Count based forest fire prediction model analysis based on climate data

Canada region

Namrata Dakua

29 April 2024

1. Abstract

Forest fire activities have intensified in recent decades globally; and Canada is one of the worst impacted countries for forest fires considering majority of population is dependent on forests for livelihood. This paper is aimed to evaluate count based statistical analysis and identify the climate parameters significant in predicting fire ignition using historical fire occurrence data and reanalysed climate dataset. Predictions on count-based data is performed using Poisson Regression model but real-world count-based-data do not always follow conditional mean and variance equivalence hence Quasi-Poisson and Negative-Binomial Regression model are also implemented. The accuracy of all three models is compared with root mean square error (RMSE) and variance on the test dataset. The results indicate that Quasi-Poisson regression model with square root link function gives better accuracy.

Keywords Canada, wildfires, climate impact, fire frequency, Poisson regression model, Quasi-Poisson regression model, Negative-Binomial regression model, square link function, over-dispersion, under-dispersion

2. Introduction

Forest fire activities have intensified in recent decades globally; Canada being a country with a vast landmass covered with forests, wildfires have been a serious natural disturbance for its boreal forests, and there is abundant evidence suggesting that fire conditions could worsen in the next century (Wang, Y. 2024). According to "Natural Resources Canada" annual 2023 report[1], Canada has 367 million hectares (ha) of forest, corresponding to 9% of the world's forest and 25% of the world's boreal forest. Canada's boreal forest is crucial to the national economy because of the available timber and non-timber products, mineral and energy resources, and hydroelectric potential of regional rivers. The boreal forest[2] provides food and renewable raw materials to Canadians. Weather and climate – including temperature, precipitation, wind, and atmospheric moisture – are critical aspects of fire activity (Flannigan et al. 2009). Gillett et al. (2004) suggest that the increase in area burned in Canada over the past four decades is due to human-caused increases in temperatures. Fire activity including area burned, fire occurrence, fire season, fire intensity, and fire severity respond dynamically to the climate–weather, fuels, and people. Recently, our climate has been warming because of increases of radiatively active gases (CO2, CH4, etc.) in the atmosphere caused by human activities (IPCC 2007). Such warming is likely to have a rapid and profound impact on fire activity (Scholze et al. 2006; Krawchuk et al. 2009a), as will potentially changes in precipitation, atmospheric moisture, wind, and cloudiness (Flannigan et al. 2006; IPCC 2007). Long-term fire activities are heavily influenced by climatic conditions; simulating the future climate through earth system models makes it possible to analyse future fire danger and severity for the Canadian boreal forests (Wang, Y. 2024). With the rapidly changing climate and global warming it has become important to monitor the weather and environmental factors contributing to forest fires, so that fire management strategies can be prepared before a fire occurs. Predicting the possibility of fire occurrence can help intake measures to reduce the amount of area burned during the fires (Robin et al. 2021).

This paper addresses the prediction of forest fire by exploring Poisson regression, Quasi-Poisson regression and Negative-Binomial regression models by estimate the number of fire ignition based

[1] Natural Resources Canada Report -https://natural-resources.canada.ca/our-natural-resources/forests/sustainable-forest-management/boreal-forest/13071

[2] Natural Resource Canda – Boreal Forest Report - https://natural-resources.canada.ca/our-natural-resources/forests/sustainable-forest-management/boreal-forest/13071

on weather forecasts and soil conditions. The dataset constitutes of historical fire activities from FIRMS (Fire Information for Resource Management System) and climate data from Copernicus Climate Change Service (C3S) Climate Data Store (CDS).

The subsequent sections of this paper is structured as follows: Section 3 describes the dataset variables which includes daily fire occurrence and weather reanalysis from CDS. The section 4 describes the notations, equations for different models and model validation criteria. Section 5 focuses on the results of model comparisons and significant independent variables for the prediction. Section 6 is conclusion, about importance of prediction model and future implications.

3. Data

3.1 Study areas

The area of this study is Canada country spanning between latitude 80°N in north to 41°N in south and 141°W in west to -52°W east.

According to Canadian Interagency Forest Fire Center (CIFFC) (Canada), there have been 4883 fires in 2022 and 6551 fire in 2023 burning area 1467976 hectares and 18496057 hectares, respectively. These figures indicate that Canada had its worst wildfire season in year 2023 burning around 19 million hectares of forest. Figure 1 from National Forestry Database website indicates the number of fires and area burned from year 1990.

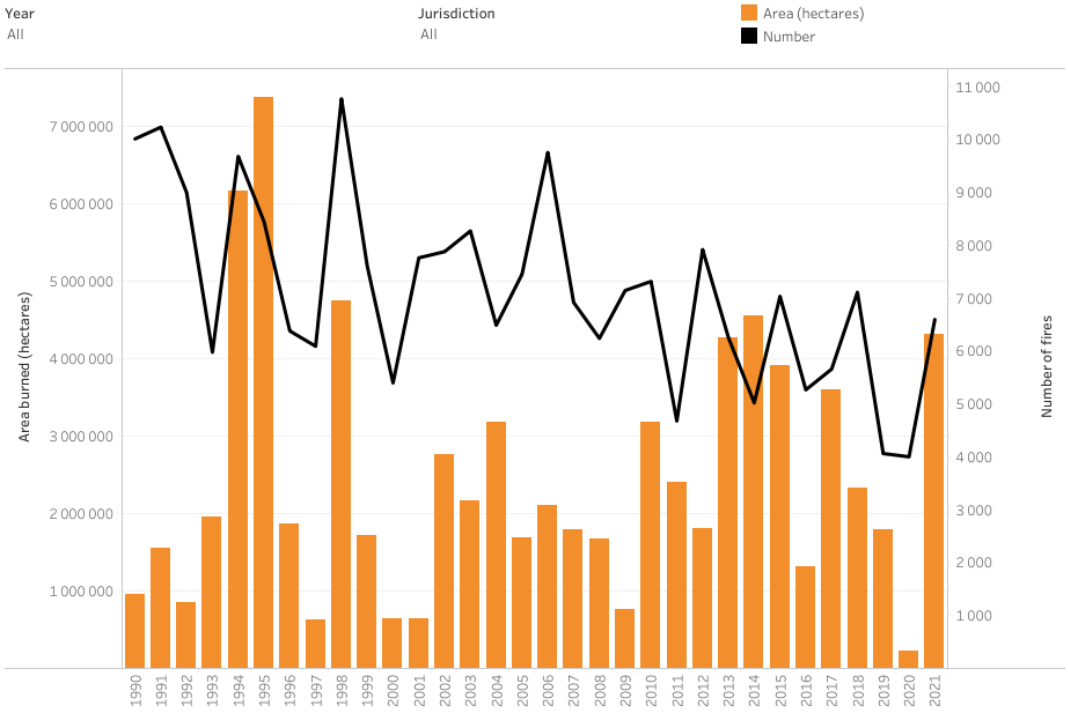

Figure 1. shows the high variability in both number of fires and area burned in Canada per year. This data graph is taken from National Forestry Database website - http://nfdp.ccfm.org/en/data/fires.php

3.2 Independent variables

The independent variables are comprised of historical fire data and climate data for past14 year between 2010 to 2023.

3.2.1 Historical Fire Dataset

The historical fire data has been obtained from FIRMS (Fire Information for Resource Management System) which is "Fire Information for Resource Management System (FIRMS)" distributes Near Real-Time (NRT) active fire data from the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard the Aqua and Terra satellites, and the Visible Infrared Imaging Radiometer Suite (VIIRS) aboard S-NPP, NOAA 20 and NOAA 21 (formally known as JPSS-1 and JPSS-2). Globally these data are available within 3 hours of satellite observation, but for the US and Canada active fire detections are available in real-time.

Table 1. shows the attributes of the fire dataset

Attribute	**Description**
Latitude	Center of 1 km fire pixel, but not necessarily the actual location of the fire as one or more fires can be detected within the 1 km pixel.
Longitude	Center of 1 km fire pixel, but not necessarily the actual location of the fire as one or more fires can be detected within the 1 km pixel.
Brightness	Channel 21/22 brightness temperature of the fire pixel measured in Kelvin.
Scan	The algorithm produces 1 km fire pixels, but MODIS pixels get bigger toward the edge of scan. Scan and track reflect actual pixel size.
Track	The algorithm produces 1 km fire pixels, but MODIS pixels get bigger toward the edge of scan. Scan and track reflect actual pixel size.
Acq_Date	Acquisition Date - Data of MODIS acquisition.
Acq_Time	Time of acquisition/overpass of the satellite (in UTC).
Satellite	A = Aqua and T = Terra.
Confidence	This value is based on a collection of intermediate algorithm quantities used in the detection process. It is intended to help users gauge the quality of individual hotspot/fire pixels. Confidence estimates range between 0 and 100% and are assigned one of the three fire classes (low-confidence fire, nominal-confidence fire, or high-confidence fire).
Bright_T31	Channel 31 brightness temperature of the fire pixel measured in Kelvin.
FRP	Fire Radiative Power - Depicts the pixel-integrated fire radiative power in MW (megawatts).
Type	Inferred hot spot type - 0 = presumed vegetation fire 1 = active volcano 2 = other static land source 3 = offshore
DayNight	Day or Night: D= Daytime fire, N= Nighttime fire

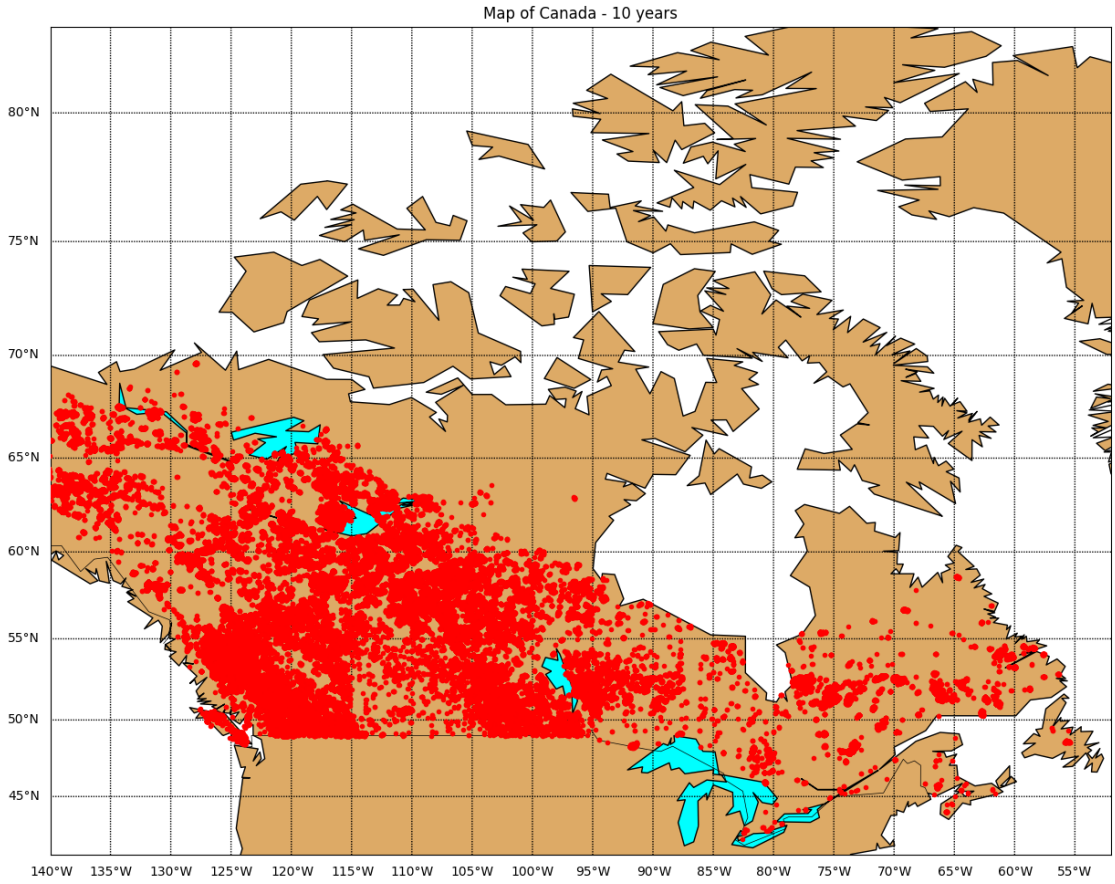

Figure 2. shows the location of all the fire ignition between 2010 to 2023 with confidence greater than 90% which means the quality of hotspot, higher value means high-confidence of fire.

3.2.2 Climate Dataset

The climate dataset has been obtained from Copernicus Climate Data Store (CDS) - https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form, an online platform that provides access to a wealth of climate data past, present, and future. It is part of the Copernicus Climate Change Service (C3S) led by the European Union. The climate data can be downloaded from the web form as well as using their api. I have downloaded the climate data for each fire record by calling the api which required date, time, area coordinates (north, west, south, east) and list of climate parameters needed in the response. The response of the api is ".nc" file which are Network Common Data Form files are, utilized as a standardized format for storing climate data in multi-dimensional arrays. These arrays enable users to access specific elements based on dimensions such as Latitude (Lat), Longitude (Lon), Time, and others
Table 2 shows the climate and weather variables used in the analysis. These explanatory variables represent temperature, precipitation, wind and soil conditions.

Github link for implementation for downloading climate data - https://github.com/namratadakua/2816863D-CountBasedForestFirePrediction/blob/main/download_climate_data/final_climate_data_file_only.ipynb

Github link for implementation for converting the climate data in .nc files to csv file -

https://github.com/namratadakua/2816863D-CountBasedForestFirePrediction/blob/main/download_climate_data/final_convert_to_csv.ipynb

Table 2. Climate and weather predictors

Factor	Name of variable	Variable description	Value description/Units
Temperature and pressure	2m temperature	This parameter is the temperature of air at 2m above the surface of land, sea or inland waters.	Kelvin (K)
Wind	10m u-component of wind	This parameter is the eastward component of the 10m wind. It is the horizontal speed of air moving towards the east, at a height of ten metres above the surface of the Earth, in metres per second.	m s^{-1}
	10m v-component of wind	This parameter is the northward component of the 10m wind. It is the horizontal speed of air moving towards the north, at a height of ten metres above the surface of the Earth, in metres per second.	m s^{-1}
Precipitation and rain	Total precipitation	This parameter is the accumulated liquid and frozen water, comprising rain and snow, that falls to the Earth's surface.	The units of this parameter are depth in metres of water equivalent.
Soil	Soil type	This parameter is the texture (or classification) of soil used by the land surface scheme of the ECMWF Integrated Forecasting System (IFS) to predict the water holding capacity of soil in soil moisture and runoff calculations.	The seven soil types are: 1: Coarse, 2: Medium, 3: Medium fine, 4: Fine, 5: Very fine, 6: Organic, 7: Tropical organic. A value of 0 indicates a non-land point
	Soil temperature level 1	This parameter is the temperature of the soil at level 1 (in the middle of layer 1).	Layer 1 - surface is at 0 - 7cm. Kelvin (K)
	Soil temperature level 2	This parameter is the temperature of the soil at level 2 (in the middle of layer 2).	Layer 1 - surface is at 7 - 28 cm. Kelvin (K)
	Soil temperature level 3	This parameter is the temperature of the soil at level 3 (in the middle of layer 3).	Layer 3 – surface is at 28 – 100cm. Kelvin (K)

	Soil temperature level 4	This parameter is the temperature of the soil at level 4 (in the middle of layer 4).	Layer 4 - surface is at 100-289cm. Kelvin (K)
	Volumetric soil water layer 1	This parameter is the volume of water in soil layer 1 (0 - 7cm, the surface is at 0cm).	$m^3 m^{-3}$
	Volumetric soil water layer 2	This parameter is the volume of water in soil layer 2 (7 - 28cm, the surface is at 0cm).	$m^3 m^{-3}$
	Volumetric soil water layer 3	This parameter is the volume of water in soil layer 3 (28 - 100cm, the surface is at 0cm).	$m^3 m^{-3}$
	Volumetric soil water layer 4	This parameter is the volume of water in soil layer 4 (100 - 289cm, the surface is at 0cm).	$m^3 m^{-3}$

3.3 Dependent variables

The dependent variable for Poisson model was generated using the latitude and longitude coordinates which are the centre of 1 km fire pixel, but not necessarily the actual location of the fire as one or more fires can be detected within the 1 km pixel. These coordinates are mapped to 1x1km grid cell and number of fire points calculated for each cell.

The distance between two latitudes or longitude point is approximately 111km. The grid is created by range of parallels with latitude 41 to 70 and distance 1/111; similarly range of meridians with longitude –142 to 55 and distance 1/111.

The count is assigned to each cell by mapping the fire points to cell as shown in Figure 4 flowchart.

Steps for grid creation:

1. Read each fire record one by one from year 2010 to 2023.
2. Find the grid cell based on latitude and longitude.
3. If the confidence of the fire record is greater than 90% and it is considered as a hotspot and included in the fire count
4. If the confidence is not less than 90% then it is added to the non-fire list on the cell.

A cell can have both fire ignition point list as well as non-fire because the climatic condition can be different for different year, month and day combinations. Some climatic conditions can lead to fire ignition, and some may not hence both kind of data are mapped for the grid.

Figure 4 describes the month-wise number of fire hotspots indicating summer season is more prone to forest fires.

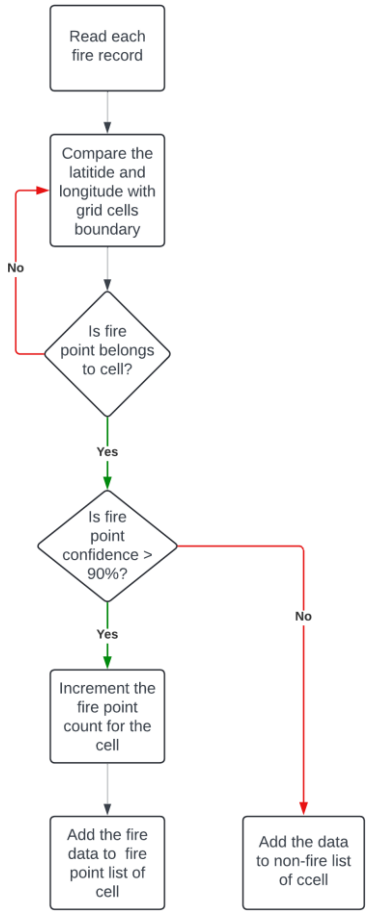

Figure 3. Flow chart for grid data preparation.

Github link for above flowchart - https://github.com/namratadakua/2816863D-CountBasedForestFirePrediction/blob/main/final_model_data_preparation.py

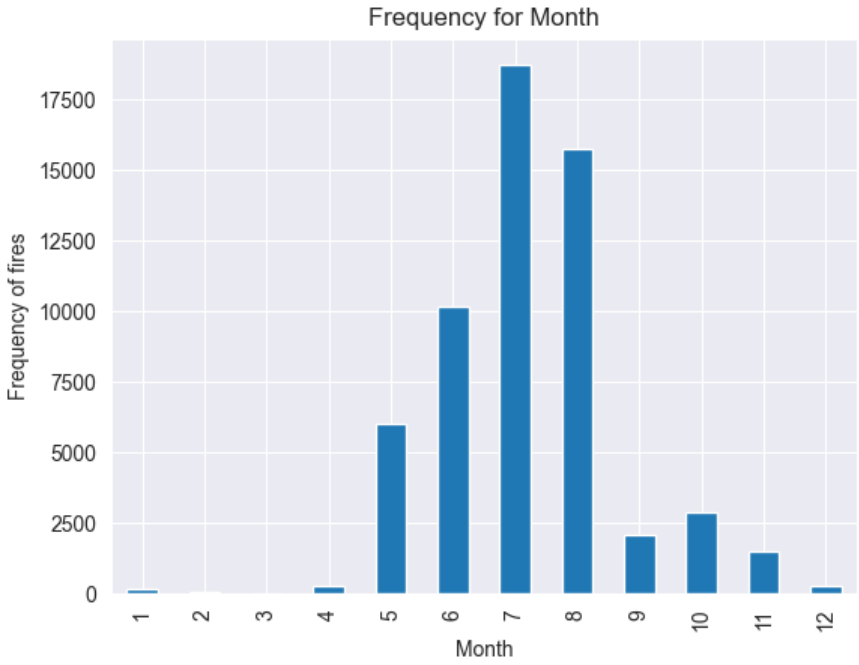

Figure 4. Number of fire points month-wise

4. Methods

The fire ignition data considered here is in the form of counts and we would like to relate these data to climatic and environmental conditions. The most commonly used count models are Poisson and Negative Binomial (Cameron and Trivedi, 2023). But in practice count data are often overdispersed, with the conditional variance exceeding conditional mean (Cameron and Trivedi, 2023).

A common way to deal with overdispersion for counts is to use a generalized linear model framework (McCullagh and Nelder 1989), where the most common approach is a ‘‘quasi-likelihood,’’ with Poisson-like assumptions (that we call the quasi-Poisson from now on) or a negative binomial model [9]. According to Wilson et al., 2021, the Quasi-Poisson can also be used for under-dispersed dataset as well.

In this report, a comparison of Poisson regression, quasi-Poisson regression and negative binomial model is performed to identify the significant parameters for fire ignition counts with respect to climate and environmental conditions.

4.1 Notation and model structure

The spatial domain is divided in equal-sized 1x1km grid.

The response $y_c$represent the number of active fire points in a grid on a given unit time; it takes natural values$y = 1, 2, 3 \ldots y = 1, 2, 3 \ldots$

Each cell (centroid) is represented by $c \in \{(i,j)\}$, where $i = 1,2, \ldots M$ and $j = 1, 2, \ldots P$, where $C = M \times P$ is the total number of locations, with M representing meridians and P representing parallels; concretely, M = 9657 and P = 3219.

x_c is a p-dimensional vector of explanatory covariates observed on time *t* in grid cell *c*. The list of covariates is shown in Table 2, including temperature, pressure, soil condition and soil water volume.

4.2 Poisson Regression Model

The Poisson regression model specifies that y_c given x_c is Poisson distributed with density:

$$f(y_c|x_c) = \frac{e^{-\mu_c}\lambda_c^{y_c}}{y_c!} \quad (1)$$

and mean parameter

$$\mu_c = \exp(X_c'\beta) \quad (2)$$

Counting process theory provides a motivation for choosing the Poisson distribution; taking the exponential of $X_c'\beta$ in (2) ensures that the parameter μ_c is nonnegative.

This model implies that the conditional mean is given by,

$$\mathrm{E}[y_c|x_c] = \exp(X_c'\beta) \quad (3)$$

$$E(y_c) = e^{\beta_0 + \beta_1 x_{c1} + \beta_2 x_{c2} + \cdots} \quad (4)$$

where expected $E(y_c) = \mu_c$ is the expected number of fire points for a given cell c with covariates x_1, x_2, x_3 ... and regression coefficients β_0, β_1, β_2, β_3 ... respectively.

where the offset term represents the natural log of the time taken for the y events to occur.

4.3 Quasi-Poisson Model

Poisson regression is a commonly used tool for analysing rate data; however, the assumption that the mean and variance of a process are equal rarely holds true in practice. When this assumption is violated, a quasi-Poisson distribution can be used to account for the existing over- or under-dispersion (Wilson et al., 2021).

Quasi-Poisson model is one of the generalized linear models with expected value and variance equation as:

as:

$$E(y_c) = \mu_c \quad (5)$$

$$Var(y_c) = \theta\mu_c \quad (6)$$

where $E(y_c)$ is the expectation of y, $var(y)$ is the variance of y, $\mu > 0$ and $\theta > 1$ or $\theta < 1$. Thus θ allows the variance to be larger or smaller than the mean.

Likelihood function for quasi-Poisson (quasi-likelihood) does not require a specific probability density function to estimate regression parameter except for response variable assumption (McCullagh and Nelder, 1989).

In Eq. (6), θ is an over or under dispersion parameter. The close relationship between Eq. (6) and the expectation and variance of a Poisson distribution, along with the use of a link function, justify calling this a "quasi-Poisson" model of two parameters μ and θ, denoted as $Y \sim Poi(\mu, \theta)$

The quasi-Poisson model formulation has the advantage of leaving parameters in a natural, interpretable state and allows standard model diagnostics without loss of efficient fitting algorithms [9].

4.4 Negative Binomial Model

Negative binomial model is also a generalised linear model whose distribution denoted as

$$Y \sim NB(\mu, \kappa)$$

with parameterization such that

$$E(y_c) = \mu \quad (7)$$

$$Var(y_c) = \mu_c + \kappa\mu_c^2 = (1 + \kappa\mu_c)\mu_c \quad (8)$$

where $\mu > 0$ and $\kappa > 0$. The overdispersion (the amount in excess of μ) is the multiplicative factor $1 + \kappa\mu$, which depends on μ.

4.5 Model Validation

The acquisition date (acq_date) attribute from the fire record is split into year, month, and day.

After assigning the fire ignition points to the respective 1x1km grid, a dataset of mean temperature, pressure, soil condition and fire point count is created by aggregating each cell's fire points on DayNight, year, month, day.

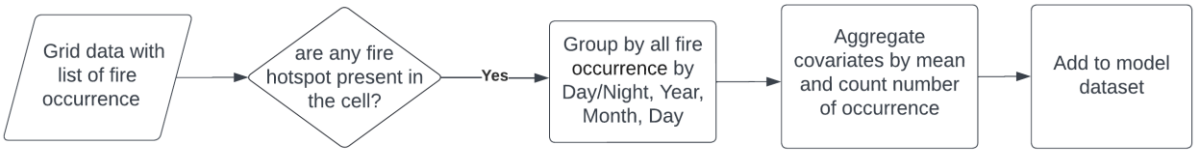

Figure 5. Flow chart for model dataset preparation.

As the data considered is the last 14 years (2010-2023), there is a difference between climatic conditions of prior years to recent time. Hence aggregation is done at day level. Each record in the model dataset represents number of fire points on a day during day/night with mean value of independent variables. The grid and combined data set of fire record with corresponding climate data is implemented in python.

The dataset is randomly sample to create 70% training data and 30% test data using "sample" method in R which extracts a random subset of elements from the provided data with or without replacement. I have taken without replacement approach. The number of records in training data are 137650 and in test data are 58993.

The variables present in the model data are: "daynight", "month", "10m u-component of wind", "10m v-component of wind ", "2m temperature", "soil temperature level 1", "soil temperature level 2", "soil temperature level 3", "soil temperature level 4", "soil type", "total precipitation", "volumetric soil water layer 1", "volumetric soil water layer 2", "volumetric soil water layer 3", "volumetric soil water layer 4", "number of fire"

From acq_date parameter only month field is included in the model creation as the weather conditions can remain similar for a particular month in different years.

The training dataset is applied to all the above-mentioned models; RMSE and variance is calculated on the predicted test data.

Poisson regression and Quasi-Poisson regression with log link function produces same coefficient estimates and standard error. Hence Quasi-Poisson is implemented with square root link function because the square root link function focuses on the expected square root of the response variable, ensuring non-negative predictions and potentially addressing specific issues in count data modelling. The square root link function assumes that y is the squared function of the predictors:

$$y = [\beta_0 + \beta_1 x_1 + \beta_2 x_{2\ldots}]^2$$

The square root link function can help stabilize the variance in such cases as it transforms the predicted counts onto a square root scale, effectively compressing the higher values and reducing the impact of large counts on the overall variance.

This can lead to a more accurate fit and improved model performance when the variance-mean relationship violates the assumptions of the standard Poisson model.

5. Results

Root mean square and variance is used to compare the models.

$$RMSE = \sqrt{\frac{(actual\ count - predicted\ count)^2}{number\ of\ samples}}$$

Usually For any given data set, information theoretic approaches such as Akaike information criteria (AIC; Akaike 1973) or Bayesian information criteria (BIC; Schwarz 1978) might be considered to choose between a quasi-Poisson model and a negative binomial. These approaches depend on a distributional form and a likelihood; however, quasi models are only characterized by their mean and variance, and do not necessarily have a distributional form (Ver Hoef et al. 2007). Hence here for model evaluation RMSE is used. These values are shown in Table 4. Based on this comparison, Quasi-Poisson regression model gives better accuracy than Poisson regression and Negative-Binomial regression.

Table 3 and Table 4 shows the coefficient estimates, standard errors and p-value for each of the model independent variable.

According to Patrick et al. (2022), 94% of nighttime fire radiative power (FRP) detected by Moderate Resolution Imaging Spectroradiometer (MODIS) from 2003–2020 was emitted from wildfires. The positive coefficient estimates for night indicates that there are more chances of fire ignition in night due to higher temperatures leading to nighttime higher temperatures and lower humidities. The coefficient for summer months for Canada – June to September, are approaching 0 or positive value, indicating, these months are more prone to forest fire.

For Quasi-Poisson with square root link function, one-unit increase is temperature by exp(-0.002) = 0.998002. Northward (10v component) and eastward (10u component) wind movement also have positive impact on fire ignitions. Wind speed facilitate fire spread and intensity by improved heat transfer to adjacent fuels and by providing oxygen for combustion (Patrick et al. 2022).

The different soil types are - 1: Coarse, 2: Medium, 3: Medium fine, 4: Fine, 5: Very fine, 6: Organic, 7: Tropical organic. Coefficient of soil type 1(Coarse) that do not hold water is highest, again indicating that dryer soil can result in fire ignition.

All three models indicate that Soil temperature level 4 and Volumetric soil water layer 4 (that is 100-289 cm below surface) is not significant. The rest of the independent variables are significant in predicting the count for a particular weather condition variable.

Dispersion parameter (θ) for Quasi-Poisson is 0.7055879. This indicates that dataset is under-dispersed. One of the reasons for under-dispersion due to grouping and aggregation of fire ignition data based on date (and then getting the climatic data) and other is having more records with zero number of fire occurrence.

Github link for model validation implementation - https://github.com/namratadakua/2816863D-CountBasedForestFirePrediction/blob/main/model_validation.R

Table 3. Coefficient estimates, standard error and p-Value for Poisson, Quasi-Poisson (both log and link functions) and Negative Binomial regression. All the models have been implemented with log-link function

	Coefficient			
	Poisson	Quasi(log)	Quasi(sqrt)	NB
(Intercept)	-1.763	0.008	0.687	-1.763
daynightN	0.74	0	0.208	0.74
month2	-0.38	0.079	-0.09	-0.38
month3	-0.843	0	-0.144	-0.843
month4	-0.851	0	-0.168	-0.851
month5	0.094	0.334	0.026	0.094
month6	0.097	0.322	0.032	0.097
month7	0.088	0.373	0.033	0.088
month8	0.173	0.079	0.054	0.173
month9	0.055	0.584	0.018	0.055
month10	0.041	0.672	0.018	0.041
month11	0.196	0.045	0.054	0.196
month12	0.17	0.146	0.045	0.17
10m u-component of wind	0.016	0	0.004	0.016
10m v-component of wind	0.021	0	0.006	0.021
2m temperature	-0.011	0	-0.003	-0.011
Soil temperature level 1	-0.014	0.001	-0.003	-0.014
Soil temperature level 2	0.051	0	0.013	0.051
Soil temperature level 3	-0.03	0	-0.008	-0.03
Soil temperature level 4	0.003	0.169	0.001	0.003
Soil type1	0.054	0.161	0.008	0.054
Soil type2	0.014	0.726	-0.003	0.014
Soil type3	0.04	0.377	0.003	0.04
Soil type4	-0.055	0.26	-0.018	-0.055
Soil type5	-0.002	0.978	-0.007	-0.002
Soil type6	-0.158	0.003	-0.051	-0.158
Total precipitation	-94.815	0	-24.712	-94.815

Volumetric soil water layer 1	0.641	0.001	0.18	0.641
Volumetric soil water layer 2	-1.912	0	-0.496	-1.912
Volumetric soil water layer 3	1.937	0	0.511	1.937
Volumetric soil water layer 4	-0.154	0.375	-0.047	-0.154

Table 4. Coefficient estimates, standard error and p-Value for Poisson, Quasi-Poisson and Negative Binomial regression.

	Standard Error				p-Value			
	Poisson	Quasi (log)	Quasi (sqrt)	NB	Poisson	Quasi (log)	Quasi (sqrt)	NB
(Intercept)	0.661	0.555	0.144	0.661	0.008	0.001	0	0.008
daynightN	0.023	0.019	0.005	0.023	0	0	0	0
month2	0.217	0.182	0.036	0.217	0.079	0.037	0.013	0.079
month3	0.228	0.192	0.031	0.228	0	0	0	0
month4	0.117	0.098	0.02	0.117	0	0	0	0
month5	0.098	0.082	0.019	0.098	0.334	0.25	0.174	0.334
month6	0.098	0.083	0.019	0.098	0.322	0.239	0.097	0.322
month7	0.099	0.083	0.019	0.099	0.373	0.288	0.085	0.373
month8	0.099	0.083	0.019	0.099	0.079	0.037	0.005	0.079
month9	0.1	0.084	0.02	0.1	0.584	0.514	0.37	0.584
month10	0.096	0.081	0.019	0.096	0.672	0.614	0.349	0.672
month11	0.098	0.082	0.019	0.098	0.045	0.017	0.005	0.045
month12	0.117	0.098	0.024	0.117	0.146	0.084	0.055	0.147
10m u-component of wind	0.002	0.002	0.001	0.002	0	0	0	0
10m v-component of wind	0.003	0.002	0.001	0.003	0	0	0	0
2m temperature	0.003	0.002	0.001	0.003	0	0	0	0
Soil temperature level 1	0.004	0.003	0.001	0.004	0.001	0	0.001	0.001
Soil temperature level 2	0.004	0.004	0.001	0.004	0	0	0	0
Soil temperature level 3	0.004	0.003	0.001	0.004	0	0	0	0
Soil temperature level 4	0.002	0.002	0.001	0.002	0.169	0.102	0.262	0.169
Soil type 1	0.038	0.032	0.008	0.038	0.161	0.095	0.298	0.161
Soil type 2	0.041	0.034	0.009	0.041	0.726	0.677	0.702	0.726

Soil type 3	0.045	0.038	0.01	0.045	0.377	0.293	0.739	0.377
Soil type 4	0.049	0.041	0.01	0.049	0.26	0.179	0.076	0.26
Soil type 5	0.055	0.046	0.012	0.055	0.978	0.974	0.57	0.978
Soil type 6	0.054	0.046	0.012	0.054	0.003	0.001	0	0.003
Total precipitation	22.714	19.073	5.181	22.715	0	0	0	0
Volumetric soil water layer 1	0.189	0.159	0.043	0.189	0.001	0	0	0.001
Volumetric soil water layer 2	0.31	0.261	0.069	0.31	0	0	0	0
Volumetric soil water layer 3	0.259	0.217	0.057	0.259	0	0	0	0
Volumetric soil water layer 4	0.174	0.146	0.04	0.174	0.375	0.291	0.232	0.375

Table 5. below for RMSE for test data of 58993 size indicates that Quasi-Poisson provides better fit that log link functions.

Table 5. Root mean square error and variance of the prediction on test data

	RMSE	Variance
Poisson	1.648	0.106
Quasi-Poisson (log link)	1.648	0.106
Quasi-Poisson (Square root link)	0.507	0.008
Negative-Binomial	1.648	0.106

The monthly cross-table plot of actual prediction versus Quasi-Poisson predictions is not exact but following the pattern month-wise that we could expect more fire count during the summer months.

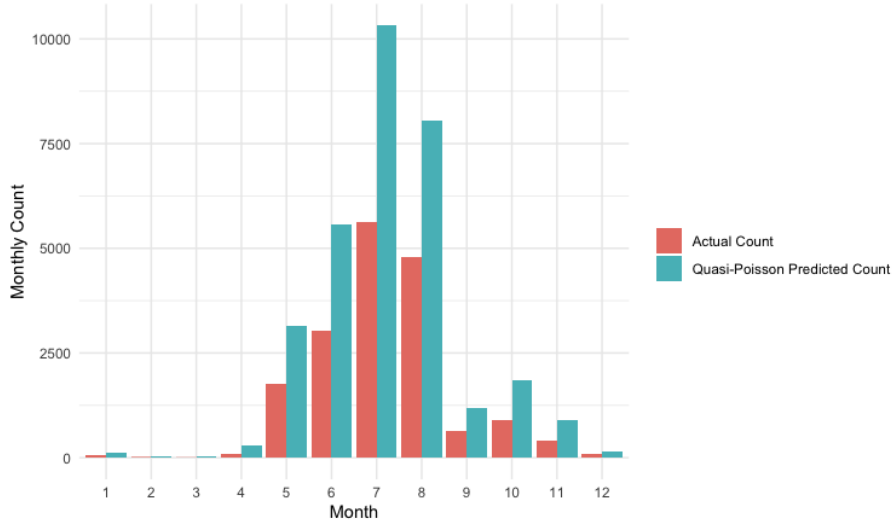

Figure 7. Cross-tab plot for test data, monthly actual count vs. monthly Quasi-Poisson Predicted Count

6. Conclusion

Global warming and climate change is disturbing our eco-system, and one of the major impacts is frequent wild forest fires. In recent years, forest fires have become more frequent, impacting human health, livelihood and contributing to global warming. This analysis proposes a machine learning model to identify the climatic and environment parameters contributing to fire occurrence and a way to predict the future occurrences. The Poisson Regression model considers the conditional mean and variance to be equal but real count-based data are over-dispersed or under-dispersed, hence Quasi-Poisson regression model is implemented. The analysis result indicate that the data is under-dispersed, and reasons are grouping and aggregation. For further studies, I can collect climate data for each fire occurrence by time without grouping for the day and instead of considering only fire data with confidence parameter more than 90% for the count; I can include fire data with confidence parameter more than 70% in the count, which will increase the variance and balance the number of records with count zero. For under-dispersed dataset Conway-Maxwell-Poisson can also explored. With this dataset, a time-series prediction can be implemented, based on last 10 days climate condition whether there can be fire ignition on current day. This model can be extended to include more granular climate parameters for rain, heat, evaporation, and vegetation to get more accurate predictions.

REFERENCES

[1] C. A. Graff, S. R. Coffield, Y. Chen, E. Foufoula-Georgiou, J. T. Randerson and P. Smyth, "Forecasting Daily Wildfire Activity Using Poisson Regression," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 7, pp. 4837-4851, July 2020, doi: 10.1109/TGRS.2020.2968029.

[2] Cameron AC, Trivedi PK. Basic Count Regression. In: Regression Analysis of Count Data. Econometric Society Monographs. Cambridge University Press; 2013:69-110.

[3] Carcaillet C, BergeronY, Richard PJH, Fréchette B, Gauthier S, Prairie YT (2001) Change of fire frequency in the eastern Canadian boreal forests during the Holocene: does vegetation composition or climate trigger the fire regime? Journal of Ecology 89, 930–946. doi:10.1111/J.1365-2745.2001.00614.X

[4] Flannigan MD, Amiro BD, Logan KA, Stocks BJ,Wotton BM (2006) Forest fires and climate change in the 21st century. Mitigation and Adaptation Strategies for Global Change 11, 847–859. doi:10.1007/S11027-005-9020-7

[5] Flannigan, Mike & Krawchuk, Meg & Wotton, Mike & Johnston, Lynn. (2009). Implications of changing climate for global Wildland fire. International Journal of Wildland Fire. 18. 483-507. 10.1071/WF08187

[6] Gillett NP,WeaverAJ, ZwiersFW, FlanniganMD(2004) Detecting the effect of climate change on Canadian forest fires. Geophysical Research Letters 31, L18211. doi:10.1029/2004GL020876

[7] Hector, Andy, 'Generalized Linear Models', *The New Statistics with R: An Introduction for Biologists*, 2nd edn (Oxford, 2021; online edn, Oxford Academic, 19 Aug. 2021), https://doi.org/10.1093/oso/9780198798170.003.0015

[8] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2023): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.adbb2d47 (Accessed on DD-MMM-YYYY)

[9] Holodinsky JK, Yu AYX, Kapral MK, Austin PC. Comparing regression modeling strategies for predicting hometime. BMC Med Res Methodol. 2021 Jul 7;21(1):138. doi: 10.1186/s12874-021-01331-9. PMID: 34233616; PMCID: PMC8261957

[10] IPCC (2007) Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the FourthAssessment Report of the Intergovernmental Panel on Climate Change. (Eds RK Pachauri,AReisinger) (Geneva, Switzerland)

[11] Krawchuk MA, Moritz MA, Parisien M-A, Van Dorn J, Hayhoe K (2009a) Global pyrogeography: macro-scaled statistical models for understanding the current and future distribution of fire. PLoS ONE 4, e5102

[12] National Forestry Database - http://nfdp.ccfm.org/en/data/fires.php

[13] Natural Resources Canada - https://natural-resources.canada.ca/our-natural-resources/forests/state-canadas-forests-report/how-do-forests-benefit-canadians/16509

[14] Patrick H. Freeborn, W. Matt Jolly, Mark A. Cochrane, Gareth Roberts, Large wildfire driven increases in nighttime fire activity observed across CONUS from 2003–2020, Remote Sensing of Environment, Volume, 268, 2022, 112777, ISSN 0034-4257, https://doi.org/10.1016/j.rse.2021.112777.

[15] Piyush Jain, Sean C.P. Coogan, Sriram Ganapathi Subramanian, Mark Crowley, Steve Taylor, and Mike D. Flannigan. 2020. A review of machine learning applications in wildfire science and management. *Environmental Reviews*. 28(4): 478-505. https://doi.org/10.1139/er-2020-0019

[16] Robin Singh Bhadoria, Manish Kumar Pandey, Pradeep Kundu, RVFR: Random vector forest regression model for integrated & enhanced approach in forest fires predictions, Ecological Informatics, Volume 66, 2021, 101471, ISSN 1574-9541, https://doi.org/10.1016/j.ecoinf.2021.101471

[17] Scholze M, KnorrW, Arnell NW, Prentice IC (2006) A climate-change risk analysis for world ecosystems. Proceedings of the National Academy of Sciences of the United States of America 103, 13 116–13 120. doi:10.1073/PNAS.0601816103

[18] Vegetation Zones of Canada: a Biogeoclimatic Perspective (Information Report GLC-X-25). Natural Resources Canada, Canadian Forest Service, Great Lakes Forestry Centre. Information Report GLC-X-25. 172 p

[19] Ver Hoef, Jay M. and Boveng, Peter L., "QUASI-POISSON VS. NEGATIVE BINOMIAL REGRESSION: HOW SHOULD WE MODEL OVERDISPERSED COUNT DATA?" (2007). Publications, Agencies and Staff of the U.S. Department of Commerce. 142. https://digitalcommons.unl.edu/usdeptcommercepub/142

[20] Wang, Y. (2024). The effect of climate change on forest fire danger and severity in the Canadian boreal forests for the period 1976–2100. Journal of Geophysical Research: Atmospheres, 129, e2023JD039118. https://doi.org/10.1029/2023JD039118

[21] Wilson SR, Leonard RD, Edwards DJ, Swieringa KA, Underwood M. Inference for Under-Dispersed Data: Assessing the Performance of an Airborne Spacing Algorithm. Qual Eng. 2018;30(4):546-555. doi: 10.1080/08982112.2018.1482339. Epub 2018 Oct 18. PMID: 33442200; PMCID: PMC7802820.