**Logo

Description automatically generated**

**Fire Weather Data Visualization**

**Midterm Report**

**STAT 683**

Prof. Sutanoy Dasgupta

**Team 1**

Abdullatif Alnuaimi, Ben Klein

# Introduction

Wildfires occur often around the United States, especially in hotter and drier climates and can cost either millions of dollars in damages or catastrophic natural damages to our environment. Many local governments are tasked with trying to figure out ways to dampen these financial and environmental losses by assessing large datasets over long periods of time and making decisions that provide relief to the overall negative effects that wildfires can cause. This is where our team comes into the picture.

# Problem Background

The problem our team is trying to solve is how can we visualize spatially the spread of wildfires in the US and specifically the state of Texas. We were tasked with generating a method(s) to find either trends or relationships among the wildfire locations. These trends that we find can ultimately lead to either an infographic or visualization that shows local government officials where action needs to be taken.

# Goal

Our goal with this project and analysis is to be able to use interactive visualizations and predict the time spent on putting out individual wildfires based on their location using a gaussian process spatial regression model. We want to come up with a precise and comprehensive model that is able to take into account not only location, but time as well, similar to the process of time series modeling, so we can identify any type of seasonal trend or stationarity in the data.

# Motivation

Both Abdullatif and I are fairly skilled in terms of spatial and temporal modeling in statistics so approaching this part of the project provided by Dr. Tao was going to complement both of our programming strengths. The challenge was also included in constructing a spatial prediction model with data of this size which is a new step for both of us. We also believe that our strengths could provide a benefit to helping forecast these wildfires and potentially lower the time spent to put out each individual fire. Using an accurate prediction model could provide local government the necessary evidence to place more fire departments and fire station hubs in the proximity of wildfire hotspots, which can effectively lower the natural and economical damage that wildfires usually cause.

**Data Description**

The objectives of this step are to identify datasets that help us understand the problem statement and give us an idea on what models we should use. Three datasets were collected:

**1- Weather\_Stations\_Data\_US**

* From: National Interagency Fire Center
* Link: <https://data-nifc.opendata.arcgis.com>
* Period: Dec,2021 - Aug,2022

**2- Wildland\_Fire\_Locations\_US**

* From: National Interagency Fire Center
* Link: <https://data-nifc.opendata.arcgis.com>
* Period: April,2009 - Sep,2022

**3- Texas\_Fire\_Departments\_and\_Substations**

* From: Texas A&M Forest Service
* Link: <https://public.tfswildfires.com>

**Exploratory Data Analysis (EDA)**

1. **Data Preparation (Data cleansing and imputation)**

We used Python to perform data cleaning to the three datasets including the following operations:

1. Removing duplicates
2. Filtering unwanted outliers
3. Selecting specific features
4. Splitting / Rename columns
5. Changing data types
6. Handling missing data

The results of these operations were the following:

1- Weather\_Stations\_Data\_US

* # of data points: **2951**
* Dataset features:







2- Wildland\_Fire\_Locations\_US

* # of data points: **136,868**
* Dataset features:





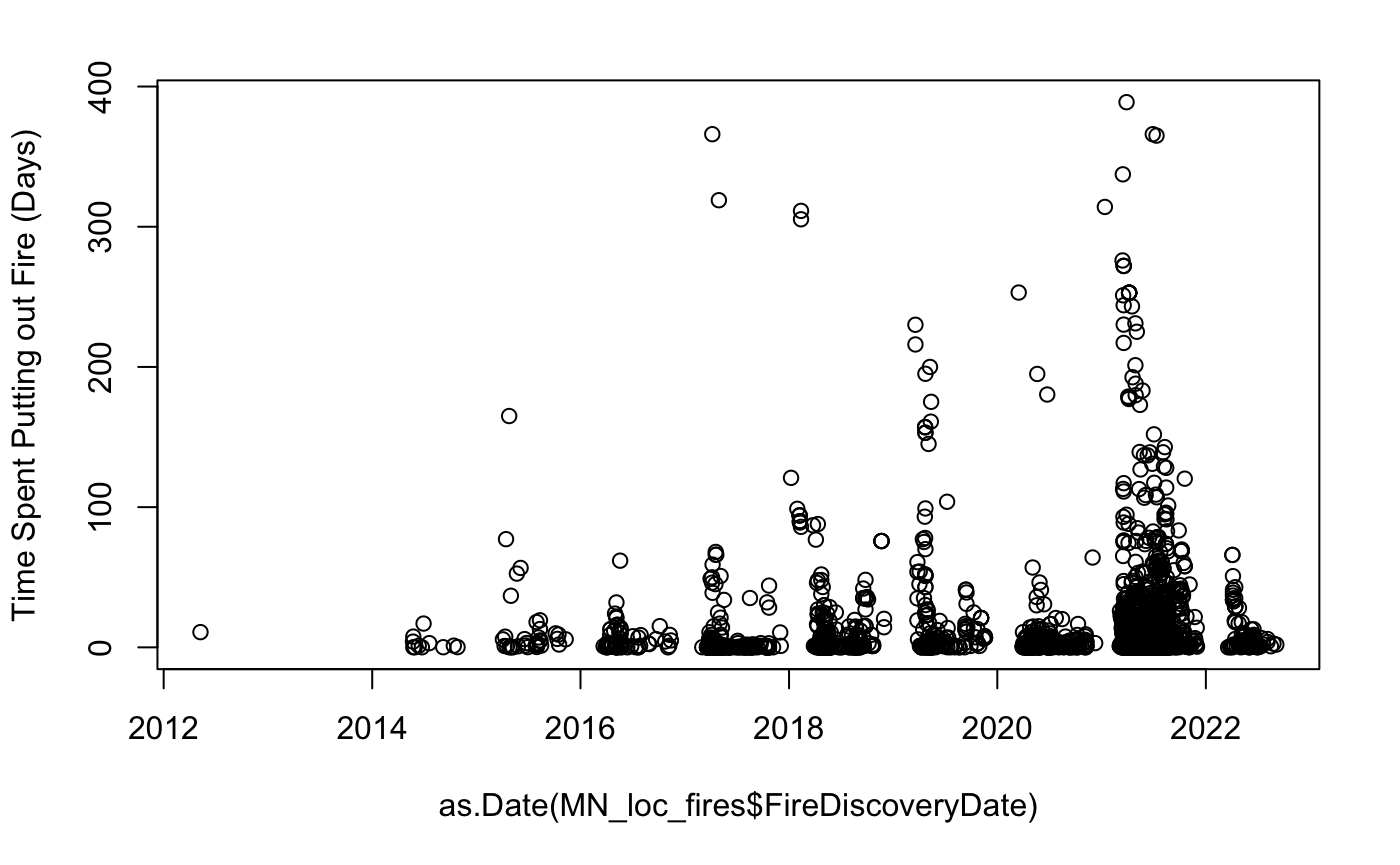
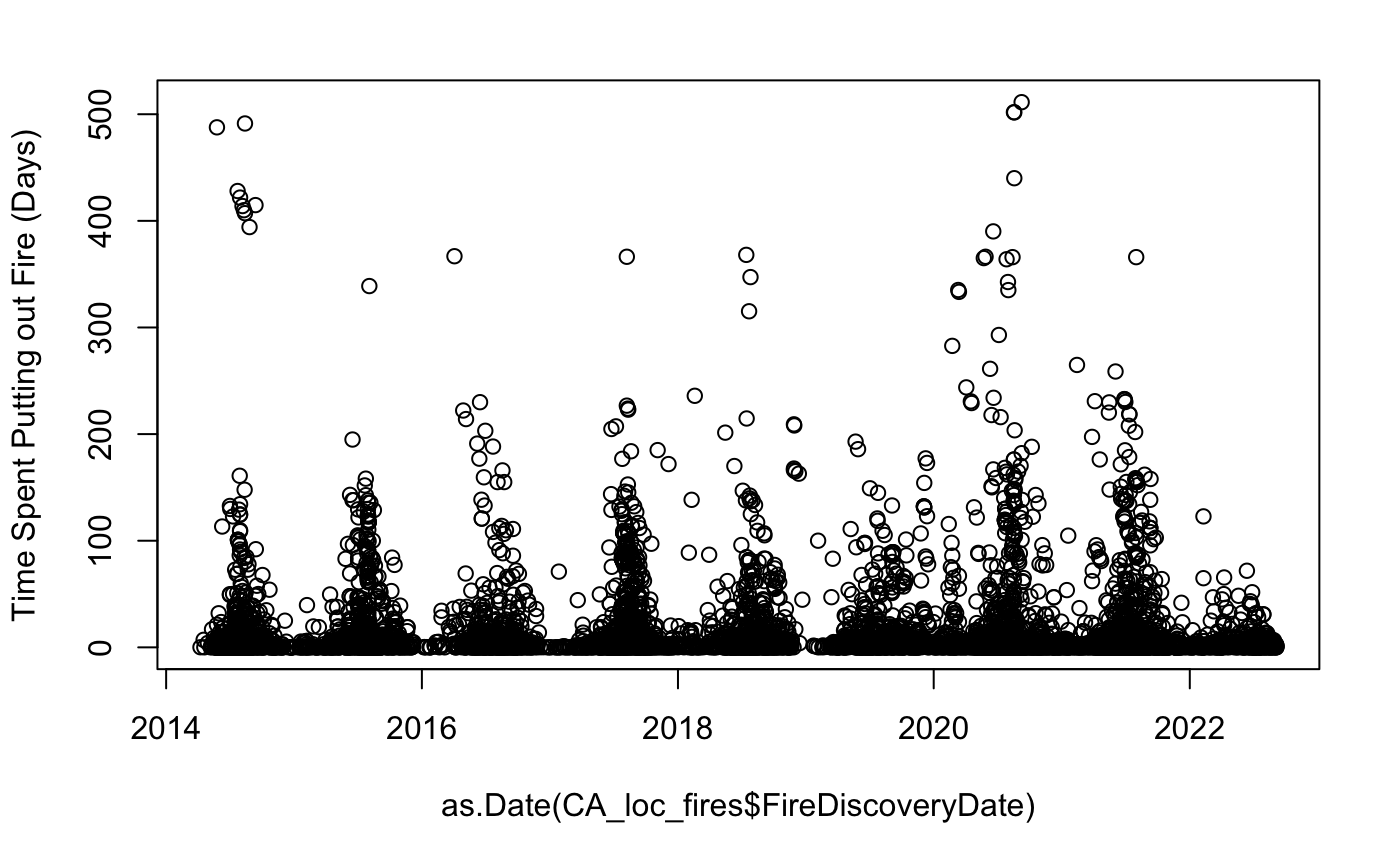
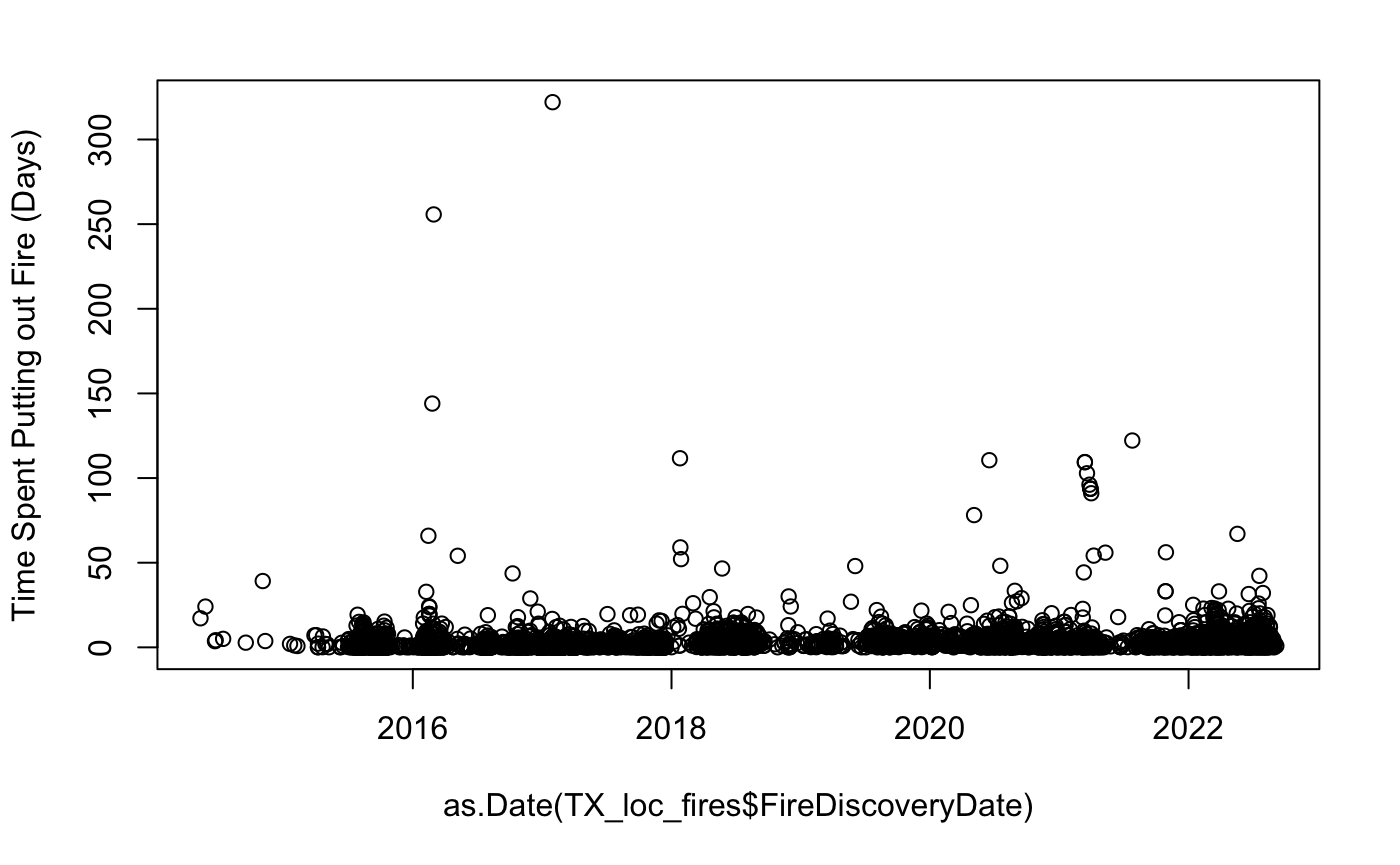
3- Texas\_Fire\_Departments\_and\_Substations

* # of data points: **3189**
* Dataset features:

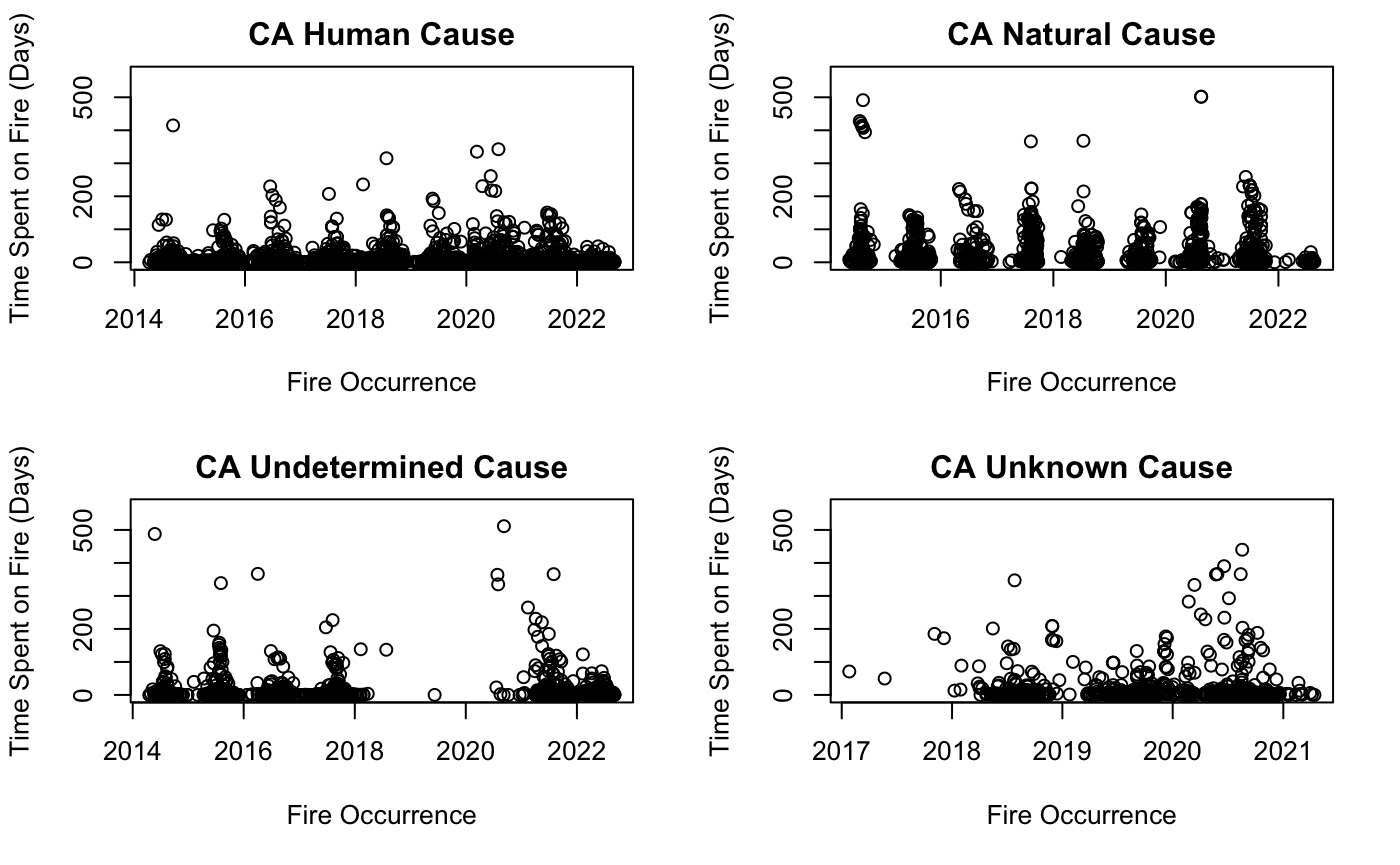
| FireDepartmentName | SubstationName/PhysicalCity | County | Latitude | Longitude |
| --- | --- | --- | --- | --- |

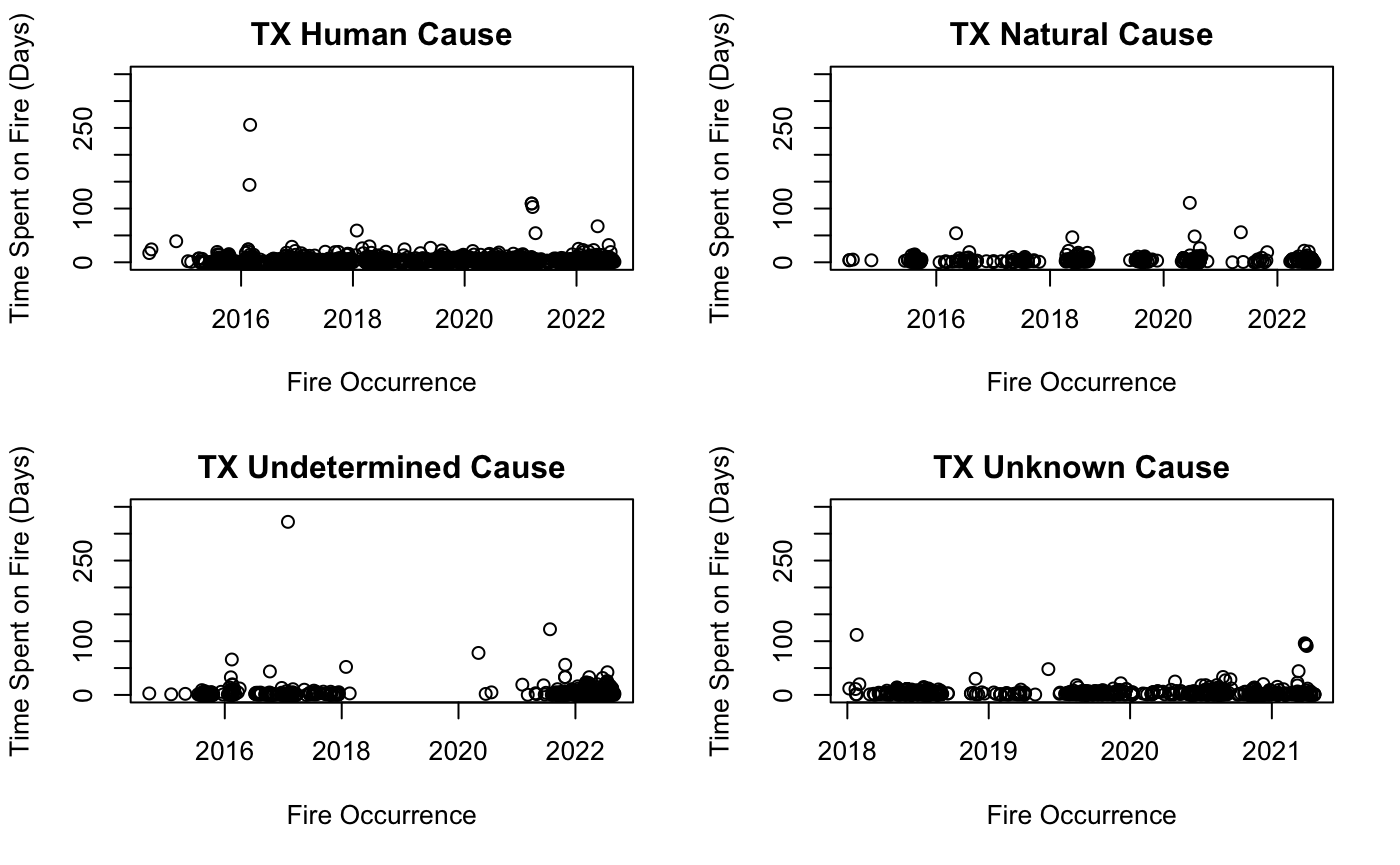
1. **General Data Analysis**

The important part of our general/exploratory data analysis is to get a baseline idea of what kind of data we’re working with, how it is distributed in terms of time and space, and what initial results we can extract using visualization and heat mapping. First off the important variable we really want to pay specific attention to in all of the datasets is the **Time Spent Putting out Fire (Days)** variable. Initially, we chose 3 states (California, Texas, Minnesota) to show three different environments and how long it takes to put out fires in each of those different locations. We first showed the density of each state’s time to put out fires over the time in which each fire was discovered (~2012-2022):



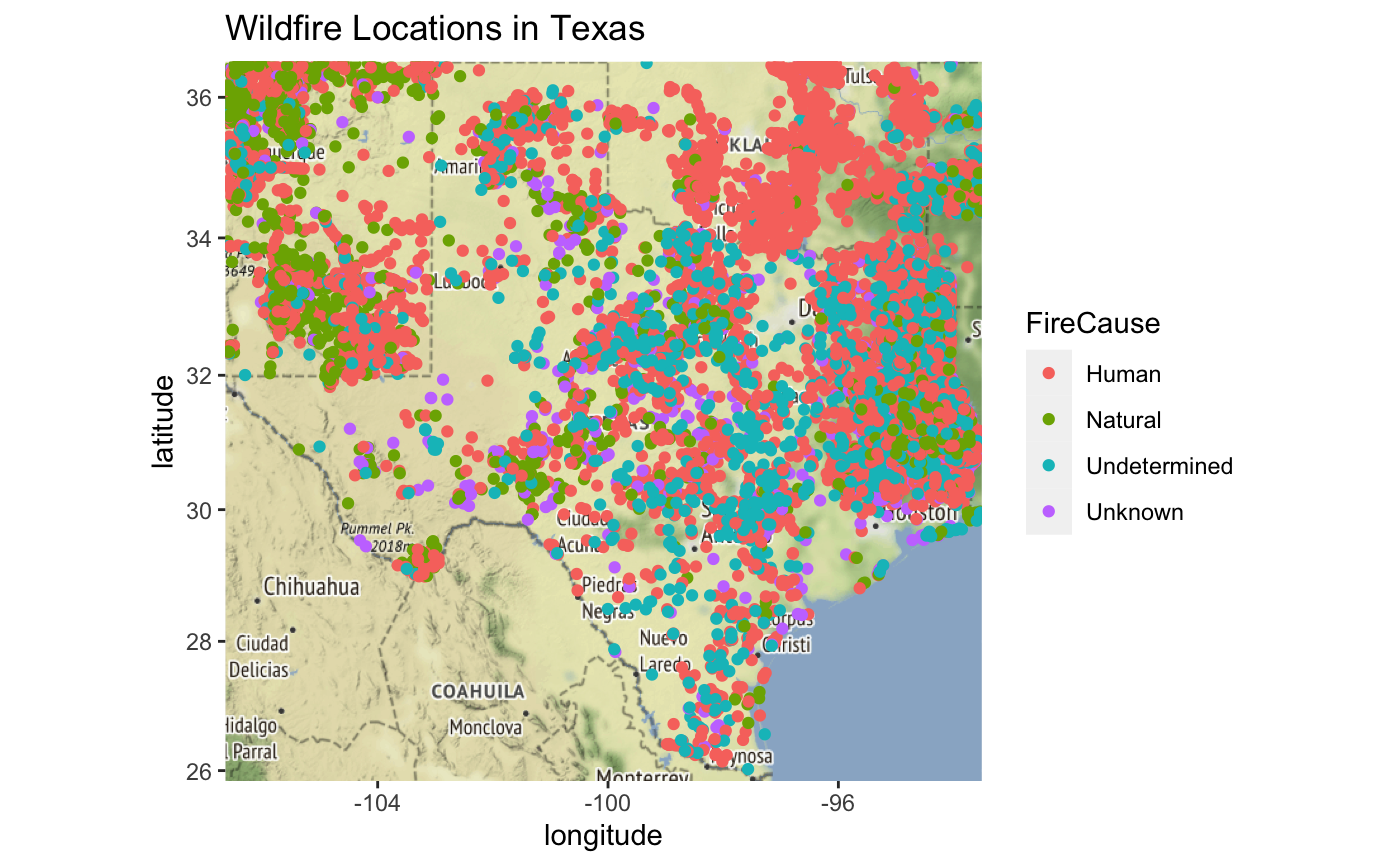
Our initial thoughts on the graphs above were that Texas seems to have the best ability to put out fires quickly (at least in this sample size) compared to California (CA) and Minnesota (MN). Both MN and CA seem to have a sort of seasonality where every year at some point, the amount of wildfires is far higher than any other time in the year. This makes sense to our general knowledge that CA has always had a big natural wildfire problem in the summer months (May-August). MN on the other hand also seems to have the problem of an increasing variation in time spent on putting out the fires, clearly the amount of fires is increasing over time which may be causing the local fire departments to strain in terms of getting people to put out numerous fires at the same time.

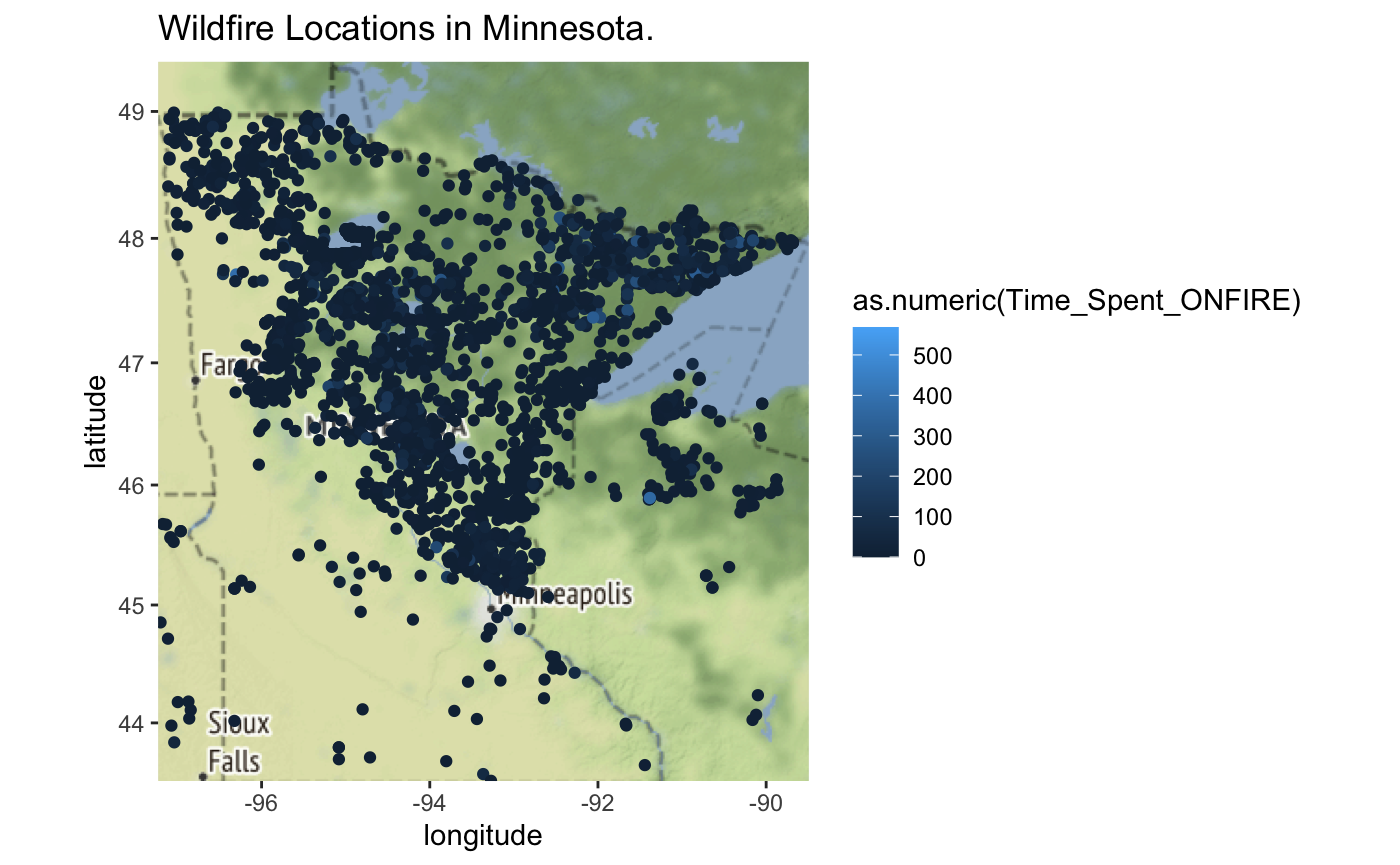
The second part of the density analysis is to look at another important variable in our dataset which is the **Fire Cause**. The plots below show the different densities of fire causes **(Human, Natural, Undetermined, and Unknown)**, and their respective times spent on putting out that type of fire. Again, we included the three states from the previous graphs: 

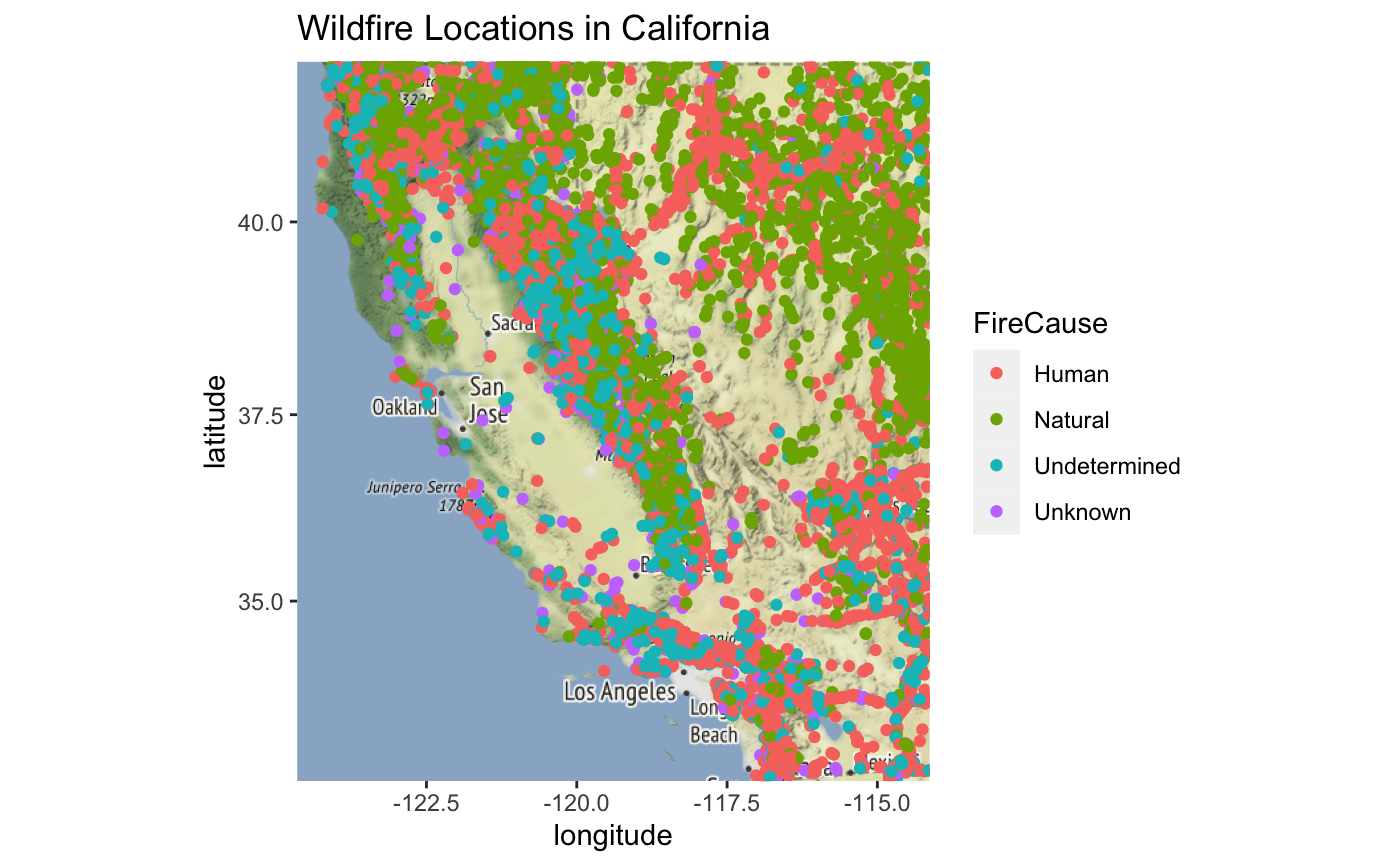
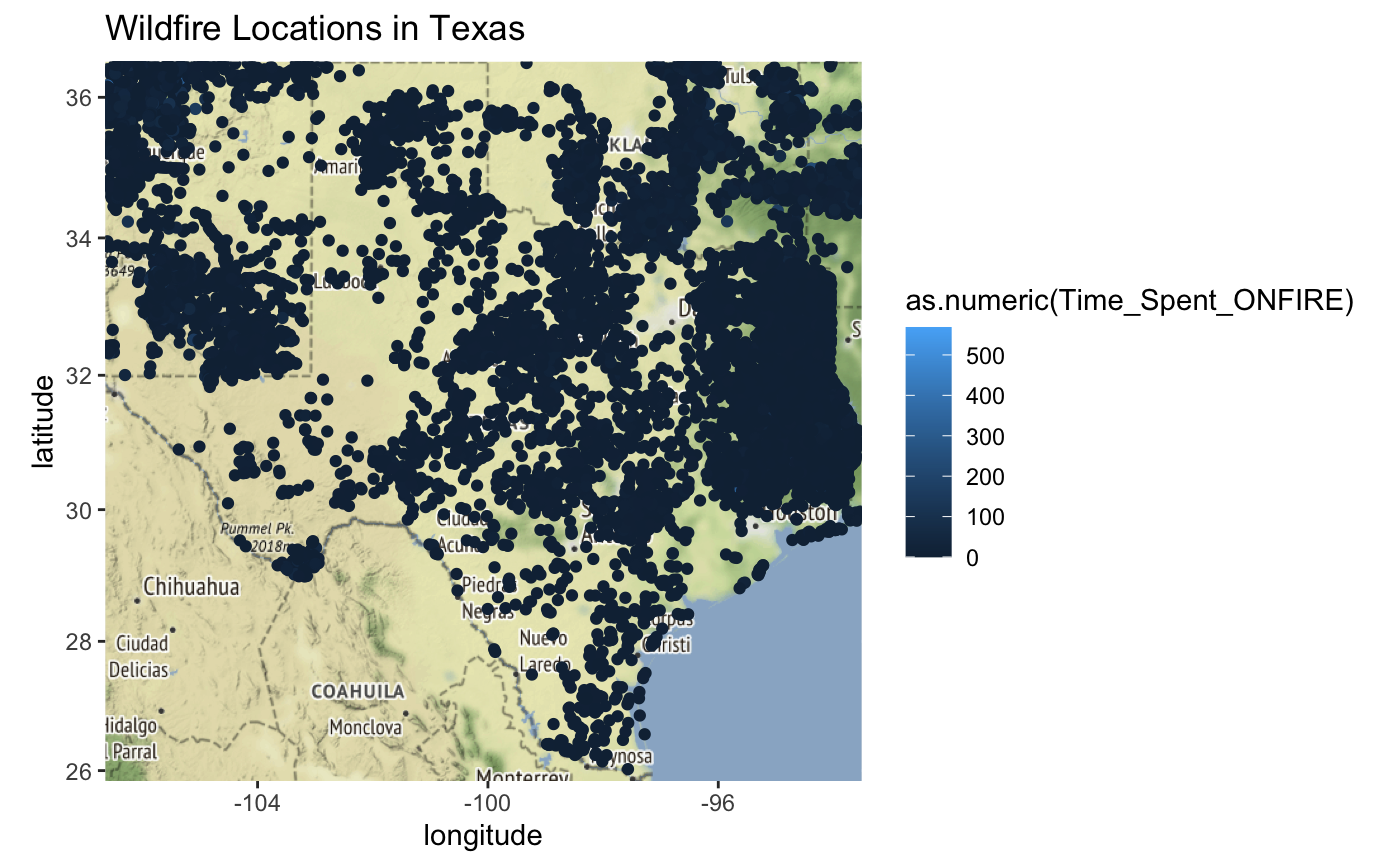


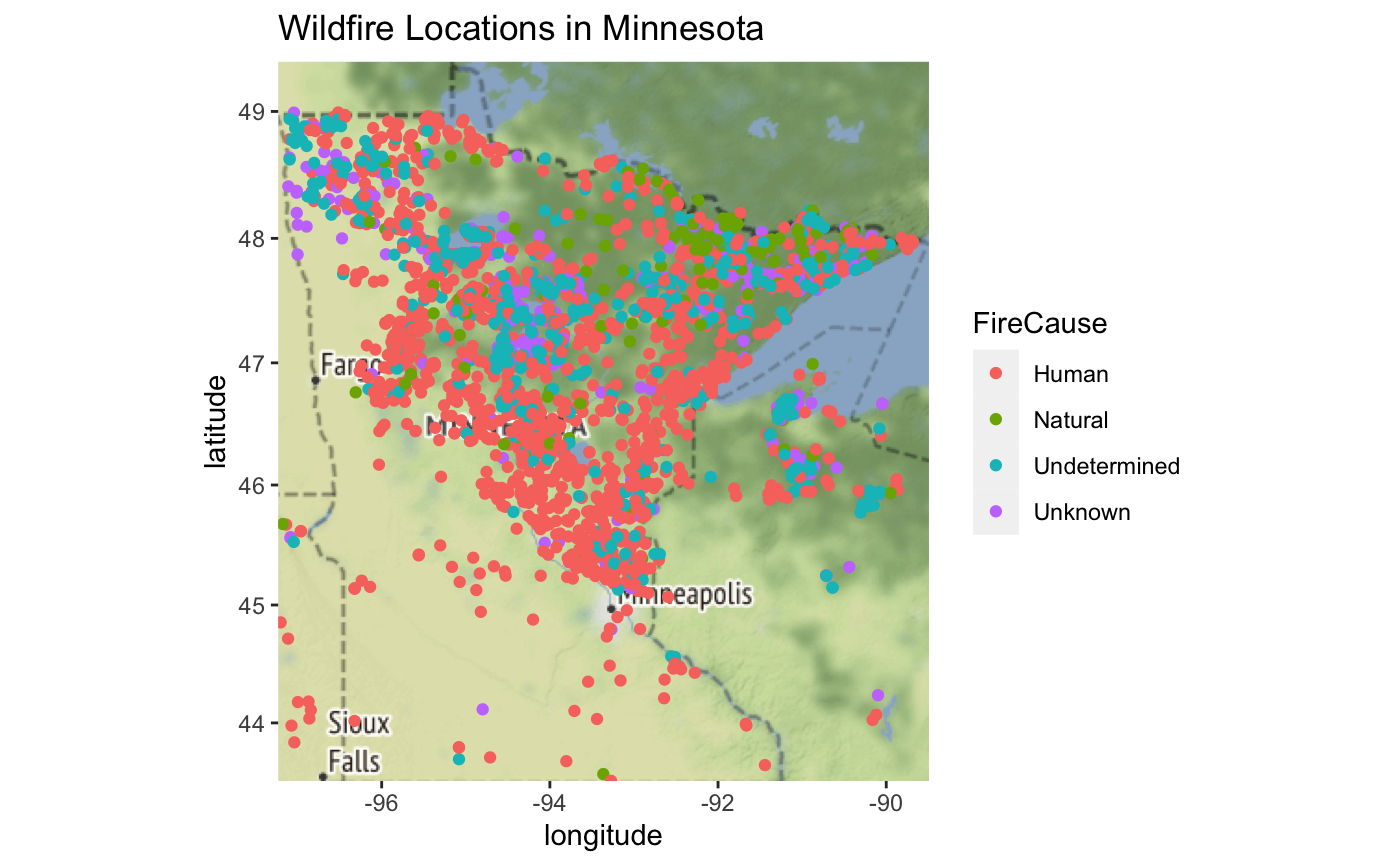
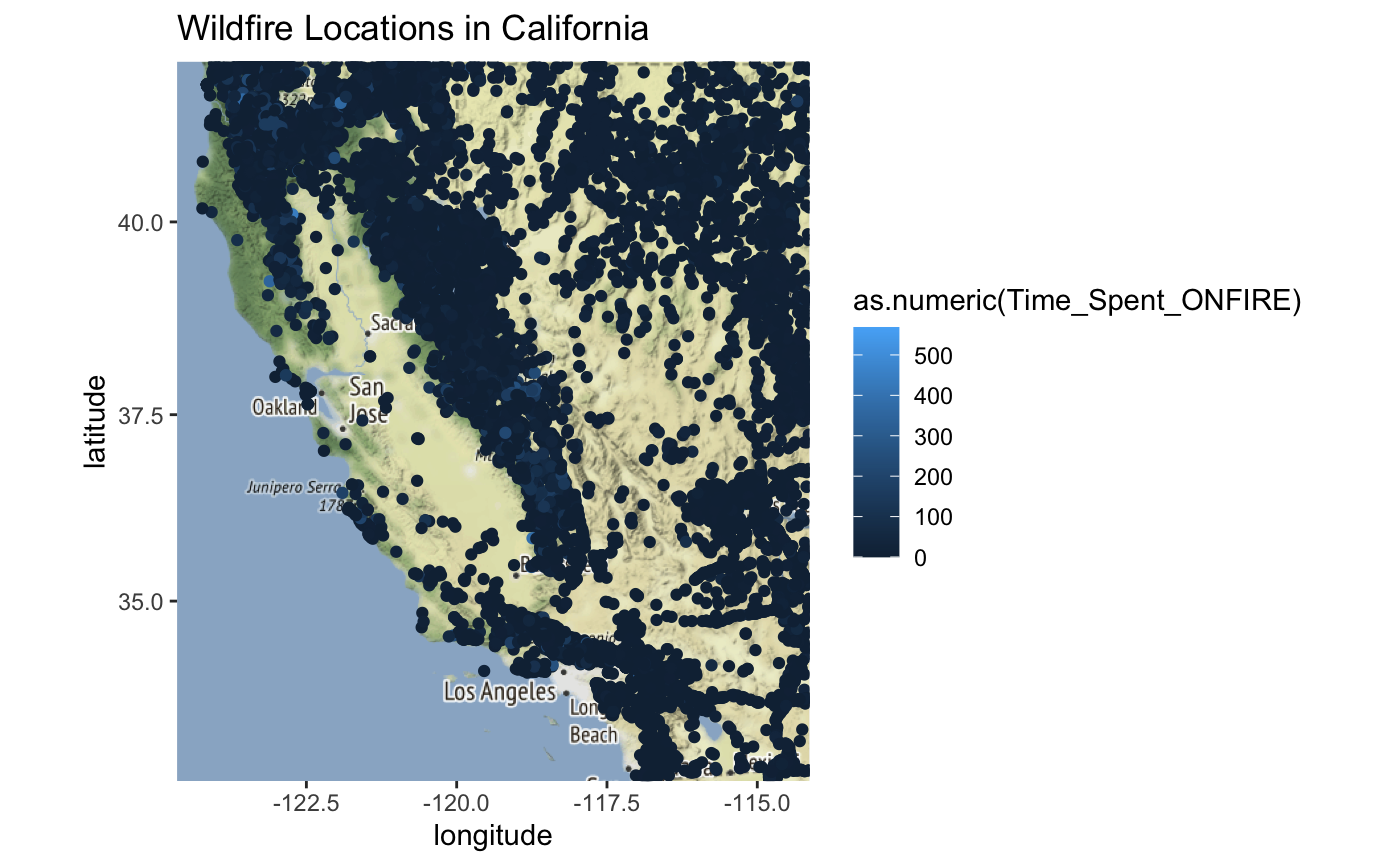
We can clearly see here that human and natural fires are the main causes of wildfires in the dataset, which is what we will choose to look at more clearly in our modeling and further analysis. As of now, we do not have enough information on undetermined/unknown fire causes to justify their results, so we will hold off on adding them into the model. If anything changes on that front and we do get some better information, they will be added in if needed.

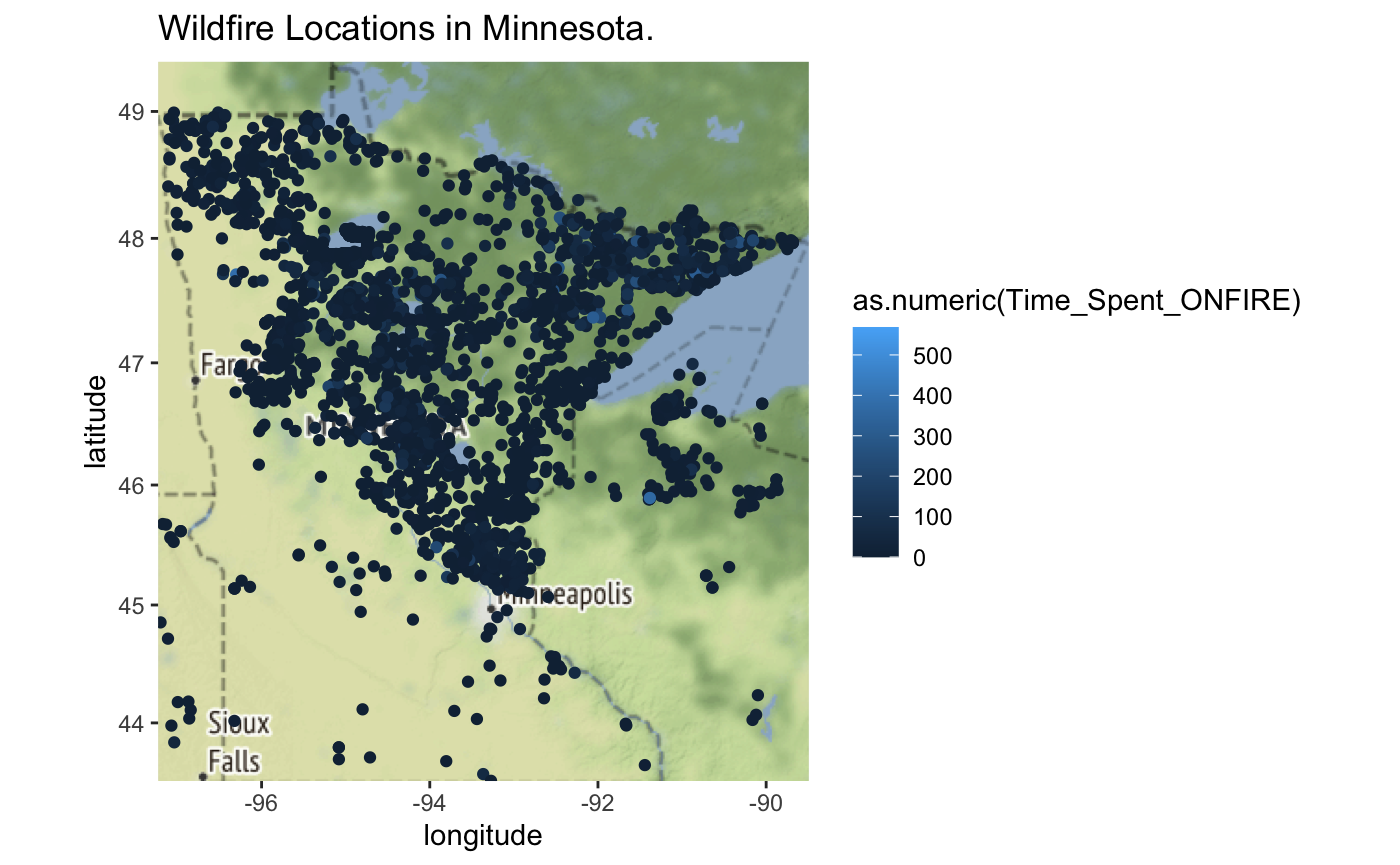
The second part of our general data analysis would be to find the spatial distribution of our data in terms of wildfires. The idea here was to simply visualize how and where each fire cause occurs and how and where the more intense fires are distributed spatially. These are the graphs we extracted for each of the three states below:











We can see how each state has its own unique terrain/environmental effects on where each type of fire might occur. This gives us enough information to go into modeling on how we can maybe predict the time spent on a fire based on its location (longitude and latitude) and possibly what the cause may be. For both models, we will continue using only the Texas data for all three datasets to simplify the strain.

**Data Modeling**

In the modeling part, we use a Gaussian Process Spatial Regression model. The objective is to predict how much time the fire will last in a specific location (S). The below figure shows the model formulation and variables.

**A picture containing text, night sky

Description automatically generated**











For W(S), we use: Gaussian Process with mean = 0 and covariance = 𝝈^2 \* 𝙥 (r), where 0 < 𝙥 (r) < 1

1. We estimated 𝝈^2: applying a simple linear regression model (fit.lm) not considering the data as a spatial data and take the variance of the output
2. We estimated 𝙥 (r): By applying exponential covariance function with the following parameters:
   1. r = The distance between two locations
   2. 𝜙 = Spatial range parameter estimated as (the maximum of distance matrix / 3 \* 2)

Text

Description automatically generated

**Implementation**

We design the model using R-studio. We load necessary libraries needed to apply the model. We choose our data set to be the natural fire in the state of Texas with total data points of 530.Then, we split the data set into 80% training and 20% testing. Then we plot the data in a spatial coordinate for a visualization check as follows:

**Chart, scatter chart

Description automatically generated**

After that we fit the spatial model using the following code:

fit\_mle <- likfit(data=Y.train,coords=coords.train,

trend = ~ X.train[,1]+X.train[,2],

fix.nugget=F,cov.model="exponential",ini = c(ini.sigma.sq,ini.phi))

**Results**

The result of fitting the model comes as follows:

## likfit: estimated model parameters:

## beta0 beta1 beta2 tausq sigmasq phi

## "-6.0596" "-0.2497" "-0.4386" "53.6561" " 0.0000" " 0.0000"

## Practical Range with cor=0.05 for asymptotic range: 0.0001159668

##

## likfit: maximised log-likelihood = -1446

We have estimated for the covariant variables (Lon,Lat) and also we construct the confidence interval to measure the significance of the variables and this is the result :

## lower.bond upper.bond

## intercept -18.5885878 6.4694772

## covar1 -0.3611500 -0.1383101

## covar2 -0.6549354 -0.2222032

So since the 0 is not in the range of the confidence interval for the two covariant variables hence we can conclude that the coordinates (Lon,Lat) are statistically significant to the model and the model has a spatial dependence.

**Validation**

To validate how good our model is, we utilize the model to predict on the testing data Then compare it with the real data and then calculate the Root mean square error (RMSE) as shown in the following code:

*# Doing prediction on the testing data*

krigecontrol=krige.control(type.krige = "OK", trend.d = ~ X.train[,1]+X.train[,2], trend.l = ~ X.test[,1]+X.test[,2], obj.model = fit\_mle)

prediction<-krige.conv(data=Y.train,coords=coords.train,locations=coords.test,krig=krigecontrol)

*# Calculating the root mean square error between predicted VS real*

sqrt(mean(prediction$predict-Y.test)^2)

## [1] 0.1043633

As we can see, the RMSE is 0.10 which is very low which means our model performed well in predicting the real values.

**Second Model (GAM - Generalized Additive Model)**

* Showing initial densities of the raw data compared to the GAM prediction densities.
* Talk about flexdashboard instead of R shiny for better utilization, R shiny having problems with ggplot() output.
* Show basic images from the slides
* Provide link to flexdashboard URL
* For the report, add in screenshots for each with small explanations (1-2 sentences).
* Come to the conclusion that both space and time have a clear effect on prediction, but we can clearly see that time has a larger impact visually, spatial effect has a much more analytical effect
* Include tons of visuals for both report and slides.

We are thinking of doing an interactive visualization on the model above using the R-Shiny which has recently implemented the use of Python code to push out visualization apps. Our hope is that any user can use the link to interact with the visualization which will act as a sort of infographic on our model and general data analysis. The below image is an example of how we would like our visualization to look at launch:

We think having a concise product like an R-Shiny app would be very informative and clean for the consumer. The process of tweaking and updating the app is quite simple too which means we could also concurrently change the visualizations with up-to-date data for all three datasets.

The other model we have that is still currently being tested out and fine-tuned is our spatial-temporal model. This is the model we speak about in the above sections that will essentially give us an idea of how the data trends over time and location and see which aspect (time or space) has a larger impact on the time spent putting out wildfires. This is in the early stages of development but should be done and applied to our “R-Shiny” app implementation once it is completed. It will hopefully be a combination of our general density plots like the one below and a time series plot similar to the first graphs we showed covering time density.