmental variables on large fires, controlled quickly, are of particular importance for fire suppression, while identifying the conditions leading to small fires with little need to suppress is also crucial for the management of suppression resources. Hence the focus here on a simple four-component model. Alternatively, approaches that estimate the number of subpopulations in the outcomes may be developed by extending current mixture model methodologies for a single outcome (e.g. McLachlan and Peel, 2000). Such approaches would be theoretically and computationally complex yet would provide a more elegant solution and could be considered in the future.

Note that the marginal distributions of the outcomes can be replaced by other location-scale distributions, specifically, the Weibull or the loglogistic distribution. In a univariate analysis by Xi et al. (2020), the lognormal models demonstrate the best fit using the deviance statistic. Unsurprisingly, the lognormal models truncated at duration of 2 days and size of 4 hectares yield a slightly better fit, but the corresponding joint mixture model is more complicated to estimate and needs more detailed investigation in the future in order to resolve identifiability problems and other issues with regards computations.

In the second stage of the analysis, roughly 50% of the p_{ij} are close to one and zero, but none of them is exactly one or zero. In component 1, 2 and 4, the p_{ij} and the transformed p_{ij} differ by no more than (-0.06, 0.015). In component 3, the difference is a bit larger (greater than 0.4) for 20% of the fires. The transformation does not appear to have a strong impact.

The Coast and Columbia Mountains are major topographic features in the Coastal and Southeast Regions of B.C., respectively, whereas the Fraser Plateau is a dominant feature in the Cariboo Region. The finding that fires of extreme duration are smaller in the Coastal and Southeast Regions and larger in the Cariboo Region maybe due to the influence of rugged topography on constraining fire size in the western Cordillera of North America (Krawchuk et al., 2016).

Fire weather conditions influencing fire spread and duration vary daily to seasonally as well as spatially across British Columbia. Forest floor moisture contents are typically higher in the spring following snowmelt, decreasing in July and August, with an opposite trend in DMC and DC. During September and October decreasing day lengths and temperatures and increasing dewpoint overnight limits the daily period for active fire growth. Fires of extreme size and duration would be expected to be more frequent in mid fire season with peak burning conditions as discussed in the previous section. Higher probability of fires of average size and duration, or average size and long duration may be due to the more limited burning conditions. Fire centres may also change their management strategies to less

aggressive suppression actions as cooler temperatures and the end of the fire season approaches, contributing to longer duration fires.

Spring fires are typically wind driven and can result in a large size during a short time. In such cases, fire containment is only effective accompanied with rain events, which limit the size and the duration of the fires. As we saw in the previous section, this is reflected in Figure 5 where wind has the most dominant effect in component 2 (displayed in yellow) while precipitation has the most dominant effect in component 4 (displayed in red).

The consumption of surface organic matter is an important factor in achieving the critical surface fire intensity for crown fire initiation (Van Wagner, 1977). The association between increasing fire size and DMC may reflect increasing surface fuel consumption and probability of crown fire occurring over the duration of the fire, which favours fire growth - fire spread rates increase by about an order of magnitude when a fire transitions from a surface to crown fire.

The finding of increasing fire duration with Drought Code is consistent with high DC values being associated with smouldering combustion in deeper organic layers (Lawson et al., 1997); when smouldering combustion persists in deep organic layers, fires are more difficult or more time consuming to fully extinguish. There are very likely correlations between DC and seasonal effects; DC, in particular, typically increases throughout the fire season, whereas DMC varies more throughout the fire season in response to wetting and drying weather systems.

An important consideration when considering the influence of weather and fire danger variables on fire size and duration is that these measures are interpolated to a fire location from observations at a network of weather stations (with elevation and modeled weather as covariates) that could be from several kilometres to 100 kilometres distant in more remote parts of BC. Temperature, relative humidity, and precipitation (and so DMC and DC) have more spatial correlation over longer distances, and so interpolated values are more accurate than for wind speed.

The two stages of the analysis can be considered differently by incorporating covariates in the mixture model as a direct relationship with the outcomes, for example, in a model for the probability of component membership. This approach would be conceptually more elegant but would not permit ease of computation, as covariate and model selection would be quite computationally intensive. Note that instead of conducting a variable selection, we employed the selected variables in Xi et al. (2020). However, similar variable selection could be employed for the two-stage analysis of the mixture model with relative ease. Non-parametric estimation methods for modeling density functions and regularization methods in variable selection may also be useful in this context and are potential future research directions.

References

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modelling: Three-step approaches using M plus. *Structural Equation Modelling: A Multidisciplinary Journal, 21*(3), 329-341.

Bayham, J. (2013). *Characterizing incentives: an investigation of wildfire response and environmental entry policy.* PhD Thesis, Washington State University

Cumming, S. (2001). A parametric model of the fire-size distribution. *Canadian Journal of Forest Research, 31*(8), 1297-1303.

DaCamara, C. C., Calado, T. J., Ermida, S. L., Trigo, I. F., Amraoui, M., & Turkman, K. F. (2014). Calibration of the Fire Weather Index over Mediterranean Europe based on fire activity retrieved from MSG satellite imagery. *International Journal of Wildland Fire, 23*(7), 945-958.

Dean, C., Nathoo, F., & Nielsen, J. (2007). Spatial and mixture models for recurrent event processes. *Environmetrics.* The official journal of the International Environmetrics Society, 18(7), 713-725.

Douma, J. C., & Weedon, J. T. (2019). Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods in Ecology and Evolution, 10*(9), 1412-1430.

Duchateau, L., & Janssen, P. (2007). *The frailty model*: Springer Science & Business Media.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62*(2), 355-366.

Feng, C., & Dean, C. (2012). Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models. *Environmetrics, 23*(6), 493-508.

Filmon, G. (2004). *Firestorm 2003 provincial review*: Firestorm 2003 Provincial Review.

Finney, M., Grenfell, I. C., & McHugh, C. W. (2009). Modelling containment of large wildfires using generalized linear mixed-model analysis. *Forest Science, 55*(3), 249-255.

Fried, J. S., & Gilless, J. K. (1989). Notes: Expert opinion estimation of fireline production rates. *Forest Science*, 35(3), 870-877.

Gopi, E. (2019). *Pattern Recognition and Computational Intelligence Techniques Using*

Matlab: Springer.

Henderson, R., & Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika, 90*(2), 355-366.

Holmes, T. P., Huggett, R. J., & Westerling, A. L. (2008). Statistical analysis of large wildfires *The Economics of Forest Disturbances* (pp. 59-77): Springer.

Huang, Y., Yan, C., Yin, P., & Lu, M. (2016). A mixture of hierarchical joint models for longitudinal data with heterogeneity, non-normality, missingness, and covariate measurement error. *Journal of biopharmaceutical statistics, 26*(2), 299-322.

Juarez-Colunga, E., Silva, G., & Dean, C. (2017). Joint modelling of zero-inflated panel count and severity outcomes. *Biometrics, 73*(4), 1413-1423.

Krawchuk, M. A., Haire, S. L., Coop, J., Parisien, M. A., Whitman, E., Chong, G., & Miller, C. (2016). Topographic and fire weather controls of fire refugia in forested ecosystems of northwestern North America. *Ecosphere, 7*(12), e01632.

Lawson, B. D., Frandsen, W. H., Hawkes, B. C., & Dalrymple, G. N. (1997). Probability of sustained smoldering ignition for some boreal forest duff types. *Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre, Forest Management Note, 63*.

Lundy, E. R., & Dean, C. (2018). Analyzing Heaped Counts Versus Longitudinal Presence/Absence Data in Joint Zero-inflated Discrete Regression Models. *Sociological Methods & Research*, 0049124118782550.

Maier, M. J. (2014). DirichletReg: Dirichlet regression for compositional data in R.

McLachlan, G. J., & Peel, D. (2004). *Finite mixture models*: John Wiley & Sons.

Morin, A. A., Albert-Green, A., Woolford, D. G., & Martell, D. L. (2015). The use of survival analysis methods to model the control time of forest fires in Ontario, Canada. *International Journal of Wildland Fire, 24*(7), 964-973.

Murphy, K., & Murphy, T. B. (2019). Gaussian parsimonious clustering models with covariates and a noise component. *Advances in Data Analysis and Classification*, 1-33.

Nadeem, K., Taylor, S. W., Woolford, D. G., & Dean, C. B. (2020). Mesoscale spatiotemporal predictive models of daily human-and lightning-caused wildland fire occurrence in British Columbia. *International Journal of wildland fire*, 29(1), 11-27.

Renouf, E., Dean, C. B., Bellhouse, D. R., & McAlister, V. C. (2016). Joint survival analysis of time to drug change and a terminal event with application to drug

failure analysis using transplant registry data. *International Journal of Statistics in Medical Research, 5*(3), 198-213.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods, 11*(1), 54.

Sun, C. (2013). Bivariate Extreme Value Modeling of Wildland Fire Area and Duration. *Forest Science*, 59(6), 649-660.

Taylor, S. W., Woolford, D. G., Dean, C., & Martell, D. L. (2013). Wildfire Prediction to Inform Management: Statistical Science Challenges. *Statistical Science*, 586-615.

Tremblay, P.-O., Duchesne, T., & Cumming, S. G. (2018). Survival analysis and classification methods for forest fire size. *PloS one, 13*(1).

Vink, M. A., Berkhof, J., van de Kassteele, J., van Boven, M., & Bogaards, J. A. (2016). A bivariate mixture model for natural antibody levels to human papillomavirus types 16 and 18: baseline estimates for monitoring the herd effects of immunization. *PloS one, 11*(8).

Wagner, C. V. (1969). A simple fire-growth model. *The Forestry Chronicle, 45*(2), 103-104.

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

Xi, D. D., Dean, C., & Taylor, S. W. (2020). Modelling the duration and size of extended attack wildfires as dependent outcomes. *Environmetrics,* e2619.

Xi, D. D., Taylor, S. W., Woolford, D. G., & Dean, C. (2019). Statistical models of key components of wildfire risk. *Annual review of statistics and its application, 6*, 197-222.

Yoder, J., & Gebert, K. (2012). An econometric model for ex ante prediction of wildfire suppression costs. *Journal of Forest Economics, 18*(1), 76-89.

Chapter 5

5 Joint Modeling of Hospitalization and Mortality of Ontario Covid-19 Cases

5.1 Introduction

In epidemiology, various empirical methods have been developed to quantify the outbreak of infectious diseases. One approach models public health data as time series processes (Zeger et al. 2006), which is typically suitable when an outcome is observed for a long period of time. Time series models generally assume that the observation today is linearly related to the observations lagged several days prior, with additive error terms independently and identically distributed (i.e. i.i.d.) from a normal distribution with a mean of zero and an unknown variance. The average, the trend, and the seasonality of the outcome process can then be specified in the model.

Time series models have previously been used in public health studies of infectious diseases. Examples of the diseases and study regions where time series models have been applied are Campylobacter and measles in Montreal, Canada (Allard, 1998), diarrhoea in Peru (Checkley et al., 2000), and Covid-19 in Italy (Ding et al., 2020). Time series models are prominently studied in fields outside of public health, such as econometrics, where a technique, termed cointegration analysis, can further assess whether there is correlation in the long run between two processes (Pfaff, 2008). For example, cointegration analysis was applied to various processes of stock prices to examine if the SARS outbreak in 2003 had an impact on them (e.g. Chen et al., 2018). As hospitalization data regarding Covid-19 are collected and become available, several authors have indeed studied the relationship between the daily number of cases and stock prices (e.g. Zeren and Hizarci, 2020; Şenol and Zeren, 2020).

Another potential approach for studying the relationship between two outcomes is through joint-outcome modeling (Dunson, 2000; Henderson et al., 2000). One approach links the outcomes through a latent variable, a shared error term that is incorporated in the models for each of the outcomes, which then induces an underlying correlation between the outcomes. The method has been utilized in linking, for example, various outcomes that are count data (Feng and Dean, 2012; Juarez-Colunga et al., 2017), survival data (Tsiatis and Davidian, 2004), and presence/absence data (Lundy and Dean, 2018), where the latent variable is shared among the outcomes. Such methodology has not been considered in linking time series data, and it may provide a novel perspective for understanding the long-run relationship between two time series processes.

In this chapter, we analyze the daily number of new hospitalizations and the daily number of new deaths from Covid-19 in Ontario as autoregressive processes. In infectious disease studies, these two processes are key indicators in an outbreak (e.g. Trivedi et al., 2012). We chose to model hospitalized cases instead of the number of new infections because testing was initially limited to the sickest patients or those recently returned from travel, so that case counts did not reflect the true progression of transmission. Section 2 outlines two frameworks for assessing the relationship between hospitalizations and deaths, where a cointegration analysis and a joint modeling framework are used to understand and model the long-run relationship between these two outcomes. Section 3 presents results of the analysis on the Ontario data using each framework, identifying the unique perspective that each framework provides. Section 4 closes with a discussion of the utility of each of the frameworks and potential ways that the models can be extended.

5.2 Models and Methods

5.2.1 Cointegration Analysis

We assume that the time series process, $y_t, t = p+1, \ldots, n$ follows an autoregressive model with lag p, termed an AR(p) model, defined as

$$y_t = \mu + \theta_1(y_{t-1} - \mu) + \cdots + \theta_p(y_{t-p} - \mu) + \varepsilon_t,$$

where μ is the intercept; $\theta_s = \sigma_s/\sigma_0, s = 1, \ldots, p$ such that $\sigma_s = \mathrm{COV}(y_t, y_{t+s})$, the covariance between y_t and y_{t+s}, is the autocorrelation coefficient associated with lag s; ε_t is the random error assumed to be distributed as i.i.d. $N(0, \sigma^2), t = p+1, \ldots, n$. Inference on the model is straight forward when the time series process is stationary, that is, if the intercept and the autocorrelation are both fixed and do not depend on t. This is equivalent to stating that $|\theta_s| < 1$. A non-stationary process can often become stationary by differencing y_t with respect to time d times, and such process is denoted as $y_t \sim I(d)$. For example, if $y_t \sim I(1)$, then $y_t^* = \Delta y_t = y_t - y_{t-1}$ is stationary; whereas if $y_t \sim I(2)$, then $y_t^{**} = \Delta y_t^* = y_t^* - y_{t-1}^*$ is stationary. The value d is often referred as the *order of integration*.

Two outcome processes $y_{kt}, k = 1, 2$ are cointegrated with other, if there exists an integer constant b such that $y_{1t} \sim I(d)$, $y_{2t} \sim I(d)$ and $z_t \sim I(d-b)$, where z_t is a linear combination of y_{1t} and y_{2t}. In other words, given two processes that are stationary after differencing d times, if their residuals z_t are stationary by differencing less than d times, the two processes are related in a unique long-run relationship and they are termed cointegrated. Heuristically, the processes will deviate, but in a random or stochastic and stationary fashion.

We first need to determine the value of d that supports stationarity in the two outcome processes. Several tests can be used to determine if a process, y_t is stationary. For example, an AR(1) process with intercept zero can be written as

$$y_t = \theta y_{t-1} + \varepsilon_t.$$

For testing that y_t is non-stationary, the Dickey-Fuller test (Dickey and Fuller, 1979) tests that $\pi = 0, \pi = \theta - 1$, in the rearranged model framework

$$y_t - y_{t-1} = \theta y_{t-1} - y_{t-1} + \varepsilon_t$$

$$\Delta y_t = (\theta - 1) y_{t-1} + \varepsilon_t$$
$$\Delta y_t = \pi y_{t-1} + \varepsilon_t.$$

When the null hypothesis is true, the stochastic error term accumulates over time and hence the process is unstable. Hence the null and the alternative hypotheses can be written as

$$H_0: \pi = 0$$
$$H_1: \pi < 0.$$

If there is significant evidence to reject H_0, we conclude that y_t is stationary. More generally, assuming an AR(p) process analogously yields the augmented Dickey-Fuller (ADF) test, utilizing the same null and alternative hypotheses, here using the modeling framework

$$\Delta y_t = \pi y_{t-1} + \sum_{s=1}^{p-1} \gamma_s \, \Delta y_{t-s} + \varepsilon_t.$$

The parameters can be estimated using least squares. The test statistic follows a Dickey-Fuller distribution whose p-value is computed through Monte Carlo methods (i.e. Park, 2002; Wei, 2014; Chang et al., 2017). Although the Dickey-Fuller test is a standard in the literature, we note that alternative tests, such as the Phillips-Perron (PP) test, the Elliott-Rothenberg-Stock (ERS) test, and the Schmidt-Phillips (SP) test may also be used; see Pfaff (2008) for a description of these tests.

5.2.2 Joint Modeling

Let $y_{kt}, k = 1, 2, t = p + 1, \ldots, n$ be two time series processes, each process with lag p_k and $p = \max(p_1, p_2)$. This model assumes that the processes quantify outcomes measured at the same values of t. The model is defined by

$$y_{kt} = \mu_k + \theta_{k1}\left(y_{k,t-1} - \mu_k\right) + \cdots + \theta_{kp_k}\left(y_{k,t-p_k} - \mu_k\right) + b_{kt} + \varepsilon_{kt},$$

where, associated with outcome k: μ_k is the intercept; $\theta_{ks} = \sigma_{ks}/\sigma_{k0}, s = 1, \dots, p_k$ such that $\sigma_{ks} = \mathrm{COV}(y_{kt}, y_{k,t+s})$, the covariance between y_{kt} and $y_{k,t+s}$, is the autocorrelation coefficient with lag s; ε_{kt} are random errors distributed as i.i.d. $N(0, \sigma_k^2)$; and $\boldsymbol{b}_t = (b_{1t}, b_{2t})^T$ is a 2×1 vector of random effects, independent from ε_{kt}, used to model the shared variability between the outcomes. The distribution of $\boldsymbol{b}_t, t = p + 1, \dots, n$, is assumed i.i.d. $Q(\boldsymbol{b}|\boldsymbol{D}) = N_2(\boldsymbol{0}, \boldsymbol{D})$, with a 2×1 mean vector $\boldsymbol{0}$ and a 2×2 symmetric and positive definite variance-covariance $\boldsymbol{D}$. Each outcome is of an order of integration d_k. In other words, the outcomes will need to be differenced d_k times before model development in order to achieve stationarity in the transformed outcomes.

It is convenient to represent the framework in matrix notation. The response $\boldsymbol{Y}_k$, the design matrix $\boldsymbol{X}_k$, and the associated vectors of parameters and random effects are specified as follows

$$\boldsymbol{Y}_k = \boldsymbol{X}_k \boldsymbol{\theta}_k + \mu_k \left(1 - \sum_{s=1}^{p_k} \theta_{k,s} \right) + \boldsymbol{B}_k + \boldsymbol{\epsilon}_k,$$

where

$$\boldsymbol{Y}_k = \begin{bmatrix} y_{k,p+1} \\ \vdots \\ y_{k,n} \end{bmatrix}, \boldsymbol{X}_k = \begin{bmatrix} y_{k,p} & \cdots & y_{k,p+1-p_k} \\ \vdots & \ddots & \vdots \\ y_{k,n-1} & \cdots & y_{k,n-p_k} \end{bmatrix}, \boldsymbol{\theta}_k = \begin{bmatrix} \theta_{k,1} \\ \vdots \\ \theta_{k,p_k} \end{bmatrix}, \boldsymbol{B}_k = \begin{bmatrix} b_{k,p+1} \\ \vdots \\ b_{k,n} \end{bmatrix}, \boldsymbol{\epsilon}_k = \begin{bmatrix} \varepsilon_{k,p+1} \\ \vdots \\ \varepsilon_{k,n} \end{bmatrix}.$$

The joint posterior distribution is expressed as

$$p(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{D}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{b}) Q(\boldsymbol{b}|\boldsymbol{D}) p(\boldsymbol{\mu}) p(\boldsymbol{\theta}) p(\boldsymbol{D}) p(\boldsymbol{\sigma}),$$

where $\boldsymbol{y} = (\boldsymbol{y}_{p+1}, \dots, \boldsymbol{y}_n)$, $\boldsymbol{y}_t = (y_{1t}, y_{2t})$, $\boldsymbol{\mu} = (\mu_1, \mu_2)$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $\boldsymbol{b} = (\boldsymbol{b}_{p+1}, \dots, \boldsymbol{b}_n)$, and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$. The first term on the right-hand side is the conditional likelihood

$$p(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{b}) \propto \prod_{t=p+1}^{n} f(\boldsymbol{y}_t|\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{b}_t),$$

where $f(\boldsymbol{y}_t|\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{b}_t)$ is the joint density function of $\boldsymbol{y}_t$. We impose different constraints on the term b_{kt} and ε_{kt} to create four joint outcome models and these are shown in Table 5.1. For instance, Model B defines the vectors of error terms as $(b_{1t}+\varepsilon_{1t}, b_{2t}+\varepsilon_{2t})^T = (\gamma b_t, b_t+\varepsilon_{2t})^T$. This model assumes that all the variability in y_{1t} is explained by the term b_t which follows i.i.d. $N(0,\sigma_b^2)$ and is scaled by the factor loading parameter γ; as well, that all the variability in y_{2t} is explained by the sum of b_t and the additive error term ε_{2t}, where ε_{2t} is $N(0,\sigma_{\varepsilon 2}^2)$. Since the outcomes are independent, given the shared random effect, we have the joint density expressed as

$$\prod_{t=p+1}^{n} f(\boldsymbol{y}_t|\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{b}_t) = \prod_{t=p+1}^{n}\prod_{k=1}^{2} f_k(y_{kt}|\mu_k,\boldsymbol{\theta}_k,b_{kt}),$$

where $f_k(y_{kt}|\mu_k,\boldsymbol{\theta}_k,b_{kt})$ is the marginal density function of y_{kt}. Finally, the product of the prior distributions is given by

$$p(\boldsymbol{\mu})p(\boldsymbol{\theta})p(\boldsymbol{D})p(\boldsymbol{\sigma}) = \prod_{k=1}^{2}\left[p(\mu_k)p(\theta_{k1})\dots p\left(\theta_{kp_k}\right)p(\sigma_k)\right]p(\gamma)p(\sigma_b).$$

Choices of the distributions of the priors will be discussed more fully in the next section.

5.3 Results and Analysis

5.3.1 Ontario Data

We obtained data from the daily epidemiological summaries provided by Public Health Ontario. To study the delayed effect of hospitalization on mortality, the daily number of new hospitalizations 6 days prior and the daily number of new deaths are defined as the outcomes of interest. We shifted the time between these two outcomes by 6 days because

Table 5.1: Parameterization of the joint models, where $\boldsymbol{u_k} = \boldsymbol{\mu_k} + \boldsymbol{\theta_{k1}}(\boldsymbol{y_{kt-1}} - \boldsymbol{\mu_k}) + \cdots + \boldsymbol{\theta_{kp_k}}(\boldsymbol{y_{kt-p_k}} - \boldsymbol{\mu_k})$

Model Form	$\begin{bmatrix} b_{1t} + \varepsilon_{1t} \\ b_{2t} + \varepsilon_{2t} \end{bmatrix}$	$f_k(y_{kt}\|\mu_k, \boldsymbol{\theta}_k, b_{kt})$	$f(\boldsymbol{y}_t\|\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{b}_t)$
A	$\begin{bmatrix} \gamma b_t + \varepsilon_{1t} \\ b_t \end{bmatrix}$	$y_{1t} \sim N(u_1, \gamma^2\sigma_b^2 + \sigma_{\varepsilon 1}^2)$ $y_{2t} \sim N(u_2, \sigma_b^2)$	$\boldsymbol{y}_t \sim N_2\left(\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} \gamma^2\sigma_b^2 + \sigma_{\varepsilon 1}^2 & \gamma\sigma_b^2 \\ \gamma\sigma_b^2 & \sigma_b^2 \end{bmatrix}\right)$
B	$\begin{bmatrix} \gamma b_t \\ b_t + \varepsilon_{2t} \end{bmatrix}$	$y_{1t} \sim N(u_1, \gamma^2\sigma_b^2)$ $y_{2t} \sim N(u_2, \sigma_b^2 + \sigma_{\varepsilon 2}^2)$	$\boldsymbol{y}_t \sim N_2\left(\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} \gamma^2\sigma_b^2 & \gamma\sigma_b^2 \\ \gamma\sigma_b^2 & \sigma_b^2 + \sigma_{\varepsilon 2}^2 \end{bmatrix}\right)$
C	$\begin{bmatrix} b_t + \varepsilon_{1t} \\ \gamma b_t \end{bmatrix}$	$y_{1t} \sim N(u_1, \sigma_b^2 + \sigma_{\varepsilon 1}^2)$ $y_{2t} \sim N(u_2, \gamma^2\sigma_b^2)$	$\boldsymbol{y}_t \sim N_2\left(\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} \sigma_b^2 + \sigma_{\varepsilon 1}^2 & \gamma\sigma_b^2 \\ \gamma\sigma_b^2 & \gamma^2\sigma_b^2 \end{bmatrix}\right)$
D	$\begin{bmatrix} b_t \\ \gamma b_t + \varepsilon_{2t} \end{bmatrix}$	$y_{1t} \sim N(u_1, \sigma_b^2)$ $y_{2t} \sim N(u_2, \gamma^2\sigma_b^2 + \sigma_{\varepsilon 2}^2)$	$\boldsymbol{y}_t \sim N_2\left(\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} \sigma_b^2 & \gamma\sigma_b^2 \\ \gamma\sigma_b^2 & \gamma^2\sigma_b^2 + \sigma_{\varepsilon 2}^2 \end{bmatrix}\right)$

recent research shows that a reasonable estimate of the median time from hospitalization to death for Covid-19 varies between 4 days (i.e. Richardson et al., 2020; Ontario Agency for Health Protection and Promotion, 2020) to 7.5 days (i.e. Zhou et al., 2020). Additional evidence for using a 6-day lag is that a basic generalized additive model examining the relationship between hospitalizations at various lags and deaths gives the highest deviance explained at a 6-day lag period. There were $n = 78$ observations from March 29 to June 14. Figure 5.1 provides an illustration of the data. The left panel is the base 10 logarithm of the cumulative number of hospitalizations (black solid lines) and deaths (red dashed lines) and the right panel provides the daily number of these outcomes. On the right panel, both processes demonstrate a downward trend starting in May, while their residual, defined as their difference (blue dotted line), appears stationary. We define the daily number of new hospitalizations and new deaths by y_{1t} and y_{2t} and identify here the potential long-run relationship between them, if any.

5.3.2 Cointegration Analysis of Ontario Data

To assess if y_{1t} and y_{2t} are cointegrated, we first need to identify an appropriate model for each process, respectively denoted as AR(p_k), $k = 1,2$. For $p_k = 1, \dots, 10$, the Akaike information criterion (AIC) of the models yield a minimum at $p_1 = 5$ for y_{1t} and at $p_2 = 3$ for y_{2t}. We apply the ADF test to y_{1t} and y_{2t} to determine if they are non-stationary under the models selected with the minimum AIC. The p-values for the tests are 0.258 and 0.193, respectively, suggesting that y_{1t} and y_{2t} are not stationary. Taking the first order difference of each process and reapplying the above procedure on $y_{1t}^* = \Delta y_{1t}$ and $y_{2t}^* = \Delta y_{2t}$ yields a minimum AIC for each model at $p_1 = 9$ for y_{1t}^* and $p_2 = 8$ for y_{2t}^*. Under the models with the minimum AIC, the p-values for the tests are 0.027 and 0.052, respectively, suggesting that y_{1t}^* and y_{2t}^* are stationary. Identifying an appropriate model for $z_t = y_{1t} - y_{2t}$ as an AR(p) process, yields a minimum over $p = 1, \dots, 10$ of the AIC at $p = 3$. The corresponding p-value for testing non-stationarity of z_t, distributed as AR(3), is 0.019, suggesting that z_t is stationary. This evidence indicates that there is long-

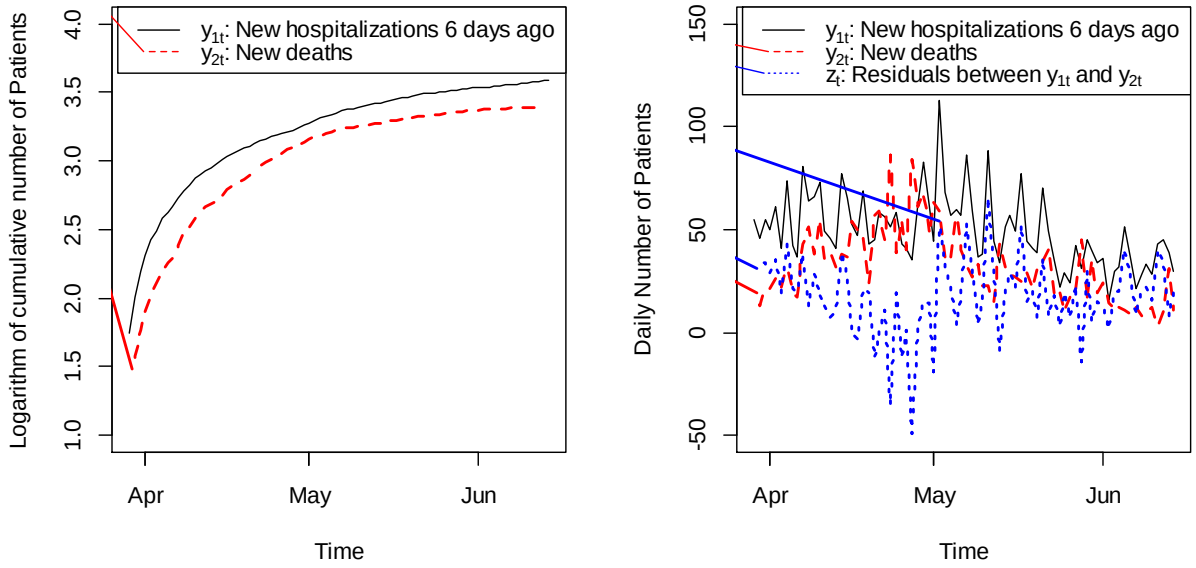


Figure 5.1: The left panel illustrates the logarithm of the cumulative number of hospitalizations 6 days prior (black) and the cumulative number of deaths (red). Hospitalizations and deaths grow with a decreasing rate over time. The right panel plots the daily number of these quantities and their residuals (blue) against time. The processes are identified as having a long-term correlation through the cointegration analysis described in the text.

run correlation between y_{1t} and y_{2t}; that y_{1t} and y_{2t} are cointegrated such that $y_{1t} \sim I(1)$, $y_{2t} \sim I(1)$ and $z_t \sim I(0)$.

5.3.3 Joint Modeling of Ontario Data

The joint model is fitted by the adaptive Markov Chain Monte Carlo (MCMC) method described in Xi et al. (2020). We assume vague priors commonly used in the literature: for $k = 1, 2$ and $s = 1, \dots, p_k$, $p(u_k)$ and $p(\theta_{ks})$ follow i.i.d. $N(0, 10000)$; $p(\gamma)$, $p(\sigma_b)$ and $p(\sigma_{\varepsilon k})$ follow i.i.d. half. $N(0, 10000)$. Credible intervals are obtained as the lower and upper 2.5% quantiles of the posterior density. The goodness of fit of the models are assessed by their deviance information criteria (DIC) with models having low DIC considered to offer a good fit to the data (Spiegelhalter et al., 2002).

We consider model parameterization in three ways. Four forms of the joint model as provided in Table 5.1 are considered; four choices of order of integration based on the result of the cointegration analysis above and additionally exploring the use of the responses themselves as well as first differences: $(d_1, d_2) = (0,0), (0,1), (1,0), (1,1)$; a hundred combinations of the number of lags: $p_k = 1, \dots, 10$ for each of $k = 1,2$. Hence a total of 1600 models are estimated. We first select the models with the optimal number of lags using the DIC criterion under each of the joint models and the forms of the order of integration, then choose an overall model that provides the best fit.

Table 5.2 lists the 16 optimal models along with their DIC. Including an outcome-specific variability term, ε_{kt}, in modeling the outcome death (i.e. as in models B and D) yields a much better fit than incorporating such a term in modeling the outcome hospitalization (i.e. as in models A and C). Models with a factor loading on hospitalization (i.e. B) have slightly better fit than those with a factor loading on death (i.e. D). Both of the outcomes hospitalization and death are best fitted with an order of integration $d_k = 1$. This is consistent with the results from our cointegration analysis. We note that for all models omitting the additive error term, ε_{kt}, yields that the maximum number of lagged terms

Table 5.2: Statistics assessing model fits for the candidate models

Form	d1	d2	p1	p2	DIC
A	0	0	4	10	518
A	0	1	4	10	518
A	1	0	5	10	516
A	1	1	3	10	520
B	0	0	10	4	493
B	0	1	10	3	487
B	1	0	10	2	485
B	1	1	10	3	482
C	0	0	5	10	521
C	0	1	4	10	517
C	1	0	5	10	517
C	1	1	3	10	518
D	0	0	10	4	494
D	0	1	10	3	486
D	1	0	10	2	486
D	1	1	10	3	484

needs to be considered. We note that a better fit may be produced by incorporating an even higher number of lagged terms in the model, but a model with high number of lagged terms is not conducive to model parsimony.

The parameter estimates along with the 95% credible intervals for model B, with the lowest DIC, are presented in Table 5.3. The intercepts of the model are non-significant with estimates of μ_k respectively as $1.560\ (-0.171, 3.350)$ and $-1.631\ (-0.280, 1.080)$ for $k = 1,2$; recall the responses here are first order differences since $d_k = 1, k = 1,2$. Although model B incorporated 10 lagged terms for the outcome hospitalization, only the credible intervals for the coefficients of the first two lagged terms do not include zero. The values of the coefficients of the leading lagged terms, θ_{k1}, are estimated respectively as $-0.295\ (-0.410, -0.165)$ and $-0.741\ (-0.977, -0.506)$, suggesting that the correlation of death $(k = 2)$ with observations on the previous day is stronger than the corresponding correlation for hospitalization $(k = 1)$. The standard deviation of the shared

Table 5.3: Posterior estimates of the model parameters

	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$
μ_1	-0.171	1.560	3.350
μ_2	-1.631	-0.280	1.080
θ_{11}	-0.410	-0.295	-0.165
θ_{12}	-0.341	-0.172	-0.046
θ_{13}	-0.053	0.020	0.094
θ_{14}	0.056	0.143	0.283
θ_{15}	-0.042	0.056	0.225
θ_{16}	-0.006	0.058	0.161
θ_{17}	-0.170	-0.077	0.010
θ_{18}	-0.101	-0.038	0.044
θ_{19}	-0.093	0.054	0.161
θ_{110}	-0.071	0.023	0.116
θ_{21}	-0.977	-0.741	-0.506
θ_{22}	-0.598	-0.318	-0.035
θ_{23}	-0.489	-0.254	-0.017
γ	11.129	13.644	14.876
σ_b	1.151	1.416	1.841
$\sigma_{\varepsilon 2}$	10.610	12.533	15.131

variability, σ_b, has an estimate of 1.416 (1.151,1.841). The factor loading parameter, γ, and the standard deviation of the outcome-specific variability in modelling death, σ_2, have estimates of 13.644 (11.129,14.876) and 12.533 (10.610,15.131), respectively. The variance of the outcomes is parameterized as $\gamma^2\sigma_b^2$ for y_{1t} and $\sigma_b^2 + \sigma_{\varepsilon 2}^2$ for y_{2t}, 373.06 (258.57,537.15) and 159.08 (114.48,229.82). These estimates suggest that although there is dependence between hospitalizations and deaths, much of the variability in these outcomes is unexplained as outcome specific random error.

The left panel of Figure 5.2 illustrates the posterior estimates of the shared random effect, b_t, along with their 95% credible intervals, plotted against time. The peak value of 5.38 on May 02 reflects the peak of hospitalization six days prior in Figure 5.1. The right panel of

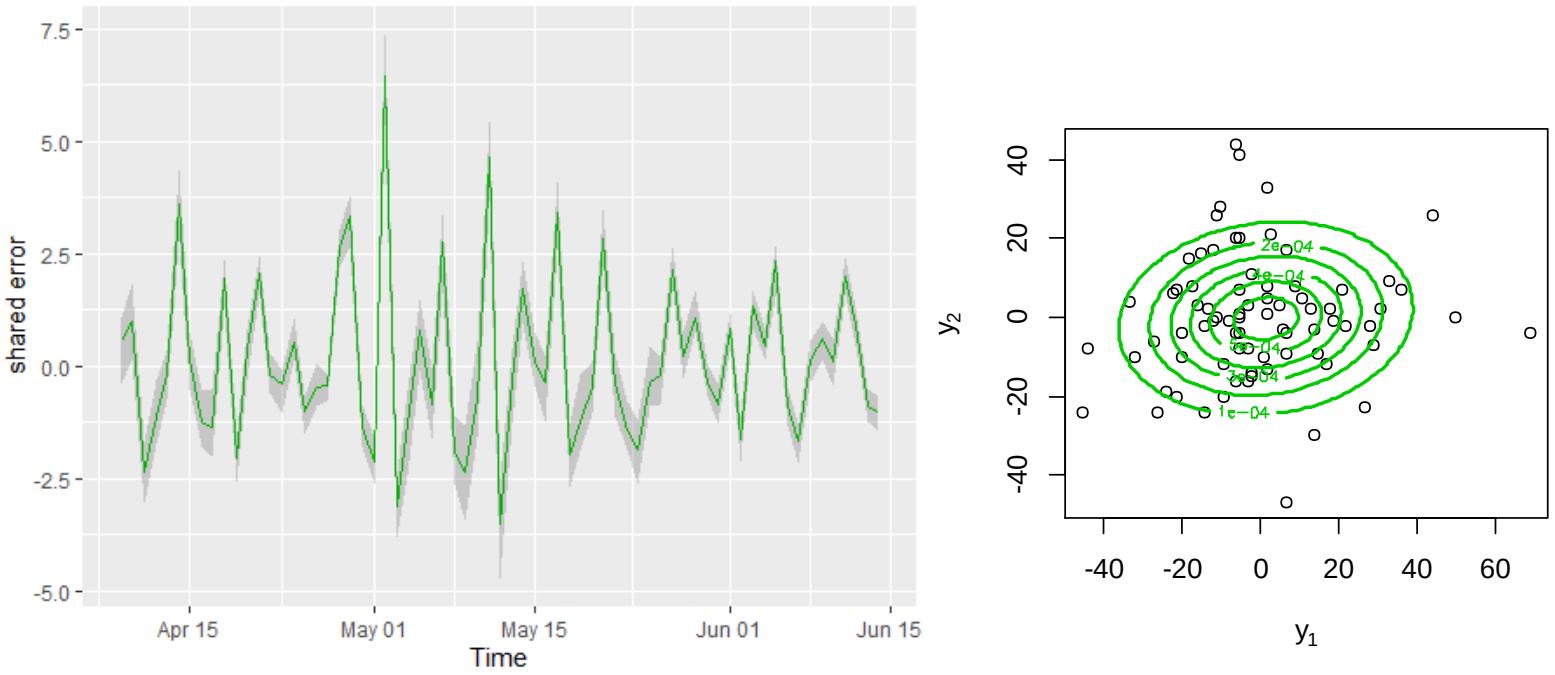


Figure 5.2: Posterior estimates of the shared random effect (left panel) and the estimated joint distribution (right panel) of the outcomes, $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ being the first order difference of hospitalizations six days prior and deaths, respectively. The posterior estimates of the shared random effect have a peak on May 02, reflecting the peak in daily hospitalizations. The estimated joint distribution of the outcomes demonstrates a weak dependence between the outcomes.

Figure 5.2 provides the estimated joint distribution of the outcomes, reflecting the positive, weak correlation in these outcomes as discussed earlier.

5.4 Discussion

The co-integration analysis identified a long-run relationship between hospitalizations and deaths subsequently modeled through a joint outcome autoregressive model with a shared latent random effect. The first order differences of hospitalizations 6 days prior, and deaths, in the joint outcome model are autoregressively correlated with the observations two and three days ago respectively. The autocorrelation could be a result of the reporting schedule by public health units as many of them do not report on weekends. The weak dependence between the outcomes may be due in part to reporting lags in both hospitalizations and deaths in the Ontario data. The data are reported to Public Health Ontario by 34 different public health units, and while the reporting lag is not currently quantified, it likely varies by health unit and by outcome. In future work we hope to be able to adjust for the lags in both outcomes.

The framework can be extended in several ways to reduce the unexplained variability, enhance predictability, and sharpen linkages across the outcomes. An ARIMA model that has moving averaging error terms may better describe the structure of the variability, and it may also be useful to incorporate autoregressive structures in the latent random effect. Comparisons with multivariate time series frameworks may help identify the benefits of using shared random effect for modeling joint outcomes beyond ease of interpretation. Environmental data associated with each day, such as temperature and humidity (i.e. Chan et al., 2011; Sajadi et al., 2020), as well as geographical information, if available, may be included into the model as explanatory covariates. As the uncertainty in the model is reduced and with stronger linkages evidenced across the outcomes, given any current increment in hospitalization, more accurate predictions of future mortality may be obtained through the estimated joint distribution of the outcomes.

References

Anderson R, May R. 1992. Infectious Diseases of Humans. Oxford: Oxford University Press

Albulescu C. 2020. Do COVID-19 and crude oil prices drive the US economic policy uncertainty? *arXiv preprint arXiv:2003.07591*

Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. 2020. Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in brief*: 105340

Breban R, Vardavas R, Blower S. 2005. Linking population-level models with growing networks: a class of epidemic models. *Physical Review E* 72: 046110

Breban R, Vardavas R, Blower S. 2007. Theory versus data: how to calculate R 0? *PLoS One* 2: e282

Chan K-H, Peiris JM, Lam S, Poon L, Yuen K, Seto WH. 2011. The effects of temperature and relative humidity on the viability of the SARS coronavirus. *Advances in virology* 2011

Chang Y, Sickles RC, Song W. 2001. Bootstrapping unit root tests with covariates. *mimeographed, Department of Economics, Rice University*

Dickey DA, Fuller WA. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74: 427-31

Ding G, Li X, Shen Y, Fan J. 2020. Brief Analysis of the ARIMA model on the COVID-19 in Italy. *medRxiv*

Ontario Agency for Health Protection and Promotion (Public Health Ontario). 2020. Weekly Epidemiologic summary: COVID-19 and Severe Outcomes in Ontario. Toronto, ON: Queen's Printer for Ontario

Park JY. 2003. Bootstrap unit root tests. *Econometrica,* 71: 1845-95

Pfaff B. 2008. *Analysis of integrated and cointegrated time series with R*: Springer Science & Business Media

Richardson, S., Hirsch, J. S., Narasimhan, M., Crawford, J. M., McGinn, T., Davidson, K. W., ... & Cookingham, J. (2020). Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. JAMA : the Journal of the American Medical Association, 323(20), 2052–2059.

Sajadi MM, Habibzadeh P, Vintzileos A, Shokouhi S, Miralles-Wilhelm F, Amoroso A.

2020. Temperature, Humidity, and Latitude Analysis to Estimate Potential Spread and Seasonality of Coronavirus Disease 2019 (COVID-19). *JAMA Network Open* 3: e2011834-e34

ŞENOL Z, ZEREN F. 2020. Coronavirus (COVID-19) and Stock Markets: The Effects of The Pandemic on the Global Economy. *Eurasian Journal of Researches in Social and Economics (EJRSE)* 7: 1-16

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. Statistica Sinica, 809-834.

Wei J. 2014. *On Bootstrap Evaluation of Tests for Unit Root and Cointegration*. Acta Universitatis Upsaliensis

Zhao S, Lin Q, Ran J, Musa SS, Yang G, et al. 2020. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International journal of infectious diseases* 92: 214-17

Chapter 6

6 Future Work

As the second chapter suggests, the field of wildland fire science has many open questions that could benefit from statistical and analytical methods. The following subsections identify preliminary ideas on two areas that are of importance to collaborators at the Pacific Forestry Centre.

6.1 Future Work as Identified in the Articles Integrated to Form the Thesis

This section identifies research directions discussed in Chapter 2 to Chapter 5 of the thesis but not developed in later chapters. The key ideas are summarized as:

- The work in Chapter 2 identified the need for the development of quantitative risk assessment methods that combine mathematical and statistical models for the estimation of complex fire dynamics. For example, fire management agencies and property insurers are interested in the probability that a fire may be ignited and spread into a nearby town on a given day. A hybrid approach utilizing fire occurrence models (Section 2.2) and burn probability modeling (Section 2.7) could provide analytic results at an appropriate temporal scale, reported as point estimates associated with standard errors.

- In Chapter 4, we used a fixed number of components in the mixture model because of the computational complexity of the model even with the number of subpopulation fixed. More generally, it will be useful to formulate computationally efficient approaches that estimate the number of subpopulations in the outcomes. Such approaches might build on algorithms developed for mixture model methodologies for a single outcome. Adding covariates in the mixture model as a direct relationship with the outcomes, applying non-parametric estimation methods for modeling density functions, and introducing regularization

methods in variable selection may also be useful in identifying the effect of environmental variables.

There are several extensions that can be considered for the analysis in Chapter 5. Indeed, this work is at an early stage of development. One key issue is adjusting for the reporting lag in the number of hospitalizations and deaths. As well, the joint time series models can be extended by incorporating environmental and geographical covariates, and importantly, autoregressive structures in the latent random effect, and moving average error terms. Comparisons with multivariate time series frameworks may help identify the benefits of using shared random effect for modeling joint outcomes beyond ease of interpretation.

The context and the methods discussed in the fire science context may be applied in the analysis of Covid data and vice versa. For instance, the log-transformed number of fires over time as well as the log-transformed area burned can be regarded as two processes evolving over time, which may be cointegrated. As well, if the Covid data arise from latent subpopulations, a mixture model may provide a better fit in that analysis.

6.2 A Framework for Predicting Daily Fire Load

This section identifies an important research project that is currently under development and that considers methods for efficient resource allocation for fire suppression activities.

Statistical and machine learning methods for predicting the arrival of extreme fires have also been utilized in the development of fire science in recent decades (e.g. Mitsopoulos and Giorgos Mallinis. 2017; Rodrigues et al. 2019; Nadeem et al. 2020). Under the fire risk modeling framework discussed in Chapter 1, we propose an integration of statistical methods for modeling fire duration and fire size, and machine learning methods for predicting fire arrivals.

The proposed work will be a component in the fire prediction system that is currently under development by Natural Resources Canada. The system will contain historical fire records up to the present, together with their associated environmental variables, updated on a daily basis. Daily weather predictions for the next two weeks provided by Environmental Canada will also be included in the system. We propose that machine learning techniques will determine the probability of a fire arrival, and survival models will be used to predict the day-to-day fire behavior. The forecasted number of fires, referred to as the *fire load*, for each day will be presented in five fire size classes. The term fire load is a managerial term reflecting the suppression resource allocation in the province of British Columbia. A stochastic model for fire load is discussed in Morin (2014).

Let $\hat{B}_{c,t}, c = 1, \dots, 5, t = 1, \dots, 14$ be the t-day ahead forecast of the predicted fire load in fire size class c at the end of day t, such that

$$\hat{B}_{c,t} = \hat{E}_{c,t} + \hat{V}_{c,t}$$

where $\hat{E}_{c,t}$ is the predicted number of fires currently active that will continue to day t in class c; $\hat{V}_{c,t}$ is the predicted number of new arrivals that will still be active on day t in class c. Hence $\hat{E}_{c,t}$ is calculated as

$$\hat{E}_{c,t} = \sum_{i=1}^{n_c} \hat{S}\left(t + a_{i,c} | a_{i,c}, \boldsymbol{x}_{i,c,t}\right)$$

where i indexes fires currently active, $i = 1, \dots, n_c$ in class $c, c = 1, \dots, 5$; $a_{i,c}$ is the age of active fire i in class c; $\boldsymbol{x}_{i,c,t}$ are covariates associated with fire i in class c on day t; $\hat{S}\left(t + a_{i,c} | a_{i,c}, \boldsymbol{x}_{i,c,t}\right)$ is the conditional survivor function, the probability that an active fire of age a lasts t more days. We refer to this survival probability as the residual survival probability at day t, calculated as:

$$\hat{S}\left(t + a_{i,c} | a_{i,c}, \boldsymbol{x}_{i,c,t}\right) = \frac{\hat{S}\left(t + a_{i,c} | \, \boldsymbol{x}_{i,c,t}\right)}{\hat{S}\left(a_{i,c} | \, \boldsymbol{x}_{i,c,t}\right)}.$$

Survivor probabilities are estimated based on models derived from analyses of historical data.

To obtain $\hat{V}_{c,t}$, we estimate arrivals on a grid of M cells that are of size 20km by 20km over the province using machine learning techniques such as random forest. This allows the prediction of arrivals based on historical data with about the same covariates as at t, including environmental covariates as well as seasonality. Let

$$\hat{R}_{c,t} = \sum_{j=1}^{m} \hat{R}_{j,c,t} ,$$

where $\hat{R}_{j,c,t}, j = 1, \dots, m, c = 1, \dots, 5, t = 1, \dots, 14$ is the predicted number of arrivals in cell j, in class c, at day t, estimated using a fire occurrence model. Then

$$\hat{V}_{c,t} = \sum_{k=1}^{t} \hat{R}_{c,k} S\left(t - k + 1 | \boldsymbol{x}_{i,c,t}\right)$$

is the predicted number of new arrivals to the system that will still be active in class c on day t.

By estimating the two-week ahead forecast of fire load, $\hat{B}_{c,t}$, we are then able to predict the fire suppression resources required provincially and hence whether there is an excess of resources available for sharing with other provinces or a deficit requiring the borrowing of resources that may not be utilized elsewhere. Extending this model to encompass all provinces could lead to a Canadian resource allocation framework that could optimize how resources move across Canada for fire suppression purposes.

References

Mitsopoulos, I., & Mallinis, G. (2017). A data-driven approach to assess large fire size generation in Greece. Natural Hazards, 88(3), 1591-1607.

Morin, A. A. (2014). *A Spatial Analysis of Forest Fire Survival and a Marked Cluster Process for Simulating Fire Load.* The University of Western Ontario.

Nadeem, K., Taylor, S. W., Woolford, D. G., & Dean, C. B. (2020). Mesoscale spatiotemporal predictive models of daily human-and lightning-caused wildland fire occurrence in British Columbia. International journal of wildland fire, 29(1), 11-27.

Rodrigues, M., Alcasena, F., & Vega-García, C. (2019). Modeling initial attack success of wildfire suppression in Catalonia, Spain. Science of the total environment, 666, 915-927.

Appendices

Appendix 3A: Full Conditional Posterior Distributions of The Models Used in The Analysis

Model 1n:

Let $\Omega = \{\mu_k, \beta_{kr}, \sigma_k, \delta\}$ where $i = 1, .., n$, $k = 1,2$, $r = 1, \dots, R_k$. Define $\varepsilon_{ik} = \frac{\log t_{ik} - \mu_k - \boldsymbol{\beta}_k^T \boldsymbol{x}_{ik}}{\sigma_k}$

$$p(\mu_k | \boldsymbol{t}, \Omega_{-\mu_k}) \propto \left\{ \prod_i \exp\left[-\frac{1}{2}\left(\frac{\varepsilon_{ik}^2 - 2\delta\varepsilon_{i1}\varepsilon_{i2}}{1-\delta^2} \right) \right] \exp\left(\frac{\varepsilon_{ik}^2}{2} \right) \right\} \exp\left(-\frac{\mu_k^2}{2} \right)$$

$$p(\beta_{kr} | \boldsymbol{t}, \Omega_{-\beta_{kr}}) \propto \left\{ \prod_i \exp\left[-\frac{1}{2}\left(\frac{\varepsilon_{ik}^2 - 2\delta\varepsilon_{i1}\varepsilon_{i2}}{1-\delta^2} \right) \right] \exp\left(\frac{\varepsilon_{ik}^2}{2} \right) \right\} \exp\left(-\frac{\beta_{kr}^2}{2(10^2)} \right)$$

$$p(\sigma_k | \boldsymbol{t}, \Omega_{-\sigma_k}) \propto \left\{ \prod_i \exp\left[-\frac{1}{2}\left(\frac{\varepsilon_{ik}^2 - 2\delta\varepsilon_{i1}\varepsilon_{i2}}{1-\delta^2} \right) \right] \exp\left(\frac{\varepsilon_{ik}^2}{2} \right) \right\}, 0 < \sigma_k < 100$$

$$p(\delta | \boldsymbol{t}, \Omega_{-\delta}) \propto \left\{ \prod_i (1-\delta^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\frac{\varepsilon_{i1}^2 + \varepsilon_{i2}^2 - 2\delta\varepsilon_{i1}\varepsilon_{i2}}{1-\delta^2} \right) \right] \right\}, 0 < \delta < 1.$$

Model 2a:

Let $\Omega = \{\mu_k, \beta_{kr}, \sigma_2, \gamma, \sigma_b\}$ where $i = 1, .., n$, $k = 1,2$, $r = 1, \dots, R_k$. Define $b_{i1} = b_i$, $b_{i2} = \gamma b_i$, $\boldsymbol{b}_i = (b_{i1}, b_{i2})$, $\sigma_1 = 1$.

Let $\boldsymbol{D} = \begin{bmatrix} \sigma_b^2 & 0 \\ 0 & \gamma^2\sigma_b^2 \end{bmatrix}$ be a 2×2 symmetric and positive definite variance-covariance

$$p(\mu_k|\boldsymbol{t},\Omega_{-\mu_k}) \propto \left\{\prod_i \exp\left[-\frac{1}{2}\left(\frac{\log t_{ik} - b_{ik} - \boldsymbol{\beta}_k^T \boldsymbol{x}_{ik} - \mu_k}{\sigma_k}\right)^2\right]\right\} \exp\left(-\frac{\mu_k^2}{2}\right)$$

$$p(\beta_{kr}|\boldsymbol{t},\Omega_{-\mu_k}) \propto \left\{\prod_i \exp\left[-\frac{1}{2}\left(\frac{\log t_{ik} - b_{ik} - \beta_{kr} x_{ikr} - \mu_k}{\sigma_k}\right)^2\right]\right\} \exp\left(-\frac{\beta_{kr}^2}{2(10^2)}\right)$$

$$p(\sigma_2|\boldsymbol{t},\Omega_{-\sigma_2}) \propto \prod_i (\sigma_2^2)^{-1/2} \exp\left[-\frac{1}{2}\left(\frac{\log t_{i2} - \gamma b_i - \boldsymbol{\beta}_2^T \boldsymbol{x}_{i2} - \mu_2}{\sigma_2}\right)^2\right], 0 < \sigma_2 < 100$$

$$p(\gamma|\boldsymbol{t},\Omega_{-\gamma})$$
$$\propto \left\{\prod_i \exp\left[-\frac{1}{2}\left(\frac{\log t_{i2} - \gamma b_i - \boldsymbol{\beta}_2^T \boldsymbol{x}_{i2} - \mu_2}{\sigma_2}\right)^2\right]\right\} |\boldsymbol{D}|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{b}_i^T \boldsymbol{D}^{-1} \boldsymbol{b}_i\right] \exp\left(-\frac{\gamma^2}{2(10^2)}\right)$$

$$p(\sigma_b|\boldsymbol{t},\Omega_{-\sigma_b}) \propto \prod_i |\boldsymbol{D}|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{b}_i^T \boldsymbol{D}^{-1} \boldsymbol{b}_i\right], 0 < \sigma_b < 100.$$

Model 2m:

Let $\Omega = \{\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{D}\}$ where $\boldsymbol{t}_i = (t_{i1}, t_{i2})$, $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$, $\boldsymbol{\mu} = (\mu_1, \mu_2)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, $i = 1, .., n$, $k = 1,2$, $r = 1, \dots, R_k$.

Let $\boldsymbol{D} = \begin{bmatrix} \sigma_{b11}^2 & \sigma_{b12}^2 \\ \sigma_{b12}^2 & \sigma_{b22}^2 \end{bmatrix}$ be a 2×2 symmetric and positive definite variance-covariance, $\boldsymbol{R} = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.1 \end{bmatrix}$.

Define $\boldsymbol{\varepsilon} = \sum_{\boldsymbol{i}} (\log \boldsymbol{t}_i - \boldsymbol{\beta}^T \boldsymbol{x}_i - \boldsymbol{\mu})(\log \boldsymbol{t}_i - \boldsymbol{\beta}^T \boldsymbol{x}_i - \boldsymbol{\mu})^T$

$$p(\boldsymbol{\mu}|\boldsymbol{t},\Omega_{-\boldsymbol{\mu}}) \propto \left\{|\boldsymbol{D}|^{-n/2} \exp\left[-\frac{1}{2} Tr(\boldsymbol{D}^{-1}\boldsymbol{\varepsilon})\right]\right\} \prod_k \exp\left(-\frac{\mu_k^2}{2}\right)$$

$$p(\boldsymbol{\beta}|\boldsymbol{t},\Omega_{-\boldsymbol{\beta}}) \propto \left\{|\boldsymbol{D}|^{-n/2}\exp\left[-\frac{1}{2}Trace(\boldsymbol{D}^{-1}\boldsymbol{\varepsilon})\right]\right\}\prod_{k}\prod_{r}\exp\left(-\frac{\beta_{kr}^{2}}{2(10^{2})}\right)$$

$$p(\boldsymbol{D}|\boldsymbol{t},\Omega_{-\boldsymbol{D}}) \propto |\boldsymbol{R}||\boldsymbol{b}|^{-1/2}\exp\left[-\tfrac{1}{2}Tr(\boldsymbol{RD})\right].$$

Appendix 3B: Comparison of Fit of Candidate Models Based on DIC and WAIC for the Fire Data

The table below displays the goodness of fit of the candidate models. Under the copula model framework, models with the static covariates generally have better measures of fit than those with all covariates or no covariate. Models with a Normal copula form generally fit better than using other copula forms. Under a joint model framework, models with the full covariates fit better than those with static covariates or the null model. Importantly, note that joint modeling always outperforms modeling the two outcomes separately. Note that the copula models, the joint form and the multivariate form of the joint models are not nested, and hence the goodness of fit metrics for those models are not directly comparable (Gelman et al., 2014). Hence the normal copula model (1n), the factor loading model (2a), and the multivariate model (2s) are chosen for discussion. Static covariates are considered in depth, while a forward selection procedure is employed for each of the four categories of derived covariates.

		Covariate Structure					
Framework	Form	All Covariates		Static covariates		No Covariate	
		DIC	WAIC	DIC	WAIC	DIC	WAIC
Copula Model (1)	Normal (n)	18237907	506	17858429	218	18239100	282
	Clayton (c)	18237937	560	17858422	243	18239036	242
	Gumbel (g)	18237932	530	17858463	266	18239153	340
	Frank (f)	17857496	562	17858489	342	17858687	416
Joint Model (2)	factor loading (a)	14704	16122	14770	16946	17635	17154
	separate form (s)	16874	16496	18062	17659	17735	17735
	multivariate form (m)	4668	4673	5798	5664	5979	5835

Appendix 4C: Full Conditional Posterior Distributions of the Models in the Analysis

Model FMJM:

Let $\Omega = \{\mu_{jk}, \sigma_{jk}, b_{ij}, \gamma_j, \sigma_{bj}, \pi_j, z_i\}$ where $i = 1,..,n$, $k = 1,2$, $j = 1,...,J$

Define $b_{ij1} = b_{ij}$, $b_{ij2} = \gamma_j b_{ij}$, $\boldsymbol{b_{ij}} = (b_{ij1}, b_{ij2})$, $\sigma_{j2} = \sigma_2$,

Let $\boldsymbol{D}_j = \begin{bmatrix} \sigma_{bj}^2 & 0 \\ 0 & \gamma_j^2 \sigma_{bj}^2 \end{bmatrix}$ be a 2×2 symmetric and positive definite variance-covariance

$$p\left(\mu_{jk}|\boldsymbol{t}, \Omega_{-\mu_{jk}}\right) \propto \left\{\prod_i \exp\left[-\frac{1}{2}\left(\frac{\log t_{ik} - b_{ijk} - \mu_{jk}}{\sigma_{jk}}\right)^2\right]\right\} \exp\left(-\frac{\mu_{jk}^2}{2}\right)$$

$$p\left(\sigma_{jk}|\boldsymbol{t}, \Omega_{-\sigma_{jk}}\right) \propto \prod_i \left(\sigma_{jk}^2\right)^{-1/2} \exp\left[-\frac{1}{2}\left(\frac{\log t_{ik} - \gamma_j b_{ijk} - \mu_{jk}}{\sigma_{jk}}\right)^2\right], 0 < \sigma_{jk} < 100$$

$$p\left(\gamma_j|\boldsymbol{t}, \Omega_{-\gamma_j}\right) \propto \left\{\prod_i \exp\left[-\frac{1}{2}\left(\frac{\log t_{i2} - \gamma_j b_{ij2} - \mu_{j2}}{\sigma_{j2}}\right)^2\right]\right\} |\boldsymbol{D}_j|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{b}_{ij}^T \boldsymbol{D}_j^{-1} \boldsymbol{b}_{ij}\right] \exp\left(-\frac{\gamma_j^2}{2(10^2)}\right)$$

$$p\left(\sigma_{bj}|\boldsymbol{t}, \Omega_{-\sigma_{bj}}\right) \propto \prod_i |\boldsymbol{D}_j|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{b}_{ij}^T \boldsymbol{D}_j^{-1} \boldsymbol{b}_{ij}\right], 0 < \sigma_{bj} < 100$$

$$p\left(\pi_j|\boldsymbol{t}, \Omega_{-\pi_j}\right) \propto \text{Dir}(\boldsymbol{1})$$

$$p\left(z_i|\boldsymbol{t}, \Omega_{-\sigma_{bj}}\right) \propto \text{Multinomial}(\boldsymbol{1}, \pi).$$

Model FMBM:

Let $\Omega = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \pi_j, z_i\}$ where $\boldsymbol{t}_i = (t_{i1}, t_{i2})^T$, $\boldsymbol{\mu}_j = (\mu_{j1}, \mu_{j2})^T$, $i = 1, .., n$, $k = 1,2$, $j = 1, \dots, J$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} \sigma_{N1}^2 & \rho_1 \sigma_{N1} \sigma_{N2} \\ \rho_1 \sigma_{N1} \sigma_{N2} & \sigma_{N2}^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} \sigma_{N1}^2 & \rho_2 \sigma_{N1} \sigma_{E2} \\ \rho_2 \sigma_{N1} \sigma_{E2} & \sigma_{E2}^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_3 = \begin{bmatrix} \sigma_{E1}^2 & \rho_3 \sigma_{E1} \sigma_{N2} \\ \rho_3 \sigma_{E1} \sigma_{N2} & \sigma_{N2}^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_4 = \begin{bmatrix} \sigma_{E1}^2 & \rho_4 \sigma_{E1} \sigma_{E2} \\ \rho_4 \sigma_{E1} \sigma_{E2} & \sigma_{E2}^2 \end{bmatrix}$$

be 2×2 symmetric and positive definite variance-covariance matrices. Define $\boldsymbol{\varepsilon}_j = \sum_i (\log \boldsymbol{t}_i - \boldsymbol{\mu}_j)(\log \boldsymbol{t}_i - \boldsymbol{\mu}_j)^T$

$$p\left(\boldsymbol{\mu}_j | \boldsymbol{t}, \Omega_{-\boldsymbol{\mu}_j}\right) \propto \left\{ |\boldsymbol{\Sigma}_j|^{-n/2} \exp\left[-\frac{1}{2} Tr\left(\boldsymbol{\Sigma}_j^{-1} \boldsymbol{\varepsilon}_j\right) \right] \right\} \prod_k \exp\left(-\frac{\mu_{jk}^2}{2} \right)$$

$$p\left(\boldsymbol{\Sigma}_j | \boldsymbol{t}, \Omega_{-\boldsymbol{\Sigma}_j}\right) \propto \left\{ |\boldsymbol{\Sigma}_j|^{-n/2} \exp\left[-\frac{1}{2} Tr\left(\boldsymbol{\Sigma}_j^{-1} \boldsymbol{\varepsilon}_j\right) \right] \right\} \prod_k \left(\sigma_{jk}^2\right)^{-1/2}, 0 < \sigma_{jk} < 100$$

$$p\left(\pi_j | \boldsymbol{t}, \Omega_{-\pi_j}\right) \propto \text{Dir}(\mathbf{1})$$

$$p\left(z_i | \boldsymbol{t}, \Omega_{-\sigma_{bj}}\right) \propto \text{Multinomial}(\mathbf{1}, \pi).$$

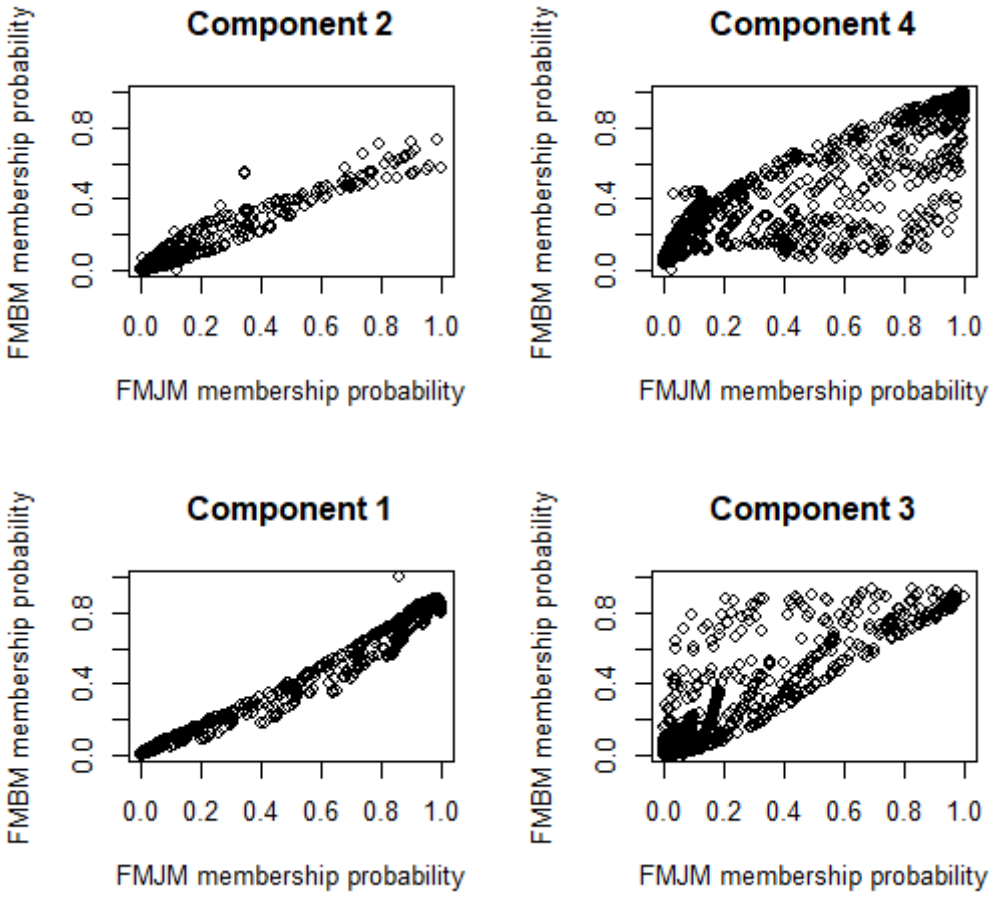
Appendix 4D: Sensitivity to Priors in the FMJM

Alternative priors for the variance parameters of the FMJM, including half.$N(0,10)$, $U(0,10000)$, $U(0,10)$, $IG(0.00001,0.00001)$ and $IG(0.1,0.1)$, are used to determine the robustness to the choice of the priors. The table below provides the posterior estimates of the parameters and model fits in terms of the Deviance Information Criteria (DIC) under these choices of priors. The results suggest that compared to using half.$N(0,10000)$, alternative choices of prior do not provide strikingly different estimates. Convergence issue arose under the $U(0,10000)$ prior.

	half.$N(0,10000)$				half.$N(0,10)$			
	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$	DIC	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$	DIC
γ_N	0.516	2.951	6.410		0.436	1.92	4.083	
γ_E	4.247	5.933	8.681		3.914	5.368	7.833	
σ_{bN}	0.077	0.136	0.247		0.094	0.182	0.274	
σ_{bE}	0.200	0.285	0.385	36470	0.216	0.308	0.411	28956
σ_{N1}	0.196	0.291	0.354		0.137	0.259	0.325	
σ_{E1}	0.718	0.772	0.827		0.711	0.767	0.823	
σ_2	0.782	1.001	1.158		0.883	1.032	1.186	
	$U(0,10000)$				$U(0,10)$			
	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$	DIC	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$	DIC
γ_N	0.898	2.019	4.38		0.213	2.54	6.901	
γ_E	2.411	7.711	47.254		4.145	5.885	9.923	
σ_{bN}	0.334	0.49	0.732		0.033	0.079	0.219	
σ_{bE}	5.627	11.539	20.449	30501	0.182	0.284	0.393	44121
σ_{N1}	0.261	0.502	0.637		0.181	0.308	0.37	
σ_{E1}	1.256	25.967	64.811		0.716	0.772	0.826	
σ_2	0.771	1.327	1.773		0.786	1.031	1.183	
	$IG(0.00001,0.00001)$				$IG(0.1,0.1)$			
	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$	DIC	$Q_{.025}$	$Q_{.500}$	$Q_{.975}$	DIC
γ_N	0.148	0.215	0.282		0.259	2.8	5.241	
γ_E	1.86	2.261	2.791		3.372	4.515	6.246	
σ_{bN}	0.447	0.704	0.856		0.25	0.35	0.462	
σ_{bE}	0.585	0.672	0.759	30082	0.341	0.43	0.536	30877
σ_{N1}	0.264	0.38	0.499		0.247	0.365	0.457	
σ_{E1}	0.264	0.38	0.499		0.619	0.696	0.771	
σ_2	0.911	1.067	1.228		0.594	0.946	1.112	

Appendix 4E: Comparison of Estimated Component Membership Probabilities from FMJM and FMBM

The estimated component membership probabilities from FMJM (x-axis) and FMBM (y-axis) are plotted for each component $j = 1, \ldots, 4$. The correlations are estimated as 0.99, 0.97, 0.90, and 0.90 for each of the components, respectively.



Appendix 4F: Additional Covariates Not Discussed in Detail in Section 3

For completeness, the plots below provide the estimated transformed component membership probabilities by the covariates not discussed in detail in section 3. There are no obvious trends in the posterior probability estimates by slope, elevation, ground attack size, BUI intercept, BUI slope, and ground attack size.

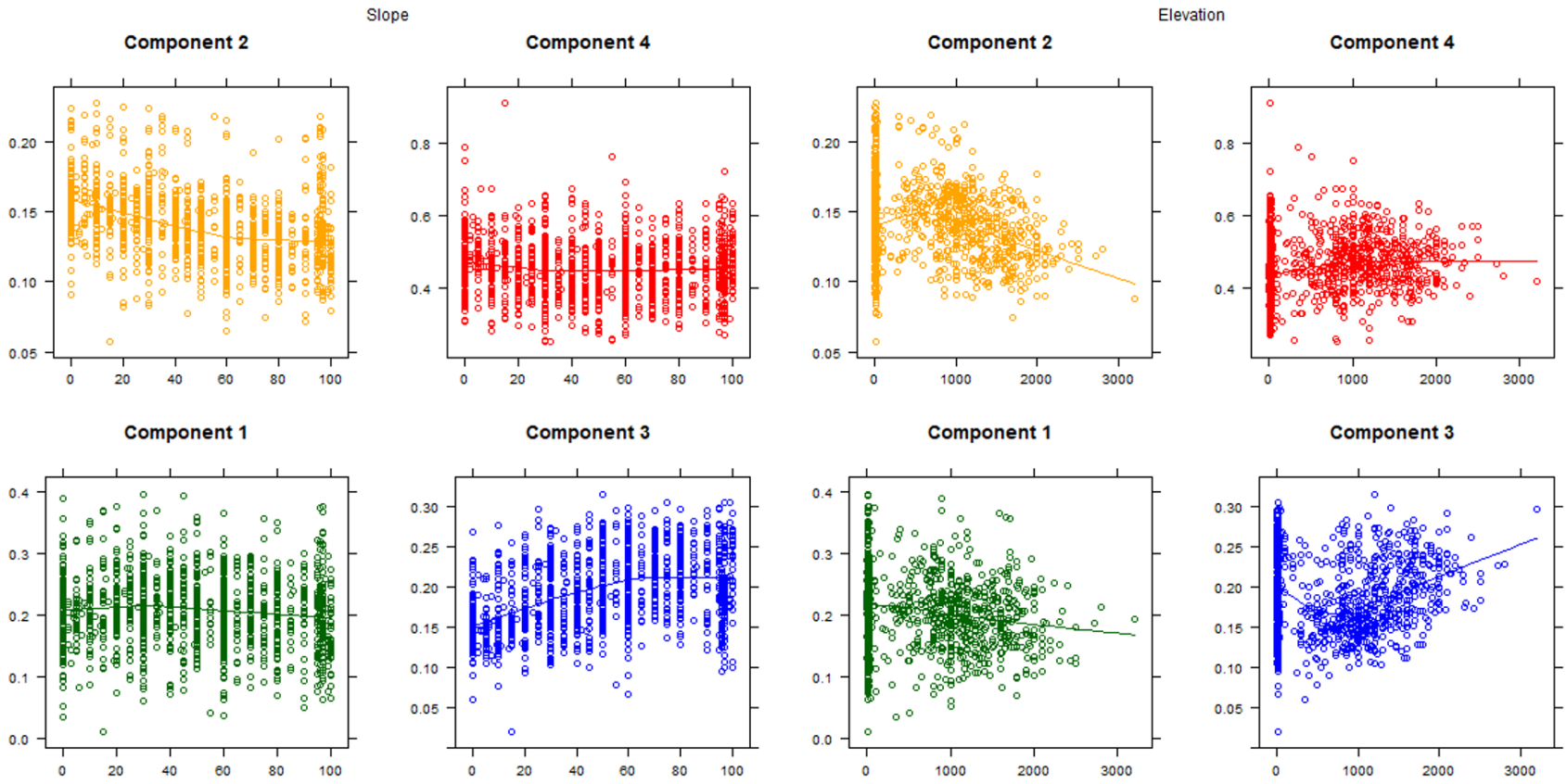


Figure B.6: Posterior estimates of component membership by slope and elevation

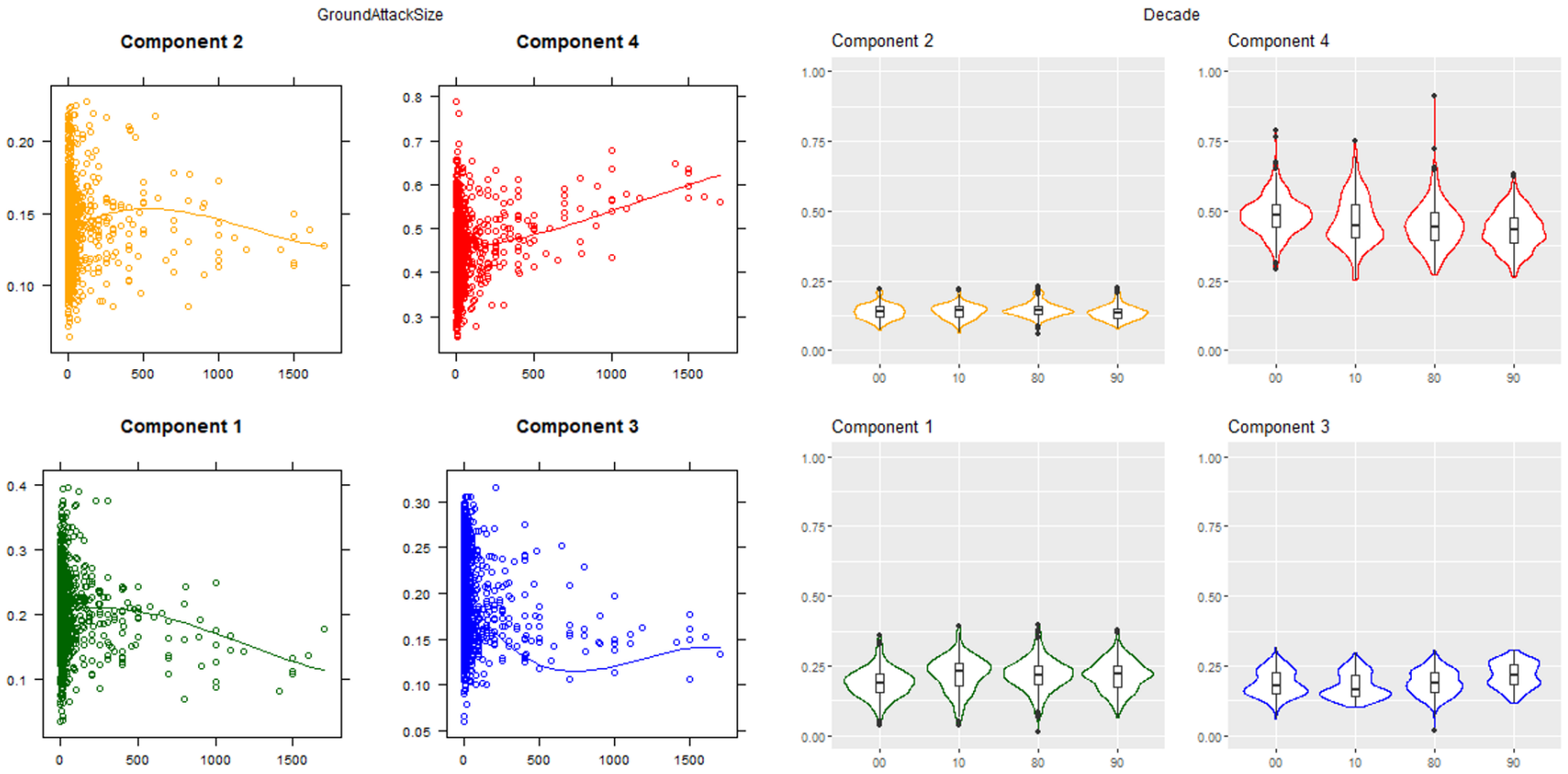


Figure B.7: Posterior estimates of component membership by ground attack size and decade

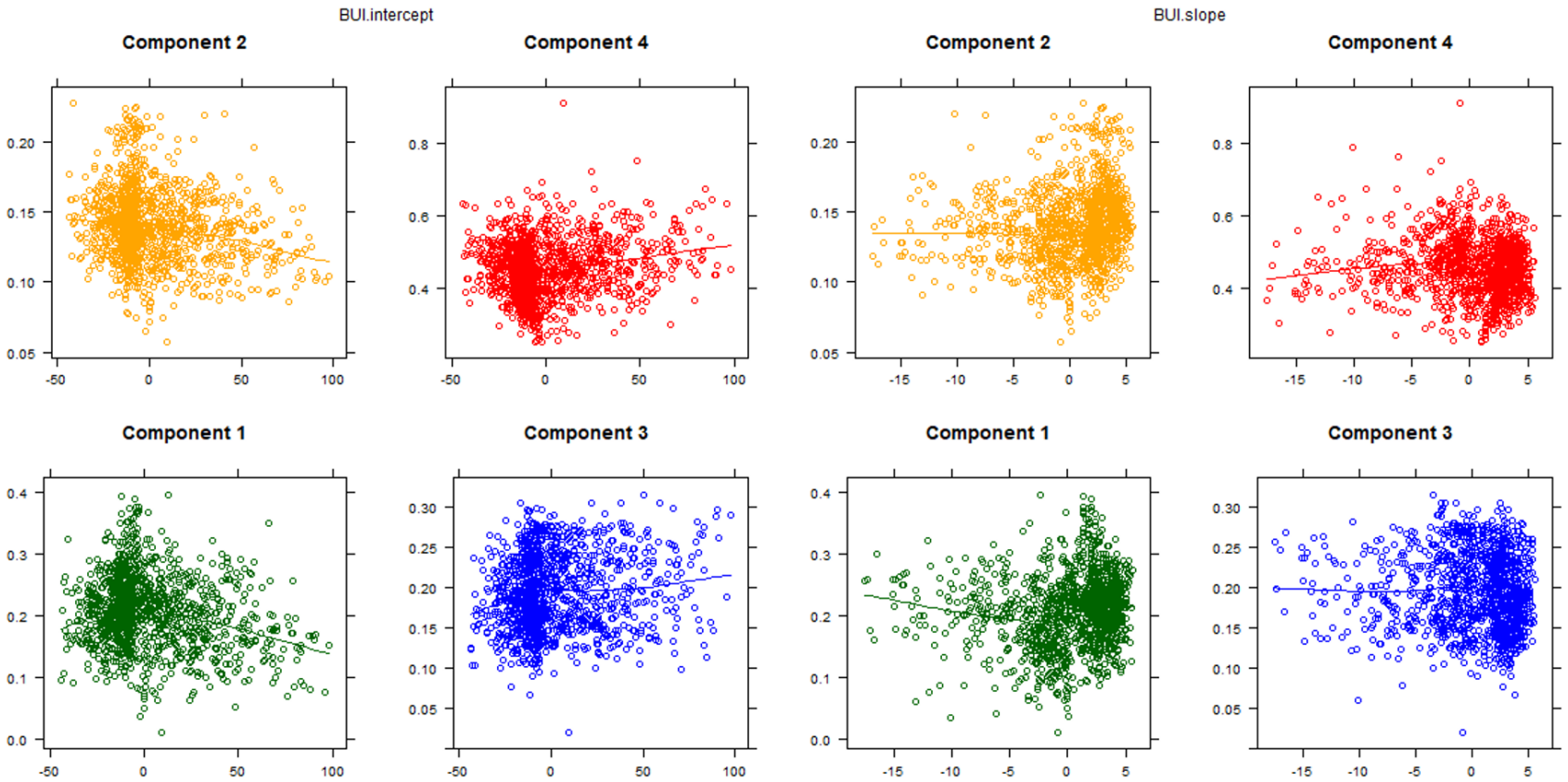


Figure B.8: Posterior estimates of component membership by BUI intercept and BUI slope

Curriculum Vitae

Dexen Da Zhong Xi

Education

2014–2020	**Ph.D., Statistics**, University of Western Ontario
2013–2014	**M.Sc., Statistics**, University of Western Ontario
2008–2013	**B.Sc., Statistics**, Simon Fraser University

Appointments

2014–2019	**Research Assistant**
	Wildland Fire Science Laboratory, University of Western Ontario
2014–2015	**Statistical Consultant**
	Western Data Science Solutions, University of Western Ontario
2014 Summer	**Online Course Designer**
	Department of Statistical and Actuarial Sciences, University of Western Ontario
2011 Fall	**Co-op position: Methodologist Assistant**
	Methodology Branch, Statistics Canada

Teaching Assistant

2015 Spring	SS3848, Introduction to Study Design	University of Western Ontario
2014 Fall	SS1024, Introduction to Statistics	University of Western Ontario
2014 Spring	SS2244, Statistics for Science	University of Western Ontario
2013 Fall	SS2244, Statistics for Science	University of Western Ontario

2011 Spring Statistics Workshop Simon Fraser University

Scholarships and Awards

2017 Queen Elizabeth II Graduate Scholarship in Science and Technology

2016 Queen Elizabeth II Graduate Scholarship in Science and Technology

2016 Statistical Society of Canada Student Poster Award, based on Ph.D. thesis

2016 SSC 2016 Student Travel Award

2013 Tuition Scholarship WGRS – STATISTICS

2011 Native Education College Tutor Certificate

2008 Summit Entrance Scholarship

Publications

Journals

[1] **Xi, D. D. Z.**, Taylor, S. W., Woolford, D. G., & Dean, C. B. (2019). Statistical Models of Key Components of Wildfire Risk. *Annual Review of Statistics and Its Application*, 6(1), 197–222. https://doi.org/10.1146/annurev-statistics-031017-100450

[2] **Xi D. D. Z.**, Dean, C.B., & Taylor, S.W. (2020). Modeling the duration and size of extended attack wildfires as dependent outcomes. *Environmetrics*. 31(e2619). https://doi.org/10.1002/env.2619

[3] **Xi, D. D. Z.**, Dean, C. B., & Taylor, S. W. *Modeling the Duration and Size of Wildfires Using Joint Mixture Models*. Submitted for publication.

[4] **Xi, D. D. Z.**, Dean, C. B., & Renouf, E. *Joint Modeling of Hospitalization and Mortality of Ontario Covid-19 cases*. In preparation.

Technical reports

[1] **Xi, D. D. Z.** (2011). Methodology Research and G-Confid Development: Research on the Shuttle Algorithm in the macro AUDIT of G-Confid. Statistics Canada internal document. Retrieved from http://www.statcan.gc.ca/pub/12-206-x/2012000/papers- documents-eng.htm

Conference Presentations

Invited Presentations

[1] **Xi, D. D. Z.**, Taylor, S. W., & Dean, C. B. (2016). *Dependent Models for the Duration and Size of BC Fires.* Invited talk presented at the International Environmetrics Society 2018, Guanajuato, Mexico, Jul. 16-21 http://ties2018.eventos.cimat.mx/

[2] **Xi, D. D. Z.**, Taylor, S. W., & Dean, C. B. (2016). *Joint Models for the Duration and Size of BC Forest Fires.* Poster session presented at the Population Models in the 21st Century, The Mathematical Biosciences Institute, Columbus, Ohio.

Contributed Presentations

[1] **Xi, D. D. Z.**, Taylor, S. W., & Dean C. B. (2018). *Joint Models for the Duration and Size of BC Fires*. Contributed talk presented at the Statistical Society of Canada Annual Meeting 2018, McGill University, Montreal, Québec.

[2] **Xi, D. D. Z.**, Taylor, S. W., & Dean C. B. (2017). *Dependent Models for the Duration and Size of BC Fires*. Contributed talk presented at the Statistical Society of Canada Annual Meeting 2017, The University of Winnipeg, Winnipeg, Manitoba.

[3] **Xi, D. D. Z.**, Taylor, S. W., & Dean C. B. (2017). *Joint Models for the Duration and Size of BC Forest Fires.* Poster session presented at the 2017 Fallona Family Interdiscipinary Showcase, University of Western Ontario, London, Ontario.

[4] **Xi, D. D. Z.**, Taylor, S. W., & Dean C. B. (2016). *Joint Models for the Duration and Size of BC Forest Fires.* Poster session presented at the Statistical Society of Canada Annual Meeting 2016, St. Catharines, Ontario.

Professional Activities

- Workshop Presenter, Crash Course in Introductory Inferential Statistics, 2015 Summer
- Student Volunteer, UWO R Workshop, 2015 Spring

Student Volunteer,