Blaze Abaters
Spring 2023 DS Competition Final Report

April 4, 2023
Texas A&M University

Sreekar Annaluru

Manuel Sojan

Luis Loo

Matt Byrom

Allison Moore

Executive Summary

Wildfires are not only devastating in real life but also difficult to model in the digital world. Many scholars and industry experts have been working relentlessly to predict and mitigate wildfires effectively. In response to the Texas A&M University Spring 2023 Data Science Competition prompt, our team developed machine learning models targeted at answering our research question: **What features lead to increased fire occurrences in the Denver region, as compared to the Columbus region?** This executive summary provides an overview of our approach to answering this question, followed by the innovative insights and recommendations.

During our preliminary research, we identified that the Columbus area experienced a significantly lower occurrence of wildfires compared to the Denver area. This discovery led to our decision to pursue a Most Similar System Design methodology. Columbus and Denver are two important metropolitan areas in the United States - with comparable populations - that exist on the same latitude and are both bordered by a mountain range. Understanding why a specific region experiences less wildfires is as important as predicting the occurrence of wildfires in that area. By developing machine learning models to uncover these important features, we can aid stakeholders in their search to better understand wildfire behavior with more confidence.

Our approach to tackling this problem included collecting diverse data from multiple sources such as NASA Giovanni (NASA 2023a), NASA FIRMS (NASA 2023b), Visual Crossing (VCC 2023), et al. Most of the data was easily accessible and available in a structured format; however, some sources required the use of Application Programming Interfaces (APIs) to extract data. After filtering out data corresponding to our region, we used forward and backward selection, as well as lasso regularization, to perform variable selection. We then built Neural Network, XG Boosting, SVM, and Logistic Regression models. We trained these models on one region and tested it on the other region.

Following a completely data-driven approach, we successfully validated our problem statement. Each region experiences its own set of important features when predicting the 'brightness' associated with the occurrence of a fire. For example, a model built using Denver data includes the Normalized Differential Vegetation Index (NDVI), an important variable which represents vegetation. However, this variable is not identified as important for the Columbus model. Additionally, large values in the features with high correlation with the occurrence of fires (such as solar radiation and wind) in the Denver region indicate that the Denver region is more susceptible to wildfires than the Columbus region.

Although there are several features unique to each region, there are some that were identified as important to both regions. These variables, such as dust scattering and transpiration - which were identified by lasso and forward/backward selection in both models - suggest great importance in identifying the probability of a wildfire. However, these features and insights warrant additional research. Based on our results, we strongly believe ensemble models yield pragmatic results. Wildfire prediction is a complex, iterative process which requires improvements with each iteration. We recommend that scholars consider utilizing a similar approach in future research.

Methodology and Data

In our approach to tackle the 2023 Texas A&M University Student Data Science Competition prompt, our team used the six-step methodology shown in Figure 1.

Figure 1. Displays the Blaze Abaters six-step methodology for the Texas A&M University Data Science Spring 2023 Competition on the Behavior of Wildfires.

In order to construct the most accurate model possible, independent research was the foundation of our methodology. Approximately 70% of the allotted time was spent researching previous models, potential variables and datasets, and gaining industry knowledge. Once our team had a holistic understanding of the problem, we generated several potential research questions. Ultimately, we narrowed these questions to:

1. *What features lead to increased fire occurrences in Denver, as compared to Columbus?*
2. *What features demonstrate synergic effects?*
3. *Are there any peculiar variables that are affecting the wildfire?*

These questions arose from the popular political science methodology of Most Similar Systems Design (MSSD), where the sample space is selected based on the similarity of specific characteristics but differ in "one crucial respect (related to the hypothesis of interest)" (Halperin and Heath 2017, 239).

Utilizing the MSSD approach - and considering the data discovered in our preliminary search - we selected two largely populated cities within the continental United States for our analysis: Denver, CO and Columbus, OH. These two cities are comparable in terms of population, they are both located on the same latitude, and they are both uniquely situated on the interior side of a mountain range (i.e., Denver is

east of the Rocky Mountains and Columbus is west of the Appalachian Mountains). The crucial aspect where these two locations differ is the number of wildfires which occurred in our specified time period (2014-2022). Of the 623 fires which occurred, 438 (~70%) of these were in the Denver area.

Next, we finalized our data sources and collected the raw data on our selected dependent and independent variables. Our variables fall into four categories:

1. Fire Incident Data. This includes identifying fire information such as the geolocation of individual incidents, the date, and the size of the fire.
2. Weather Data. Weather data comprises a large portion of our features and includes features such as wind speed, temperature, and dust scattering.
3. Vegetation Data. Only two variables fall into this category: Landsat Normalized Difference Vegetation Index (NDVI) and Leaf Area Index (LAI).
4. Geospatial Data. This data includes ArcGIS geospatial files such as terrain slope, electrical power lines, and national park areas.

For more information on the specific variables, please refer to Appendix A: Data Dictionary. We extracted all of our data based on the boundaries of the latitude (Lat) and longitude (Lon) of our regions. For the Denver region, our boundary was Lat: 39.25^{o} to 40.25^{o} and Lon: -104.56^{o} to -105.56^{o}. For the Columbus region, our boundary was Lat: 39.25^{o} to 40.25^{o} and Lon: -82.05^{o} to -83.05^{o}.

Once all data was collected, data wrangling was conducted to format and compile the data into a size and format appropriate for answering our research questions. The data was combined into monthly observations from January 2014 to December 2022, with the total number of fires in a month recorded as the 'Occurrences' variable, for each location separately. Every month is represented, even if no fires occurred. The final data set consists of 4 identifying features, 36 independent/dependent variables, and 216 observations (108 from Denver and 108 from Columbus). Including the "Occurrences" column, there were 41 columns in our final data set for each region.

Following the compilation of data, we conducted exploratory data analysis (EDA) and data visualization to familiarize ourselves with the basic relationships between our variables. The insights gained from this process enabled us to narrow down our features and select appropriate machine learning models for answering our research questions. These insights and the selected models are discussed in detail in the next section. The primary response variable selected for modeling is 'brightness,' which aids in the detection of a fire (i.e., whether or not a fire has occurred). However, 'confidence' was used as the response variable in order to determine the best fit for logistic regression. Each model uses a training and validation set which consists of training on the Denver region data and testing on the Columbus region, and vice versa.

Lastly, after modeling our data, we assessed the performance of each model and developed insights for discussion. We close the paper by offering some recommendations for future study, as well as possible improvements to our methodology.

Modeling and Analysis

To model our data, we generated and compared the accuracy of four machine learning algorithms. Before fitting any models, we utilized Forward/Backward Selection and Lasso to determine which features to include in each model.

Forward/Backward Selection & Lasso

We first performed forward and backward selection (F/B) to identify the significant variables in the Denver, CO and Columbus, OH regions. These variables then gave us an idea on which input variables to include in the XGBoost and Neural Network models, for each region. For Logistic Regression and SVM, we

used the output of lasso regression for deciding which variables to include, as it performed better in terms of accuracy than F/B selection. We performed both the F/B and lasso variable selection - and implemented Logistic Regression - in R.

Our final dataset had 41 variables and after performing lasso regression and F/B selection, independently, we ended up with 21 variables for the Denver region and 11 variables for the Columbus region (refer to Appendix B). The lists in Appendix B assume a response variable of 'brightness'. Since we used 'confidence' as our response for logistic regression, we swapped out 'confidence' with 'brightness' for the lasso regression variables when fitting the logistic model.

Gradient Boosting (XGBoost).

We built a regression boosted tree, using "brightness" as our response variable. We had to tune several parameters for fitting this model. (Refer to XG Boost script to see list of tuning parameters.) The optimal parameters chosen for the Denver and Columbus model, using cross validation, for the 9th iteration are listed below:

1. max_depth: 20
2. min_child_weight:10
3. colsample_bytree: 0.9
4. n_estimators: 600
5. reg_alpha: 0.5
6. reg_lambda: 2
7. gamma: 1

Figure 2 reveals the test error rate for each iteration of the model, for each region. The RMSE indicates the average distance between the predicted values from the model and the actual values in the dataset, in Kelvin. As mentioned earlier, the Denver model was trained using Denver data and tested using Columbus data, and vice versa for the Columbus model. Figure 3 shows the feature importance plot for the 9th iteration of the model, using the Columbus and Denver data.

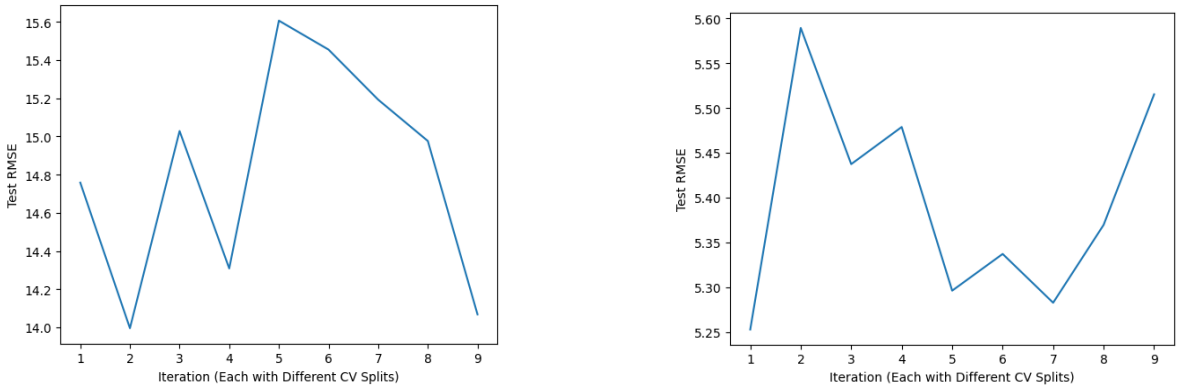

Figure 2. RMSE: trained on Columbus, tested on Denver (left) and trained on Denver, tested on Columbus (right).

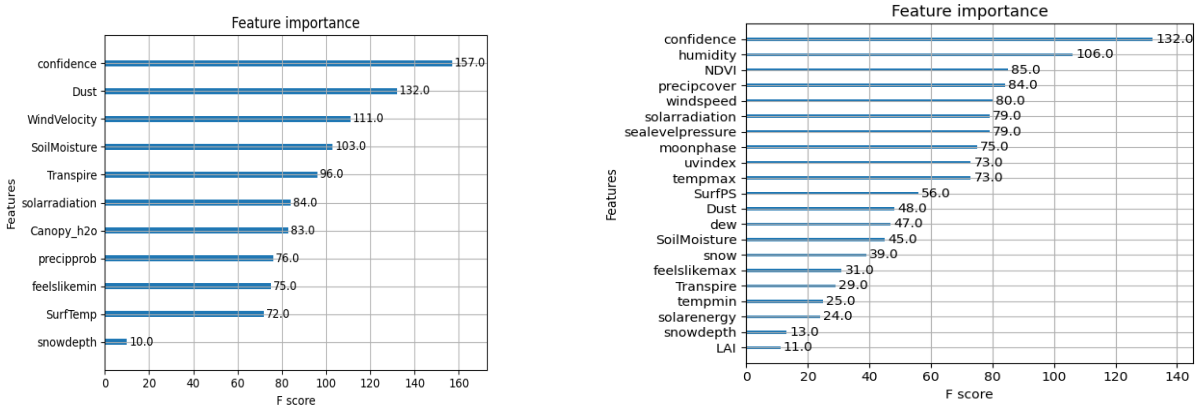

Figure 3. Feature importance: trained on Columbus, tested on Denver (left) and trained on Denver, tested on Columbus (right).

Support Vector Machine (SVM)

SVM was rigorously tested utilizing a grid search and cross validation for each region. The parameters used in the grid search include: the regularization parameter, C; the kernel coefficient, gamma; and the kernel type. Utilizing the features selected during lasso regression, several models were fit for Columbus and Denver with the response variable 'brightness.' To better understand how the variables impact the response variable, each region was treated as a training set, with the alternative region acting as the validation set.

For each model, the best results occurred with non-linear kernels, with various other parameter selections unique to each model. Figure 4 displays the learning curves for the two variations of training and testing on opposite regions. From these results, we conclude that SVM is best when training on Denver and testing on Columbus, as this model exhibits the lowest RMSE. To determine feature importance, each model was forced into a linear kernel. Because most models performed best on non-linear kernels, the feature importance shown in Figure 5 are only estimates.

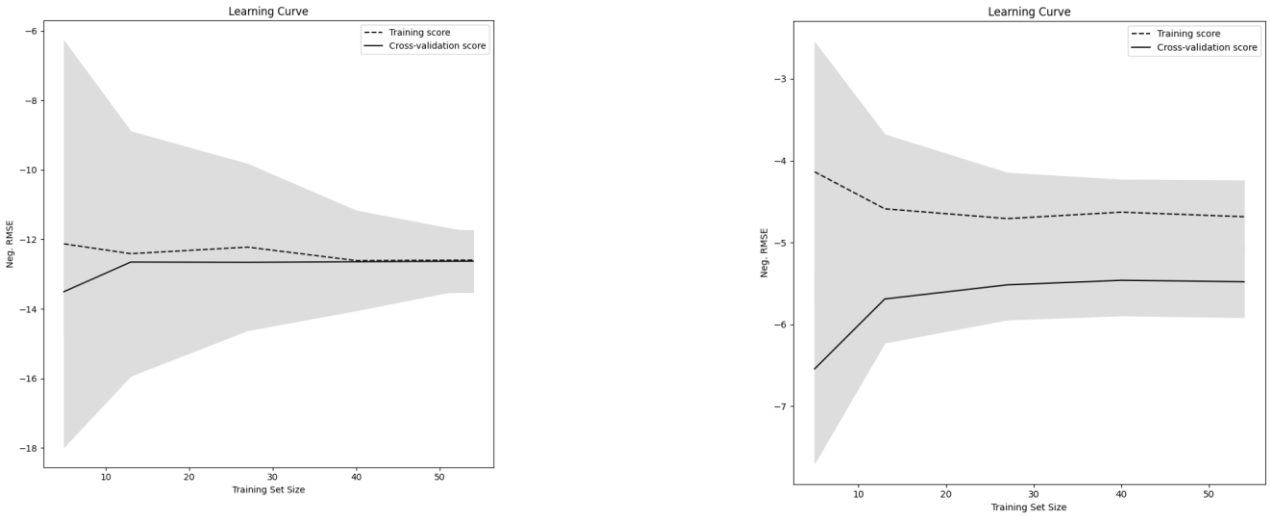

Figure 4. SVM Learning Curves. (Left) Trained on Denver, tested on Columbus, optimal RMSE=8.86. (Right) Trained on Columbus, tested on Denver, optimal RMSE=16.2.

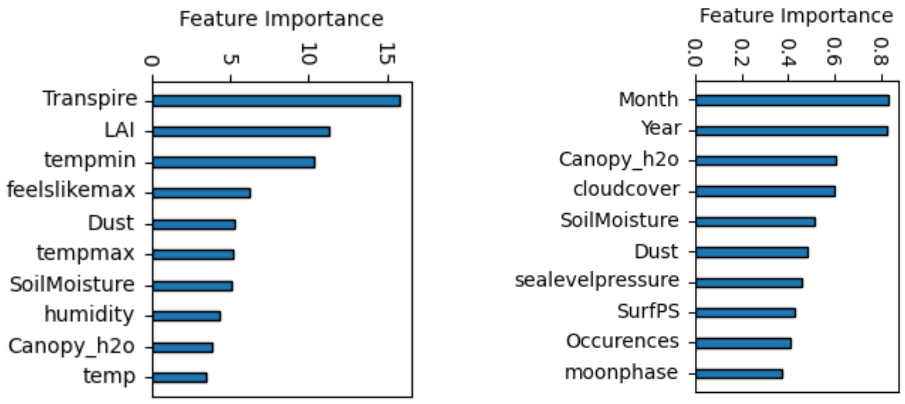

Figure 5. Top 10 Feature importance. (Left) Trained on Denver, tested on Columbus and (Right) Trained on Columbus, tested on Denver.

Logistic Regression.

For building logistic regression, we used 'confidence' as our response, so we converted the confidence column into binary format with a threshold of 70%. In other words, if the confidence is greater than 70% then it is coded as 1, else 0. Both regions' models have shown pretty decent actual test accuracy of around 67%. Hence, we strongly believe Logistic Regression is one of the best models to understand the wildfire.

ROC plots for the two models provide us solid evidence that these results are reliable but can be improved.

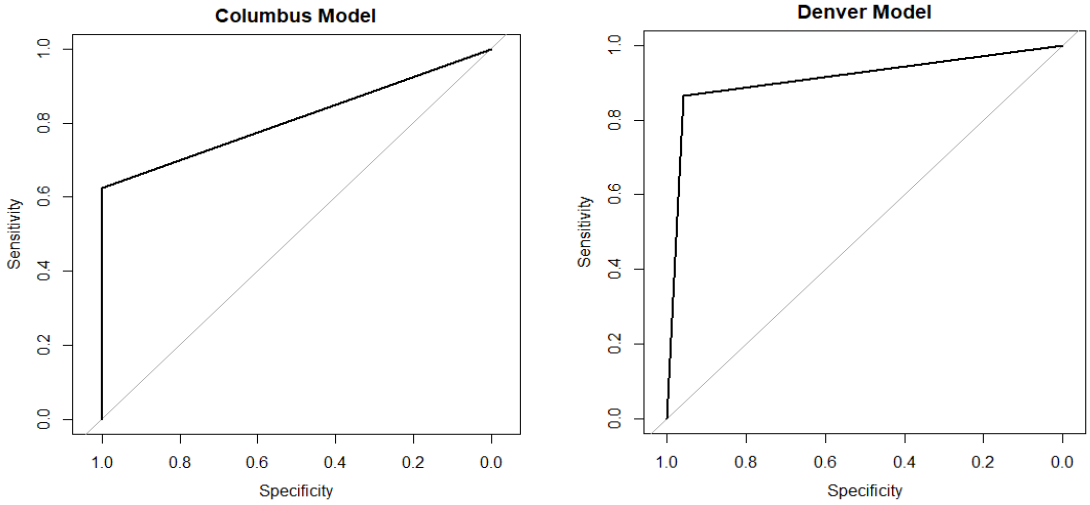

Figure 6. ROC plots for Columbus and Denver models.

Neural Network (NN).

We built a 3-layer neural network with "ReLU" activation layer and "adams" optimizer. Input layer has 32 neurons followed by 16 and 8 to predict brightness of the spot. We generated mean squared error

and mean absolute error as the outputs. We normalized the variables and reverse normalized the outputs. Models were fit with 25 epochs and gradients of batch size 32 were used. Table 1 lists the results from our Neural Network model. The model trained with the Columbus region resulted in the best performance.

Table 1. Neural Network Results

Region/Parameters	Test MSE	Test MAE	Parameters
Columbus	70.81	6.18	11
Denver	168.85	13.76	21

The most important takeaway from this model is that Neural Networks are quite reliable in predicting the wildfire. Our model is learning after each epoch and the actual test results are pretty reliable which implies we are going in the right direction.

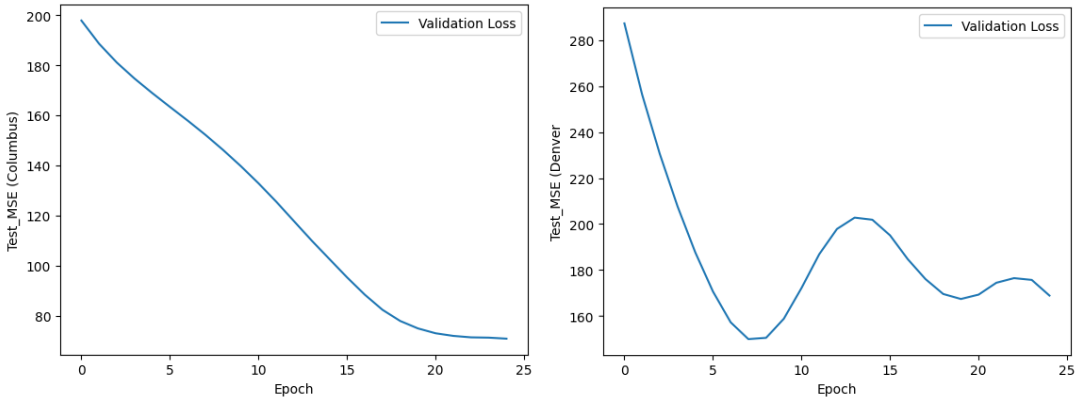

Figure 7. Test MSE vs Epoch for Columbus (left) and Denver (right).

Results and Summary

The goal of developing the models discussed above is to validate our problem statement. Below, Table 2 summarizes our results. You can find some general facts and intuition around our problem statement. The Facts and Intuition column lists our key findings and discoveries. The majority of these findings have been fully validated by the machine learning models. Our reasoning and/or supporting evidence for these findings is listed in the comments section. There remain a few items that are currently still in development, such as uncovering the importance or necessity of implementing an ensemble method.

Table 2. Compilation of Blaze Abaters machine learning models and research. Many goals have been validated; however, there are some that remain in progress.

Facts/Intuition	Validated by model(s)	Comments
More fires in Denver area compared to Columbus area	Yes	Occurrences Variable, and Dashboard
Multiple factors are responsible for the wildfire	Yes	Variable Selection
Denver area is more susceptible to wildfires	Yes	Brightness and Confidence, Dashboard
More healthy vegetation in Columbus area	Yes	NDVI Index
Interaction of variables is crucial for understanding wildfire	Yes	Difference in models for the two regions, Optimal variable selection
Indirect variables might affect wildfire	Yes	Inclusion of moon phase in the models
Reliability of results	Yes	Actual test error
Importance of ensembling	Pending	Work in progress
Accuracy of results	Pending	Work in Progress
Going in the right direction	Yes	Our entire work.

Conclusions and Recommendations

After iterating through our selected models, we have come to several conclusions. First, Logistic Regression demonstrated the best performance for the response variable ‘confidence,’ which was coded for classification. However, the remaining models provided similar results when predicting the continuous values seen in ‘brightness.’ Although Logistic Regression performed the best overall, the other models contributed to uncovering some additional insights, as well as identifying areas which constitute additional research.

The importance of the lunar phase (‘moonphase’) in XGBoost and SVM in the Denver models provides for an interesting point of discussion. Within current scientific literature, there is little content on how the lunar phase can impact environmental factors, such as humidity and rainfall. Kohyama and Wallace (2016) suggest a strong relationship between lunar phase and rainfall; however, their work is focused on how this impacts humidity within the tropics. We strongly believe that further research is needed to better understand how the lunar phase impacts the occurrence of wildfires. There may be further implications of the lunar phase, such as increased surface temperatures on Earth during periods of night. Additional research should be conducted on how these changes impact wildfires at night.

Transpiration also appears to play an important role in our models. A brief review of current scientific literature indicates that most research on the relationship between transpiration and wildfires is one directional (i.e., focused on how fires impact transpiration). However, our study has indicated that additional research may be necessary on how natural fluctuations in transpiration can affect the occurrence and severity of wildfires.

Our ArcGIS Blaze Abaters map also provides additional insights on fire behavior in the Denver and Columbus regions:

1. There is a higher density of fires nearer diverse topographical locations (e.g., greater occurrence of fires within the specified regions on the side closest to the mountains).
2. Columbus sees much lower wind speeds than Denver.
3. Columbus fires tend to occur within the confines of national park boundaries.

Users can explore these insights by toggling various overlays on the Blaze Abaters ArcGIS Dashboard. As seen on these maps and in our feature importance plots, wind velocity played an important role in most of our models. We believe targeted studies should be conducted in our specified regions to better understand how air flow in ridges and draws may impact fires. The line of demarcation along a ridgeline is a well-known danger zone for aircraft, as this area results in substantial turbulence and disrupted air (Day 2016). It is possible that these areas of disrupted air flow could lead to increased risk of wildfires.

This research offers a unique approach to the study of wildfire behavior. Our Most Similar Systems Design uncovered several interesting differences between Denver and Columbus that likely impact the occurrence and severity of wildfires in the two regions. However, due to time constraints, there are several improvements to our model that we recommend. First, future research should consider how the use of prescribed fires affects naturally occurring wildfires. Using the data provided by the National Interagency Fire Center (NIFC), researchers could incorporate the fire classification codes in their analysis to determine if there are less occurrences or less severity in wildfires following prescribed burns (NIFC 2022). Second, time of day should be incorporated into future studies. This information is also available in the NIFC dataset.

Bibliography

Day, Joshua. 2016. “The Wind Zone Model.” *Risk Management Magazine.* U.S. Army Combat Readiness Center. https://safety.army.mil/MEDIA/Risk-Management-Magazine/ArtMID/7428/ArticleID/5514/The-Wind-Zone-Model.

Esri. 2023. “Terrain: Slope Map.” https://elevation.arcgis.com/arcgis/rest/services/WorldElevation/Terrain/ImageServer.

Esri Data and Maps. 2022. “USA Parks.” https://services.arcgis.com/P3ePLMYs2RVChkJx/arcgis/rest/services/USA_Parks/FeatureServer/0.

Federal_User_Community. 2022. “US Electric Power Transmission Lines OGC.”. https://services2.arcgis.com/FiaPA4ga0iQKduv3/arcgis/rest/services/US_Electric_Power_Transmission_Lines_OGC/OGCFeatureServer.

Halperin, Sandra and Oliver Heath. 2017. “Political Research: Methods and Practical Skills.” Oxford University Press: Oxford.

Kohyama, Tubasa, and John M. Wallace. 2016. “Rainfall Variations Induced by the Lunar Gravitational Atmospheric Tide and their Implications for the Relationship Between Tropical Rainfall and Humidity.” *Geophys. Res. Lett.*, 43, 918– 923, doi:10.1002/2015GL067342.

National Aeronautics and Space Administration. 2023a. “Giovanni:The Bridge Between Data and Science v 4.38.” *EarthData*. https://giovanni.gsfc.nasa.gov/giovanni/.

National Aeronautics and Space Administration. 2023b.”Fire Information and Resource Management System: FIRMS Fire Map.” https://firms.modaps.eosdis.nasa.gov/map/#d:24hrs;@0.0,0.0,3z.

National Interagency Fire Center. 2022. “Wildland Fire Incident Locations.” https://data-nifc.opendata.arcgis.com/datasets/nifc::wildland-fire-incident-locations/about.

Visual Crossing Corporation. 2023. “Total Weather Data.” https://www.visualcrossing.com/weather-data.