Analysis of air quality and asthma data in California, 2013-2017

https://github.com/aliciaszhao/finalproject

Alicia Zhao

Contents

List of Tables

List of Figures

1 Rationale and Research Questions

In the United States, the National Ambient Air Quality Standards from the 1970 Clean Air Act help monitor air pollutant levels. While significant reductions in air pollution have been made since the standards were put in place, there are still many areas that do not meet the standards. California, in particular, has been cited as a leader in air pollution, with cities such as Bakersfield, Fresno, Los Angeles and San Francisco having some of the highest recorded levels of ozone levels in the country.

From a health perspective, exposure to higher levels of air pollutants such as ozone (O3), nitrogen dioxide (NO2), and particulate matter (PM2.5) is associated with reduced lung function, asthma exacerbations, increased hospital visits, and death (Schraufnagel et al., 2019). In fact, asthma is the leading chronic condition in children, affecting 1 in 12 children in the United States (Zahran, 2018). According to a recent Center for Disease Control and Prevention (CDC) study, children have higher rates of hospital and emergency department visits associated with asthma compared to adults (Moorman et al., 2012). The health impacts of asthma are not distributed equally among children, however. Prevalence of asthma in children can differ by age, family history, racial and ethnic group, and socioeconomic status (U.S. EPA, 2013).

As such, my study seeks to answer two main questions:

- **Question 1:** What are the trends of O3, NO2 and PM2.5 in some of the most polluted cities in California over a five-year period, from 2013 to 2017 ?
- **Question 2:** How is asthma incidence in California impacted by O3 and PM2.5 levels, accounting for socioeconomic variables? Is this relationship different in children compared to adults?

2 Dataset Information

2.1 EPA air quality datasets

Air quality data were collected using EPA's Download Daily Data tool (Table 1).

Table 1: Summary of selections made for daily air quality datasets.

Option	Selection
Pollutant	PM2.5, Ozone and NO2
Year	2013-2017
Geographic Area	California
Download	Download CSV (spreadsheet)

The downloaded files, which were accessed on 2020-04-11, were saved in the project folder path `./Data/Raw/Air quality/` as `EPAair_[Pollutant]_CA_[Year]_raw.csv`. For example, the O3 dataset for 2017 was saved as `EPAair_O3_CA_2017_raw.csv`.

2.1.1 PM2.5 dataset content information

This datset contains daily mean PM2.5 concentrations and the corresponding air quality index (AQI) value in California over the years 2013-2017.

The number of sites and counties in the dataset varied slightly by year (Table 2).

Table 2: Summary of sites and counties for PM2.5 dataset.

Year	Sites	Counties
2013	149	52
2014	152	52
2015	151	52
2016	151	52
2017	152	51

2.1.2 O3 dataset content information

This dataset contains daily maximum 8-hour O3 concentrations and the corresponding air quality index (AQI) value in California over the years 2013-2017.

The number of sites and counties in the dataset varied slightly by year (Table 3).

Table 3: Summary of sites and counties for O3 dataset.

Year	Sites	Counties
2013	183	49
2014	182	49
2015	182	49
2016	182	49
2017	181	49

2.1.3 NO2 dataset content information

This dataset contains the daily maximum 1-hour NO2 concentration and the corresponding air quality index (AQI) value in California over the years 2013-2017.

The number of sites and counties in the dataset varied slightly by year (Table 4).

Table 4: Summary of sites and counties for NO2 dataset.

Year	Sites	Counties
2013	102	33
2014	105	33
2015	108	33
2016	109	33
2017	106	33

All three datasets contain 20 variables (Table 5). Variable names without descriptions are self-explanatory.

Table 5: Summary of variables in air quality datasets.

Variable	Description
Date	Month/day/year
Source	AQS (Air Quality System) or AirNow
Site ID	A unique number within the county identifying the site
POC	Parameter Occurrence Code used to distinguish different instruments that measure the same parameter at the same site
Daily Mean PM2.5 Concentration	
Daily Max 8-hour Ozone Concentration	
Daily Max 1-hour NO2 Concentration	
Units	Units for concentration
Daily_AQI_Value	Air quality index (range 0-500)
Site Name	

Variable	Description
DAILY_OBS_COUNT	Number of observations per day
PERCENT_COMPLETE	
AQS_PARAMETER_CODE	
AQS_PARAMETER_DESC	
CBSA_CODE	The FIPS code of the metropolitan area
CBSA_NAME	
STATE_CODE	The FIPS code of the state
STATE	
COUNTY_CODE	The FIPS code of the county
COUNTY	
SITE_LATITUDE	
SITE_LONGITUDE	

Table 6: Air quality index (AQI) values and corresponding levels of health concern.

AQI.Values	Levels.of.Heath.Concern
0-50	Good
51-100	Moderate
101-150	Unhealthy for Sensitive Groups
151-200	Unhealthy
201-300	Very unhealthy
301-500	Hazardous

2.1.4 Data wrangling

Air quality datasets for different years were combined using `rbind` to form one dataset for each pollutant.

The following columns were selected:

- **Date**
- **Site.ID**
- **Daily.Max.1.hour.NO2.Concentration**
- **DAILY_AQI_VALUE**
- **Site.Name**
- **COUNTY**

Additionally, columns for **Month** and **Year** were added using the **Date** column.

2.2 Asthma datasets

Asthma data were collected using Tracking California's Asthma Data Query tool (Table 7).

Table 7: Summary of selections made for daily asthma datasets.

Option	Selection
Type of event	Emergency department visits due to asthma
Age sub-group	Age 0-17, AGe 18 & over
Year	2013-2017
How event is measured	Age-adjusted rates per 10,000
Race/ethnicity	All races/ethnicities
Gender/sex	Both sexes
Type of information	Conventional
Type of geography	Zip codes

The downloaded files, which were accessed on 2020-04-12, were saved in the project folder path `./Data/Raw/Asthma/` as `TrackingCA_Asthma_ERVisits_[Age Group]_[Year]_raw.csv`.
For example, the dataset for adults in 2013 was saved as `TrackingCA_Asthma_ERVisits_Adults_2013_raw.`

2.2.1 Adult asthma dataset content information

This dataset contains the annual rates of asthma-related ER visits for adults in California over the years 2013-2017.

2.2.2 Children asthma dataset content information

This dataset contains the annual rates of asthma-related ER visits for children in California over the years 2013-2017.

Both datasets contain 2 variables. Variable names without descriptions are self-explanatory.

Table 8: Summary of variables in asthma datasets.

Variable	Description
Zip code	
Incidence	Age-adjusted rate of emergency department visits due to asthma per 10,000 California residents

2.2.3 Data wrangling

A **Year** column was added to all asthma datasets. Datasets for 2013-2017 were combined using `rbind` to form one dataset for each age group (Adults, Children).

Since the asthma datasets provide only zip code information but not county information, an online search was done to find zip codes and their corresponding counties in California. This information was scraped into a data frame. This dataframe was then combined with the adult dataset and the children dataset using `left_join`. The incidence rates were then grouped by county and the average incidence rate for each county was calculated.

Finally, adult and children asthma datasets were combined using `full_join`.

2.3 Demographics dataset

Demographic data were collected from County Health Rankings & Roadmaps.

Since demographics are assumed to remain somewhat constant over the five-year period, only one dataset was chosen. The 2019 dataset, which uses data published in 2017, was chosen.

The xls file, was accessed on 2020-04-12, was saved in the project folder path `./Data/Raw/Demographics/` as `CountyHealthRankings_CA_2019_raw.xls`. Since there multiple tabs in the xls file, relevant information from the file was taken and converted into a csv file.The csv file was saved as `CountyHealthRankings_CA_2019_filtered_raw.csv`.

The following information is contained in the dataset:
* **FIPS**
* **State**
* **County**
* **Median Household Income**
* **Population**
* **Percent African American**
* **Percent Rural**
* **Percent Smokers**

3 Exploratory Analysis

3.1 Air quality datasets

In examining the air quality index (AQI) values for the three pollutants across all sites, there is not much temporal variation (Figure 1). However, the frequency distrbution does differ by pollutant. NO2 generally has lower AQI values, which aggregate well below 50, the cut-off number for the "Good" air quality range. O3 and PM2.5 have higher AQI values, which concentrate closer to 50, and also contain more values that are above 50.

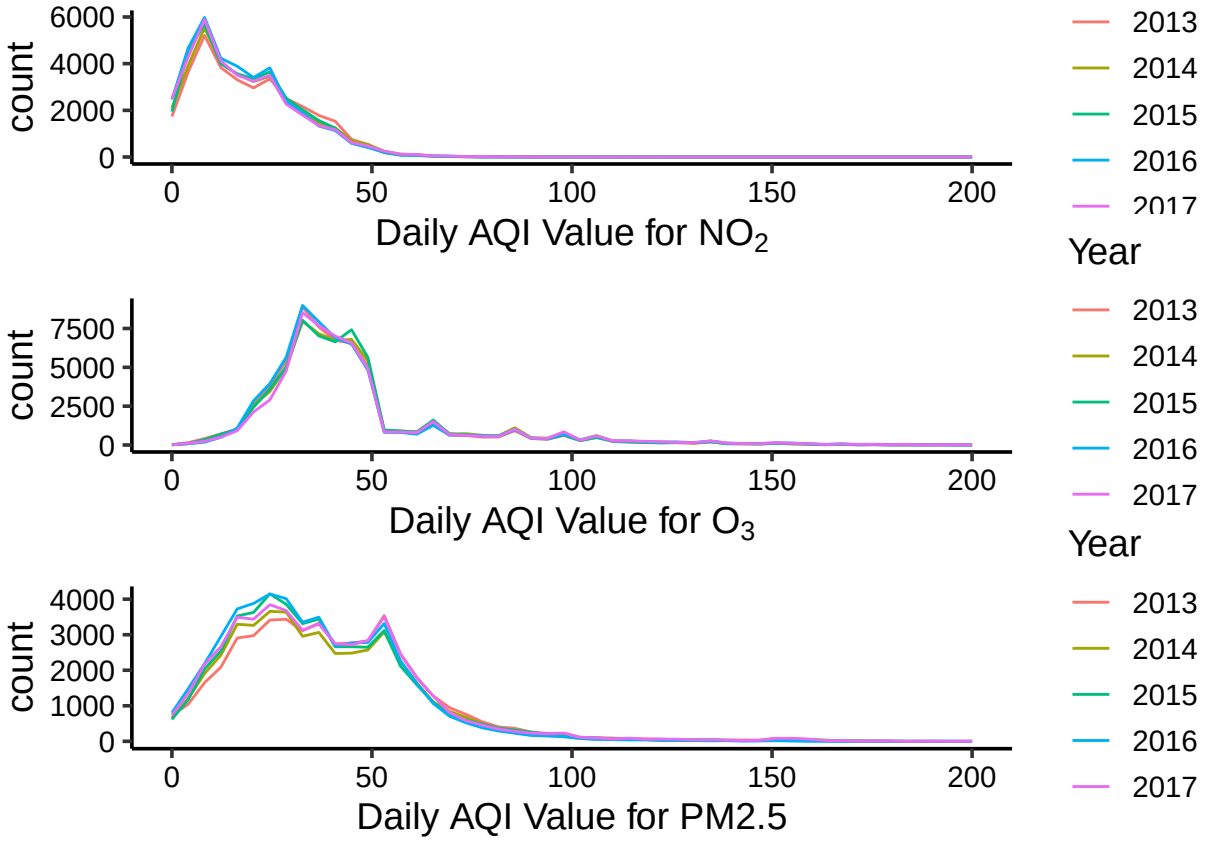

Figure 1: Frequency of daily AQI values in California from 2013 to 2017.

Among the four target sites (Bakersfield, Fresno, Los Angeles and San Francisco), Bakersfield appears to have higher AQI values than the other sites across all pollutants (Figure 2). In contrast, San Francisco generally has lower AQI values compared to the other sites. Although AQI values for NO2 stay below the "Unhealthy for sensitive populations" range (denoted by the dashed line) for all sites, the AQI values for O3 and PM2.5 are generally in this range.

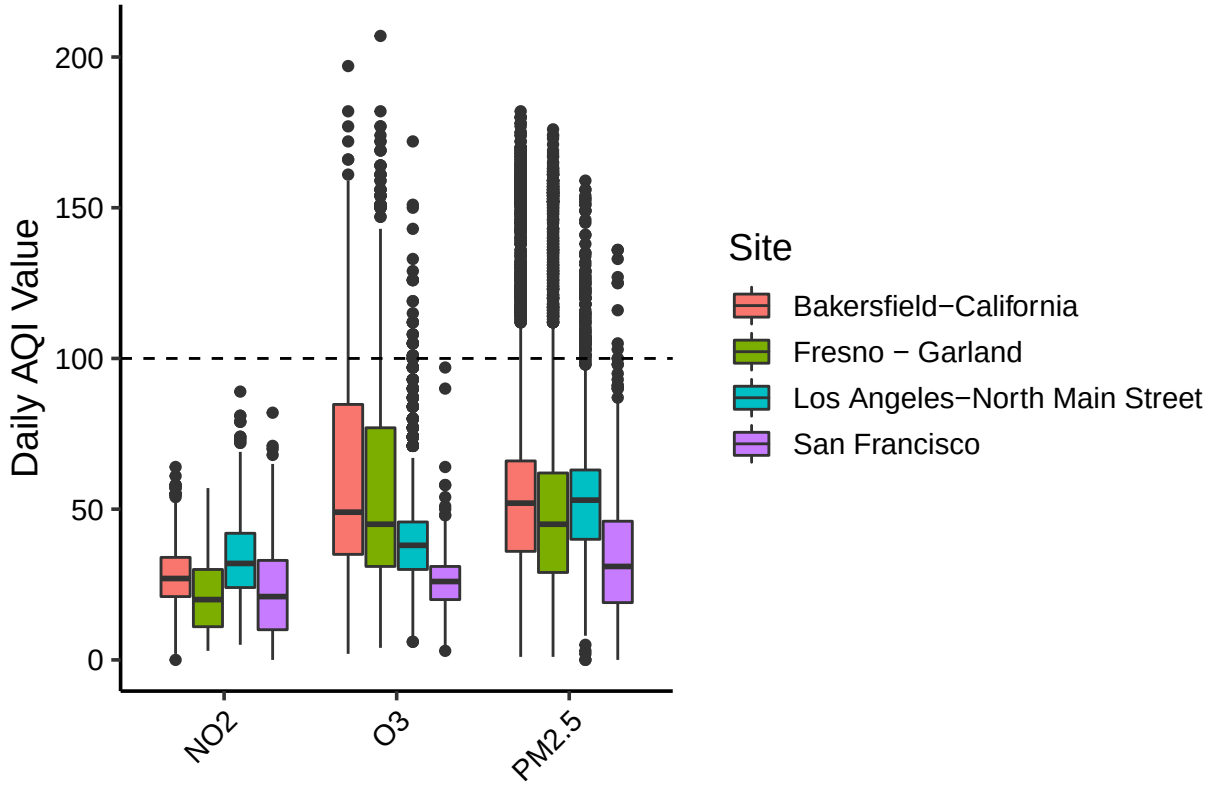

Figure 2: Boxplots of daily AQI values in four key sites from 2013 to 2017.

3.2 Asthma datasets

In examining the asthma-related ER visits for adults and children across 58 counties in California, children appear to have more asthma-related ER visits than adults (Figure 3). Distributions for both age groups show a right skew.

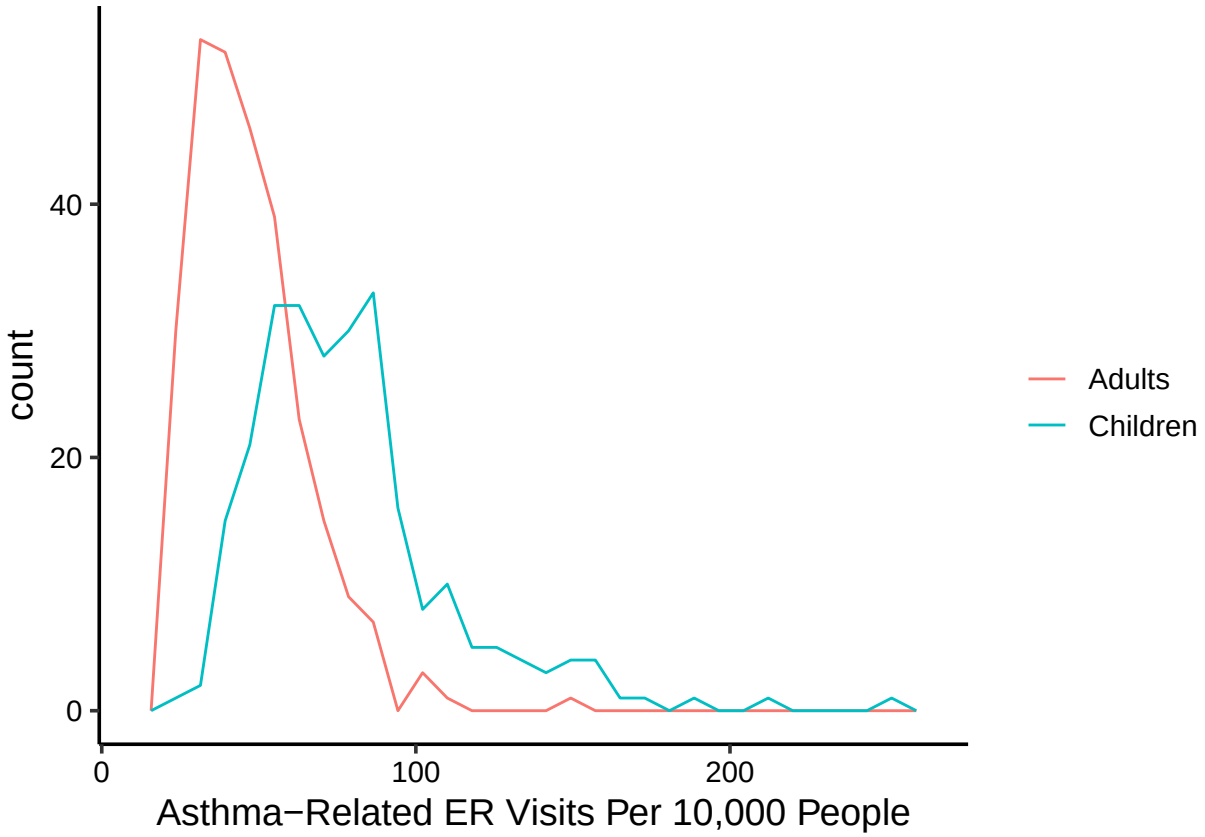

Figure 3: Frequency of asthma-related ER visitation rates for children and adults in California from 2013 to 2017.

3.3 Demographics dataset

In examining demographics for 52 counties in California, counties generally have a low percentage of African American residents and a much higher percentage of Hispanic residents; a median household income clustering between 40,000 and 75,000; and more urban areas than rural areas (Figure 4).

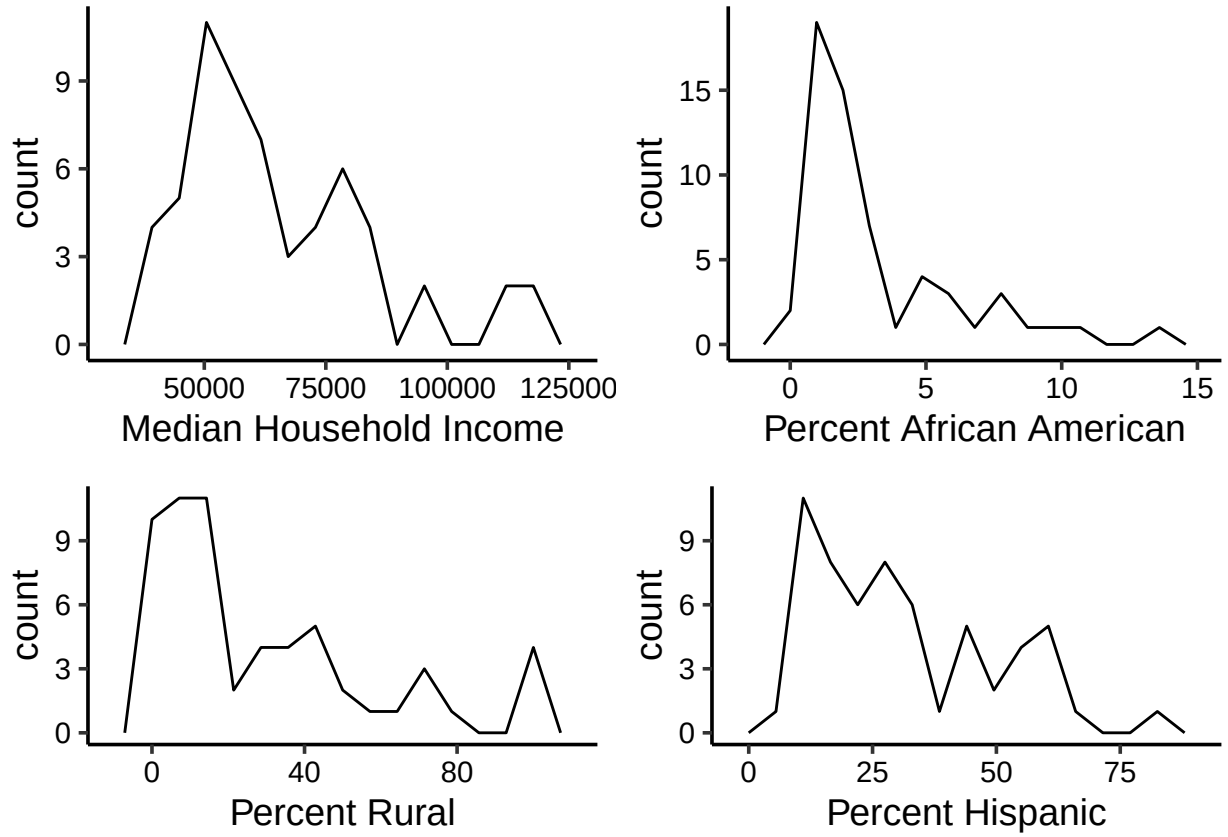

Figure 4: Frequency of median household income, percent African American, percent rural, and percent Hispanic across counties in California from 2013 to 2017.

4 Analysis

4.1 Time Series Analysis

The time series analysis method was used to determine trends of O3, NO2 and PM2.5 in some of the most polluted cities in California from 2013 to 2017 (Figures 5-8). The four sites chosen were: Bakersfield, Fresno, Los Angeles and San Francisco.

The dashed lines in the figures separate the AQI categories into **'Good'**, **'Moderate'**, and **'Unhealthy for sensitive groups'**. San Francisco appears to have the lowest levels of pollution, whereas O3 and PM2.5 levels for both Bakersfield and Fresno have reached the unhealthy level.

In performing seasonal Mann-Kendall tests to these time series, it was observed that there are some trends for the pollutants beyond seasonal trends.

4.1.1 Bakersfield

There is a decreasing overall trend for NO2 (seasonal Mann-Kendall, z= -2.41, p-value = 0.016), a decreasing overall trend for PM2.5 (seasonal Mann-Kendall, z = -3.46, p-value < 0.001) and no significant trend for O3 beyond seasonal trends over the period 2013-2017.

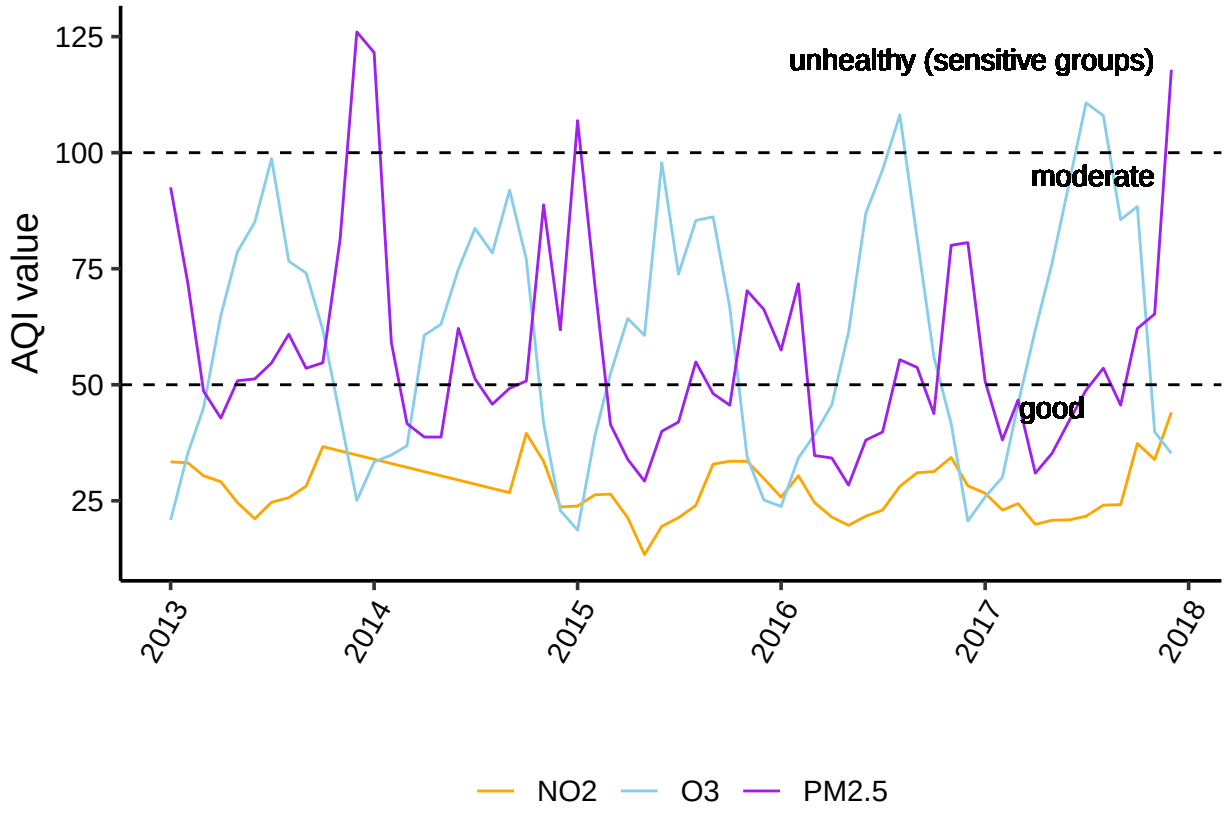

Figure 5: Time series analysis of monthly AQI values in Bakersfield, CA from 2013 to 2017.

4.1.2 Fresno

There is no significant trend for NO2, O3, or PM2.5 beyond seasonal trends over the period 2013-2017.

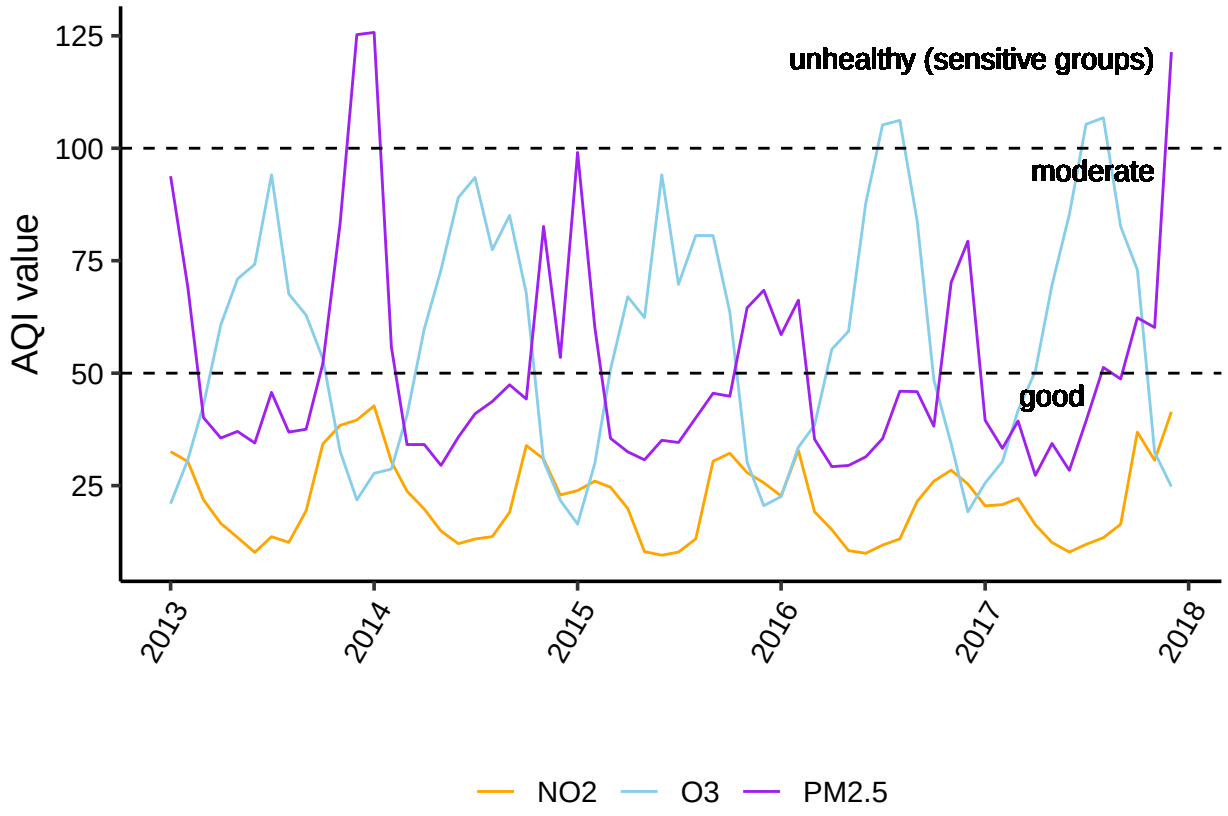

Figure 6: Time series analysis of monthly AQI values in Fresno, CA from 2013 to 2017.

4.1.3 Los Angeles

There is a no significant trend for NO2 beyond seasonal trends, an overall increasing trend for O3 (seasonal Mann-Kendall, z = 3.18, p-value =.0001), and an overall decreasing trend and a decreasing trend in April for PM2.5 (seasonal Mann-Kendall, z = -2.05, p-value = 0.040; z = -2.205, p-value= 0.027) over the period 2013-2017.

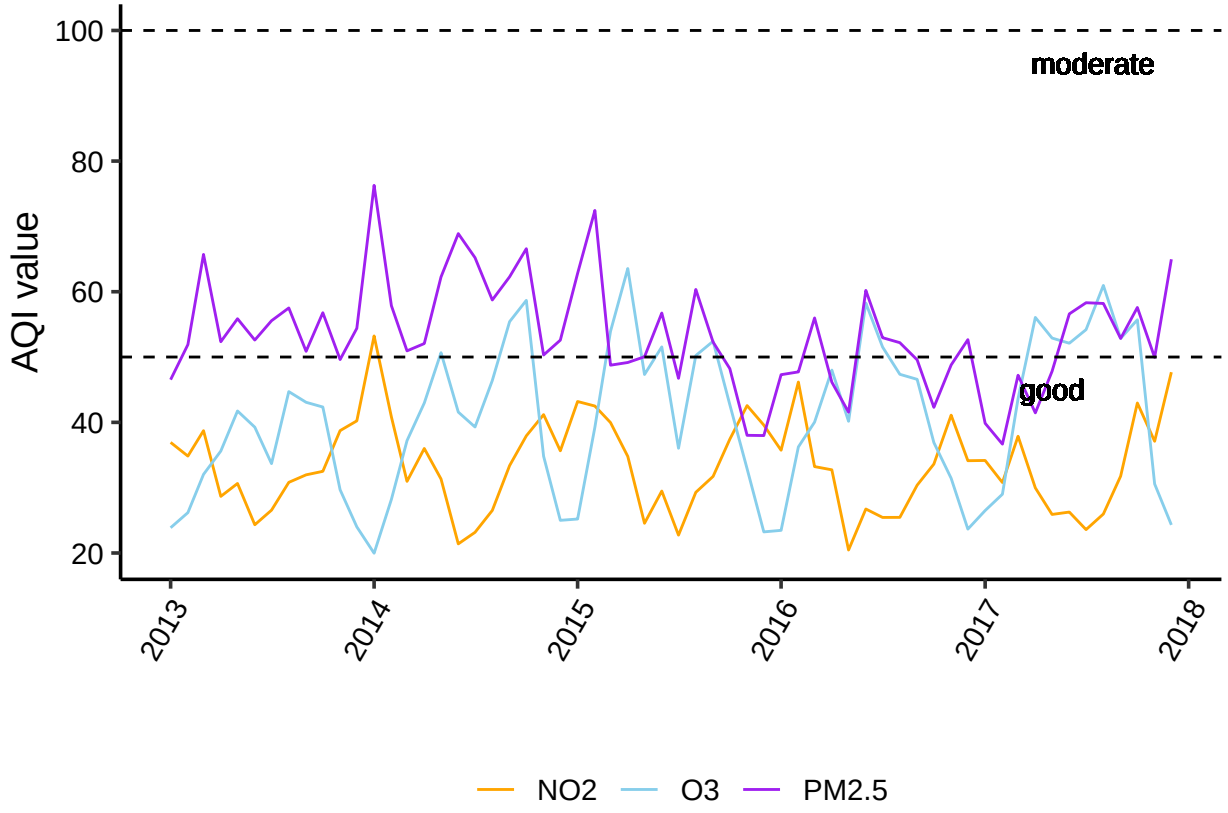

Figure 7: Time series analysis of monthly AQI values in Los Angeles, CA from 2013 to 2017.

4.1.4 San Francisco

There is an overall decreasing trend for NO2 (seasonal Mann-Kendall, z = -2.192, p-value = 0.028), and no significant trend for O3 or PM2.5 over the period 2013-2017.

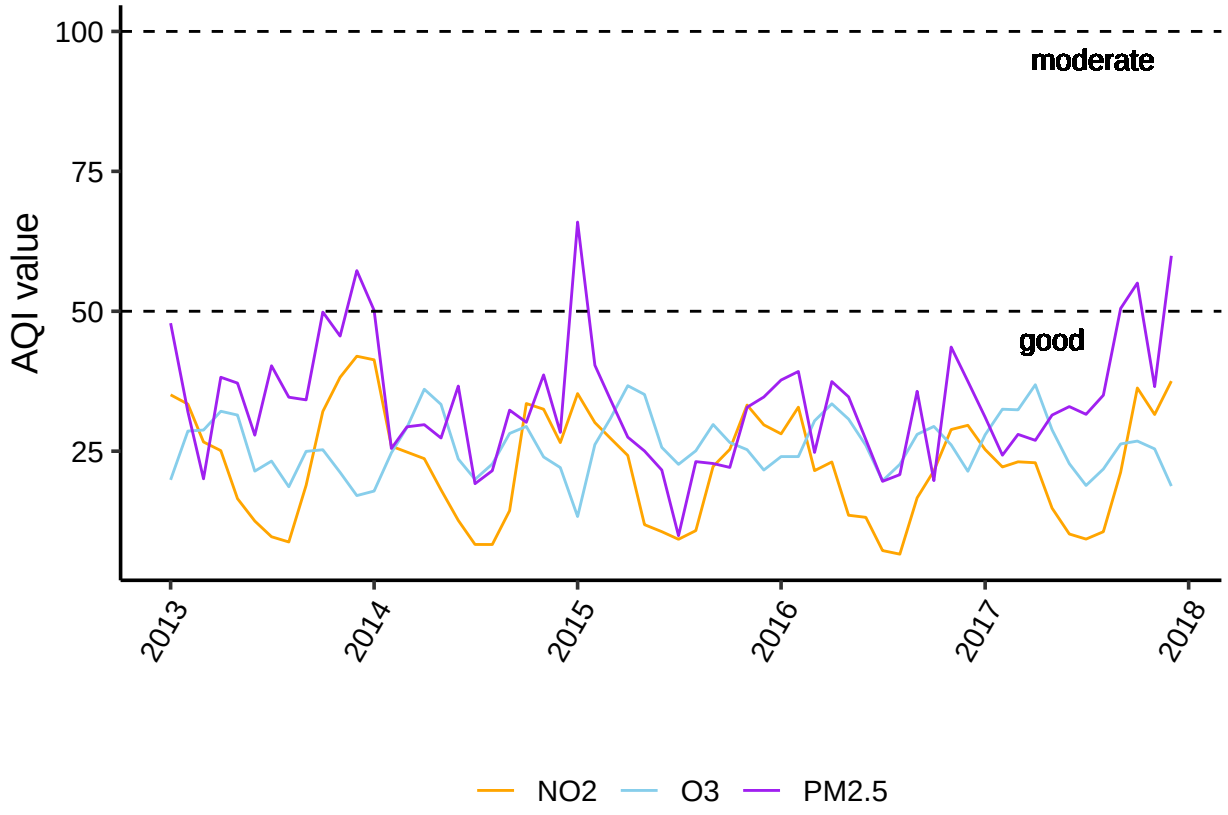

Figure 8: Time series analysis of monthly AQI values in San Francisco, CA from 2013 to 2017.

4.2 General Linear Model

The general linear model (GLM) method was used to answer how asthma incidence in children and adults is impacted by O3 and PM2.5 levels and socioeconomic variables.Data were evaluated on a county-level across the years 2013 to 2017.

4.2.1 Dependent variables

The dependent variable is asthma incidence, measured by using age-adjusted asthma-related ER visitation rates per 10,000 residents. This rate was evaluated separately for adults and children.

4.2.2 Explanatory variables

The explanatory variables are:
* **Annual average O3 AQI value**
* **Annual average PM2.5 AQI value**
* **Median household income**
* **Percent African American**
* **Percent Hispanic**
* **Percent rural**
* **Percent smokers**

O3 and **PM2.5** were included to evaluate the impact of air quality on asthma incidence. **NO2** was not chosen because as the exploratory analysis showed, NO2 levels are not as high as O3 and PM2.5 levels, and also there were more missing data for NO2.

Additionally, a few demographic variables were also evaluated. **Median household income** was chosen to account for socioeconomic factors. **Percent African American** and **Percent Hispanic** were chosen to account for racial factors. **Percent rural** was chosen to account for differences in air quality and healthcare access. **Percent smokers** was chosen to account for other health factors that could impact asthma.

4.2.3 Asthma incidence in children

A multiple linear regression was run to predict asthma-related emergency room visitation rates in children while accounting for the explanatory variables described previously.

The most parsimonious model includes the variables O3, percent African American, percent Hispanic, percent rural and percent smokers. This model has an adjusted R2 of .66, with df = 159 and p <.001.

This model specifies that holding all other variables constant, for every unit increase in the AQI for O3, the ER visitiation rate decreases by .67. For every percent increase in African American populations, ER visitation rate increases by 5.3. For every percent increase in Hispanic populations, the ER visitation rate increase by 1.2. For every percent increase in smokers, the ER visitations rate increases by 2.9.

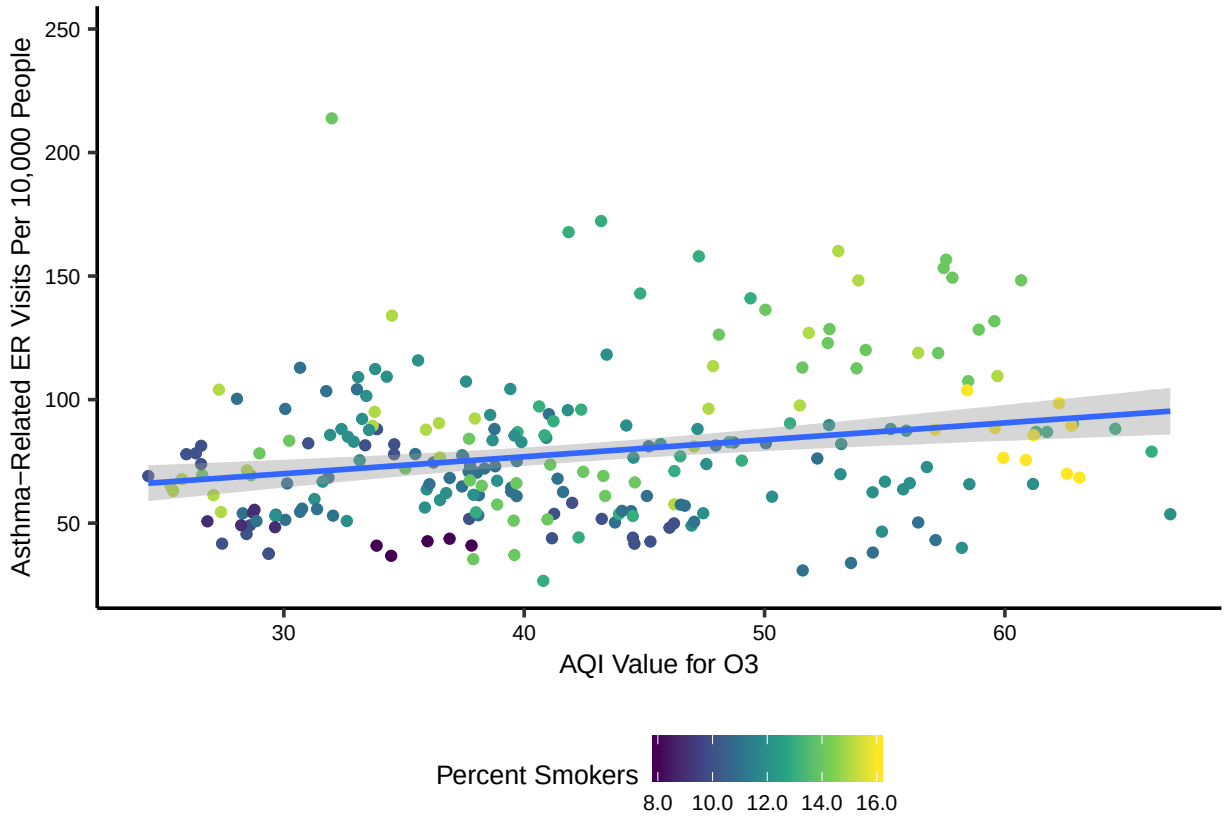

Figure 9: Annual average O3 AQI values, percent smokers, and asthma-related ER visitation rates for children across 52 counties in California from 2013 to 2017.

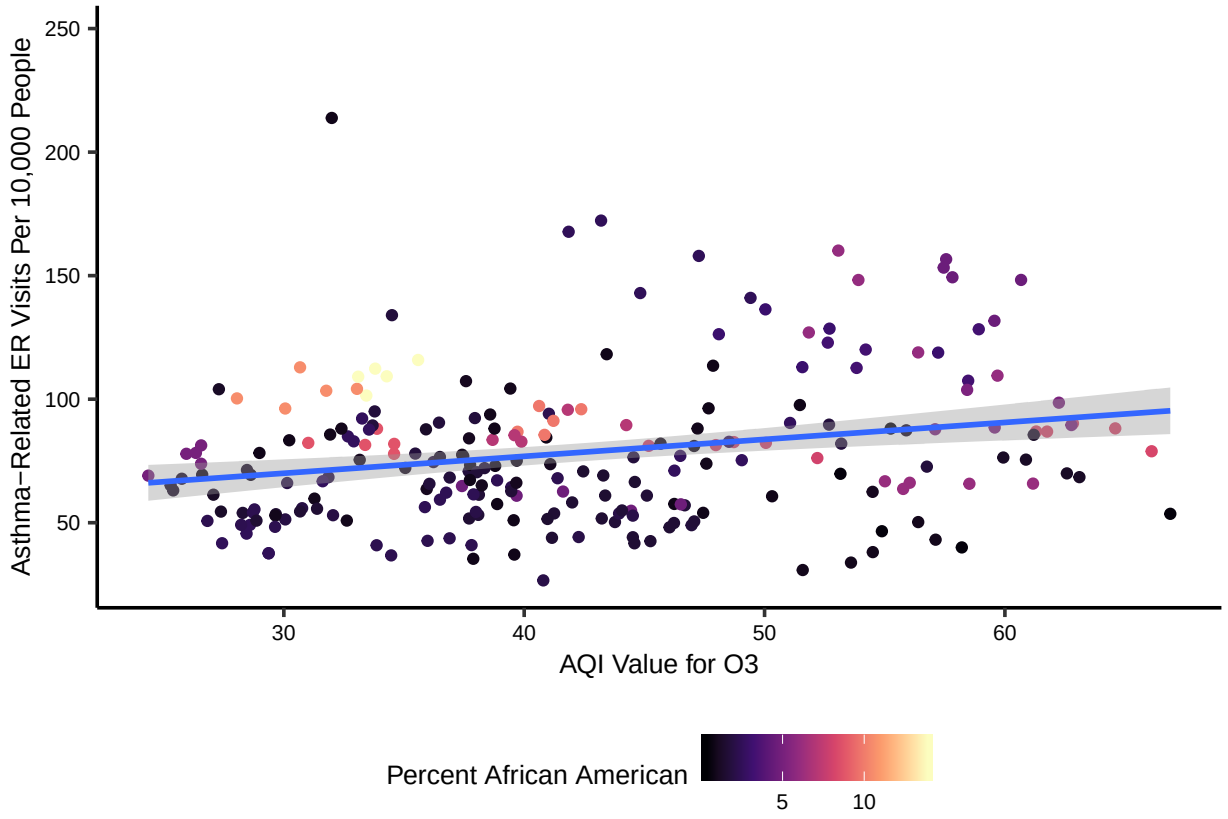

Figure 10: Annual average O3 AQI values, percent African American, and asthma-related ER visitation rates for children across 52 counties in California from 2013 to 2017.

4.2.4 Asthma incidence in adults

A multiple linear regression was run to predict asthma-related emergency room visitation rates in adults while accounting for the explanatory variables described previously.

The most parsimonious model includes the variables O3, percent African American and percent smokers. This model has an adjusted R2 of .59, with df = 161 and p <.001.

This model specifies that holding all other variables constant, for every unit increase in the AQI for O3, the ER visitiation rate decreases by .41. For every percent increase in African American populations, the ER visitation rate increases by 2.4. For every percent increase in smokers, the ER visitations rate increases by 5.3.

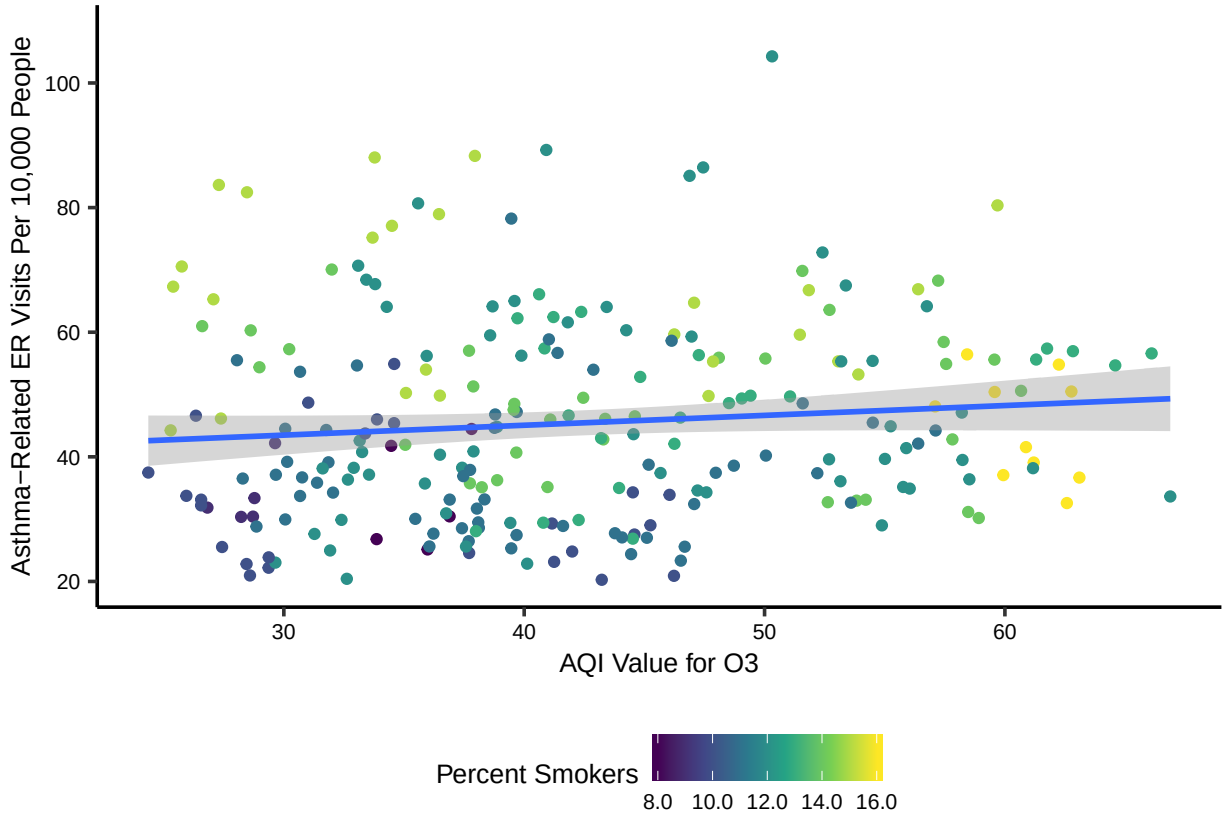

Figure 11: Annual average O3 AQI values, percent smokers, and asthma-related ER visitation rates for adults across 52 counties in California from 2013 to 2017

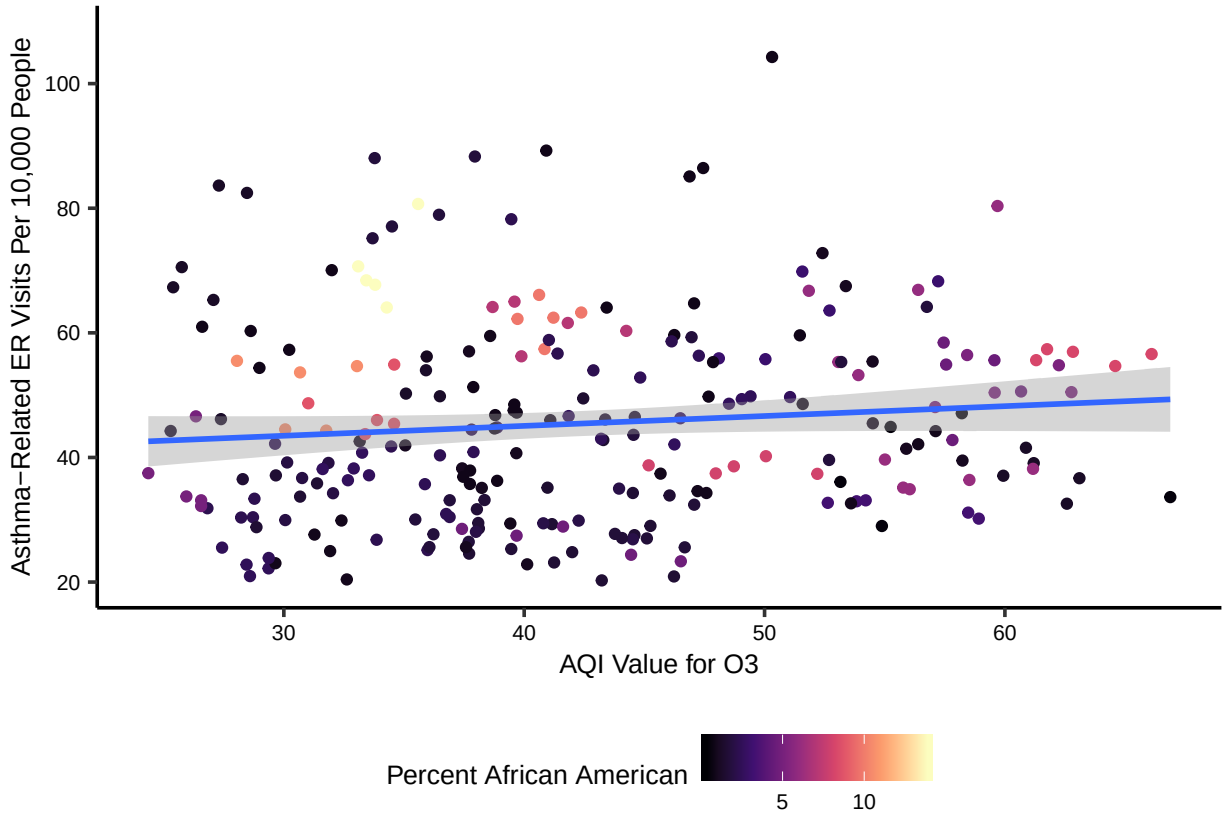

Figure 12: Annual average O3 AQI values, percent African American, and asthma-related ER visitation rates for adults across 52 counties in California from 2013 to 2017

5 Summary and Conclusions

5.1 Question 1: What are the trends of O3, NO2 and PM2.5 in some of the most polluted cities in California over a five-year period, from 2013 to 2017 ?

The time series analysis from this study suggest that the most part, there is either no trend or a decreasing trend for all pollutants across the sites, which is encouraging to find. The exception is O3 in LA, which has shown an increasing trend. However, this is not too much of a concern, as ozone levels are only bordering the 'Moderate' range but mostly still in the 'Good' range.

5.2 Question 2: How is asthma incidence in California impacted by O3 and PM2.5 levels, accounting for socioeconomic variables? Is this relationship different in children compared to adults?

Findings from this study suggest that asthma incidence in both children and adults are impacted by *O3 concentration levels*, *percent African Americans*, *percent rural* and *percent smokers* in each county. However, incidence in children is more impacted by race (such as *percent African American* and *percent Hispanic* in each county) when compared to incidence in adults. On the other hand, asthma incidence in adults is more sensitive to *percent smokers* in each county.

Surprisingly, PM2.5 was not found to be a significant variable in our models. Additionally, although O3 was a significant variable, it has a slightly inverse relationship with asthma incidence, which is contrary to what was expected. One explanation is that because air pollution levels are generally in the healthy range in California, their impacts on asthma incidence are negligible when compared to the impacts of socioeconomic and racial variables. There also are likely temporal, spatial and other types of variables which were left out in the model, which can cause omitted variable bias.

Based on our results from 2013-2017, California counties with a higher percentage of African American populations and a higher percentage of smokers are more likely to have higher asthma incidence rates in both children and adults. These results reveal that policies targeting better health outcomes may not achieve desired results by only reducing air pollution. Pre-existing disparities associated with socioeconomic and racial status may play a larger role in impacting asthma incidence, and should be addressed in conjunction with air quality policies.

6 References

Moorman JE, Akinbami LJ, Bailey CM, Zahran HS, King ME, Johnson CA, et al. National surveillance of asthma: United States, 2001–2010. (2012). Vital Health Stat 3 (35):1–58.

Schraufnagel, D.E. et al. (2019). Air pollution and noncommunicable diseases: a review by the forum of International Respiratory Societies' Environmental Committee, part 1. Chest. 155(2), pp.409-416.

U.S. EPA. (2013). America's Children and the Environment, Third Edition.

Zahran, H.S., Bailey, C.M., Damon, S.A., Garbe, P.L. & Breysse, P.N. Vital signs: Asthma in children – United States, 2001-2016. (2018). Morb Mortal. Wkly Rep. 67, pp.149-155.