Exploring Human-Caused Fire Occurrence Prediction

Name: Ruyi Jin

Supervised by Dr. Douglas Woolford & Dr. Kevin Granville

Date: 2022.08.15

Introduction

One aspect of Wildland Fire Science concerns predicting forest fire occurrence. In detail, it investigates factors that can contribute to heightened wildland fire danger to predict when they are more likely to ignite. Wildland fire ignition sources can be separated into two groups, human-caused (e.g., campfires, industry fires, railway fires, etc.) and natural-caused (e.g., lighting). In this research, we investigated human-caused fire ignitions in a region of Ontario. Our study included exploratory data analysis of historical wildland fire records and developing fire occurrence prediction models.

Some important fire attributes in our data set include variables from the Canadian Forest Fire Weather Index (FWI) System (Van Wagner, 1987). Natural Resources Canada (2022) explains the FWI System as having "six components that account for the effects of fuel moisture and weather conditions on fire behavior." In this study, we mainly focus on the Fine Fuel Moisture Code (FFMC), which is defined as "a numeric rating of the moisture content of litter and other cured fine fuels." (Natural Resources Canada, 2022) It is an indicator of surface-level fuel flammability and relative ease of ignition.

Data

The data sets we used in our study are historical fire records and weather station data that were observed over a 30-year period for a region of within the Province of Ontario. These data were provided by the Ontario Ministry of Natural Resources and Forestry (MNRF). These data are copyright and are used under the terms of their Electronic Intellectual Property Licence facilitated through a Collective Research Agreement between the MNRF and the University of Western Ontario.

In the historical fire archive, several attributes describing fires were available including the date and time of a fire's ignition, the location of its ignition, what fuel type was the fire ignited in, and which type of human-caused action resulted in the fire. FWI System variables for an ignition's time and location are also included. The weather station data set combines once daily observations collected from different weather stations at 1 pm local time, including the weather variables of relative humidity, temperature, wind speed, and rain, which are accompanied by the FWI System variables.

Due to the confidential nature of those data, I am unable to freely share the data or my exact results in this report. Therefore, pseudo data was generated in such a way that the displayed results are representative of the results discovered throughout the course of our research during this program without revealing specifics of the raw data provided by the MNRF.

Methods

For modeling fire occurrence, we employ Generalized Linear Models (GLMs) (e.g., Hosmer et al., 2013). These models are an extension to linear regression models, which allow for the response variables to operate under different distributional assumptions (e.g., Logistic, Poisson, Negative Binomial). Within these models, we also test the importance and influence of other predictor covariates (e.g., in which month a fire ignites).

This analysis is restricted to one specific district in Ontario which a high frequency of human-caused fire ignitions over the course of the study period (we refer to this region as “District A”). The main predictor of daily fire danger used in occurrence modeling is the average FFMC value across weather stations in this district. If on a given day, there is no recorded FFMC value at any of these weather stations, then we omit it from this analysis, regardless of whether one or more fires ignited on that day.

Another covariate used in our model is “month”, which can be extracted from a fire’s time of ignition. In our models, because weather station data tends to have no records for months outside of the fire seasons (operationally defined by the Forest Fires Prevention Act to be April 1 through October 31 in Ontario, Government of Ontario, 1990), we restrict the data only includes fire season months, from April to October.

The response variable of interest is daily fire counts which are calculated using the historical fire archive. After filtering, only human-caused fires ignited in District A are kept. Then, after extracting the ignition dates from each fire, we compute how many fires were ignited on each date which become the counts on those days. Days without any fire ignitions are assigned counts of 0.

Results

Visualizations

In the interest of preserving data confidentiality, we limit our focus to two of the most important results we found in our exploratory data analysis related to modeling. They mainly investigate the following two questions: When (in which month) do different types of human-caused fire ignitions happen, and at which levels of FFMC do different types of human-caused fires tend to ignite?

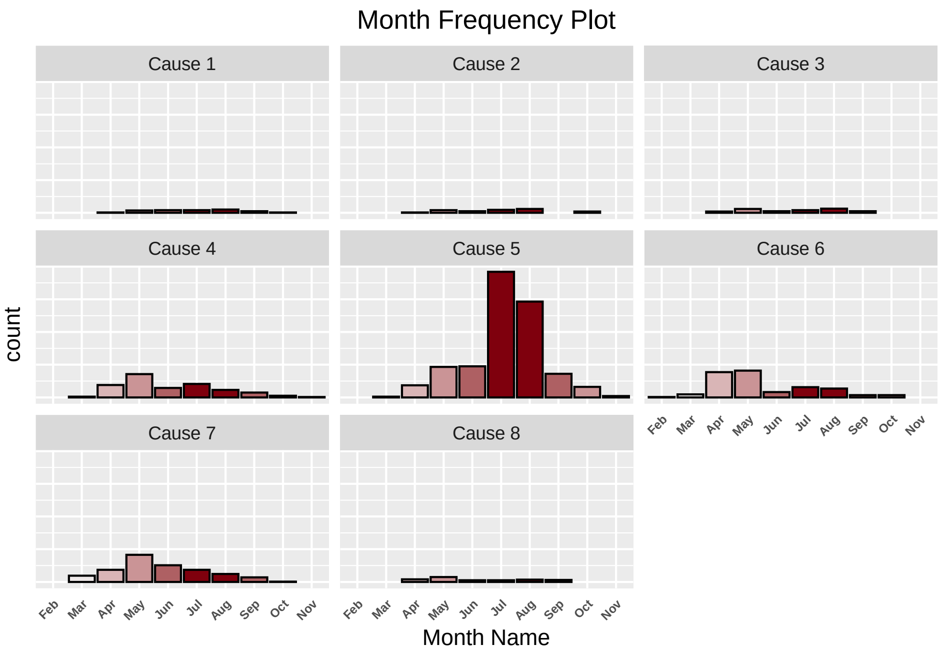

Figure 1 Histograms of monthly frequencies of human-caused fires grouped by type of cause.

Figure 1 above shows the relationship between month and ignition counts for different types of human-caused fires. The causes of ignition are censored, along with the scale of counts. To enable direct comparisons, a consistent y-axis scale is used for all plots. From this figure, Cause 4, Cause 6, Cause 7, and Cause 8 has their peak value in May but Cause 5 reaches its peak in July and August. So, these peak values happen most in summer seasons, but there is some variability depending on the types of human activities that result in these ignitions. This is a justification for us to consider month as a covariate in our models.

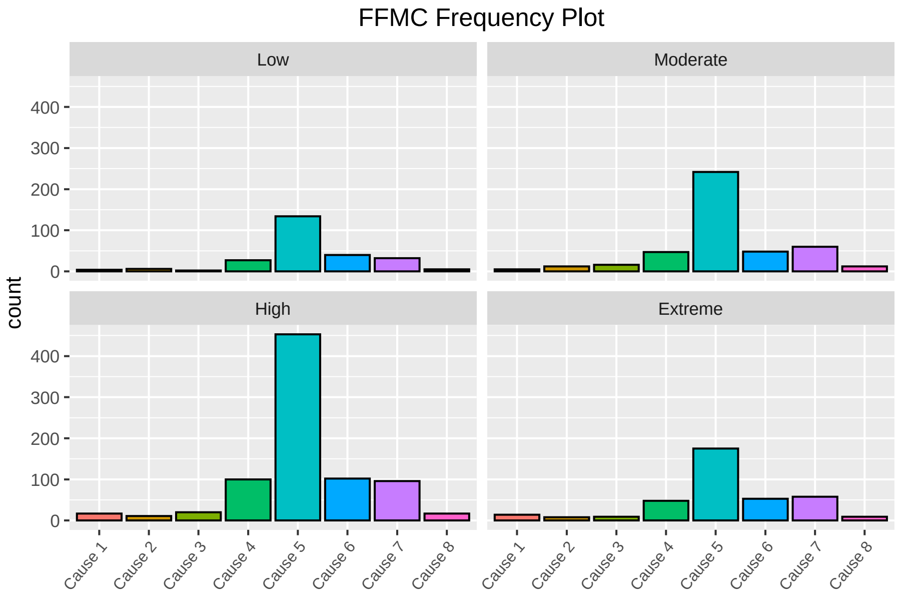

Figure 2 Histograms of frequencies of type of human-caused fires grouped by level of FFMC and cause.

Figure 2 compares the different types of human-caused fires with FFMC operational risk classifications. Here, FFMC values are classified into 4 levels: Low, Moderate, High, and Extreme, with numeric ranges as defined by Stocks (1971, 1974). The frequencies in these plots are only approximate values, as the FFMC values have been jittered through the addition of random noise. However, this approximation still roughly shows the relationships observed in the raw data. The recorded fires mostly happen when FFMC is at the High level. When the values of FFMC are increasing in level, the frequency of fires igniting also increases. Although the Extreme level has fewer counts than High level, this does not mean that the probability of a fire igniting when FFMC is at its Extreme level is less than at High. It just shows that in the fire records, we have fewer fires happen under Extreme FFMC conditions (e.g., since Extreme fire weather conditions are less common). When the FFMC values are in Extreme classification, there are still very high levels of fire danger.

Modeling Fire Occurrence

In our modeling, we began with a Poisson mode which is a commonly used framework for modelling counts. When the response variable we want to study is a count number, we may model the random process using the Poisson distribution. More details defining Poisson regression models can be found in the book "Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regression Models" written by Faraway in 2016. A similar modeling process by using a Poisson model to fit the average number of fires happening on each day depend on FFMC value can be found in Cunningham and Martell (1973).

After fitting the Poisson model, we obtain a linear function representing relationship between daily fire counts and average FFMC value. That is,

$$\hat{y} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1) = \exp(-8.424 + 0.098\, x_{FFMC\_AVG}) .$$

The table below showing the summary of the Poisson model we fit.

	Beta Coefficient	**Standard Error**	**z value**	**p-value**
Intercept	-8.424	0.203	-41.411	$< 2 \times 10^{-16}$
x_{FFMC_AVG}	0.098	0.002	41.651	$< 2 \times 10^{-16}$

Table 1 Summary of the fitted Poisson model.

After fitting the Poisson model, we detect overdispersion, with a dispersion ratio equal to 2.073. Saputro et al. (2021) explained that overdispersion occurs because "the presence greater variance of response variable caused by other variables unobserved heterogeneity". A Poisson model, which has only one parameter may be restrictive for empirical fitting purpose, can easily result in overdispersion.

When overdispersion occurs, it will influence the dependency of covariates to the response variable and may result in fitting a wrong model. To better deal with overdispersion, we consider the use of a more flexible Negative Binomial model.

In a Poisson model, the mean is equal to variance, E(Y) = Var(Y) = μ. This assumption places a strict restriction on the variance that may not be true in practice. In a Negative Binomial model, the mean is still equal to μ (i.e., E(Y) = μ), but variance is allowed to differ. Here, Var(Y) = $\mu + k\mu^2$, where $k \geq 0$ is usually referred as the dispersion parameter. This assumption gives more freedom to the model's fitted variance and can help deal with overdispersion. Therefore, we fit the Negative Binomial model and compare it against the Poisson model.

We obtain a new linear function after fitting the Negative Binomial model,

$$\hat{y} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1) = \exp(-8.506 + 0.099\, x_{FFMC\_AVG}) \ .$$

Table 2 shows a summary of this model. Both the beta coefficients and standard errors are changed relative to the Poisson model.

	Beta Coefficient	**Standard Error**	**z value**	**p-value**
Intercept	-8.506	0.258	-32.98	$< 2 \times 10^{-16}$
x_{FFMC_AVG}	0.099	0.003	32.39	$< 2 \times 10^{-16}$

Table 2 Summary of the fitted Negative Binomial model.

In Figure 3, we plot the differences between observed values and predicted values for each model. The red color represents differences between the real data and predicted value from a Negative Binomial model, while the blue color show difference between the observed value and predicted value from a Poisson model.

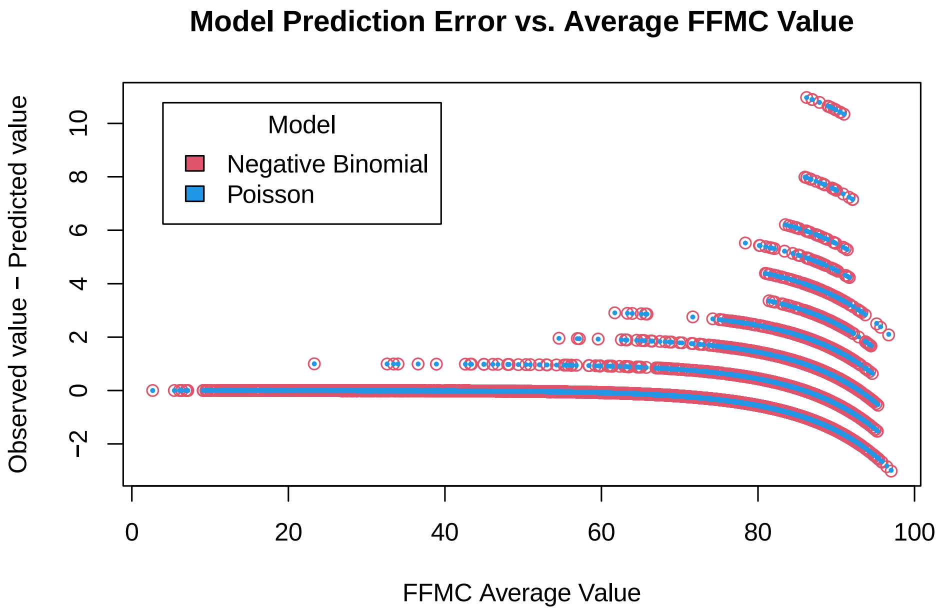

Figure 3 Scatter plot comparing differences between observed and predicted values for the fitted Poisson and Negative Binomial models.

By plotting a standardized residual plot, we find that the Negative Binomial model resulted in more condensed plots than the Poisson model.

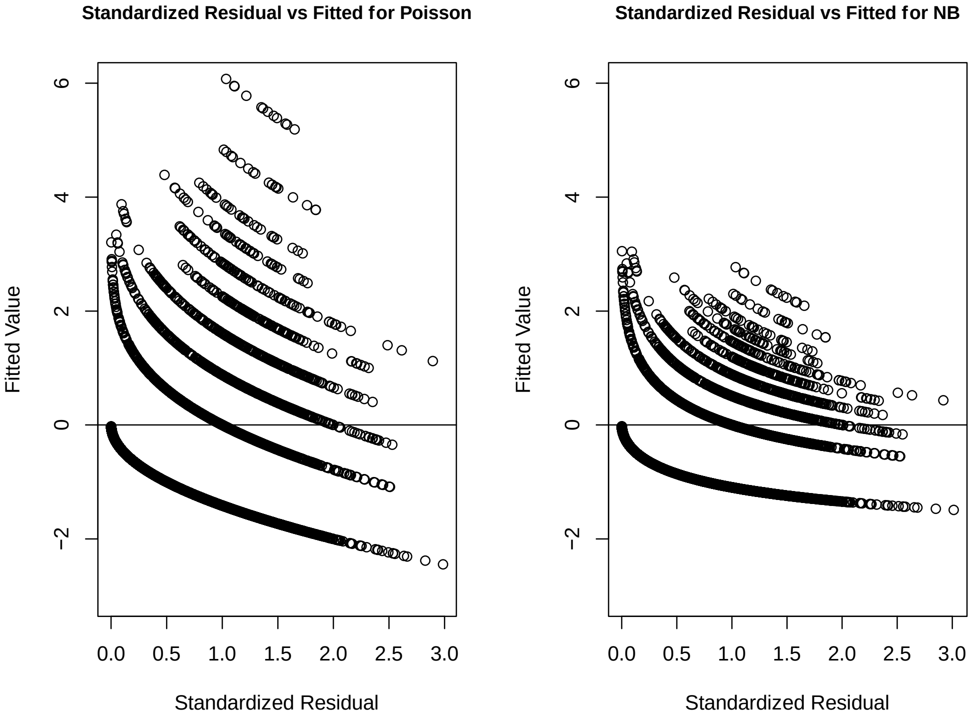

Figure 4 Standardized residual plot comparing Poisson and Negative Binomial models.

In order to compare predictive capabilities of both types of models, we perform a k-fold Cross Validation. The process of Cross Validation is defined as when we randomly separate the data into a training set and a validation set, with m and n-m observations respectively, then we fit our model to the training test and use the result to predict the response variable in validation set (Matloff, 2017). We use this to approximate the predictive ability of that model. In k-fold Cross Validation, we separate data into k groups and systematically alternate letting one group at a time be the training set while using the other k-1 groups as the validation set. In our data, we let $k = 10$ and randomly split 30 years of data into 10 groups of three years each.

Our metric of prediction error is Root Mean Square Error (RMSE), which is defined as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{N}}[\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}]^2}{\mathrm{N}}},$$

where N is the number of observations, $\mathrm{y_i}$ is the i-th measurement, and $\hat{\mathrm{y}}_\mathrm{i}$ is its corresponding prediction (C3.ai, 2021). Table 3 shows the resulting means and standard deviations RMSE values from the 10-fold Cross Validation.

10-Fold Cross Validation RMSE	Poisson Model	NB Model
Mean	1.2379	1.2380
Standard Deviation	0.3598	0.3597

Table 3 10-fold Cross Validation results contrasting Poisson and Negative Binomial models.

In addition to average FFMC, we consider models with and without the additional covariate “month”. Month is considered as a categorical variable, treating April as the baseline. A new function is obtained after we fit the Poisson model,

$$\begin{aligned}\hat{y} &= \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2\right) \\ &= \exp\,(-7.644 + 0.089\, x_{FFMC\_AVG} - 0.050\, I_{Month\_May} - 0.603\, I_{Month\_Jun} \\ &\qquad + 0.454\, I_{Month\_Jul} + 0.189\, I_{Month\_Aug} - 0.788\, I_{Month\_Sep} \\ &\qquad - 1.345\, I_{Month\_Oct}\,)\,,\end{aligned}$$

where the letter *I* represent the indicator function, such that I_{Month_Jun} will equal to 1 if the fire ignites in June and equals 0 otherwise. Table 4 shows the summary of this expanded model.

	Beta Coefficient	Standard Error	z value	p-value
Intercept	-7.644	0.214	-35.656	$< 2 \times 10^{-16}$
x_{FFMC_AVG}	0.089	0.002	37.118	$< 2 \times 10^{-16}$
Month_May	-0.050	0.059	-0.845	0.39835
Month_Jun	-0.603	0.069	-8.807	$< 2 \times 10^{-16}$
Month_Jul	0.454	0.053	8.507	$< 2 \times 10^{-16}$
Month_Aug	0.189	0.057	3.296	0.00098
Month_Sep	-0.788	0.083	-9.478	$< 2 \times 10^{-16}$
Month_Oct	-1.345	0.137	-9.856	$< 2 \times 10^{-16}$

Table 4 Summary of the fitted Poisson model incorporating month as a predictor.

Unfortunately, overdispersion is again discovered, now with a dispersion ratio equal to 1.969. Therefore, we again fit a Negative Binomial model to see if the model performance improves. The new function is

$$\begin{aligned}\hat{y} &= \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2\right) \\ &= \exp\,(-7.747 + 0.090\, x_{FFMC\_AVG} - 0.024\, I_{Month\_May} - 0.593\, I_{Month\_Jun} \\ &\qquad + 0.463\, I_{Month\_Jul} + 0.225\, I_{Month\_Aug} - 0.775\, I_{Month\_Sep} \\ &\qquad - 1.299\, I_{Month\_Oct}\,)\,.\end{aligned}$$

Table 5 presents the summary of this fitted Negative Binomial model.

	Beta Coefficient	Standard Error	z value	p-value
Intercept	-7.747	0.267	-28.972	$< 2 \times 10^{-16}$
x_{FFMC_AVG}	0.090	0.003	29.898	$< 2 \times 10^{-16}$
Month_May	-0.024	0.091	-0.258	0.7962
Month_Jun	-0.593	0.099	-6.000	1.98×10^{-9}
Month_Jul	0.463	0.086	5.411	6.26×10^{-8}
Month_Aug	0.225	0.089	2.522	0.0117
Month_Sep	-0.775	0.111	-6.978	3×10^{-12}
Month_Oct	-1.299	0.159	-8.184	2.75×10^{-16}

Table 5 Summary of the fitted Negative Binomial model incorporating month as a predictor.

In Figure 5, we again plot and compare the differences between the observed values and predicted values. The red color still represents differences between the real data and predicted value from a Negative Binomial model, while the blue color show difference between the observed value and predicted value from a Poisson model.

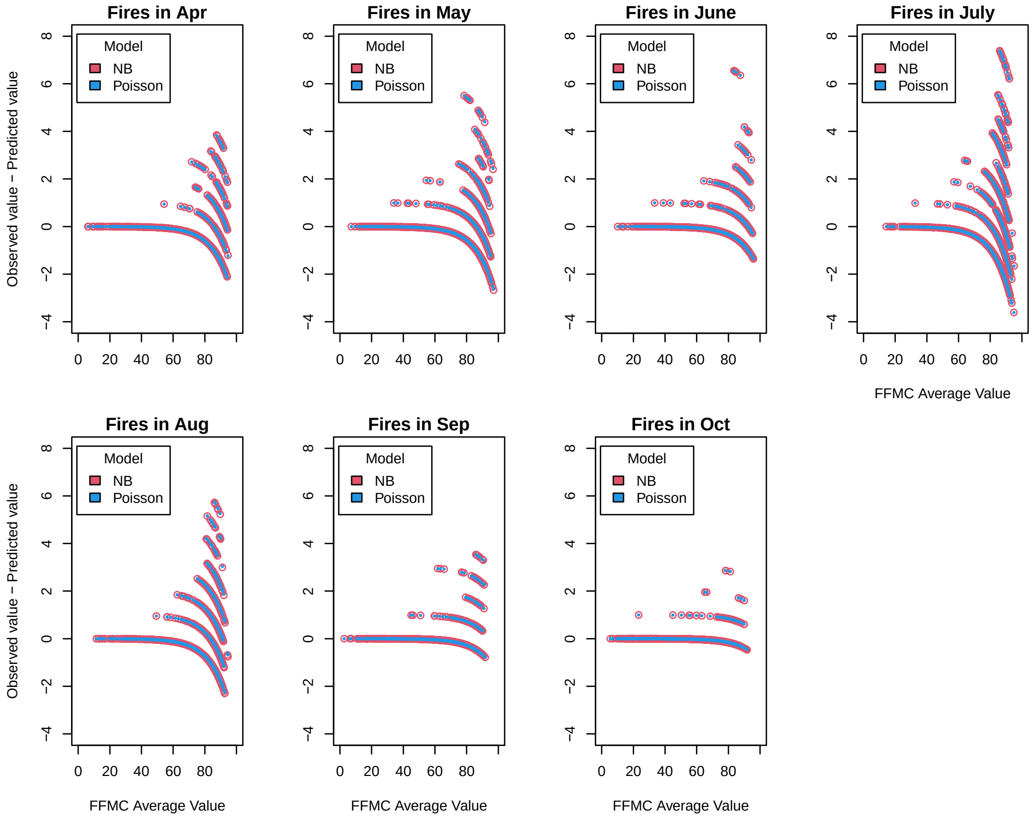

Figure 5 Scatter plots comparing differences between observed and predicted values for the fitted Poisson and Negative Binomial models incorporating month as a predictor.

Figure 6 is the standardized residual plot comparing Poisson and Negative Binomial model with month. Similarly, the Negative Binomial model resulted in more compacted plots than the Poisson model.

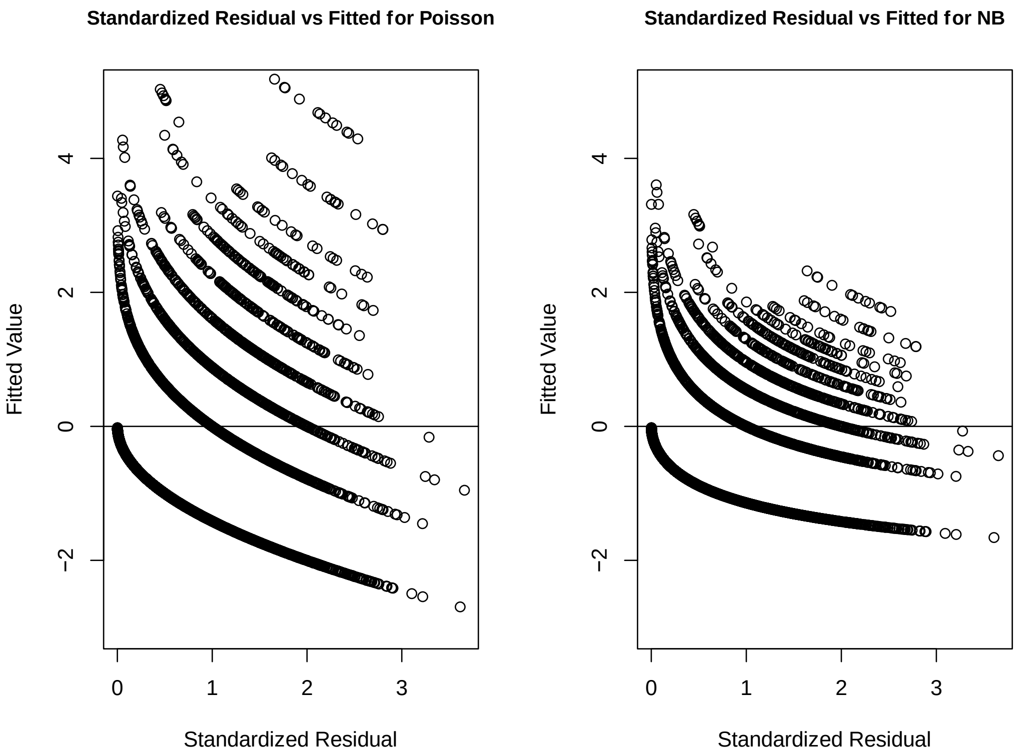

Figure 6 Standardized residual plot comparing Poisson and Negative Binomial models incorporating month as a predictor.

Table 6 summarizes the results of 10-fold Cross Validation using both expanded models. Similar to Table 3, all these results will be used later in the discussion of these models.

10-Fold Cross Validation RMSE	Poisson Model	NB Model
Mean	1.2120	1.2112
Standard Deviation	0.3326	0.3318

Table 6 10-fold Cross Validation results contrasting Poisson and Negative Binomial models incorporating month as a predictor.

Discussion

Poisson vs. Negative Binomial model performance

Investigating the difference in performance between the Poisson and the Negative Binomial Models, Table 1 and Table 2 show that they have very close beta coefficients, with a few changes in standard errors and z values. Figure 3 and Figure 4 comparing the differences between the predicted and observed values agree with this similarity, as they show the red and blue dots almost precisely overlapping each other. This indicates that the differences between the corresponding Negative Binomial and Poisson models are small.

Based on Figure 4 and Figure 6, the residuals are the very similar in the Poisson and Negative Binomial models, but the standardized residuals are notably different. When we have differences in standard errors, they are accompanied by differences in the estimates of prediction standard errors, impacting the standardized residuals. Plots for Negative Binomial models are more compact because these residuals are divided by larger values (bringing the standardized residuals closer to 0). These plots show clear differences between Poisson model and Negative Binomial model, but do not show which model is better in fitting the data. Cross Validation may help compare their performances.

The Cross Validation results are shown in Table 3 and Table 6. Table 3 shows that the mean RMSE values obtained are really close for both models, 1.2379 and 1.2380. Similarly, the standard deviation values of RMSE are also very close. In Table 6, mean values of RMSE are 1.2120 and 1.2112, and again, the standard deviation values are close to each other. In theory, the Negative Binomial model has an advantage in terms of accounting for overdispersion; however, there are no practical differences in RMSE observed here. This shows that by testing the ability of model prediction, Negative Binomial models do not show a very obvious advantage over the Poisson models. Based on the current test and visualization processes we have used, we cannot say that the Negative Binomial models perform better.

No Month vs. Month model performance

To compare the goodness of fit, I use the Akaike Information Criterion (AIC), which is a "fined technique based on in-sample fit to estimate the likelihood of a model to predict/estimate the future values" (Akaike, 1974). It is defined to equal

$$AIC = -2\ln(L) + 2\,p\,,$$

where L is the value of the likelihood and p is the number of estimated parameters. A smaller AIC value is generally understood to accompany a better model.

After calculating AIC for the two Poisson models, we find that without month, AIC is equal to 14,046.61, and the model with month has an AIC of 13,226.11. This means that with respect to goodness-of-fit, the model with the month covariate is better. For Negative Binomial models, without month and with month have AIC values of 12,237.25 and 11,866.19, respectively, supporting the same conclusion that the model with month is better. However, we also consider the results of our Cross Validations to measure differences in predictive capabilities between these models.

By looking at Cross Validation results in Table 3 and Table 6, the mean of RMSE values when fitting a Poisson model without month and with month are 1.2379 and 1.2120, and the standard deviations are 0.3598 and 0.3326, so we conclude that the Poisson model with month has a better ability to predicting this data. Similarly, for the

Negative Binomial models, the means of RMSE are equal to 1.2380 and 1.2112, with the model using month observing the lower error value. The standard deviation of RMSE for the Negative Binomial model with month is also smaller than the model without month. Hence, in terms of their ability of prediction, we find that the model with month will perform better.

In summary, with better abilities in both fitting the data and in making predictions on validation sets, we can conclude that for both model families, including the month covariate is preferable.

Conclusion

In conclusion, based on the result and discussion, we found out that using a Negative Binomial model did not have a very obvious advantage over the Poisson model. Additionally, in the Negative Binomial models, we assumed that the response variable follows a Negative Binomial distribution, which models the number of trails until we get a specified number of successes. Unlike a Poisson distribution, this is not as directly interpretable in explaining wildland fire counts. Thus, we are inclined to think that a Poisson model is more appropriate to model these data.

A Quasi-Poisson model has also been used to address the overdispersion problem. It directly incorporates a dispersion parameter when fitting models, so it is another approach for dealing with overdispersion (Faraway, 2016). However, it does not change the beta coefficients relative to the corresponding Poisson model, but rather only changes the standard errors. Therefore, its plot of difference values (between observed and predicted) are identical to the plot of the Poisson model. Also, Cross Validation gives identical results for both mean and standard deviation values of RMSE. So, while these results have been omitted, in order to deal with overdispersion a Quasi-Poisson model may be considered a preferrable alternative to using a Negative Binomial model.

For future work, we plan to consider more modeling approaches, in addition to investigating more flexible modeling techniques to handle temporal effect (e.g., splines).

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada, the Ontario Ministry of Natural Resources and Forestry, and the University of Western Ontario’s USRI program.

Bibliography

Akaike, H. (1974, December). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 716–723. https://doi.org/10.1109/tac.1974.1100705

C3.ai. (2021, September 28). *Root Mean Square Error (RMSE)*. C3 AI. Available at https://c3.ai/glossary/data-science/root-mean-square-error-rmse/ [Verified August 16, 2022]

Canadian Wildland Fire Information System | Canadian Forest Fire Weather Index (FWI) System. (n.d.). Natural Resources Canada. https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi

Centre, G. L. F. R., Stocks, B. J., & Great Lakes Forest Research Centre. (1974). *Wildfires and the Fire Weather Index System in Ontario*. Academic Service.

Cunningham, A. A., & Martell, D. L. (1973, June). A Stochastic Model for the Occurrence of Man-caused Forest Fires. *Canadian Journal of Forest Research*, 282–287. https://doi.org/10.1139/x73-038

Faraway, J. J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)* (2nd ed.). Chapman and Hall/CRC.

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

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013, March). Applied Logistic Regression. *Wiley Series in Probability and Statistics*. https://doi.org/10.1002/9781118548387

Matloff, N. (2017). *Statistical Regression and Classification: From Linear Models to Machine Learning (Chapman & Hall/CRC Texts in Statistical Science)* (1st ed.). Chapman and Hall/CRC.

Saputro, D. R. S., Susanti, A., & Pratiwi, N. B. I. (2021, February). The handling of overdispersion on Poisson regression model with the generalized Poisson regression model. *THE THIRD INTERNATIONAL CONFERENCE ON MATHEMATICS: Education, Theory and Application*. The Third International Conference on Mathematics. https://doi.org/10.1063/5.0040330

Stocks, B. J. (1971). *Analysis of the Fire Weather Index in Ontario (1963 to 1968)*. Internal Report 0–25. https://cfs.nrcan.gc.ca/publications?id=33695

Wagner, C. E., Canadian Forestry Service, van Wagner, C. E., & Canadian Forestry Service. (1987). *Development and Structure of the Canadian Forest Fire Weather Index System*. Academic Service.