Spread of Wildfire Pollutants in California

Akwasi Owusu-Akyaw, Richard Li, Courtney Moran

December 14, 2019

1 Introduction

Californian wildfires have become significantly more violent in recent years compared to past wildfires in the state's history. The 2018 Mendocino Complex Fire in Northern California, for example, became the "largest fire in state history with 459,123 acres burned"[1]. In addition, the Camp Fire, which occurred during the same year, was considered the "deadliest and most destructive fire on record in the state"[1]. Given this significant intensity in wildfire damage, we thought that it would be useful to create a tool that could advise the public during a major wildfire. For that reason, we propose a model which predicts the CO concentration of Californian counties near a major wildfire outbreak. Specifically, this model does the following:

1. Utilizes historical data for major wildfire events and air pollutant levels across California.
2. Predicts CO concentration levels in counties near the vicinity of the wildfire.
3. Provide guidance to California citizens on the expected duration and intensity of air pollution when a wildfire breaks out.
    - For example: We can expect a 70,000 acre wildfire in Sonoma county to cause 5-7 days of unhealthy air quality (AQI>100) in Santa Clara county, if the event occurs in October with eastward Diablo wind patterns.

2 Related Work

To place this problem in a larger context, researchers such as Wu, Winer, and Delfino have looked into predicting PM concentration levels in [2] at the zip code level during wildfires in Southern California. However, the limitation of their work was the recording of PM concentration every 3rd or 6th day for U.S. air quality measurement systems.

In addition to this, other works utilize satellite imagery in order to record PM concentration during the wildfire's time period. For instance, Falke [3] describes the use of satellite imagery to qualitatively point out fire events. From a quantitative perspective, Engel-Cox[4] and Husar[5] use data from the Moderate Resolution Imaging Spectroradiometer (MODIS) and the Geostationary Operational Environmental Satellite East (GOES). However, these tools have limitations such as incorrect classification of regular clouds as smoke from fires, or a low correlation between Aerosal Optical Depth (AOD) measurements and actual PM concentrations (about .13).

One work that focuses on using machine learning to predict PM 2.5 levels for a California wildfire is from Reid [6]. This work finds an optimal model for PM concentration using 10-fold cross validation of various algorithms. In conclusion, this model was able to achiever a GBM model with a CV-R^2 value of 0.803. We view our work as another approach to modeling wildfire pollution, but with a stronger application for public awareness.

3 Datasets and Features

We used the following sources for the train/test data for this project:

1. EPA Air Quality Data [7]
    - Contained daily CO readings for various US locations (in latitude and longitude)
2. News sites with information about major historical wildfires in California
    - Contained information such as latitude and longitude data for the approximate center of the fire as well as the size (in acres) and dates

From these sources, we obtained data for 24 fires to use in our modeling. There are a number of factors that affect wildfire propagation and how pollutants will spread in the air. We based our predictions on the following features that we thought were most essential:

1. Fire size (acres)
2. Direction relative to fire (degrees)
3. Distance from fire (meters)

Given that the air quality dataset and fire specific dataset originated from different sources, we had to match the time period of each fire with the air quality measurements of that same period. Once we did this, we were able to formulate qualitative assessments for our data, which are discussed in the next section.

4 Methodology

As mentioned in our previous section, in order to create our models, we obtained data from the US EPA website [7] on daily air quality. Specifically, we examined the daily concentration of CO in various locations in California:

- Target variable: Concentration of CO in parts per million:
    - for each day (2006-2018)
    - for each sensor location (latitude and longitude)

In addition, we performed research into major fires that occurred within the past couple of years [8], and for each fire gathered data about:

- location (latitude and longitude)
- duration of fire (start date and end date)
- size of fire (in acres)

With this information, we wanted to visualize how CO pollution tends to spread spatially from the wildfire origin. Thus, we created some simple intensity plots for CO concentration in surrounding areas during four wildfire events (Fig. 1). The CO levels shown indicate the maximum CO reading for that sensor over the entire duration of the fire.

By visual inspection, we can identify some trends. All the fires seem to impact adjacent counties more than distant counties. The magnitude of air quality impact seems to vary from fire to fire, presumably due to fire intensity. The Woolsey and Thomas fires seem to have impacted counties Southeast of the fire most significantly, while there appears to be less directional bias with the other two fires.

This provided guidance for our feature selection, as listed below. We provide some commentary on how we hypothesize each feature will affect CO readings:

- Distance from the fire center (measured in miles, based on latitude and longitude of sensor and fire)

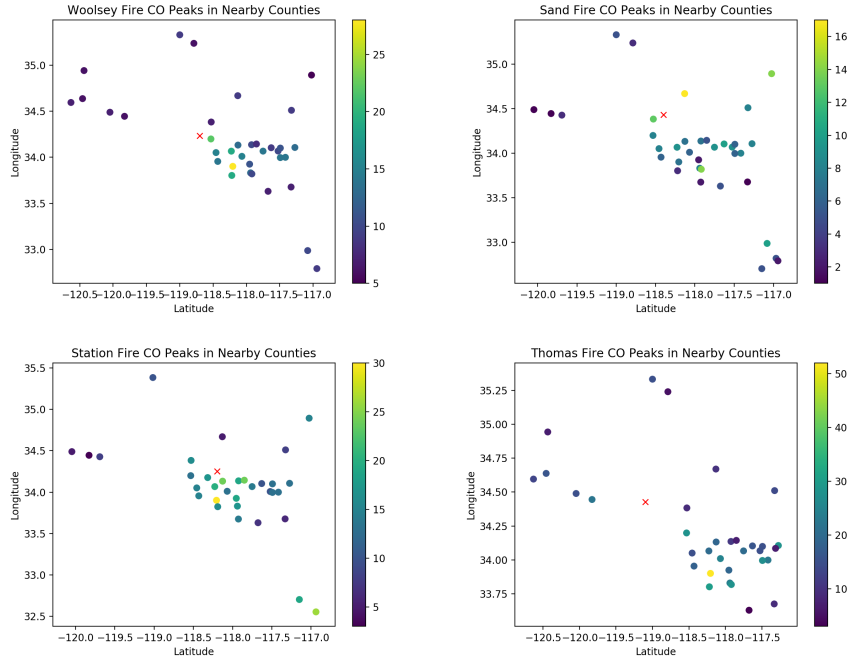

Figure 1: Max CO readings (AQI) in areas surrounding wildfires (marked by red x)

  - We expect a r^2 relationship between wildfire distance and air impacts based on physical properties of air dispersion.

- Direction from the fire center (measured in geometric degrees, based on latitude and longitude of sensor and fire)
  - We expect the directionality of air pollution to be related to both wind patterns and geography. We will not explicitly include wind patterns into our features, but our model should be able extract seasonal wind patterns from the time of year.

- Size of the fire (total acres burned)
  - We expect that larger fires will have proportionally higher air quality impacts.

- Time of year (early or late season)
  - We hope to extract information about seasonal wind patterns (i.e. coastal winds, Diablo winds) from the time of year.

As previously described, our target variable we want to predict is CO concentration (ppm) in each county, for every day during a wildfire's duration.

We constructed this model in piecewise, iterative fashion – starting with a simple model, evaluating its performance, then adding other features and complexity to improve the model.

4.1 Preliminary Model

Our preliminary model only incorporates two features: distance from fire and size of fire. We will ignore directionality for now, and also remove the temporal aspect by only considering the single maximum CO reading over the entire duration of the wildfire.

The base model was trained on data from 11 wildfires. A linear regression and a RANSAC regression were used. In Fig 5, the first two plots below show the linear regression model plotted with the data from 2 fires in the training set. The last two plots show the models plotted over test data (from a fire not in the training set). It seems that the RANSAC model performs slightly better because it is less sensitive to outliers.

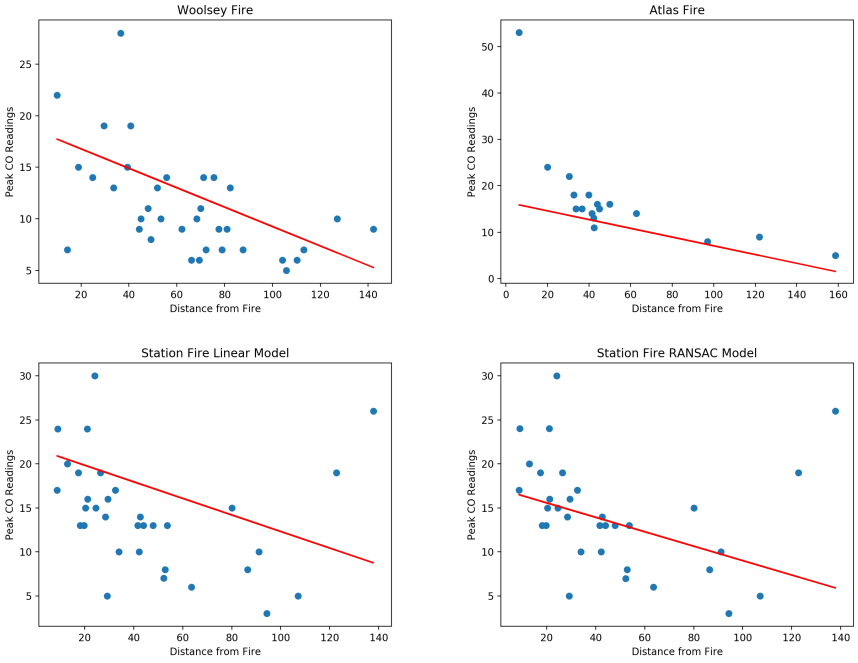

Figure 2: Top: linear regression plotted over training data. Bottom left: linear model plotted over testing data (MSE = 48.22). Bottom right: RANSAC model plotted over testing data (MSE = 38.14).

The mean squared errors are reasonable for this dataset, and likely dominated by the two outliers on the right. These two counties are likely downwind of the fire, and thus more greatly impacted than areas that are equidistant, but in different dirctions from the fire.

5 Current Model

5.1 New Feature: Direction

The next iteration of our model incorporated the direction of the fire relative to the sensor. This additional feature would presumably account for wind effects, which would not spread fire pollution equally in all directions.

Adding bearings as a feature improved the accuracy of the model in predicting all fires except for a handful (Fig. 4). As a reminder, the MSE for each row was calculated by training on all fires except for that fire, and testing on that fire.

The fires that did not benefit from bearings information had directionality trends that deviated from the other fires. This suggests the fires should be grouped into categories depending on their directionality trends.

Fire	With Bearings	Without Bearings
Woolsey	24.87	30.03
Roberanes	20.24	14.28
Erskine	8.23	11.46
Chimney	10.90	5.58
Sand	12.44	13.1
County	9.08	10.67
Rocky	8.64	9.53
King	22.41	10.47
La Brea	15.98	21.8
Basin Complex	16.27	22.14
Indians	19.99	23.12
Zaca	122.45	294.28
Lick	21.18	28.25
Day	49.07	55.96

Figure 3: Mean squared error (MSE) of predicted CO concentrations with and without bearings (direction) as a feature. Predictions in red were worse with directionality added.

5.2 Fire Categorization by Seasonality

California experiences characteristic wind patterns at different times of year. For most of the year, coastal air from the ocean (west) blows inland. From October-November, however, warm winds from the northeast (Diablo winds) can dominate. This has historically exacerbated many wildfires, as the wind can blow smoke from the forested regions where fires typically originate (in the east) towards the urban areas near the coast (west).

As a result, we started by grouping fires based on their time of year. Fires that finished burning before October 1 of any given year were designated season 1, while fires that started burning after October 1 were designated season 2. Fires that spanned both seasons were empirically placed in season 1 or 2 depending on which model better fit the data.

In Figure 5, predictions for each fires are made with and without seasonality groupings. In the first column, the model is only trained on fires that were initially classified as season 1. In the second column, the model is trained on all fires. As shown, several of the fires were better predicted by training on all fires rather than the subset in season 1. This suggests that either those fires are misclassified, or that the dataset is so small that having more examples is more beneficial than segmenting the data.

The fires in red were manually toggled between the two seasons to determine which grouping resulted in the best predictive accuracy for the other fires.

5.3 Fire Selection

To augment our data, we added fires dating back to 2006, increasing our dataset from 11 to 24 fires.

Fire	With Seasonality	Without Seasonality
Woolsey	24.87	19.07
Roberanes	20.24	16.15
Erskine	8.23	12.93
Chimney	10.90	5.65
Sand	12.44	60.26
County	9.08	8.87
Rocky	8.64	17.2
King	22.41	8.39
La Brea	15.98	37.29
Basin Complex	16.27	15.44
Indians	19.99	15.53
Zaca	122.45	61.88
Lick	21.18	24.39
Day	49.07	60.85

Figure 4: Mean squared error (MSE) of predicted CO concentrations with and without groupings based on seasonality. The predictions in red were significantly worse with seasonality groupings, suggesting they should be grouped differently.

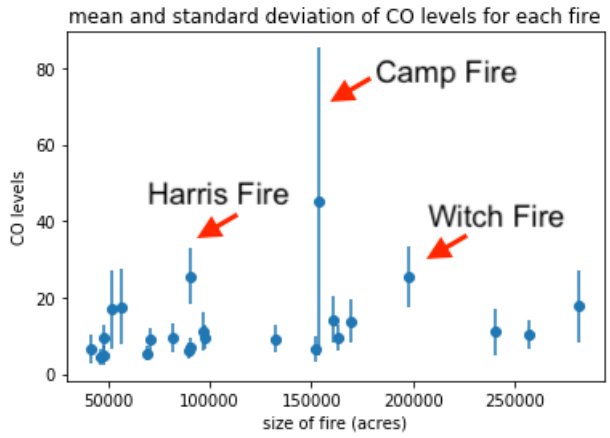

Figure 5: Mean and standard deviation of CO levels for each fire. Some fires had significantly more variation than the rest of the fires, which skewed the model.

As shown in Figure 6, there are several fires with large deviations in CO levels. These fires were skewing the model and impacting the accuracy of other fire predictions. As a result, the Camp Fire from 2018 was removed from the training dataset.

We found that MSE was significantly improved for most fires with the removal of the Harris and Witch fires (Figure 7), which also had relatively large variability.

5.4 Polynomial Features

We expanded our linear regression to include polynomial features, as we hypothesized that the CO intensity should fall off with distance with an R^2 relationship.

We observed that 3rd order polynomial features were better able to describe our training data than 2nd order features, without overfitting or sacrificing generalizability to our test set. Thus, we determined that 3rd order polynomials represented an acceptable balance between bias and variance.

Fire	Camp, Witch, Harris Removed	Camp Removed
Woolsey	24.87	22.04
Roberanes	20.24	36.79
Erskine	8.23	20.54
Chimney	10.90	3.91
Sand	12.44	66.75
County	9.08	18.91
Rocky	8.64	21.66
King	22.41	11.51
La Brea	15.98	56.90
Basin Complex	16.27	55.19
Indians	19.99	14.92
Zaca	122.45	75.05
Lick	21.18	26.79
Day	49.07	37.73

Figure 6: MSE was significantly improved for most of the fires after removing Camp, Harris, and Witch fires. Thus, we remove these three fires from the training set.

6 Results

The optimal categorization of each fire into early or late season is shown in Figure 8. Fires that spanned into October-November were initially sorted into late season, with all others early season. Two fires, Woolsey and Butte, were manually switched into the other groups as that significantly improved model accuracy for the other fires.

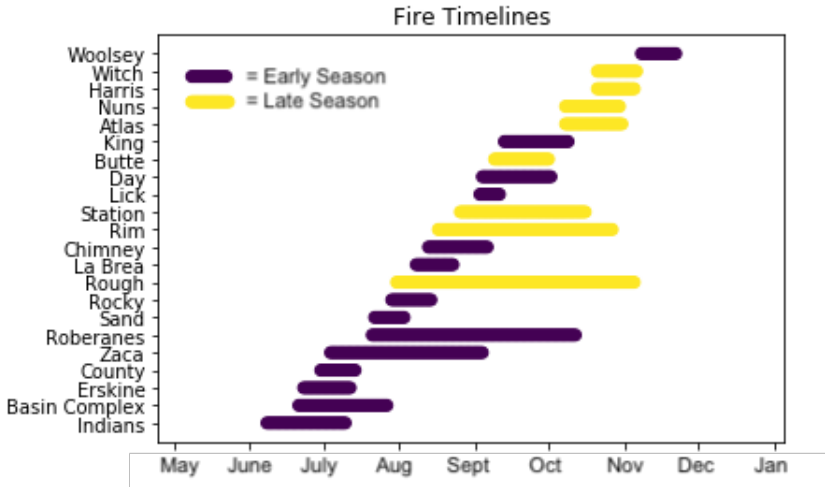

Figure 7: Predictive accuracy of model using each fire as test set. 65% of normalized mean square errors were below 0.3.

The directional dependence of early season vs late season fires is shown in Figure 9. As expected, maximum air quality impacts of early season fires are found east of the fire, due to coastal winds coming from the ocean. For late season fires, maximum air impacts are found southwest of the fire, which is consistent with Diablo winds coming from the northeast.

We are pleased to find that the directional dependence of CO intensity as determined by the model is consistent with our understanding of seasonal wind patterns in California. One issue we need to address in the future is ensuring that the output values correctly wrap around from -180 degrees to +180 degrees (instead of the discontinuity).

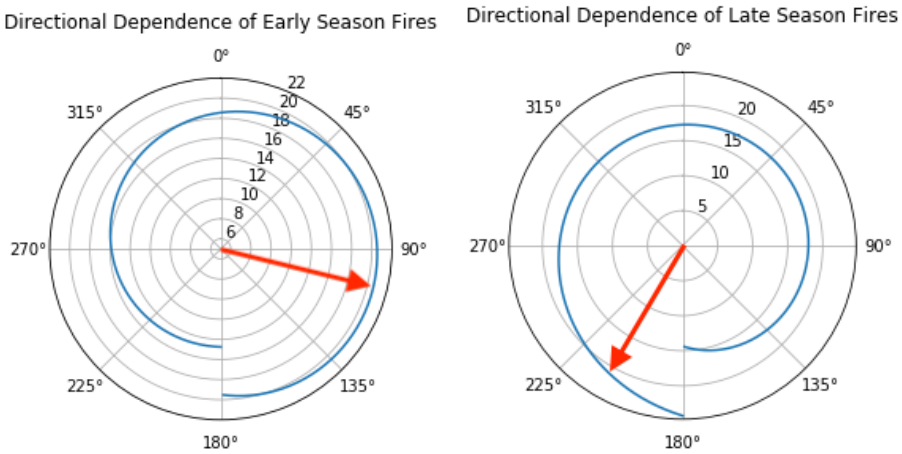

Figure 8: Directional dependence of early season (left) and late season (right) fires. As expected, early season fires tend to spread to the east (due to coastal winds), while late season fires tend to spread to the southwest (due to Diablo winds).

The predictive accuracy of the model is shown in Figure 10. In each case, the model is tested on one fire and trained on all other fires. The MSE was normalized by the square of the average value Figure 10. 65% of the NMSEs were below 0.3. The fires with high NMSE were disproportionately late season fires, likely due to the small sample size. Presumably, this error would decrease if the model were trained on more late season fires.

It is also possible that the fires with high error had atypical wind patterns for that time of year, or were located in areas with unique topology (i.e. mountain ridges) that obscured or facilitated the spread of pollutants. A more refined model could layer a topological map on top of the current model to capture these effects.

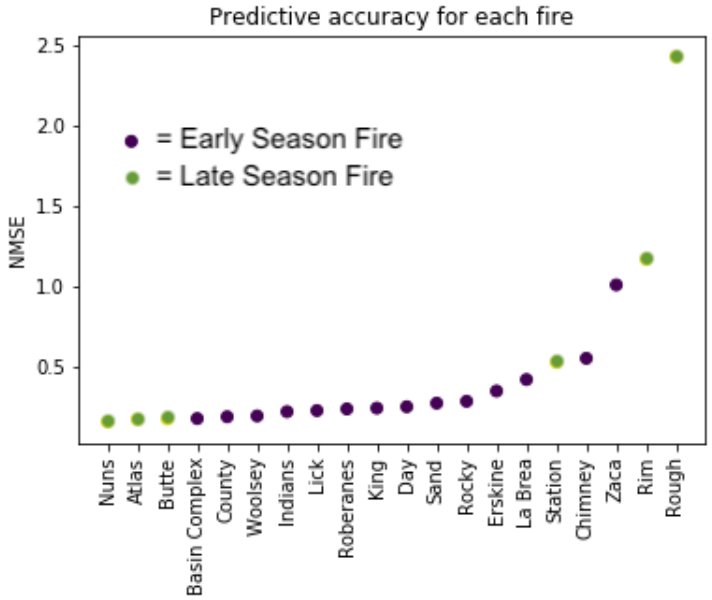

Figure 9: Predictive accuracy of model using each fire as test set. 65% of normalized mean square errors were below 0.3.

The dependence of CO levels on distance and direction is shown for one example fire in Figure 11. As shown, the CO intensity falls off with distance, which is modeled with an R^2-like relationship. Further, note that the purple dots, which represents sensors located northwest (-50 degrees) of the fire, are relatively low at a given distance. This directional dependence is captured by the

model, where the purple points are also disproportionately low.

At the bottom of Figure 11, CO intensity appears highest around 100-150 degrees (east–southeast), which is consistent with coastal winds during the early season. This trend is captured in the model. Again, notice that the sensors closest to the fires (blue points) are disproportionately higher than the sensors far from the fires (yellow points).

This demonstrates that the model is simultaneously accounting for both distance and direction in predicting CO intensity at a given sensor.

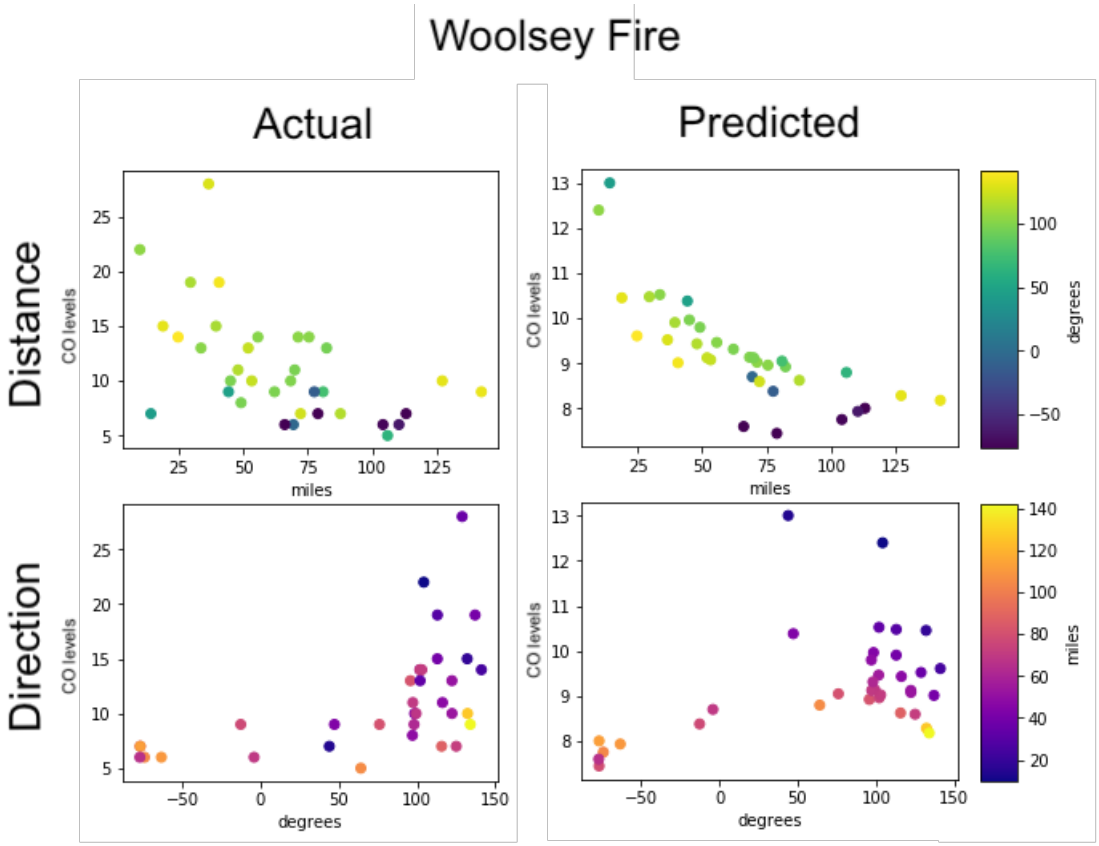

Figure 10: Dependence of CO intensity on distance (top) and direction (bottom) for Woolsey Fire. Actual data on left, model predictions on right. CO intensity falls of with distance with R^2 dependence, and is highest at 100-150 degrees (east-southeast), consistent with coastal winds.

The final feature used in the model is fire size, which is shown in Figure 12. Each point represents the average of all nearby CO readings for one fire. As expected, the average CO readings increase with larger fire sizes. This trend is captured by the model.

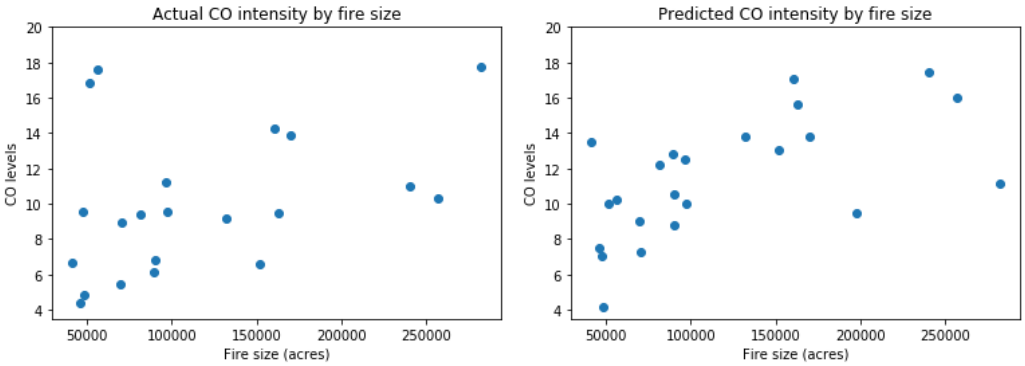

Figure 11: Dependence of CO intensity on fire size, actual (left) and predicted (right). Each point represents average of all nearby CO readings for one fire.

7 Conclusions

After exploring and modelling the data, we are able to conclude that CO concentrations in areas surrounding wildfires are (as expected) dependent on size of the fire, distance from the fire, and relative direction to the fire. The first two were intuitive, while the direction feature was a bit less obvious, but is clear to see in the initial plots of the data (Figure 1). After more research, the dependence on direction to the fire was attributed primarily to seasonal wild patterns. This pattern was reflected in the model. While the model represented most fires well, there were a few that did not. We hypothesize that these inconsistencies can be attributed to irregular wind patterns, geological/topological features, and/or insufficient data points.
As to the last point, some counties had little to no sensor data. Generally these were the counties with small populations. Because of this, we selected fires that were mostly near more urban areas, but with some counties missing data, the model might not have enough information to capture the effect of the wind patterns.

8 Future Work

Some future improvements to this model would be:

- Resolve the bearings wrapping bug
- Implement time-series model
- Gather more data for less populated counties
- Include wind data in the model

An issue with the direction feature in the model is that it exhibited a discontinuity at $\pm 180^o$. In future iterations this should be resolved.
A time-series model would give more precise representation of how the wildfires affect air quality in surrounding areas since fires can burn for several weeks or even months. Our current model is able to predict well which areas are most affected, and what the peak values in CO emissions will be. However, it does not predict when these peaks will occur (other than sometime in the duration of the fire). It would be most useful to predict which days will have the worst air quality.
To implement the time-series model, we would need information on how the fires spread since the size is not constant over its duration. This data is not readily available online, but perhaps the spread could be estimated with a model.
Another way to improve the model would be adding in data points for counties that were missing in our dataset. Many of the larger fires in recent years have burned in northern California in less populated counties. Having sensor readings from these counties would allow us to add more wildfires to train the model on, and improve the accuracy of models that were already tested on.
Finally, the model could be improved by incorporating data on wind speed and direction. While the data for most fires exhibits behavior consistent with seasonal winds, a few do not. This makes

sense since wind does not always strictly follow seasonal patterns, and days with higher wind speeds could increase CO levels in areas further away.

9 Contributions: Needs edits

Courtney: Researched California wildfire data. Wrote scripts to plot the wildfire data as a map of max CO readings. Wrote scripts to run linear regression models and plot the results. Wrote functions that were used to organize the training and test data so that we could plot higher order regression models more quickly.

Akwasi: Refined and scoped new project idea. Gathered wildfire data. Explored using decision tree regression as a potential model to predict air quality index (parallel to using higher order models). Organized the initial framework of the final report and the poster.

Richard: Generated and scoped original project idea. Visualized air quality evolution over time due to wildfires. Explored using higher order regression models and plotting the results for these models. Analyzed the trends and findings from our model.

10 References

[1] "Facts Statistics: Wildfires," Facts Statistics: Wildfires, 2019. [Online]. Available: https://www.iii.org/fact-statistic/facts-statistics-wildfires. [Accessed: 10-Dec-2019].

[2] Wu, J., Winer, A. M, Delfino, R. J. (2006). Exposure assessment of particulate matter air pollution before, during, and after the 2003 Southern California wildfires. Atmospheric Environment, 40(18), 3333-3348. Retrieved from https://escholarship.org/uc/item/8s10r6ct

[3] Falke, S.R., Husar, R.B., Schichtel, B.A., 2001. Fusion of SeaWiFS and TOMS satellite data with surface observations and topographic data during extreme aerosol events. Journal of the Air and Waste Management Association 51, 1579–1585.

[4] Engel-Cox, J., Holloman, C., Coutant, B., Hoff, R., 2004. Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality, Atmospheric Environment 38, 2495-2509.

[5] Husar, R.B., Tratt, D.M., Schichtel, B.A., Falke, S.R., Li, F., Jaffe, D., et al., 2001. The Asian dust events of April 1998. Journal of Geophysical Research-Atmospheres 106 (D16), 18317–18330.

[6] Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning Colleen E. Reid, Michael Jerrett, Maya L. Petersen, Gabriele G. Pfister, Philip E. Morefield, Ira B. Tager, Sean M. Raffuse, and John R. Balmes Environmental Science Technology 2015 49 (6), 3887-3896 DOI: 10.1021/es505846r

[7] EPA, "AirData website File Download page," EPA, 13-2019. [Online].
Available: https://aqs.epa.gov/aqsweb/airdata/download$_f iles.html.[Accessed: 10-Dec-2019]$.

[8] "2018 California wildfires," Wikipedia, 01-Dec-2019. [Online].
Available: https://en.wikipedia.org/wiki/2018$_C alifornia_w ildfires.[Accessed: 10-Dec-2019]$.