

PPOL 564 GROUP 2 FINAL PROJECT :

Exploring the relationship between asthma, race, and toxic release facilities

Cecil John¹, Fanni Varhelyi², Keya Sengupta³, and Zhongxian Liu⁴

^aGeorgetown University, McCourt School of Public Policy

This manuscript was compiled on December 7, 2022

This research tries to explore the relationship between asthma, the presence of facilities releasing toxic air pollution, and racial demographics in the United States. We analyzed the data on how proximity to toxic waste facilities affects people's health with regards to asthma and how this effect potentially differs based on the ethnic / racial demographics in the immediate region, while also taking into account poverty. Our results show that there indeed seem to be a relationship between race and the positioning of the toxic release facilities - we found in the two states we investigated, Louisiana and North Carolina, a higher density of toxic release facilities in counties where the population were 30% or more non-white than in counties where the population was less than 30% non-white. Furthermore, our results indicated that the presence of a toxic waste facility in any given census tract was positively correlated with a higher percentage of asthma cases in the population. On the other hand, the results indicated a minor but negative relationship between asthma and race/ethnicity. The strongest indicator in our study that affected asthma was poverty, which potentially leads to follow-up questions to investigate the interaction between poverty and the other variables in our study.

Data Science | Asthma | Air Pollution | Environmental Racism

1. Introduction and related work

The adverse impact of air pollution on everyday lives is well-known. Health issues caused by air pollution impact pulmonary, cardiac, vascular, and neurological systems and have become one of the most severe health issues in the United States. In 2005, a CDC report showed that 130,000 PM2.5-related deaths and 4,700 O3-related deaths were because of air pollution. Asthma is one of the major air pollution-related health issues costing, on average, 56 billion US dollars annually. It is a life-threatening chronic respiratory disease, and affects both pediatric and adult populations. It affects 24 million people in the United States, including more than 6 million children. (1) Asthma has social and economic impacts and causes three in five patients to limit their physical activity or miss days at school or work. Two primary air pollutants related asthma are ozone and particle pollution, that can originate from a number of sources, notably toxic waste disposal sites or traffic.

Research has long been focused on not just on sources of air pollution and the cause of asthma, but racial disparities and related environmental justice. A pivotal study in 1987, Toxic Waste and Race in the United States, established a strong correlation between race and the location of Toxic Waste facilities.(2) Subsequent research shows that this relationship still exists. As an example, Black people have a higher chance to visit the emergency department because of asthma. Furthermore, African Americans are three times more likely to die from asthma than white people in the United States.(3) In general, people of color breathe more particulate air pollution on average. Despite the constant public awareness of this issue, decades of environmental justice-related policies on both state and federal level, and sustained activism, Mascarenhas et al., amongst others, found a similar correlation in recent years as well.(4). This corroborates with findings reported by the World Health Organization which has reported that worldwide, people who live in overpopulated and developing countries experience the burden of outdoor (ambient) air pollution disproportionately with over 90 percent of the 4.2 million premature deaths in 2016 occurring in low- and middle-income countries of the South-East Asia, Central Africa and Western Pacific regions where the exposure is higher according to the WHO Ambient (Outdoor) Air Pollution. (5)

This paper seeks to build on this existing research by investigating further on the relationship between toxic waste, ethnicity / race, and asthma, using recent data from 2019. While the origin of air pollution can be from a multitude of sources, this study will limit its scope to toxic waste facility-related air pollution, as reported to the EPA. Our geographical focus will be the continuous mainland of the United States, and in some cases particular states to investigate the relationship in detail. To count and measure population, ethnic and racial minorities, and cases of asthma, data from the CDC and the American Census Bureau will be leveraged. By combining data from these

three sources, we hope to find more connections, or find local outliers. We will also investigate poverty as a related factor.

Our research question is thus:

Are Toxic Release facilities with high emissions more likely to be located in counties with larger non-white population?

To understand the underlying relationships better, we will also investigate the following:

Sub-question 1: Are Toxic Release facilities with high emissions more likely to be located in counties with larger non-white population?

Sub-question 2: Are there more asthma cases in the counties around the Toxic Release Facilities?

2. Data

This paper uses information from three sources: Toxic Release Inventory, CDC Places, and American Community Survey

1. Toxic Release Inventory

- The data is a publicly available data set from the United States Environmental Protection Agency containing self-reported toxic emissions from industrial facilities. The time of data set is from 2019, and covers the total pollution reported for that year, measured in pounds. The data contains a set of chemical pollutants that has been found harmful for human health, outlined by the EPA as mandatory to report based on federal laws. Both federal and industrial facilities report the data, and the EPA only consolidates this, it does not collect it based on any central measurement.
- To account for the temporal factor of air pollution (i.e., the fact that longer exposure might be needed to develop asthma, and that exposure decades ago might contribute to illness registered in 2019) we also used the list of already operating facilities from 1999 to subset the 2019 data. By this measure, we aimed to establish a list of facilities that have been operating over a longer period of time; these facilities have of a greater likelihood to have influence the health of a larger segment of the population.
- A key limitation of the TRI dataset is that it is based on self-reported data; a lack of reporting diligence could lead to emissions missing. A further limitation is that the dataset focuses on larger sources of toxic waste facilities; smaller industrial or business units are not part of the database but might still contribute to pollution-related adverse health effects.

2. CDC Places

- The CDC Places data was collected by the Centers for Disease Control and Prevention with the help of the Robert Wood Johnson Foundation and CDC foundation. The dataset estimates 29 measures, including health outcomes, preventive services use, chronic disease-related health risk behaviors, and health status in 3142 counties within 50 more states and Washington DC. The data not only has county level but also place, local area-level, census tract, and ZCTA level.
- To coordinate with the data from Toxic Release Inventory, we also focused on using the 2019 dataset that contains asthma cases for the population aged 18 or above.
- A key limitation of this data is that demographics, such as age, or exact address and severity of illness are not included. The smallest unit of analysis is census tract, which can in some cases cover a large geographical territory. It also relies on the American Community Survey for some information on the population and thus inherits its reliability issues.

3. American Community Survey, 5-year data, 2019

- Census Bureau provides two types of population-related data: decennial census, released every 10 years, and the American Community Survey, released every year using 1-year and 5-year estimates. We opted to use the 2015-2019 data contained in the American Community Survey 2019 5-year Estimates. This data relays information on 2019, but uses a method to include 5-year averages to improve reliability.⁽⁶⁾
- The American Community Survey (ACS) is an comprehensive source for the American population and housing information. It has over 3.5 million addresses in the annual sample, which is a significant subset of the population, though overall small compared to the entire population. While the sample is small compared to the decennial census, the frequency is high, and the data collected includes a large number of demographics information, including metrics such as poverty, income, mobility, that the decennial census does not measure.

- The data source was explored to identify best-fitting variables for identifying all ethnic and racial groups within the population. As the Census Bureau evaluates Hispanic as an ethnicity and other U.S. minorities such as Asians as races, we decided to use the variables in the ACS that first subset the population into Hispanic and non-Hispanic, and then separated the non-Hispanic into further races. Subsequently, when discussing ethnicity or race, when referring to any specific race, such as white or Black, it should be understood as non-Hispanic white or non-Hispanic Black.
- Beyond the non-Hispanic racial demographics, total population and population living below the poverty level was also accessed from the database to provide a basis for calculating percentages and an important control variable for our study.
- While the ACS data is available as granular as block-group level, to match the smallest available analysis unit in the CDC Places dataset, census tract was chosen as the unit of analysis.
- An important limitation of this data is due to the fact that it is based on a relatively small sample of the population, which can result in some unreliability. To counteract this, we used the 5-year estimate and not the 1-year estimate, and we also calculated percentages from the absolute value and did not use the population counts as the basis of our analysis.⁽⁷⁾

3. Methods

3.1. Data Acquisition.

To acquire the CDC Places and Toxic Release Inventory datasets, we downloaded them from their respective websites (cdc.gov and EPA.gov) and worked from local copies as this was the fastest method to acquire the data, and it could be subset easily afterwards.

To acquire the ACS census data, we used API call to already subset it to the fields relevant to us, as described in 2.3, using an API key. In order to achieve this, we first used the ACS documentation to identify the relevant fields, and then created a list called `vars_to_retrieve`.

To automate the process and make it easily reproducible in the future, all steps have been defined in functions, and outlined in a separate .py file called `CensusMethods`. The `CensusMethods` functionality is outlined as the following regarding data acquisition:

- Instantiates a `Census` object using an API key, passed to the constructor as a parameter
- Retrieves a table of all of the census variables
- Subsets the table of all census variables based on the passed `vars_to_retrieve` list, which contains all the relevant fields
- Wraps the `Census` library `acs5.state_county_tract` function, such that it can be used with a state abbreviation, rather than a FIPS code
- Retrieves census data for a list of states, initially used for the entirety of the US. While later the focus of the analysis was narrowed to certain states, we wanted to have the entire US in case anything specific was needed and for potential further analysis
- Retrieves, concatenates and maps to the EPSG 4326 coordinate system the census tract-level shape files

As a result, census is created from the relevant list of variables for all states.

3.2. Data cleaning and merging.

3.2.1. Initial cleaning and merging of the census data.

The ACS census data, stored in the dataframe we named `census`, was already subset for the analysis, but it needed to be merged with the compiled shape files and relabeled. To merge census with the shape files, exact match inner merge was used based on tract-level GEOID found in the tract shapes files. To create the corresponding GEOID in the census data table, state, county and tract values were concatenated (the entire process has been automated within the `CensusMethods` .py file). The merged resulted in `full_us`.

To re-label the `full_us` data, the descriptive table of all variables was used with `relevant_values` as a dictionary. To this table, shortened custom column labels were added so that it can be used as the dictionary output (this process could be more automated if the existing labels were used, but once the variable list became final, working with the full column names became inefficient in the output tables).

3.2.2. Cleaning and merging the CDC Places and TRI datasets.

Two automated and generalized processes were created to clean the CDC Places and TRI datasets:

- CDC.py enables subsetting the CDC dataset based on year and sickness type (in this case, 2019 and asthma) and subsetting to the relevant columns for the analysis. Furthermore, it calculates the absolute value of asthma cases from the Data_Value and TotalPopulation columns to ensure a mathematical basis for calculations even if the data is summarized on a different level (e.g., on a county level). It also prepares for merging the LocationID (i.e. GEOID) column by changing the type to string and padding it where necessary, as in some cases, the GEOID starts with a 0 and this 0 was missing from the CDC Places LocationID. New column named GEOID. Finally, the data was subset to the relevant columns, resulting in in cdc_analysis_merge
- TRI.py uses regex to remove the numbers and periods from the column names, then uses a function to subset the TRI dataset based on the TRIFD column. Then it subsets the data to the mainland USA - as TRI data also includes Alaska, Hawaii, Puerto Rico, and other territories - and filters and subsets for air pollution (both for stack air and fugitive air), and subsets to the largest polluters (top 10% of polluters). It also creates a tuple pair to represent the coordinates from the separate longitude and latitude, such that each location can create a shapely.geometry.Point type.

Finally, the cdc_analysis_merge and full_us datasets were merged using inner join on the GEOID as key based on exact match. Out of 72 359 census tracts in the list, after the merge, we could match 69 676 asthma values, so our match rate was above 96%. Asthma cases on the census tract level might be missing for various reasons. Most likely it is due to some census tracts not having any asthma values, or any estimates on asthma values, and thus missing from the original CDC dataset (the data only contains 70 338 rows of observations). Another explanation could be some mismatch between the GEOID and LocationID that we did not identify and correct.

3.3. Analysis.

As the starting point of the analysis, in addition to the merged CDC Places ACS dataset, us_cdc_census we used the TRI functionality on the TRI 1999 and 2019 datasets ensure we're looking at relevant sites and to create tri_2019_clean. We did this to account for the temporal limitations of our study by using the 1999 TRI database was used to subset the 2019 list of facilities. With this simplification, we're making an assumption that any TRI facility that was operating in 1999 has been operating over a long time horizon and is considered as a constant source of pollution, relevant for the entire population. 1999 was chosen as the control time as the last major expansion of the TRI database was in 1997 and we wanted to account for as many types of facilities as possible (8). In addition, state-specific files were created containing subsets of the census data as part of an automated map creating process.

To ensure statistical accuracy and due to data limitations, all files contain absolute values and not percentages. This ensures that files will be usable at any geographical aggregation level. But due to the inherent limitations of the ACS data, in all cases, these absolute values will be translated into percentages, as the percentages are more reliable indicators than the absolute counts.

To answer our research question, we were looking for a relationship between asthma, race and toxic air pollution. To visualize and quantify this relationship, we used a spatial, map-based method and also ran a regression analysis.

3.3.1. Spatial analysis.

We decided to use maps and visualization as one of our main research outputs due to the nature of the datasets(9)(10). While we had access to tract-level population data, we did not have very precise estimates on where each individual with asthma lives (e.g., zip-code or address-based data). This is a potential limitation as the results of air pollution might be measured on a smaller scale than this level of geographical aggregation allows. The exact distance to the source of pollution that is necessary to analyze the relationship is also unclear: for example, Chakraborty et al. notes research has ranged from 100 yards to 3 miles regarding effects of TRI facility pollution, but with the pollution plume potentially extending out up to 44 miles,(11) defining exactly which TRI facilities should be looked at for each census tract's asthma cases is questionable.

Thus, we decided to rather create state-level visualizations where the location of TRI facilities and the amount of pollution are easily identified, and where race and asthma are also visible due to color-coding. To simplify the output and make it easier to visually interpret, we used the .dissolve function from geopandas to aggregate the tracts on the county level.

To further refine the analysis and simplify the output, we defined the ethnic/racial demographics a subset of two separate groups:

- 'White' refers to census tracts (or counties) where more than 70% of the population is white. This information could be directly derived from the census data by dividing the absolute number of white population and the total population. As discussed in the data section, it refers to non-Hispanic white population only.
- 'Non-white' refers to census tracts of counties where more than 30% of the population is non-white. This was calculated from the census data by dividing the absolute number of white population and the total population, and subtracting this number from 1. This includes Hispanic population as well.

To simplify the calculation of the above percentages, Census Method contains functions to automate this process which can be used at every level of geographical aggregation. As the last step, we decided to subset the analysis into two states based on three criteria: 1) relatively high density of population, especially minority population (12) (13) 2) High density of toxic release facilities based on the TRI list of facilities and 3) historically larger ethnic/racial disparities. Thus, we selected North Carolina and Louisiana.

The output of the analysis is two separate tables, one containing the asthma cases as %, and the white non-white population as % linked to county shapes. The other table contains all TRI facilities, their coordinates and the total air pollution linked to them.

Below is a sample of the US-CDC merged dataframe, containing the asthma cases, white and nonwhite populations, and spatial data:

	STATEFP	COUNTYFP	white_percentage	non_white_percentage	asthma_percentage	asthma_non_white	asthma_white
	1834	37	001	0.640586	0.359414	0.093614	0.093614
	1835	37	003	0.867896	0.132104	0.096814	0.000000
	1836	37	005	0.868823	0.131177	0.101887	0.000000
	1837	37	007	0.445988	0.554012	0.114061	0.114061
	1838	37	009	0.925176	0.074824	0.095050	0.095050

This dataframe varies across states in terms of how many records it contains. The key fields here are the geometry, used to map out the counties; the white_percentage and non_white percentage, used to determine which colour heat map to use to represent the different demographics; and the asthma_percentage, which is what the heat map is based on.

The second of the two datasets used to generate our final maps for our geospatial analysis was the state level subset of our TRI data, as can be seen below:

	YEAR	STACK AIR	FUGITIVE AIR	geometry
	447	2019	40456.0	21958.00 POINT (-78.03138 35.25875)
	1828	2019	41831.0	1778.00 POINT (-78.90656 36.52123)
	1839	2019	19977.0	74.00 POINT (-78.90656 36.52123)
	1938	2019	23334.0	266.68 POINT (-81.94629 35.37798)
	2011	2019	570104.0	0.00 POINT (-79.04634 35.60067)

While we did not restrict our TRI dataset solely to these columns, as we had initial intentions of exploring other fields, this subset of columns contains the key relevant columns for our geospatial analysis are STACK AIR, FUGITIVE AIR, and geometry. These were used to compute the relative size of the points overlaying the heatmap, such that we could use marker size to represent the level of emissions from the facility from both chemicals released through stacks and fugitive emissions, as well as locate where to place them on the map.

To create the final maps, MapGenerator functionality was created in a separate .py file. Within MapGenerator, create_asthma_population_plot_for_state_county_with_race_and_facility function uses matplotlib mapping to layer on top of each other asthma cases split by white and non_white, and the place and emissions of TRI facilities. Using a for loop, a single loop creates as many separate state-level maps as specified earlier in the input field of string_format_list.

4.3.2. Regression.

Keeping the limitations on distance, type of air pollution, and not knowing the exact location of the individuals with asthma in mind, we decided to try and quantify the relationship between asthma, race and pollution using a simple binary indicator that determines if there is a toxic waste facility in the given census tract or not. This 0/1 indicator has been created by matching the coordinates of the TRI facilities to the tract geometry, returning 1 to indicate the presence of a facility within that geographical area and 0 otherwise. To limit the distance from the facility and try to understand better the effect of pollution on the nearby population, we used tract-level data, and recreated the white, non-white and asthma percentages as outlined with the mapping process. Furthermore, to introduce one more control variable, we included a variable on poverty as well (as % of population living below the poverty line in the given tract). All the variables take values from the 0 to 1 scale, with the toxic release facility indicator having values of either 0 or 1.

While using the county level subset of the data yielded more understandable visualizations, for our regression analysis, we chose to use the smaller, tract-level information, as it provides more detail.

	non_white_percentage	pov_percentage	tri_facility_in_polygon	asthma_percentage_on1
0	0.167679	0.091911	0	0.097
1	0.618722	0.396304	0	0.121
2	0.244013	0.143485	0	0.102
3	0.573722	0.307372	0	0.118
4	0.044048	0.073214	0	0.088

This subset of data contains four fields: the three independent variables we hypothesized to have an impact on asthma prevalence, as well as asthma prevalence as measured as a percentage of the population. The subset, being on the tract level with NA values excluded, contains 69, 537 rows.

Our regression model is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

$$\% \text{ of asthma cases} = \beta_0 + \beta_1 \% \text{ of nonwhite population} + \beta_2 : \% \text{ of population living below the poverty line} + \beta_3 \text{ presence of a toxic facility} + \epsilon$$

We used an Ordinary Least Square (OLS) regression to understand the effect of race, poverty and the presence of a toxic facility on asthma. The statsmodels.api and sklearn packages were used to evaluate the model.

4. RESULTS

4.1. Spatial analysis.

We can see the results of the spatial analysis for Louisiana on Figure 1 and the results for Georgia on Figure 2. Multiple relationships can be investigated using the maps created for our focus states:

1. Investigating racial disparities in toxic facility placement

Our research question was: Are Toxic Release facilities with high emissions more likely to be located in counties with larger non-white population?

In Louisiana, this seems to be the case - almost no TRI facilities are located in counties classified as white. In North Carolina, there still seem to be more facilities in non-white counties, but there is also a number of them in white counties. Investigating the individual towns or cities the facilities are located in might help further analyze the relationship, but the current we're using data does not allow this depth.

In summary, there seems to be some linkage between race and the placement of the TRI facilities.

2. Investigating the relationship between asthma and Toxic Release Facilities

Our research question was: Are there more asthma cases in the counties around the Toxic Release Facilities?

Asthma cases seem to be higher in most counties around TRI facilities, especially in Louisiana, but there are counties with smaller asthma populations where TRI facilities are located. Based on the visualization, it is hard to definitely tell if there is a strong relationship. This highlights the needs to run a regression to find potential correlations. Furthermore, there are several counties where there is a high percentage of individuals who have asthma, but there are no TRI facilities. Asthma can be caused by multiple factors beyond pollution from TRI facilities, such as traffic-related air pollution, that fall outside of the scope of this research. These counties might be located e.g, close to highly polluting traffic, or are close to other pollution sources.

3. Investigating racial disparities in asthma and toxic waste

Our research question was: Does race seem to be significant for Toxic Release Facility-related asthma cases, as far as we can establish a relationship based on visuals only?

Since there are more TRI facilities in non-white counties, and in multiple cases asthma seems to be more prevalent in counties around TRI facilities, race seems to be an indicator for TRI-related asthma cases. The relationship is more visible in North-Carolina, where there is both a higher number of TRI facilities, and more associated emission in the north-east part of the state, which also seem to be surrounded by higher percentage of asthma cases. On the other hand, this does not hold true in Louisiana, where a high density of TRI facilities in the southern / middle part of the state does not seem to correlate with asthma percentages. Overall, the results regarding the relationship of these three variables are inconclusive, and thus the results of the regression analysis might provide more insight.

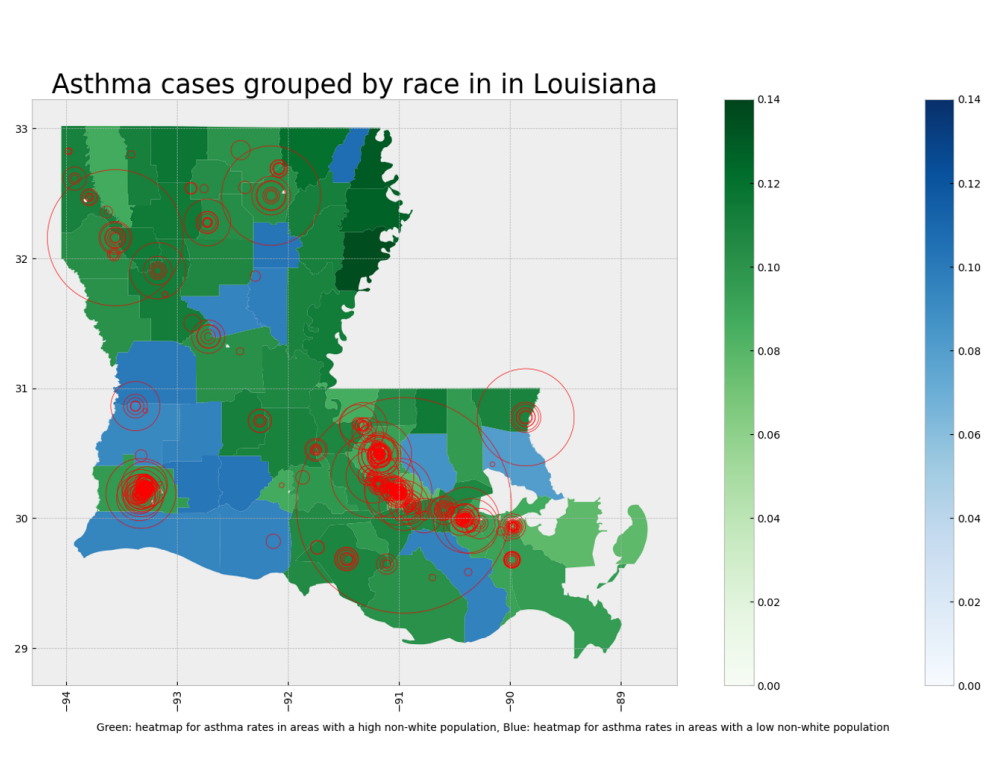


Fig. 1. Output map for analysis in Louisiana

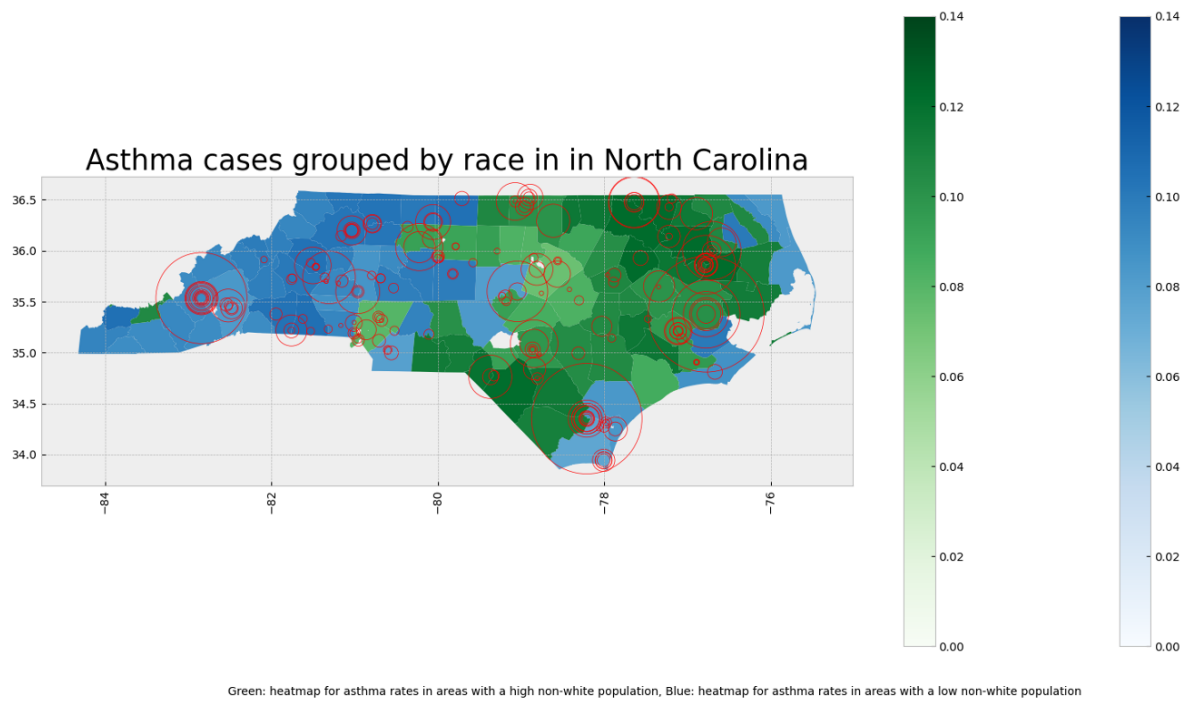


Fig. 2. Output map for analysis in North-Carolina

5.2. Regression analysis.

Table 1. Regression Results

Dep. Variable:	asthma_percentage_on1	R-squared:	0.462
Model:	OLS	Adj. R-squared:	0.462
Method:	Least Squares	F-statistic:	1.989e+04
Date:	Tue, 06 Dec 2022	Prob (F-statistic):	0.00
Time:	21:19:09	Log-Likelihood:	2.1232e+05
No. Observations:	69537	AIC:	-4.246e+05
Df Residuals:	69533	BIC:	-4.246e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	0.0857	7.86e-05	1089.686	0.000	0.086	0.086
non_white_percentage	-0.0067	0.000	-40.829	0.000	-0.007	-0.006
pov_percentage	0.0978	0.000	232.819	0.000	0.097	0.099
tri_facility_in_polygon	0.0016	0.000	6.397	0.000	0.001	0.002

Omnibus:	1715.313	Durbin-Watson:	0.763
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3660.433
Skew:	0.130	Prob(JB):	0.00
Kurtosis:	4.093	Cond. No.	10.8

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The regression results show a statistically significant relationship between the proportion of non-white population, the proportion of people living below the poverty line and the presence of a toxic release facility on the prevalence of asthma, at the unit of a census tract.

The presence of a toxic release facility in the census tract shows a positive association with an increase in asthma in the population. The proportion of population below the poverty level shows the strongest relationship with the prevalence of asthma. This points towards an uneven burden on the poor, who are more likely to be affected than more well-off sections of society. It raises concerns about environmental justice and equity where such impacts may pose additional stressors on existing social disparities and cleavages. As described earlier, the population was divided into white and non-white segments, where all non-white ethnic and racial minorities were consolidated together. While we expected a similar positive regression coefficient for our race variable - non white percentage - it shows a slight negative relationship. This might be due to some minorities being less predisposed to asthma, while others are more - for example, investigating the correlation solely for African American or Hispanic minorities might have a different outcome.

The above graph plots the effect of each variable on the dependent variable, asthma. This emphasizes the findings from reading the numbers of the regression table, as it highlights that there is a positive correlation between the poverty percentage and the asthma percentage, and a small negative correlation between the non white percentage and the asthma percentage.

5. DISCUSSION

5.1. Conclusion.

Our geospatial analysis and regression model helped make clear that different factors may be of more relevance in different locations, through demonstrating the difference between Louisiana and North Carolina in terms of the relative ratios of TRI facilities in white and non-white counties, or differing densities of facilities in white vs non-white areas.

From our regression model, we learned that poverty rates have a positive correlation with asthma rates, while rates of non-white populations have a smaller negative correlation. This provides insight into potential future areas to explore: what is the relationship between toxic release facilities and other forms of health condition? How would subsetting by income level and race affect these relationships? What impact does the density of TRI facilities have?

5.2. Limitations and next steps.

There are several limitations we must bear in mind that resulted from both the data sources we utilized, and the types of analysis we chose. Regarding the mapping process, one key limitation is looking at facilities at the state level. As there might be other nearby polluting facilities that impact the population in the area, looking at nearby states might provide additional insight. Similarly, the county-level aggregation, while helpful for understanding the output, increased our unit of analysis and

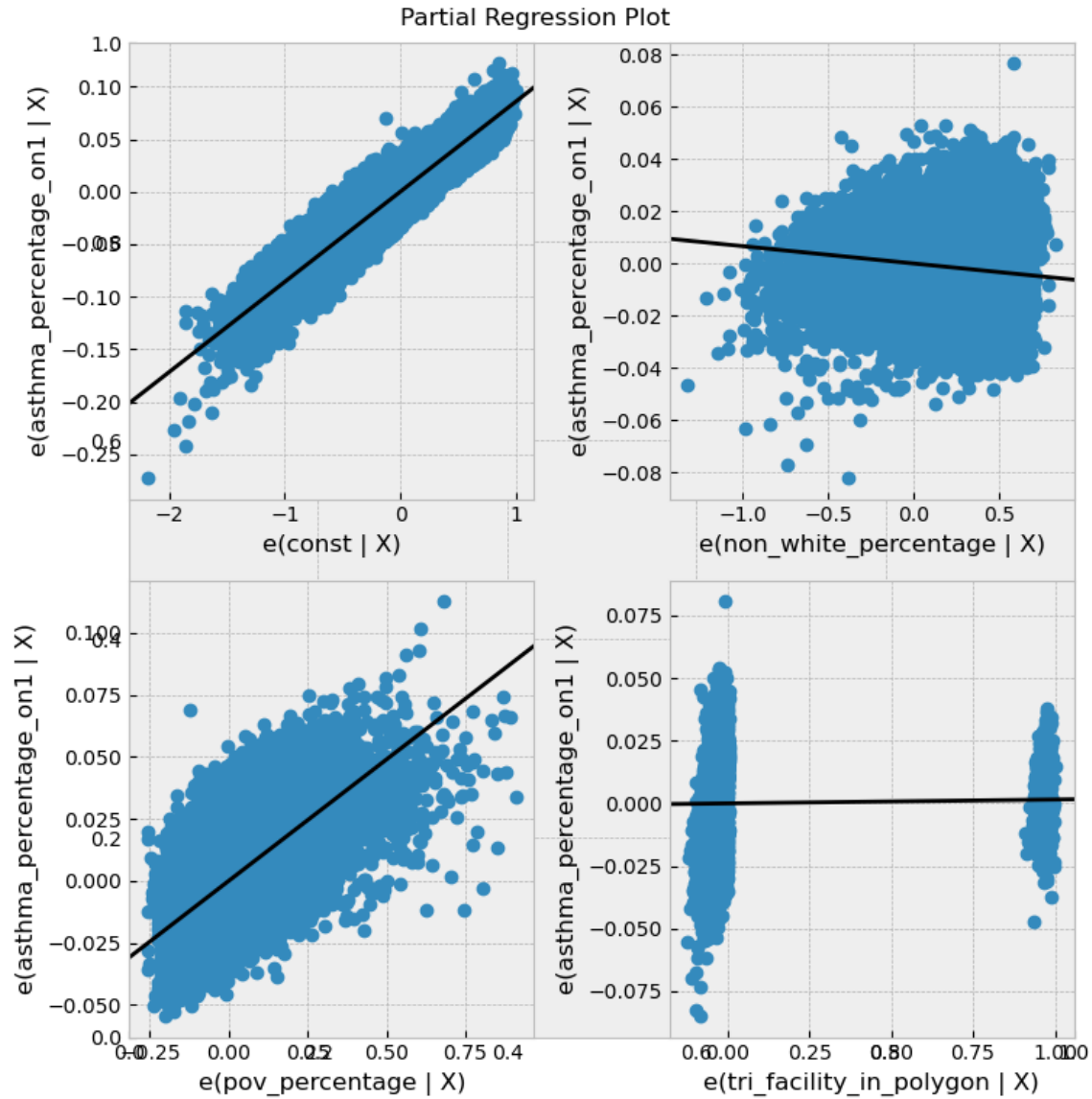


Fig. 3. Partial regression plot

thus aggregated results to a level where nuances are less visible. Adding further details to the map, such as traffic, population density, or elevation could help further refine the visualization and could help in interpreting the result better.

Regarding the regression, a simple binary classifier of whether or not a TRI facility exists in a tract is likely not the best measure of impact. This does not account for magnitude of the impact, as there may be instances where within a tract, multiple facilities exist. A potential alternative would be to count the number of TRI facilities present in each census tract. It may also be possible to do some further geospatial analysis to determine the maximum possible distance a person within a census tract can be from a TRI facility.

While we chose to use a binary classifier in hopes of accounting for our lack of knowledge of the locations of people with asthma within their census tracts, rather than measuring from, for instance, the center of the census tract, this has negative repercussions on the accuracy of measuring where facilities are likely to have an impact. As facilities are located near tracts other than the one they are in, they can be closer to adjacent tracts than to various points within their own, with an adverse impact on people in other locations. With this in mind, weighting facilities by distance from the centers of tracts, therefore accounting for facilities in neighboring tracts when analyzing a tract's asthma rates, may be a step towards improving our model.

In addition to our issues with geolocation of TRI facilities and the individuals with asthma, we also may need to reconsider our initial data processing steps. We chose to limit to the TRI facilities with the highest emissions for multiple reasons. The

full TRI dataset is extremely lengthy. There are nearly 80, 000 records in 2019. This makes it extremely difficult to visualize, on top of the fact smaller facilities are less useful for our research question of emissions have an impact on asthma rates. While this was a reasonable approach for our initial analysis, in future steps, it may be wise to use a larger subset of TRI data, as we are currently not accounting for the density of facilities in our analysis. Similarly, our equal weighting of stack emissions and fugitive emissions may not have been the best approach. Stack emissions and fugitive emissions are released at different intervals and may have different concentrations, implying they have different impacts on health. Therefore, it would have potentially led to better data to research this more and weigh them differently to derive our subset.

Finally, there are limitations to the data itself, outside of our wrangling. Our data regarding asthma patients is aggregate data; therefore, we do not have information regarding fields like resident ages. This matters because emissions from a TRI facility will have different impacts on residents of the same census tract, as residents have different lengths of exposure to those emissions thanks to age and movement. While the law of averages means that deviation within a census tract should balance out, this has implications for cross tract consistency. Since some areas of the country typically have less movement than others (eg: urban vs rural), it becomes difficult to isolate true correlations when comparing tracts to each other.

Regarding the American Census data, there are also some inherent limitations due to the nature of the survey itself. The ACS data collection is based on a relatively small sample of the population, collected at more frequent intervals. This leads to lower reliability. While we used percentages instead of counts to limit this effect, in some cases the results might be unreliable due to an inherent issue with the ACS count itself, which was used to calculate asthma, poverty, and ethnic/racial percentages as well.

Using the already existing code, this research could be taken further in several ways. Limitations regarding the regression could be addressed - e.g., a distance-based model could be devised, using the complete list of facilities and their distance from the center of the tract polygon. The visualization could be also adapted to include more details, and further granularity could be explored, e.g., tract-level maps. The rest of the states could be investigated and results could be compared on a national level. Finally, an important next step would be investigating not just white / non-white demographics, but all minorities, as some communities might more more affected than others. A race-poverty interaction model could also be used to explore some further potential connections.

1. Centers for Disease Control, Prevention, Be alert about asthma (2020).
2. Commission for Racial Justice, Toxic wastes and race in the united states: A national report on the racial and socio-economic characteristics of communities with hazardous waste sites. (1987).
3. A Biggers, What to know about asthma in african americans (2021).
4. BJH Neal Fann, Charles M. Fulcher, The influence of location, source, and emission type in estimates of the human health benefits of reducing a ton of air pollution. (2009).
5. WH O, Ambient (outdoor) air pollution (2021).
6. the U.S. Census Bureau, Understanding and using acs single- year and multiyear estimates (2018).
7. A Attia, Pitfalls of the american community survey (2016).
8. EPA Website, Toxic release inventory program history (2022).
9. the Statista, Population density in the u.s. by federal states including the district of columbia in 2021 (2021).
10. the U.S. Census Bureau, Hispanic or latino origin by race (2020).
11. BJ Chakraborty J., Maantay J.A., Disproportionate proximity to environmental health hazards: methods, models, and measurement. *Am J Public Heal.* **101 Suppl 1**, S27–36 (2011).
12. World P opulation Review, Us states by race 2022 (2022).
13. World P opulation Review, United states by density 2022 (2022).