Inference and Prediction of PM2.5 Exposure on Youth's Mobility Outcomes

Lan Dinh **Faith Dennis** **Huy Dao** **Anand Shankar**
University of California, Berkeley

1 Data Overview

Our analysis utilized multiple data sources to investigate mobility outcomes and their relationships with environmental and socioeconomic variables in California counties. Air quality data was sourced from the CDC's publicly available datasets (Centers for Disease Control and Prevention (CDC), 2014). Specifically, we downloaded a PM2.5 dataset spanning from 2001 to 2014. This dataset was filtered to include only California counties using an online filtering tool provided by the data portal. This data was selected because it represents the oldest and most comprehensive air quality information available for California counties.

Additionally, mobility outcomes and covariates were obtained from the Opportunity Atlas project, a collaboration between Opportunity Insights and the U.S. Census Bureau (Opportunity Insights, 2024). The Opportunity Atlas provides county-level data related to economic mobility and its determinants. We selected datasets labeled *County-Level Trends in Outcomes (1978-1992 Cohorts) by Parental Income, Race, and Gender* and *County-Level Covariates*. These datasets were downloaded from the Opportunity Atlas website in CSV format, accompanied by detailed codebooks and methodological descriptions.

We chose to integrate these datasets because they offer complementary information. The PM2.5 dataset provided environmental quality indicators, while the Opportunity Atlas data included socioeconomic metrics that are critical for understanding mobility. By combining these sources, we aimed to capture a more holistic view of factors influencing mobility in California counties.

The datasets represent a mix of sample-based and census-based data. For instance, the Opportunity Atlas data combines near-census-level records, such as federal income tax returns, with sample-based sources like the American Community Survey. The granularity of the data is at the county level, with each row representing aggregated metrics for a specific county. This aggregation limits our ability to analyze individual-level variation but provides valuable insights into regional patterns. While this granularity allows for robust comparisons across counties, it may obscure intra-county disparities or the experiences of specific demographic groups.

Our analysis is affected by several data challenges. The PM2.5 data and Opportunity Atlas datasets do not perfectly represent the entire population due to potential systematic exclusions. For instance, undocumented populations or individuals not filing tax returns may be underrepresented in the Opportunity Atlas data. These exclusions could introduce selection bias, particularly in understanding outcomes for marginalized groups. Moreover, some variables in the datasets were missing or incomplete. For instance, data for certain years or demographic subgroups were unavailable, which limited the scope of our analysis. Missing data entries were treated as non-informative and excluded during preprocessing, which could introduce bias if the missing data were not randomly distributed.

In the course of data cleaning and preprocessing, we applied several steps to ensure relevance and consistency. For the Opportunity Atlas data, we focused exclusively on California counties and selected columns relevant to the 1992 cohort. We retained only aggregated data across race and gender, indicated by *pooled* in the column names, and standardized column names by removing redundant substrings. For example, variables like `kfr_pooled_pooled_p50_1992` were renamed to `kfr_p50`, representing the mean of the 50th percentile rank. Irrelevant columns, such as those related to parental employment or state-level identifiers, were dropped. For the air quality dataset, we mapped county codes to county names and grouped data by mean predicted PM2.5 values.

The cleaning and preprocessing decisions were driven by the need to improve interpretability, reduce noise, and enhance computational efficiency. By narrowing the focus to California counties and relevant cohorts, we ensured that our models addressed specific research questions without being confounded by extraneous variables.

Our datasets lack important variables that could provide additional context. For example, data on housing quality, crime rates, and education metrics such as teacher-student ratios or standardized test scores are unavailable. These variables could have allowed us to investigate how living conditions and educational resources influence mobility outcomes. Similarly, economic indicators like local unemployment rates and economic growth trends, as well as intersectional data combining race, socioeconomic status, and immigration status, could have provided insights into unique barriers faced by specific groups.

Another limitation of the datasets is the missing historical data for earlier years, which constrains our ability to understand long-term trends. The reliance on aggregated county-level data also introduces challenges in interpretation, as it reflects average trends and outcomes rather than individual-level variations. This aggregation may obscure the experiences of underrepresented groups within counties and limits the generalizability of our findings to smaller populations or subgroups.

Finally, our analysis must contend with potential biases in the data. Selection bias could arise due to systematic exclusions, and measurement error might exist in variables such as air quality predictions. Despite these challenges, the preprocessing steps and data integration strategies enhance the relevance and consistency of the datasets for our research objectives.

2 Exploratory Data Analysis

Figure 1, a histogram of PM2.5 levels in 2001, illustrates the distribution of the fine particulate matter concentrations across different counties in California. The majority of PM2.5 levels fall within the range of 8 to 12, with fewer occurrences in higher ranges (14 and above). This suggests that most counties have good air quality, with a few experiencing worse pollution. This information establishes the context for examining how varying levels of air pollution might influence long-term socioeconomic outcomes.

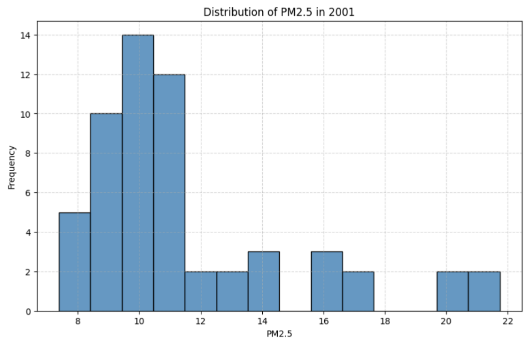

Figure 1: Distribution of PM2.5 in 2001.

Figure 2, a geographic map of PM2.5 levels, displays spatial variability in air quality across counties, with urban and industrial zones experiencing higher concentrations of particulate matter. For instance, South California and parts of the Central Valley show significantly worse air quality compared to Northern California and coastal areas. This geographic disparity aligns with known patterns of industrialization and urbanization, which often correlate with socioeconomic inequalities. It also motivates our questions of whether air quality is linked to socioeconomic outcomes, as it identifies areas where environmental conditions may disproportionately affect economically disadvantaged populations. Understanding where poor air quality is concentrated allows for a more targeted analysis of its potential long-term effects on residents' economic mobility.

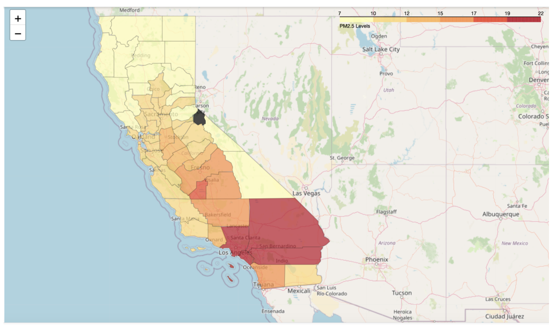

Figure 2: Geographic Distribution of PM2.5 Levels across California Counties in 2001.

Figure 3, a geographic map of income ranks by country, demonstrates a relationship between childhood neighborhoods conditions and adult economic outcomes. Regions with higher PM2.5 levels, as shown in Figure 2, often overlap with areas where children from lower-income families experience lower average income ranks in adulthoods. For example, parts of the Central Valley and Southern

California show both poor air quality and reduced socioeconomic mobility, while regions with better air quality, such as coastal Northern California, exhibit higher income ranks. This suggests that childhood air quality is a potential predictor of future socioeconomic outcomes, particularly for those from disadvantaged neighborhoods.

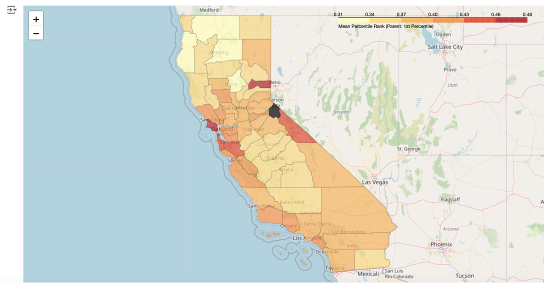

Figure 3: Geographic Distribution of Mean Income Percentile across California Counties in Adulthood.

Figure 4 examines the relationship between PM2.5 levels and average income ranks for individuals from various parental income percentiles. There exists a consistent shape across all parental income percentile groups, suggesting a shared underlying relationship between air quality and future socioeconomic outcomes. White the general shape of this relationship remains similar, the magnitude of the effect varies by parental income percentile. For lower-income families (e.g., 1st and 25th percentiles), there is a slight clustering of lower income ranks with higher PM2.5 levels. In contrast, for higher-income families (e.g., 75th and 100th percentiles), the same shape persists, but the income ranks are generally higher than the other parental income groups.

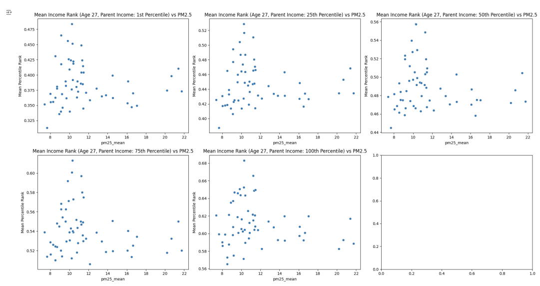

Figure 4: Relationship Between PM2.5 Levels and Mean Percentile Rank by Parental Income.

In addition, Figure 5 provides a comparative analysis of income ranks across different parental income groups. Children from families in the lowest income percentiles (1st and 25th) show significantly lower median income ranks. Combining these two insights, it becomes evident that more nuanced factors beyond air quality affect mobility across parental income percentiles, which will be explored in later sections.

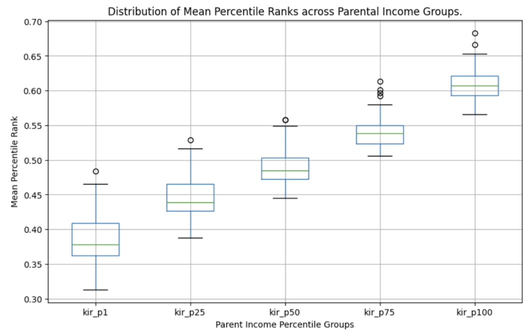

Figure 5: Distribution of Mean Percentile Ranks across Parental Income Groups.

3 Research Questions

Research Question 1: Is there a causal relationship between air quality at the county level and the socioeconomic mobility of the county's residents?

Answering the above question could help inform public policy decisions. Specifically, knowing that poor air quality in a county leads to worse economic outcomes could be used to implement environmental regulations to improve air quality, increase job opportunities and expand access to educational opportunities.

In order to answer this question, we plan to use causal inference techniques since we are interested in whether poor air quality causes lower economic mobility, not just the correlation between those two variables. Additionally, using causal inference allows us to explore this relationship in a meaningful way and account for some of the confounding variables which could misinform our results.

The limitations of using the causal inference techniques of propensity scores and outcome regression include the inability to account for all confounding variables and unobserved relationships between variables. When we are not able to model all confounders in the model, the causal effect of our chosen variable (in this case, air quality) might be distorted and we would only get a biased estimate of its effect.

Research Question 2: Can we use the air quality data of a county to predict the growth in income of the county's future residents?

This research question is essentially asking if exposure to poor air quality as a child affects future financial outcomes and economic mobility. Knowing the answer to this question can tell authorities where to direct medical resources and environmental legislation. It can also inform residents about the outcomes for their kids and give them an opportunity to factor this into their decision to live in a given county.

To build this predictive model, we will be comparing Generalized Linear Models (GLMs) and non-parametric models. Of the techniques we learned in the class, these are the ones best suited for predictive models and can be trained with several features. Additionally, these models are quite interpretable, which is useful in answering and analyzing the research questions.

The limitations of using GLMs for prediction include assuming that a linear (or nearly linear) relationship exists between the variables of interest. If the relationship was more complex, we are usually not able to model that using only GLMs. In the case of the non-parametric methods we used such as decision trees and random forests, there are concerns regarding the model overfitting to the training data as well as the model not being very interpretable.

4 Prior Work

Source 1: ”Childhood PM2.5 Exposure and Upward Mobility in the United States." (Kingsbury Lee et al., 2024)

In the research article published by PNAS, researchers seek to understand the relationship between childhood air pollution exposure and intergenerational mobility (Kingsbury Lee et al.). The justification for this research is that air pollution's impact on economic opportunities are not as well understood as air pollution's impact on health outcomes. Our research question seeks to understand this relationship as well for similar reasons. Like our research, they use PM2.5 levels and data from the Opportunity Atlas to estimate county-level associations, but use a different cohort. Researchers in the article look at children born in 1980 considering exposure in infancy and looking at outcomes for these children measured in 2015. We instead consider mobility from a 1990 cohort considering childhood exposure from 8-18 years old (multi-year air quality data) and measuring outcomes in 2019.

In the article, the researchers cross-validate their results using entropy balancing, inverse probability of treatment weighting, and generalized propensity score matching for continuous treatments, and they use a larger set of covariates for which they fit hierarchical models and perform a sensitivity analysis for unmeasured confounding to validate the robustness of their results. Our methods were more simple, using propensity weighting to address confounding, but with limited confounders available to us. We also attempted to cross-validate results by utilizing outcome regression. We did not consider sensitivity analysis. They find that there is a statistically significant negative relationship between PM2.5 exposure during infancy and earnings in adulthood.

Source 2: "Prenatal Exposure to Air Pollution and Intergenerational Economic Mobility: Evidence from U.S. County Birth Cohorts." (O'Brien et al., 2018)

Another source we used is a published research article considering how exposure to air pollution measured by TSP in the birth year affects intergenerational mobility outcomes of children in adulthood with a focus on how this difference varies based on income of families (O'Brien et al.). TSP is believed to follow PM2.5 making it relevant to our research (we look at PM2.5). They link economic mobility at the county level for children born between 1980 and 1986 with TSP particulates for their birth year. Adulthood outcomes are measured at age 26. We also look at air pollution effects on mobility outcomes of children in adulthood at the county level, but consider exposure from ages 8-18, outcomes measured at 27, and a later cohort year.

Both of our studies control for parent income percentiles. Researchers in the article estimate multivariate linear regression models for each cohort, regressing economic mobility on TSP levels. They adjust their models for confounders but do not specify the exact method used. They do however state that they grouped confounders into county economic characteristics and characteristics of the birth cohort. They also performed a sensitivity analysis. Like them, we used separate models for each birth cohort based on parent income similar to this article, but only had access to economic characteristics of the county and estimates with high variance of racial backgrounds. We performed outcome regressions, linear regressions, general linear regression models to estimate mobility based on air quality by year, but instead considered the years when children were 8-18.

Researchers in this article found that higher levels of TSP produced less upward economic mobility for children from low-income families whereas children from high-income families were not affected by TSP levels. This suggests that parental income may act as a moderator on this effect.

5 Causal Inference

5.1 Methods

We use the county level mean percentile individual income controlling for mean parent income percentile for our mobility outcomes. This data is collected in 2019 when children in the study are 27. Parent income percentile is taken when the children are born in 1992.

We use air quality from the CDC for our treatment. PM2.5 measurements from the CDC are provided on a linear scale at a daily level for each county. We rolled up this data to get a yearly mean of the median air quality for each county. To binarize air quality based on PM2.5 levels, we classify values below 12 as 1 (indicating good quality) and values of 12 or higher as 0 (indicating poor quality), as defined by the U.S. Environmental Protection Agency's Air Quality Index (United States Environmental Protection Agency (EPA), 2016). We considered air quality from 2001 because this was the oldest data we had available. This data corresponds to when the children are 8-9.

For confounders, we considered neighborhood characteristics aggregated at the county level from 2000 data. We believe these characteristics have the most direct influence on county level air quality in 2001 as opposed to older data and also encompass the information and influence of previous neighborhood characteristics. We tested 7 characteristics based on availability. Available characteristics included employment rate, fraction college graduation, fraction of foreign population, gini coefficient, median household income, per capita income, fraction below poverty level, and single parent fraction. We compared the distributions between each county's characteristics by air quality, and selected 4 whose distributions were visibly different by air quality group. We used these as confounders for propensity weighting based on the significance of their relationship. We selected fraction college graduates, foreign share, per capita income, and fraction poor as our confounders. Based on research, we do not believe that any colliders exist in our datasets.

We used logistic regression to calculate propensity scores for the effect of our confounders on the treatment, and applied inverse weighting to account for the effect of these confounders in our ATE scores.

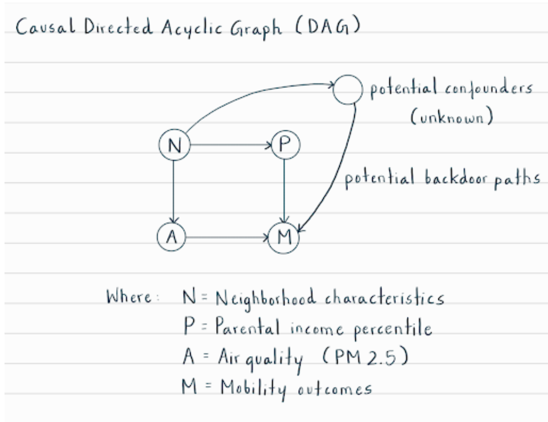

Figure 6: Directed Acyclic Graph

5.2 Results

The method of **Inverse Propensity Score Weighting** revealed a negative relationship between poor air quality (measured as PM2.5 levels) and upward mobility across all parental income percentile groups. As shown in Table 1, the Average Treatment Effect (ATE) ranged from -0.11 for the 1st percentile group to -0.19 for the 100th percentile group. While these estimates indicate a consistent negative trend, the magnitudes are relatively small, suggesting that the relationship between air quality and upward mobility is weak.

Percentile Group	ATE
1st	-0.1134383984643135
25th	-0.136193979766845
50th	-0.14781010990141114
75th	-0.16395000749596048
100th	-0.18699107997999328

Table 1: Estimated Treatment Effects for Different Percentiles using Inverse Propensity Score Weighting

Before proceeding with the **Outcome Regression** method, we calculated the Pearson correlation coefficients to evaluate the linearity between air quality (PM2.5 levels) and mobility outcomes (mean percentile rank) across different parental income groups. The results, summarized in Table 2, indicate weak or negligible linear relationships across all income percentiles. For instance, the

Pearson correlation coefficient for the 1st percentile group was -0.0426, while for the 100th percentile group it was -0.1506, both suggesting a lack of strong linear association.

Given these findings, we concluded that the assumption of linearity required for the Outcome Regression method is not satisfied. As a result, we decided not to draw any conclusions using this method, as it could lead to inaccurate or misleading results in the absence of a linear relationship.

Percentile Group	Correlation Coefficient
1st	-0.0425597351687
25th	-0.0151300552488
50th	-0.108009689938
75th	-0.16907848207
100th	-0.150576847671

Table 2: Pearson Correlation Coefficients for PM2.5 and Mean Percentile Rank by Parental Income Percentile

In conclusion, the use of Inverse Propensity Score Weighting indicates that children raised in areas with poorer air quality are, on average, slightly less likely to experience upward economic mobility as adults compared to those from areas with better air quality. However, the observed effect is small and not strongly conclusive, suggesting that while air quality may play a role, its impact on economic mobility is limited.

5.3 Discussion

Currently, the results from our analysis support the need for further investigation, as the observed weak negative causal relationship between air quality and upward mobility may be influenced by limitations in our methods and data. One significant limitation is the potential presence of unaccounted confounders, such as differences in access to quality education, healthcare, regional economic opportunities, or systemic factors like incarceration rates, which disproportionately affect certain demographics and may independently influence socioeconomic mobility. Incorporating these factors into the analysis is essential, as incarceration rates, particularly among marginalized groups, can significantly impact family stability, economic prospects, and access to opportunities, potentially confounding the observed relationship between air quality and mobility.

Additionally, our data lacks the granularity needed to capture variations within counties or neighborhoods, where individual-level factors and local policies might significantly affect the observed relationships. The assumption of linearity in the Outcome Regression method was also not satisfied, further reducing our confidence in some of the conclusions drawn from this method. These methodological challenges underscore the importance of incorporating additional contextual and demographic variables to refine the analysis.

Given these limitations, we are only moderately confident in asserting that there is a weak negative causal relationship between air quality and upward mobility. The small magnitudes of the observed effects suggest that the relationship, if present, is weak and potentially overshadowed by other factors.

6 Prediction

6.1 Methods

The goal for our prediction problem is to determine how well county level air quality exposure can predict county mobility outcomes controlling for parental income. Years from 2001-2010 were considered. These are the years for which the children for the study are 8-18 years old. Children sampled for the outcomes measured in our dataset are identified as having lived in their representative county until at least 18, so measured county air qualities after 2010 may not accurately represent exposure. The mean PM2.5 concentration for each of these years were treated as the features for each county and the mean percentile of individual income at 27 was treated as the response variable for our models. We had about 52 counties worth of samples after processing for each parent percentile, and we procured a 30% validation set from these. We assume parental income is the major modifier of the relationship of exposure on outcomes, so we control for this by using different models based on parental income. We consider parametric and non-parametric models and compare the results of both across all parent percentile groups.

For our parametric models we set up a generalized linear model. Because our outcome (percentile of individual income) is a continuous random variable, we considered a Gaussian likelihood. Using a Gaussian likelihood also made sense because our plots of the distribution of outcomes for each parent percentile are roughly Gaussian matching our assumption . We used the identity function as our link and inverse link taking on the assumption that the relationship between our mean outcomes and air quality are linear with Gaussian noise.

For non-parametric methods, decision trees, random forests, and neural networks were considered. Neural networks were ruled out as a good fit due to their complexity and tendency to overfit. Decision trees and random forests both make no assumptions of the data and can be used for regressing on continuous outcomes making them suitable for this task. Random forest offers the added benefit of ensembling to add uncertainty to the model and reduce overfitting. We compared all models with a naive estimator consisting of the mean outcome for each parent percentile group.

We evaluated both models on the basis of root mean square error, or RMSE.

6.2 Results

Parametric models: The root mean squared error was calculated for all five models, revealing a comparable error for all of them and the highest error for the 1st percentile income earners data.

The effects of pollution in different years varied significantly, and the trends in these variations were consistent across the five different income groups. Specifically, the air quality in 2004, 2009 and 2013 were much more consequential to the outcome variable compared to the other years. This suggests that environmental legislation during those years potentially had significant impacts on the air quality.

Non-parametric models: The random forests using the naive feature set performed best when predicting outcomes for the children born in the 1st, 25th, 50th, and 100th percentiles of parent income. This is likely due to decision tree models yielding a higher variance. The decision tree model predicting outcomes for the 75th percentile of parent income performed better than both random forests, but the difference was statistically insignificant.

The RMSE of the best model for each percentile group was around 0.02 (+/- 0.01) percentile points. The 1st parent percentile of income had the highest RMSE. All models performed better than the naive estimator with statistical significance (30-70% improvement), suggesting that air quality concentration offers predictive value. This confirms that air quality does have predictive value, though it does not speak to whether air quality may be a proxy for some other variable that directly impacts outcomes.

Below we plot the distributions of the outcome predictions against the distribution of actual outcomes. Figure 7 indicates success in the models to capture the general mode of the outcome distribution. The prediction distributions are concentrated around this mode.

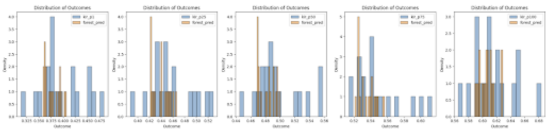

Figure 7: Distribution of the Random Forest Predicted Outcomes against Real Outcomes by Parent Income Percentile.

The size and sparsity of the dataset may have resulted in the models failing to learn the whole range of the domain, resulting in a failure to predict the full range of available outcomes and a tendency to underestimate higher outcomes. This is supported by the plots of the residuals. Residuals were plotted for the random forest models and indicated a variable positive bias across the models. This suggests a tendency across the models to underestimate outcomes. This could be caused by nonlinearity in the data that are not captured by the features, but also the above mentioned reasons.

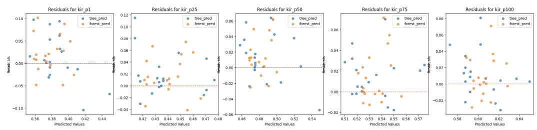

Figure 8: Residual Plots for Random Forests Models by Parent Income Percentile.

Overall, the results indicate that county air quality can be used to predict children mobility outcomes, but this predictive power may not be homogenous across different populations, having the least predictive power for children of top earners

6.3 Discussion

Both models revealed comparable findings. GLMs with the right assumptions and non-parametric models are worth applying again to cross-validate findings in future datasets.

To evaluate the goodness of fit of the best performing non-parametric models, the random forests, on the data, we used 5-fold cross-validation to produce 5 RMSE scores for each model and computed the average RMSE overall. This score is 0.025. This score should be used for comparative evaluation against other models and against the variance, with a smaller value indicating a better fit. Because the score is obtained via predictions on withheld data (cross-validation), it accounts for overfitting.

Using the fitted Gaussian generalized linear model to visualize the effects of each year on the income percentile revealed that the effect of air quality in 2012 had a large negative effect on income percentiles. On the other hand, the effect of 2011 air quality on the outcome was negative and small. Across the features visualized by importance using the fitted decision trees, no undisputed single year was identified as most influential. No relationship between influential years was identified either (i.e. if exposure earlier on is more predictive than later on).

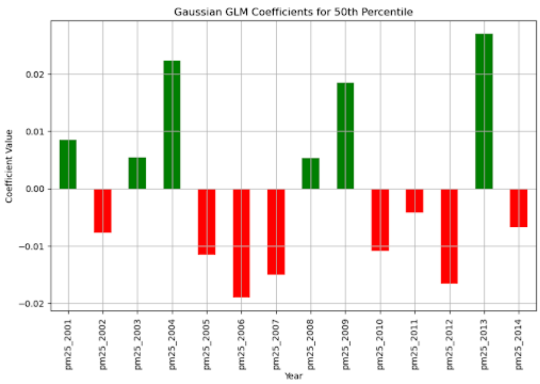

Figure 9: Gaussian GLM Coefficients for 50th Percentile.

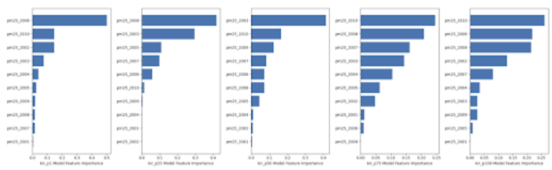

Figure 10: Plots of Feature Importance for Each Decision Tree Modeling by Parent Percentile.

The Gaussian GLM is limited as it relies on the assumptions we made. After the analysis, the Gaussian assumption still stands out as a reasonable option, but rather than taking a frequentist approach, adding a correct, informative Bayesian prior could improve predictive power. Both the GLM and the non-parametric models are also limited by a lack of data. Additional data and more complex feature engineering can improve the ability of the random forests to capture complex relationships. Without enough samples, the random forest models used cannot capture the full range of outcome values. Further, the predictive power may be limited by the use of naive features used to capture an otherwise complex relationship between air quality and mobility outcomes. This may have resulted in higher uncertainty in our results.

7 Conclusions

7.1 Outcomes summary

Our exploration of the first research question, we considered the causal impact of air quality on socioeconomic mobility. Our causal inference methods suggest a negative relationship between the two. This may support the conclusion that lower county air quality can lower children's future incomes.

In our exploration of the predictive value of county air quality across youth's developmental years on their mobility outcomes, we confirmed that this information can be used to predict outcomes, but the predictive power is weakest for the highest income earners in the county. We also find that the pollution levels of different years have non homogeneous associations with the outcome variable, but with no identifiable pattern.

7.2 Critical evaluation

Some limitations of the data we use include:

- The dataset is not very granular. There is great variation in air quality within counties, so reducing each county to one data point prevents us from seeing how specific neighborhoods might be affected differently. This may have negative implications for invoking the SUTVA assumptions.

- We use air quality data from 2001–2014 and income data from 2014, which only models the effects of air quality on income for one generation of a county's residents. In order to generalize the results and view historical trends, we would need to complete such analyses for several generations.

- It is not possible to account for all possible confounders given our limited dataset for the given counties and timelines. Based on similar research, a large set of confounders is needed for this problem because a variety of environmental and social factors are believed to affect mobility outcomes. Given further time and resources, this is something that should be considered in future studies.

The domain knowledge we are missing includes a more thorough model of the relationship between air quality and economic outcomes which includes all related confounder, collider and intermediate variables. One question that we might ask a domain expert would be to identify some important confounding variables that we have missed, and how we would access the data for those variables. Given that information, we can account for it in our causal inference models and obtain more accurate results.

Our results remain fairly consistent across different methods for both questions. One thing we might change is the number of outcome variables we are looking at. Since socioeconomic mobility could be characterized by many other variables such as high school graduation rates, college graduation rates, net income increase and change in income percentile, we might be able to infer more about our questions by using these different outcome variables. A user interpreting this data should keep in mind that we are using only one outcome variable currently, and using a few more would be more informative.

We see a general trend in our results of poorer air quality leading to worse economic outcomes. This feels plausible since poorer air quality correlates with lower income demographics, lesser developed neighborhoods and lesser access to education, opportunities and mobility. If we were to use other outcome variables that serve as a proxy for socioeconomic mobility, the results would look similar and show a negative relationship between air quality and mobility.

7.3 Recommendations

One follow-up study could look at more granular data at the city or ZIP code level and analyze the relationship between air quality and graduation rates or change in income percentile. This way, we could look at variations across different neighborhoods in a county and not just compare counties to each other. Such a study could also incorporate multiple outcome variables that serve as proxies for socioeconomic mobility and examine whether the negative relationship between air quality and mobility presents itself in all the different cases of outcome variables.

Actions that could be taken include passing laws that monitor air quality in high-risk areas and levy taxes on the companies creating this situation. The government could also designate certain areas as industrial areas that would be zones where factories can operate and people do not reside, so that they do not experience the negative effects of pollution.

- Such an idea might be feasible in large states/counties with free government land to shift industries to. However, it is not possible to enact this in densely populated areas without room to shift people or industries. The government might be able to implement such a plan, but this relies on public support and research-backed approval. The companies that would be taxed for polluting the environment would object to this idea, as would the residents who are asked to relocate to a cleaner neighborhood.

- Individuals that would have to relocate would face great inconveniences to do so, but might benefit in the long run because of better financial outcomes for their children. Environmental advocacy groups would appreciate the government prioritizing the health of citizens over the profit of companies. However, the corporations that stand to lose from such a plan would be affected financially. Speaking to an individual who objects to the plan might entail informing them of the long-term benefits of living in a cleaner environment in terms of health, outcomes and mobility.

- The values guiding this recommendation include setting up an equitable system where communities that are most affected by poor air quality are given an option to remedy that solution, and the companies that cause the pollution are held accountable for their actions. The proposal seeks environmental justice that helps affected communities as well as takes steps to preserve nature and natural resources.

7.4 Github Code: (PM25InferenceCode)

References

Centers for Disease Control and Prevention (CDC). 2014. Daily county level pm2.5 concentrations, 2001-2014. Accessed: 2024-12-13.

Sophie-An Kingsbury Lee, Luca Merlo, and Francesca Dominici. 2024. Childhood pm2.5 exposure and upward mobility in the united states. *Proceedings of the National Academy of Sciences*, 121(38):e2401882121.

Rourke L. O'Brien, Tiffany Neman, Kara Rudolph, Joan Casey, and Atheendar Venkataramani. 2018. Prenatal exposure to air pollution and intergenerational economic mobility: Evidence from u.s. county birth cohorts. *Social Science & Medicine*, 217:92–96.

Opportunity Insights. 2024. The opportunity atlas. Accessed: 2024-12-13.

PM25InferenceCode. 2024. Inference and prediction of pm2.5 exposure on youth's mobility outcomes. Accessed: 2024-12-13.

United States Environmental Protection Agency (EPA). 2016. 2012 aqi fact sheet. Accessed: 2024-12-13.