Analysis of Wildfire Susceptibility and Prediction

California leads the nation in wildfires and acres burned. We assess important variables for occurrence and predictive risk analysis.

John Hemmer | Arno Abrahamian | SSCI 575 Fall 2023
November 30, 2023

Why Study Wildfires in California?

The state of California leads the nation and most parts of the world in wildfires, continuing to be one of the most dangerous natural hazards to West Coast communities. It has been estimated that over 80% of California wildfires are avoidable or easier to control. The first step in figuring this out is understanding what variables contribute to wildfires and where the wildfires are primarily located. This will give us an understanding of how these disasters form, what regions they affect, the length to which they spread, and what we can predictive mitigation efforts we can implement.

What areas of California are wildfires more concentrated and what variables affect the perimeter spread of wildfires? How can we use this information to predict future wildfires in California?

This is a map outlining our study area, the state of California.

The Data

What kind of data did we consider?

We used datasets that were reflective of common variables that affect wildfires. To begin, we used a California NDVI dataset. We used a dataset showcasing the perimeter of all California wildfires in 2022. We decided to use 2022 because it would provide the most recent data and was the last fully completed year. We then utilized a shapefile showcasing all the major roads in California. Finally, we considered two important factors in wildfires: the precipitation in the total region and the mean windspeed for 2022. Before analysis, all data was converted to the NAD 1983 California coordinate system. We also utilized a California census tracts dataset. In

ArcGIS Pro, many of these datasets were combined to create new data that would be utilized for future analysis including

Name	Attributes	Link
California NDVI	WGS 1984. Red, green and blue band data for California	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
Windspeed	WGS 1984. Windspeed mean for 2022	Link

Data table for the datasets utilized in the analysis.

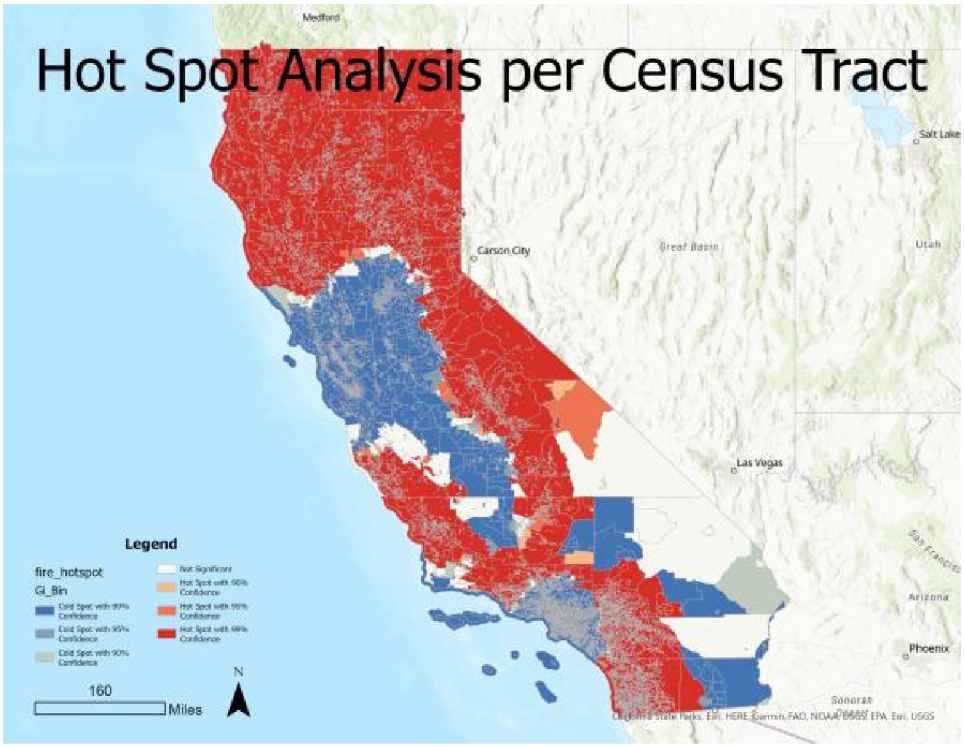
Finding a way to prepare our data in a way that could be put into a supervised learning model proved to be quite challenging. Our input training features was a background of census tracts with the fire perimeters unioned on top (vector). Before creating the union, we added a new field to each layer called 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 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. Since Random Forest also allows raster input layers as explanatory variables, 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. For other processes, the NDVI values had to be extracted using the extract bands property tool. This is due to the NDVI only

representing the red band when using raster to points (points were needed for some models).

Our Analysis

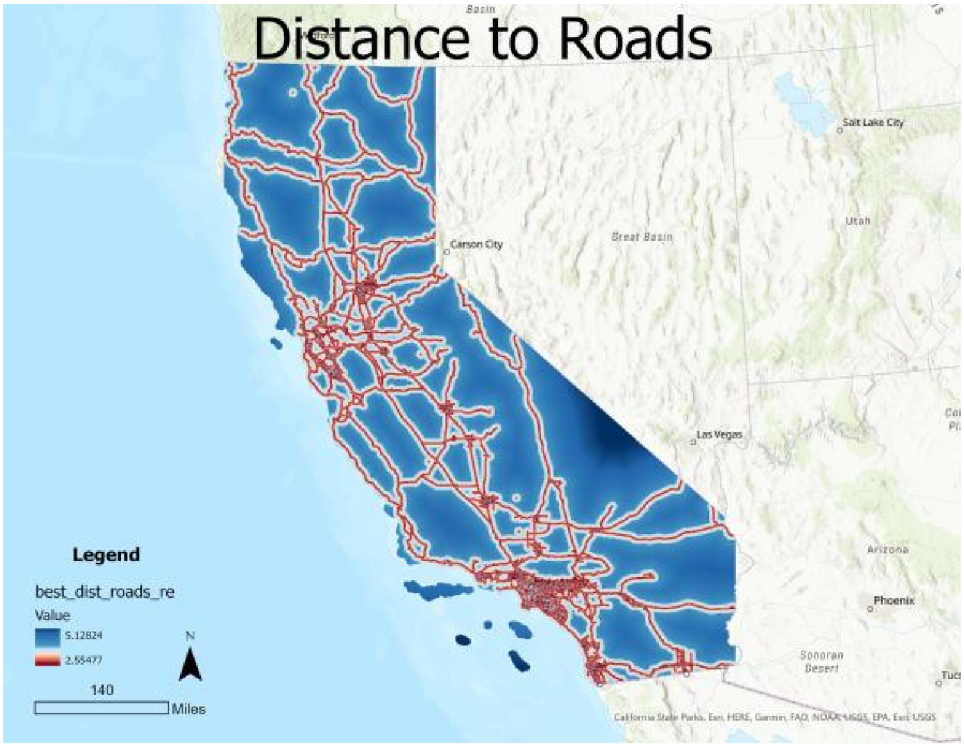
Hot Spot Analysis

We utilized a hot spot analysis to identify areas within the state of California that are prone to wildfires. Within ArcGIS Pro, we created a joint layer of census tracts and the 2022 wildfires. The hotspot analysis was to be done based on census tracts. We utilized a fixed distance band and Euclidean distance method and input the fire column.



Wildfire hot-spot analysis per California census tract.

We wanted to create a visualization of the distance of wildfires to roads. In addition to this, we wanted to directly compare it to the hot-spot analysis completed above.

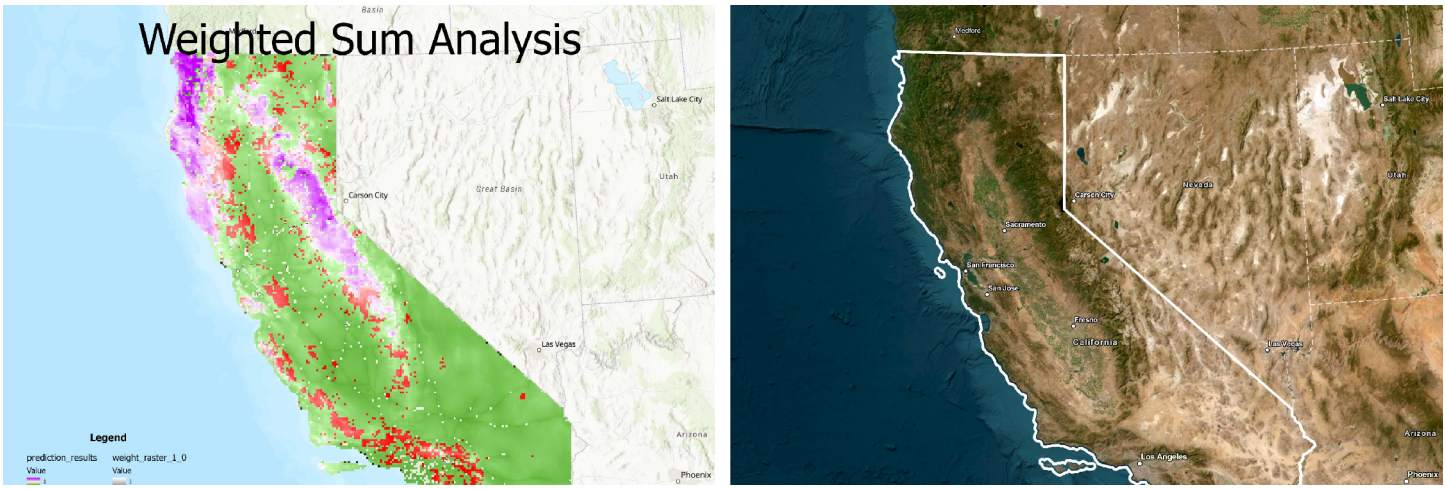


Distance of wildfires to roads in California.

Weighted Sum Analysis

We use a weighted sum analysis as a predictor. We utilize the precipitation, distance from roads, and windspeed datasets as inputs. Weights are assigned based on the value field. Precipitation: 0.002487, roads: 0.25278, windspeed: 0.28658. The dependent variable is a raster of the fire dataset. It is assigned a value of 1 for fire and 0 for no fire. The fires are larger in areas of higher elevation where windspeed can be assumed to be higher. The weighted sum an be an explanatory raster for the random forest model.

Other methods: The initial plan was to use a Geographically Weighted Regression (GWR) but it failed. There was a large amount of multicollinearity and not enough variation in the dependent variable. This issue seems to arise due to the binary nature of the dataset.



This slide can show the weighted sum analysis against the terrain.

Random Forest Model

Our first step in the analysis was to utilize a forest-based model. We started with the precipitation data, converting it to the NAD 1983 coordinate system. The California roads dataset was utilized to calculate the distance to the perimeter of the fire. Next, the California NDVI was utilized and entered all bands at once. We finally inserted the windspeed dataset.

We selected a random forest as our predictive model due to its ensemble learning, meaning it aggregates predictions of various models together. This tends to lead to more robust models that can capture non-linearity and complexity in the data while also reducing overfitting. We also have imbalanced data, which is covered well by random forest, there is even an option in the dialogue box for sparse data. Although we initially thought Geographic Weighted Regression would work well, it is not amazing for binary classification and our variables have some levels of multicollinearity that caused the model to fail.

Random forest model of our study area.

There was a total of 100 trees and a maximum tree depth of 1351. Given the analysis, we found the NDVI to be the most important variable, followed by windspeed then precipitation. The tree was split into two categories: 0 and 1. Respectively, the F1-score was 0.77 and 0.30. The sensitivity was 0.67 and 0.54. The accuracy was 0.65 for both.

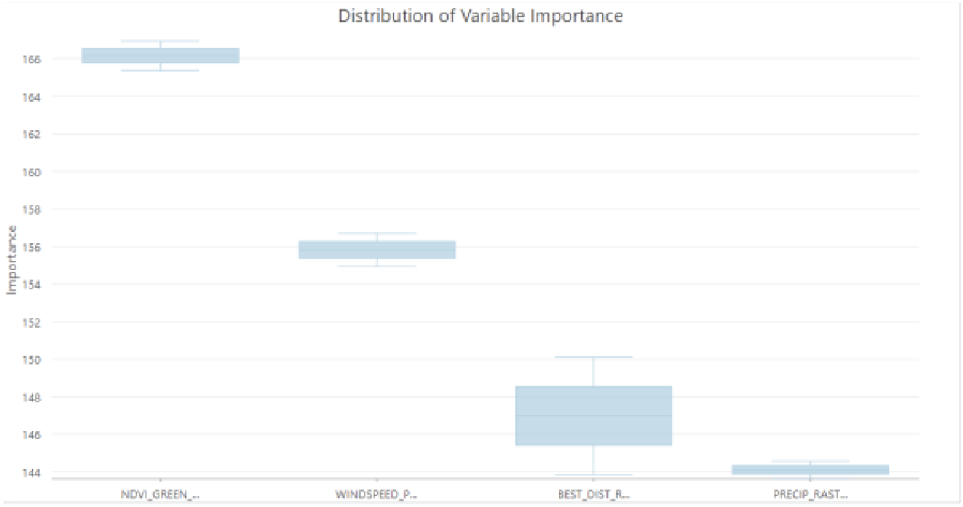
Other variables: Utilizing a land cover dataset was attempted by did not work; there were too many categories (500). The forest model failed to handle a large amount of categorial data and we couldn't group them because the numbers were arbitrary.

Parameters Environments Messages (11)

Prediction Type	PREDICT_RASTER
Input Training Features	final_union
Variable to Predict	final_union.Fire_1
Treat Variable as Categorical	CATEGORICAL
Explanatory Training Variables	
Explanatory Training Distance Features	
Explanatory Training Rasters	precip_raster_clip_project #;NDVI_green_Clip #;best_dist_roads_recalc3 #;windspeed_ProjectRaster_Clip #
Input Prediction Features	
Output Predicted Features	
Output Prediction Surface	G:\575\Final_project\Final_project.gdb\Forest_6
Match Explanatory Variables	
Match Distance Features	
Match Explanatory Rasters	precip_raster_clip_project precip_raster_clip_project;NDVI_green_Clip NDVI_green_Clip;best_dist_roads_recalc3 best_dist_roads_recalc3;windspeed_ProjectRaster_Clip windspeed_ProjectRaster_Clip
Output Trained Features	G:\575\Final_project\Final_project.gdb\Output_features6
Output Variable Importance Table	G:\575\Final_project\Final_project.gdb\output_variable_importance_table6
Convert Polygons to Raster Resolution for Training	TRUE
Number of Trees	100
Minimum Leaf Size	
Maximum Tree Depth	
Data Available per Tree (%)	100
Number of Randomly Sampled Variables	
Training Data Excluded for Validation (%)	20
Output Classification Performance Table (Confusion Matrix)	G:\575\Final_project\Final_project.gdb\confusion_matrix6
Output Validation Table	
Compensate for Sparse Categories	TRUE
Number of Runs for Validation	2
Calculate Uncertainty	FALSE

Parameters utilized for the random forest model.

Conclusion



Distribution part of variable importance

The state of California's most common natural disasters are wildfires. Given the state's dry climate, flammable vegetation, and recent changes in drought and climate change, we see a continuous increase in wildfires throughout the entire state. We did various probability and risk analysis tests, demonstrating that NVDI and windspeed represented the leading variables in wildfires in California. Additionally, we mapped trends of fires, utilizing random forests, hot-spot analyses, and weighted sum testing to do probability testing and predictive analysis. This helped us locate areas of higher

fire concentration throughout the state and allowed us to conclude future fires based on the varying trends of the analyzed data.

John concentrated primarily on the Machine Learning models, specifically the Weighted Sum Analysis, Random Forest model, and Distance to Roads. Arno worked on making layouts for these models and working on the Hot-Spot Analysis. Additonally, John concentrated on writing about the data, methodology, and analysis while Arno concentrated on explaining the study area, introducing the project and speaking on the results of the analysis. Both Arno and John contributed to writing the report, analysis, and developing the story maps.