Investigating Distributions of Epochs in Wildland Fire Lifetimes

By Xinlei Wang

Supervised by Dr. Douglas Woolford and Dr. Kevin Granville

August 15, 2022

Abstract

The objective of my research project is to explore the relationship between variables related to wildland fire and to model distributions of epochs in wildland fire lifetimes. Several distributional families are considered for modeling these epochs, including the exponential distribution, gamma distribution, Weibull distribution and continuous phase-type distribution. I explain each of these distributions in short terms and illustrate how they are fit. Visual results of my exploratory data analysis are illustrated in two parts, data visualization and data modeling, along with my interpretation of each. Since this work is preliminary, I conclude the report with a discussion on what I have learned from the overall research experience.

1. Introduction

1.1 Understanding Wildland Fire Science

The operational fire season in Ontario is defined to be the period from April 1st to October 31st (Government of Ontario, 1990). From 1997 to 2022, approximately 7,300 forest fires have ignited each year in Canada, consuming about 2.5 million hectares of total area burned on average (Government of Canada, 2022). Most fires are caused by lightning or as an unintended result of human activities, while others could be set intentionally for the purpose of renewing and maintaining a healthy ecosystem. It is also notable that not every fire needs to be suppressed, as many are just left to burn themselves out naturally if they are remote from valuable assets like human settlements and infrastructures (Government of Canada, 2020). In Ontario, the Ministry of Natural Resources and Forestry (MNRF) is responsible for fire management decision making through the evaluation of various factors with the support of the Canadian Forest Fire Danger Rating System (Stocks et al., 1989). Wildland fire science is essential for people to understand the causes, risks, and potential benefits of wildland fires and helps inform decision making by fire managers concerning prevention and mitigation.

1.2 Research Objectives

The objective of my research project was to explore statistical methods to model the distributions of several epochs (i.e., periods of time) in wildland fire lifetimes. I investigated the relationships between several variables of interest and distributions of time lags (response lag, report lag, etc.) for wildland fires ignited in a district of Ontario, Canada. Through this investigation, I was able to see how different variables impacted the variation of multiple time lags.

1.3 Outline

The rest of my report is structured as follows. In Section 2, I discuss the data set, who has provided us with this data (and our data sharing agreement), as well as how it's adjusted for use within this report for the sake of confidentiality. Section 3 introduces different distributions that I used to fit the data, including common distributions and more flexible approaches. Then, Section 4 shows the results and graphs from my exploratory data analysis, which are separated into two subsections, data visualization and fitted distributions. Finally, I conclude with some comments on my work, discuss some avenues for future work, and end with my acknowledgments.

2. Data

The data used for my USRI research is copyright. It was provided by the Ontario Ministry of Natural Resources and Forestry and is used under the terms of their Electronic Intellectual Property Licence facilitated through a Collaborative Research Agreement between that MNRF and the University of Western Ontario. I also signed a data sharing agreement to gain access to this data. For the purposes of this report, pseudo data was used to protect the confidential nature of the data provided by the MNRF. Such pseudo data was generated in such a way that the results of this report are representative of what was found over the course of our research without revealing specifics of the actual raw data provided by the MNRF.

3. Methods

3.1 Considered Continuous Distributions

To start, several common continuous distributions were considered for modeling distributions of fire lifetime epochs, including exponential, gamma, and Weibull distributions. Following these, I investigated a more flexible approach to better fit to the data, which was the continuous phase-type distribution. In this section, some basic notations of these distributions are presented.

3.1.1 Gamma Distribution

The gamma distribution is a continuous probability distribution that can be used to model variables that have distributions that are positive and skewed (Sengupta, 2020). It is specified by two parameters, the shape parameter $\alpha > 0$ and the rate parameter $\beta > 0$. The corresponding probability density function (pdf) is

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-\beta x} \beta^{\alpha}}{\Gamma(\alpha)}\ , \quad for\ x > 0,$$

where $\Gamma(\alpha)$ is the gamma function satisfying

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx\ .$$

Moreover, if the shape parameter is a positive integer, the gamma distribution becomes an Erlang distribution, where the shape parameter represents the number of sequential events we are waiting to complete, each having independent and identically distributed exponential distribution with rate β. The Erlang distribution can be used to model the waiting time in the process, and was used to assist in estimating the components of continuous phase-type distributions introduced later in the report.

I used the “fit distribution” function in the “MASS” package of RStudio and the maximum likelihood estimation (MLE) method to estimate the shape and the rate parameters. Then, using these values, it was possible to calculate the theoretical density for each data value if the data was assumed to follow gamma distribution.

3.1.2 Exponential Distribution

The exponential distribution is a special case of a gamma distribution with the shape parameter α being equal to 1. In that circumstance, exponential distribution only has one parameter, the rate parameter, denoted here by λ. The corresponding pdf is

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad for\ x \geq 0.$$

The rate parameter tells us how quickly decay of the exponential function occurs (Glen, 2016). Exponential distributions are often used to model the time intervals between independent events (Frost, 2021).

To fit these distributions, I used the reciprocal of the data’s mean as the theoretical rate parameter since the MLE of the exponential distribution’s rate parameter is the inverse of the sample mean.

3.1.3 Weibull Distribution

The Weibull distribution is also a continuous probability distribution that can model both left and right skewed positive data with a variety of distribution shapes. In addition, it can approximate other distributions due to its flexibility and is frequently applied in life data, reliability analysis, capability analysis and so on (Frost, 2021). The probability density function of Weibull distribution is

$$f(x;\lambda,k) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1}e^{-(x/\lambda)^k} \quad , \qquad for\ x \geq 0,$$

which is composed of two parameters, the shape parameter $k > 0$ and the scale parameter $\lambda > 0$. An alternative third parameter, the location parameter, may also be defined but can be omitted and set to zero.

To fit Weibull distributions, I used the "eweibull" function in the "EnvStats" package of RStudio that can estimate the shape and scale parameters using a MLE approach.

3.1.4 Continuous Phase-Type Distribution

To fit more flexible parametric models to non-negative distributions, continuous phase-type distributions (CPH distributions) are often used in system evaluation and modeling (Reinecke et al., 2012). A CPH random variable represents the time until absorption in a continuous-time Markov chain (CTMC) having at least one absorbing state, with each transient state corresponding to one phase of the CPH distribution. In general, the distribution has two components, the initial probability vector and the transition rate matrix. Most importantly, it can be used to approximate any non-negative-valued distribution given a sufficiently high number of phases (Asmussen, 2000). More details about the definition of CTMCs and the properties of CPH distributions can be found in the book *Fundamentals of Matrix-Analytic Methods* by He (2014).

I used the Expectation-Maximization (EM) algorithm to estimate the components of CPH distributions to observed data (Asmussen et al., 1996). The EM algorithm is an iterative method that performs maximum likelihood estimation and concerns the presence of missing or latent data (Brownlee, 2019).

4. Results

4.1 Data Visualization

Data visualization is a crucial part of the research process during which information in data is translated into visual elements like graphs. All the visualizations in this research are done with the application of ggplot2 found in the set of packages “tidyverse” in RStudio.

There are several epochs of interest during the lifetime of a fire, such as durations between the ignition of fires and their discovery time, the report time and the attack time, the time fires were under control and the time they were declared out, and so on. My investigation started with one epoch of interest (Epoch A) and explored its relationship with independent variables of interest. As our fire data spanned the years from 1990 to 2019, it was convenient to group data into their respective decades to investigate changes in distribution over time.

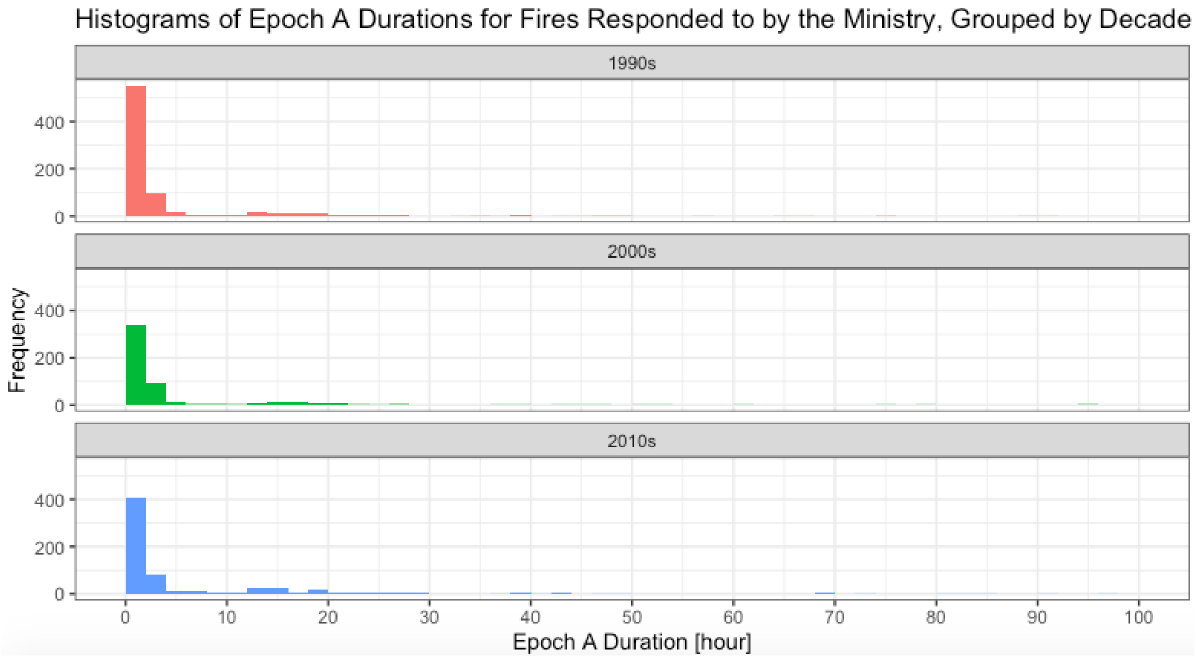

Figure 1. Histograms of observed time epoch values from 1990 to 2019, grouped by decade. As the majority of the values are concentrated on the far left of the histograms, I used the "coord_cartesian" function in the package of ggplot2 in RStudio to adjust the range of the x-axis so that the left part was zoomed in for more precise visualization.

Using information in the data concerning what month a fire was ignited in, months are grouped into quarters of a year and fires were grouped based on these quarters. In Figure 2, it can be noticed that there does not exist a first quarter, which corresponds to January, February, and March. This is due to the fact that I restrict my analysis to the operational fire season which is from the first of April to the end of October (Government of Ontario, 1990). Consequently, there are notably fewer fires in quarter four, due to the omission of November and December. In addition, during July and August, the frequency of fires is higher and responses tend to be slower than in quarter two, as the proportion of durations that are greater than 2 hours in quarter two is 28.2%

while in quarter three it is 36.6%, which could be the result of more fires competing for a limited number of suppression resources.

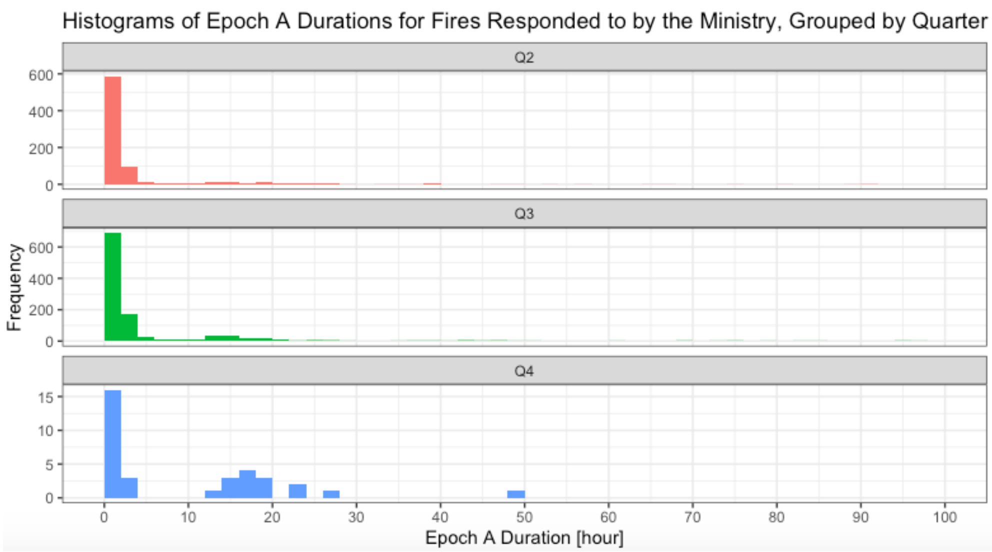

Figure 2. Histograms of observed time epoch values from 1990 to 2019, grouped by quarter.

4.2 Fitted Distributions

4.2.1 Gamma Distribution

In the perspective of spatial variables, I chose to investigate on the relationship between the epoch values and the distance from the ignition location to the nearest road (referred to as "distance from road").

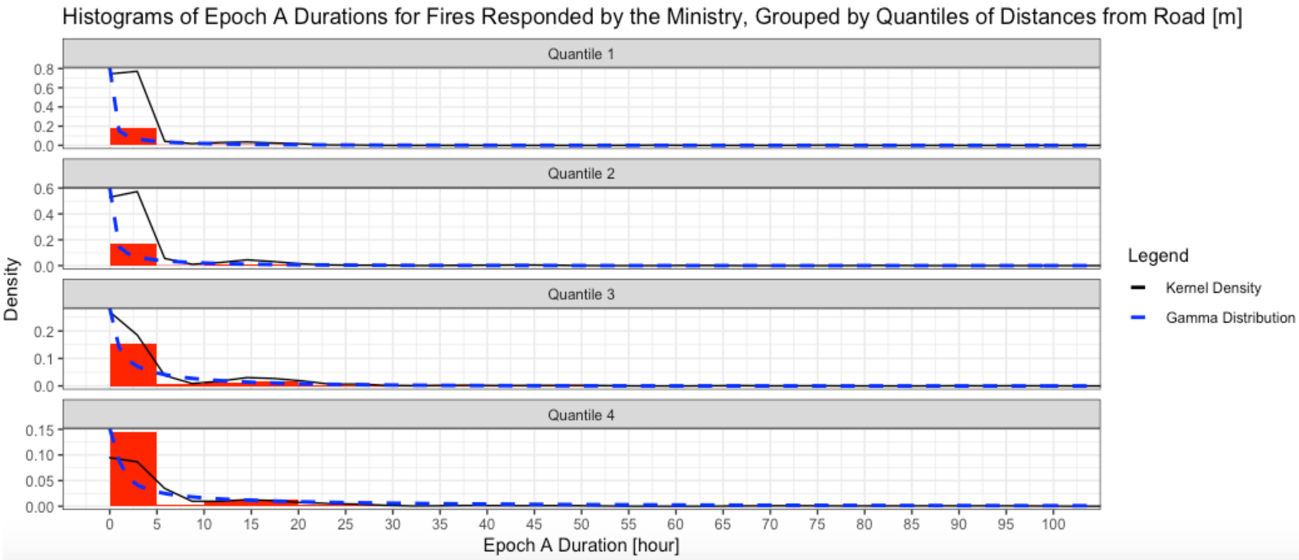

Figure 3. Histograms of observed epoch values grouped by quantiles of distances from road, with fitted Kernel and gamma distributions densities. The solid black lines correspond to Kernel density estimates and the blue dashed lines correspond to gamma distributions.

For comparison, I stratified the epoch durations with respect to quantiles of distance from the road, meaning that the distances were arranged from the minimum to the maximum, so that "Quantile 1" corresponded to the first 25% of the distances and so on. For each group, I first used Kernel density estimation to model the distributions of response lag. Next, I fit gamma distributions to compare with the Kernel densities. It is noticeable that some of the black lines have a small bump after the decay of the highest peak, making the distribution bimodal, which cannot be replicated by the gamma distribution.

4.2.2 Weibull and Exponential Distributions

Other parametric distributional modeling approaches were also attempted. This time another epoch was investigated (Epoch B). Fine Fuel Moisture Code (FFMC) is one of the six components of the

Canadian Forest Fire Weather Index System that is useful to measure forest fire danger in Canada (Van Wagner, 1987). Here, the epoch was divided by four FFMC operational risk categories, ranging from "low" to "extreme", and by two different causes of fire — human-caused and lightning-caused (i.e., "Natural"). Next, Weibull distributions were fit to human-caused fires while exponential distributions were fit to lightning fires. In this case, the Kernel density estimates and the fitted distributions roughly agreed with each other.

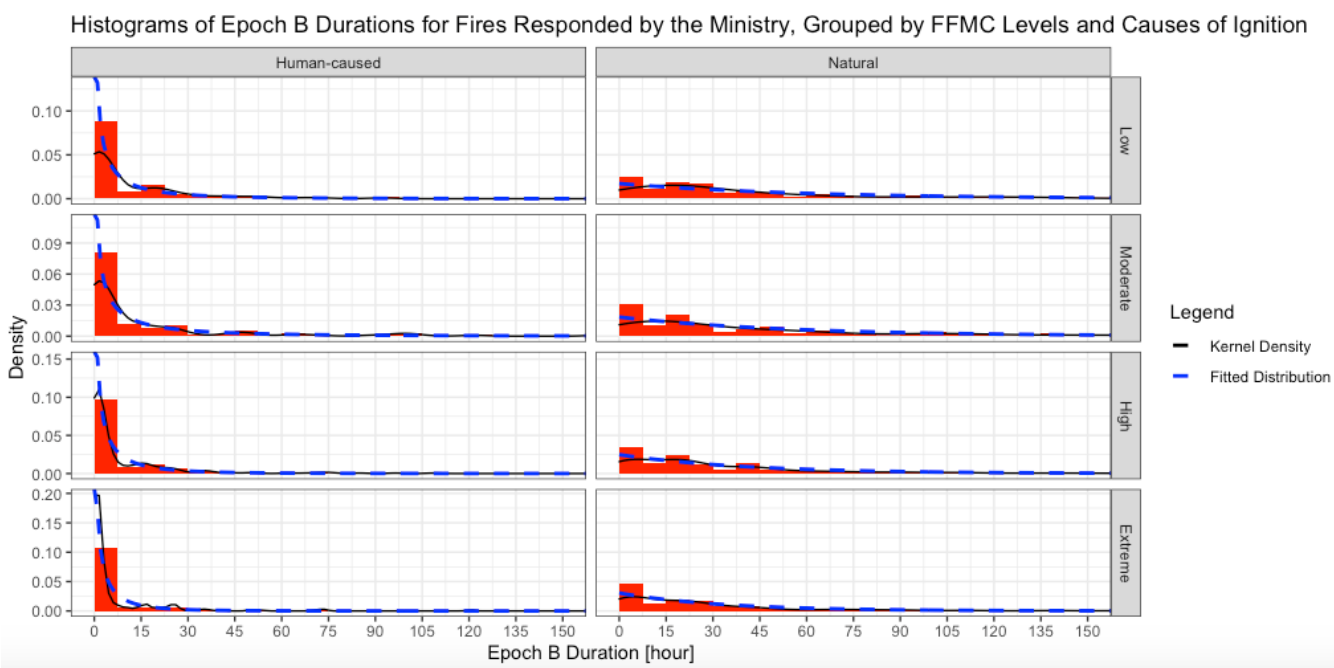

Figure 4. Histograms of observed epoch values grouped by FFMC risk classifications and two causes of ignition, with either fitted Weibull or exponential distributions. The solid black lines correspond to Kernel density estimates and the blue dashed lines correspond to parametric distributions.

4.2.3 Continuous Phase-Type Distribution

An important consideration when using the EM algorithm to fit CPH distributions is the choice of initial guesses for the distribution's components. As observed in Section 4.2.1, we desire the ability to fit both the high peak and the little bump observed in the Kernel density. Therefore, I divided the density line into two parts: the first part with the high peak and the latter part consisting of the bump. Next, two continuous CPH distributions were chosen to roughly approximate each part with separate probability vectors and rate matrices. For the high peak, I estimated that it had one phase, which simplifies to an exponential distribution. For the small bump, I applied the knowledge of the Erlang distribution whose shape parameter represents the number of phases in the CPH distribution. Since adding more phases would increase the accuracy but lead to a longer runtime, I tried to use 12 phases. The mixture of two CPH distributions is also CPH so I was able to combine these into a single distribution whose components could be used as the initial guesses for the EM algorithm. Below is the result after thirty runs.

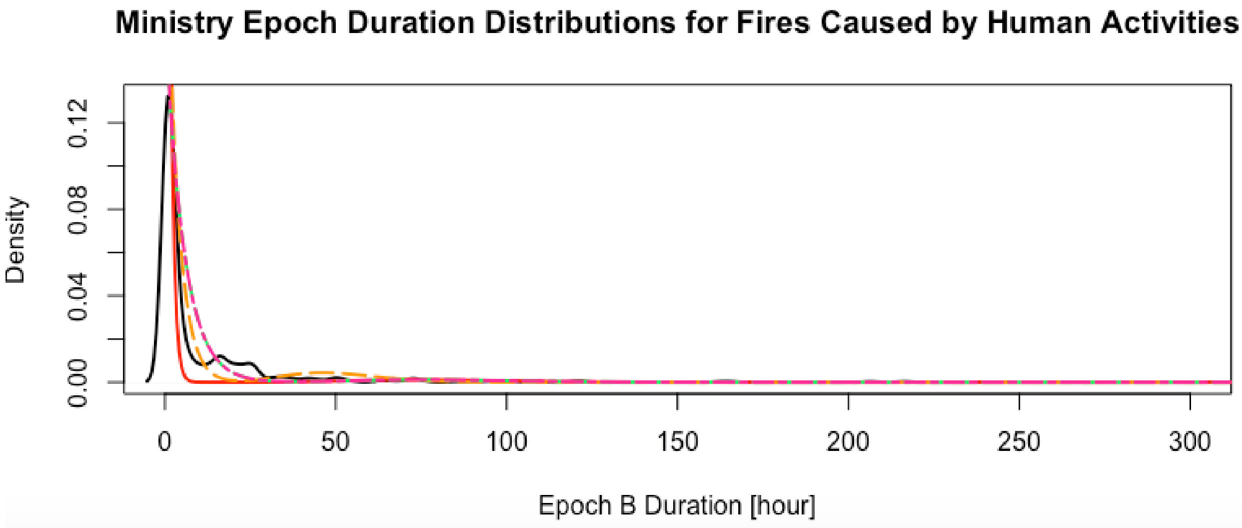

Figure 5. Plot of iteratively fit CPH distribution densities using the EM algorithm to values of Epoch B for human-caused fires. The Kernel density estimate is plotted in black. The red, yellow, green, and purple dashed lines correspond to the CPH density lines after 1, 5, 20, 30 iterations, respectively.

Based on the plot in Figure 5, I infer that it is possible that adding more phases and running more iterations of the EM algorithm would be needed to narrow the gap between the densities at the small bump.

5. Conclusion and Future Work

As my research progressed, I got to know a lot of terms used in wildland fire science. Wildland fire science truly plays an important role in helping the MNRF to understand past fires and develop systems and tools to aid in managing future wildland fires. Through the exploratory data analysis, I was able to learn and apply both parametric (common distribution families) and non-parametric (Kernel density) models to different subsets of data. Although some of the parametric models failed to capture certain features of the data (e.g., bimodality), I learned what variables could influence the MNRF's response time, whether to attack by ground, by air or both, what specific attack tools the MNRF used, what could influence their decision making, and much more. In the future, I look forward to applying the modeling methods and research skills I learned to more social or environmental issues, including further research on wildland fire science.

Acknowledgements

I would like to express my special thanks of gratitude to my supportive co-supervisors Dr. Douglas Woolford and Dr. Kevin Granville who were willing to supervise me after our very first meeting, introduced me to the brand-new field of wildland fire science, added me to the Wildland Fire Science Laboratory and guided me throughout the program. The knowledge of wildland fire science, statistical methods and research skills they taught me are extremely valuable for both my future research and career paths. I would also like to thank all the senior fire lab members for being willing to share their knowledge and experience, listening to my presentations and giving me useful advices.

Lastly, I would like to thank USRI program coordinators for giving me the incredible opportunity to conduct a research project during my undergraduate period and for providing me with invaluable professional development sessions.

Bibliography

Asmussen, S. (2000). *Matrix-analytic models and their analysis*. Scandinavian Journal of Statistics, vol. 27, no. 2, pp. 193-226.

Asmussen, S., Nerman, O., & Olsson, M. (1996). *Fitting phase-type distributions via the EM algorithm*. Blackwell Publishing on behalf of Board of the Foundation of the Scandinavian Journal of Statistics, vol. 23, no. 4, pp. 419-441.

Brownlee, J. (2019). *A gentle introduction to expectation-maximization (EM algorithm)*. Machine Learning Mastery. Retrieved August 10, 2022, from https://machinelearningmastery.com/expectation-maximization-em-algorithm/

Frost, J. (2021). *Exponential distribution: uses, parameters & examples.* Statistics by Jim. Retrieved August 9, 2022, from https://statisticsbyjim.com/probability/exponential-distribution/

Frost, J. (2021). *Weibull distribution uses, parameter & examples*. Statistics by Jim. Retrieved August 9, 2022, from https://statisticsbyjim.com/probability/weibull-distribution/

Glen, S. (2016). *Rate Parameter*. Statistics How To. Retrieved August 13, 2022, from https://www.statisticshowto.com/rate-parameter/

Government of Canada. (2020). *Forest fire danger rating tool*. Retrieved August 8, 2022, from https://www.nrcan.gc.ca/our-natural-resources/forests-forestry/wildland-fires-insects-disturban/forest-fire-danger-rating-tool/14470

Government of Canada. (2022). *Forest fires*. Retrieved August 8, 2022, from https://www.nrcan.gc.ca/our-natural-resources/forests/wildland-fires-insects-disturbances/forest-fires/13143

Government of Ontario. (1990). *Forest Fires Prevention Act, Revised Statutes of Ontario. c. F.24*. Retrieved August 13, 2022, from https://www.ontario.ca/laws/statute/90f24

He, Q.M. (2014). *Fundamentals of matrix-analytic methods*. New York: Springer, vol. 365.

Reinecke, P., Bodrop, L., Danilkina, A. (2012). *Phase-type distributions.* Resilience Assessment and Evaluation of Computing Systems. Springer, Berlin, Heidelberg, pp. 85-113. https://doi.org/10.1007/978-3-642-29032-9_5

Sengupta, S. (2020). *Gamma distribution explained: what is gamma distribution*. Great Learning. Retrieved August 9, 2022, from https://www.mygreatlearning.com/blog/gamma-distribution/

Stocks, B.J., Lawson, B.D., Alexander, M.E., Van Wagner, C.E., McAlpine, R.S., Lynham, T.J., Dubé, D.E. (1989). *Canadian Forest Fire Danger Rating System: An Overview.* The Forestry Chronicle, vol. 65, no. 4, pp. 258-265.

Van Wagner, C.E. (1987). *Development and structure of the Canadian Forest Fire Weather Index System*. Canadian Forest Service, Forest Technical Report 35. Ottawa, ON, Canada.