Data 102 Final Project

Nikil Sunku, Akshay Patel, Natalia Ramirez

Spring 2023

1 Data Overview

Our project utilized datasets from the American Census Survey, the EPA, and California's Open Data Portal. We did not use any datasets provided in the project spec.

Note: Our causal inference question relies on asthma hospitalizations in 2015, and the CDC did not provide census-tract level PM2.5 estimates for 2015 so we could not use the provided dataset. Described below in *PM2.5 Daily Averages*, we found the EPA website that is likely where the CDC gets its data from, and downloaded the 2015 data for our analysis.

1.1 PM2.5 Daily Averages

PM2.5 is a fine inhalable pollutant, which the EPA classifies as generally 2.5 micrometers and smaller. The dataset was generated using a Bayesian space-time downscaling fusion model to combine monitoring data from State and Local Air Monitoring Stations, and the Community Multiscale Air Quality model, to create estimates for PM2.5 levels across the contiguous United States. The largest purpose of this model is to provide estimated levels for county tracts, which serves as a common identifier to be merged onto other county-specific data.

Daily averages for the year 2015 were downloaded from the EPA's Fused Air Quality Surface Using Downscaling Files website, under *Output files*. Each row of the data represents a daily average PM2.5 estimate ($\mu g/m^3$) at a specific coordinate in the contiguous U.S. on a specific day. This will impact our findings, as each coordinate point is not necessarily an actual sensor location, but a virtual point generated through a model we do not have control over. Therefore, there will be areas in California where there may be more geographic representation of a county, and areas where there is less. When constructing a model that might explain geographical variation of PM2.5 readings across California, this can alter our results and bias the significance of regions.

Much of these concerns arise from possible sampling bias and convenience sam-

pling. The State and Local Monitoring Stations are unlikely to be perfectly uniformly distributed in California. Some ares are harder to reach and are simply less important when funding is limited. Therefore, more rural counties and/or areas where the terrain is difficult to traverse may have less coordinate PM2.5 daily average estimates. Some areas of California have also been inhabited for longer periods of time, and by more people. This can affect how long sensors have been collecting data, and how much time scientists have had to build up a dense network of sensors in the region. Measurement error is a concern when estimating values, and the dataset contains a column of the standard error ($\mu g/m^3$) for each estimate. However, we do not believe this will play an important factor in our research, since the predicted PM2.5 values can still be modeled for geographical trends with an appropriate level of certainty.

To understand the how PM2.5 levels vary geographically in California, latitude and longitude are not the only features that contain valuable information. We wish that with every coordinate, the dataset also contained features related to climate. Although we added climate data onto our final dataset, its granularity is at a county level and was only temperature. If we had variables like wind speed, humidity, rainfall, and temperature at a finer level, we could answer how PM2.5 changes throughout all space, instead of being limited to county-by-county information.

There were no missing values in the dataset. We cleaned the data to only contain coordinate points inside California, and have coordinate points mapped to counties. We did this to use the PM2.5 readings alongside county-level climate data, and county-level hospitalization data.

Furthermore, we cleaned the data to only contain information from January. We noticed after our EDA that the average PM2.5 level varied from month to month, and geographic trends shift from seasons. Since month is a confounder, we decided to fix it to January.

1.2 County Geospatial Data

To perform both of our research questions, we needed to group data into California counties. To do this, we used California Geographic Boundaries from the state's Open Data Portal. The Shapefile of county boundaries was generated using the US Census Bureau's 2016 MAF/TIGER database. The TIGER/Line Shapefile contains the current geography of California and various variables such as the county federal processing code (which is standardized through many government datasets), the area land in squared meters, and the area of water in squared meters. Each row of data is a unique county with identifying features and characteristics, with some named above.

Convenience sampling and sampling bias are not relevant for this dataset, but measurement error can affect the boundaries of a county and the areas of land and water. This could affect our modeling and causal inference approaches in our research questions, since some variables such as climate and hospitalizations

are by county. PM2.5 estimates incorrectly mapped into one county instead of another may result in error, albeit likely small.

The columns *CSAFP*, *CBSAFP*, and *METDIVFP* have missing values. The *CSAFP* is the Census Bureau's Combined Statistical Area (CSA) Federal Information Processing Standard (FIPS) code. A CSA is a geographic region that consists of two or more adjacent metropolitan statistical areas that have significant economical and social connections. The *CBSAFP* code is the Census Bureau's Core Based Statistical Area (CBSA) FIPS code. It is a geographic region that consists of one or more counties that have strong economic and social ties based on commuting patterns. Lastly, the *METDIVFP* stands for Metropolitan Division (METDIV) FIPS code. The Census Bureau assigns smaller sub-areas within Metropolitan Statistical Areas that have distinct demographic and economic characteristics a code. We found this information in the official TIGER documentation. We do not believe that these values missing have any relevance to our research questions or modeling approach, since we did not use these columns.

We would have liked to have more information per county, optimally related to the geographical location. This logic is very similar to what we wanted in the *PM2.5 Daily Averages* dataset. In addition, however, it would have been useful to have more data about each county's demographic composition to better help us draw relationships between PM2.5 and county for the causal inference research question. This includes features relating to age, income, ethnicity, and labor force. These features would also allow us to answer whether geographical variation is a significant factor to PM2.5 versus demographic covariates.

We did not clean this data. We merged it with *PM2.5 Daily Averages* (mapping values to counties).

1.3 Asthma Hospitalizations

Asthma hospitalization rates by county (per 10,000 residents) in California were derived from the Department of Health Care Access and Information Patient Discharge Data. This data is generated through collecting the total number of asthma hospitalizations from all licensed hospitals in California, and can therefore be reasonably treated as a census. This data was downloaded from California Open Data Portal.

Under the assumption that all licensed California hospitals represent the population of asthma hospitalizations, our dataset is not a sample. However, we decided to compare the 2015 year's data with a composite average across 2015 to current to analyze the differences of our selected study year.

The dataset is stratified by group. The age is stratified as follows: all ages, 0-17, 18+, 0-4, 5-17, 18-64, 65+. The race/ethnicity is stratified as follows: white, black, Hispanic, Asian/Pacific Islander, American Indian/Alaskan Native. We did not determine that any groups were systematically excluded from

the data based on age and race/ethnicity. However, the portal does specify that "these data are based only on primary discharge diagnosis codes". This may imply that a sub-population of hospitalizations that did not have a primary discharge diagnosis code has been excluded, assuming such a population exists. Secondly, we also rest our statement that this dataset is a census on the assumption that licensed hospitals in California are overwhelmingly or entirely the case. It would be unsurprising to discover asthma hospitalizations of underprivileged and undocumented immigrants to go unnoticed or treated in unlicensed hospitals, which may cause these groups to be excluded.

We cannot say whether the patients were or were not aware that their data would be shared in this capacity. The granularity of the data is rather large, with each row corresponding to a county, year, stratification, and the corresponding number of hospitalizations and age-adjusted hospitalization rate. There is also a comment which specifies that a rate is missing due to statistical instability or too small of data.

Convenience sampling does not play a role in this hospitalization dataset, since the state collects data from all licensed hospitals. Measurement error may play a role, but based on the simplicity of the recorded data, we assume it is small. Lastly, sampling bias has the potential to play the largest role, since some patients may have a higher likelihood to be recorded than others. As mentioned earlier, Californians who do not go to licensed hospitals for asthma hospitalizations will go unrecorded.

There are several steps that California has taken to ensure differential privacy. Most importantly, the data is not the count of unique individuals who were hospitalized, but rather the number of total hospitalizations. This preserves the privacy of a unique patient. The data is also per county - if the granularity was finer, it may pose a risk to a patient's anonymity. Lastly, the state redacted hospitalization rates and counts if the value was too small, and left a comment that said so in the row. This is common across public health to ensure differential privacy.

We do not believe that we wanted other columns in this dataset, as we searched for it precisely for hospitalization rates. There are many missing entries in the comments column, but that means that there was no issues in that row's values, so it does not affect our analysis. In fact, the comment existing is more problematic since it meant a rate or count was null, but there were few comments and the reduction in rows did not heavily affect our dataset.

Before merging the asthma dataset onto our final DataFrame, we removed the comment column and dropped all rates and/or counts that were null. Then, we selected the year 2015 and chose the total population, age-adjusted hospitalization rate (AAHR) for each county. The decision to use the total population, AAHR instead of stratified statistics will impact our casual inference study, since some ages and races may have a stronger or weaker causal relationship between PM2.5 and asthma hospitalizations when they are analyzed indepen-

dently.

1.4 Climate Data

We accessed California county-level data from the NOAA National Centers for Environmental information on average temperature over 2015-2019. Specifically, we obtained the statewide time-series data for average temperature per county. We used this dataset to explore the potential counding effect that temperature could have on PM2.5 levels.

One drawback is that this data presents the average temperature over a five-year period, which may not adequately represent the temperature fluctuations across seasons, months, or years. This could be limiting when determining the correlation between temperature and PM2.5 levels, since PM2.5 can also fluctuate considerably on a daily, weekly, and seasonal basis. That being said, average temperature per county over a 5 year period can be indicative of general climate and biome conditions of that county, which could certainly help predict PM2.5. Therefore, we determined this feature to be an adequate potential counfounder for our model.

1.5 Census Data

We obtained California county-level data for several of our relevant confounders from the American Community Survey (ACS), which is conducted annually by the United States Census Bureau. We accessed the ACS data through the Census Bureau's website (https://data.census.gov). We combined the ACS data with the PM2.5 data to explore the relationships between several features and PM2.5 levels in California.

The datasets we obtained from this source included:

- DP05 *ACS DEMOGRAPHIC AND HOUSING ESTIMATE*S: This dataset contains 2019 demographic data such as population, age, race, ethnicity, education, income, housing per county in California.
- DP03 *SELECTED ECONOMIC CHARACTERISTICS*: This dataset contains 2019 data including employment status, industry, occupation, insurance status, and class of worker per county in California.
- S1901 *INCOME IN THE PAST 12 MONTHS (IN 2021 INFLATION-ADJUSTED DOLLARS)*: This dataset contains 2019 data on income distribution per county in California.

One potential weakness of the ACS datasets is that it is subject to sampling error, since it is based on a sample of the population. Additionally, there may be some reporting bias since the data is self reported. Some individuals may make errors in their reports, or may deny participating. Additionally, most of these datasets are missing some of the smaller counties, which may introduce bias.

2 Research Questions

With our completely combined dataset, we aim to answer the following questions:

2.1 Does PM2.5 vary geographically in California?

By understanding what geographical features are most important in determining PM2.5 levels in California, policymakers can approach reducing PM2.5 pollutants in an efficient manner. For example, if accurately modeling PM2.5 reveals that the area of water in a county is negatively correlated to PM2.5 in a significant way, policymakers may choose to target drier counties first, or add bodies of water near pollution hotspots.

Modeling using Frequentist and Bayesian GLM's and a non-parametric model is a good method because coefficient estimation will allow us to determine the significance of a feature and its correlation with Pm2.5 with a certain level of uncertainty. Ultimately, we are trying to uncover the strength of a relationship, not whether it is causal or correlated.

GLM's have several limitations. Firstly, they assume that the relationship between PM2.5 and the predictor features can be modeled using a specific probability family (we will be using Normal). If our assumption is wrong, our model can go wrong. Next, the relationship between PM2.5 and the features has to be linear. If this is not true, then our model will not predict well. Lastly, GLM's are sensitive to outliers and multicollinearity. If, for example, latitude and longitude are highly collinear, their coefficients may be inaccurate and our interpretations can be wrong.

2.2 Does increased PM2.5 levels cause higher county asthma hospitalization rates?

PM2.5 pollutants are more dangerous than PM10 pollutants because they are finer and can get into deeper areas of the lungs, and even the bloodstream [1]. Long-term exposure to PM2.5 can be attributed to health effects not limited to ischemic heart disease, lung cancer, strokes, and type 2 diabetes [2].

It is important to determine whether PM2.5 pollution is the cause of diseases and hospitalizations, as evidence of damage can propel decisions to curtail the danger. The prevalence of PM2.5 in society is reminiscent to common smoking in urban areas throughout the 20th century. A causal study by Hammond and Horn in 1952 was a landmark discovery that found men who smoked regularly had a considerably higher death rate than men who had never smoked

[1] https://www.cdc.gov/air/particulate$_matter.html$
[2] https://www.stateofglobalair.org/health/pm

Figure 1: This image shows how average PM2.5 level changes by month from 2011-2014, with a bar to represent the standard deviation of each average

Figure 2: This image shows how average PM2.5 level varies by month throughout 2015, with a bar to represent the standard deviation of each average

(or smoked with cigars or pipes only) [3]. Hammond and Horn's work began a massive shift in American opinions on tobacco and cigarettes. It is possible that a breakthrough study establishing that PM2.5 levels cause asthma hospitalizations can be part of a shift too. Determining this causal relationship at a county level is not optimal for individual-level analysis, but it is good for policymakers. For example, the state of California can take our study and then start directing health funding to certain counties more than others.

We are using causal inference to answer whether increases PM2.5 levels causes higher county asthma hospitalization rates, since we want to uncover a causal relationship between PM2.5 levels and asthma hospitalizations. Depending on this relaitonship, we may be able to see the overall implications PM2.5 level has on public health.

We will be using inverse propensity scoring to manage the difference in covariates across counties and develop an unbiased average treatment effect.

IPW assumes unconfoundedness to generate an unbiased ATE. If we do not find all confounders, then our causal study may be biased. Another limitation is that IPW can be sensitive to high weights. If some counties have very high or low propensity scores, our causal relationship might be unstable. Lastly, IPW relies on accurate propensity scores, which means our model ahs to be good. We chose to use a logistic regression like in Homework 4, but that may not be performing well and we could get biased IPW's.

3 EDA

This visualization shows how average PM2.5 levels vary across each month. In this example, we are visualizing the Catgorical variable of month, and comparing it to the quantitative variable Average PM2.5 Level across 4 years. It illustrates how year is a confounding variable, and how the month within each year is a confounder as well. This deviation prompted us to include temperature as a factor, as winter months being colder may influence the average PM2.5 level.

This visualization helps show the difference between 2015 and the previous 4

[3] https://www.cancer.org/research/acs-research-news/the-study-that-helped-spur-the-us-stop-smoking-movement.html

Latitude vs. PM2.5:

Quantitative: *Average PM2.5 Level, Latitude*

Qualitative: *Time Period*

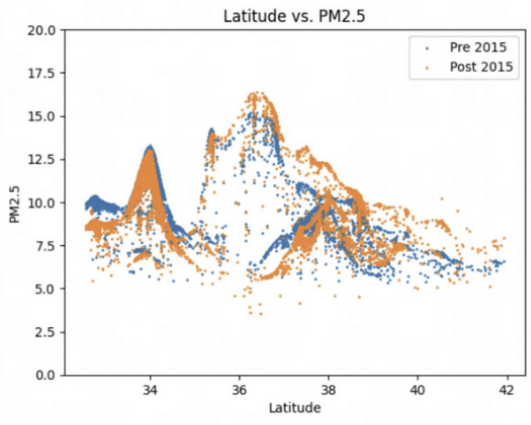

Figure 3: This image compares the relationship between Latitude and PM2.5 Pre and Post 2015

years. In this case, the categorical variable is once again month except we are not comparing across 4 years. With the quantitative variable being Average PM2.5 level in 2015. The reason for this visualization was to see how month's influence on average PM2.5 level changes depending on the year. Relative to the previous 4 years, the trends look pretty similar, however December seems low. January however seems the most consistent between all of the years, so we decided to stick with January for most of our analysis.

These 2 plots compares PM2.5 levels vs. Latitude and Longitude between two time periods, pre-2015 and post-2015. They are both qualitative via Time Period, and quantitative via Average Pm2.5 level, and Latitude/Longitude respectively. The original dataset from the project guidelines did not include hospitalization rates for asthma per county. We discovered this data for 2015 onwards, so we decided to search for the same 2011-2014 data, but in 2015. We performed this comparison to ensure that the data we found is from the same repository, since it is in an obscure location. For the pre 2015 data points, at higher points of latitude, PM2.5 is lowest. Around latitude level of 36 is when PM2.5 tends to be highest, peaking at around 15. The post 2015 points do seem to follow a similar trend as the pre 2015 group, with deviations that we assign to variation from unexplained confounders. Additionally, we only compare one year (2015) with 2011-2014, which also explains the differences alongside general variation that may have occured. For longitude both 2015 and pre-2015 have peaks around -119 and follow a relatively similar distribution. Although

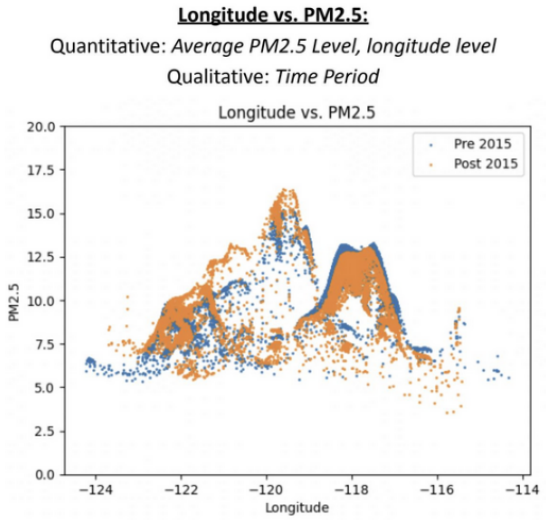

Figure 4: This image compares the relationship between Longitude and PM2.5 Pre and Post 2015

clear distinctions exist for both the latitude and the longitude graph, the trends are similar enough for us to justify using the 2015 hospitalization rates in our analysis.

This information can help us identify appropriate confounding variables and select statistical models and techniques to control for these confounding factors, ultimately leading to more accurate and reliable data analysis results. Coastal regions may have different topography, meteorology, and sources of air pollution compared to inland regions, and these variations could potentially affect PM2.5 levels. Therefore, measuring PM2.5 levels across different distances from the coast could provide valuable insights into the factors that contribute to high levels of PM2.5, and help us better understand the relationship between PM2.5 levels and geographical or meteorological factors.

4 Research Question 1: Do PM2.5 Levels Vary Geographically?

4.1 Methods

We are predicting average daily PM2.5 levels, and determining whether PM2.5 levels have a geographical trend.

Our features are:

- Latitude

- Longitude
- Percent of total California population
- Population density
- Average temperature between 2015 and 2019
- Anomaly in temperature (1901 - 2000)
- Average temperature between 1901 and 2000
- Area that is land (in meters squared)
- Area that is water (in meters squared)
- Percent of people within a county who make less than 10,000 U.S. dollars
- Percent of people within a county who make more than 200,000 U.S. dollars
- Mean income (U.S. dollars)
- Percent population under 5 years old
- Percent population white

To initially choose features, we used our EDA, domain knowledge, and research. We needed to include a variety of geographical features to ensure we can draw conclusions to answer our research question. Therefore, we included *latitude, longitude*, climate features such as the average temperature historically and recently, and the area of the county that is land and water: *ALAND* and *AWATER*. We feature engineered *population density* to increase the significance between larger, rural counties and larger, urban counties.

We further refined our feature selection using a correlation matrix, alongside trial and error with our random forest model and the statistical significance given by our GLM models. To improve accuracy, we decided to include demographic features such as *mean income, percent population white* and *percent population who make less than 10,000 dollars*.

4.1.1 Frequentist GLM

We used a Frequentist GLM with the identity function as the inverse link function and a Gaussian distribution family. We made this choice because after plotting our outcome variable, *daily average PM2.5 levels*, the distribution is relatively normal.

Based on our choice, we are assuming several things about our outcome variable.

1. PM2.5 average daily levels are continuous and normally distributed.
2. The variance of PM2.5 average daily levels is constant.

3. The relationship between the predictor variables and the outcome variable is linear.
4. The errors of our model are I.I.D.

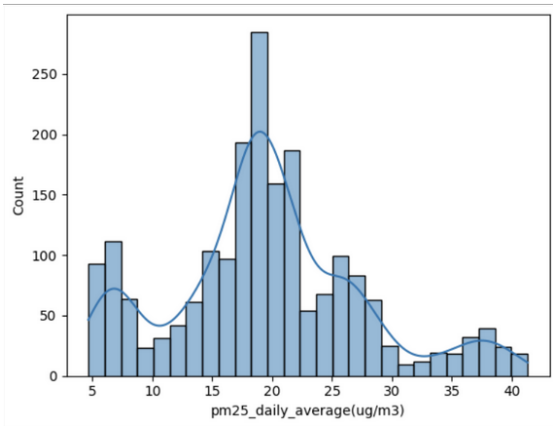

Figure 5: Distribution PM25 Daily Average in January 2015

4.1.2 Bayesian GLM

Since we are based our approach for the Frequentist GLM on our assertion that the daily average PM2.5 levels distribution is approximately normal, we carried that thinking over to the Bayesian GLM. Therefore, we decided to use a normal prior and an identity inverse link function.

For each parameter, we decided to use a normal prior with a mean of 0 and a standard deviation of 10. This normal prior is extremely loose. After experimentation with various standard deviations (1, 5, 10), we determined that 10 was best. This is beneficial for us, since we are not confident of the true parameter values for our features. After analyzing distributions of each feature, it was difficult to determine distribution hyper-parameters that would effectively capture the variance of the respective feature. Therefore, we decided to avoid as much under-fitting as possible.

The assumptions we are making in the Bayesian GLM is similar to the Frequentist GLM, since our prior for the outcome variable and the inverse link is the same. An additional assumption is:

1. We are using a HalfCauchy prior for sigma. We are assuming that the noise of PM2.5 daily averages follows a heavy-tailed distribution. Since pollution can skyrocket or dip heavily due to factors like storms and fires, we opted to allow for the possibility of extreme outliers, which a HalfCauchy is perfect for.

4.1.3 Random Forest

We used a random forest as a non-parametric approach for several reasons. The random forest allows us to identify the most important predictors in the model. Since our goal is to identify the importance of geography in the variation of PM2.5, it is necessary to order the importance of features. The random forest is also capable under high-dimensional data and non-linear relationships, both of which exist in our data. Geospatial relationships are often not linear, and it can be seen in our visualizations that the distribution of the observed variable and the predictor variables are often multi-modal.

Assumptions we have to keep in mind include the fact that some of our features have a level of correlation (ie. longitude and latitude) so we have to add both of them as inputs.

Multicollinearity is another thing that exists within our data however it is not much of an issue since we are doing a random forest model

4.1.4 Evaluating Performance

To evaluate the Frequentist GLM's performance, we will be analyzing the Root Mean Squared Error (RMSE) of the trained model on a test set, the confidence intervals of each coefficient, their magnitudes, their standard errors, and their 95 percent confidence intervals.

To evaluate the Bayesian GLM's performance, we will generate posterior predictive samples and calculate the RMSE of the posterior predictive samples. Then, we will plot the posterior predictive samples, the mean posterior predictive distribution, and the observed data together to visually analyze the fitness of our model. Finally, we will analyze the credible intervals of each coefficient, each coefficient's magnitude, and their similarity to the Frequentist GLM.

To evaluate the Random Forest's performance, we will use a combination of RMSE, R-squared, and feature importance.

4.2 Results

4.2.1 Frequentist GLM

Our model had an test RMSE of 2.816, which means our model was on average 2.816 $\mu g/m^3$ incorrect. When analyzing specific geographical features, *Latitude* had a coefficient of -2.9371, *Longitude* had a coefficient of -12.850, *AWATER* had a coefficient of 2.305, and *ALAND* had a coefficient of 6.236. Therefore, based on our model, *Latitude* and *Longitude* had a negative correlation with PM2.5, while *AWATER* and *ALAND* had a positive correlation with PM2.5. *Population density* was also a relevant geographical variable, and had a positive association with PM2.5.

Average Temperature 2015-2019 and *1901-2000 Mean* had the largest impact

Features	Value	P ¿ |z|	[0.025	0.975]
const	19.6244	0.000	19.499	19.750
Latitude	-2.9371	0.003	-4.903	-0.971
Longitude	-12.8495	0.000	-13.814	-11.885
percent_pop	-82.4761	0.000	-94.705	-70.247
pop_density_m2	16.5313	0.000	13.657	19.405
Average Temperature 2015-2019	337.2966	0.000	254.692	419.902
Anomaly (1901-2000 base period)	-50.1321	0.000	-60.651	-39.613
1901-2000 Mean	-310.7139	0.000	-386.961	-234.467
ALAND	6.2363	0.000	4.859	7.614
AWATER	2.3049	0.001	0.968	3.642
Less than $10,000	-0.4406	0.016	-0.798	-0.083
$200,000 or more	8.0693	0.000	3.963	12.176
Mean income (dollars)	-7.2380	0.000	-11.049	-3.427
Under 5 years	53.2397	0.000	42.100	64.379
White	22.8273	0.000	19.159	26.496
Worked from home	0.2569	0.411	-0.355	0.869
With Social Security	-3.9724	0.000	-4.931	-3.014
Unemployment Rate	-2.9003	0.000	-3.766	-2.034
Not in labor force	2.9404	0.001	1.179	4.702
Agriculture, forestry, etc.	6.6305	0.000	5.816	7.445
Manufacturing	0.1783	0.705	-0.744	1.100

Table 1: Figure 5: Frequentist GLM Results.

on predictions, although the former had a large positive correlation, while the latter had a large negative correlation.

For each feature, we are 95% confident that the true population parameter lies within the feature's 0.025 and 0.975 columns in Table 1. Most features do not have 0 within the 0.025 and 0.975 columns, which means we are 95% confident that the true population parameter has a negative or positive correlation with PM2.5. Exceptions are *Worked from home* and *Manufacturing*, which means we cannot say with 95% confidence whether the true population parameter is positive or negative.

When looking at the P ¿ |z| column, it can be seen that all values are less than 0.05 except *Worked from home* and *Manufacturing*. This means that there is not enough evidence to suggest that the features have a statistically significant effect on PM2.5 at 5% significance .

4.2.2 Bayesian GLM

Our model had an test RMSE of 2.880. When analyzing specific geographical features, *Latitude* had a coefficient of -6.911, *Longitude* had a coefficient of -13.142, *Average Temperature 2015-2019* had a coefficient of 15.964, *1901-2000*

Features	**mean**	**hdi_3%**	**hdi_97%**
Constant	19.623	19.495	19.750
Latitude	-6.911	-8.636	-5.264
Longitude	-13.142	-14.064	-12.202
percent_pop	-84.680	-92.676	-77.376
pop_density_m2	16.379	13.935	18.747
Average Temperature 2015-2019	15.964	2.059	29.041
Anomaly (1901-2000 base period)	-6.454	-8.747	-4.348
1901-2000 Mean	-15.101	-26.858	-1.971
ALAND	4.720	3.382	5.976
AWATER	0.584	-0.536	1.774
Less than $10,000	-1.009	-1.311	-0.708
$200,000 or more	0.629	-2.989	4.331
Mean income (dollars)	0.566	-2.749	3.903
Under 5 years	55.436	47.436	62.349
White	22.930	19.641	26.432
Worked from home	0.116	-0.469	0.641
With Social Security	-4.012	-4.906	-3.116
Unemployment Rate	-1.367	-2.178	-0.620
Not in labor force	5.493	4.245	6.800
Agriculture, forestry, etc.	3.890	3.217	4.529
Manufacturing	0.846	0.102	1.589
Sigma	2.947	2.860	3.037

Table 2: Figure 6: Bayesian GLM Results.

mean had a coefficient of -15.101, *AWATER* had a coefficient of 0.584, and *ALAND* had a coefficient of 4.720. Therefore, based on our model, *Latitude*, *Longitude* and *1901-2000 mean* had a negative correlation with PM2.5, while *AWATER*,*ALAND* and *Average Temperature 2015-2019* had a positive correlation with PM2.5. *Population density* was also a relevant geographical variable, and had a positive association with PM2.5.

Percent population and *Under 5 years* had the largest impact on predictions, although the former had a large negative correlation, while the latter had a large positive correlation.

For each feature, there is a 94% probability that the true value of our parameters lies within the hdi_3% and hdi_97% columns of Table 2. Most features do not have 0 within the hdi_3% and hdi_97% columns, which means that there is a 94% probability that the true population parameter has a negative or positive correlation with PM2.5. Exceptions are *Worked from home*, *mean income (dollars)*, *$200,000 or more*, and *AWATER*, which means we cannot say that there is a 94% probability that the true value of the parameters are positive or negative.

Features	Importance
Latitude	0.134
Longitude	0.131
percent_pop	0.004
pop_density_m2	0.003
Average Temperature 2015-2019	0.014
Anomaly (1901-2000 base period)	0.001
1901-2000 Mean	0.005
ALAND	0.191
AWATER	0.427
Less than $10,000	0.004
$200,000 or more	0.003
Mean income (dollars)	0.038
Under 5 years	0.010
White	0.000
Worked from home	0.001
With Social Security	0.004
Unemployment Rate	0.003
Not in labor force	0.007
Agriculture, forestry, etc.	0.004
Manufacturing	0.017

Table 3: Feature importance for Random Forest model.

4.2.3 Random Forest

The random forest model had a train RMSE of 0.313 and a test RMSE of 0.961. When analyzing specific geographical features seen in Table 3, *Latitude* had an importance of 13.4%, *Longitude* had an importance of 13.1%, *Average Temperature 2015-2019* had an importance of 1.4%, *1901-2000 mean* had an importance of 0.5%, *AWATER* had an importance of 42.7%, and *ALAND* had an importance of 19.1%. Therefore, based on our model, *AWATER* was the most important geographical feature to determining PM2.5 daily averages across California. *White*, the proportion of people who are White in a county, had an importance of 0%. Temperature *Anomaly (1901-2000 base period)* had an importance of 0.1%.

4.3 Discussion

Our Frequentist GLM fits our data relatively well when we analyze several metrics froms the statsmodels library, shown in Table 4. Considering the high variance of PM2.5 across California and how easily it is impacted by other variables, we aimed to capture the most variance of the data. When looking at the Pseudo R-squ. value, our model explains almost all the variance of the data, roughly 99.9%. The deviance is high at 16479, but the log-likeihood divided by Df residuals is roughly -0.5, which is close to 0. Also, the Peasrson chi2 value

Performance	
No. Observations:	2012
Df Residuals:	1991
Df Model:	20
Scale:	8.2765
Log-Likelihood:	-4970.4
Deviance:	16479.
Pearson chi2:	1.65e+04
Pseudo R-squ. (CS):	0.9989

Table 4: Frequentist GLM performance.

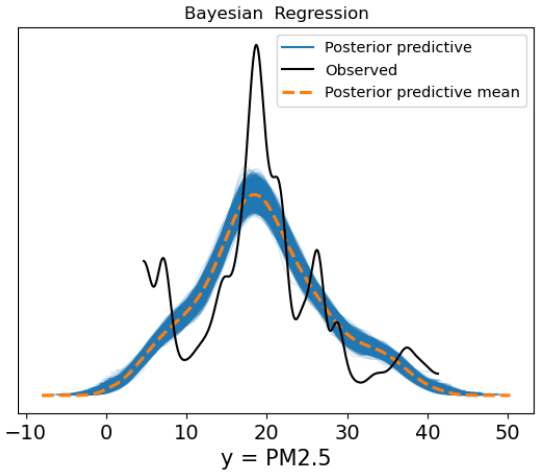

Figure 6: Average posterior predictive distribution over observed data.

is close to the Df residuals which means the model predictions are close to the actual data.

The random forest model fit the data extremely well. Not only was the RMSE small, but the R^2 was 0.986. This means that the random forest captured 98.6% of the patterns in the data and is additional evidence that the model is a good fit.

Visualizing our Bayesian GLM shows that our model does not fit our data well, but is good for inference. The RMSE was 2.880, but when looking at Figure 10, our model's posterior predictive mean does not fit the observed data well. No posterior predictives reached the peak of the observed data, and the distribution only barely captured the waves in the observed data.

However, when analyzing the summary output after the Bayesian GLM ran as shown in Table 5, all ESS Tails were above 1000, and all R-hats were 1. The

Parameter	ESS Tail	R-hat
β_0	2508.0	1.0
β_1	3006.0	1.0
β_3	2221.0	1.0
β_5	2400.0	1.0
β_6	2513.0	1.0
β_7	2678.0	1.0
β_8	2751.0	1.0
β_9	2721.0	1.0
β_{10}	2439.0	1.0
β_{11}	2366.0	1.0
β_{12}	2144.0	1.0
β_{13}	2278.0	1.0
β_{long}	3134.0	1.0
σ	2888.0	1.0
β_2	1656.0	1.0
β_4	2138.0	1.0

Table 5: ESS Tail and R-hat Values

ESS Tail values mean that we have effective samples, and our R-hats mean that all the chains have converged.

This leads us into the comparison between the Frequentist and Bayesian GLM's. Although we assert that the Frequentist GLM fits the data better, both RMSE's are similar: 2.816 and 2.880. We observed that the Bayesian GLM assigned higher importance to features that the Frequentist GLM did not. For example, the Frequentist GLM had a coefficient value of -2.937 for *Latitude*, while the Bayesian GLM had the value -6.911. In the Frequentist GLM, the highest magnitude for a coefficient was 337.297 for *Average Temperature 2015-2019* while it was 15.964 for Bayesian. Overall, most coefficient values were a smaller magnitude for the Bayesian GLM.

After comparing each model, the random forest model performed better because the RMSE was smallest and the data was fit well. We are confident in applying the random forest model in future datasets, because the RMSE of the test data was higher than the training data, but not by a large amount. This makes us believe that although a random forest could be overfitting the data when compared to GLM's, the test RMSE is still likely to be small enough for it to be useful for stakeholders. In other words, we are confident that the random forest is balancing variance and bias.

5 Research Question 2: Does increased PM2.5 levels cause increased asthma hospitalizations?

5.1 Methods

Our treatment variable is:

- High or low average annual daily PM2.5 mean levels in 2015, decided based on median value across California counties.

Our outcome variable is:

- Age-Adjusted Hospitalization Rate (AAHR) in 2015

Confounders are (granularity of county):

- Average temperature from 2015-2019
- Anomaly in temperature (1901-2000 base period)
- Average temperature from 1901-2000
- Population density, in squared meters
- Proportion of males
- Proportion of females
- Proportion of age:
  - Under 5 years old
  - From 5 to 9 years old
  - From 10 to 14 years old
  - From 15 to 19 years old
  - From 20 to 24 years old
  - From 25 to 34 years old
  - From 35 to 44 years old
  - From 45 to 54 years old
  - From 55 to 19 years old
  - From 60 to 64 years old
  - From 65 to 74 years old
  - From 75 to 84 years old
  - Over 85 years old
- Proportion of races:

  - Proportion White
  - Proportion Black or African American
  - Proportion American Indian and Alaska Native
  - Proportion Asian
  - Proportion Hispanic or Latino
- Proportion of people with income:
  - Less than $10,000
  - $10,000 to $14,999
  - $15,000 to $24,999
  - $25,000 to $34,999
  - $35,000 to $49,999
  - $50,000 to $74,999
  - $75,000 to $99,999
  - $100,000 to $149,999
  - $150,000 to $199,999
  - $200,000 or more
- Median income
- Household income in the past 12 months
- Proportion of people working in each industry:
  - Manufacturing
  - Agriculture, forestry, fishing and hunting, and mining
  - Not in Labor Force
- Proportion Working from home
- Proportion with Social Security
- Unemployment Rate

Our causal DAG is Figure 6. We had trouble making the images go where we want them to go, so they may above or below the relevant section.

There are no colliders in our dataset. To adjust for our confounders, we will use inverse propensity weighting. To determine the propensity scores, we will use logistic regression.

We believe the unconfoundedness assumption holds, since we were exhaustive in identifying confounders at a county level. We included health, demographic,

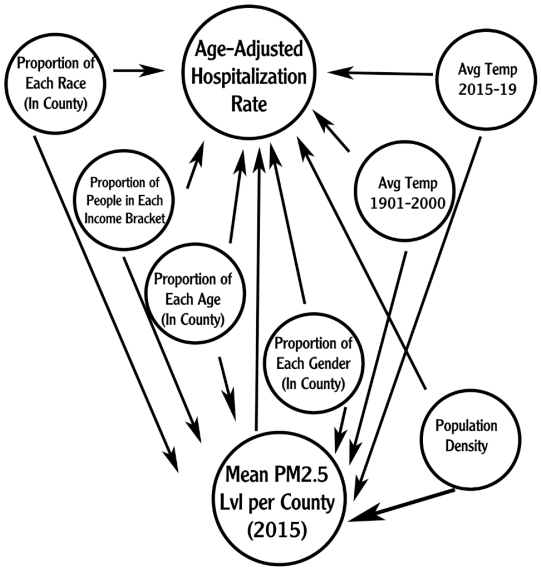

Figure 7: DAG Diagram

geographical, and climate data to account for all sorts of confounders. We did research to identify common factors of asthma, which pointed us towards some of the variables we included. For example, asthma can be exacerbated by hot days, and different age groups are more likely to be hospitalized.

5.2 Results

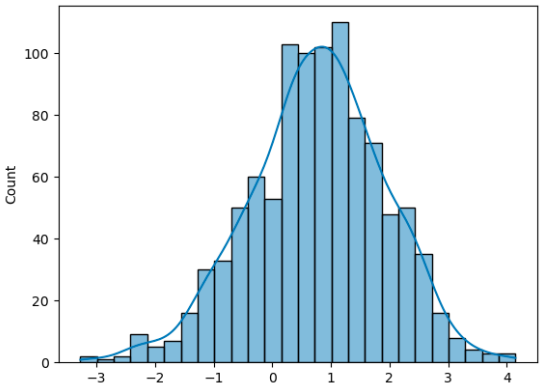

Figure 8: 95% Confidence Interval of Causal Study

We found inconclusive evidence that increased PM2.5 daily average levels in a county causes higher asthma hospitalization rates. Assuming unconfoundedness, we are 95% confident that the true average treatment effect lies within [-1.58, 3.09] which includes positive and negative values and the null value 0 (seen in Figure 7). Therefore, although the average ATE is 0.76, it could have been positive, negative, or nothing. In terms of magnitude, the confidence interval implies that the age-adjusted hospitalization rate per 10,000 people could decrease at a smaller magnitude than it could increase.

5.3 Discussion

The biggest limitation we faced is that we could not run a randomly-controlled trial experiment. Therefore, we had to attempt to account for every confounder to assume unconfoundedness and perform inverse-propensity weighting for our confounders. During analysis, we compared the standardized mean differences of each confounder, and determined that many of them were above the conventional 0.1 limit. Most hovered around 0.3 to 0.4, which was much lower than after inverse weighting but still above 0.1. This means that the counties that are receiving the treatment or not may be unbalanced.

Additionally, by looking at this causal relationship at a county level, the granularity fails to account for a lot more localized changes in PM2.5. Census data was per county, but we had PM2.5 daily averages for coordinate points. Therefore, in most counties we lost detail in PM2.5 readings for the sake of county-level information. County-level socioeconomic factors also ignores neighborhood-level changes which can affect hospitalizations.

As such, additional data relating counties to each other in terms of socioeconomical factor would have helped us identify confounders more accurately. For example, the Census Bureau identifies metropolitan areas that are similar to adjacent ones, or different by some capacity. Something like this for income, poverty, health insurance, etc., would have allowed us to develop more features.

As mentioned in the results, we cannot be confident that there is a causal relationship between the treatment and outcome because the null value is in our 95% confidence interval. There is a high magnitude of effect on both sides of the interval,

6 Conclusions

6.1 Summary of Findings

Our first research question asked whether there are geographical trends to PM2.5 levels in California. Our parametric and non-parametric models assigned importance to geographic features such as latitude, longitude, temperature, population density, and the area of water and land when predicting daily average

PM2.5 in a county. Area of water, area of land, and population density had a positive association with PM2.5. Latitude and longitude each had a negative association with PM2.5. Our parametric and non-parametric models performed well and captured most of our data's variance. The random forest regressor assigned the most importance to the area of water and areas of land in the county, and the least importance to the proportion of home workers and citizens who are White.

Our second research question asked whether higher PM2.5 levels in a county caused increased asthma hospitalizations. Our causal inference using inverse propensity weighting produced inconclusive results. We are 95% confident that the true average treatment effect lies within -1.58 to 3.09, which contains the null value.

6.2 Generalizability

Results for geographical trends are generalizable to an extent. Higher temperature and higher population density causing higher PM2.5 daily averages is likely to be applicable anywhere in the country. However, when extrapolating the results for longitude and latitude, we cannot say that they are negatively associated with PM2.5 on a continental or worldwide scale. It is safe to say then that our findings are narrow, since our model was accurate based on California, and our geographical features had coefficients assigned with confidence and 99% variance based on California county PM2.5.

The causal study was inconclusive, but we can make educated assumptions about the generalizability and scope of our findings. Even if the causal study was conclusive, it is difficult to generalize the findings beyond California in the year 2015, since we did not account for the confounders necesssary to do so. The findings would also be narrow in this case, for the same reasoning and because California is not the same as other states in terms of regulation, culture, preferences, and others variables.

6.3 Calls to Action

Being more southeast in the state is negatively correlated to PM2.5, higher population density is positively correlated, and more water is strongly correlated to higher PM2.5. When looking at all these factors together, lawmakers should target areas of the state that fit the geographical descriptions for higher PM2.5, in addition to what we already know about pollution. Resources to address pollution are limited (money, manpower, etc.), and optimal allocation can be based on our models.

When viewing the inconclusive causal study, two options should be considered. If the true ATE value is positive, then public health officials should immediately start advocating for more funding to prevent PM2.5 pollution and/or provide more services for citizens with asthma. If the true ATE value is negative and

high PM2.5 pollution actually causes fewer hospitalizations, authorities should officially begin allocating funding to uncovering the true cause and determining how the behavior of asthma patients changes with increasing pollution.

6.4 Data Merging and Limitations

We merged different data sources to create a data set with county-level information along with daily average PM2.5. A benefit from merging was that we were able to include many confounders and information to model the outcome variable. A consequence was the extreme amount of data cleaning required for us to have a usable DataFrame. Also, it as we merged data, we were forced to drop counties that did not have values across every feature.

As mentioned throughout this report, we used county-level data from the Census Bureau and California Open Data. Both of these repositories ensure differential privacy, and some of our data was masked or removed because of too few samples. These fewer samples would have actually been beneficial for our modeling and causal study, since it indicates a stronger correlation/causation. By having to remove them, we could not account for those values and the variance of our data changed.

6.5 Future Research and What We Learned

Future studies could focus on places where the area of land and water are controlled. Or even apply similar logic to different states. The granularity can change too, potentially from county to state or country level, or smaller to a county tract. Also, researchers could focus on other aspects of asthma like prevalence, deaths, and length of hospitalization. Ultimately, there are many avenues to take for both treatment and outcome.

We learned many different things while doing this project. We became better at using Python libraries like PyMC3 and statsmodels, and learned new domain knowledge on pollution and county-level census data. Area of water being one of our strongest predictors for PM2.5 level is not something we expected to see. Additionally, being able to predict PM2.5 level from the features we had seemed unlikely when we began, but we were able to develop high-accuracy models. Furthermore, it was surprising how the winter months had higher PM2.5 levels than the summer months, however we reasoned that with more artificial heating sources during winter, higher PM2.5 levels likely occurred. Outside data science, we learned how to assemble a report in Latex by using Overleaf, which was really useful knowledge going forward.