Final Report

Fire Watch – Wildfire & Air Quality Analysis

Team 17 : Wendell Hom, Po Hsien Hsu, Yeong Jer Tseng, Samuel Stentz, Srikesh Srinivas, Lauren Vossler

Introduction

For our project, we looked into US wildfires and air quality databases to visualize major US hotspots and to forecast air quality. We find this project interesting and valuable as recent wildfires in the US have created large issues for our country. Financially, annual losses from these fires can range up to 285 billion dollars. Health wise, the smoke from the fires can worsen heart and lung diseases. This is why we believe visualizing the spread of wildfire and smoke in the air could beneficially impact all of society.

Problem Definition

Our project consists of three parts: visualization, forecasting, and correlation analysis. For visualization, we overlayed the data onto a map of the US and displayed the sizes of ongoing fire by region and daily air quality levels. A dropdown for time range and a toggle for fire data and air quality data is also present. For the forecasting portion, we used air quality and wildfire data to predict the air quality of the next week. Lastly for our correlation analysis, we visualized air quality and wildfires in several aspects to determine if a correlation exists.

The bulk of our analysis is daily prediction of air pollution levels in the US. The most advanced current systems in use consider current and past air pollution levels, meteorological data, fuel source locations, controlled burn positions, and wildfire locations to forecast future air pollution levels [1][2][3][4]. Virtually all models used to predict air quality are some form of Monte Carlo simulation combined with a physical assimilated geological observations to predict both pm and ozone concentration [4]. Cetegen et al. (1982) studies the geometry of fires to improve models [5]. Dios et al. (2011) use geometric measurements, image processing, geo-location data, but encountered measurement errors [18]. O'Neill et al. (2008) describe incorporating fuel consumption models for wildfire analysis [9]. All of these approaches are limited not by their data sources but by their analysis methods.

Survey

We used machine learning models to forecast air quality based on data from previous US wildfires and air quality data. We applied regularized linear regressions, support vector regressors, and random forest regressors on previous air quality readings and fire locations to try to predict air quality for 7 days[11]. We believe this is an appropriate tactic as physics simulations for air quality already have shown predictive power and rely on underlying assumptions of fire patterns. Our models make no attempt to directly model physical phenomena and instead directly map the supplied input data to a value for pm, which we expect to improve accuracy. While a few previous works have attempted to predict air quality using machine learning algorithms, previous models only incorporated air quality readings into their algorithms [6][11]. Because fires are known to effect air quality [7] and wildfires are only expected to grow in frequency and intensity due to global warming [8], we expect providing the location of fires will give our model an advantage over previous models.

This project focuses on data from the US, so people living in the US are our target users. This includes people who are interested in studying patterns in the history of wildfires and air quality, people with are at risk [12][13][16], and government and health professionals who may want to provide early warnings to people living in areas which may become affected by bad air quality[13][18]. Wain et al. (2009) shows how dispersion model forecasts are used to aid land

developers according to smoke impacts [10]. The UCOP report discusses the effects on business planning and mental health effects on individuals.

For the interactive visualization, success was measured by user feedback on their experience with the application. For the predictive portion, we measured success by measuring the accuracy of our model in forecasting air quality for a given week and then comparing it with the actual values from our air quality database.[14]

Dataset Description:

- Air Quality Dataset <Source: United States Environmental Protection Agency (EPA)>[15]:

  Annual AQI data in each country is available on EPA. The annual AQI data from 1992 to 2015 was collected to match wildfire dataset. Important features including AQI, Category (Good, Moderate, Unhealthy), Defining_Parameter (Ozone, PM 2.5, PM 10, etc), State_Name, Country_Name.

- Wildfire Dataset <Source: National Wildfire Coordinating Group (NWCG) & Kaggle>[15]:

  It contains the wildfires that occurred in the US from 1992 to 2015. It also contains over 1.8 million data entries. Within the dataset, Important features including fire_dates, state, county, fire_cause, fire_size, fire_duration, fire_code, etc.

Proposed Methods

- Intuition:

Our proposed methods for visualization are valuable despite what currently exists as they incorporate air quality and wildfire data in an easy-to-understand format. Our methods for analysis improve upon current tactics for air quality prediction as we incorporate wildfire data along with air quality data.

- Description of Approaches:
  1. Determine correlation between yearly air quality and wildfire incidents in the US

  ➔ We are utilizing the air quality dataset from United States Environmental Protection Agency (EPA) & the organized wildfire dataset originated from National Wildfire Coordinating Group (NWCG) to see the correlations. The metric for air quality is AQI, and the metric used for wildfire is wildfire occurrence in a county. We will display if AQI is positively correlated to wildfire occurrence in several time spans (day, month, year) by heatmap color scale and additional plots besides choropleth.

  2. Utilize D3 to customize data visualization for more accurate message delivery

  ➔ To visualize our analysis of the dataset, we are implementing choropleth of the United States with D3. The US map will be displaying color scaling through each county of each state. The dropdowns and scroll bar alongside the map would be used to progress through year and days. We will use this to visualize the AQI datasets with the wildfire datasets. The map will provide a comprehensive view of the correlation between the two datasets. The choropleth will be our main visualization alongside with view charts to illustrate our analysis. The charts will be in a story telling manner to explain the reasons and steps of our analysis.

  3. Predict air quality with wildfire occurrences

  ➔ PCA/K-Means clustering of wildfire features, as well as yearly and monthly AQI to fire size correlation scatterplots were created to visualize the relationship between wildfires and AQI. The predictive models were trained with the following features: fire occurrence, size, previous AQI values, month, and state. Our objective is to predict if the future week's AQI > 100 for each county each day using these features. We used logistic regression with

equal weighting for both classes in sklearn. Data was scaled to std. dev. 1 for each feature. Lastly, we used 5-fold cross validation to evaluate the regression.

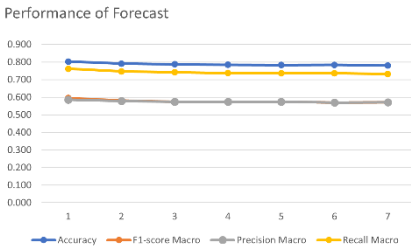
Package/Tools

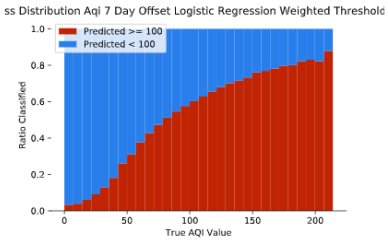
- Data Preprocess: Jupyter notebook, SQLlite, pandas
- Analysis: Jupyter notebook, numpy, seaborn, sklearn, matplotlib, pandas
- Visualization: D3, Bootstrap, Ajax, JQuery

Experiments/Evaluation

1. 7 Day Forecasting Analysis:

➔ Forecast Classification:
For every day and county reporting AQI (7,243,237 unique entries), a classifier to predict whether AQI x days later would be above a threshold of 100 was made for x in [1,2,3,4,5,6,7]. Weighted logistic regression was used, with previous 7 days of AQI, fire size and count in the county and state 0, 1, and 2 weeks prior, and a one hot encoding of State and Month used as features. Results shown are for a 5-fold cross validation on the data, and predictions for counties w/out a recording x day out was done by training on the whole dataset. Missing values for previous 7 days of AQI was filled using a backfill. Below we can see that as the AQI prediction is for further out, the fraction of counties correctly classified worsens but the forecast remains meaningful. This is reflected in summary statistics for classification (Accuracy: D1 0.803, D2 0.791, D3 0.787, D4 0.784, D5 0.783, D6 0.784, D7 0.780). Importantly, the logistic regressor increases in accuracy as the True AQI value furthers from our threshold (100), indicating for extreme AQI days the classifier is extremely reliable.



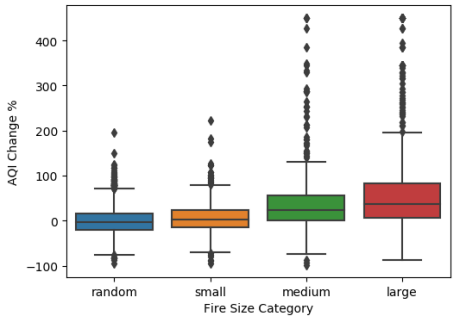


2. Wildfire & AQI Correlation:

➔ Influence of Large Wildfire on 14-day AQI average:
From our intuition and previous wildfire news coverage, we expected large wildfires to have a major impact on air quality. To test this hypothesis, we measured the AQI change over a 14-day window for wildfires starting from the time the fire was discovered. Here, AQI change is the percent change compared to the AQI average over the year for that county.

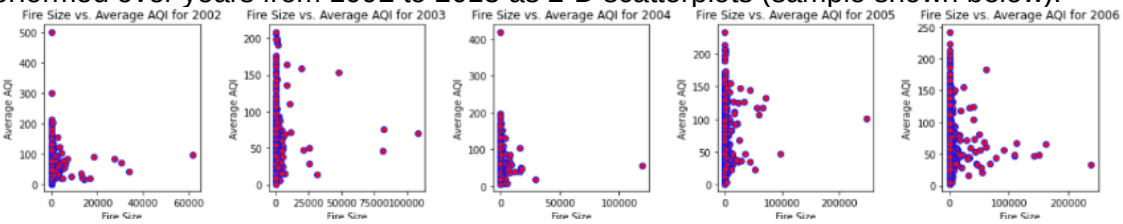
Our findings show that small fires (< 5,000 acres burned) had a minor impact on the AQI with an average increase of 7.1%. Medium-sized fires (between 5,000 and 10,000 acres burned) had an average AQI increase of 42.8% while large fires (> 10,000 acres burned) had an average AQI increase of 65.0%. For random days, the average AQI change over a 14-day window had an average increase of -0.6% which is very close to 0 and is what we expected. While this trend clearly shows that larger wildfires have more impact on the AQI, we noticed that there is also a larger variance in the AQI change for larger wildfires as seen in the boxplot below. This implies that there may be other factors that

can heavily affect the AQI besides the presence of wildfires (e.g., wind, rain, fog, size of county, distance from fire to sensor, etc.)



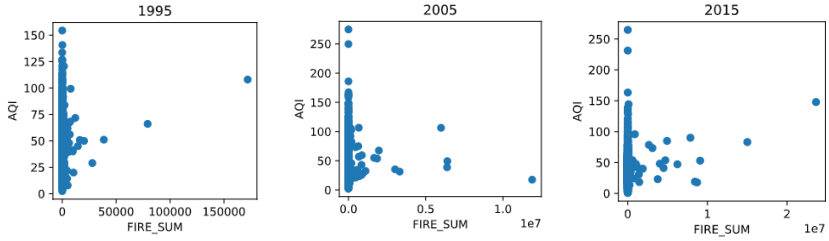
➔ Fire Size vs. Average AQI (Yearly Correlation)

Analysis plots were generated to show to relationship between the size of each individual fire size (FIRE_SIZE) and the average AQI (AVG_AQI) measure over the entire duration of that fire in that county. Insignificant fire sizes (i.e. <30) were discarded, and yearly plots for roughly 3000 fires were generated. This does reveal a slightly visible positive correlation (observe the “left-to-right" motion of the data points), but suggests that more valid comparisons are formed with more features (see K-Means section), most likely because the effect of a single attribute on another is not sizeable enough. This was performed over years from 1992 to 2015 as 2-D scatterplots (sample shown below).



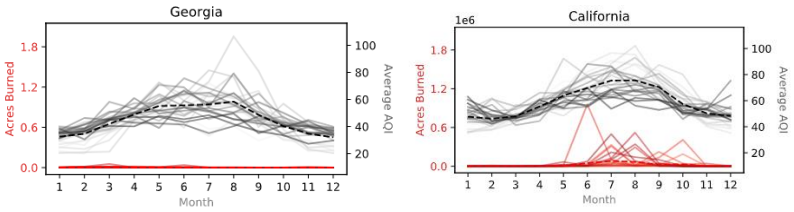
➔ Monthly AQI Average vs Monthly Wildfire Cases:

Considering AQI isn’t changing simultaneously with the occurrence of wildfires, we measure the monthly average of AQI with respect to wildfire sum. Wildfire sum (FIRE_SUM) equals to fire size * the duration of the fires for better addressing the delay effect of wildfire on AQI. From the parity plots, it is shown that most wildfires have no significant effect on AQI. And there should be other main factors increasing AQI such as gasoline combustion and factory gas emission.



➔ Line Plots:

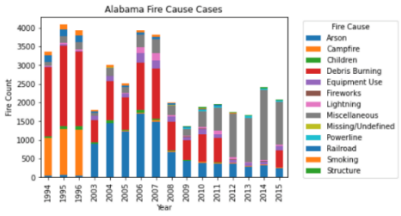
Within each state each month from 1998-2015, the total size of fires which were discovered in that month is shown, as well as the average AQI reported in the state. Each individual year is shown as a separate line, with an average for each month for both metrics shown as the dotted trend lines. These line chearts were based on fire size sum per day, average AQI, and acres burned.



3. Wildfire Cases & Causes (acres burned):

➔ Fire Causes:

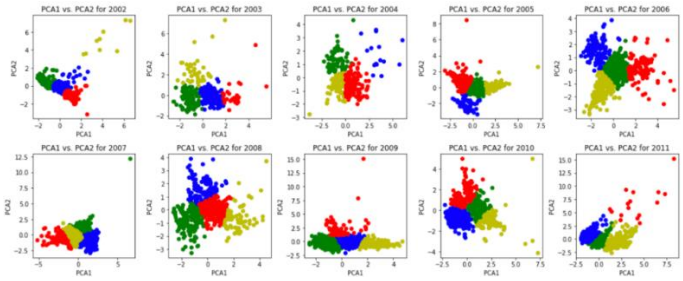
From this experiment, we would like to learn from the dataset about the major causes of each fire for each state and each year. The reason we want to investigate the causes of wildfire is that we think it can potentially be a factor to our classification models of predicting the AQI. From the results and charts, we've learned that debris burning would be the main causes of wildfire. With closer examination, we stacked the count number of each year from 1992-2015 and each cause together to have a better feel of the trends. The following bar chart utilized the STAT_CAUSE_DESCR (I.e. cause of fire) feature of our fires dataset.



4. Clustering & Classification:

➔ PCA / KMeans:

PCA/K-Means analysis was performed on 2500 fires for each year over the range of dates (i.e.,1992-2015). Data was first normalized with StandardScaler. Histograms of the explained variance for the new components and the elbow method plots were used for the yearly analysis as well. This is one instance of dimensionality reduction utilized in our actual analysis to map the data into more informative z-space vectors. The yearly clustering for a subset of the years is shown in the figures below. For visualization. scatterplots suffice. The features used in the PCA reduction were fire size (FIRE_SIZE), AQI level averaged over the entire duration of the fire (AVG_AQI), suppression/containment time of the fire (CONT_TIME), location variables (LATITUDE, LONGITUDE), and the length of each fire measured in days (DAYS). These incorporate more features (in z-space) and so provide insight on the similarity of each fire using more data. Notice the data points form roughly equal clusters, and they are not crowded in a corner, suggesting stronger accuracy in the grouping. Packages used for this analysis were sklearn (I.e. cluster, preprocessing, KMeans, PCA), seaborn, matplotlib, numpy, and pandas.



Conclusion

We ran many different experiments to find a correlation between our wildfire dataset and AQI dataset. We also tested acreage burning and clustering unsupervised learning methods. In quite a few models, we could not show that there is a strong correlation that would affect the AQI of a state or a county in a duration of time. There is simply not enough linear correlation between those two features. Our logistic regression models demonstrate moderate predictive power for all 7 days, implying our feature set has predictive power for AQI, but likely relies on features other than wildfires for prediction. However, we did find slight positive correlation when we calculate the AQI changes with a threshold of fire size and fire duration. This proves that our experiments and assumptions weren't entirely wrong, and our forecasting model could be used as a rough estimation of the AQI prediction within a week. We were able to obtain some more optimistic results in terms of the acreage analysis. Acreage burning across all years, as well as annual trend lines, revealed two things: (1) that yearly fires tend to be most intense in the summer months, and (2) that fire intensity escalated in a linear slope from 1992 to 2015, peaking in 2005 and 2012 (see Analysis Dropdown -- Acres Analysis (Aggregate) discussion on the webpage). Our choropleth accurately reflects these observations. Additionally, our 7-day forecast model provides a meaningful indicator of hazardous AQI even 7 days out. It is noteworthy that unsupervised learning algorithms did represent equal groupings of fires using more features (with proper dimensionality reduction, of course), validating that a syncretizing of more data was needed to form valid correlations/groupings. Scatterplot renderings were provided in this case. Finally, our visualization application provides any user a clear view of wildfires and AQI for any date and most counties between 1992 – 2015. The choropleth can act as an indicator of the air quality of the specific region the user might be interested in.

Work Distribution

Work Distribution		Wendell Hom	Po Hsien Hsu	Yeong Jer Tseng	Samuel Stentz	Srikesh Srinivas	Lauren Vossler
Visualization	Choropleth	O				O	
	User Interface	O	O			O	
Analysis	Data preprocess	O	O	O	O	O	
	Forecast				O		
	Correlation	O	O	O	O	O	
	Clustering				O	O	
Presentation	Final Report/ Poster		O	O			O

Literature References

[1] O'neill, S. M., Larkin, N. S. K., Hoadley, J., Mills, G., Vaughan, J. K., Draxler, R. R., ... & Ferguson, S. A. (2008). Regional real-time smoke prediction systems. *Developments in environmental science*, *8*, 499-534.

[2] Byun, D., & Schere, K. L. (2006). Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system.

[3] JGR Atmospheres. Volume (106, IssueLarkin, N. K., O'Neill, S. M., Solomon, R., Raffuse, S., Strand, T., Sullivan, D. C., ... & Ferguson, S. A. (2010). The BlueSky smoke modeling framework. *International Journal of Wildland Fire*, *18*(8), 906-920.

[4] Bey, I., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B. D., Fiore, A. M., ... & Schultz, M. G. (2001). Global modeling of tropospheric chemistry with assimilated meteorology: (Model description and evaluation. *Journal of Geophysical Research: Atmospheres*, *106*(, 23073-23095.

[5] Cetegen, B. M. (1982). *Entrainment and flame geometry of fire plumes* (Doctoral dissertation, California Institute of Technology).

[6] Kang, G. K., Gao, J. Z., Chiao, S., Lu, S., & Xie, G. (2018). Air quality prediction: Big data and machine learning approaches. *International Journal of Environmental Science and Development*, *9*(1), 8-16.

[7] Running, S. W. (2006). Is global warming causing more, larger wildfires?. *Science*.

[8] Phuleria, H. C., Fine, P. M., Zhu, Y., & Sioutas, C. (2005). Air quality impacts of the October 2003 Southern California wildfires. *Journal of Geophysical Research: Atmospheres*, *110*(D7).

[9] O'neill, S. M., Larkin, N. S. K., Hoadley, J., Mills, G., Vaughan, J. K., Draxler, R. R., ... & Ferguson, S. A. (2008). Regional real-time smoke prediction systems. *Developments in environmental science*, *8*, 499-534.

[10] Wain, A., Mills, G., McCaw, L., & Brown, T. (2008). Managing smoke from wildfires and prescribed burning in southern Australia. *Developments in environmental science*, *8*, 535-550.

[11] Zhu, D., Cai, C., Yang, T., & Zhou, X. (2018). A machine learning approach for air quality prediction: Model regularization and optimization. *Big data and cognitive computing*, *2*(1), 5.

[12] Holm, S. M., Miller, M. D., & Balmes, J. R. (2020). Health effects of wildfire smoke in children and public health tools: a narrative review. *Journal of Exposure Science & Environmental Epidemiology*, 1-20.

[13] Black, C., Tesfaigzi, Y., Bassein, J. A., & Miller, L. A. (2017). Wildfire smoke exposure and human health: Significant gaps in research for a growing public health issue. *Environmental toxicology and pharmacology*, *55*, 186-195.

[14] Knorr, W., Dentener, F., Lamarque, J. F., Jiang, L., & Arneth, A. (2017). Wildfire air pollution hazard during the 21st century. *Atmospheric Chemistry and Physics*, *17*(14), 9223.

[15] University of CA Wildfire Smoke and Air Quality Report. https://www.ucop.edu/safety-and-loss-prevention/_files/systemwide_aqwg_report_final_20190925.pdf

[16] Miller, N., Molitor, D., & Zou, E. (2017). Blowing smoke: Health impacts of wildfire plume dynamics. *Retrieved April*, *23*, 2020.

[17] Liu, J. C., Wilson, A., Mickley, L. J., Dominici, F., Ebisu, K., Wang, Y., ... & Anderson, G. B. (2017). Wildfire-specific fine particulate matter and risk of hospital admissions in urban and rural counties. *Epidemiology (Cambridge, Mass.)*, *28*(1), 77.

[18] Martínez-de Dios, J. R., Merino, L., Caballero, F., & Ollero, A. (2011). Automatic forest-fire measuring using ground stations and unmanned aerial systems. *Sensors*, *11*(6), 6328-6353.