Modelling the Extent of Wildfires Using Machine Learning

Gao Yu
Kimberly Clemen Yap
Tan Guang Yuan
Varun Vaitheeswaran
Wong Kar Mun Shermaine

Abstract

This project made use of historical fire occurrence data in the United States to develop a predictive model that can accurately determine the probability and extent of wildfire occurrences. Other climate and socio-economic data were included to create additional features for the machine learning prediction models. Random Forest, Support Vector Machine and Neural Network models were evaluated based on the prediction accuracy and recall. It was found that while the models were able to achieve relatively high accuracies of up to 64%, most were unable to predict the larger classes of fire. As such, we simplified the models to a binary classification task by categorizing the various classes of fire into “small” or “large” fires. The accuracy of all the supervised learning models improved to above 70%, with the Neural Network model achieving up to 72.8% and with a recall for the larger fires at 78%.

1. Introduction

Climate change is one of the most pressing challenges of the 21st century. The increasing availability of data and computing power have allowed governments, activists and scientists to leverage on machine learning to resolve many of these challenges. This study uses data collected on wildfire occurrences in the United States (US) to predict the extent of wildfires globally in an attempt to mitigate their impact and deploy timely and adequate resources when necessary.

Wildfires, especially when not controlled in time, can be devastating disasters; destroying precious green forests, damaging properties, claiming wildlife and even human lives. Large-scale wildfires are known to be highly destructive and can spread quickly. The frequency and impact of wildfires have been increasing over the years. From the 1990s to the 2000s, the acres of land scorched by wildfires have doubled, with about 72,400 wildfires occurring each year (Congressional Research Service, 2019). The year 2015 recorded the largest wildfire season in US history, burning over 10 million acres of land (Washington Post, 2016).

Our study leverages six years of geo-referenced wildfire data (2010 to 2015) in the US to develop a model that predicts the size of wildfire occurrences based on climate and location-specific conditions. The outcome of the model could be used to recommend the size of firefighting forces needed to combat a wildfire as well to provide insights to firefighting departments in terms of deployment of resources. The data and codes used in this study can be accessed at https://github.com/vision20-20/BT5153-Group-Project.

2. Literature Review

The importance of wildfire prediction can be seen from the significant amount of literature and research work available. Sayad et al (2019) collected data through remote sensors and combined it with meteorological indicators to perform machine learning using Neural Network and Support Vector Machine (SVM) models, while Subramanian and Crowley (2018) modelled wildfire spread as a Markov Decision Process and applied deep learning algorithms. A similar study by Neves et al (2007) incorporates fire indices in their analysis and applied SVM and tree-based models on northeast Portugal fire data with the objective of optimising firefighting resources.

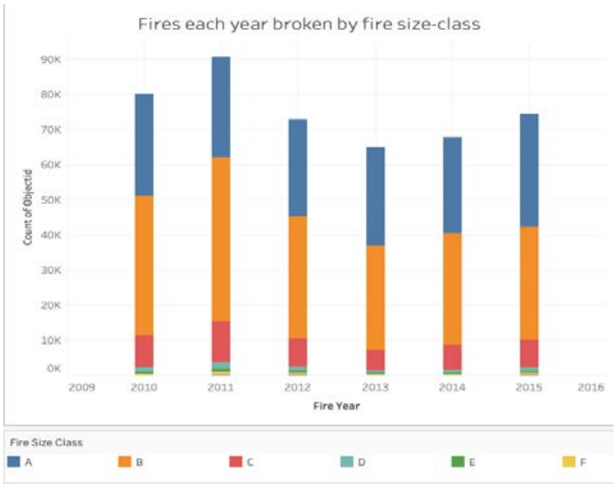
A study done by Vilar et al (2016) on Central Spain wildfires also showed that factors, such as population and urban planning, can be included in performing fire risk assessments. The analysis performed used linear and Maximum Entropy (Maxent) models. A comparable study in South Korea combined both socio-economic and environmental variables to estimate fire probability in forests by using a Random Forest model in addition to Maxent (Kim et al, 2019). Assuming the fire is due to human-related factors, socio-economic variables are relevant in building the prediction model.

3. Exploratory Data Analysis

The main dataset used in the study is obtained from Kaggle (Short, 2017) and records the wildfire occurrences in the US supporting the Fire Program Analysis (FPA). The features in the dataset include date, duration, location, cause and severity (land area affected by the fire) of each wildfire incident. Unlike other existing works focusing mainly on environmental conditions, this study aims to explore other location-based factors to gain useful insights on wildfire drivers that can determine their occurrences and severity. It has been noted that certain states and cities in the US have more severe and frequent occurrences of wildfires.

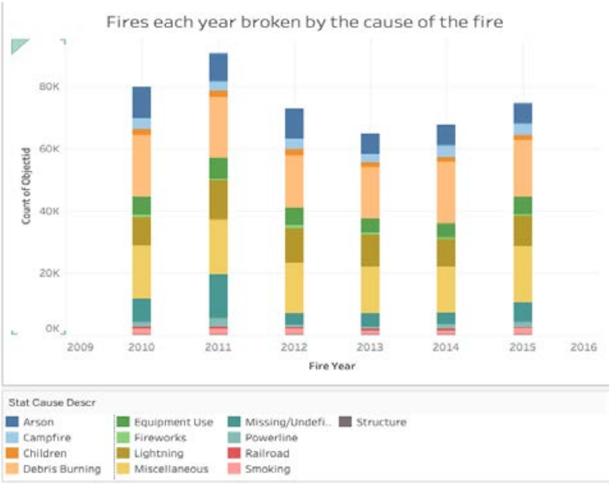
Additional variables suggested in related works have been explored in this study, such as environmental and socio-economic factors. For environmental factors, both temperature and instances of drought in various US locations were used. These datasets were obtained from Kaggle originating from the US Drought Monitor (2020). Socio-economic factors, specifically median income and unemployment rate by county, were included as well. These data were extracted from the US Census Bureau (2020) and Bureau of Labor Statistics (2020), respectively.

Figure 1. Fire occurrences summarized by year and by the size of the fires



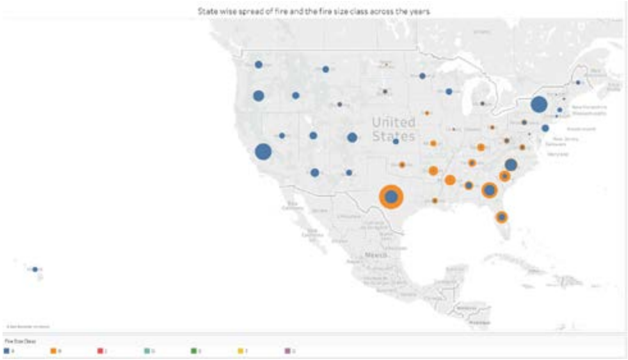
Initial findings from the visualization showed us that 2011 saw the highest number of fires with Class B fires (those that have more than one-fourth of an acre but less than 10 acres) being the biggest contributor (refer to Figure 1). Across the years, debris burning and arson (besides other miscellaneous reasons) are the biggest causes for fire with contribution from lightning-induced fires increasing over the years (see Figure 2).

Figure 2. Fire occurrences summarized by year and by the cause of the fires



From a geolocational perspective, Texas has the largest number of fires (of both Class A and B) in the USA. Arkansas and Mississippi have had fewer but more severe fires and the severity of fires was higher in the southern region where temperature is generally higher than the eastern or colder states, such as Washington.

Figure 3. Fire occurrences by fire size by state


State wise spread of fire and the fire size class across the years
United
States

4. Methodology

4.1 Data Extraction, Pre-processing and Integration

Given that data were obtained from various sources, we needed to do some pre-processing of the datasets to properly join them with our main 'Fires' dataset. The datasets that we have uploaded have already been cleaned by the team members separately and could be joined easily.

We started by cleaning up the 'Fires' dataset since it would be the main dataset that we joined the other datasets to. The pre-processing steps could be found in the Jupyter notebook attached. In brief, these were the key steps taken:

- Standardising the naming of the identifying keys, such as changing the column name ‘FIPS_NAME’ to ‘county_name’ to facilitate easy joining to the different datasets subsequently.
- Using a mapping table with high degree of granularity at state, county and city levels as an intermediate key to join the different datasets; with various degrees of granularity and naming conventions.
- Standardising all column names before performing the joins. As each dataset has different granularity, it is inevitable that certain data points will be lost for each dataset joined. We checked the loss of data points after each join and made a conscious decision to discard those datasets that resulted in too much loss.

Some of the additional variables that we used to run in our model includes:

- Drought - A binary variable that determined if there was a drought occurring for that county during the occurrence of the fire.
- Median Income - The median household income of the county for that year when the fire occurred.
- Unemployment rate - The unemployment rate of the county for that year when the fire occurred.

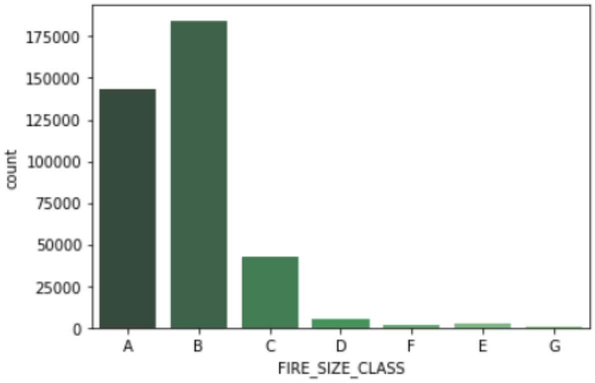
We initially had a dataset on monthly temperature of the different US cities, which would have been useful for prediction of fires. Unfortunately, the dataset was only available up till the year 2013. More importantly, the dataset only has the information of 147 cities in the US, while the “Fires” dataset contains 1,652 counties. The disparity in terms of the scale of the data available was such that if we were to join the two datasets, we would end up losing more than 95% of the data points (from 380,000 to around 17,000). Hence we decided to drop the temperature dataset and used ‘Months’ as a proxy to capture changes in season and thus temperature. The other datasets that we joined to the main “Fires” dataset only resulted in a loss of less than 1% of data points, which was acceptable.

4.2 Balancing the Data by Sampling

From data exploration, it was observed that the different classes of fire had varying numbers of occurrences, with Class A and B fires making up the majority of the total number of fires (refer to Figure 4 below), while the sum of fires under Class E, F and G only made up less than 0.65% of the total occurrences. This imbalance in the frequency of occurrences will have an impact when it comes to classification of the various fires, which we will explain in the Results section.

To overcome this, we attempted both oversampling and undersampling to balance out the different classes of fire in order to achieve better classification outcome. To further refine our model classification, we also tested a separate set of models with the various Fire Sizes converted to a binary variable of small fires (A & B) vs larger fires (C, D, E, F & G). As there were sufficient data points, with more than 50,000 in each set, we could run a balanced model without the need to oversample or undersample. In practice, this will allow the government to predict a likelihood between large and small fires and dedicate the appropriate amount of resources.

Figure 4. Distribution of fire size classes



4.3 Feature Engineering and Dimensionality Reduction

The categorical variables were converted using one-hot encoding, resulting in a sparse matrix. The dataset was then scaled using Min-Max Scaler to ensure that all the variables were given similar weight in the modelling, instead of placing more importance on variables with higher values. After which, the dataset was further processed using Principal Component Analysis (PCA) to reduce feature dimensionality.

Not all variables were selected to be run in the model eventually. We experimented with variations to be input into the final models. The choice of variables will significantly affect the dimensionality of the input data due to one hot encoding. For example, if we were to include “State” as a variable, the hot encoding will result in 46 additional variables. On the other hand, when we include “county” as a variable, the hot encoding will result in more than 1,000 additional variables. The number of PCA components to use in the modelling would be dependent on the selection of variables.

The final dataset used in the machine learning model was split into training and test sets at a 90-10 proportion, respectively.

4.4 Models Considered and Evaluation Criteria

The target variable used in the study is the extent of each wildfire. We used classification models to determine this based on the Fire Size Class. In accordance with the definition by the National Wildlife Coordinating Group, there are 7 Classes of Fire Size, which were used in the dataset:

- Class A - one-fourth acre or less;
- Class B - more than one-fourth acre, but less than 10 acres
- Class C - 10 acres or more, but less than 100 acres
- Class D - 100 acres or more, but less than 300 acres
- Class E - 300 acres or more, but less than 1,000 acres
- Class F - 1,000 acres or more, but less than 5,000 acres
- Class G - 5,000 acres or more.

We first explored multi-label classification models whereby we attempted to predict each Fire Size Class. Similar to other studies noted in the literature review, we tested with Random Forest, Support Vector Machine, Multilayer Perceptron (MLP) models using the Scikit-Learn package. We also tuned our own MLP model built using the Keras package. In general, although all models could reach an accuracy of about 63% after tuning, it was found that the models could predict Classes A and B well, but could not predict Classes E, F and G well. In fact, only the MLP models were capable of predicting all seven Fire Size classes while Random Forest and Ridge Classifier models could only predict three and two classes respectively. Even though prediction accuracy for these models was within a reasonable range, the objective of the predictive model was not achieved as it was far more important to be able to accurately predict larger fires than smaller ones.

Hence, we subsequently modified our model into a binary classification model. As mentioned in the earlier section, Classes A and B were grouped under the “small fire” category, while Classes C, D, E, F and G were classified as “large fires.” A binary classification was then done to predict whether the fire is likely to be large or small, based on geo-locational, socio-economic and climate factors. The main evaluation criteria for the binary classification models was classification accuracy, although we also looked at Recall especially for the larger fires which were of greater interest to us. We decided to use Recall as an accuracy metric because the ratio of correctly predicted classes of fires to the total number of actual fires was important for our use case, which is to derive the amount of resources to deploy to the fire to better manage our scarce resources. As such, we would need the Recall to be high so that we predict more classes of fires accurately.

Overall, due to substantial data points available, we performed random sampling from the data available to create a pool of equal number of “small” vs “large” fires, without the need to worry about the imbalanced classes in the dataset. As such, we were more confident of a well-rounded outcome with higher accuracy. The accuracy of the binary classification model was indeed higher at about 72%.

5. Models Explored

In general, the models attempted for multi-label and binary classifications were similar, although the required pre-processing, number of data points used and variable selections were slightly different. The following section describes the different models used for both multi-label and binary classification.

5.1 Model Selection and Hyperparameters Tuning

Ridge Classifier: This serves as a basic baseline model for prediction. As the relationship of the data points should be highly non-linear, this classifier is not expected to give the best performance. We explored different values for alpha (the regularisation strength) to attempt to reduce the variance of the estimates.

Random Forest: This model aggregates various decision trees and thus reduces model variance. While this model is less interpretable as compared to linear models, we can still use partial dependence plots to understand the impact of each predictor on the model if required, making it suitable for our purpose. The Gini and entropy criteria were both tested, along with balancing the weights of the classes.

Support Vector Machine: The use of SVM with a Radial Basis Function (RBF) Kernel, which can transform non-linear relationships into higher dimensional spaces with linearly-separable boundaries could be a potentially good model to adopt. Both linear and radial SVM were tested.

Multilayer Perceptron (SK Learn): The multilayer perceptron (MLP) model from SK Learn Library serves as a baseline for Neural Network model with 100 hidden layers, ReLU activation function and Adam optimiser with a learning rate of 0.0001.

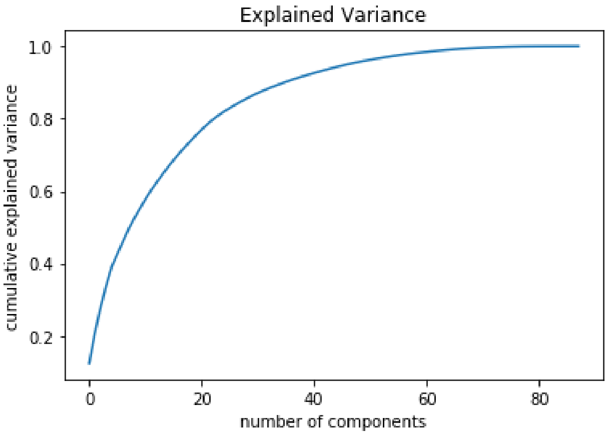
Multilayer Perceptron (Keras): Finally, we tuned our own MLP network using the Keras library. Several parameters were tested such as the number of hidden layers, the number of hidden units, batch size and learning rate. In general, it was found that overfitting can occur quite easily, hence the number of hidden layers and hidden units have to be kept at a moderately low value and epoch should not exceed 200. The learning rate using Adam optimiser was also kept low at 0.00005.

5.2 Multilabel Classification

We ran the full dataset for the multi-label models. As mentioned, we tried using both oversampling and undersampling, but neither gave improved results. This might be because oversampling created too many artificial data points that could not capture the true distribution patterns of data, while undersampling caused a dramatic reduction of data points if we were to match the number of higher classes of fires with the lower classes.

We selected “Latitude” and “Longitude” as the approximation for locations of the fires and “state_id” to control for variation of data across the various states. After one hot-encoding and fitting with PCA, we chose 40 PCA components to be input into the models as that approximately accounted for 90% of the variance in data.

Figure 5. Principal components plot for multilabel model with corresponding cumulative explained variance from PCA



5.3 Multilabel Classification - Results and Evaluation

As observed from the results in Table 1, all five models gave an overall accuracy of about 60%. However, this was because all models were capable of predicting the large number of Class A and B fires but the Ridge, Random Forest and Support Vector Machine models could only predict up to three classes of fire. Only the MLPs were able to predict more classes and generally have a higher accuracy rate. We could also see that the recall for the higher classes were very poor even for the MLPs, which shows that the models were unable to effectively predict bigger fires.

Table 1. Summary of results from multilabel classification

Type of Model	Accuracy	Recall	Time Taken to Train Model
Ridge Regression	60.3%	0.55 for Class A 0.76 for Class B 0 predictions for Classes C-G	5 seconds
Random Forest	52.2%	0.56 for Class A 0.04 for Class B 0.91 for Class C 0.01 for Class D 0 predictions for Classes E-G	10 min
Support Vector Machine (linear)	58.7%	0.53 for Class A 0.80 for Class B 0 predictions for Classes C-G	6 hours
Support Vector Machine (RBF)	63.1%	0.52 for Class A 0.61 for Class B 0 predictions for Classes C-G	12 hours
Multilayer Perceptron (SK Learn)	56.7%	0.64 for Class A 0.68 for Class B 0.01 for Class C 0.01 for Class F 0 predictions for Classes D, E, G	20 min
Multilayer Perceptron (Keras)	65.2%	N/A	23 min

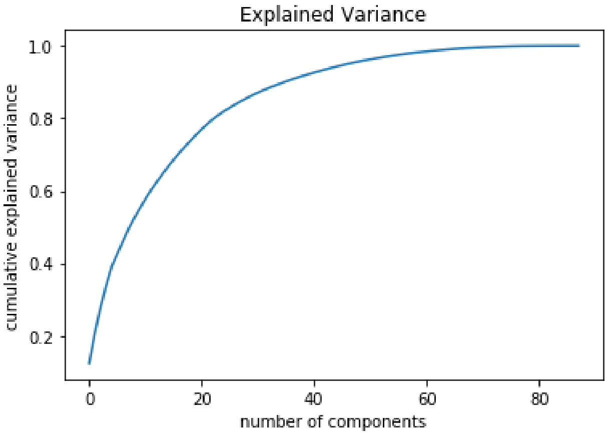
This could happen due to imbalanced data. However, even after undersampling and oversampling, the results still did not improve. In fact, when the sample size was reduced such that all 7 Classes have a similar number of data points, the models made some predictions of the higher classes of fire but the accuracy went down to only about 30%. Likewise, we tried class rebalancing in the Random Forest model and while it allowed the model to predict all seven classes, the accuracy scores were significantly inferior at 43%. It could well be that it was just difficult for the models to predict the higher classes of fire due to limited variations in the data.

Hence, while we derived interesting results from the multilabel classifications, it did not fulfil our project objective well as the bigger fires were the ones that we were more interested in. We therefore tweaked our model by classifying the fires only into two main groups: small fires (Class A and B) and large fires (Class C, D, E, F and G).

5.4 Binary Classification

For binary classification, there were sufficient data points for the two groups of fires; there were about 50,000 points for each group. In addition, instead of solely using "Latitude" and "Longitude", we made use of the "state-county" variable to control for variation across counties. One hot-encoding created more than 2,000 variables for this case. We then used PCA with 600 components, which could explain about 90% of the variance.

Figure 6. Principal components plot for binary model with corresponding cumulative explained variance from PCA



5.5 Binary Classification - Results and Evaluation

The results of the various models are shown in Table 2. We see that the accuracy of all the models improved when using binary classification, with most models able to reach above 70% accuracy. Furthermore, the recall for the bigger fires have also improved tremendously. In fact, the models seem to be able to predict bigger fires better now, with the recall for bigger fires consistently performing better than for smaller fires. As before, the MLP tuned using Keras gave the best prediction and SVM using RBF kernel also performed well, although it took several hours just to run the SVM model.

Table 2. Summary of results from binary classification

Type of Model	**Accuracy**	**Recall**	**Time Taken to Train Model**
Ridge Regression	71.1%	0.75 for big fire 0.67 for small fire	5 seconds
Random Forest	71.6%	0.77 for big fire 0.66 for small fire	10 min
Support Vector Machine (linear)	69.0%	0.79 for big fire 0.59 for small fire	2 hours
Support Vector Machine (RBF)	72.8%	0.77 for big fire 0.68 for small fire	4 hours
Multilayer Perceptron (SK Learn)	70.1%	0.70 for big fire 0.70 for big fire	20 min
Multilayer Perceptron (Keras)	72.8%	0.78 for big fire 0.68 for small fire	100 seconds (after all the parameters tuning have been done)

6. Conclusion

For the purpose of predicting the size of wildfires, the models utilised in the study prioritised prediction accuracy and recall over model interpretability since the research objective was to be able to make sound predictions to deploy adequate and timely resources to deal with wildfires. As such, model interpretability is compromised in order to obtain more accurate results. Therefore, SVM, Neural Network and Random Forest were the key predictive models explored in this study, similar to related works performed in this area.

In addition, as multi-label classifiers were unable to predict larger fires well, we simplified the final model into a binary classifier. Although this model simplification was not as ideal since the fire sizes varied over a large range over our "small" and "large" fire categories, the modification allowed us to have a higher probability of predicting larger fires, which was crucial to meeting our research objective. Both Neural Network using Keras MLP model and SVM using a RBF kernel yielded similar results, with an accuracy of 72.8% and a recall on "large" fires of about 75%.

6.1 Research Contributions

This study would benefit government agencies, citizens, environment activists, wildlife, and other countries. With this model, given the climate and the socio-economic situation of a county in the United States, we can make a reasonably accurate prediction on the occurrence of a fire and the likely extent. With that in mind, authorities and environment activists could dedicate sufficient resources to areas with higher likelihood of larger fires. The appropriate resource levels would also impact the society in terms of their safety and also the protection and preservation of wildlife in affected forests. The earlier the fire is extinguished or kept under control, the better it is for the safety of the neighbourhoods and ecosystems near the location of the fire. Other countries with high occurrences of wildfires can also implement a similar prediction model with altered parameters to suit their circumstances. This will help to benefit a wider group of people and the environment as a whole.

6.2 Limitations

In our attempt to create a good predictive model, one of the key challenges was the sourcing of relevant climate-related data at the required granularity. Many variables such as rainfall, humidity and wind speed are important to predict forest fires, as noted in similar works in this topic. As mentioned in Section 4, a key dataset that we found but were unable to utilise was the temperature dataset, as it only covered 147 cities in the United States. If these datasets were available, it would have made our prediction more robust.

Also, the dataset used in the study is a subset of the US data, specifically data from the years 2010 to 2015. Due to global warming and rapid climate change, the trained model may not be able to perform accurate predictions on recent data since the variables may have changed significantly since then.

The study was designed to focus on county-level as our most granular level for analysis. This may have restricted the predictive ability of our model since certain variables such as temperature or wind speed may occur at a higher or lower geographic level. For example, certain factors may be applicable at a country-level (aggregated) or at a city-level (more granular).

6.3 Further Works

Further work includes sourcing for other relevant datasets that could explain fire occurrences, collecting real-time data, and constructing models that can dynamically adapt to the current circumstances of the fire. Small fires can go unnoticed for days and the usage of satellites to transmit timely data can aid in wildfire predictions. Other relevant datasets include temperature, wind speed, and humidity as these have been noted as important variables in other work done by researchers. Other environmental factors that contribute to wildfires, such as types of vegetation and surrounding land quality can also be studied. Such data could be obtained through closer partnership with government agencies handling environmental and meteorological fields. The models can then be designed to incorporate dynamic data to have an updated prediction based on the current circumstances. Furthermore, the inclusion of appropriate data in the modelling may also facilitate multi-label classification with the objective of enhancing model predictions, as initially designed in this study.

References

Bureau of Labor Statistics (2020). *State and County Employment and Wages.* Retrieved from https://www.bls.gov/data/

Congressional Research Service (2019). *Wildfire Statistics.* Retrieved from https://fas.org/sgp/crs/misc/IF10244.pdf

Cortez, P., & Morais, A.J. (2007). *A data mining approach to predict forest fires using meteorological data.*

Kim, S. J., Lim, C.-H., Kim, G. S., Lee, J., Geiger, T., Rahmati, O., … Lee, W.-K. (2019). *Multi-Temporal Analysis of Forest Fire Probability Using Socio-Economic and Environmental Variables.* Remote Sensing, 11(1), 86. doi: 10.3390/rs11010086

Sayad, Y. O., Mousannif, H., & Moatassime, H. A. (2019). *Predictive modeling of wildfires: A new dataset and machine learning approach.* Fire Safety Journal, 104, 130–146. doi: 10.1016/j.firesaf.2019.01.006

Short, K. (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

Subramanian, S. G., & Crowley, M. (2018). *Using Spatial Reinforcement Learning to Build Forest Wildfire Dynamics Models From Satellite Images.* Frontiers in ICT, 5. doi: 10.3389/fict.2018.00006

Thiessen, M. (2019, October 25). *California fires are raging: Get the facts on wildfires.* Retrieved from https://www.nationalgeographic.com/environment/natural-disasters/wildfires/

United States Census Bureau (2020). *Income and Poverty.* Retrieved from https://data.census.gov/cedsci/

United States Drought Monitor (2020). *Comprehensive Statistics.* Retrieved from https://droughtmonitor.unl.edu/Data/DataDownload/ComprehensiveStatistics.aspx

Vilar, L., Gómez, I., Martínez-Vega, J., Echavarría, P., Riaño, D., & Martín, M. P. (2016). *Multitemporal Modelling of Socio-Economic Wildfire Drivers in Central Spain between the 1980s and the 2000s: Comparing Generalized Linear Models to Machine Learning Algorithms.* Plos One, 11(8). doi: 10.1371/journal.pone.0161344

Washington Post (2016). *U.S. wildfires just set an amazing and troubling new record.* Retrieved from https://www.washingtonpost.com/news/energy-environment/wp/2016/01/06/2015-wildfire-season-just-set-an-amazing-and-troubling-new-record/