Predicting air pollution based on social-demographic attributes

Data Science Project

Fanni Varhelyi
May 10, 2023

Executive Summary

This paper investigates the relationship between residential areas of racial minorities (on a census tract level) and PM2.5 air pollution. Air pollution has been linked to adverse health effects, even death, and if certain elements or minorities of the population are bearing a disproportionate burden related to air pollution, it is important to leverage this knowledge in policy making. Currently, an vital U.S. policy, Justice40, does not regard race and ethnicity as a factor when deciding if an area is disadvantaged or not. The goal of this paper is to show that there is an explicit relationship and race and ethnicity should be taken into account when looking at pollution, such as PM2.5 aerial levels.

The paper leverages three machine learning techniques in an attempt to show this relationship. It includes sources of pollution and poverty as control factors and finds that the presence of both Non-Hispanic Black and Hispanic minorities can be used to predict air pollution and these are more important variables than poverty. While more research is needed to definitely show how these minority residential areas bear a disproportionate burden, this is a promising start and a potential argument for the inclusion of race & ethnicity in the Justice40 policy model.

I. Introduction

Air pollution has well-known adverse effects: it has been linked to health issues and environmental damage. The World Health Organization estimates that annually 7 million deaths are associated with air pollution globally (Roser, 2021). Reliably being able to predict air pollution can thus help save lives and affect policy decisions. Academic research has been using machine learning modeling to predict air quality based on physical attributes such as weather patterns and pollution sources (e.g., Castelli et al., 2020). Research conducted in the United States typically uses EPA data, collected by ground-based air quality monitors.

Another avenue of research is identifying who is affected the most by air pollution. Rather than looking at how air quality changes over a short time frame, it aims to identify patterns related to areas where average air quality is significantly worse than elsewhere. In the United States, studies have linked pollution, waste sites, and air quality to race and poverty. For example, a 1987 study, reaffirmed in 2021, found that Toxic Waste Facility locations are correlated with residential areas of racial minorities (United Church of Christ, 1987; Mascarenhas et al., 2021). Other research linked air quality to race and ethnicity, finding it a much better indicator than poverty (Miranda et al., 2011). Overall, the idea behind these research questions, that certain communities bear disproportionate burdens and identifying them is imperative, also led to a recent initiative called Justice40.

Justice40 is a policy initiative of the current administration that aims to identify disadvantaged communities, and ensure adequate funding is allocated to them (The White House, 2022). One of its design principles, however, is that it does not include race & ethnicity as an attribute. It also only aims to identify disadvantaged communities, but it does not establish or investigate correlation or causality.

Based on the above, my project aims to leverage the existing research that links socio-demographic factors to air pollution and investigate the existing predictive models to see if any learnings can be leveraged from them. At the same time, I aim to use the attributes identified by Justice40, supplement them with metrics on race and ethnicity, and see if these can predict one of the major air pollutants in the U.S., PM2.5. This research aims to address a shortcoming in the existing Justice40 research, which does not include race and ethnicity as an attribute, though academic research has identified a correlation between it and pollution. If race can be used predict air pollution, that might be an argument for its inclusion in the attributes used to identify disadvantaged communities.

My explicit research question is as follows:

How does socio-economic background, more specifically, race and ethnicity, relate to PM 2.5 air quality? Can these indicators be used to predict air quality on a census tract level?

II. Data

The primary data source for this research is the data behind the Justice40 tool. The tool has been developed by the US Digital Service in partnership with the Council on Environmental Quality (Climate Economic Justice Screening Tool, GitHub). The dataset is organized on a census tract level and contains 602 attributes, with N = 74,160. It contains data on numerous air quality indicators.

Since this data does not contain information on race and ethnicity, additional data was acquired through an API from the US Census Bureau: the American Community Survey (ACS) includes data on race, ethnicity, and poverty on a census tract level. ACS data (5-year, 2018) was merged with the Justice40 dataset to provide a complete set of relevant indicators.

A potential alternative source of data could be combining satellite measurements with decennial census data and look at even smaller (census block group or census block) level of resolution. It has been shown in some cases that satellite data can also be a reliable source and might be better as air quality monitors do not cover all areas equally well (Li et al., 2011, Castelli et al., 2020). However, I decided to use the Justice40 data as the main motivation behind my research was proving the Justice40 dataset could benefit from using variables on racial and ethnic makeup of the population when classifying census tracts as disadvantaged. Nevertheless, satellite data could be a good source to verify the findings of this study, and could address some of the limitations of this data, most notably the inability to investigate the relationship on a more nuanced resolution.

As the Justice40 and ACS datasets both contain many irrelevant variables, the following selection was made:

- **Outcome variable:** *PM 2.5 level in the air*. I choose this indicator as PM 2.5 is one of the most dangerous air pollution source, and this pollutant is often investigated in literature (e.g., Miranda et al., 2011)

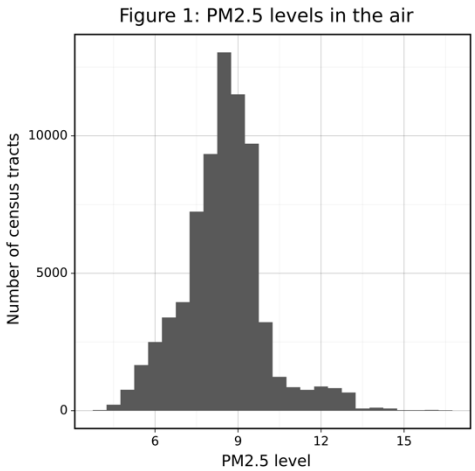
- **Socio-economic factors:** To answer my research question, I will focus on the presence of two minorities: *non-Hispanic Black population as percentage of the total population*, and *Hispanic population as a percentage*. I also included *percentage of population below the poverty line* as a control variable. I created these three variables using the ACS data by dividing the absolute numbers with the total population. I will not investigate other demographical factors as these variables are highly correlated.
- **Pollution-level determining factors:** To control for sources of pollution, two additional variables are included from the Justice40 dataset: *Proximity to hazardous waste sites*, and *Proximity and volume of traffic*.

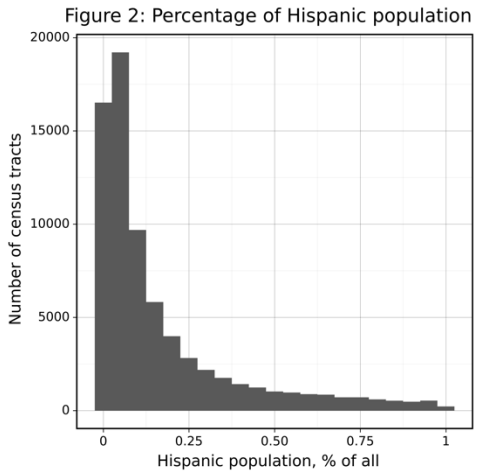
Table 1 provides descriptive statistics for the outcome variable and the other important features outlined above. Out of 74 160 census tracts, 72 877 could be matched between the two data sources, and after removing tracts with one or more missing values, the final number of census tracts in the model is N = 69 688.

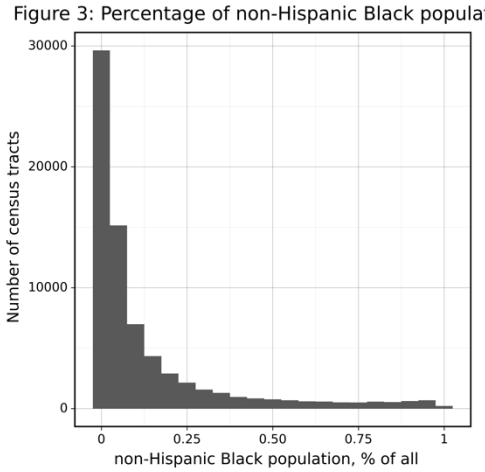
Variable	Mean	Std. Dev.	Min	Median	max
Hispanic percentage of population	0.17	0.21	0	0.08	1.00
Non-Hispanic Black percentage of pop.	0.13	0.21	0	0.04	1.00
Percentage of pop. below poverty line	0.15	0.11	0	0.12	1.00
Aerial PM2.5 level	8.49	1.49	4.00	8.53	16.46
Proximity to waste sites	5.16	21.16	0	1.07	434.08
Traffic proximity and volume	823.62	1,641.74	0	298.43	31,282.37

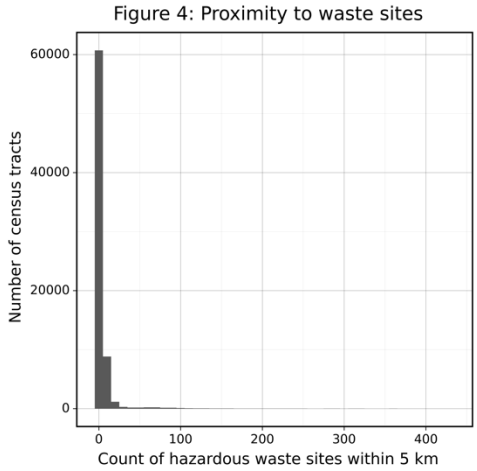
Table 1: Descriptive statistics. N = 69 688

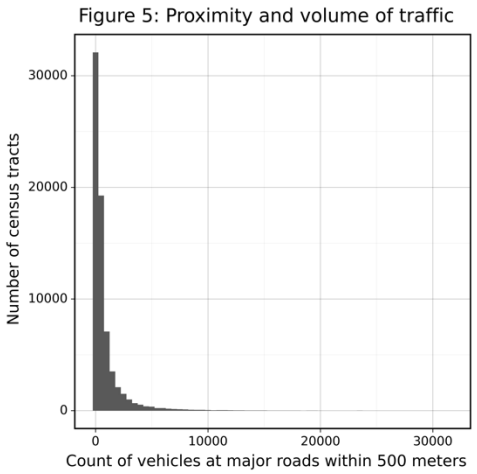
Figure 1-6 describes the distribution of these variables. First, the outcome variable appears to have a somewhat bell-shaped distribution, with some skewness to the left. *Proximity to waste sites* and *Traffic proximity and volume* are both skewed right, and there seem to be outliers for both variables with high values and low frequency. The *Hispanic, Black* and *% of poverty* variables are also skewed right, but does not seem to have extreme outliers.

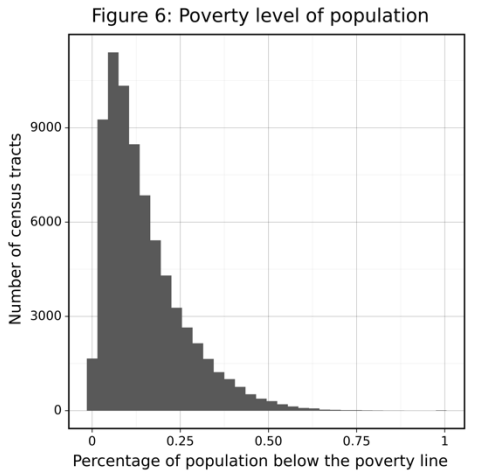




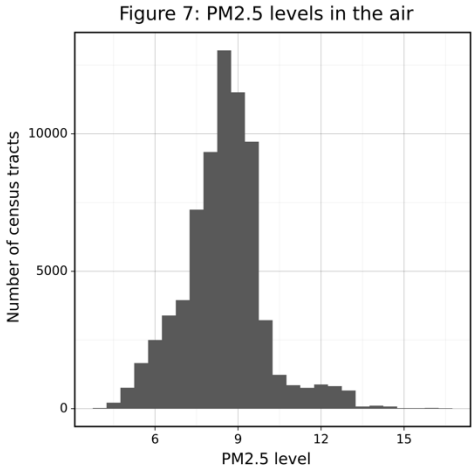


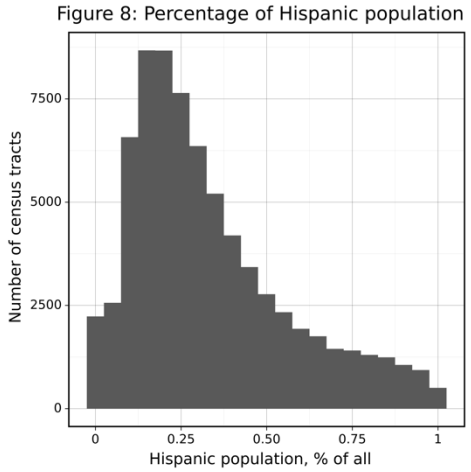


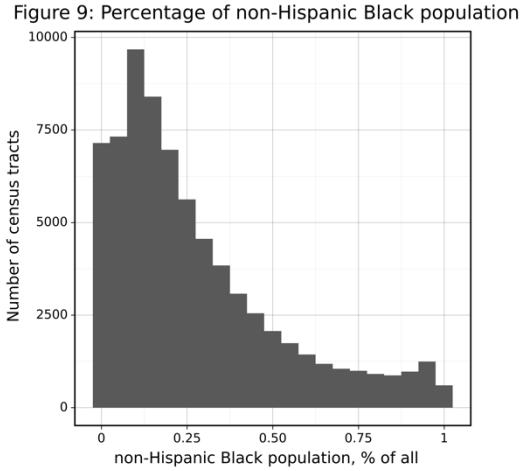


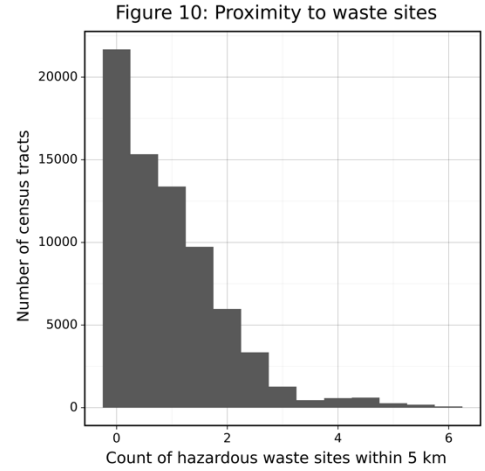


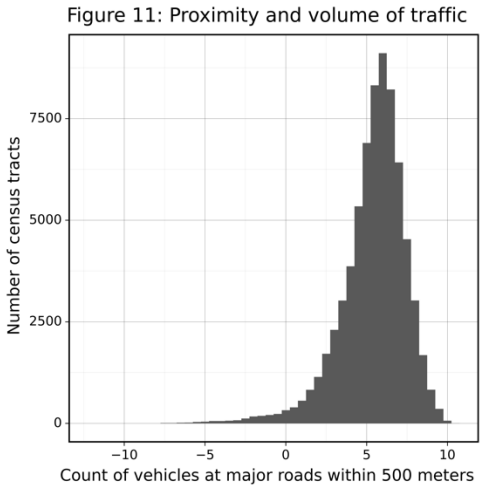
To address the skewness of the variables, I decided to use the most appropriate transformation per feature, a combination of logarithmic and square root, respectively, and not drop any outliers. The reason behind this decision is preserving outlier values to better capture extremities and whether these extreme values are related to racial and ethnic minority presence or not. After the transformations, the skewness was not fully addressed, but it was improved for all variables. Figure 7-12 displays the final distribution of the variables.

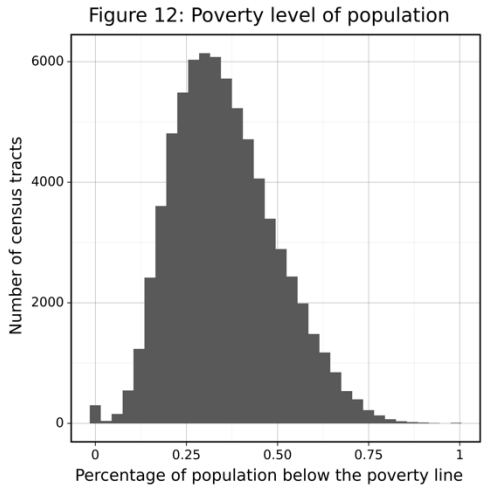












Additionally, Table 2 describes the relationship between each variable and the outcome variable. The outcome variable doesn't seem to be strongly correlated with any individual variables, but there seem to be some connection with traffic proximity and Hispanic population. Similarly, there seem to be some correlation between some of the variables, most notably between waste sites, traffic, and Hispanic / Black population, and between poverty and Black / Hispanic population.

Variable	Hispanic % of population	Non-Hispanic Black % of pop.	% of pop. below poverty line	Aerial PM2.5 level	Proximity to waste sites	Traffic proximity and volume
Hispanic % of population	-	0.25	0.06	0.14	0.28	0.11
Non-Hispanic Black % of pop.	0.25	-	0.30	0.06	0.24	0.13
% of pop. below poverty line	0.06	0.30	-	0.03	0.10	0.05
Proximity to waste sites	0.28	0.24	0.10	-0.10	-	0.26
Traffic proximity and volume	0.11	0.13	0.05	0.44	0.26	-
Aerial PM2.5 level	0.14	0.06	0.03	-	-0.10	0.44

Table 2: pair-wise correlation between the variables, and the variables and the outcome

III. Methodology

I used two different versions of the dataset to test my hypothesis. First, I tested all the below described models on the full dataset to capture the potential relationship between the variables and the outcome. Then, I took a specific sample. Based on the work of Miranda et al., 2011, I selected the 20% of census tracts that had the best, and the 20% of census tracts that had the worst PM2.5 air quality. The goal of this sampling was to see if my findings are different if I disregard the average census tracts, and only focus on the outliers. This does not replicate Miranda et al.'s 2011 study as it tested a multivariable logistic regression model, but it replicates their sampling logic. I will refer to this sampled data as the subset onwards.

I leveraged two types of models to ensure accurate results. First, I made an assumption about the relationship between the variables and the outcome. I used linear regression and LASSO to test the relationship. Gray et al. (2013) used linear regression to investigate the relationship between race, socio-economic status, and air in North Carolina, and found a significant relationship. Linear regression is a straightforward method: it assumes the shape of the relationship to be a line (or a plane) between air quality and the other variables. To put it simply, if my hypothesis is true, as the percentage of non-Hispanic Black and / or Hispanic population increases, the PM 2.5 level also increases. The LASSO model builds upon this by applying a penalty to less influential variables, thus clarifying which variables are important for explaining or predicting the outcome. (James et al., 2021) I used validation curves to find the optimal alpha parameter for the LASSO models.

Linear models are potentially susceptible to skewed data. To ensure the best results, I tested both the original and the preprocessed version of the dataset and found that the preprocessed version produced slightly better results. In addition, I rescaled the variables using the MinMax scaler before implementing the LASSO model.

The potential weakness of a linear approach is that in reality, the relationship between variables is often nonlinear. To test for this, I also utilized a nonparametric model, random forest. Decision trees split the data using if-else logic to find subsets of the data in a way to be able to determine the relationship between the variables and the outcome – i.e., to be able to predict the outcome with as little error as possible. Adejare et al. (2020) used this method when investigating racial disparities related to asthma. I used Decision Tree as a tool to understand the underlying logic in the dataset and as a tool for hyperparameter tuning for the random forest.

I utilized random forests as my main nonparametric model, a method that was also used to investigate racial disparities and air quality (Cheeseman, 2022). The random forest method builds multiple decision trees using a subset of observations and variables. Random forest is more difficult to interpret than the linear regression and LASSO models, but it has the potential for making more accurate predictions. Additionally, it is less susceptible to skewed data, and thus I used the original dataset when fitting the model. Based on the results of the Decision Tree, I decided to use a validation curve to find the optimal depth for the random forest models, as the tree showed depth is potentially an issue when fitting this data.

For all models, I used R^2 as the metric to decide between their performance. I fit linear regression, LASSO, and Random Forest on both the original dataset and the subset dataset, compared the results, and investigated feature coefficients and feature importance in order to interpret the results. For all models, I used 80% of data as training set and 20% as test set. All hyperparameter tuning was conducted through validation curves, with cross validation folds of 5. The random forest consisted of 1000 trees, and depth was limited to 9 and 21, respectively.

IV. Findings

Table 3 compares the different models using R^2 and MSE on both the original and subset data. All results refer to test scores. The results clearly indicate that Random Forest is the best model when using only the top/bottom 20% subset data with an R^2 score of 0.77 and an MSE score of 0.37.

	Metric	**Linear regression (preprocessed)**	**LASSO**	**Random Forest**
Original dataset	R^2	0.14	0.13	0.30
	MSE	*1.89*	*1.92*	*1.54*
Subset data	R^2	0.27	0.26	0.77
	MSE	*1.12*	*1.10*	*0.37*

Table 3: Model comparison

The results clearly indicate that there is a relationship between these variables, but this relationship is not well approximated using linear methods. Furthermore, the random forest performed significantly better on the subset data, indicating that these variables are more accurate for prediction when only the best and worst areas (from a PM 2.5 air quality perspective) are investigated. The best R^2 of 0.77 is a relatively good result, and the MSE is also an improvement compared to the linear models.

The poor performance of the linear model somewhat limits the usability of these findings for policy purposes. The random forest is difficult to interpret and understand and would be probably more difficult to use in a policymaking environment. However, we can still use the coefficient tables to understand the magnitude and importance of the variables from the linear model – with the limitation that the model had poor performance and thus should not be used as the sole source of interpretation. Table 4 displays the different coefficients form the linear regression and LASSO models.

	Original dataset		**Subset data (top/bottom 20%)**	
Variable	Linear reg.	LASSO	Linear reg.	LASSO
Hispanic % of population	1.31	1.19	1.24	1.25
Non-Hispanic Black % of pop.	0.85	0.69	-0.91	-0.62
% of pop. below poverty line	-0.38	0	0.24	0
Proximity to waste sites	0.21	1.29	0.35	1.16
Traffic proximity and volume	0.08	0.59	0.02	0

Table 4: Linear and LASSO regression coefficients

In all cases, traffic proximity and volume as well as poverty seem like as less important features, and Hispanic population seems like the prominent one. However, both poverty percentage and non-Hispanic Black percentage changes from positive to negative between looking at the entire U.S. and the subset of best / worst census tracts. Given literature has well established both poverty and non-Hispanic Black as having a positive relationship with pollution (e.g., Miranda et al., 2011), the results of these models are questionable and do not compare well with other research.

The random forest models' variable importance plots can also be used to better understand how prominent each feature is. Figure 13 and 14 compares the feature importance of the two random forest models. Interestingly, the Hispanic percentage of the population becomes the most important feature when only focusing on the subset data. This indicates that the presence of Hispanic minorities in an area serves as a better indicator on whether that area has good or bad air quality for the best and worst census tracts than sources of pollution or poverty. Similarly, the presence of non-Hispanic Black population is also a more important feature than poverty or traffic proximity.

Figure 13: RF Variable Importance on full dataset

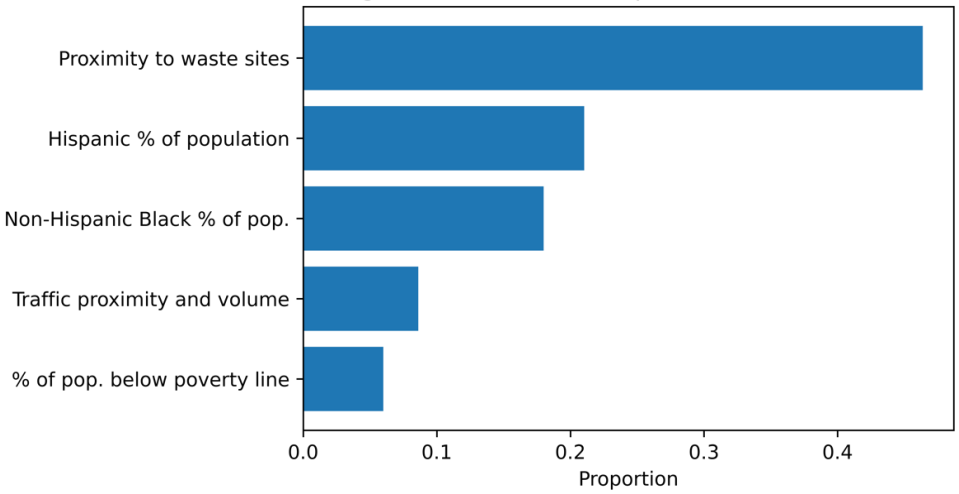
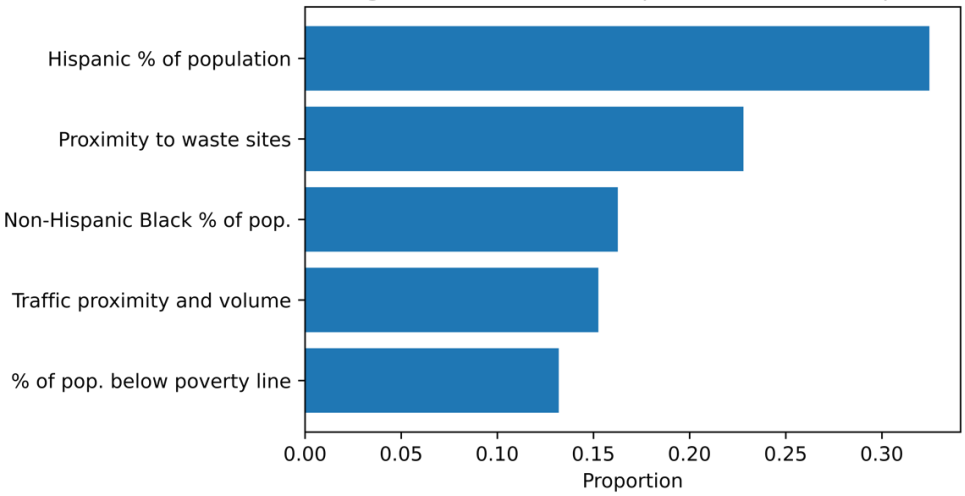


Figure 14: RF Variable Importance on subsample



These results indicate that it might be beneficial if the Justice40 platform included race and ethnicity in their evaluation when classifying certain census tracts disadvantaged. In all models, including the linear ones, the presence of minorities in the population either had a more important / more prominent coefficient, or was a more important feature.

V. Conclusion and Limitations

The linear models failed to produce accurate predictions, but the random forest model performed relatively well on the subset of the data that contained only the best and worst census tracts by levels of PM 2.5 in the air. This result indicates that while understanding the exact nature of the relationship between race & ethnicity and air quality might require additional studies, there is a relationship and a random forest model can use it to predict air quality relatively well. It also indicates this relationship is most prevalent when looking at census tracts with the best and worst air quality. In any case, this study provides at least an argument that race and ethnicity should be

potentially considered as a factor for Justice40 when qualifying census tracts disadvantaged, and it seems to be a better predictor than poverty.

There were several limitations that potentially impacted these results. Most notably, census tract level might be a too big geographical area to fully capture the nuances of air quality. Studies using satellite data (e.g., Li et al., 2011, Castelli et al., 2020) could produce air quality measurements on a more precise level than studies using EPA data. Furthermore, researchers also found that certain minority areas lack even adequate number of air quality monitors (e.g., Miranda et al., 2011), which is a further limitation of this study as well.

The data analysis methodology could also be a potential limitation and an area of further consideration when thinking about follow-up research designs. I leveraged existing literature, most notably Miranda et al., 2011, when deciding to select 20% of census tracts with the best and worst air quality, and to look at non-Hispanic Black and Hispanic population. Results might be different when looking at 10% or 15% of census tracts, and other minorities could also be investigated. Understanding the correlation between variables, some of which might be non linear in nature, could also help fine-tune the results. For example, toxic waste site placements have been shown to be correlated with minority population residential areas, most notably with Black population (United Church of Christ, 1987; Mascarenhas et al., 2021).

Finally, while the models relied on multiple data sources and tried to incorporate sources of pollution, it did not include all predictors potentially affecting air quality. Weather patterns or some extreme weather phenomenon might affect air quality in a given period of time. Additional emission sources, or tree canopy coverage could be also included. Repeating the analysis using different time periods could also strengthen the validity of the results by showing this is not a one-off phenomenon but a pattern persistent over years regardless of small changes in weather or emission sources. Finally, linking the results to adverse health effects of bad air quality could strengthen the policy argument that action is necessary not only to improve air quality metrics, but to prevent further illness or death.

Implementation appendix

To be able to analyze how race and ethnicity relate to air quality, I had to augment the Justice40 data with demographics attributes. As also mentioned in the Data chapter, I did this using American Community Survey 5-year data from 2018. To access this dataset, I used the Census API and pulled specific variables by perusing the full list of variables in the ACS dataset and selecting the ones relevant for my study (Census Data API, 2018). Leveraging my literature review (e.g., Miranda et al., 2011), I decided to investigate non-Hispanic Black and Hispanic population, as well as poverty to provide a control variable. I also included all other racial counts in case there was a need to include another minority group in the analysis, and the total population so that I can calculate the percentages. I pulled the following variables from the data:

- B01003_001E, B03002_012E, B03002_003E, B03002_004E, B03002_005E, B03002_006E, B03002_007E, B03002_008E, B03002_009E, B17001_002E

I used tract-level GEOIDs to merge this dataset with the Justice40 data. I acquired the Justice40 data through the Environmental Impact Data Collaborative's Redivis data portal API (EIDC, Redivis). I subset the data to only include the variables of interest for my research.

To minimize skewness of the attributes, I checked each variable's distribution using the skew method. This identified skewness in all features, but not in the outcome variable. I implemented a log transformation to account for this skewness, but it did not yield a good result for all variables. Then, I tried two more methods, square root transformation and box cut transformation, and selected the most appropriate one for each variable. Table 5 displays the differences between the transformations and the final results that was used for linear regression and LASSO models. The log transformations contain a numeric correction of adding one for the variables where 0 was also a value is the dataset (proximity to waste sites, and the demographic variables).

Variable	Original skewness	Log transform	Square root transform	Boxcut transform	Final dataset
Hispanic % of population	1.94	1.62	0.96	n/a	0.96
Non-Hispanic Black % of pop.	2.29	1.97	1.19	n/a	1.19
% of pop. below poverty line	1.45	1.14	0.45	n/a	0.45
Proximity to waste sites	10.34	1.63	4.43	n/a	1.63
Traffic proximity and volume	5.36	-1.32	1.98	167.14	-1.32
Aerial PM2.5 level	0.49	n/a	n/a	n/a	0.49

Table 5: skewness of variables before and after specific transformations

When implementing the models, I used the preprocessed dataset for all the linear models, as this dataset seemed to perform slightly better measured by R^2. In addition, I used cross validation to

define the best alpha parameters for the LASSO models. In general, the LASSO models performed similarly to the linear regression models and could not improve performance through shrinkage. The optimal alphas were 0.01 and 0.015 for the original dataset and the subset, respectively. Figure 15 and 16 shows the two validation curves. The poor model performance is apparent from these curves as well. The optimal alphas were chosen not on best model performance, but by choosing a model that already shrinks some coefficients to 0 while still performing relatively well.

Figure 15: LASSO Validation Curve on full dataset

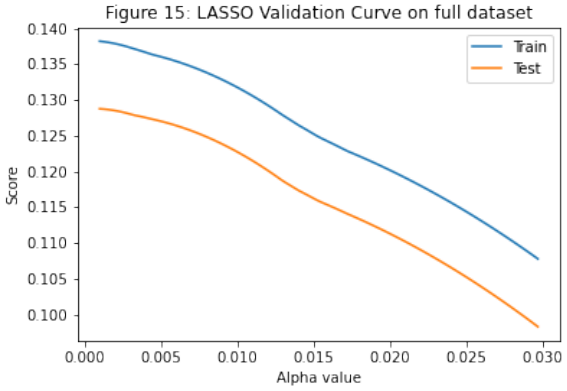
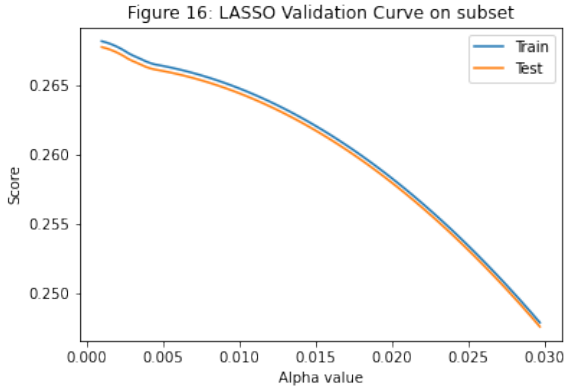


Figure 16: LASSO Validation Curve on subset



Similarly, validation curves were used to select the best performing random forest models. Figure 17 and 18 displays the random forest validation curves. The best maximum depth hyperparameter for the models were 9 and 21, respectively.

Figure 17: Random Forest Validation Curve on full sample

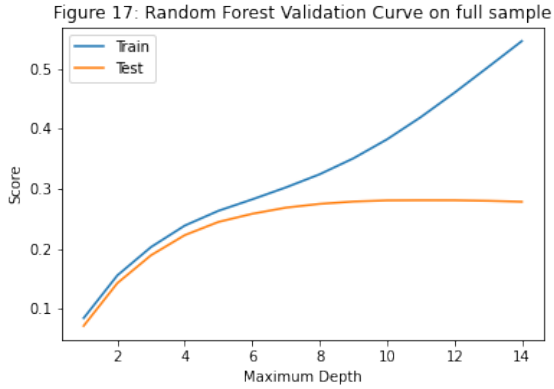
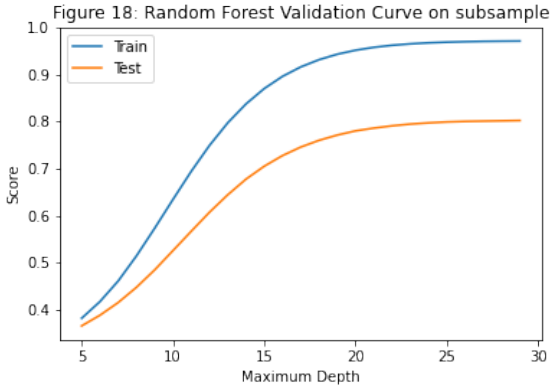


Figure 18: Random Forest Validation Curve on subsample



Bibliography

Adejare, A. A., Gautam, Y., Madzia, J., Mersha, T. B. (2022) *Unraveling racial disparities in asthma emergency department visits using electronic healthcare records and machine learning*. Journal of Asthma, 59:1, 79-93, DOI: 10.1080/02770903.2020.1838539

Castelli, M., Clemente, F. M., Popovič, A., Silva, S., Vanneschi, L. (2020). *A Machine Learning Approach to Predict Air Quality in California.* Complexity, vol. 2020, Article ID 8049504, 23 pages, 2020. https://doi.org/10.1155/2020/8049504

Cheeseman, M. J. (2022). *Investigating the enhancement of air pollutant predictions and understanding air quality disparities across racial, ethnic, and economic lines at US public schools* (Order No. 29060492). Available from ProQuest Dissertations & Theses Global. (2672375452). Retrieved from http://proxygt-law.wrlc.org/login?url=https://www.proquest.com/dissertations-theses/investigating-enhancement-air-pollutant/docview/2672375452/se-2

Gray, S. C., Edwards, S. E., Miranda, M. L. (2013). *Race, socioeconomic status, and air pollution exposure in North Carolina.* Environmental Research, Volume 126, Pages 152-158, ISSN 0013-9351. https://doi.org/10.1016/j.envres.2013.06.005.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.) [PDF]. Springer.

Li, C., Hsu, C. N., Tsay, S. (2011). *A study on the potential applications of satellite data in air quality monitoring and forecasting.* Atmospheric Environment, Volume 45, Issue 22, 2011, Pages 3663-3675, ISSN 1352-2310. https://doi.org/10.1016/j.atmosenv.2011.04.032.

Mascarenhas, M., Grattet, R., & Mege, K. (2021). *Toxic Waste and Race in Twenty-First Century America,* Environment and Society, 12(1), 108-126. DOI:10.3167/ares.2021.120107

Miranda, M. L., Edwards, S. E., Keating, M. H., & Paul, C. J. (2011). *Making the environmental justice grade: the relative burden of air pollution exposure in the United States.* International journal of environmental research and public health, 8(6), 1755–1771. https://doi.org/10.3390/ijerph8061755

United Church of Christ. (1987) *Toxic wastes and race in the United States: a national report on the racial and socio-economic characteristics of communities with hazardous waste sites.* Public Data Access: Inquiries to the Commission, New York, N.Y.

Roser, M. (2021, November 25). *Data Review: How many people die from air pollution?* Our World in Data. URL https://ourworldindata.org/data-review-air-pollution-deaths

The White House. (2022). Justice40. *The White House*. URL https://www.whitehouse.gov/environmentaljustice/justice40/

Data sources:

Climate Economic Justice Screening Tool. *Justice40 map.* Screeningtool.geoplatform. URL https://screeningtool.geoplatform.gov/en/#3/33.47/-97.5

Environmental Impact Data Collaborative. (2022). *Justice40 Tool* (Version 2.0) [Data set]. Redivis (RRID:SCR_023111). URL: https://redivis.com/datasets/mwa0-b9v8xcbzk?v=2.0

Justice40 Initiative team. Justice40tool. *GitHub*. URL https://github.com/usds/justice40-tool

United States Census Bureau. (2018) *American Community Survey 5-year estimates.* United States Census Bureau. URL https://data.census.gov

United States Census Bureau. (2018) *American Community Survey list of all attributes.* Census Data API. URL https://api.census.gov/data/2018/acs/acs5/variables.html