Datanators

Rohit Dube, Tushar Pandey, Arpan Pal, Mansi Bezbaruah, Benjamin Warren

Executive Summary

Github: https://github.com/ManaswineeB/datanators

This project focuses on analyzing wildfires through historical data (from 2019 to 2022). Wildfires pose a great threat to the loss of wildlife, people as well as money and require a significant amount of planning to stop the spread from one area to an entire county. More than half of wildfires are caused due to human activities and the number of incidents seems to be increasing every year. From 24,722 reported incidents in 2019, it went up to 37012 in 2022. Furthermore, the area burnt in 2022 was 1 Million acres as compared to 0.46 Million acres in 2019. As much as prediction models are required to be better prepared for wildfires, there is a need for better planning as well.

While measures are taken to reduce or contain the wildfires caused by human activities, it still amounts to more than 50% of the wildfires. In this project, we focus on the aspect of resource allocation and local station assignments within a state and a county respectively. The historical data is sampled from 2019 to 2022 and patched with weather and population data. We also, in turn, produce a tabular format of weather data for these years. As much as the agencies required to control the fire are important, it's also important to record the fire and take preliminary actions to avoid more life losses. This is where the local scheme of things comes in.

For each county, the location of wildfire data, along with time, rainfall, temperature, and population density is analyzed. An integer programming model is used to find out suitable placement of "First Response Stations". This model is relatively simple and uses binary variables to find the optimal number of stations as well as their location. For example, 'Cass County in Texas recorded more than 180 wildfire incidents within the last four years. The mixed integer program suggested 9 stations as well as the locations. None of these locations are within a few miles of each other which provides stability and robustness to the system. We create an index WFL (Wildland Fire Likelihood) based on the area burned, weather conditions, population density, time, and geographical location to account for the level of threat it imposes. This also provides us with a county-level WFL, with higher WFL implying more frequent wildfires and therefore more resources required to contain them as soon as possible.

To further analyze the resource constraints, we looked at the existing fire resource locations in Texas and ranked them according to another index called WWF (Weighted Wildfire Frequency). This index characterizes the intensity of cumulative wildland fires using its area and the time required to go from an existing station to the wildfire location. This index suggests that "Cass" county should receive the most volume of resources whereas "Fredericksburg" needs the least. This also aligns with the existing resources where "Cass" county has 7 Dozers whereas "Fredericksburg" has 1. This analysis suggests some changes in the resource allocation in comparison to the current placement across Texas.

The analysis and the modeling resulted in a paper (Work in Progress) that we plan on submitting to an IISE conference. The IP model for local station placement and the analysis for state-wise resource allocation provide a new method to look at resource planning in order to minimize the damages caused by wildfires.

Problem Statement

Wildfires have a huge impact on vegetation, air quality, and wildlife. The aim is to look into the depth of how these wildfires happen and what can be done to prepare ourselves for it. As an analyst, we try to observe patterns, check the effectiveness of existing resource allocation, and provide inferences as well as recommendations as to where to invest more resources for the next wildfire season.

Data Collection and Preprocessing

The analysis has been made on four groups of data: Wildfire Locations, Weather, County demographics, and Fire Resource locations. We created a dataset including the first three to analyze county-level requirements of resources. We use the first, third, and fourth datasets to analyze the resource planning question for the state of Texas. The following links have been used to collect all four sets of data for the last four years.

- https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/mapping/110-pcp-202201-1.csv
- https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/mapping/110-tmax-202201-1.csv
- https://data-nifc.opendata.arcgis.com/datasets/nifc::wildland-fire-incident-locations/explore?location=-0.000000%2C0.000000%2C2.97&showTable=true
- https://www2.census.gov/programs-surveys/popest/geographies/2021/state-geocodes-v2021.xlsx
- https://simplemaps.com/static/data/us-counties/1.73/basic/simplemaps_uscounties_basicv1.73.zip
- https://public.opendatasoft.com/api/explore/v2.1/catalog/datasets/us-county-boundaries/exports/csv?lang=en&timezone=America%2FChicago&use_labels=true&delimiter=%3B

Weather data over the years has been collected and consolidated into a single datasheet. Data associated with the counties of interest are merged and analyzed. At first, the features were cleaned and the null entries of the fire area were set to zero. Further, the data was merged such that for each county in the wildfire dataset, we augment it with the weather and demographics data.

For the data corresponding to different stations in Texas was collected from

- https://www.kristv.com/news/local-news/federal-state-fire-crews-in-coastal-bend-ready-to-assist-local-departments
- https://www.dshs.texas.gov/regional-local-health-operations/public-health-regions
- http://router.project-osrm.org/route/v1/car/

Data Exploration

The historical data was a big dataset with lots of missing values and features. We extract some features, namely County, Location, State, Cause, Area, the start time of the fire as well as the time when it was under control. We look at the states with the frequency of wildfire occurrences as well as the month of the year when wildfires are prevalent. Here is a plot for the states with the most occurrences of wildfires.

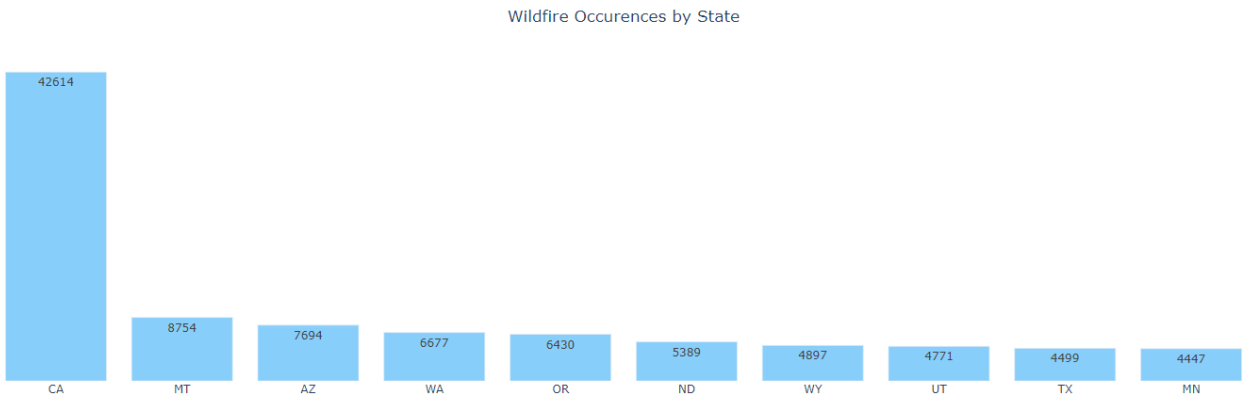

Figure 1: The number of wildfires recorded in different states in a descending order

Figure 2: Fires in New Mexico over different months. This shows some trend in the area burnt over different months. For the months of May, June and July, the area burnt is more in general, as well as more densely burnt patches inside the state.

After the exploratory analysis, we added more natural conditions to the environment, aiding with rainfall and temperature data. We see that the correlation between the number of wildfires in each month and the average precipitation is 0.15. Wildfires caused by humans depend less on natural conditions. Therefore, we try to study another problem, i.e. the resource placement problem. This also motivated the question about the resource-sharing problem which we solve in the second part.

Methodology

- **Feature Engineering**: After cleaning the dataset, some more features were constructed which includes the population density per county. For the first part of the problem, the WFL (Wildland

Fire Likelihood) score is created for each wildfire incident reported. For the second part of the problem, the distance between different county and wildfire locations are added as a feature in the dataset.

- **Integer Programming**: Once suitable features were engineered, we moved to an integer programming model. We model the problem into an integer program in the following way:

$$Minimize\ \Sigma_{i,j}\, d_{ij} x_{ij}$$

Subject to:

$$\Sigma_j\, x_{ij} = 1\ \forall\, i\ ;\quad \Sigma_i x_{ij} < k\, y_j\ ;\quad \Sigma_j\, y_j \leq max\ stations$$

x_{ij} is a binary variable which equals 1 if location i looks over station j,

y_j is a binary variable as well which takes the value 1 when station is built at location j,

d_{ij} is the distance between location i and station j,

and each station has a capacity k.

- **Dynamic Visualization**: Different wildfire locations are plotted alongside the stations created using integer programming over the course of multiple years. The graph is plotted for any given county fips code. The graphs of wildfire occurrences as well as state-wise wildfires are also plotted. For the final problem, the WWF score is plotted against counties that describe the need for resources in different regions of Texas.

Modeling, Analysis, and Interpretation

Cause Analysis

The important part of any problem lies in the identification of the problem itself. Most of the time, given a data set, the new natural thing that comes to mind is predicting a feature. While most people try to find the correlation to predict an outcome, there is a crying need to use these predictions to take action. We first looked at the cause of different wildfires, only to realize that it's half human. This changes a lot about predictions already because a non-natural and spontaneous element is added to it.

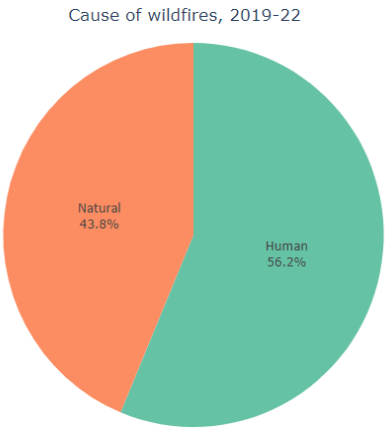

Figure 3: Proportion of wildfires caused by human activities vs natural causes (extrapolated from unknown causes 50-50 towards both)

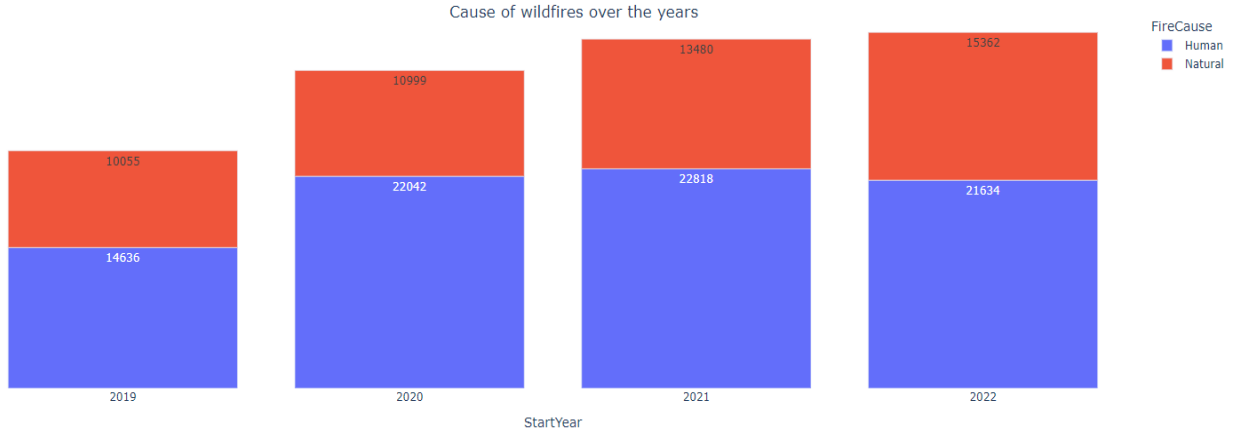

Figure 4: The natural wildfires are sharing more percentage in the total wildfire caused. This suggests the problem of resource management

In Figure 4, the number of fires due to natural causes seems to be increasing which would mean the prediction models will start performing better. However, with better prediction models, that would mean the model can be fed into our IP model to get the locations of local stations.

Plot Overview: County-level Stations

For different (and potentially all) counties which have fires with an area of more than 1 Acre, the model spits out the number of stations as well as the locations. In this plot, Cass County is selected. The model takes into account these different wildfire locations for all four years and creates a parameter space with 16 options of locations for stations. Then based on constraints, it minimizes the cost as a function of distance. Cass County has a high volume of wildfires, therefore 9 stations are decided and selected. It captures more than one year's data, which makes sense in selecting the stations.

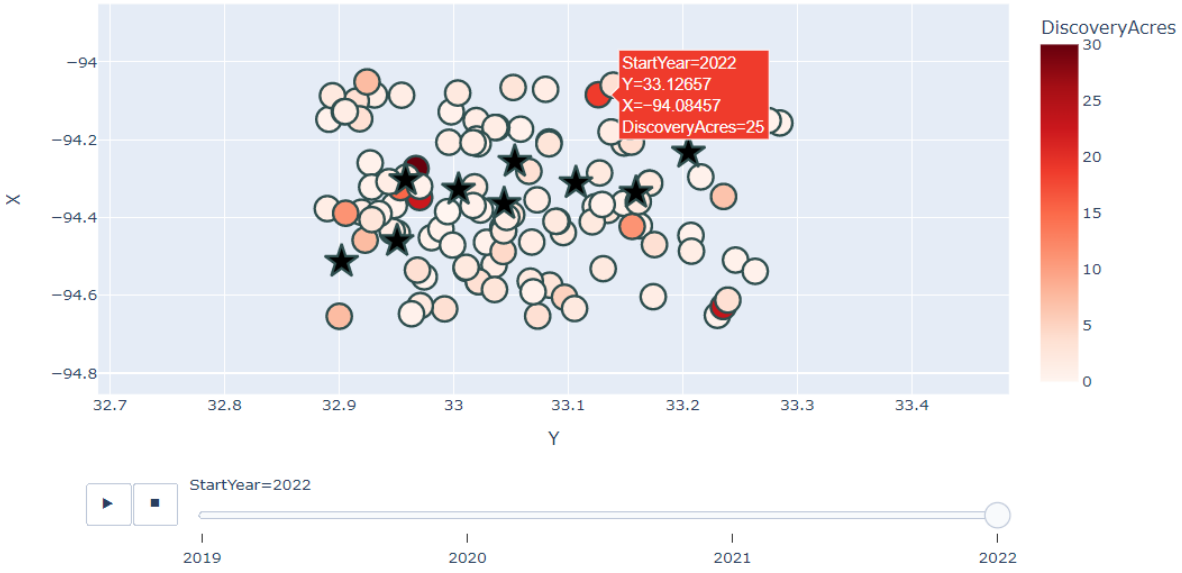

Figure 5: Cass County: very high frequency of wildfires, and therefore we see 9 stations. The stations stay stationary over the years and maximize the impact of available resources towards identifying the fire, its cause and potential impacts

In another county, Butte, in Idaho, there are not too many wildfires, though there are some big ones every year. Therefore, we get one station closer to the bigger fire. Usually, a forest isn't burnt out after a wildfire, which means there is still a possibility of wildfire in the area. Since the data is not limited to one year, the location of the station is only mildly dependent on the higher-intensity data.

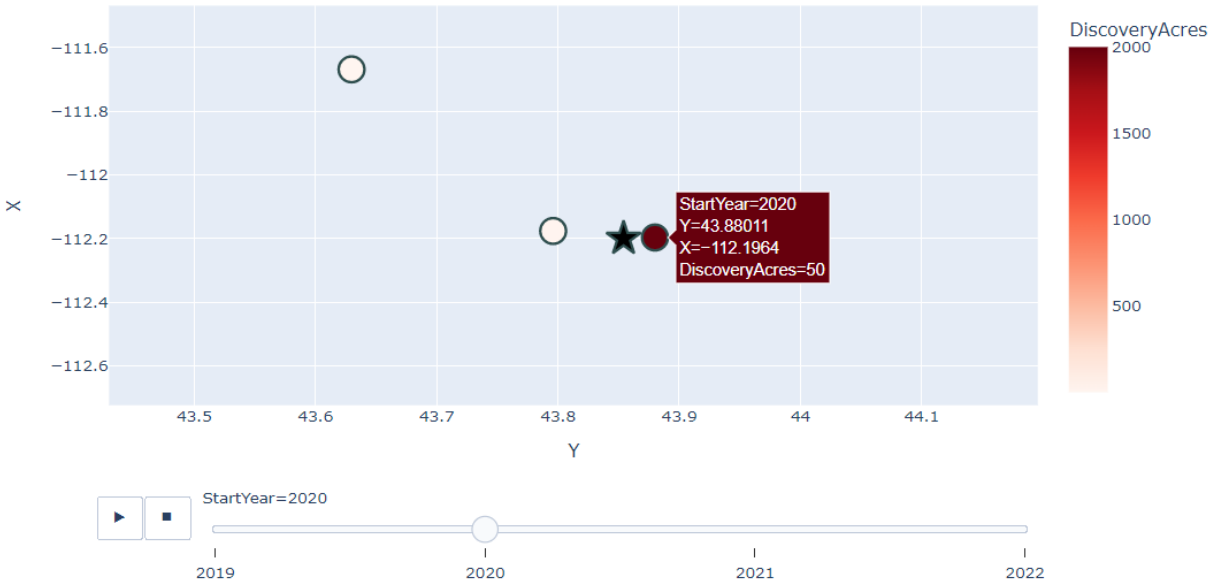

Figure 6: Butte County, Idaho. The number of incidents is much lesser than in Cass County, but the intensity for a few of them is high enough for it to be considered in the model. We see that due to the lesser number of incidents as well as the area, we get only one station assigned to the county. It's also close to one of the locations where the area burnt was large

WFL: Wildland Fire Likelihood

For each incident we assign a score based on different features in the dataset.

$$\mathcal{WFL} = \frac{10 - \hat{A} - \hat{R} + \hat{T} - 0.01(M-12)(4+M) + 0.1(PD)^{-1} + 0.01(\mathcal{M}ax(L) - L)}{100}$$

L: Latitude of the location
M: Month of the incident
PD: Population density
$\hat{A}$: Normalized Area burnt for that instance
R: Normalized Rainfall for a month in the county
$\hat{T}$: Normalized Max Temperature in a month in the county

The index is assigned by taking into contributions from weather, time, county demographics, location and the intensity of the fire.

1. Burnt area reduces the chance slightly of the area burning again
2. Increase in moisture content reduces the near term burning chances as well
3. The higher the temperature, the more likely it is to have possibility of natural burn
4. There is a clear correlation between month and number of incidents, which achieves its maximum around July, Aug and min in December. January.
5. Low population density means more natural area available. This increases the chances of wildfire in the vicinity. The correlation is small, therefore, an additional factor of 0.01 is included.
6. The higher the latitude, the lower chances it has for burning, however the difference is slightly small, therefore an additional factor of 0.01 is included.
7. The score is normalized by 100, to have a reasonable score of wildfires within a county.

The scores are accumulated for different counties and after sorting, it turns out that Los Angeles County has the highest WFL score. This is rightfully so because there are more than 12000 incidents of fire recorded in four years in this county. Since the state of California records the highest number of wildfires, most of the top WFL counties are from California.

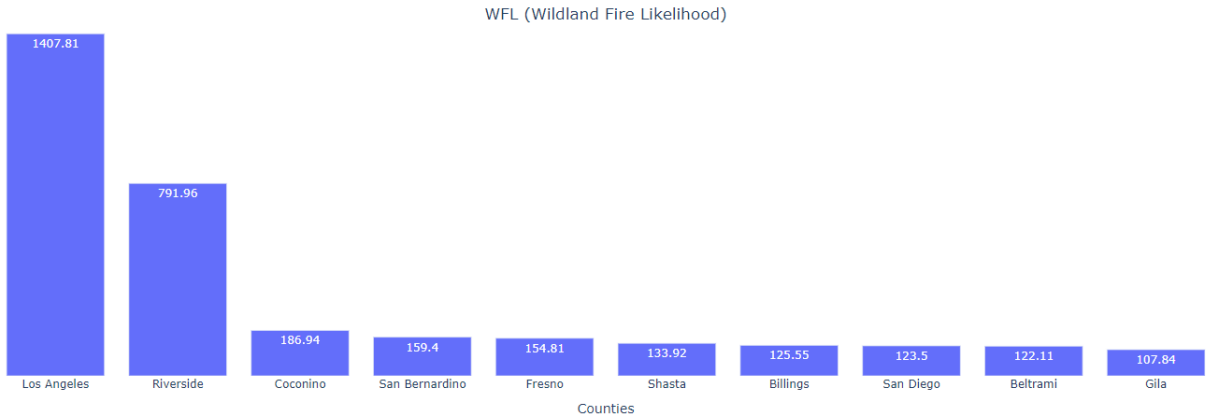

Figure 7: Los Angeles county has over 12000 wildfires in four years, which is the reason why its WFL score is very high. California being one of the most “burnt” states in terms of wildfires, more than fifty percent of the top scoring counties are from CA.

Resource allocation among Stations

Further investigation on the resources currently allocated at the various wildfire stations and the actual necessity of resources is done. The resources needed by wildfire stations to contain a nearby wildfire are directly proportional to the number and intensity of wildfires the station encounters. Due to the changing pattern of wildfires across a state, it is important to come up with an informative function that would be helpful in classifying high-priority wildfire stations needing better resources to fight wildfires. We form such a metric based on the duration it took to travel from the wildfire station to the wildfires that occurred in the last 4 years. The study is done for the state of Texas where there are 19 wildfire stations currently shown in Figure [8]. The 19 stations belong to 9 regions of Texas to which they cater to.

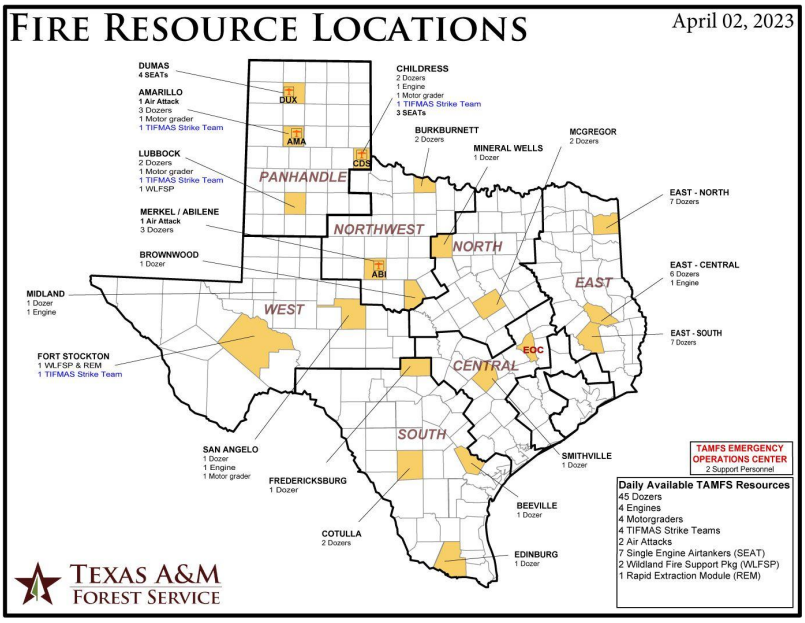

Figure 8: Wildfire Stations in Texas

We start the analysis by first allotting the County Wildfire Score (CWS) for each county using data from the last 4 years.

$$CWS = \frac{3 \times \hat{I} + \hat{A}}{4}$$

where, $\hat{I}$: Normalized number of wildfire occurrences in a county

$\hat{A}$: Normalized Area (acres) burnt in a county

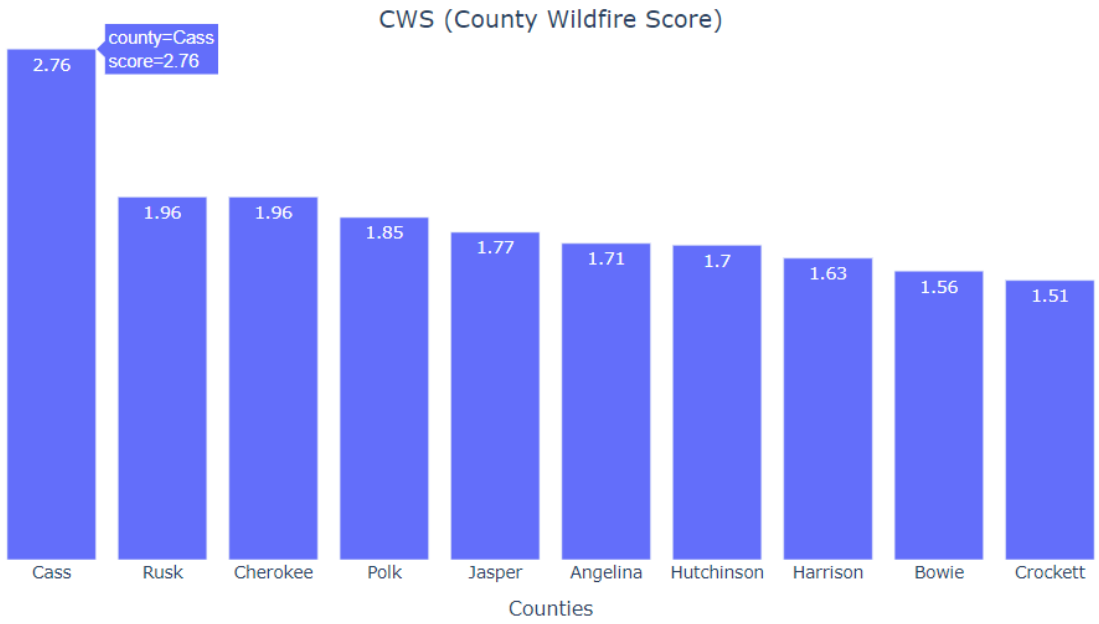

Figure 9: CWS for top 10 counties in Texas

CWS helps us to identify the counties which were highly affected by wildfires, the score will be higher for the counties with repeated occurrences of wildfires in the past and will help in identifying which wildfire stations across all the counties need more resources. The duration it takes from all 19 wildfire stations to transport resources from the station to the wildfire locations is calculated. The Weighted Wildfire Frequency Score for each station is calculated using this duration of travel weighted by the CWS of the county in which the wildfire was located.

$$\mathcal{WWF}(stn) = \sum_{L_i} d(L_{stn}, L_i) \times \mathcal{CWS}_i$$

Where, *WWF(stn)* is the Weighted Wildfire Frequency Score for the wildfire station

i : index of all wildfire locations in the region of wildfire station

L_{stn}: The location of wildfire station

L_i : The location of wildfire in the region of wildfire station

$d(L_{stn}\ L_i)$: Driving duration between two locations in hours

CWS_i : County Wildfire score of the county of the wildfire location

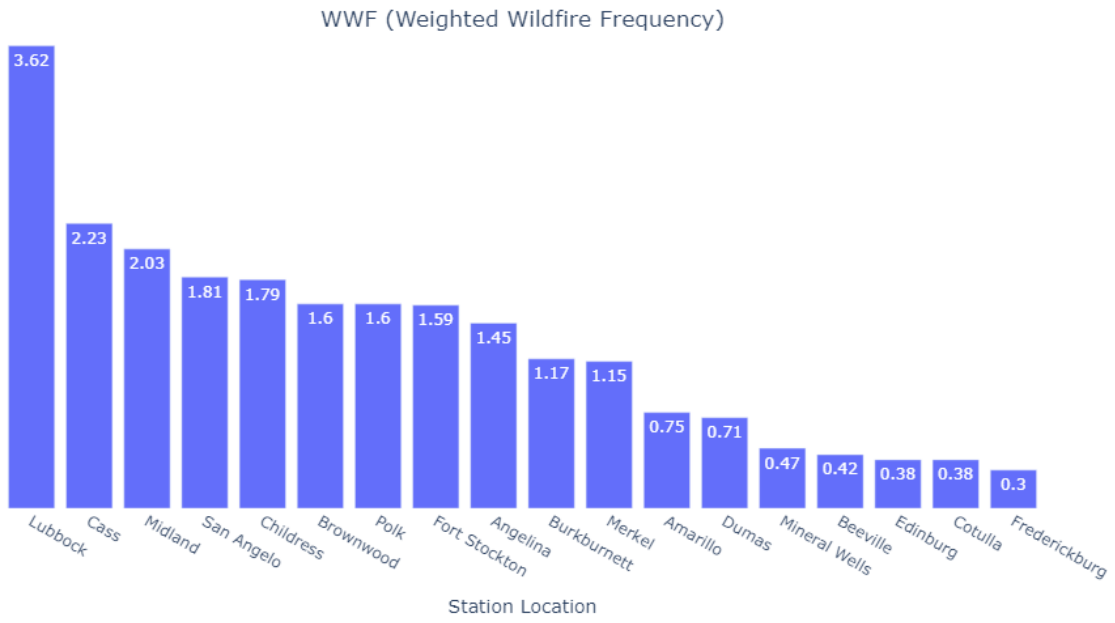

Figure 10: Weighted Wildfire Frequency for all the wildfire stations in Texas, based on time duration from different wildfire locations, taking into account a score for every county. This matches with the existing resource allocation for different stations, up to an extent, but needs some changes. Fredericksburg being at the bottom has 1 dozer, and so does Edinburg.

The WWF for the present wildfire stations in Texas almost matches the resource allocation shown in Figure [8]. Lubbock station has one of the highest current resources in the state and is almost equal to the Childress station, both located in the same region. However, the WWF for Lubbock is almost twice that of Childress indicating the need for better resource distribution in favor of Lubbock or an additional station in

the region. Similarly *WWF* score can be used to redistribute the resources among the stations based on the historical wildfire occurrences.

Conclusions and Recommendations

As much as the prediction of wildfire is important, it still needs to be paired up with strategic resource planning in order to reap the maximum reward from it. We work on this strategy and make it dynamic based on different time intervals. The first conclusion is to create first response stations or add resources based on either recent wildfires or predicted wildfires if there is any accurate prediction model. Assigning scores to different wildfires, be it historic or predicted, helps us better prepare for the future, both in terms of reporting as well as minimizing the damages.

The second conclusion we draw from our analysis is that it's better to be dynamic about the placement of bigger/more expensive resources within a state. This means, assigning an index to different available resource locations within a state every year, and permuting them according to the proposed idea. After assigning the scores, it is possible to gather more data about the cost of different resources, including the working cost, productivity/efficiency, the time required to arrive, etc, and add them to another MIP model in order to find the allocation of resources in real-time for different time frames (if possible).

Some possible complexification of models based on additional data include:
1) Changing the IP model to include the cost associated with extinguishing the wildfire based on time.
2) Adding a budget constraint to find the suitable number of local first response stations.
3) Acquiring the cost sheet for different resources along with their depreciation cost, and making a MIP model that allocates the resources into different stations every year based on a fixed budget.
4) Adding the cost of transfer of resources to verify the extent of practicality in it, as well as the half life of these resources at a station.

References:

1. Budget Justification, Forest Service, US Dept of Agriculture, https://www.fs.usda.gov/sites/default/files/2022-03/FS-FY23-Congressional-Budget-Justification.pdf, Mar 2022
2. The costs and losses of wildfires, a literature review,: https://doi.org/10.6028/NIST.SP.1215 https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1215.pdf
3. An Integer Programming Model to Optimize Resource Allocation for Wildfire Containment, Geoffrey H. Donovan and Douglas B. Rideout, https://www.fs.usda.gov/pnw/pubs/journals/pnw_2003_donovan002.pdf