Investigating the Relationship Between Drought Severity and Wildfires in California

Amita Anand

EAS 4480 Final Project

Abstract

The goal of this paper is to identify any statistically significant relationships between drought severity, average wind speed, and wildfires. We also would like to know the cycle length of droughts by various severity levels. We did not find a statistically significant relationship between the stated variables, however, there exists a positive weak correlation between average wind speed, wildfire occurrence, and acres burned by wildfire.

Introduction

California has become increasingly prone to wildfires in recent years. The wildfires in 2018, 2020, and 2021 were the most deadly and destructive to date. In 2018 alone there were over 100 fatalities, 24,226 structures damaged, and 1.9 million acres of land burned (1). Wildfires not only take lives and damage the environment but cause long lasting air and water quality decline. Further, the extreme temperatures of wildfires cause forest soils to lose essentially all nutrients as the organic material on top of the soil burns away (2). This then makes it difficult for vegetation to return to the area.

Wildfires do not have exclusively one cause, however, drought and high wind are contributing factors. The state of California is prone to both drought and windstorms making it the focus of this paper. As climate change gets worse, unpredictable severe weather events become more frequent making it imperative to understand the relationship between environmental factors and such events. The goal of this paper is to identify a statistically significant relationship between drought, wind, and wildfires. Further, using time series analysis, we would like to know if wildfires can be predicted based on the period of drought occurrence.

Data

1. Sources

The data used in this paper comes from NOAA's Storm Events Database and CalFire's Incident Reports. The Storm Events Database contains historical records for drought and high wind events dating back to 1970. In this study we use data in the period 2000-2020 for the state of California. The drought severity data is reported on a scale from D0 – D4 where D0 indicates abnormally dry and D4 indicates extreme drought. The data is organized into percentage of total area covered in each drought severity level. This data is reported weekly. The high wind event data gives dates of recorded wind speeds of 40 mph or greater, lasting 1 hour or longer or 58 mph and greater for any duration. Associated with each date is the highest wind speed recorded in mph. CalFire reports the total number of wildfires per year as well as the total acres burned by wildfires per year which will serve as our metrics for the analysis.

2. Processing

Since the drought severity is reported weekly and separated into various levels, it is necessary to aggregate by year. To do this an average percentage was calculated for each year by the drought severity type. Then, an average total percentage area covered in drought was calculated resulting in 21 data points. For the wind data, a similar method was used to aggregate the data by year. Again, the resulting data consists of 21 points, each the average wind speed of high wind events for that year.

Preliminary Analysis

1. Histograms

To understand the distributions of drought severity, the following histograms were plotted. We can clearly see that as the severity of drought increases the total percentage area tends to decrease, which is expected.

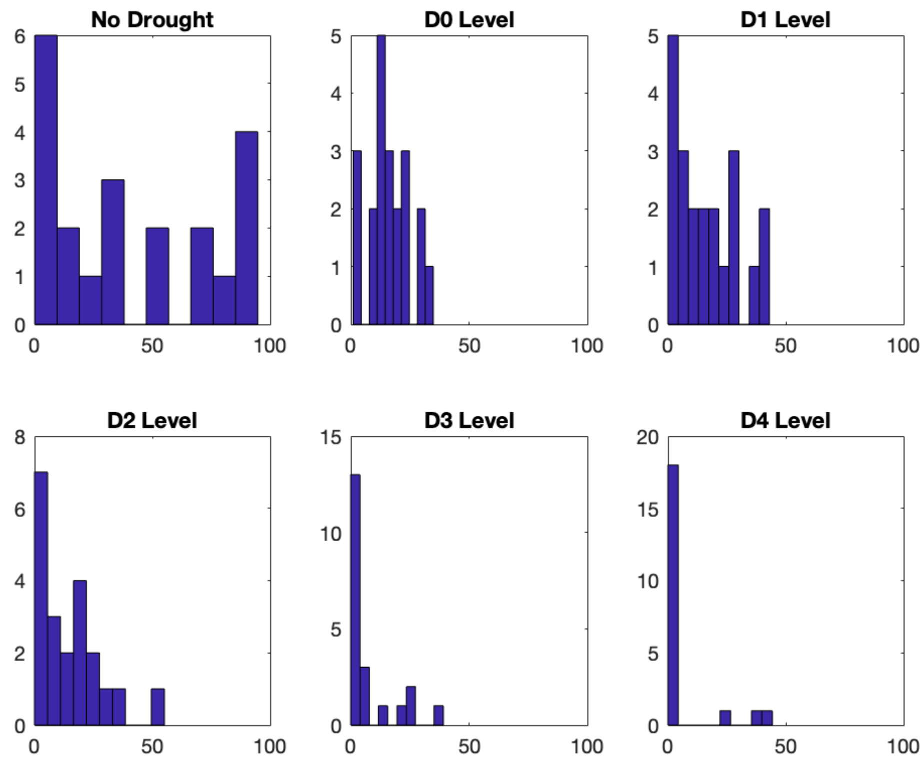

Figure 1 Histograms of Drought Severity

2. Box Plot

To further visualize the drought data, a box plot of the drought severity levels was computed. From this plot we see the mean percentage area of land covered in drought decreases as severity level increases which is consistent with the histograms. Further, we see the spread of the data decreases.

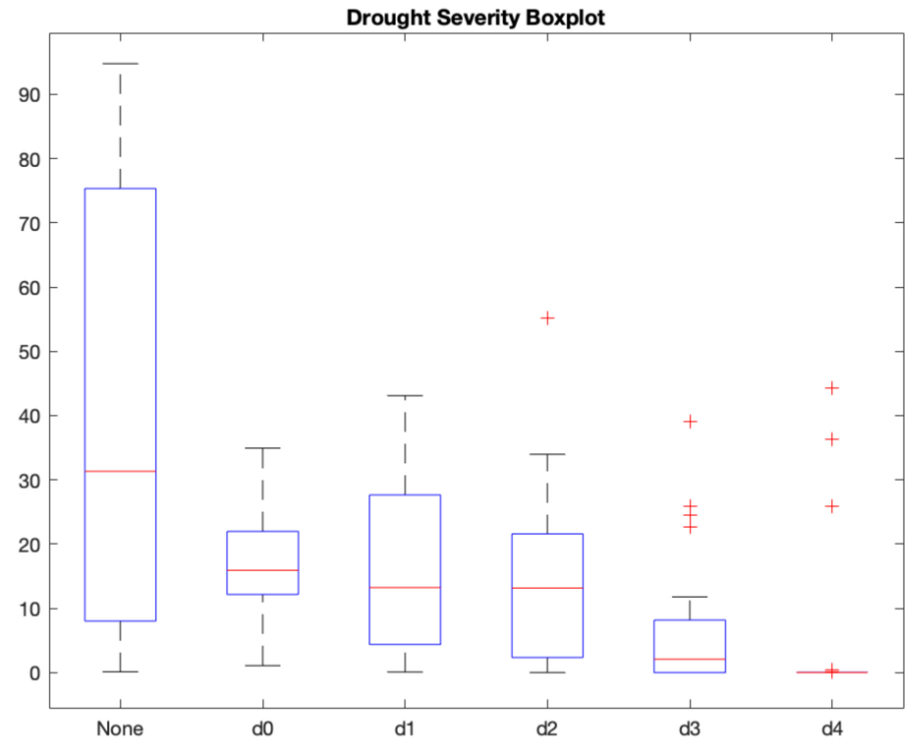

Figure 2 Box Plot of Drought Severity

Least Squares Regression

We want to identify how the response variable (wildfire occurrence and acres burned by wildfires) are affected by the input variables (drought severity and wind speed). This requires a least squares regression. Using the MATLAB polyfit and polyval functions a least squares regression line was calculated for drought severity vs wildfire occurrence, drought severity vs acres burned by wildfire, average wind speed vs wildfire occurrence, and average wind speed vs acres burned by wildfire. The results of the least squares regression are shown in fig 3-6.

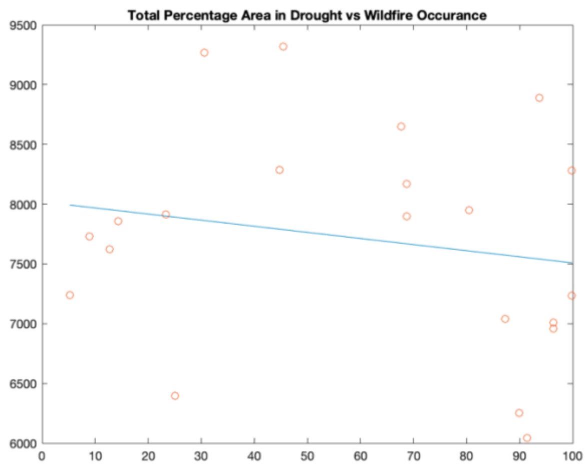

Figure 3

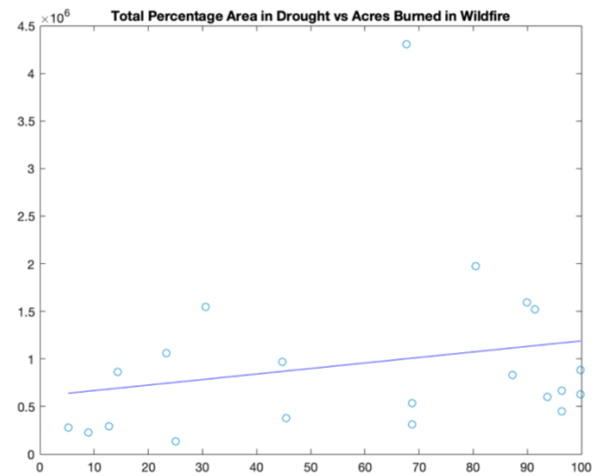

Figure 4

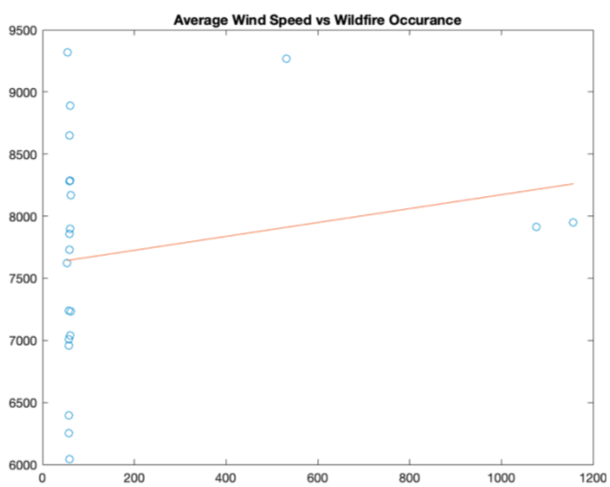

Figure 5

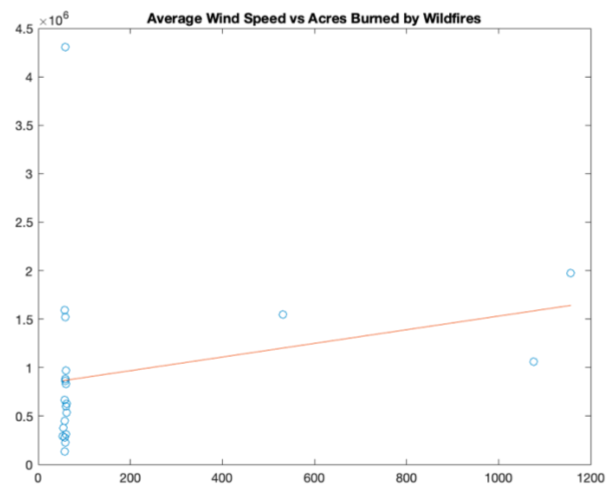

Figure 6

Correlation Analysis

To find statistically significant correlations, we use the corrcoef function which returns a correlation matrix and an associated p-value matrix. Each entry in Table 1 is the correlation coefficient between the variables in that column and row. Each entry in Table 2 is the probability of getting the corresponding correlation by sampling randomly. A p-value less than 0.05 is considered statistically significant. The correlation and p-value matrix are given in Table 1 and 2.

	Drought	Average Wind	Wildfire Occurrence	Acres Burned
Drought	1	-0.1196	-0.1926	0.2183
Average Wind	-0.1196	1	0.1993	0.2505
Wildfire Occurrence	-0.1926	0.1993	1	0.1773
Acres Burned	0.2183	0.2505	0.1773	1

Table 1 Correlation Matrix

	Drought	Average Wind	Wildfire Occurrence	Acres Burned
Drought	1	0.6057	0.4029	0.3417
Average Wind	0.6057	1	0.3869	0.2735
Wildfire Occurrence	0.4029	0.3863	1	0.4421
Acres Burned	0.3417	0.2735	0.4421	1

Table 2 p-Value Matrix

From the table we see none of the p-values are less than 0.05, but the two lowest values correspond to the correlation between average wind speed and wildfire occurrence and average wind speed and acres burned by wildfires.

Error Analysis

1. Bootstrap

To determine the validity of the least squares regression, we apply a bootstrap analysis to access regression reproducibility. Using the bootstrp function in MATLAB with sampling 1000 times. The histograms of the bootstrap analysis are seen in fig 7-10.

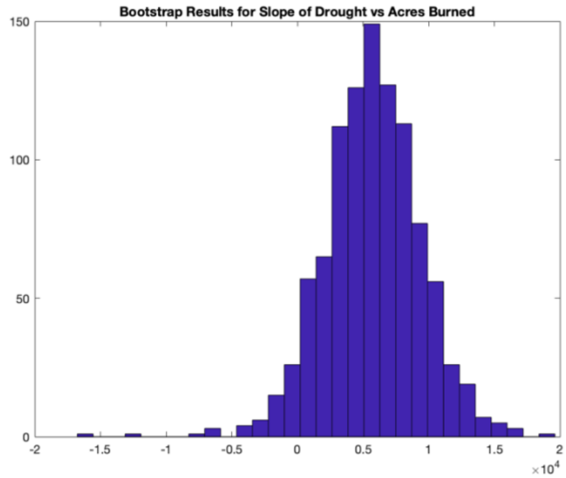

Figure 7

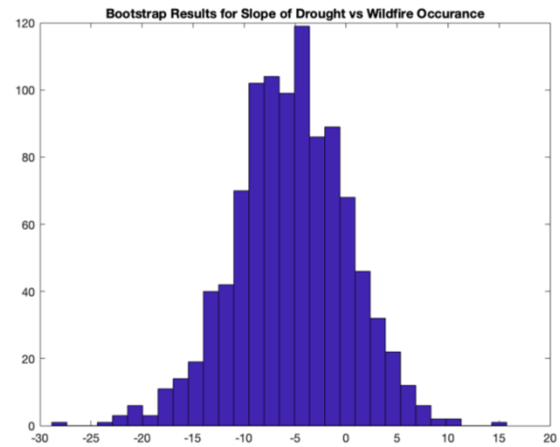

Figure 8

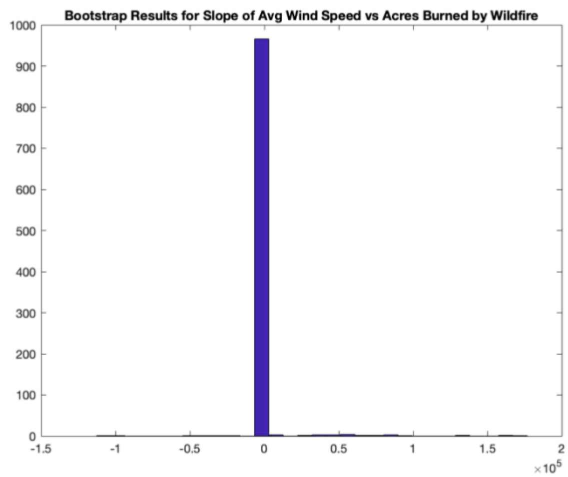

Figure 9

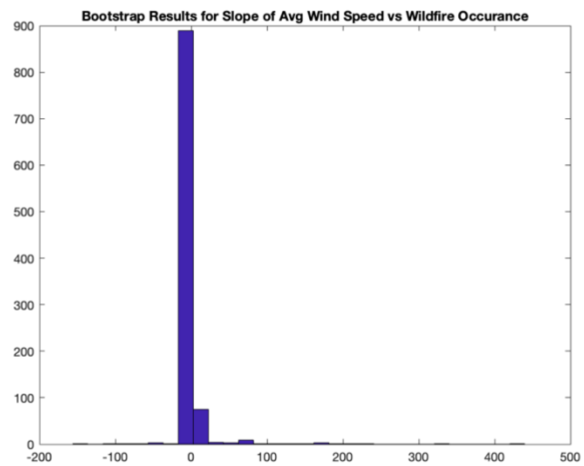

Figure 10

Using the results of the bootstrap analysis we can compute 95% confidence intervals for the slope of the regression. First we sort the output of the bootstrp function, then use the prcntile function to find the 2.5 and 97.5 percentiles. The results are summarized in the Table 2.

Drought vs Wildfire Occurrence	[-17.7652, 5.2814]
Drought vs Acres Burned	[-0.1279*10^4 , 1.2983*10^4]
Average Wind vs Wildfire Occurrence	[-0.0262, 34.4524]
Average Wind vs Acres Burned	[-0.0076*10^4, 1.711*19^4]

Table 3

2. Chi Squared Test for Normality

The chi squared test is used to test normality in the residual terms. If the residuals are normally distribution around zero, the least squares regression is valid. The stem plots of the residuals are shown in fig 11-14.

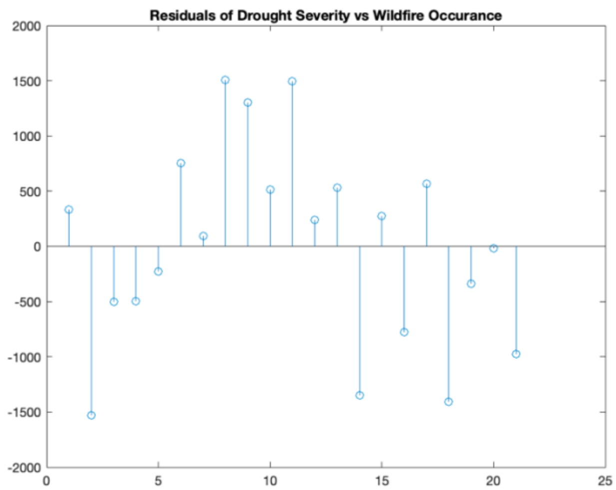

Figure 11

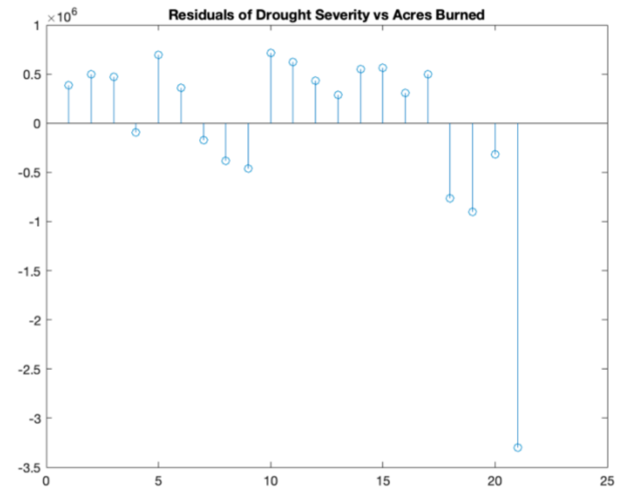

Figure 12

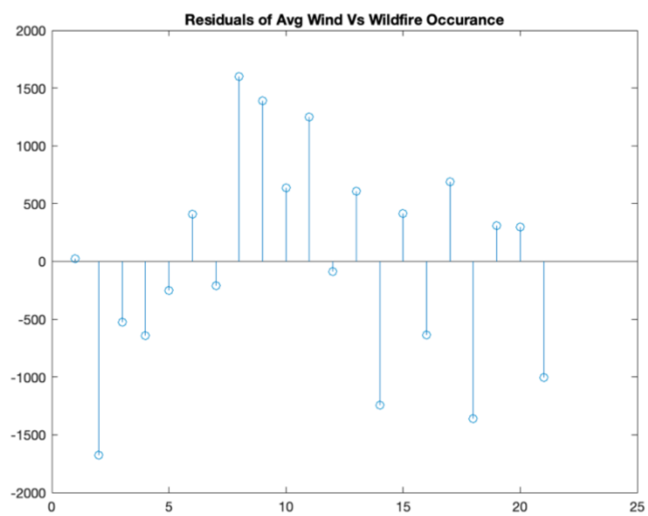

Figure 13

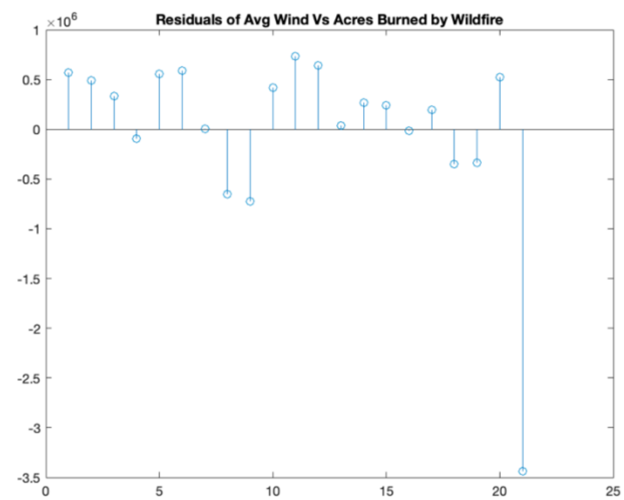

Figure 14

The null hypothesis of the test is that the residuals follow a normal distribution. The alternate hypothesis is that the residuals do not follow a normal distribution. The degree of freedom for the test is the number of bins minus the number of variables plus 1, so we get DOF =

2. Finally, for the test to have significant level equal to 0.05 we need to find the value of the inverse chi squared function with 2 degrees of freedom at 0.95. We reject the null hypothesis of the critical value is less than the chi2 value. The results of the test are summarized in the table below.

	Drought vs Wildfire Occurrence	Drought vs Acres Burned	Wind vs Wildfire Occurrence	Wind vs Acres Burned
Chi2	1.4496	22.1950	2.2984	40.5864
Critical Value	5.9915	5.9915	5.9915	5.9915
Result	Accept	Reject	Accept	Reject

Table 3 Chi Square Test Results

Time Series Analysis

Using the drought data taken weekly, we have 1092 data points. To find the period of drought occurrence we first detrend the data then use the periodogram function to calculate the power spectral density. From the graph we can identify the maximum period. The results of the periodogram and plot for each drought severity level are given side by side in fig 15-19.

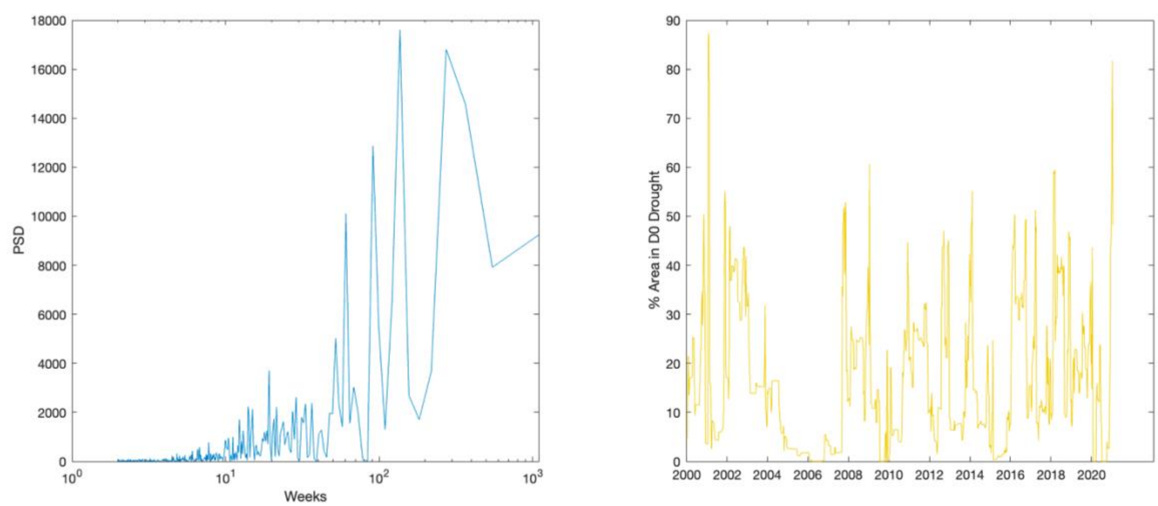

Figure 15

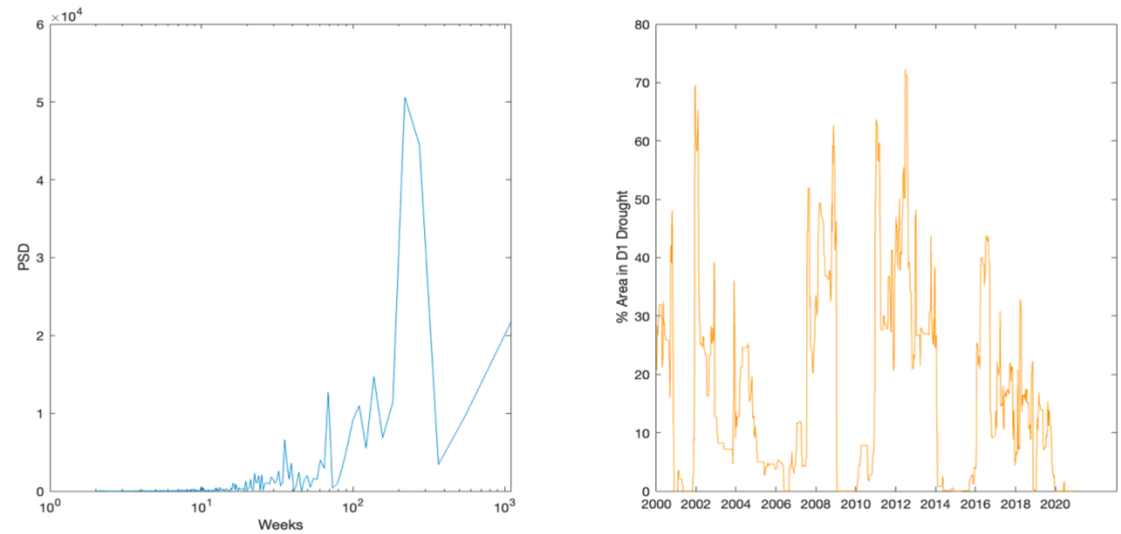

Figure 16

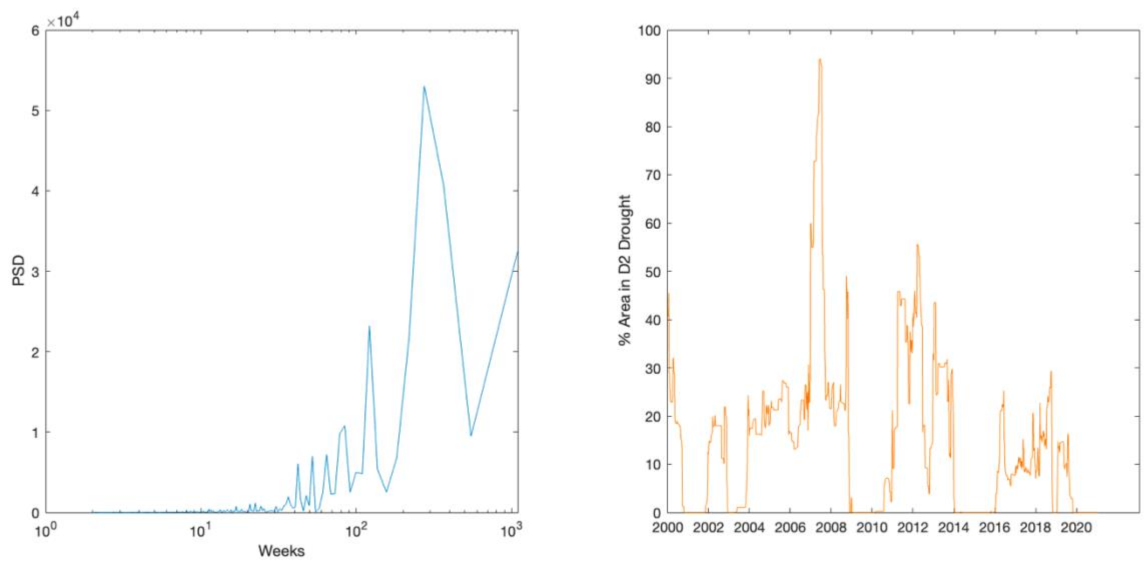

Figure 17

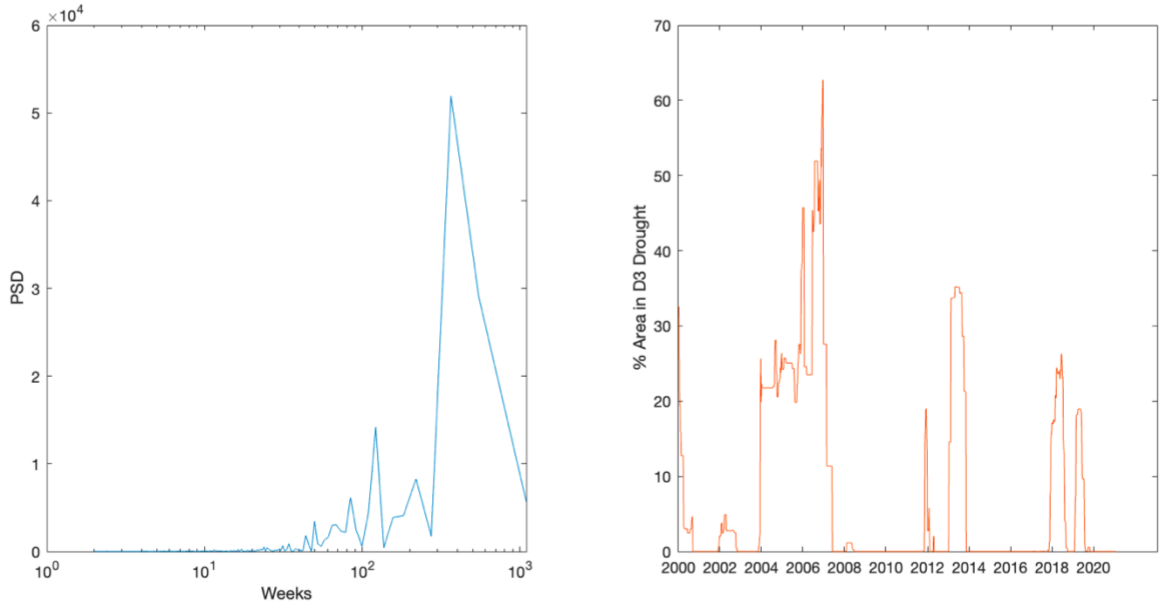

Figure 18

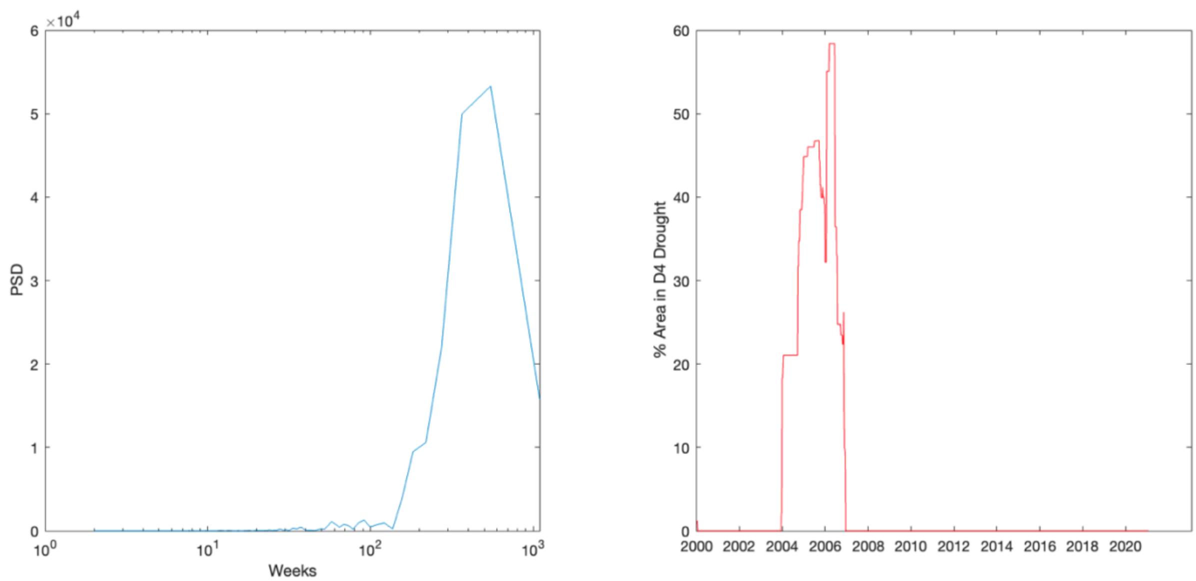

Figure 19

From the plots above we can identify the period, summarized in the Table 4.

D0	2.63 years
D1	4.21 years
D2	5.27 years
D3	7.02 years
D4	10.53 years

Table 4 Drought Severity Periodicity

Discussion

The least squares regression indicated a negative relationship between drought and wildfire occurrence which is surprising. This could be explained by many outliers in the data and how spread out the data is. Taking data from a longer time period than 20 years, say 40 years, could indicated a positive relationship. For the other regressions, there was an overall positive weak relation. Again, this is most likely caused by the outliers and lack of data.

The correlation analysis is consistent with the least squares regression, there is a weak positive relationship between drought and wind speed with wildfire occurrence and size. Since none of the p-values are less than 0.05, based on this data we cannot say there is a statistically significant relationship between the variables. At best, the lowest p-value is 0.1993 corresponding to average wind speed and wildfire occurrence which is still an order of magnitude greater than 0.05. The high p-values are likely caused by the spread of the data. Due to limitations of computing power and time, we had to aggregate the data over the month then over the year. With more time and memory, data collected weekly would likely lead to a stronger correlation as there is more information.

If the chi square test for normality fails, the least squares regression is not valid. Since the test failed for drought severity vs acres burned and average wind speed vs acres burned, we know these regressions are not valid. This is consistent with the correlation test since we saw neither of these variables have a statistically significant correlation. Similar to the least squares regression, if more data is used over a longer period, the chi square test would likely succeed.

The time series analysis gave interesting results. We see the time period per drought severity level decreases as the severity level increases. This is expected since an area in D4 drought would have had to have been in all the subsequent levels to get to D4. For D3 and D4 especially, we cannot confidently state the time period as 7 and 10 years respectively because of spectral leakage. In other words, the periods of these events are too long to be fully observed in 20 years. Again, we need data over a longer range to property state the time periods.

Conclusion

In this paper we compared drought severity and average wind speed to wildfire occurrence and acres burned by wildfires. We did not observe a statistically significant correlation, but this is likely due to limitations in data and computing power. We also saw the time period of droughts increase with severity and range from about 2 -10 years. For future work, average temperature and humidity could be looked at as variables predicting wildfires.

References

(1) https://www.fire.ca.gov/incidents/2018

(2) Ice, George G., Daniel G. Neary, and Paul W. Adams. "Effects of wildfire on soils and watershed processes." *Journal of Forestry* 102.6 (2004): 16-20.

(3) Littell, Jeremy S., et al. "A review of the relationships between drought and forest fire in the United States." *Global change biology* 22.7 (2016): 2353-2369.