Milestone 5

Team 20: Karla Guerrero, Qianhui Jin, KC Whitsett

11/29/2021

Problem Statement

There are many known health issues associated with air pollution, often measured by Particulate Matter (PM), a term for solid or liquid particles or droplets that are found in the air (such as dust, smog, or smoke). PM of size 2.5 micrometers or smaller are very fine inhalable particles that are particularly damaging to lung health[1]. Our team seeks to investigate the potential links between Air Quality Levels and Chronic Lower Respiratory Disease (CLRD) Mortality in California.

Currently, PM2.5 levels of under 12.0 are considered "good," and under 35.4 are considered "moderate" according to the US Environmental Protection Agency. It is not until we reach levels over 35.5 that air quality is considered "unhealthy for sensitive groups," and over 55.5 to be considered "unhealthy." Our team believes that this threshold of "safe" PM2.5 levels should be drastically lowered in California. We will investigate aggregate statewide PM2.5 measurements and CLRD Mortality counts 1) at a zip code level to determine where further research should be conducted, and 2) examine a potential correlation between zip codes with critically high PM2.5 levels and CLRD deaths from 2009-2018.

Methods

Data Sources

Exposure data: This dataset comes from CalEnviroScreen3.0 dataset from the California Environmental Protection Agency (CalEPA). It contains three columns, county, zip code and average PM2.5 per zip code.
Outcome data: This dataset contains counts of deaths for California residents by ZIP Code based on information entered on death certificates from 2009-2018.

Cleaning and Creating New Variables

We first removed the NA from our exposure data and obtained the 75th percentile (11.83) of the average PM2.5 per zip code. We then grouped our exposure data by zip code and created a new binary column to indicate whether or not the zip has PM2.5 measurement in the critical category of PM2.5 (the 75th percentile). We created a subset of our data which only included zip codes in the critical category. Duplicated zip codes were eliminated from the dataset.

---

[1] Environmental Protection Agency. (n.d.). *Particulate Matter (PM) Pollution*. EPA. Retrieved November 21, 2021, from https://www.epa.gov/pm-pollution/particulate-matter-pm-basics.

We grouped our outcome data by zip code and created a new column called total count containing the total number of CLRD deaths in each zip code from 2009 to 2018. We then filtered zip codes so that only those that are both in our exposure dataset and outcome dataset were retained. Zip codes that were not in the exposure dataset were not included in this new subset.

We combined our exposure data, zip codes and corresponding PM2.5 measurements in the critical category, and the outcome data, zip codes and corresponding CLRD deaths, into one final dataset. Columns of county, zip code, PM2.5 measurements and total number of CLRD deaths were retained. NA were removed from this final dataset. Zip codes were stored as factors. PM2.5 measurements were rounded to two decimal places.

Analytic Methods

A plot containing a bar chart and a line chart was created from our final dataset. The bar chart represents the total number of CLRD deaths in each zip code while the line chart represents the PM2.5 measurements in each zip code. Then linear regression and logistic regression were used to test the relationship between PM2.5 pollution levels and total Chronic Lower Respiratory Disease Mortality among zip codes in the critical category of PM2.5.

Results

Table 1: Chronic Lower Respiratory Disease (CLRD) Mortality by California zip codes with critically high levels of Particulate Matter 2.5 (PM2.5), 2009-2018

Zip Code	County	Average PM2.5	CLDRM Counts
90008	Los Angeles	12.23	151
90043	Los Angeles	12.09	194
90044	Los Angeles	12.10	228
90047	Los Angeles	12.05	204
90220	Los Angeles	12.02	183
90247	Los Angeles	12.05	165
90250	Los Angeles	12.05	219
90280	Los Angeles	12.05	156
90503	Los Angeles	12.05	187
90640	Los Angeles	12.05	159
90650	Los Angeles	11.87	403
90706	Los Angeles	11.98	280
91701	San Bernardino	12.79	158
91710	San Bernardino	12.89	230
91711	Los Angeles	12.26	185
91730	San Bernardino	12.89	180
91762	San Bernardino	12.89	188
91786	San Bernardino	12.81	243
91911	San Diego	11.94	312
92335	San Bernardino	12.94	199

92373	San Bernardino	12.12	222
92376	San Bernardino	12.86	233
92404	San Bernardino	12.05	396
92410	San Bernardino	12.30	171
92503	Riverside	12.99	319
92504	Riverside	12.51	265
92505	Riverside	13.35	151
92509	Riverside	13.79	265
92570	Riverside	12.41	177
93230	Kings	16.89	295
93257	Tulare	15.67	369
93274	Tulare	17.05	301
93277	Tulare	17.08	278
93292	Tulare	16.54	145
93304	Kern	18.92	254
93305	Kern	18.88	153
93306	Kern	19.21	325
93307	Kern	19.31	260
93308	Kern	18.76	469
93309	Kern	18.76	331
93312	Kern	18.76	189
93612	Fresno	15.40	208
93705	Fresno	15.40	163
93727	Fresno	15.40	212
95204	San Joaquin	13.44	196

95205	San Joaquin	13.55	159
95207	San Joaquin	13.44	257
95336	San Joaquin	12.85	255
95340	Merced	12.89	213
95350	Stanislaus	12.89	372
95351	Stanislaus	12.89	202
95355	Stanislaus	12.89	322
95380	Stanislaus	12.89	170

Legend

CLRD mortality counts highlighted in red represent death counts in the 75th percentile. Only zip codes with critical levels of PM2.5 are included in this table, and zip codes with no reported deaths between 2009-2018 were excluded.

Interpretation

Counties with the most number of zip codes with both critical levels of PM2.5 and CLRD deaths include Los Angeles (13), San Bernadino (10), and Kern (7) counties. On the other hand, zip codes that hold the highest death counts in the 75th percentile belong to many counties: Kern, Los Angeles, Stanislaus, Tulare, Riverside, San Diego, and Kings counties, with one Kern zip code (93308) holding the highest death count of 469. Kern zip codes also hold the top 7 highest PM2.5 levels ranging between 18.76-19.31.

Table 2: Chronic Lower Respiratory Disease (CLRD) Mortality by California counties with critically high levels of Particulate Matter 2.5 (PM2.5), 2009-2018

County	Average PM2.5	CLRD Mortality Counts
Los Angeles	12.22	2714
San Bernardino	12.67	2220
Kern	17.57	1981
Riverside	12.97	1177
Tulare	15.56	1093
Stanislaus	12.81	1066
San Joaquin	12.95	867
Fresno	15.24	583
San Diego	13.16	312
Kings	15.86	295
Merced	12.91	213
Imperial	13.23	0
Madera	12.67	0
Orange	12.05	0

Legend

Zip code level PM2.5 and CLRD death data are averaged by county. Only counties with critical levels of PM2.5 are included in this table, and zip codes with no reported deaths between 2009-2018 were excluded.

Interpretation

In accordance with the table above, Los Angeles, San Bernardino, and Kern counties hold the highest death counts. Counties with the highest average PM2.5 levels are Kern, Kings, Tulare, and Fresno (17.57, 15.86, 15.56, and D 15.24, respectively). Despite having critical levels of PM2.5, the counties of Imperial, Madera, and Orange had no reported CLRD deaths between 2009-2018.

Results

Figure 1: Chronic Lower Respiratory Disease (CLRD) mortality and Particulate Matter 2.5 (PM2.5) air pollution measurements per zip code among those with critical levels of 2.5PM (California 2009-2018)

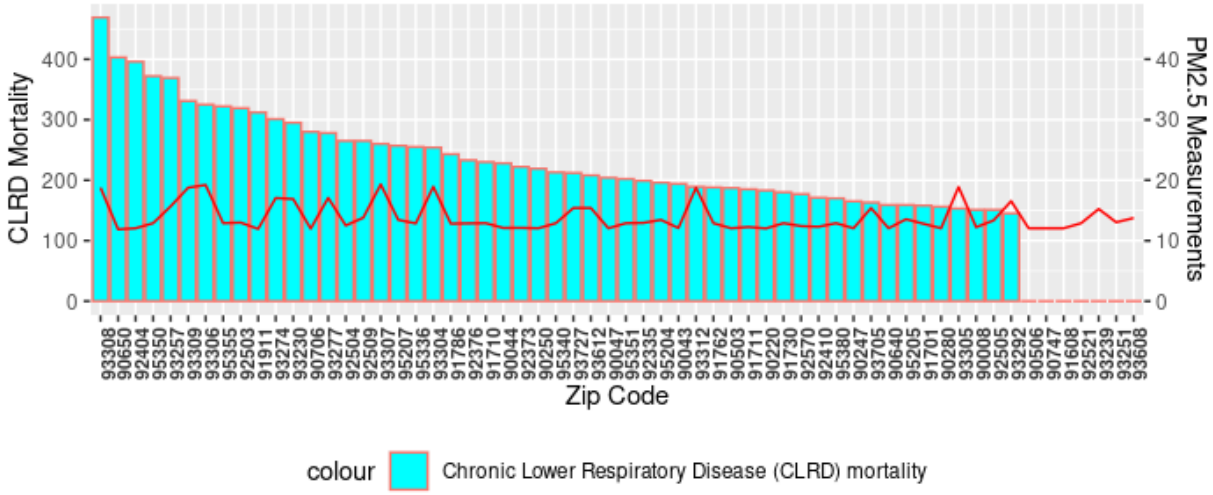

Figure 1 shows the CLRD mortality and PM2.5 air pollution measurements in zip codes where PM2.5 measurements are in the critical category (the highest quantile). The blue bars represent CLRD death counts while the rea line represents PM2.5 air pollution measurements this plot we cannot see a strong correlation between CLRD mortality and PM2.5 measurements.

Figure 2: Total Chronic Lower Respiratory Disease (CLRD) mortality by Particulate Matter 2.5 (PM2.5) air pollution per zip code among those wi critical levels of PM2.5 (California 2009-2018)

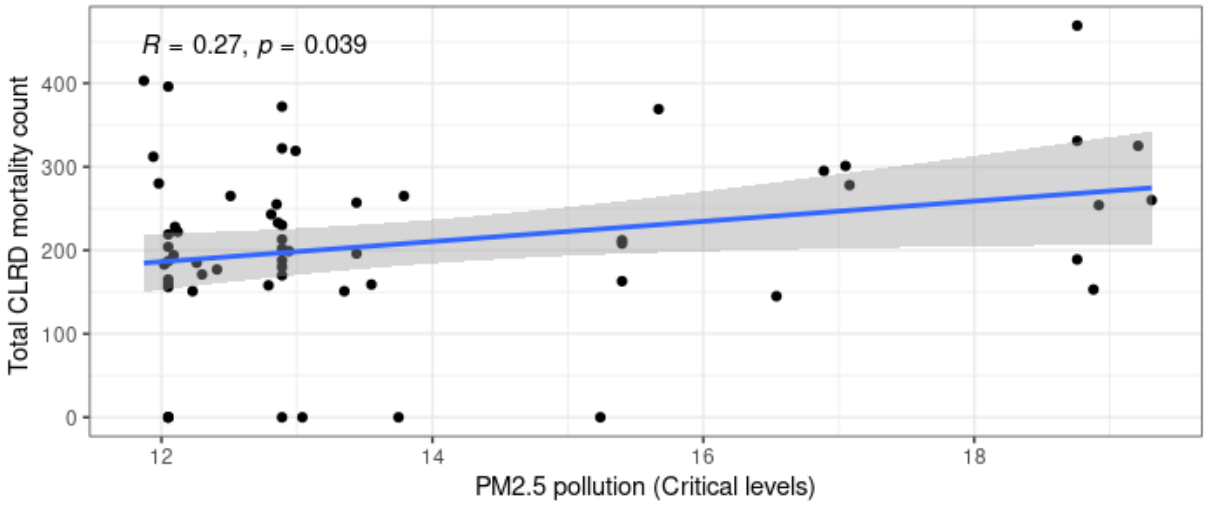

Figure 2 shows a positive association between higher levels of PM2.5 air pollution and increased levels of CLRD mortality per zip code. Our correlation coefficient value is 0.27, which indicates a small correlation but positive correlation between critical levels of PM2.5 air pollution and CLDR mortality. The p-value for our correlation is 0.039, meaning that although our correlation is small, it is still statistically significant.

We used linear regression to find the correlation between PM2.5 pollution levels and total Chronic Lower Respiratory Disease Mortality among zip codes in the quartile with highest levels of PM2.5 pollution (PM2.5 >11.83). Our intuition was that we would see CLDR mortality increase as levels of air pollution increase. If this linear association was true, we would see a correlation value greater than 0. Our logistic regression model found a correlation value of 0.27, which indicates a small association. However, the p-value was less than 0.05 so the association was shown to be statistically significant.

From our tables and visualizations, it is easy to identify the zip codes with the highest levels of PM2.5 pollution or the largest CLRD mortality counts. We believe these tables would be useful for “our non-profit” to add onto its website. However further research on all of these zip codes would be costly, so we recommend a targeted investigation of the zip codes that ranked among the top 25 most burdened for both PM2.5 pollution and Chronic Lower Respiratory Disease Mortality. These zip codes are "93307" (Kern), "93306" (Kern), "93304" (Kern), "93308" (Kern), "93309" (Kern), "93277" (Tulare), "93274" (Tulare), and "93230" (Kings). Further research can include an analysis of the sources of pollution and resident proximity to these locations for comparative analysis against state averages, for example.

Discussion

According to the US Environmental Protection Agency, levels of PM2.5 between 12.1 and 35.4 are considered “moderate”. While the EPA states that these levels of air pollution can be classified as “safe”, we argue that any level of PM2.5 above 11.84 should be reclassified as “unhealthy” in the state of California. Our group believes that truly “safe” levels of PM2.5 are those which demonstrate zero correlation between air pollution and CLRD mortality. However, our analysis suggests that there is a statistically significant correlation between “moderate” levels of air pollution and CLRD mortality (Figure 2). This is further supported by our Figure 1, which demonstrates elevated levels of CLDR mortality for all levels of PM2.5 pollution greater than 11.84. We believe further research in the zip codes most burdened by both PM2.5 and CLDR mortality will provide insight into this relationship. We believe this research will also give us information on other potential safety measures California may consider implementing, such as a minimum residential distance from high emission structures.