Data 102
Adverse Health Reactions

Kyle Atkinson, Hans Ocampo, Jeffrey Qiu, and Osvaldo Valadez

December 13, 2021

Abstract

Recent studies have attempted to prove a link between various environmental factors and adverse health effects. This paper attempts to show how environmental factors impact asthma mortality rates in the United States among women, and show if counties are more likely to experience higher levels of pollution due to having a majority non-white population.

1 Introduction: Data Overview

Our research utilizes the data presented in five distinct data sets, herein referred to by their categorized number.

1. CDC PM 2.5 Concentrations, which provides the modeled predicted mean of PM2.5 concentrations/associated standard deviation in various locations and census tracts. This is census data rather than sampled.
2. CDC Ozone Concentrations, which provides the modeled predicted mean ozone concentrations/associated standard deviations in various locations and census tracts. This is census data rather than sampled.
3. American Community Survey Census Data, of which we've extracted racial categories by county.
4. CDC U.S. Chronic Disease Indicators: Asthma, which provides reports of chronic disease data and state-specific indicator data.
5. Mapping Inequalities, which provides redlining data in the San Francisco Bay Area (Additional/Not used in research questions)

1.1 Data Sources and Download Process

While Datasets 1, 2, and 4 were provided by Data 102 course staff, Dataset 3 was acquired from external sources. Dataset 3 (the ACS Census Data) was collected from the data.census.gov official site. Additional datasets were collected to allow us to further explore how locational data relates to population distributions by race.

1.2 Systematically Excluded Groups

Several states (Alabama, Arkansas, Alaska, California, Colorado, and Connecticut) were not included in the study based on availability of relevant data in Datasets 3 and 4.

1.3 Participant Awareness

Participants of Datasets 1 and 2 were unaware of the collection/use of this data (as this data does not represent human participants). While participants of Datasets 3 and 4 were both aware of the collection/use of this data when recorded (as this is survey/medical data), the participants are not aware of the subsequent research in this report.

1.4 Data Granularity

Datasets 1 and 2's granularity is based on census tract. Dataset 3's granularity is based on county. Dataset 4's granularity is based on state.

1.5 Data Concerns

The datasets utilized in this paper were all census data collected from official government sources. Therefore, issues across selection bias and convenience sampling should not be an issue, as government census data covers virtually the entire U.S. population, rather than samples.

The data from Datasets 1 and 2 was not measured directly (as its based on a predictive model) and was missing data from several locations.

1.6 Unavailable Features

Unavailable data features for this project include:

- PM2.5 and Ozone concentrations by neighborhood. These features would have allowed us to cover redlined neighborhoods, and further allow research into relation/causation between redlining and adverse health effects due to pollution.
- Data from excluded states.
- County-level Asthma data.

2 Research Questions

This paper aims to answer the following research questions:

2.1 Controlling for confounders, does a county being non-white lead to higher levels of pollution?

This question is based on the concept of "redlining", which is the discriminatory practice of denying services (financial and otherwise) to residents of certain areas based on their race or ethnicity. In this case, the "service" would refer to clean air (low pollution).

Conclusions from this research could inspire whether or not policy decisions need to be made to reduce pollution-causing factors in largely non-white areas.

For this question, we used causal inference methodology. We believe causal inference to be the best fit here as we are aiming to establish racial demographics as a leading contributor to factors that lead to pollution.

2.2 How do environmental factors impact asthma prevalence among women in the United States in 2012-2014?

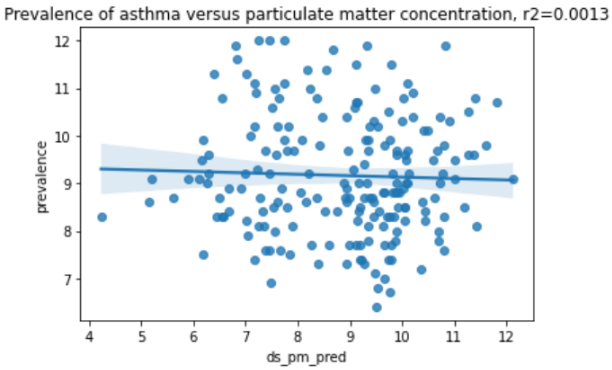
This question was inspired by similar factors to the previous question: whether or not a minority group is disproportionately affected by pollution.

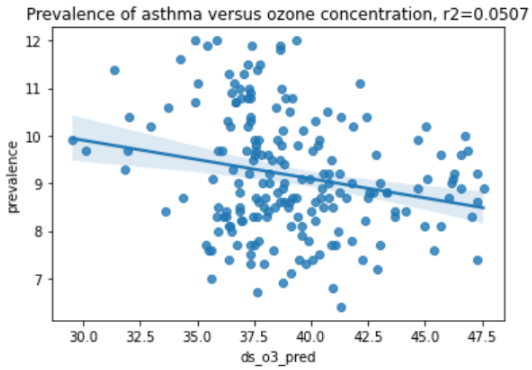
While saying one group is more affected than another likely couldn't motivate to the same types of policy decisions as in Question 1 (as geographical gender distribution is much more uniform than racial distribution), the research could be beneficial to women deciding on which geographic areas they should live (based on level of health concerns).

3 Exploratory Data Analysis Summary

3.1 Asthma and Ozone: Datasets 2 and 4

To clean this data, we looked at the cross section of crude prevalence of asthma among adults 18 and older, expressed as a percentage. This data was grouped by state and year, from 2011-2014, to match up with the pollution data. To remove null values and for consistency across datasets, we decided to look only at the contiguous 48 states plus District of Columbia. Next, we merged the three datasets by first mapping the state names in the asthma dataset (Dataset 4) to FIPS codes, then dropping and reordering the appropriate columns after. Finally, we merged the data sets.

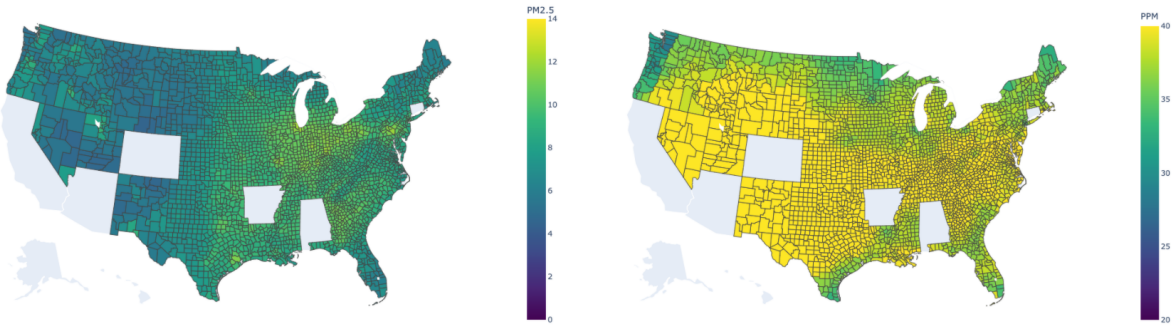




From the graphs, we observe that the link between air pollution and asthma prevalence is not as straightforward as a linear regression: specifically, we see that very little of the variance is captured by the ozone and PM2.5 concentrations (r2 = 0.054). Though we do observe a highly significant link between pollution and asthma prevalence, it is in the negative direction. This is counterintuitive, and something that would benefit from additional exploration. Furthermore, from this preliminary analysis, while ozone appears to have a highly significant relationship with asthma, PM2.5 does not. This result is dubious, and suggests that the current model is insufficient. This graph motivates our future research because we now know that we will likely have to turn to other, potentially non-parametric or GLM models in order to demonstrate this relationship more clearly.

3.2 Pollution Concentrations: Datasets 1 and 2

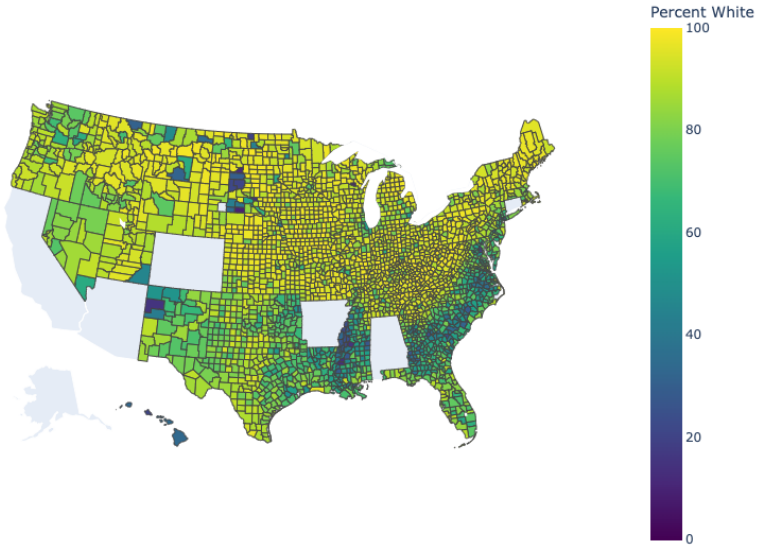
The initial Data in these Datasets was already fairly clean. To aggregate Datasets 1 and 2, we took the mean of each concentration, grouped by county and the year the data represented.



Based on the heatmaps, we notice several states and counties not represented in the datasets, shown in grey. The PM2.5 Concentration heatmap (left) shows that the PM2.5 concentration increases from the West Coast to the East Coast, with PM2.5 at its highest concentration around the Midwest. The Ozone Concentration heatmap (right) has a much clearer trend, as concentrations are lower at the coasts and north/south borders of the United States. Observing both maps simultaneously, there are high relative Ozone and PM2.5 concentrations in the coastal South, which are predominantly made up of non-white populations.

3.3 Majority White vs. Majority Non-White Counties: Dataset 3

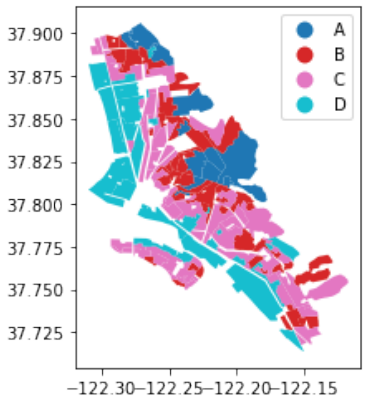
To organize our counties based on racial dispersion, Dataset 3 required a bit of cleaning. Estimating that NaN values were mainly rural counties with very little racial variance, we replaced the NaN values with 0's. Furthermore, as the ACS data includes hundreds of irrelevant features, we eliminated all columns except for FIPS, state, county, population, median income, and population proportions by race.



Similarly to the previous example, we are missing the same states. The heatmap used here shows most middle-America counties have a majority white population, and major population centers have a majority non-white population. As more populated areas tend to be more industrialized/polluted, this may suggest a link between non-white areas and pollution.

3.4 Redlining: Dataset 5 (Additional Data)

To further expand on the ideas of redlining defined in Section 2, we explored redlining in the Bay Area to motivate our research questions. In terms of data cleaning, we restricted the data points to those in the Berkeley/Oakland area. The reasoning for this was visibility. Having a larger data set would have made the redlining separations harder to see for a larger area, so because of that we decided to restrict it to this area. Similarly to Section 3.2, we start with a very clean dataset.



From the visualizations above, we can see that the least desired and "hazardous areas" were very much those closer to the bay to the west, while the more desired areas were those east towards the Berkeley/Oakland Hills

As redlining had major impacts in the housing market, non-white/poorer populations were restricted to undesirable/hazardous areas. We believe that looking at redlining would be useful so we can examine if those areas that we seen as undesirable are now areas that are most affected by air pollution and whose population experiences more adverse health effects

4 Controlling for confounders, is a county being nonwhite lead to higher levels of pollution?

Our question is based on the concept of "redlining", which is the discriminatory practice of denying services (financial and otherwise) to residents of certain areas based on their race or ethnicity. In this case, the "service" would refer to clean air (unaffected by pollution). To answer this question, we utilized Causal Inference methodology.

4.1 Methods

1. Variables
    - Treatment: White and Non-White
    - Outcome: Pollution Level (as measured by PM2.5 and ozone concentration
2. Confounding Variables
    - Socioeconomic status of a county. Non-whites may be more likely to live in areas with higher pollution levels due do geographic confounders, including propensity to live in industrial areas.

- County income. Non-white groups have historically lived in lower income/inexpensive areas, which may have more more pollution. To mitigate this, we need to control for median income.
- Propensity for wood burning/fireplace use. Burning wood is a major contributor to PM2.5 concentration, which may have an association with wealth (i.e. lower income households using fireplaces as an alternative to central heating).

We've shown in class that in order to apply propensity score matching, unconfoundedness must hold. In this case, unconfoundedness is not likely to hold, even though we include many control variables (as there are an inestimable number of confounding variables that could lead to racial dispersion in counties, or levels of pollution.

3. Confounder Adjustment Methods We will use propensity score matching to match similar counties with each other, adjusting for a large number of variables included in the census datasets.
4. Colliders As the variables we condition on are all confounders rather than colliders, we do not estimate any colliders in the dataset.

4.2 Results

In this section, we answer the question of how a county being majority non-white (the treatment) affects the concentrations of PM2.5 particulate and ozone in the county (the outcome).

Hypothesis: A county being majority non-white does not cause higher levels of pollution, as measured by concentration. Our hypothesis is based on the assumption that although confounders exist, there is no direct mechanism between race and pollution.

In order to estimate the causal effect, we apply propensity score matching. First, we estimate propensity scores by applying logistic regression on a slew of 20 covariates encompassing geography, demographics, and economics. Using the fitted values, we match on one, three, five, and ten matches, then compute the mean difference in outcomes between non-white counties and their matched white counties. The table below summarizes our model:

	Mean difference in PM2.5	Mean difference in Ozone
One match	-0.102189	0.111217
Three matches	-0.275258	0.273373
Five matches	-0.258093	0.243363
Ten matches	-0.321982	0.237915

On average, we estimate the causal effect of a county being majority non-white on PM2.5 as a 0.1 $\mu g/m^3$ to 0.3 $\mu g/m^3$ increase in concentration. We estimate the causal effect of a county being majority non-white on ozone as a 0.1 $\mu g/m^3$ to 0.3 $\mu g/m^3$ decrease in concentration.

We add multiple matches as a robustness check, and see that across the different numbers of matches, the general direction and magnitude hold.

The above analysis assumes that unconfoundedness holds (specifically, it assumes that in the estimation of the propensity scores, we included every possible confounding factor). As mentioned in section 2.1.2, the sheer number of possible confounding factors here is incalculable. Although we included many covariates, we're still possibly mising crucial confounders (i.e. smoking rates,

specific geographic data, and number of industrial buildings at the county level which could all effect our analysis).

Furthermore, concentration values of 0.1 $\mu g/m^3$ to 0.3 $\mu g/m^3$ are largely insignificant in proportion to the mean values of ozone and PM2.5 concentrations (at 41 $\mu g/m^3$ and 9 $\mu g/m^3$, respectively), as they represent a 0.4% and 0.2% difference.

If we assume our model is "close" to unconfoundedness, we can reason that "omitted" confounders like industrial buildings may lead to overestimating the effect of race on pollution. Therefore, taking into account these omitted variables, we condlude that it is unlikely there is a causal relationship between race and pollution.

4.3 Discussion

As discussed in the previous section, unconfoundedness is unlikely. Additionally, further limitations to this analysis lie in the data:

- The number of majority nonwhite counties in the US (220) is proportionally very low, which makes estimation of propensity scores much more difficult.
- Inability to use standard error estimation with propensity score matching due to Data C102 course scope. However, we believe these errors to likely be large given the small dataset and two-stage regression process.

5 How do environmental factors impact asthma mortality rates among women in the United States in 2012-2014?

For this question, our methodology utilized was prediction with GLMs and non-parametric methods.

5.1 Methods

In this research question, we are aiming to predict asthma prevalence among women in the United States using data between the years 2012-2014. We will be utilizing environmental factors as features of our models, such as ozone and PM2.5 concentrations to predict asthma prevalence.

The Generalized Linear Model that we will be using is a Gaussian Model. We chose this model because we seeking to predict real-valued outputs, asthma prevalence.

The non-parametric method we will be using is a Random Forest model. When estimating asthma prevalence using several non-parametric models, we found the Random Forest model to be most accurate and effective.

To evaluate the effectiveness of the Random Forest model, and compare its effectiveness to other non-parametric models, we calculated the root means squared error for both the training and test datasets' model predictions. Initially we found the Decision Tree model to have a very low training error relative to its test error, indicating over-fitting in the training set. With the Random Forest model, the training error is also low relative to the test error, suggesting slight over-fitting and poor generalization. Compared to the Decision Tree, the Random Forest model was more accurate and less over-fitted.

5.2 Results

The Gaussian model was used twice for two features, Ozone and PM2.5. The Ozone model, a coefficient of -0.0081 was produced while for the PM2.5 model a coefficient of 0.0069. This suggests that the higher the Ozone concentration, the less prevalent Asthma is in a state, which differs than what was expected. PM 2.5 had a small, statistically insignificant effect on Asthma prevalence. In general, Random Forest models are difficult to interpret. The model did find that the Ozone and PM2.5 features were almost equally important in predicting asthma prevalence among women.

The Bootstrap Standard deviation error of the coefficients Ozone and PM2.5 are both fairly low. These std errors are used to measure uncertainty within the model. The error for Ozone was .022, while the error for PM2.5 was .057, making Ozone the better predictor

5.3 Discussion

The GLM had a higher level of performance, as it has a lower error when compared to the non-parametric model. The significant difference in errors leads us to greater confidence when applying this to future datasets.

One limitation of our model is the data is not very recent (2012-2014). However, these years were chosen because they were the only years that contained the entirety of the necessary features.

6 Conclusions

6.1 Question 1: Race and Pollution

Based on our analysis in section 4.2, given that the confounding variables tend to cause us to overestimate the difference in pollution levels between white and non-white groups, we conclude that it is unlikely that there is a causal relationship between race and pollution.

Our results are not necessarily generalizable. As our census data is already generated from very large categories (most counties in the U.S., barring excluded counties as mentioned in Section 1.2), there are few broader categories that we'd be able to generalize to (such as continental or worldwide levels). Furthermore, this type of methodology would become very difficult in geographic regions with extremely low or high racial dispersion.

As our results show that a county's racial profile does not have a significant impact on level of pollution, our government policy examples as mentioned in Section 2 would not hold/be difficult to argue. As all groups suffer similarly from pollution based on our findings, we should seek to equally address pollution in all counties.

To generate the data required to answer this question, we merged census and pollution data (Datasets 1-3). One benefit of this merge was the ability to address confounders included in the census data, including socioeconomic status (by median income) and county population.

Limitations in this data include the sheer number of confounders that are incredibly difficult to address for the unconfoundedness to hold (such as concentration of industrial buildings).

To further build on this work in the future, it would be helpful to find a relevant instrumental variable approach, as we wouldn't need to attempt to achieve unconfoundedness. However, it is similarly difficult to find a perfect instrumental variable.

6.2 Question 2: Gender and Asthma

Our analysis shows that Generalized Linear Models performed better than non-parametric models when using Ozone and PM2.5 predictions to predict asthma prevalence in women across the United States. This conclusion was discovered by fitting the models on Ozone and PM2.5 as features, with GLMs producing lower error rates than Non-parametric models.

Similarly to Question 1, these results are not incredibly generalizable, and our findings are very broad due to the abundance of multiple years of census data.

Additionally, also similar to Question 1, we found an insignificant association between our treatment and outcome variables. Given this, we would suggest focusing policy efforts outside of these two environmental factors (for example, possibly based on other forms of pollution).

To generate the data required to answer this question, we merged Datasets 1, 2, and 4 to conduct our analysis. Given that Datasets 1 and 2 were limited to the years 2012-2014, we were limited to only using asthma prevalence data from that time period.

To further build upon this research, as we were limited in this study to statewide aggregated data over the span of 3 years, additional data on Ozone and PM2.5 concentrations over time would may be sufficient to improve this model.