CS-GY 6053, Foundation of Data Science

Project Report:
Quantifying the Impact of Wildfires on Air Quality in California

Team Members: Brunnell Velazquez (bv601), Alexander Gao (awg297)

"The 2018 wildfire season was the deadliest and most destructive wildfire season on record in California, with a total of over 7,500 fires burning an area of over 1,670,000 acres"

- *California Department of Forestry and Fire Protection*

Problem

Wildfires are natural disaster events that cause destruction, pose safety and health hazards, and are possible accelerating in rate due to climate change. We wish to quantify the impact of wildfires, particularly in the context of air quality, on both a short time-scale (days to weeks), as well as over longer periods of time (months to years). In this project, we examine air quality data (various pollutants / gas particles) before, during, and after the fire.

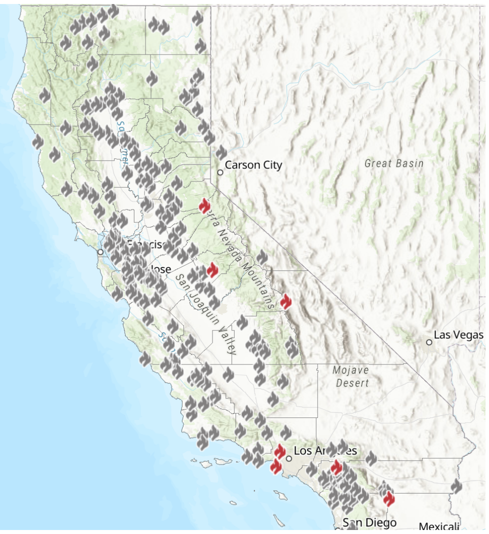
Some questions we asked:

- What specific changes do we see in the air quality as a direct result of a wildfire?
- Once the fire ends, does general air quality normalize / return back to what it was, or not? Also, what is the rate and timetable for dissipation of the polluted air?
- Over the long term, does an increased frequency of fires have a cumulative effect upon the air quality? (Is there a trend?)

Data Science Goal: Generate actionable insights for civil and government officials, and the general public, that will allow them to mitigate health risks of breathing heavily polluted air and be better equipped to navigate safely through these natural disaster events.

Background

Existing research enabled us to interpret our findings. We learned that wildfires emit gases such as nitrogen oxides (NO_x) and carbon monoxide (CO) and particles known as particulate matter that are by far the most hazardous to one's health if inhaled. Finer particles that are less than 2.5 micrometers in diameter are called pm 2.5 whereas those between 2.5 and 10 micrometer are called pm 10. Wildfires also have a secondary effect on the formation of ozone. **[1, 2]**



Map of California wildfires in 2019 (source: https://fire.ca.gov)

Data Sources

In order to perform this analysis, we aggregated data from two datasets: the OpenAQ (Open Air Quality) dataset and the California Department of Forestry and Fire Protection wildfire records.

Air Quality Dataset #1: https://registry.opendata.aws/openaq/
This dataset contains 11 features: the prominent ones and their data type are:
City
Country
UTC Time
Local Time
Value: Float
Unit (of Measurement): String
Latitude: Float
Longitude: Float

Record of California Wildfire Incidents by Year: https://www.fire.ca.gov/incidents/
These records contain information about all documented wildfires in California from the period 2013-2019. Relevant data includes: acres that burned, start date and time, end date and time, counties, latitude/longitude, duration, and structures destroyed. (Note: Not all samples contain all of the above features.)

Note: a file was not available on the website. Instead, we took the data that was displayed on the website (which came in JSON format) and transformed it into a CSV file.

Data Analysis

Features:
We are primarily used time series of particulate matter measurements during wildfire events in California as our features. We performed linear smoothing / moving average on the time series before performing other analyses / fitting our model. Measurements included pm2.5, pm10, co (carbon monoxide), so2 (sulfur dioxide), o3 (ozone), and bc (black carbon).

We used county/location as a soft feature, which mainly helped with grouping wildfires and noticing the ways in which pollutants dissipated seemed to be heavily region-dependent. Another soft feature used was Acres Burned, provided by the wildfires dataset. This was very important to visualize and cross reference against the Air Quality dataset, in order to confirm that spikes in pollutants were indeed related to occurrence of wildfires.

Target:
Our target is a prediction of the concentration of an air pollutant (pm2.5) in the air, *x* amount of time after the peak level has been reached.

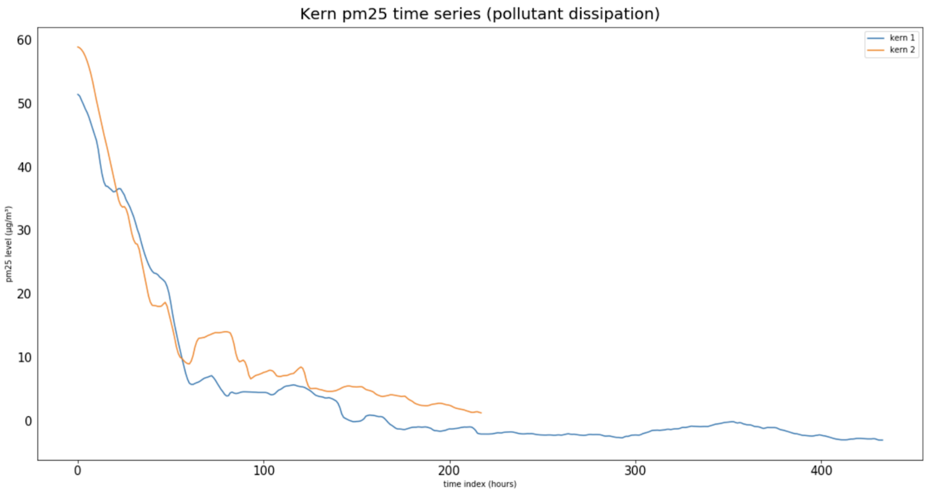
We approached analysis as a Time Series. Therefore, our X_train and X_test were simply time indexes from 0 to t. Prior to data exploration, we did not assume anything about trends, stationarity, or seasonality within the air quality data. However, after getting to know the data better, it became obvious that the wildfire events are correlated with sudden increases in levels of particulate matter. In contrast, ozone levels show seasonality and no spikes. Finally, to our surprise, we discovered that air pollution returns to a comparable level prior to a wildfire event.

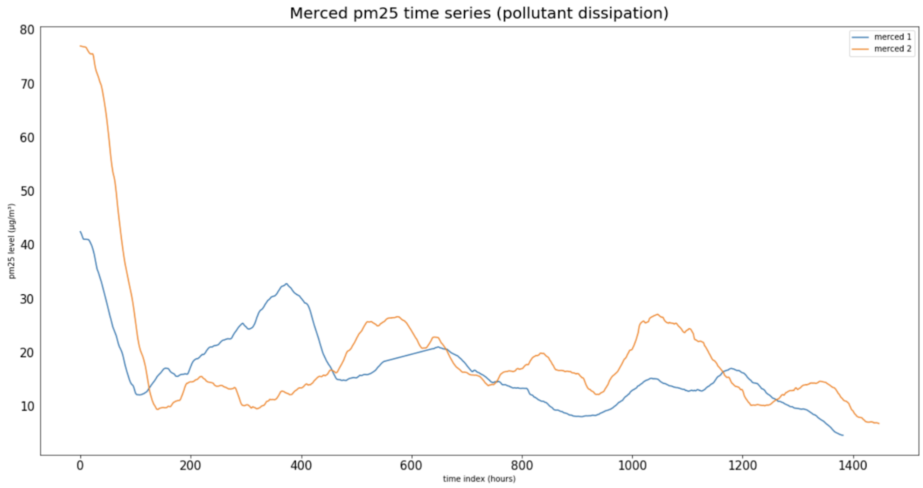
We began initial exploration of air quality time series by examining autocorrelation. This led to insights about which features would be most useful in building our model.

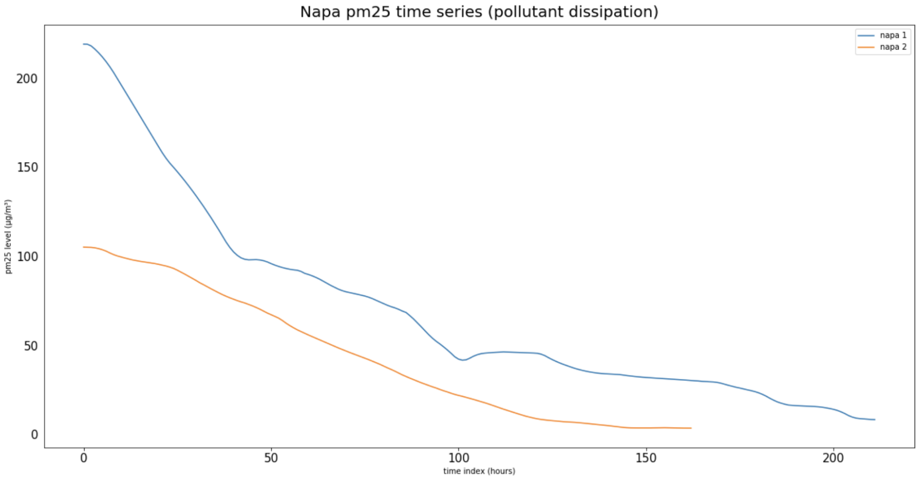
Model

We decided to build and compare Linear Regression and Exponential Regression models, and evaluate which would better fit out data, and help us to predict pollutant dissipation for future events. We used a single time series to fit our Linear and Exponential models, and evaluated the accuracy of the models on a test time series by observing the R^2 Score and the Residual Mean Square Error between test data and the fit model.

We initially assumed that the rate of dissipation of the air pollutants would all be similar—or at least those of the same type would all be similar. However, we discovered that the dissipations of pm2.5 are more similar if they occurred in the same county. The graphs below illustrate this point. We therefore, decided to focus on one specific county rather than trying to build a one-size-fits-all model to represent all counties in California.


Kern pm25 time series (pollutant dissipation)
kern 1
kern 2
pm25 level (μg/m³)
60
50
40
30
20
10
0
0
100
200
300
400
time index (hours)

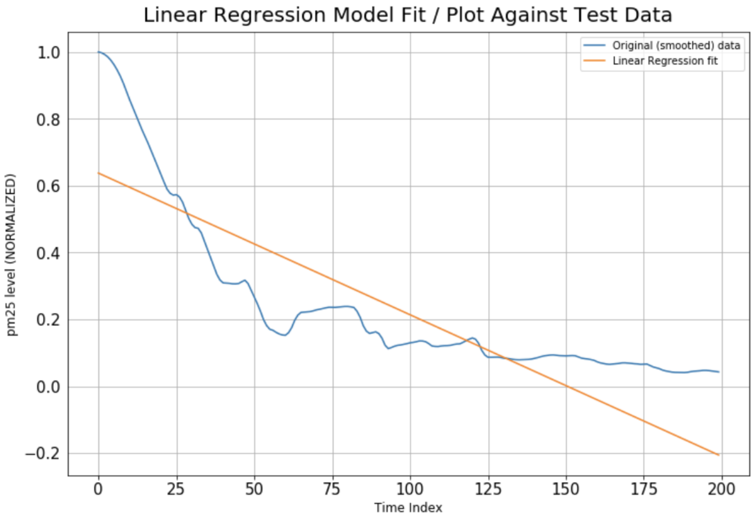

Merced pm25 time series (pollutant dissipation)
merced 1
merced 2
pm25 level (μg/m³)
80
70
60
50
40
30
20
10
0
200
400
600
800
1000
1200
1400
time index (hours)


Napa pm25 time series (pollutant dissipation)
napa 1
napa 2
pm25 level (μg/m³)
200
150
100
50
0
0
50
100
150
200
time index (hours)

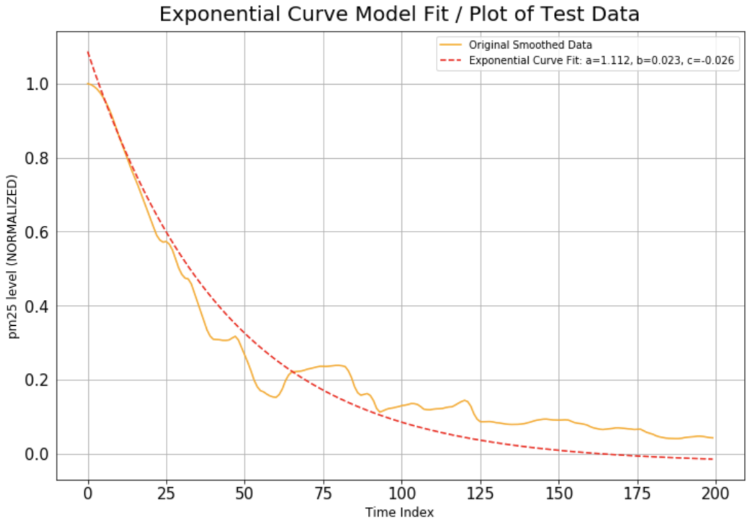
Typically, when working with time series, forecasting models are used. However, forecasting assumes an an interpretable autocorrelation in the time series such that the past measurement could predict future ones. Since the end utility of our target is to provide insight about overall level of pollutants remaining in the air, we found that a regression model suited our goals. It gave us the ability to predict a continuous value for any point in time, as opposed to a single endpoint value, and we used simple enough models that we could be confident that we would not overfit our data.

Some results (Exponential model fit our data better than Linear model):

```
R^2 of Linear Regression:  0.6377769561497821
RMSE of Linear Regression:  0.02191442139262559
```



```
R^2 of Curve:  0.9337964733634152
RMSE of Curve:  0.004005300063106849
```



Changes from the Proposal

Initially we believed that we would be able to create a model that would generalize to all possible wildfire events, but through the process of data exploration and modeling, we clearly learned that this approach would not be effective. We concluded that the most effective way to model the dissipation of air pollutants would be to build region and scale-specific models.

We decided not to pursue any geospatial analysis or research into how the air quality of neighboring regions are affected by a wildfire event, primarily because typically we've seen the air pollutant levels return to normal, which undermined our hypothesis that the wildfire events do affect the air quality of neighboring regions. We also realized there was much more in-depth analysis required simply to accomplish the modeling of pollutant dissipation.

Assumptions/Limitations

- Wildfire events are correlated with sudden increases in most air pollutants that OpenAQ dataset provides.

- While creating our time series that we would use for analysis and to build our model, we began the series at the peak level (local maximum) of air pollution during a wildfire event, and assumed this to signify that the fire had been contained.

- Dissipation of the air pollutants in a given county are more similar (and different across countries) probably because of varying differences in human resources to put out the fire and wind/weather conditions that would drive the pollutants away from the county

- Dissipation of air pollutants can indeed be modeled linearly or as a curve. We validated this hypothesis, however we believe that our data is by nature very limited, as wildfires are sparse, and not uniform, so we assume that a small amount of validation would hold for future samples.

- Not all the air pollution measurements are captured at a consistent time interval to then plot into a time series. We assume that filling in the missing values by using linear interpolation will not adverse effect on the quality of our results

- Due to the scale of the dataset, we faced a limitation in matching the wildfire events with the corresponding time series. Plotting them out helped but it still required human effort to match them.

References

- “Chemical Composition of Wildland Fire Emissions”, Shawn P. Urbanski, Wei Min Hao and Stephen Baker.
- information provided to the public by California’s government website.