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, 10 pages.

ABSTRACT

While the natural phenomenon of wildfires is not a new subject, our interaction with them as humans is majorly increasing as we continue to populate once wild and open areas of nature, introducing additional potential causes of fires to our environment. If we had an incredibly large dataset of wildfire recordings dating back to the industrial revolution, we might see a correlation between the human population and number of wildfires. Our data is a bit more localized. We are looking at wildfire data from a 23 year timeframe within the more recent timeline of our existence.

As part of our analysis, we sought to answer the following questions:

1. Have wildfires become more or less frequent over time?
2. In this dataset, what is the most common cause of fires?
3. Which cause of fire most commonly leads to the largest fire size?
4. Geographically, where are the most common places for fires to occur?

Briefly, we have found that the number of wildfires over this time period was not significantly correlated. Generally, debris burning is the most common cause of fires, providing a significant human cause. But the cause that lead to wildfires consuming the largest area of land was by far was lightning strikes. We also created an interactive map showing the wildfire activity. Depending on where you zoom in on the interactive map, fires seem to occur mostly where people live.

INTRODUCTION

Our first question, relating to frequency of fire occurrence over time, was thought to be incredibly important. From our immediate perspective, it seems that we as a nation are having wildfires and other massive environmental events more often as time progresses and there are many articles and research projects dedicated to exposing the causes and effects of global environmental changes. By looking into whether or not there has been a significant increase in the number of fires over this 23-year period, we had hoped to be able to expose the possible correlations.

We did not observe that correlation. It could be that our dataset is too recent and does not cover a long enough period to see a major trend. It

would be interesting to compare our sample from 1992 to 2015 to another time range, say 100 to 200 years ago. But, then matters of data recording and other potential areas for data misrepresentation could arise. For matters of this dataset, we decided to narrow our focus to this time range to see if there were other recent trends worth noting.

Our second inquiry, with regards to the most common cause of fires in this timeframe, related to causes of wildfires. This has the potential to be able to expose reasons why fires occur and possibly whether or not certain behaviors by humans are the cause of them. Out of all of the causes, one is non-human related (lightning) and another two are rather vague (debris burning and miscellaneous). We had hoped to be able to see if there were correlations between specific industries or criminal activity that lead to an increase in fires throughout the United States. This would have been an important question to answer as it could have led to a deeper understanding of why fires are caused and what we can do to better prevent and contain them.

Our third question was in line with the second but concerned more specifically which cause led to the largest burned areas. It would seem that smaller fires often go unreported when compared to larger fires. Larger fires would be thought to have a more dramatic effect on the environment and our human actions. For instance, there might have been a small fire in a local business caused by arson but the majority of the US population would have no knowledge of this occurrence. However, large fires that burn thousands of acres affect many people's livelihood and can change the course of nature: eliminating species, altering rain and river routes of large areas of terrain, and majorly affecting air pollution levels geographically. These large fires are significant and understanding what the most common causes of them are is important to being able to avoid damages across the board.

Our final question with regards to the most common place for fires to occur geographically within the United States also provides additional insight. This information can be used to simply pick a place to live that is generally free from wildfires, to determine further correlations between why fires occur in those locations, and to study the after effects on wildlife in those habitats.

RELATED WORK

Our dataset has been downloaded over 20,000 times, so there had been some prior work on this data. We were able to review the results of other contributors to look for new areas to analyze and explore or to question conclusions reached by others using this dataset.

Interest in wildfires has been high in recent years, particularly with respect to its interplay with climate change. There are 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] Notably, we did not observe a decrease in the annual number of wildfires during our covered period (1992-2015), as shown below.

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]

And wildfires have been discussed extensively in the popular press. In a New York Times article

from 2020 entitled “In the West, Lightning Grows as a Cause of Damaging Fires”, John Schwartz and Veronica Penney wrote about the same dataset that we analyzed and commented on the role that climate change may play in fueling lightning-strike-caused wildfires.[3] The authors assert that 44 percent of the wildfires across the United States were triggered by lightning, but those wildfires were responsible for 71 percent of the area burned during the 1992 to 2015 time period covered by the dataset.[3] The article quotes a fire expert (Park Williams) who found that between 1992 and 2015 there was a nearly fivefold increase in Western forest area burned from lightning-caused fires compared to a twofold increase for wildfires started by a human cause.[3] The article continues on to discuss some possible reasons for the discrepancy, hypothesizing that human-caused fires may occur closer to inhabited areas where they are dealt with quickly and not allowed to spread, versus lightning strikes in remote areas that may cause a blaze that lasts for a longer period before detection.[3] The more controversial assertion, which is beyond the scope of our analysis, is that climate change is a major factor in the growing impact of lightning strikes. The authors of the New York Times article posit that areas of the West are becoming more dried out: “A lightning fire that might not have spread so quickly decades ago leaps across the landscape of dry vegetation.”[3]

These are interesting areas for further research. The dataset covers only a 23-year period already within a period where climate change seems likely to have occurred. Without a longer or earlier time period to compare with, it would be hard to determine from a relatively short time period that lightning-caused fires have become more devastating due to a drier climate than in prior decades.

DATASET

1.88 Million US Wildfires - 24 years of geo-referenced wildfire records (Source: Forest Service Research Data Archive)[4]:

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

This dataset 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.

The following were the major attributes that we reviewed in preparing the data for our analysis:

- 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.

MAIN TECHNIQUES APPLIED

The first obstacle to our analysis was the sheer size of the dataset. The tabulated data contained 1.88 million rows. This caused a number of problems with our coding environment, as the team was seeking to use Python within JupyterLab notebooks to analyze the data given the team's familiarity with those tools.

The JupyterLab notebook environment struggled to handle the data. The size of the dataset caused the Python kernel to crash repeatedly on all three team members' computers before we could conduct any analysis. Additionally, the large file size meant it could not be uploaded and shared between group members on GitHub. We first needed to find ways to reduce the size of the dataset and make the relevant attributes available to everyone in our group.

We first converted the sqlite file to csv format and isolated which individual file within the archive actually contained the relevant information. This file was still way too large to share on GitHub and analyze as a single file in JupyterLab.

From there, we separated the file first into just the top 100 rows so we could get our functions and visualizations started. Next we separated the large main file into 4 smaller files with roughly 500,000 rows in each. At that point, we were able to actually look deeper into the data and begin our cleaning and preprocessing the file.

One option for making the dataset more usable was to limit the time period to a subset between

1992 and 2015. That would reduce the number of rows and allow the entire csv file to be loaded into JupyterHub without crashing the Python kernel. But that approach would have reduced the ability to track data over time by narrowing the window to a smaller number of years. As such, we decided to focus on dimension-reduction strategies instead to decrease the size of the dataset for importation and analysis in our coding environment.

During preprocessing, we found that several of the columns were missing enough data to be generally incomplete. That made them obvious candidates for dimension reduction. We ultimately decided to remove the following columns that were not useful given the lack of data for the entire set: DISCOVERY_DATE, DISCOVERY_DOY, DISCOVERY_TIME, CONT_DATE, CONT_DOY, and CONT_TIME.

In addition, we found that the dates were recorded as Julian dates (i.e. the number of days since January 1st, 4713 B.C.) This would have been great to be able to work through the project, looking at the discovery and containment information. While this would have involved converting Julian dates into something more interpretable, it became a moot point because a significant number of the rows were missing Julian dates. Thus, we decided to remove these columns from our working dataset as well.

Beyond missing data, there were a number of columns that contained classifiers that were not important to any analysis we planned to conduct and thus could be omitted from the working dataset. For example, the attributes NWCG*REPORTING*AGENCY, FIRE_NAME, OWNER_CODE, and OWNER_DESCR did not add to our analysis. Our questions were not quite aligned with the information in these columns and with the limitations on memory and processing in the kernel, we chose to remove them as well.

This left us with a more workable solution, a trimmed-down version of the original dataset still containing the full range of entries from 1992 to 2015. We ultimately retained and analyzed the following attributes: FIRE_YEAR, STAT*CAUSE*CODE, STAT*CAUSE*DESCR, FIRE_SIZE, FIRE*SIZE*CLASS, LATITUDE, LONGITUDE, and STATE.

With a smaller dataset after preprocessing, we were able to import the four trimmed-down csv files into our coding environment and combine them into Pandas DataFrames to analyze. This allowed us to work as a group on the dataset.

That is not to say it was seamless. By importing four files into Pandas DataFrames and concatenating them first into two DataFrames and finally into one, we created duplicates of the full dataset multiple times within our Python code, quickly gobbling up the available memory. The kernel would often crash midway through our code when the memory was exhausted. We eventually solved that problem with garbage-collection measures that deleted unused DataFrames behind us as we progressed through the code.

Eventually, we were able to use the Pandas DataFrames to isolate specific attributes regarding our questions above. We then set out to visualize the attributes using Matplotlib, PyPlot, and Seaborn in order to look for obvious trends in the data. Once we had an idea of what we were looking for (and discovered some unexpected results like the average size of lightning-caused fires), we set out to look for significant correlations between attributes using Python's statsmodel linear regression tools.

KEY RESULTS

In our analysis, we found no significant correlation in the amount of fires per year in this time frame. As mentioned earlier, this could be because we are only looking at a small snippet of

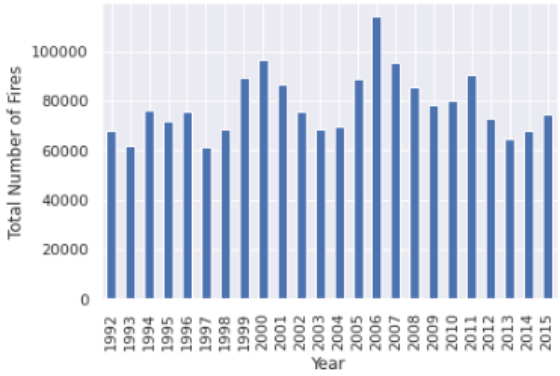
the entire time frame of the human population. It is notable, however, because a prior study of a slightly longer period (1991-2020) that encompassed our period (1992-2015) found a decrease in the annual number of wildfires.[1] As shown below, our analysis suggested a slight upward trend, although the scatterplot was diffuse and the regression results suggest that any correlation is insignificant.

Still, this difference between our results and the prior study is interesting because it shows how adding one year to the front end and five years to the back end of the data can change the conclusions drawn from the data.

The lack of correlation in our analysis could also suggest that the interaction of humans on the earth has a negligible effect on the number of annual fires, as explored by our analysis of fire causation. We found that out of all of the causes of fires, lightning strikes were by far the most common. More data would be needed as well to look further into seeing correlations between possible increases in lightning storms but for this set, lightning is the key cause of fires in the US, specifically for fires that end up burning the largest area.

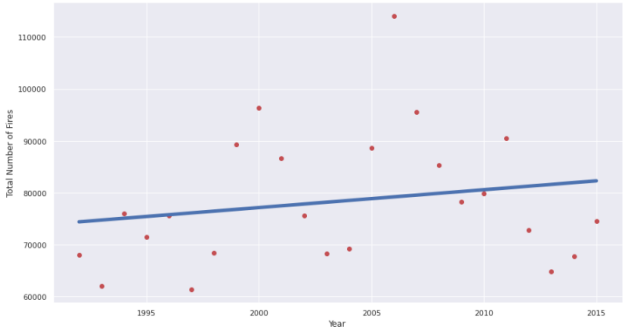
After cleaning the data and loading the dataset into Pandas DataFrames, we were able to start visualizing some of the quantitative attributes using Matplotlib and Seaborn.

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. Indeed, 2006 appears to be a bit of an outlier, as it is the only year in the dataset with recorded fires in excess of 100,000. If we ignore that year, the bar chart appears almost cyclical.

We created a linear regression model with the total number of Fires as the dependent variable over the Fire Year column. The scatterplot and linear regression line suggests that there may be a very slight positive correlation:



But that is probably not a reasonable conclusion. The scatterplot data (which mirrors the earlier bar chart) is diffused with an outlier in 2006, and the regression line does not model the data very well. This lack of fit between the model and the data is confirmed by looking at the regression results:

OLS Regression Results

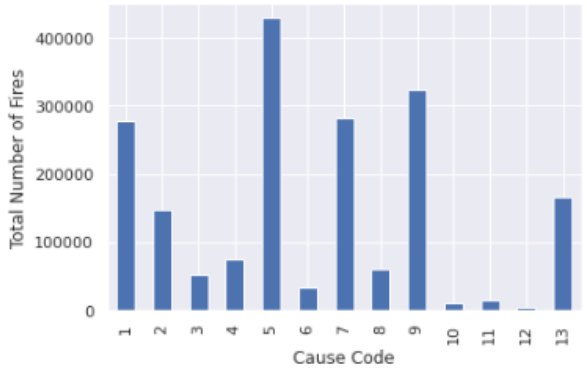
Dep. Variable:	Fires	**R-squared:**	0.036
Model:	OLS	**Adj. R-squared:**	-0.008
Method:	Least Squares	**F-statistic:**	0.8265
Date:	Tue, 29 Nov 2022	**Prob (F-statistic):**	0.373
Time:	02:55:59	**Log-Likelihood:**	-260.00
No. Observations:	24	**AIC:**	524.0
Df Residuals:	22	**BIC:**	526.4
Df Model:	1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

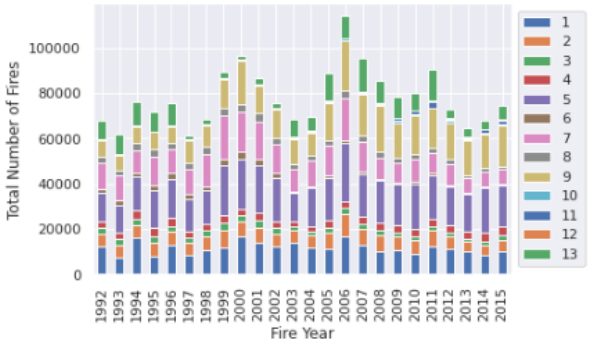
Omnibus:	5.718	**Durbin-Watson:**	0.925
Prob(Omnibus):	0.057	**Jarque-Bera (JB):**	3.859
Skew:	0.938	**Prob(JB):**	0.145
Kurtosis:	3.585	**Cond. No.**	5.80e+05

Specifically, the R-squared value is very low at 0.036. The t value is small, and the p-value is much higher than the 0.05 level for significance, suggesting that the null hypothesis that these variables are uncorrelated is likely correct. In addition, the confidence interval for the Fire Year coefficient includes 0 in its range (-439.911 to 1126.600). A coefficient of 0 would indicate that the Fire Year has no effect on Fires in this linear model. Based on this analysis, we did not see any indication that the Fire Year was a useful predictor of the number of wildfires.

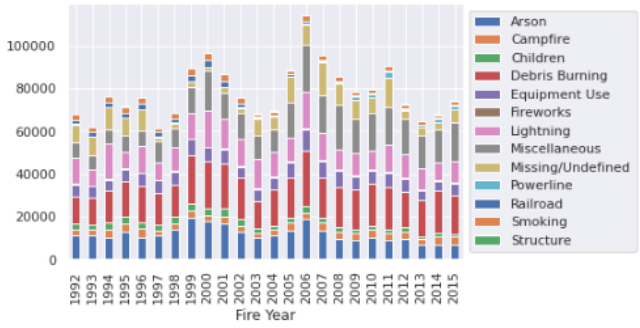
Next, we looked at causes of fire, using the Causation Code (“Stat Cause Code”) category. In visualizing the counts for each Stat Cause Code, we were able to clearly see that code number 5 (“Debris Burning”) resulted in the greatest number of fires during the relevant time period (1992-2015):



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



While those bar graphs presented interesting information, they were difficult to interpret without knowing what each code number meant. To address this issue, we prepared a similar version using the Cause Description attribute (“Stat Cause Descr”):

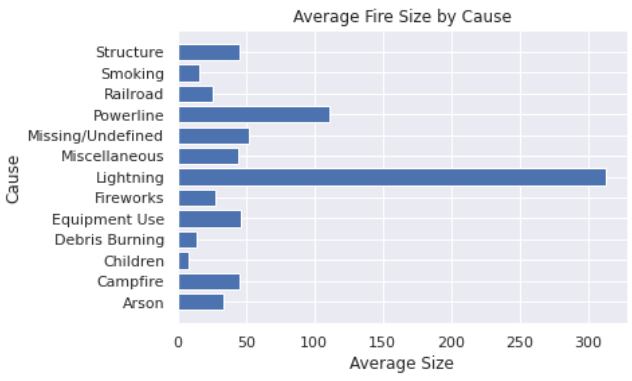


Notably, the colors for the Stat Cause Code numbers in the first stacked bar graph do not line up with the corresponding Stat Cause Descr names in the second. For example, the Stat

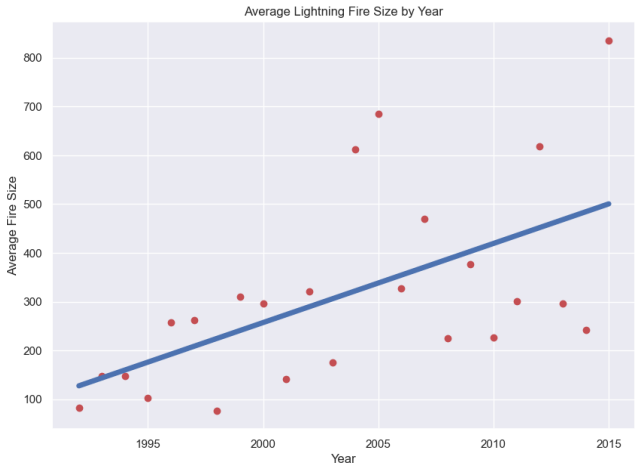
Cause Descr “Lightning” (labeled pink in the second graph) corresponds to Stat Cause Code 1 (labeled blue in the first graph). This discrepancy arises from the fact that the legend in the second bar graph is sorted in alphabetical order, whereas the legend in the first bar graph is arranged in numerical order. As one would expect, however, the overall shape of the graphs is the same. 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 in the first graph and red in the second) 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.

That led to our next question: which causes lead to the largest fires? 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.



To expand upon this discovery we decided to test if the average size of a fire caused by lighting has been increasing year over year. To do this we first calculated the average size of a lightning ignited fire for each year of our data. Next, we created another linear regression model with average lightning Fire Size as the dependent variable over the Fire Year column. Here the scatterplot and linear regression line created by this data shows a noticeable upward connection between lightning Fire Size and Year:



Here the fit between the model and the data visually noticeable and is confirmed by looking at the regression results:

OLS Regression Results

Dep. Variable:	Size	**R-squared:**	0.334
Model:	OLS	**Adj. R-squared:**	0.303
Method:	Least Squares	**F-statistic:**	11.02
Date:	Tue, 06 Dec 2022	**Prob (F-statistic):**	0.00312
Time:	14:59:32	**Log-Likelihood:**	-155.65
No. Observations:	24	**AIC:**	315.3
Df Residuals:	22	**BIC:**	317.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.217e+04	9786.229	-3.287	0.003	-5.25e+04	-1.19e+04
Year	16.2129	4.885	3.319	0.003	6.083	26.343

Omnibus:	2.779	**Durbin-Watson:**	1.975
Prob(Omnibus):	0.249	**Jarque-Bera (JB):**	2.090
Skew:	0.717	**Prob(JB):**	0.352
Kurtosis:	2.821	**Cond. No.**	5.80e+05

We see that the p-value of this model is quite small meaning that it's safe to reject the null hypothesis that these variables are uncorrelated in favor of the alternative hypothesis that there is indeed a correlation between Year and the Average Size of a lightning ignited fire. While the R-squared value is somewhat low it is enough above zero to show that there is indeed a measurable connection between Year and average lightning Fire Size. This likely points to the fact that the factors leading to a fire growing to a large size are complicated. Still, given what data we have it does appear that Fire Year is a useful predictor of the Average Size of Fires when the cause is determined to be lightning.

As discussed above in the Related Work section, two New York Times writers in 2020 posited that lightning fires cause more severe wildfires than in the past because of climate change and drier climates, particularly in the Western United States.[3] While this is an interesting hypothesis, it would not alone explain the much larger average fire size by cause shown above. If underlying conditions are promoting the spread of wildfires more today than in the past, it would seem that those same conditions would apply to any wildfire begun in a forested area, and not just lightning-caused fires. Perhaps there is an interaction between drier conditions from climate change and the remoteness of the fire origin, as fires caused by lightning strikes in isolated forest areas would have more time to spread before containment measures are advanced than a wildfire in a populated area. In that scenario, the wildfires in populated areas have as much fuel due to the dry conditions as in the isolated areas, but not the time to spread that isolation provides fires started by nature in remote areas. That would be an interesting area for further research.

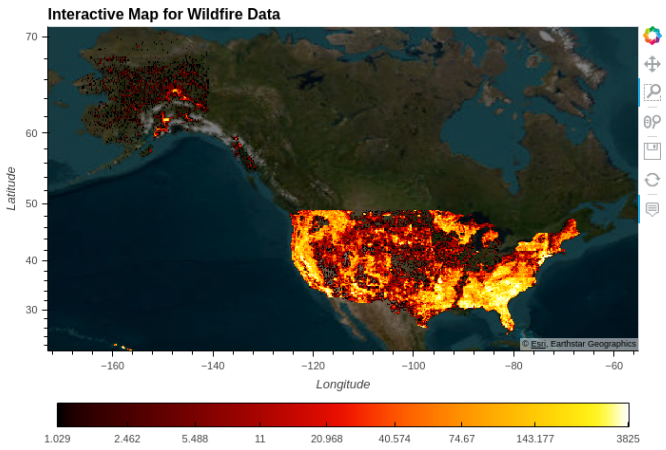
APPLICATIONS

Possibly the largest area of application of this dataset would be in learning about prevention and containment. Wildfires are actually a necessary part of our global ecosystem. They allow forests to clear out the deadfall and sick trees to allow for new growth. Our issue comes when we incorporate our lives deep within forests and become susceptible to the natural processes of nature. Wildfires are not something that should be viewed as being able to be eliminated completely but rather understood and worked with cooperatively. By identifying the common location and causes, we can position our homes and businesses outside of these zones and be prepared with barriers to protect the facilities within those zones.

The finding that lightning strikes cause the most severe wildfires is also useful in focusing efforts. Human-caused fires can no doubt be devastating. But locating fires started by lightning quickly and deploying containment measures before the wildfire spreads seems like a wise use of resources. As next steps, finding which locations within the United States have the highest percentage of lightning-caused fires could help wildfire prevention by focusing efforts on those areas.

VISUALIZATION

In order to better understand how the dataset relates to our fourth question, we decided to create an interactive visualization using the longitude and latitude data for the entire time frame along with the count of fires for that specific location. We have included the Python file that creates this visualization in our GitHub repository. Once the program runs, it will locally host this interactive map. The map allows for panning across the continental United States, along with Alaska, Hawaii, and Puerto Rico.



From a wide perspective it is interesting to note that there are definitive areas where fires occur most often and another set of areas where fires occur much less. Curiously enough, when looking at Hawaii, fires mainly occur along the coastline. When looking at Alaska, the most fires occur around Anchorage and Fairbanks, which just so happen to be the most populated areas of the state. Perhaps wildfires in Alaska's more remote regions are simply not recorded? Or perhaps the colder climate much of the year helps control the number of natural fires in non-populated areas.

Each of these questions could be an area where this project could continue further by looking for correlations between number of fires and population of a geographical region. However, after understanding that most fires are caused by lightning strikes, it is curious to think about if there is a correlation between the number of lightning strikes and human population in an area.

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] J. Schwartz, V. Penney. In the West, Lightning Grows as a Cause of Damaging Fires. The New York Times (October 23, 2020). https://www.nytimes.com/interactive/2020/10/23/climate/west-lightning-wildfires.html

[4] 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