# MIS 545 Project Report: Predicting Fire Size

## Problem description

### Problem addressed by our team

Our team project is addressing the issue of wildfires in the United States. This year, fires in California have burned 504,939 acres (California Department of Forestry and Fire Protection). The current wildfires in Northern California have caused the death of 34 people and destroyed an estimated 5,700 buildings. This includes a minimum of 2,834 homes that have been destroyed by the Tubbs fire in Santa Rosa, California (Nelson et al., 2017). This is a serious problem and our project aims to use predictive algorithms to help fire departments predict the size of a wildfire based on several attributes. The original dataset we used for stage 1 of our project contains information about wildfires that have occurred in the U.S. between 1992 and 2015. It includes wildfires reported by various government agencies and contains 1,880,465 records.

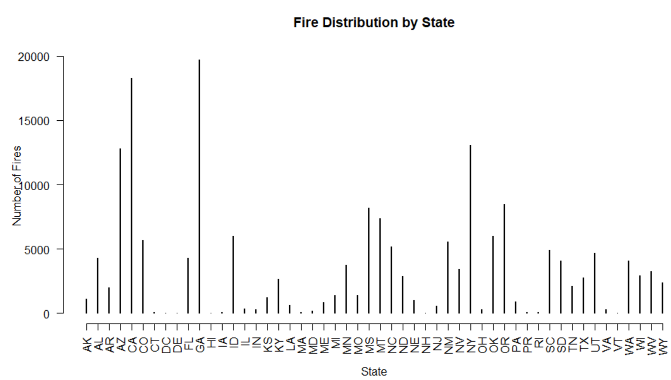
### State of the domain

Other data mining approaches have been used to predict forest fire size. These primarily relied on meteorological datasets (Shidik and Mustofa, 2014). These include studies outside of the United States that use data mining approaches on Portugal fires (Cortez and Morais, 2006). Additionally, The US Forest Service maintains a variety of models and simulation tools that predict the spread and behavior of fires once they begin.

### Stage 2 Changes

From our initial dataset, I found that California has a particularly high amount of fires, second only to Georgia as seen in Figure 1. With the current state of wildfires in California, I chose to narrow my focus for my individual project on California.

**Figure 1:** Fire Distribution by State




Wildfires are highly dependent on weather conditions, but there were no attributes in our original dataset relating to weather. For my individual project, my goal is to combine a dataset of weather attributes with our original fire dataset. I can associate weather to a fire by the year and month the fire was discovered. This will allow me to utilize additional weather attributes in my predictive algorithms in order to see if these can create a better predictor of fire size.

Our preliminary analysis utilized Naïve Bayes and decision tree algorithms. To expand my study further, I will also employ SVM, and neural network algorithms.

## Dataset Description

### Original Dataset Description (From Stage 1)

The dataset that we are using for this project contains information about wildfires that have occurred in the U.S. between 1992 and 2015. It includes wildfires reported by various government agencies and contains 1,880,465 records. Table 1 in the Appendix shows a list of the dataset's attributes as well as a description for each attribute. Table 2 includes descriptive statistics for the relevant attributes that can be used to make predictions. Some of the variables included in this dataset are different types of unique identifiers used by different agencies for the same fire and thus provide no new information or useful statistics that can be summarized in Table 2 in the Appendix.

### New Dataset

In order to utilize weather conditions in my analysis, I utilized data from the National Centers for Environmental Information, which is part of the National Oceanic and Atmospheric Administration. I limited my time frame to match my fire dataset with a start year of 1992 and end year of 2015. The data I selected was on a time scale of averages per month and corresponded to statewide California weather conditions. The following Table 3 describes the 20 weather attributes I was able to download from this source. Table 4 includes the descriptive statistics for this new set of data. It is stated from the data source: "Because these data are primarily intended for the study of climate variability and change, observations have been adjusted to account for the artificial effects introduced into the climate record by factors such as instrument changes, station relocation, observer practice changes and urbanization. Some of the more current data provided by the Climate at a Glance system are preliminary and may be modified after appropriate quality control has been performed. As a result, some values available on this site differ from the official observations."

**Table 3.** Weather Dataset Attributes

| Attribute | Description |
|---|---|
| **Date** | Date and month that the weather statistics refer to (YYYYMM) |
| **Average Temperature** | Average temperature in degrees Fahrenheit |
| **Average Temperature Anomaly** | Departure from the mean average temperature from the century (1901-2000) |
| **Cooling Degree Days** | The number of degrees that a day's average temperature is above 65 degrees Fahrenheit multiplied by the number of days in the month |
| **Cooling Degree Days Anomaly** | Departure from the mean cooling degree days from the century (1901-2000) |
| **Heating Degree Days** | The number of degrees that a day's average temperature is below 65 degrees Fahrenheit multiplied by the number of days in the month |
| **Heating Degree Days Anomaly** | Departure from the mean heating degree days from the century (1901-2000) |
| **Maximum Temperature** | Maximum temperature during the month in Degrees Fahrenheit |
| **Maximum Temperature Anomaly** | Departure from the mean maximum temperature from the century (1901-2000) |
| **Minimum Temperature** | Minimum temperature during the month in Degrees Fahrenheit |
| **Minimum Temperature Anomaly** | Degrees Fahrenheit departure from the mean minimum temperature from the century (1901-2000) |
| **Palmer Drought Severity Index** | Estimate of relative dryness based on temperature and precipitation data. From -10 (dry) to +10 (wet). It attempts to measure the duration and intensity of the long-term drought-inducing circulation patterns |

| | |
|---|---|
| **Palmer Drought Severity Index Anomaly** | Departure from the mean Palmer Drought Severity Index for the century (1901-2000) |
| **Palmer Hydrological Drought Index** | A measure of hydrological impacts of drought (e.g., reservoir levels, groundwater levels, etc.) which take longer to develop and longer to recover from. This long-term drought index was developed to quantify these hydrological effects, and it responds more slowly to changing conditions than the PDSI. |
| **Palmer Hydrological Drought Index Anomaly** | Departure from the mean Palmer Hydrological Drought Index for the century (1901-2000) |
| **Palmer Modified Drought Index** | Palmer Drought Severity Index  that takes the sum of the wet and dry terms after they have been weighted by their probabilities causing a more normal distribution |
| **Palmer Modified Drought Index Anomaly** | Departure from the mean Palmer Modified Drought Index for the century (1901-2000) |
| **Palmer Z Index** | A measure of short-term drought on a monthly scale |
| **Palmer Z Index Anomaly** | Departure from the mean Palmer Z Index for the century (1901-2000) |
| **Precipitation** | Amount of precipitation in inches |
| **Precipitation Anomaly** | Departure from the mean precipitation for the century (1901-2000) |

**Table 4.** Weather Dataset Descriptive Statistics

| Attribute | Mean | Min | Max | Standard Deviation |
|---|---|---|---|---|
| **Average Temperature** | 68.36 | 40.30 | 79.30 | 9.29 |
| **Cooling Degree Days** | 156.00 | 0.00 | 377.00 | 104.53 |
| **Heating Degree Days** | 86.54 | 0.00 | 624.00 | 137.23 |
| **Maximum Temperature** | 82.43 | 49.30 | 94.20 | 10.50 |
| **Minimum Temperature** | 54.28 | 29.5 | 64.40 | 8.13 |
| **Palmer Drought Severity Index** | -1.37 | -7.01 | 7.17 | 3.17 |
| **Palmer Hydrological Drought Index** | -1.10 | -7.01 | 7.17 | 3.39 |
| **Palmer Modified Drought Index** | -1.18 | -7.01 | 7.17 | 3.26 |
| **Palmer Z Index** | -0.64 | -5.89 | 9.60 | 1.65 |
| **Precipitation** | 0.65 | 0.01 | 12.50 | 1.024 |

## Data preprocessing activities

I started with our preprocessed dataset of fires from stage 1 of our project so there were no missing values. The original data set was then limited to those where the state was CA to focus on California fires. There was no missing data in the new weather dataset, but in order for the new data to be used, each of the weather attributes had be combined based on their date since each attribute was downloaded individually from the data source. The month and year were then derived from the date in order to compare with the fire dataset. I then merged the fire and weather where month and year matched, allowing analysis for the weather conditions during the year and month a specific fire is discovered. This left me with a new dataset of fires and weather conditions containing 18,305 records.

## Algorithms

The goal of the algorithms is to predict weather a fire will be small (less than or equal to 0.5 acres) or large (greater than 0.5 acres). As this is a classification problem, I elected to perform a comparison among the classification algorithms we have discussed in class: Naïve Bayes, Decision Trees, Neural Network, and SVM.

When evaluating feature selection, state was no longer necessary from the original dataset since the region is restricted to California. Looking at scatter plots (figures 2-4 in appendix) and evaluating a Goodman-Kruskal Tau Matrix (figure 5 in appendix) shows that the weather data does not have a strong linear correlation to fire size, but still does effect fires. The matrix also shows that the Palmer Drought Severity Index is highly correlated with every other weather attribute. Therefore, in addition to the attributes used in stage 1 of the project, I will add the Palmer Drought Severity Index and Palmer Hydrological Severity Index in order to test if weather data has an impact on the performance of our predictions as opposed to our initial analysis. Table 5 shows the attributes used.

**Table 5.** Algorithm Attributes

| Attribute Name | Scale | Implementation |
|---|---|---|
| **Cause of Fire** | Nominal | Attribute |
| **Landowner Description** | Nominal | Attribute |
| **Month the fire was discovered** | Nominal | Attribute |
| **Month the fire was contained** | Nominal | Attribute |
| **Number of days to contain the fire** | Nominal | Attribute |
| **Palmer Drought Severity Index** | Interval | Attribute |
| **Palmer Hydrological Severity Index** | Interval | Attribute |
| **Size class of the Fire** | Nominal | Class Label |

## Execution and Results

I used R to train and test my algorithms. I created training and testing sets which were 70% (12,813) and 30% (5,492) of the new dataset. The algorithms were implemented using functions provided by the caret package.

### Naïve Bayes

Confusion Matrix

| | Actual | |
|---|---|---|
| Predicted | Large | Small |
| Large | 83 | 70 |
| Small | 1416 | 3923 |

Evaluation Measures

| Baseline Accuracy | Accuracy | Small Fire Precision | Large Fire Precision | Small Fire Recall | Large Fire Recall |
|---|---|---|---|---|---|
| 73.37% (small fires) | 72.94% | 73.48% | 54.25% | 98.25% | 5.54% |

### Decision Tree

Confustion Matrix

| | Actual | |
|---|---|---|
| Predicted | Large | Small |
| Large | 273 | 147 |
| Small | 1226 | 3846 |

Evaluation Measures

| Baseline Accuracy | Accuracy | Small Fire Precision | Large Fire Precision | Small Fire Recall | Large Fire Recall |
|---|---|---|---|---|---|
| 73.37% (small fires) | 75.00% | 75.83% | 65.00% | 96.32% | 18.21% |

## Neural Network

Confusion Matrix

| | Actual | |
|---|---|---|
| Predicted | Large | Small |
| Large | 293 | 220 |
| Small | 1206 | 3773 |

Evaluation Measures

| Baseline Accuracy | Accuracy | Small Fire Precision | Large Fire Precision | Small Fire Recall | Large Fire Recall |
|---|---|---|---|---|---|
| 73.37% (small fires) | 74.03% | 75.78% | 57.11% | 94.49% | 19.55% |

## SVM

Confusion Matrix

| | Actual | |
|---|---|---|
| Predicted | Large | Small |
| Large | 44 | 51 |
| Small | 1455 | 3942 |

Evaluation Measures

| Baseline Accuracy | Accuracy | Small Fire Precision | Large Fire Precision | Small Fire Recall | Large Fire Recall |
|---|---|---|---|---|---|
| 73.37% (small fires) | 72.58% | 73.04% | 46.32% | 98.72% | 2.94% |

## Conclusion

The results of these tests demonstrate that a decision tree algorithm maintains the highest accuracy in comparison with Naïve Bayes, Neural Networks, and SVM for this particular case. However, there was not a significant improvement in accuracy in comparison to our stage 1 analysis, despite the addition of weather data. Further, the decision tree algorithm only performed slightly better (even after tuning) than the baseline case where small fires made up 73.37% of my new dataset. However, my dataset for California fires was only 18,305 records and the weather is localized to one state. Moving forward, perhaps I could continue to collect and process the data for every state, and merge that into the larger nationwide wildfire dataset. In that case, the fire sizes would not be as skewed in a larger dataset, and weather could be more of a predictor. Additionally, since a decision tree algorithm proved most accurate, I can try to use a random forest to further improve our prediction.

## Appendix

**Table 1.** Original Dataset Attributes

| Attribute | Description |
|---|---|
| OBJECT ID | Global unique identifier. |
| FOD_ID | Unique identifier that contains information necessary to locate the original record in the source dataset. |
| FPA_ID | Unique identifier that contains information necessary to locate the original record in the source dataset. |
| SOURCE_SYSTEM_TYPE | Type of source database or system that the record was drawn from (federal, nonfederal, or interagency). |
| SOURCE_SYSTEM | Name of the other identifier for source database or system that the record was drawn from. |
| 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). |
| NWCG_REPORTING_UNIT_ID | Active NWCG Unit Identifier for the unit preparing the fire report. |
| NWCG_REPORTING_UNIT_NAME | Active NWCG Unit Name for the unit preparing the fire report. |
| SOURCE_REPORTING_UNIT | Code for the agency unit preparing the fire report, based on code/name in the source dataset. |
| SOURCE_REPORTING_UNIT_NAME | Name of reporting agency unit preparing the fire report, based on code/name in the source dataset. |
| LOCAL_FIRE_REPORT_ID | Number or code that uniquely identifies an incident report for a particular reporting unit and a particular calendar year. |
| LOCAL_INCIDENT_ID | Number or code that uniquely identifies an incident for a particular local fire management organization within a particular calendar year. |
| FIRE_CODE | Code used within the interagency wildland fire community to track and compile cost information for emergency fire suppression (https://www.firecode.gov/). |
| FIRE_NAME | Name of the incident, from the fire report (primary) or ICS-209 report (secondary). |
| ICS_209_INCIDENT_NUMBER | Incident (event) identifier, from the ICS-209 report. |
| ICS_209_NAME | Name of the incident, from the ICS-209 report. |
| MTBS_ID | Incident identifier, from the MTBS perimeter dataset. |
| MTBS_FIRE_NAME | Name of the incident, from the MTBS perimeter dataset. |
| COMPLEX_NAME | Name of the complex under which the fire was ultimately managed, when discernible. |
| 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 (Julian date). |
| 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 (Julian date). |
| 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. |
| COUNTY | County, or equivalent, in which the fire burned (or originated), based on nominal designation in the fire report. |
| FIPS_CODE | Three-digit code from the Federal Information Process Standards (FIPS) publication 6-4 for representation of counties and equivalent entities. |
| FIPS_NAME | County name from the FIPS publication 6-4 for representation of counties and equivalent entities. |

**Table 2.** Original Dataset Descriptive Statistics

| Attribute | Descriptive Statistic | Value | Number of Records if applicable |
|---|---|---|---|
| **NWCG_REPORTING_AGENCY** | Top Reporting Agency (number of occurrences) | State, County, or Local Organization | 1377090 |
| **NWCG_REPORTING_UNIT_ID** | Top Reporting Unit ID (number of occurrences) | USGAGAS | 167123 |
| **NWCG_REPORTING_UNIT_NAME** | Top Reporting Unit Name (number of occurrences) | Georgia Forestry Commission | 167123 |
| **SOURCE_REPORTING_UNIT_NAME** | Top Reporting Source Name (number of occurrences) | Georgia Forestry Commission | 97844 |
| **FIRE_YEAR** | Year with the most fires (mode) | 2006 | |
| **FIRE_NAME** | Name of fire (mode) | GRASS FIRE | 3983 |
| **DISCOVERY_DOY** | Average discovery day of year (mean) | 164.7 (June) | |
| **DISCOVERY_TIME** | Time that fire was discovered(mean) | 14:53 (Military) | |
| **CONT_TIME** | Time that fire was contained(mean) | 15:35 (Military) | |
| **FIRE_SIZE** | Average Fire Size (mean) | 74.5 (Acres) | |
| **FIRE_SIZE** | Minimum Fire Size | 0.0001 (Acres) | |
| **FIRE_SIZE** | Maximum Fire Size | 606945 (Acres) | |
| **FIRE_SIZE** | Standard Deviation of Fire Size | 2497.598 (Acres) | |
| **FIRE_SIZE_CLASS** | Class size of the fire (mode) | A | 666919 |
| **OWNER_DESCR** | Name of the landowner (mode) | Small Landowners | 1050835 |
| **STATE** | State the fire occurred in (mode) | CA | 189550 |

**Figure 2.** Average Temperature by Fire Size Scatter Plot

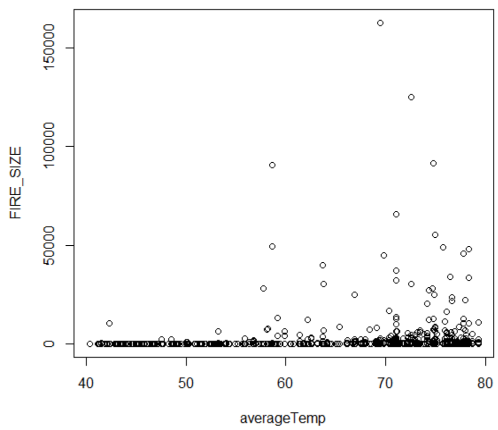



**Figure 3.** Palmer Drought Severity Index by Fire Size Scatter

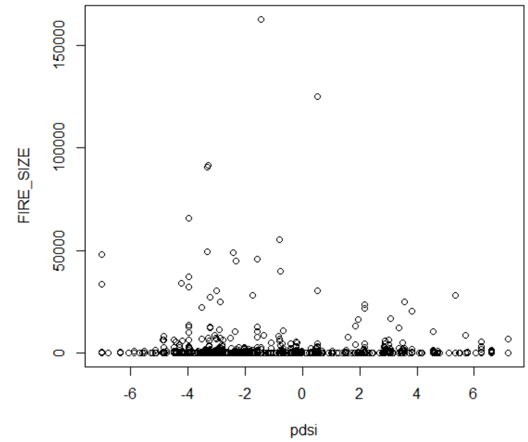



**Figure 4.** Precipitation by Fire Size Scatter Plot

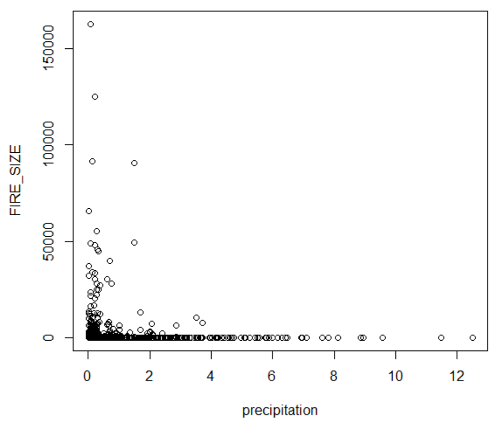



**Figure 5.** Goodman-Kruskal Tau Matrix

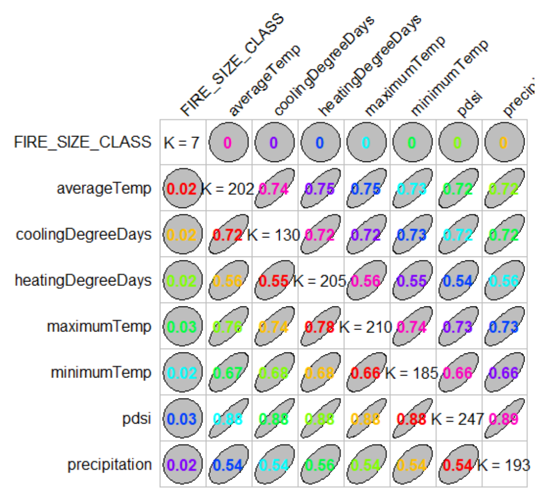
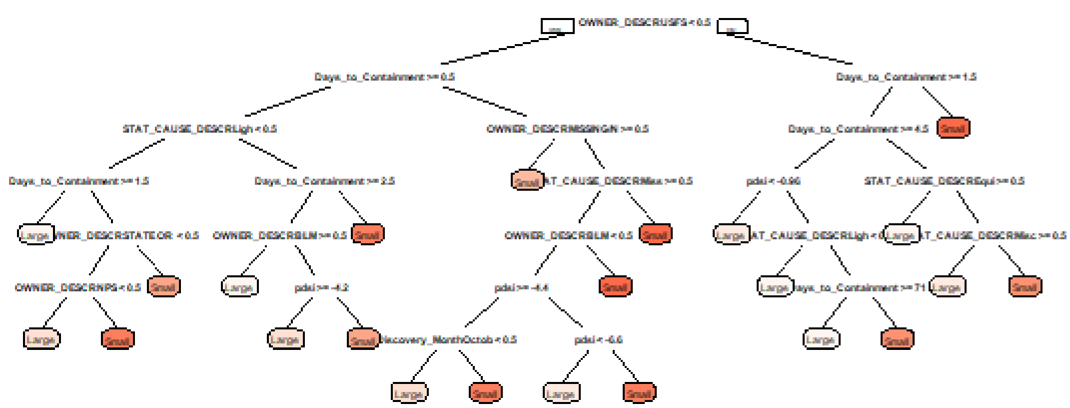



**Figure 6.** Stage 2 Decision Tree



## References


California Department of Forestry and Fire Protection (2017) NUMBER OF FIRES AND ACRES. Retrieved from http://cdfdata.fire.ca.gov/incidents/incidents_stats?year=2017

Cortez, P. and Morais, A. (2006). A Data Mining Approach to Predict Forest Fires using Meteorological Data. Department of Information Systems/R&D Algoritmi Centre, University of Minho.

Fowler, A., Teredesai, A. M. DeCock, M. (2009). An evolved fuzzy logic system for fire size prediction. Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American. Cincinnati, OH. DOI: 10.1109/NAFIPS.2009.5156419

Gregory, Matthew. (2016). Forest Fire Prediction with Support Vector Machines. www.Rpubs.com . Retrieved from https://rpubs.com/mammykins/svm_fires

National Centers for Environmental Information. “Climate at a Glance”. Retrieved from https://www.ncdc.noaa.gov/cag/

National Interagency Fire Center. “National Fire News”. Retrieved from https://www.nifc.gov/fireInfo/nfn.htm

National Oceanic and Atmospheric Administration. “US Wildfires”. Retrieved from https://www.ncdc.noaa.gov/societal-impacts/wildfires/ytd/12?params[]=fires&params[]=acres&end_date=2016

Nelson, Laura, Kohli, Sonali, St.John, Paige, Smith, Dakota, Agrawal, Nina (13 Oct. 2017). “Death toll from Northern California fires jumps to at least 34; 5,700 structures destroyed.” Retrieved from http://www.latimes.com/local/lanow/la-me-ln-fires-20171013-story.html

Shidik G.F., Mustofa K. (2014). Predicting Size of Forest Fire Using Hybrid Model. In: Linawati, Mahendra M.S., Neuhold E.J., Tjoa A.M., You I. (eds) Information and Communication Technology. ICT-EurAsia 2014. Lecture Notes in Computer Science, vol 8407. Springer, Berlin, Heidelberg

Tatman, Rachael (13 Sept. 2017). “1.88 Million US Wildfires”. Retrieved from https://www.kaggle.com/rtatman/188-million-us-wildfires