# Project Wildfire

CSPB 4502 - Fall 2022

Natalie Dreher
Applied Computer Science
University of Colorado Boulder
Boulder, Colorado, USA
nadr7654@colorado.edu

Ronald Durham
Applied Computer Science
University of Colorado Boulder
Boulder, Colorado, USA
rodu4835@colorado.edu

Grant Fairbairn
Applied Computer Science
University of Colorado Boulder
Boulder, Colorado, USA
grfa5712@colorado.edu

**CCS CONCEPTS**

Information Systems - Data Mining

**ACM Reference format:**

Natalie Dreher, Ronald Durham, and Grant Fairbairn. 2022. Project Wildfire. In CSPB 4502 - Data Mining. Boulder, CO, USA, 6 pages.

## 1 Problem Statement / Motivation

We are utilizing a data set of 1.88 million wildfires in the United States occurring between 1992 and 2015. The data set includes attributes relating to discovery date, containment date, fire size, and geographic location. This raises interesting questions about whether certain areas are more or less likely to experience wildfires and whether certain regions are more effective in fighting existing fires. We can also explore whether wildfires have increased or decreased in number or severity over time. If we find a suitable source with wind and temperature data, we can look for correlation between the size and duration of wildfires with the recorded temperature or wind speed data at the time the wildfires began.

## 2 Literature Survey

Our dataset has been downloaded over 20,000 times, so there has been some prior work on this data. We can review the results of other contributors to develop new questions to pursue using the dataset.

There are also publicly available articles discussing trends in wildfire data over multiple decades to help refine our analysis. For example, Matthew Wibbenmeyer and Anne McDarris noted in their article “Wildfires in the United States 101: Context and Consequences” that the annual number of wildfires decreased between 1991 and 2020, but the area burned in those fires increased sharply, particularly in the western United States.[1]

The link between wildfires and heavy-wind conditions has also been studied. One recent study of autumn and Santa Ana wind-driven wildfires analyzed data regarding ignition causes (*e.g.*, powerline failures), windspeed, air temperature, and precipitation prior to fires involving Santa Ana winds.[2]

## 3 Proposed Work

Our team will be addressing the major topics of preprocessing, data cleaning, and normalization by reviewing the dataset to identify conflicts and correct them. We will remove attributes that are mostly full of null values and not able to be calculated from the rest of the data. These null values will cause troubles for our project and are not useful to us. Additionally, some of the attributes are recorded in formats that are unfamiliar or uncommon in our direct application. Some of the date attributes are recorded in what is known as Julian dates, number of days since January 1st, 4713 B.C. It would be beneficial to us to use a more modern format even though many astrological data centers use this format.

After thorough cleaning, we will then look to create our test functions to begin analysis of the data. The test functions will be used to find correlations between the attributes, create multiple types of visualizations, and to look to see if we can figure out the answers to our preliminary questions as well as draw new conclusions from the data. It would be beneficial for us to trim off a sample of the dataset to use in testing our functions in order to retain computing power while bugs are being worked through. Then we will be able to apply the functions to our entire dataset and establish the project as a whole.

At this point, we may decide to include more datasets that can add historical information on air temperature and wind magnitude to be able to draw even more conclusions from the original set. This may be necessary if we think the initial design for the project looks insufficient.

In order to evaluate our findings, once we have done our analysis, we can refer to prior works using the set. These prior works found that the amount of fires reported between 1991 and 2020 decreased but the amount of area burned in those fires sharply increased especially in the western states. We do hope to discover correlations that have not yet been uncovered yet!

## 4 Data set

**1.88 Million US Wildfires** - 24 years of geo-referenced wildfire records (Source: Forest Service Research Data Archive)[3]:

https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires?resource=download

This data set covers details about wildfires that occurred in the United States from 1992 to 2015. This includes core elements such as discovery date, fire size, geographical location, etc. The data has been collected from the reporting systems of federal, state, and local fire organizations across the country. It has been preprocessed to some extent in order to remove redundant records and to conform to the standards of the National Wildfire Coordinating Group (NWCG). The data includes 1.88 million geo-referenced wildfire records that represents a total of 140 million acres burned during the 24 year period. There are 50 unique attributes used to describe the dataset within the main fires.csv file.

Major attributes that we plan to consider from the data set are:

- NWCG*REPORTING*AGENCY = Active National Wildlife Coordinating Group (NWCG) Unit Identifier for the agency preparing the fire report (BIA = Bureau of Indian Affairs, BLM = Bureau of Land Management, BOR = Bureau of Reclamation, DOD = Department of Defense, DOE = Department of Energy, FS = Forest Service, FWS = Fish and Wildlife Service, IA = Interagency Organization, NPS = National Park Service, ST/C&L = State, County, or Local Organization, and TRIBE = Tribal Organization).
- FIRE_NAME = Name of the incident, from the fire report (primary) or ICS-209 report (secondary).
- FIRE_YEAR = Calendar year in which the fire was discovered or confirmed to exist.
- DISCOVERY_DATE = Date on which the fire was discovered or confirmed to exist.
- DISCOVERY_DOY = Day of year on which the fire was discovered or confirmed to exist.
- DISCOVERY_TIME = Time of day that the fire was discovered or confirmed to exist.
- STAT*CAUSE*CODE = Code for the (statistical) cause of the fire.

- STAT*CAUSE*DESCR = Description of the (statistical) cause of the fire.
- CONT_DATE = Date on which the fire was declared contained or otherwise controlled (mm/dd/yyyy where mm=month, dd=day, and yyyy=year).
- CONT_DOY = Day of year on which the fire was declared contained or otherwise controlled.
- CONT_TIME = Time of day that the fire was declared contained or otherwise controlled (hhmm where hh=hour, mm=minutes).
- FIRE_SIZE = Estimate of acres within the final perimeter of the fire.
- FIRE*SIZE*CLASS = Code for fire size based on the number of acres within the final fire perimeter expenditures (A=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres).
- LATITUDE = Latitude (NAD83) for point location of the fire (decimal degrees).
- LONGITUDE = Longitude (NAD83) for point location of the fire (decimal degrees).
- OWNER_CODE = Code for primary owner or entity responsible for managing the land at the point of origin of the fire at the time of the incident.
- OWNER_DESCR = Name of primary owner or entity responsible for managing the land at the point of origin of the fire at the time of the incident.
- STATE = Two-letter alphabetic code for the state in which the fire burned (or originated), based on the nominal designation in the fire report.

## 5 Evaluation Methods

We will start by making several correlations: The size and number of fires with the geolocations of the fires. This correlation can be used to find areas that are more fire prone and/or are more likely to have larger fires. The geolocations will likely have to be binned longitude and latitude values. The size and number of fires with the date of the fire's discovery. This can be used both to find which dates in a year have more and larger fires as well as to show year over year trends for wildfires. The size of a fire with the fire's ignition cause can be used to find if certain fire causes are more likely to create larger fires. Beyond individual correlations we can perform a regression analysis to find the likelihood of fires happening in specific locations given specific inputs. This can be further expanded if we can add weather data to give more inputs to our model. From this analysis we can use R-squared values to test how well our models fit our data.

## 6 Tools

The tools to we are planning to use on this project include:

- Jupyter Notebook
  - For coding
- Pandas, NumPy, scikit-learn, statsmodels
  - Data processing and analysis
  - Regression analysis
- Matplotlib, Seaborn
  - Visualization

## 7 Milestones

a. Milestones Completed

Our goal was to have the dataset cleaned and be in the process of analyzing the data for patterns by November 28, 2022 (the due date for Part III for this project). We also planned to have identified and prepared a complementary set of data for wind speed and temperature data to allow us to compare the wildfire data with data about the weather conditions at the time and

allow us to create visual representations like heatmaps. We met the first milestone and made progress on the second.

We initially sampled the first 100 rows from the dataset to get a sense for the contents of the various attributes. We then split the dataset up into four files with 500,000 rows each and cleaned each subset to be ready to be recombined into a single dataframe.

In the process of cleaning the dataset, we learned that certain attributes (*e.g.*, Containment Date, Containment Time, and Discovery Date) were missing a significant amount of data. We found that we do have robust data, however, for several other attributes (*e.g.*, Fire Class, Fire Year, Stat Cause Code, Stat Cause Descr, State, Long, Lat, Fire Size, and Object ID) that we can use for our correlation analysis.

For the weather data, we are in the process of getting access to the NOAA historic climate data from the National Climatic Data Center and will be focusing on integrating that data into our dataset.

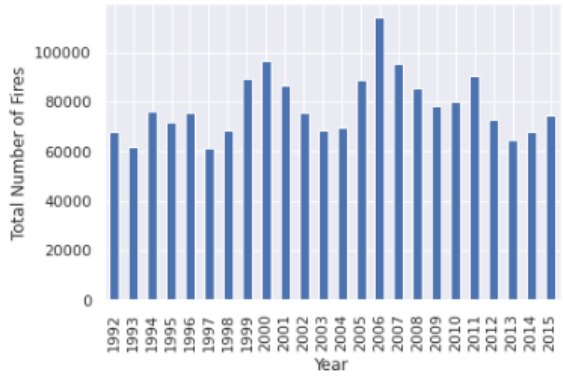
b. Milestones To Do

Now that we have the wildfire dataset cleaned and loaded into our Python file, what remains to be done is to bring in the weather data and run through our correlation analysis to look for interesting patterns.

Beyond weather data we would like to run some of the same tests we've already completed on a subset of the wildfire data containing only what are considered large wildfires, which according to the National Wildfire Coordinating Group is larger than 300 acres. This can be used to show if the changes if fires that are considered large are similar to wildfires as a whole or if large fires are being affected differently by the changes in the climate.

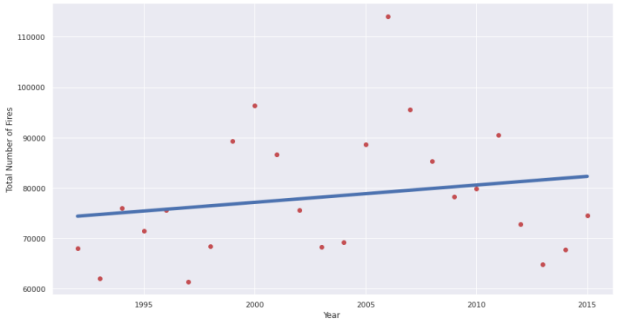
## 8 Results So Far

After cleaning the data and loading the dataset into Pandas dataframes, we were able to start visualizing some of the quantitative attributes using Matplotlib and Seaborn. First, we looked at the number of wildfires per year from 1992 to 2015 to see if there was any obvious trend in the total number of fires per year.




On visual inspection, there does not appear to be a trend either way, at least with respect to the total number of fires.

We created a linear regression model with the total number of fires as the dependant variable over the Fire Year column, and the regression line suggests a slight positive correlation:




That is deceptive, however, because the scatterplot is diffuse and the regression line does not model the data very well. This is confirmed by looking at the regression results:

OLS Regression Results

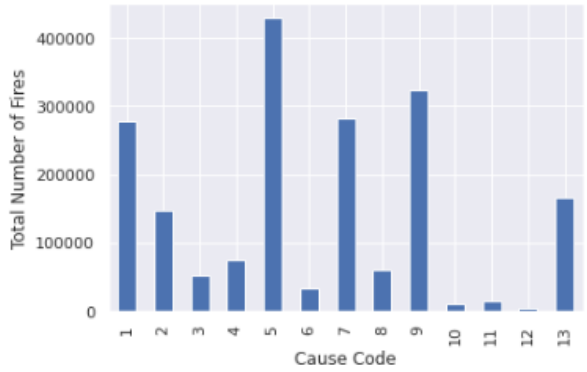
| | | | |
|---|---|---|---|
| **Dep. Variable:** | Fires | **R-squared:** | 0.036 |
| **Model:** | OLS | **Adj. R-squared:** | -0.008 |
| **Method:** | Least Squares | **F-statistic:** | 0.8265 |
| **Date:** | Tue, 29 Nov 2022 | **Prob (F-statistic):** | 0.373 |
| **Time:** | 02:55:59 | **Log-Likelihood:** | -260.00 |
| **No. Observations:** | 24 | **AIC:** | 524.0 |
| **Df Residuals:** | 22 | **BIC:** | 526.4 |
| **Df Model:** | 1 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -6.095e+05 | 7.57e+05 | -0.806 | 0.429 | -2.18e+06 | 9.6e+05 |
| **Year** | 343.3443 | 377.678 | 0.909 | 0.373 | -439.911 | 1126.600 |

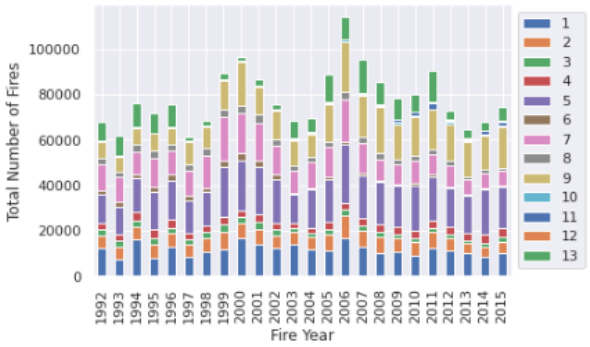
| | | | |
|---|---|---|---|
| **Omnibus:** | 5.718 | **Durbin-Watson:** | 0.925 |
| **Prob(Omnibus):** | 0.057 | **Jarque-Bera (JB):** | 3.859 |
| **Skew:** | 0.938 | **Prob(JB):** | 0.145 |
| **Kurtosis:** | 3.585 | **Cond. No.** | 5.80e+05 |

The R-squared value is very low, at 0.036. The p-value is much higher than the 0.05 level for significance. And the confidence interval includes 0. Thus, the Fire Year does not appear to be a useful predictor of the number of wildfires.

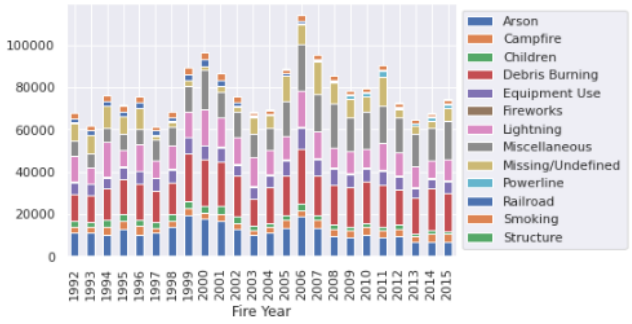
Next, we looked at the number of fires per Causation Code and saw that code number 5 (“Debris Burning”) resulted in the greatest number of fires during the relevant time period:




We then sought to visualize the Stat Cause Code attribute by year using a stacked bar graph:




In order to make that graph more interpretable, we prepared a similar version using the Stat Cause Descr attribute:

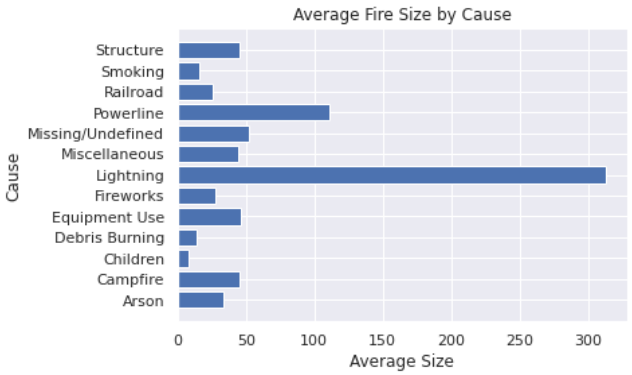



Notably, the colors for the Stat Cause Code numbers in the first stacked bar graph do not line up with the corresponding Stat Cause Descr names in the second. For example, the Stat Cause Descr “Lightning” (labeled pink in the second graph) corresponds to Stat Cause Code 1 (labeled blue in the first graph). This discrepancy arises from the fact that the legend in the second bar graph is sorted in alphabetical order, whereas the legend in the first bar graph utilized numerical order. The overall shape of the graphs is the same, because the fire counts across all causes within a year are the same no matter which order the causes are added to the stack.

With these stacked bar graphs, we can see that some causes are more frequent than others year

over year. The “Debris Burning” cause (Stat Cause Code number 5) is not only the largest culprit by aggregate number of fires during the relevant time period, it also appears to be the largest cause in most individual years as well.

Another thing we wanted to look at was what plays a part in the average size of a fire. The first step was to calculate the average fire size for each individual ignition cause. This shows “Lightning” as the cause of the largest fires on average by a fairly large margin. This fact could lead to some interesting speculation as to the cause of the disparity of the average wildfire size between “Lightning” and all other causes of ignition.




## REFERENCES

[1] M. Wibbenmeyer, A. McDarris. Wildfires in the United States 101: Context and Consequences. Explainer, Resources for the Future (2021). https://www.rff.org/publications/explainers/wildfires-in-the-united-states-101-context-and-consequences/

[2] J. Keeley, J. Guzman-Morales, A. Gershunov, A. Syphard, D. Cayan, D. Pierce, M. Flannigan, T. Brown. Ignitions explain more than temperature or precipitation in driving Santa Ana wind fires. Science Advances. Vol 7, Issue 30 (2021). https://www.science.org/doi/10.1126/sciadv.abh2262

[3] Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPA*FOD*20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4