John Hemmer and Arno Abrahamian

30 November 2023

Geospatial Analysis of Wildfire Susceptibility and Prediction in California

Abstract:

Wildfires pose a significant threat to ecosystems, human lives, and property, with a large percentage of the world's wildfires primarily concentrated in California. We utilize various forms of machine learning analysis to predict their occurrence and spread. We concentrate on a range of variables including the NDVI, distance to roads, and wind speed to identify key predictors of wildfires. Our machine learning methods, including a random forest model, a weighted sum analysis, and a hot-spot analysis allowed us to employ our variables and assess the state of California. We deduced that areas of higher elevations, higher wind speeds, increased vegetation density, and lower precipitation, primarily concentrated in Northern California, are more prone to wildfires. This research and its results aims to guide government agencies to develop wildfire prevention and mitigation efforts by utilizing predictive analysis for the California region.

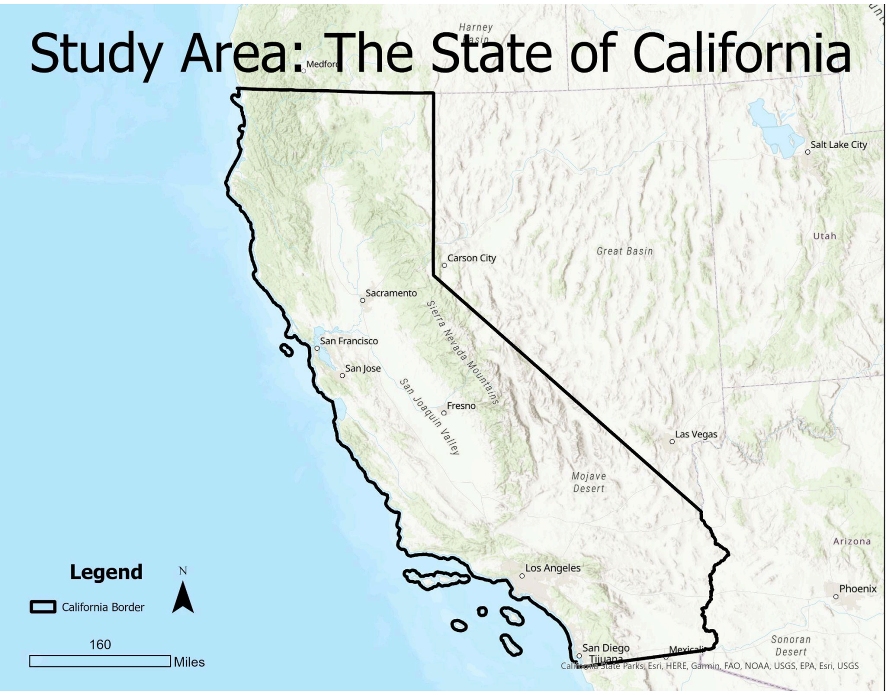
Introduction:

Wildfires alter the structure of ecosystems, affect ecological processes and functions, threaten human lives, and increase fire-suspension costs (Zhang, Lim, Sharples, 2016). Identifying fire-prone areas and understanding the underlying variables will allow us to act accordingly for fire suspension and mitigation. Wildfires are the most common natural disaster in the United States, specifically the state of California. Every year, the coverage of wildfires increases, averaging over 1.6 million acres and over 7000 wildfires over a five-year average (California Fire, 2023). So far in 2023, 10 more wildfires and over 50,000 more acres burned compared to the whole year of 2022. California leads the United States and many parts of the world in the number of fires and acres burned, double the statistics of the second-closest state which was Texas. Most wildfires are of anthropogenic origin, either deliberate or accidental, which indicates the potential relationship between fire occurrence and socioeconomic factors (Zhang, Lim, Sharples, 2016). Looking at wildfire statistics and hypothetical factors, we grew curious about the reasons why California is the leader in wildfires. Such factors included the Mediterranean climate of the state, characterized by hot, dry summers and cool, wet winters (Gabriel, Opitz, Bonneu, 2017). As current and potentially long-term residents of California, we developed concerns about the risk of wildfires in the areas we will choose to reside. This analysis is of concern to various residential dwellers throughout the numerous communities throughout California. To further analyze this, we propose the question: What areas of California are wildfires more concentrated in and what variables affect the perimeter spread of wildfires? Additionally, how can we use this information to predict the future spread and locations of wildfires? We dive into a geospatial analysis to help unpack our hypothesis.

Study Area:

As mentioned previously, our chosen study area is the state of California. It is the highest populated state with 38.9 million residents and 163,696 square miles. It is the leading state for wildfires and has the most diverse terrain out of any U.S. state. Having a considerably larger number than its neighbors, California had a mean of 8,292 wildfires and 974,980 acres burned throughout the years of 2000-2022. The climate in California varies throughout the year and by region, however, it is commonly dry in the majority of the areas within the state and has plants/trees that are considered to be highly flammable. As previously indicated that California's Mediterranean climate is a concern, we want to analyze a variety of climate-related factors to confirm that they play a role in the high number of wildfires. The first step in understanding why wildfires are so plentiful is figuring out what variables contribute the most to wildfires. Understanding the terrain and climate of the state will allow us to understand how these disasters form, their location, and predictive mitigation efforts we can implement.

Figure 1: The study area: the state of California.



Data:

The purpose of our study is to find the variables that are the best predictors for wildfire occurrence as well as predict the locations of wildfires. Therefore, our variables are going to play some role in wildfire development or spread. After consulting various wildfire prediction articles, we came up with the following explanatory variables: Normalized Difference Vegetation Index (NDVI), Distance to Roads, Precipitation, and Wind Speed. NDVI is a numerical indicator, provided via remote sensing, quantifying vegetation health and density. It is one of the most commonly used indexes for live fuel moisture content (Zhang et al. 2016, 1208). Distance to roads was also selected because firefighting resources are also further away from potential fire outbreaks, leading to larger spread and indicative of heavier forested areas (Lan et al. 2023). Precipitation is a weather factor that plays a large role in wildfires, as places without rain will be drier and more likely to ignite (Gabriel 2017, 89). Wind was included as a predictor variable because it "may fan the flames and can lead to a rapid spread of fire and sparks across large areas" (Gabriel 2017, 90).

The data was mainly in the format of WGS 1984, with the predictor variables all being raster datasets except for the California Roads, which like the fire data is a vector shapefile. The precipitation and wind speed datasets were averaged over the year 2022. Since 2022 is the most recent complete year, it was selected to be the input for our explanatory and response variable data.

Table 1: Data attributes and sources

Name	Attributes	Link
California NDVI	WGS 1984. Red, green and blue band data for California. 60cm	Link
Fire Data (firep22_1)	NAD 1983 California (Teale) Albers (Meters). Fire perimeter for all of 2022	Link
California Roads	WGS 1984 Web Mercator. Shapefile of all major roads in California.	Link
Precipitation	WGS 1984. Precipitation total, chosen rectangular region	Link
Wind speed	WGS 1984. Wind Speed mean for 2022.	Link
Census Tracts	WGS 1984. California Census tracts. Shapefile, polygons.	Link
California Boundary	WGS 1984. California boundary. Shapefile, polygons	Link

All data layers were downloaded and a folder connection was added, and the data was added to the map. Everything was projected to NAD 1983 California, as this projection has the least distortion when dealing with California, our study area. The data was then clipped to the California state boundary. For the raster datasets, a raster clip was used, and the “Use input features for clipping geometry” was selected. This makes the extent of the raster dataset cut down to the California state boundary.

Finding a way to prepare our data in a way that could be put into a supervised learning model proved to be quite challenging. Our input training features was a background of census tracts with the fire perimeters unioned on top (vector). Before creating the union, we added a new field to each layer called Fire, and gave a value of 1 on the fire perimeters and value of 0 on the census tract. After Erasing the fire perimeters from the census tracts then unioning them we get polygons of fire that have Fire value of 1 and census tracts - the perimeters with values of 0. This layer is called final_union. This fire value became our variable to predict. This is important because it gives us more instances of where there are not fires. Initially, we had the background of California be a value of 0, which would work if it were a raster layer because then every cell would be a value of 0. However, as a vector layer, there is just one polygon of 0 meaning just one large instance of no fire. This output was also exported to a raster as a raster dataset of this is needed later.

For the roads dataset, we want a raster of the distance to roads. However, the roads dataset is vector and just has the locations of roads. First a Euclidean distance was run on the California roads to get a raster that shows distance away from the roads. Using the Raster Calculator, a $\log_{10}$ was taken on the distance away from the road's raster. This gives a much lower max value for distance, and brings the data closer together. Without taking this log, or even other methods such as dividing by average, the Random Forest model was outputting lots of blank areas near roads. Taking this $\log_{10}$ transforms the data in a way that the Random Forest can handle better.

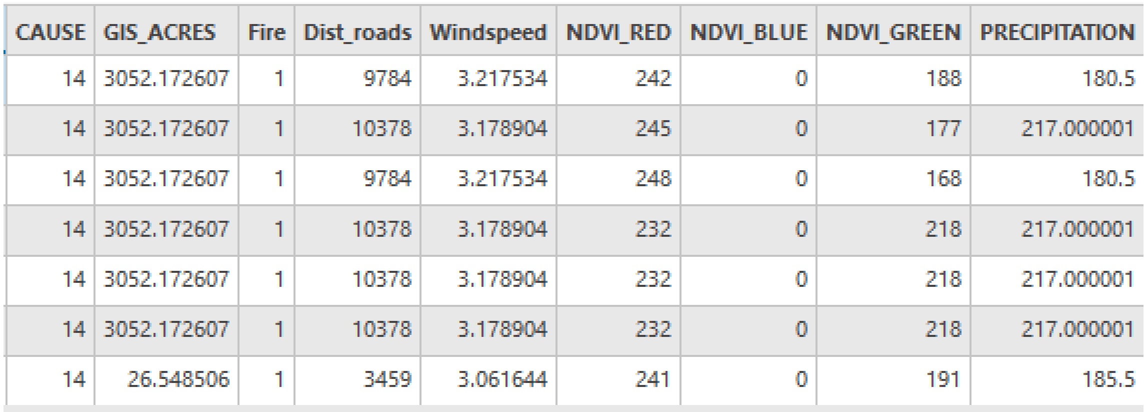
Making sure the explanatory variables were able to be the correct format for different tools was also a challenge as well. The NDVI bands were also a bit troubling to get into point data as well since raster to point only contains the value for the red band and not blue or green.

Under raster functions go to the extract band. Select the projected NDVI raster layer and delete the combination bands. Then under the band select 1. This is the red band. Repeat this but select 2 then 3. There should now be 3 new raster layers each representing the different bands. Rename layers to NDVI_red for band 1, NDVI_green for band 2 and NDVI_Blue for band 3.

Run raster to point on all three layers naming them NDVI_RED_POINTS and likewise for the other 2. Use value as the field and run. Run a spatial join on the fire_perimeters layer and add all the new raster point layers from the NDVI (i.e. NDVI_RED_POINTS).

For the other raster predictor variables (precipitation, distance to roads, wind speed), a raster to points was run. These and the individual NDVI points were spatially joined to the final_union layer. Under output fields keep only grid code and select mean for merge rule. When there are multiple points within one of the census tracts or fire perimeters it will take the average of all the points, instead of just the first point that comes (default). Under properties rename the field name and alias to the variable name (i.e. NDVI_RED_POINTS). Go to the field view and hide all of the other variables that show up with the join.

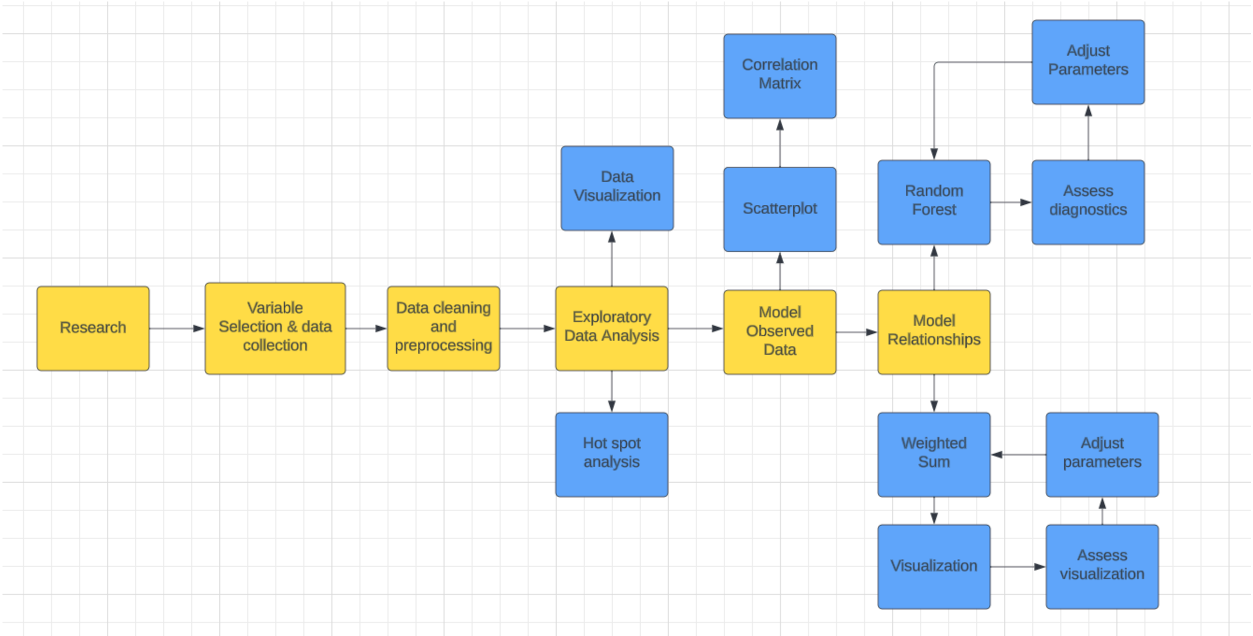
Table 2: Final_union table showing the explanatory variables and other fire data



CAUSE	GIS_ACRES	Fire	Dist_roads	Windspeed	NDVI_RED	NDVI_BLUE	NDVI_GREEN	PRECIPITATION
14	3052.172607	1	9784	3.217534	242	0	188	180.5
14	3052.172607	1	10378	3.178904	245	0	177	217.000001
14	3052.172607	1	9784	3.217534	248	0	168	180.5
14	3052.172607	1	10378	3.178904	232	0	218	217.000001
14	3052.172607	1	10378	3.178904	232	0	218	217.000001
14	3052.172607	1	10378	3.178904	232	0	218	217.000001
14	26.548506	1	3459	3.061644	241	0	191	185.5

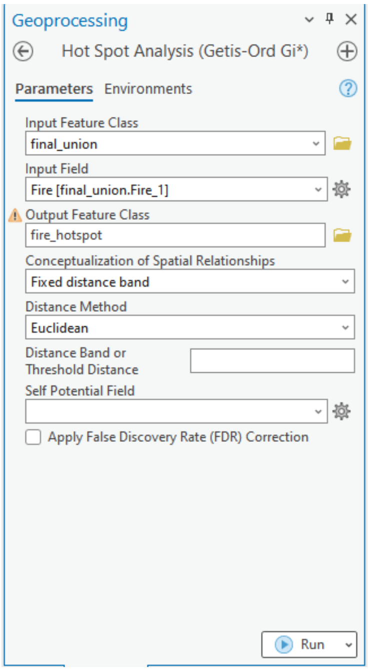
Methodology/Approach:

Figure 2: Workflow Diagram



It must be reiterated that the main goal of this analysis is to identify the key factors in wildfire occurrence as well as creating a model that can predict where wildfires will occur. One of the early steps in a machine learning workflow is exploratory data analysis. We first took a look at the distribution of wildfires via a hot spot analysis (Getis-ord Gi*). Spatial autocorrelation is a second-order effect that states there is a tendency for similar values to be clustered together or dispersed across space, and Getis-ord Gi* seeks to identify and analyze these clusters of high or low values also known as hotspots.

Figures 3 - 6: Parameter windows for hot spot analysis, scatterplot matrix and forest based regression analysis



Input Feature Class: final_union

Input Field: Fire [final_union.Fire_1]

Output Feature Class: fire_hotspot

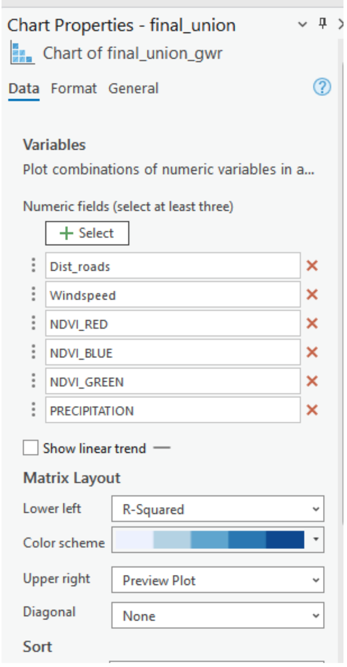
Conceptualization of Spatial Relationships: Fixed distance band

Distance Method: Euclidean

Distance Band or Threshold Distance:

Self Potential Field:

☐ Apply False Discovery Rate (FDR) Correction



Variables

Plot combinations of numeric variables in a...

Numeric fields (select at least three)

- Dist_roads
- Windspeed
- NDVI_RED
- NDVI_BLUE
- NDVI_GREEN
- PRECIPITATION

☐ Show linear trend

Matrix Layout

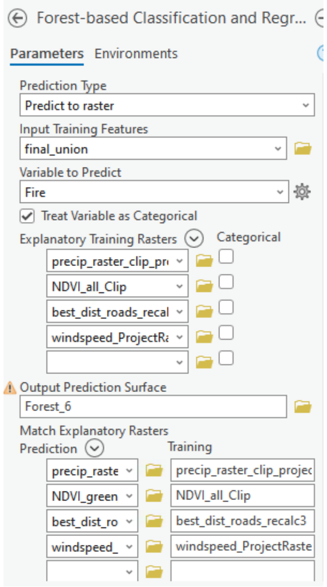
Lower left: R-Squared

Color scheme:

Upper right: Preview Plot

Diagonal: None

Sort



Prediction Type: Predict to raster

Input Training Features: final_union

Variable to Predict: Fire

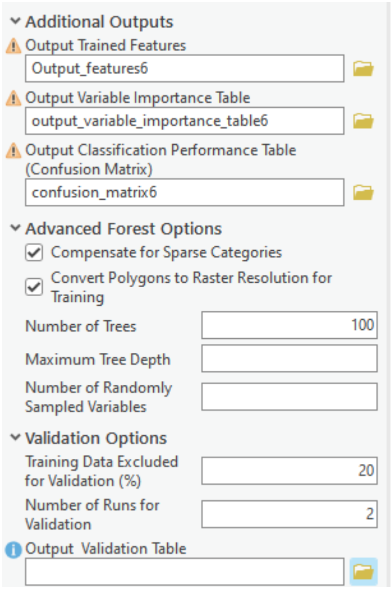
☑ Treat Variable as Categorical

Explanatory Training Rasters	Categorical
precip_raster_clip_pr	☐
NDVI_all_Clip	☐
best_dist_roads_recal	☐
windspeed_ProjectRa	☐
	☐

Output Prediction Surface: Forest_6

Match Explanatory Rasters

Prediction	Training
precip_raste	precip_raster_clip_projec
NDVI_green	NDVI_all_Clip
best_dist_ro	best_dist_roads_recalc3
windspeed_	windspeed_ProjectRaste


Additional Outputs

Output Trained Features: Output_features6

Output Variable Importance Table: output_variable_importance_table6

Output Classification Performance Table (Confusion Matrix): confusion_matrix6

Advanced Forest Options

☑ Compensate for Sparse Categories

☑ Convert Polygons to Raster Resolution for Training

Number of Trees: 100

Maximum Tree Depth:

Number of Randomly Sampled Variables:

Validation Options

Training Data Excluded for Validation (%): 20

Number of Runs for Validation: 2

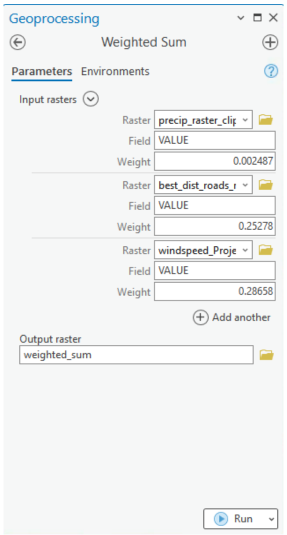
Output Validation Table:

Next we took a look at how the explanatory variables were related to each other by creating a scatterplot matrix. Although it is common for the response variable to be included in this scatterplot matrix, our response variable is binary: 0 or 1. Scatterplots are not very appropriate when one variable is categorical and the other is continuous. This was carried out by selecting the create chart, and scatterplot matrix on the layer: final_union. All the explanatory variables were then selected in the chart pane (not fire). R values were selected for bottom left with a preview plot in the top right corner.

Since Random Forest also allows raster input layers as explanatory variables, most of the explanatory variables actually did not have too much processing. The rasters were all added to the project, projected to NAD 1983 and then clipped to the California state boundary.

In order to perform the weighted sum, each weight was (1/average) in order to normalize the values. Without this, the high-valued precipitation values would be much more important even if all the variables had the same weight. Average is found by going to properties of each layer, source, statistics and mean. Afterwards a raster calculator is run on the output with (weighted_sum_output/max). This gets a value between 1 and 0, with higher values predicting higher likelihood of fires. A last raster calculator is run with the (weighted_raster_normalized - fire_results).

Figure 7: Weighted sum parameter window



Geoprocessing

Weighted Sum

Parameters Environments

Input rasters

Raster: precip_raster_clip
Field: VALUE
Weight: 0.002487

Raster: best_dist_roads_r
Field: VALUE
Weight: 0.25278

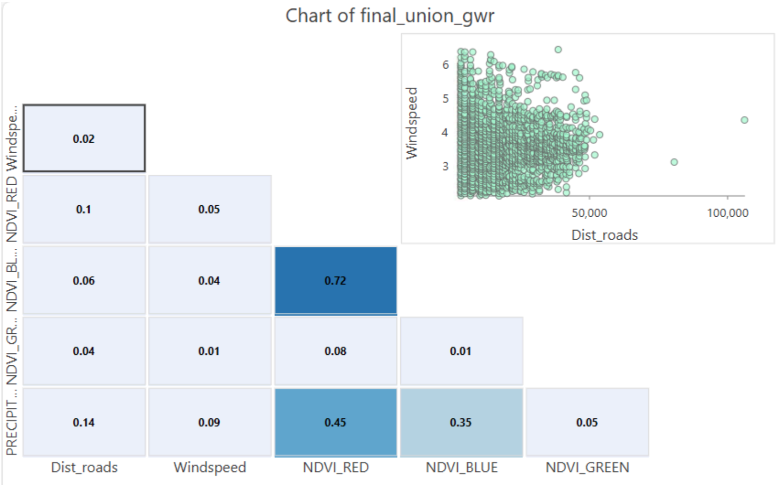
Raster: windspeed_Proje
Field: VALUE
Weight: 0.28658

Add another

Output raster: weighted_sum

Run

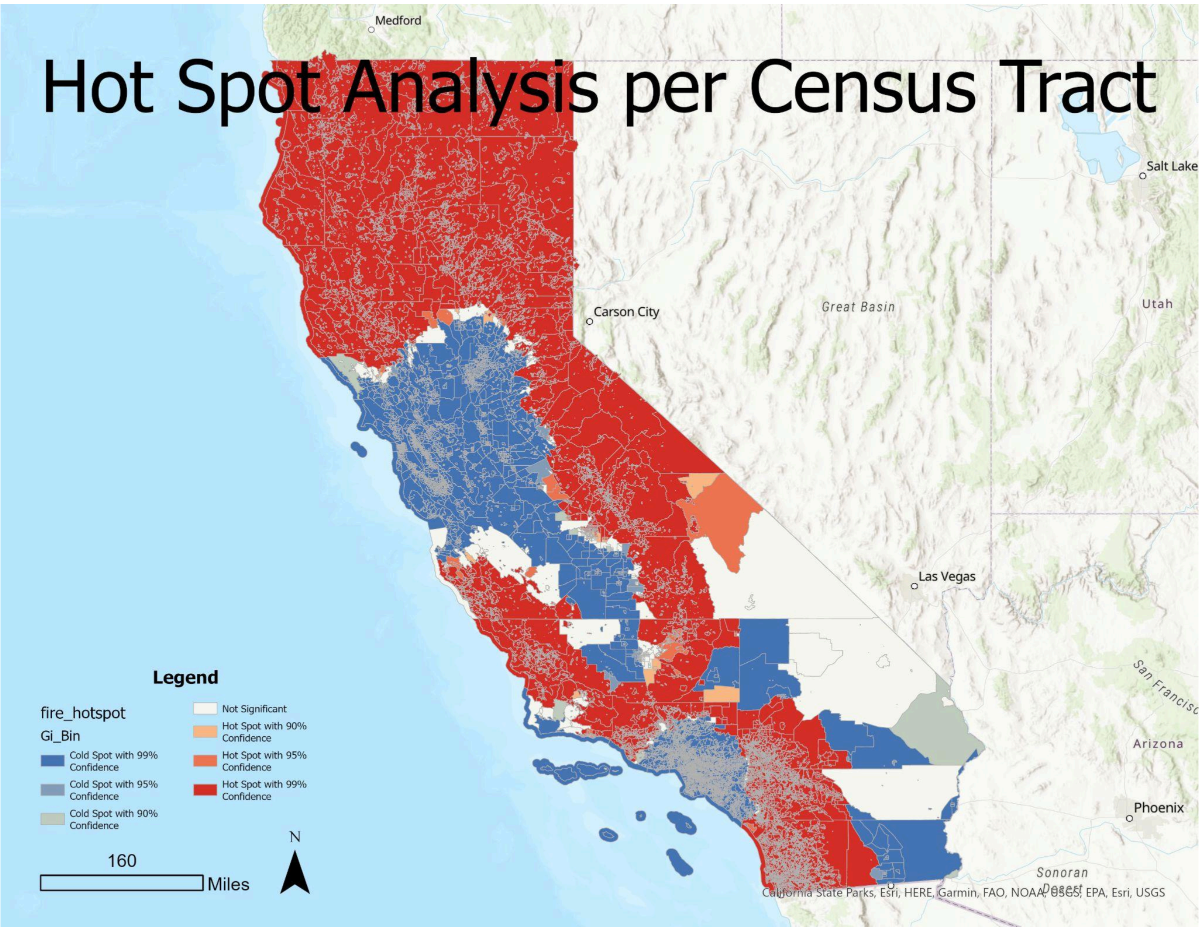
After each model was run, the diagnostic windows were checked. For the forest-based regression model, a variable importance table, and confusion matrix were outputted to analyze results. Based on the results of the diagnostic, the input parameters were adjusted to try and increase the accuracy, sensitivity, and f1 scores. The validation data classification results are more important than the training in the forest-based regression, so the focus was to increase the values there. It was found that increasing the percentage of training data excluded for validation and increasing the number of runs for validation increased the score metrics for the validation classification.



Results:

Not surprisingly there were a lot of hotspots and coldspots as the presence of a wildfire in one place means there is quite a high likelihood of it spreading to another location. Our hot-spot analysis returned hot and cold spots throughout the state of California based on census tracts. The majority of California where the elevation was higher or had mountainous terrain yielded hotter spots. We found these results were similar to other analyses we did. Figure 8, showcases our hot spot analysis.

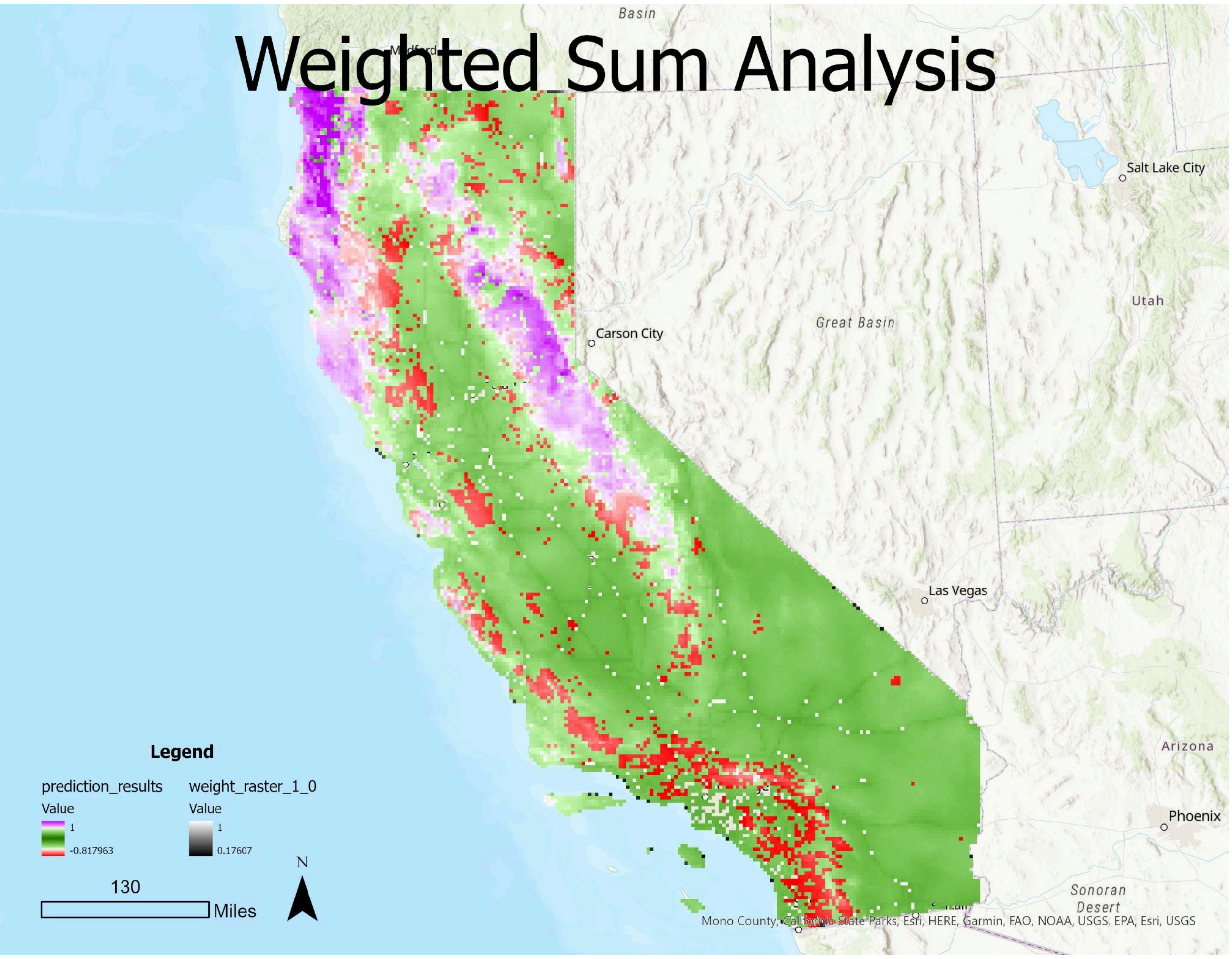
Figure 8: Hot Spot Analysis of California census tracts


Hot Spot Analysis per Census Tract
Medford
Salt Lake
Carson City
Great Basin
Utah
Las Vegas
San Francisco
Arizona
Phoenix
Sonoran Desert
Legend
fire_hotspot
Gi_Bin
Cold Spot with 99% Confidence
Cold Spot with 95% Confidence
Cold Spot with 90% Confidence
Not Significant
Hot Spot with 90% Confidence
Hot Spot with 95% Confidence
Hot Spot with 99% Confidence
160
Miles
N
California State Parks, Esri, HERE, Garmin, FAO, NOAA, USGS, EPA, Esri, USGS

Our Weighted Sum analysis was executed by utilizing the precipitation, distance to roads, and wind speed datasets as inputs. Weights are assigned based on the value field. Precipitation:

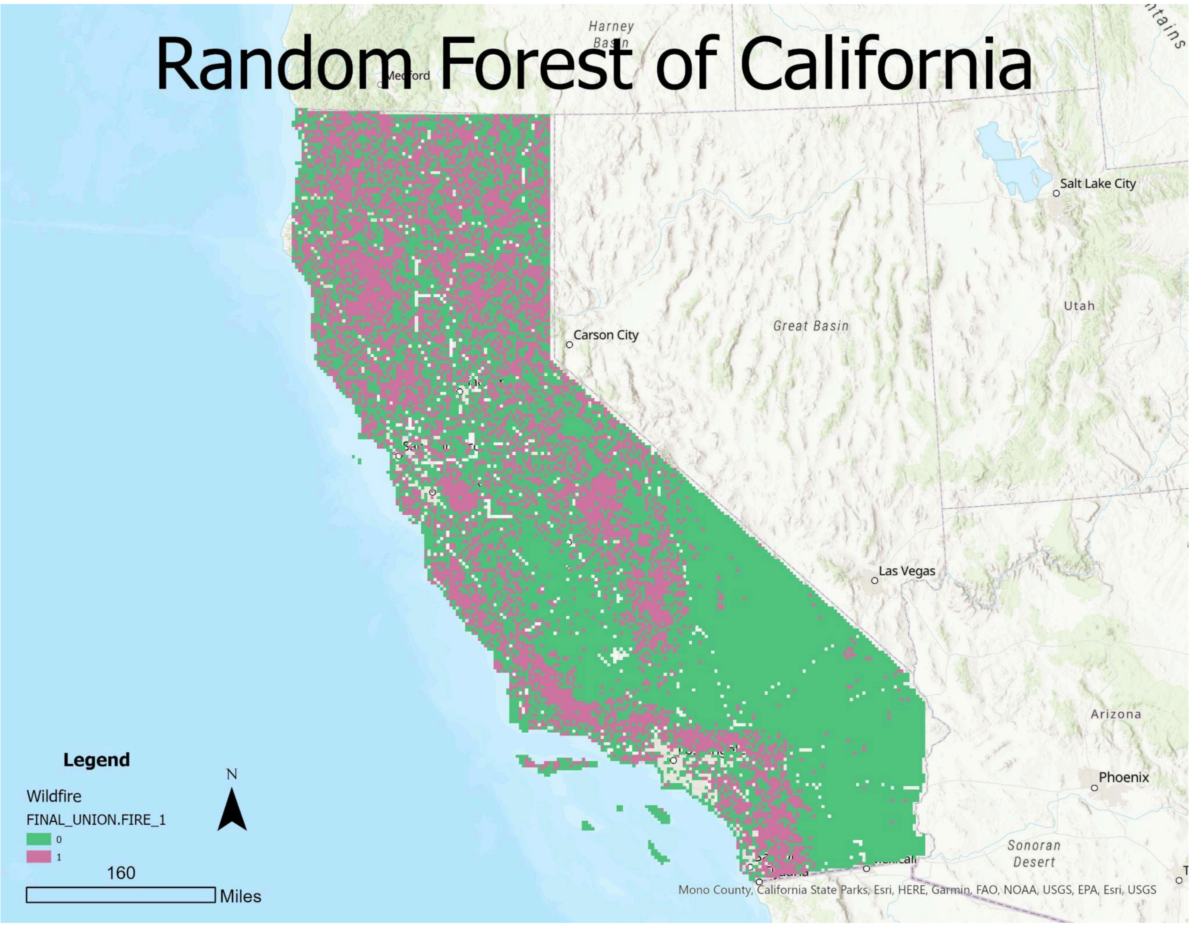
0.002487, roads: 0.25278, windspeed: 0.28658. The output was a raster layer that showcased the distance of wildfires to roads throughout California. The resulting model suggested prediction results, ranging from -0.81 to 1. The darker blue colors suggested areas where fires were more likely to occur. As seen on the map, these areas are concentrated near mountainous regions and areas with a high percentage of trees near the coasts. Additionally, we are able to suggest that areas of lower precipitation and higher wind speeds have a higher wildfire probability and perimeter spread. Figure 9 showcases the weighted sum analysis.

Figure 9: Predicting California wildfires using weighted sum analysis



For our random forest model, we added a new field to each layer called Fire, and gave a value of 1 on the fire perimeters and a value of 0 on the census tract. After Erasing the fire perimeters from the census tracts and unioning them, we got polygons of fire that have a fire value of 1 and census tracts - the perimeters with values of 0. This fire value became our variable to predict. This is important because it gives us more instances of areas where fires are not common. Conversely, this allows us to predict areas in which fires are common. Our model had 100 trees, a leaf size of 1, and a tree depth range of 1283-1449. The random forest model provided us with a top variable importance chart, suggesting that NDVI had the highest importance throughout the model (27%) followed by wind speed with an importance percentage of 25%. Our random forest model is pictured below.

Figure 10: Results of California wildfires from a random forest model.



Conclusion:

With a phenomenon that is as complex as wildfires, having a binary prediction of fire is quite limiting. Creating a continuous value from 0-1 to quantify the risk would provide more information and be more beneficial for organizations that look to mitigate risk or anyone interested. Additionally, there is not much of a temporal aspect with our data. We used averages for an entire year for our environmental data, so there are no seasonal or monthly patterns that are taken into account. Precipitation was an indicator of fire in this analysis, but if we go to a finer temporal resolution, the picture is quite different. Instead, areas with high precipitation tend to have less fires. This addresses our concern with the Mediterranean climate of California. With climate differing throughout terrain, we see that mountainous areas are typically at higher risk. Since our precipitation data is at such a coarse temporal scale, the areas with high precipitation are the areas that have plants or trees that are able to burn and the areas with low precipitation are desert areas (so no wildfires there). An improved analysis would take time of year into consideration.

Our hot spot analysis allowed us to narrow down our study area limitation to census tracts. With this analysis, we are considering the residential populations that are potentially affected by wildfires. Studying by census tract, we were able to provide a value of hot or cold spots to suggest zones of high wildfire susceptibility. We were able to deduce that the majority of Northern California was more susceptible to wildfires than Southern California.

Our weighted sum analysis provided us with insight into our hypothesis which suggested we take California's terrain into consideration. California has a drastically different landscape and the risk of wildfires changes as the elevation/terrain changes. Our model showed that areas of high elevation, specifically mountainous areas towards the east of California, and elevated areas near the northern coast, were significantly more prone to wildfires. Additionally, this model held weights to wind speed and precipitation, inferring that areas of lower precipitation and higher wind speeds lead to a higher probability of wildfires and a wider perimeter spread. This was an important step in our analysis to predict the perimeter and location of future wildfires. The analysis displayed areas of high fire probability, allowing us to make projections that wildfires are highly concentrated in Northern California.

Our random forest model primarily aided us in our prediction of higher-concentration wildfire zones throughout the entire state of California. With our previous analyses, we deduced that Northern California is more susceptible and has a higher probability of wildfire occurrences. We were able to predict that areas of higher elevation, more vegetation (suggested by the NDVI), lower precipitation, and higher wind speeds were more prone to wildfires. The random forest was the final tool in helping confirm our hypothesis and refine areas that will continue to be affected by wildfires. This model further emphasizes our discovery that wildfires are primarily located in Northern California and in areas of higher vegetation, lower precipitation, and higher wind speeds.

Project Story Map URL:

https://storymaps.arcgis.com/stories/fbd3e3e4d37c45f697d6bad2b62b29bc

References:

Gabriel, Edith, Thomas Opitz, and Florent Bonneu. "Detecting and Modeling Multi-Scale Space-Time Structures: The Case of Wildfire Occurrences." Journal de la Société Française de Statistique 158, no. 3 (2017): 86–105.

Lan, Yongcui, Jinliang Wang, Wenying Hu, Eldar Kurbanov, Janine Cole, Jinming Sha, Yuanmei Jiao, and Jingchun Zhou. "Spatial Pattern Prediction of Forest Wildfire Susceptibility in Central Yunnan Province, China Based on Multivariate Data." Natural hazards (Dordrecht) 116, no. 1 (2023): 565–586.

Radeloff, V.C, R.B Hammer, S.I Stewart, J.S Fried, S.S Holcomb, and J.F McKeefry. "Wildland-Urban Interface in the United States." Ecological applications 15, no. 3 (2005): 799–805.

Zhang, Yang, Samsung Lim, and Jason John Sharples. "Modelling Spatial Patterns of Wildfire Occurrence in South-Eastern Australia." Geomatics, natural hazards and risk 7, no. 6 (2016): 1800–1815.

California Fire. "Fire Statistics." *Cal FIRE*, www.fire.ca.gov/our-impact/statistics. Accessed 1 Dec. 2023.