



# **Understanding the Link between Environmental and Social Determinants of Health: A Deep Dive into Indicators**

Stanford University, Fall  
2018

Sneha Ayyagari and Rubi Rodriguez

---

## Table of Contents

1. Abstract	2
2. Literature Review	3
2.1 How has environmental justice been defined by the literature, and how does this shape current thoughts on health and environmental inequities?	3
2.2 What are some environmental justice issues unique to the Bay Area?	4
2.3. What are some metrics that are used to assess environmental justice and exposure, and what are some of their limitations and strengths?	5
2.3.1 CalEnviroScreen	5
2.3.2 EJSCREEN	7
2.3.3 CDC Social Vulnerability Index	7
2.3.4 CDC 500 Cities Health Data	8
3. Research Design	9
3.1 Overview of Methodology	9
3.2 Data Collection and Cleaning	10
3.3 Exploratory Data Analysis	11
3.4 Confirmatory Data Analysis	12
3.5 Spatial Analysis	15
4. Results	16
4.1 General results of Analysis	16
4.2 Analysis of Richmond, San Jose, and Morgan Hill, California	17
5. Conclusions and Future Work	22
5.1 Conclusions and Limitations	22
5.2 Future Work	22

# 1. Abstract

The Sustainable Urban Systems framework centers on using data and structuring inquiry to generate actionable results. We were interested in further exploring social and environmental metrics that have been influential in discussions about environmental justice and how those might be operationalized within the Bay Area. More specifically, created an index that can be used to understand the spatial disparities of exposure to pollutants across various groups in the Bay Area. We analyzed data that was used to develop existing metrics including CalEnviroScreen 3.0, the Social Vulnerability Index, and data from the 500 Cities Project.

We ran our own regression analysis on the variables that are inputs to each of these three categories to understand which environmental, health, and demographic variables are correlated. We cleaned the data and conducted an exploratory and confirmatory data analysis to understand the statistical significance of various environmental and social factors on health. We developed a new index using 13 variables with a P Value lower than 0.05. We created a map displaying this new health index and we also created visual representations of two of the contributing factors to our new index: median income and number of toxic waste releases. The data was obtained from Simply Analytics and the CalEnviroScreen 3.0 which used data from the Environmental Protection Agency's Toxics Release Program.<sup>1</sup> We created a dashboard to display the results. We then created a storyboard to describe the process and the project in an interactive way. Our preliminary findings support the body of literature that factors such as exposure to toxics and income affect health impacts.

We found that in general, areas with higher median income corresponded with areas of a lower health index. This confirmed our hypothesis because it is likely that people with higher incomes have access to healthcare and healthier environments which may contribute to fewer days of poor physical health. While toxic waste sites and releases did correspond to a higher incidence of poor health especially in areas such as Richmond, the trend was not as clear across the geography. For instance, near San Jose, there are several facilities documented Toxic Release Inventory, however, this does not necessarily correspond to areas of higher health burden. One potential explanation is that median household income has a greater influence than exposure to Toxic Releases. Future work could include a predictive analytics component to expand upon our exploratory observations and findings. Finally, targeting our model to a specific audience and geography can make our indicator an even more powerful tool for policymakers and stakeholders as they design and implement more effective environmental and social programs and interventions.

---

<sup>1</sup> "Environmental Protection Agency's Toxic Release Program," EPA, 2017 Available at: <https://www.epa.gov/toxics-release-inventory-tri-program>

## 2. Literature Review

Our literature review addresses three questions:

### 2.1 How has environmental justice been defined by the literature, and how does this shape current thoughts on health and environmental inequities?

As defined by the Environmental Protection Agency (EPA), environmental justice is the “fair treatment and meaningful involvement of all people regardless of race, color, national origin, or income, with respect to the development, implementation, and enforcement of environmental laws, regulations, and policies.”<sup>2</sup> Often communities of color and low income people are disproportionately affected by environmental and health issues.<sup>3</sup> Mohai and Saha conducted a thorough literature review that shows that health disparities can be traced to the siting of hazardous waste sites and industrial facilities near areas where minorities and people with low incomes live as well as demographic changes that have occurred after these facilities have been sited.<sup>4</sup> There are very few longitudinal environmental justice studies. Many environmental justice studies and data is collected at a point in time. While these collection techniques can provide helpful insights, this is inadequate to explain nuanced changes in community health outcomes.<sup>5</sup> We were mindful that the data underlying several environmental justice indicators is not longitudinal, and hence this limits the generalizability of the results.

California is a leading state on operationalizing theories of environmental justice into statute. The CalEPA Environmental Justice Task Force develops initiatives in communities that are disproportionately affected by pollution and high rates of health concerns.<sup>6</sup> For instance, the Task Force provided enforcement work and compliance assistance workshops to local governmental authorities in Oakland to address problems identified during community meetings. In particular,

---

<sup>2</sup> “EPA Environmental Justice FY2017 Progress Report,” United States Environmental Protection Agency, January 2018, *Available at:* [https://www.epa.gov/sites/production/files/2018-04/documents/usepa\\_fy17\\_environmental\\_justice\\_progress\\_report.pdf](https://www.epa.gov/sites/production/files/2018-04/documents/usepa_fy17_environmental_justice_progress_report.pdf)

<sup>3</sup> “Environmental Justice Case Study: West County Toxics Coalition and the Chevron Refinery,” *Available at* <http://www.umich.edu/~snre492/sherman.html>

<sup>4</sup> Paul Mohai and Robin Saha, “Which came first, people or pollution? A review of theory and practice from longitudinal environmental justice studies,” *Environmental Research Letters*, Volume 10, Number 12, December 2015, *Available at:* <http://iopscience.iop.org/article/10.1088/1748-9326/10/12/125011/meta>

<sup>5</sup> Paul Mohai and Robert Saha, “Reassessing racial and socioeconomic disparities in environmental justice research,” *Demography*, 2006, *Available at:* <https://link.springer.com/article/10.1353/dem.2006.0017>

<sup>6</sup> “CalEPA Environmental Justice Compliance and Enforcement Working Group,” California Environmental Protection Agency, June 2013, *Available at:* <https://calepa.ca.gov/wp-content/uploads/sites/6/2016/10/EnvJustice-IWG-2013yr-PolicyMemo.pdf>

the Task Force provided assistance on addressing air pollution, reducing exposure to toxic jewelry products, and excessive solid waste being stored near vulnerable populations.<sup>7</sup> Several state organizations including the Attorney General's office are tasked with rulemaking and environmental law enforcement related to environmental justice.<sup>8</sup>

## 2.2 What are some environmental justice issues unique to the Bay Area?

The Bay Area has experienced tremendous economic growth over the past decade. As of 2017, the Bay Area had a gross domestic product (GDP) of \$748. If it were a country, it would have the 19th largest economy in the world. Between 2014 and 2017, the GDP grew by 4.3 percent per year.<sup>9</sup> However, California is lagging in meeting its nation-leading environmental and social goals, and there is a large disparity in social and health outcomes. The California Air Resource Board determined that the state is behind the 5.2 percent of reduction in greenhouse gas emissions needed to meet its 2050 goals. One of the many challenges is that there is not adequate transit-oriented development, which could reduce vehicle miles traveled by 20-40 percent and reduce transportation emissions by 9-15 percent by 2050.<sup>10</sup> Housing and access to amenities such as transportation is a contributor to social inequality and should be part of understanding environmental justice from a holistic perspective.

However, policies intended to reduce carbon emissions can also have unintended impacts. For instance, the California's cap-and-trade carbon trading scheme could enable facilities that release large amounts of pollutants to continue to operate in locations that primarily affect low income communities. This is because it is easy to allocate allowances in disadvantaged areas. Essentially, facilities such as refineries could continue to place their operations in low income communities and could buy offsets to meet cap-and-trade requirements whereas they would not be able to do this in wealthier areas. In addition to carbon dioxide emissions, this could also increase concentrations of air pollutants including PM 2.5, PM 10, total PM, SO<sub>x</sub>, NO<sub>x</sub>, and CO that negatively impact human health.<sup>11</sup>

---

<sup>7</sup> "Oakland Initiative Report," California Environmental Protection Agency, 2017, Available at [https://calepa.ca.gov/wp-content/uploads/sites/6/2018/03/OAKEJ\\_initiative\\_FINALweb.pdf](https://calepa.ca.gov/wp-content/uploads/sites/6/2018/03/OAKEJ_initiative_FINALweb.pdf)

<sup>8</sup> "Environmental Justice & Healthy Communities," Office of the Attorney General, California Department of Justice, Available at: <https://oag.ca.gov/environment/communities>

<sup>9</sup> "Continuing Growth and Unparalleled Innovation: Bay Area Economic Profile," Bay Area Council Economic Institute, July 2018, Available at: <http://www.bayareaeconomy.org/files/pdf/BayAreaEconomicProfile2018Web.pdf>

<sup>10</sup> California Environmental Protection Agency, Air Resources Board, First Update to the Climate Change Scoping Plan, May 2014, [https://www.arb.ca.gov/cc/scopingplan/2013\\_update/first\\_update\\_climate\\_change\\_scoping\\_plan.pdf](https://www.arb.ca.gov/cc/scopingplan/2013_update/first_update_climate_change_scoping_plan.pdf). Estimate assumes a reduction of 11.4 million metric tons of CO<sub>2</sub> per year from 2020 to 2050 is necessary for California to meet its 2050 goals.

<sup>11</sup> Christa M. Anderson, Kendall A. Kissel, Christopher B. Field, Katharine J. Mach, "Climate Change Mitigation, Air Pollution, and Environmental Justice in California," Environmental Science and

In addition to cap-and-trade issues, toxics spills and industrial accidents have contributed to environmental justice and health concerns. In Contra Costa County alone, the Chevron facility had 304 accidents between 1989 and 1995 including spills, leaks, toxic gas releases, and air contamination.<sup>12</sup> In 2013, Richmond sued Chevron over a crude-oil pipeline leak in August 2012 that led 15,000 people in Richmond to be hospitalized for respiratory problems.<sup>13</sup>

## 2.3. What are some metrics that are used to assess environmental justice and exposure, and what are some of their limitations and strengths?

### 2.3.1 CalEnviroScreen

The CalEnviroScreen was developed at the University of California at Berkeley and the Office of Environmental Health Hazard Assessment at the California Environmental Protection Agency to attempt to quantify environmental health disparities. The initial CalEnviroScreen 1.1 consisted of 11 environmental indicators including concentration of ozone, PM 2.5 exposure, Diesel PM exposure, Pesticide Use, Toxic Releases, Traffic density, Cleanup sites weighted by site type and status, potential contamination sources and monitoring wells, Permitted hazardous waste facilities and generators, impaired water bodies, solid waste.

CalEnviroScreen 1.1 also explored 6 dimensions of demographic data to identify populations who were particularly susceptible to pollution including percentage of children and elderly, asthma visits per 10,000 people, low birth weight percentage, educational attainment percentage measured in % of population aged >25 years with less than a high school education, linguistic isolation measured in percent of the population aged >14 years where nobody speaks English very well. Most of this data was derived from Census Bureau's American Community Survey 5 year estimates, and estimates from California environmental agencies such as California Air Resources Board. Each zip code in California was assessed using each of these 17 indicators and the average of the percentile scores were divided by 10 to derive pollution burden and population vulnerability scores. The analysis showed that counties in the San Joaquin Valley and Southern California had the greatest proportion of communities that were affected. It also found

---

Technology, 2018, 52, 18, 10829-10838, available at:  
<https://pubs.acs.org/doi/abs/10.1021/acs.est.8b00908>

<sup>12</sup> Scott Sherman. "Environmental Justice Case Study: West County Toxics Coalition and the Chevron Refinery," University of Michigan, Available at: <http://www.umich.edu/~snre492/sherman.html>

<sup>13</sup> Henry K. Lee. "Richmond sues Chevron over refinery fire," San Francisco Gate, August 2013, available at: <https://www.sfgate.com/bayarea/article/Richmond-sues-Chevron-over-refinery-fire-4703370.php>

that the median cumulative impact score was 75 percent higher for Hispanics and 67 percent higher for African Americans than for non-Hispanic whites.<sup>14</sup>

In 2017, CalEnviroScreen was updated to version 3.0 which includes two new indicators reflecting health and socioeconomic vulnerability to pollution and removed an analysis of age. These indicators included measure of rate of emergency department visits for heart attack and high cost of living.<sup>15</sup> The weighting of each category was adjusted to more accurately quantify exposure, environmental effects, socioeconomic factors, and sensitive populations. This indicator was used to inform SB 535, a bill proposed by Senator DeLeon in 2012. This bill requires that at least 25 percent of the funds generated by the Global Warming Solutions Act of 2006 (AB 32) be invested in projects that benefit disadvantaged community with at least 10 percent of the funds going to projects located within the communities.<sup>16</sup>

Despite its large political and environmental impact, there are still several critics of the CalEnviroScreen. Firms representing project developers argue that CalEnviroScreen is not adequate in assessing “cumulative impacts” as defined by the California Environmental Quality Act (CEQA). While the CalEnviroScreen does provide the caveat that the tool is not meant as a substitute for individual project analysis under CEQA, some firms claim that there have been instances where the tool may have been misused.<sup>17</sup> CalEnviroScreen 3.0 explicitly states that this tool should not be used as a substitute for a more thorough review following the CEQA.

Others argue that the weights assigned to the socioeconomic factors are skewed towards urban populations and rural communities may not qualify for funding. For instance, the Rural County Representatives of California point out that pollution sources such as black carbon are left out of the analysis. Wildfire emissions account for more than half of California's black carbon emissions, and rural communities are the most impacted by this. For example, using the methodology of the CalEnviroScreen 3.0, Lake County, an area affected by wildfires in not eligible for benefits under SB 535 but suffers from many environmental health issues.<sup>18</sup>

---

<sup>14</sup> Lara Cushing, John Faust, Laura Meehan August, Rose Cendak, Walker Wieland, George Alexeeff, “Racial/Ethnic Disparities in Cumulative Environmental Health Impacts in California: Evidence From a Statewide Environmental Justice Screening Tool (CalEnviroScreen 1.1)”, *American Journal of Public Health* 105, no. 11 (November 1, 2015): pp. 2341-2348.

<sup>15</sup> Matthew Rodriguez, Lauren Zeise, “Update to the California Communities Environmental Health Screening Tool,” California Office of Environmental Health Hazard Assessment, January 2017, Available at: <https://oehha.ca.gov/media/downloads/calenviroscreen/report/ces3report.pdf>

<sup>16</sup> “Senate Bill No.535 Chapter 380,” California Legislative Information, September 2012, Available at: [https://leginfo.ca.gov/faces/billNavClient.xhtml?bill\\_id=201120120SB535](https://leginfo.ca.gov/faces/billNavClient.xhtml?bill_id=201120120SB535)

<sup>17</sup> Norman F. Carlin, Kevin Ashe, “CalEnviroScreen 3.0 - Still the Wrong Tool for CEQA Review,” Pillsbury Winthrop Shaw Pittman LLP, January 2017, Available at: <https://www.pillsburylaw.com/en/news-and-insights/cal-enviro-screen-still-the-wrong-tool-for-ceqa-review.html>

<sup>18</sup> Patricia Megason, “Updated Tool Continues to Ignore California’s Most Disadvantaged Communities,” Rural County Representatives of California, October 2016, Available at: <http://www.publicceo.com/2016/10/updated-tool-continues-to-ignore-californias-most-disadvantaged-communities/>

### 2.3.2 EJSCREEN

The Environmental Protection Agency (EPA) developed EJSCREEN in response to presidential Executive Order 12898. EJSCREEN combines demographic indicators with environmental indicators. The 11 environmental indicators include National Scale Air Toxics Assessment (NATA) air toxics cancer risk, NATA respiratory hazard index, NATA diesel PM exposure data, particulate matter concentration, ozone concentration, traffic proximity and volume, lead paint indicator, proximity to risk management plan sites, proximity to hazardous waste sites, and proximity to National Priorities List sites, and the wastewater dischargers indicator. The environmental indicators are analyzed on a national scale with screening level data available at the block group level. EJSCREEN uses demographic information predict susceptibility to exposures to various environmental contaminants. While this is a helpful tool in identifying disparities in exposure, it does not explain the root causes that lead to these disparities.<sup>19</sup> Unlike CalEnviroScreen 3.0, the EJSCREEN does not attempt to quantify causal links associated with pollutants and demographics. EJSCREEN is an important part of the EPA's EJ 2020 measure.

Environmental groups, trade groups, activists, city regulators and state regulators filed comments on the EPA's EJ 2020 initiative suggesting that while the tool provides some useful data, it is more symbolic than action oriented. Commenters demanded more details and called for the EPA to translate findings from tools such as EJSCREEN to extend access to resources that can mitigate some of the health and environmental disparities that are uncovered.<sup>20</sup> While the index contains helpful exposure and environmental data, we decided to focus on using data from the CalEnviroScreen 3.0 because it provides a more granular and comprehensive data set for California and the Bay Area more specifically.

### 2.3.3 CDC Social Vulnerability Index

The Social Vulnerability Index (SVI) was developed by the Geospatial Research, Analysis & Services Program (GRASP) on behalf of the Center for Disease Control (CDC). The goal is to determine social vulnerability of every Census tract across four themes: socioeconomic status,

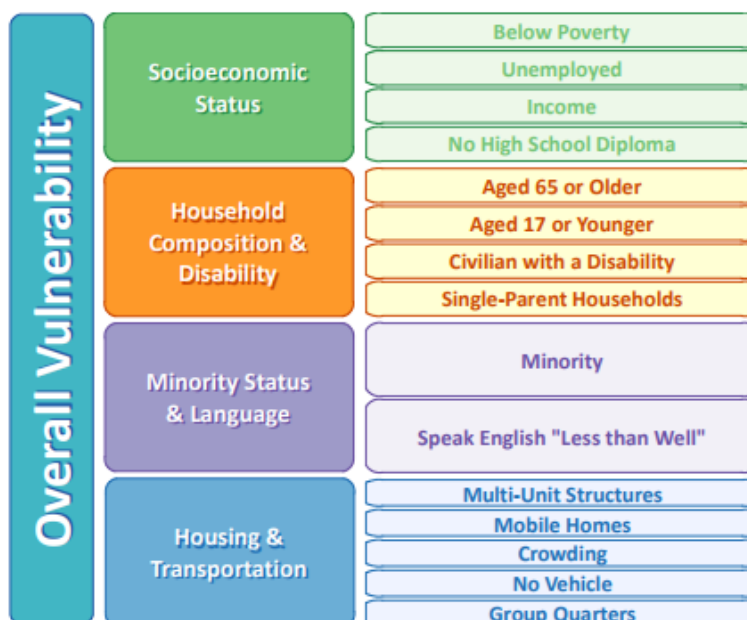
---

<sup>19</sup> "EJSCREEN Technical Information," U.S. Environmental Protection Agency (EPA), 2015, available at: [https://www.epa.gov/sites/production/files/2015-05/documents/ejscreen\\_technical\\_document\\_20150505.pdf#page=13](https://www.epa.gov/sites/production/files/2015-05/documents/ejscreen_technical_document_20150505.pdf#page=13)

<sup>20</sup> Kristen Lombardi, Talia Buford, "EPA draft plan would perpetuate environmental racism, critics say," The Center for Public Integrity, (2015), available at: <https://www.publicintegrity.org/2015/10/09/18269/epa-draft-plan-would-perpetuate-environmental-racism-critics-say>



household composition, race/ethnicity/language, and housing/transportation.<sup>21</sup> Several parameters including the ones displayed in Figure 1 were grouped into 4 main themes.



**Figure 1: the Social Vulnerability Index is Grouped Across 4 Themes**

The SVI contains many demographic indicators which have been used in aggregate to allocate resources in natural disasters. For instance, this index has been used to target aid and map vulnerability to fires in Los Angeles California.<sup>22</sup> While this indicator provides robust information about social factors, it does not provide insight into the environmental underpinnings of social conditions such as housing and transportation.

### 2.3.4 CDC 500 Cities Health Data

The 500 Cities Project data was collected by the Center for Disease Control and Prevention, Division of Population Health, Epidemiology and Surveillance Branch in partnership with the Robert Wood Johnson Foundation and CDC Foundation. Nationally this represents 28,000 census tracts that range in population from less than 50 to 28,690 people. In California, this includes tracts in San Jose, San Francisco, Oakland, Fremont, Hayward, Sunnyvale, Santa Clara, Berkeley, Richmond, Daly City, San Mateo, Napa, Union Cities, Mountain View, and Milpitas. This

<sup>21</sup> "The Social Vulnerability Index," Agency for Toxic Substances and Disease Registry, Available at <https://svi.cdc.gov/>

<sup>22</sup> Evan Lue, John P. Wilson, "Mapping fires and American Red Cross aid using demographic indicators of vulnerability," May 2016, Available at: <http://johnwilson.usc.edu/wp-content/uploads/2014/12/Mapping-fires-and-American-Red-Cross-aid-using-demographic-indicators-of-vulnerability.pdf>

data set provides information about preventative health and chronic conditions for the largest 500 cities in the United States at the census tract level. We were especially interested in a filtered dataset released by the CDC in 2016 that estimates the proportion of the population by census tract for which physical health is not good for greater than or equal to 14 days among adults. This estimate is derived a study that analyzed measures of chronic disease related to healthy behaviors, health outcomes, and use of preventative services.<sup>23</sup> We selected this metric because it may be more comprehensive than considering any one health condition in isolation.

## 3. Research Design

### 3.1 Overview of Methodology

The major goals of the study are to:

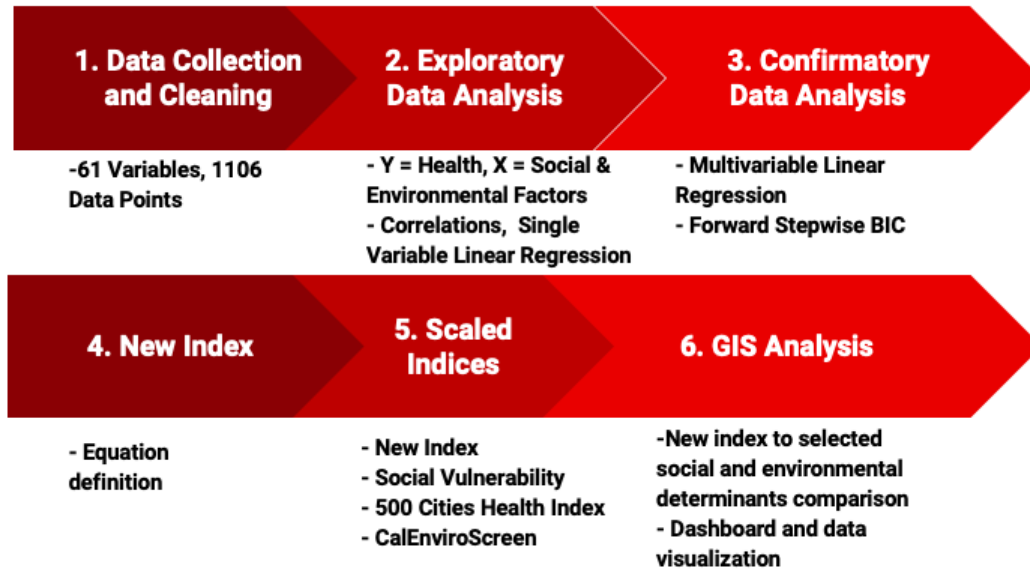
- Understand the limitations of each existing indicator
- Understand which social and environmental factors considered in each of these existing indices correlate to health and understand their statistical significance
- Understand the spatial distribution of this index across the Bay Area

To achieve these objectives, we analyzed data from the 500 Cities health dataset, environmental factors from CalEnviroScreen 3.0, and demographic information from the Social Vulnerability Index as summarized in Figure 2. We cleaned the data and conducted an exploratory and confirmatory data analysis to understand the statistical significance of various environmental and social factors on health. We developed a new index using 13 variables with a P Value greater than 0.05. We created a map displaying this new health index and we also created visual representations of two of the contributing factors to our new index: median income and number of toxic waste releases. The data was obtained from Simply Analytics and the CalEnviroScreen 3.0 which used data from the Environmental Protection Agency's Toxics Release Program.<sup>24</sup> We created a dashboard to display the results. We then created a storyboard to describe the process and the project in an interactive way.

---

<sup>23</sup> "500 Cities: Physical health not good for  $\geq 14$  days among adults," 2018 Available at: <https://chronicdata.cdc.gov/500-Cities/500-Cities-Physical-health-not-good-for-14-days-am/5w59-igre>




<sup>24</sup> "Environmental Protection Agency's Toxic Release Program," EPA, 2017 Available at: <https://www.epa.gov/toxics-release-inventory-tri-program>



**Figure 2: Overview of Project Scope**

## 3.2 Data Collection and Cleaning

Data for the state of California at a census tract level was collected from three main sources: CalEnviroScreen 3.0, Social Vulnerability Index and CDC 500 Cities Health Indicator. Each indicator had an extensive dataset, and we selected portions of these datasets to use for our analysis. For instance, CalEnviroScreen 3.0 provided data not only for environmental variables but also socioeconomic ones. We decided not to use the demographic factors because some of these overlapped with the more comprehensive socioeconomic information in the Social Vulnerability Index dataset. Therefore, we decided to use only the environmentally related data from CalEnviroScreen 3.0 and use the sociodemographic data from the Social Vulnerability Index. We also filtered these data sets by the nine counties in the Bay Area and deleted some rows that were missing information. We considered predicting the missing data by using a trained machine learning model, but we did not have enough data per county to predict this accurately. Furthermore, we decided that since each county has such a different and social environmental context, this approach to predicting outcomes may not yield meaningful results. Therefore, we decided instead to exclude tracts that were missing data from the analysis instead. Table 1 shows the data sources as well as the size of our datasets (m= data points, n= variables) before and after cleaning them. The final size of our data set that was used for our analysis was 1106 data points which represented a total of 61 variables: 40 socioeconomic, 15 environmental, 3 health and 3 geographic related.

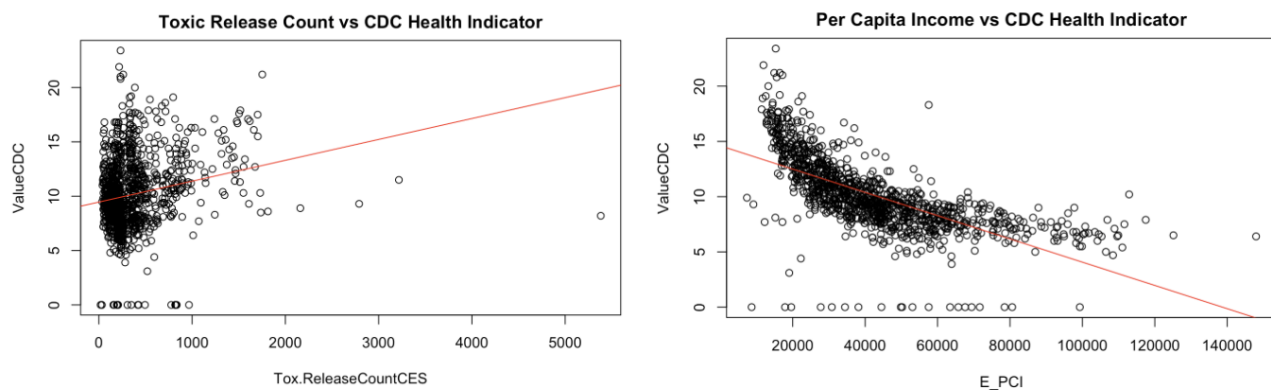
	Original DataSet (mxn)	Cleaned DataSet (mxn)
	1583 x 125	1578 x 40
	1582 x 57	1582 x 15
	1106 x 22	1106 x 6

**Final Dataset: 1106 x 61**

**Table 1: Data sets before and after cleaning process**

### 3.3 Exploratory Data Analysis

In order to start understanding our data, we performed some Exploratory Data Analysis that included single variable regression across different variables using the data from the CalEnviroScreen 3.0, Social Vulnerability Index, and CDC 500 Cities Health Indicator (CDC Health Indicator). We scaled the indices to compare how each changes across the various census tracts that we studied. For example, as shown in Figure 3, we found that while for variables such as income there is a clear trend and higher correlation to the health indicator from the 500 Cities data, there are others such as Toxic Release where correlation is not that evident.



**Figure 3: Exploratory data analysis shows higher correlation between income than toxic release count**

The exploratory data analysis process also led us to believe that we might have some variables out of our 61 that were highly correlated to each other which could inadvertently add extra noise to the model. Therefore, we created a flattened correlation matrix of 1830 pairs across all the

variables and deleted variables with a correlation coefficient more than 0.9 that were similar in their definitions. For instance, high confidence and low confidence factors for CDC health indicator were removed because we already had the actual value for that score and including the high and low confidence factors would not add to our conclusion. Another pair that was highly correlated by -0.92 was Housing Units and PM 2.5. In this case, we decided to keep both variables even though their correlation is high. Keeping mind that correlation is not necessarily causation, we decided that it would be useful to keep both variables since both could inform distinct and potentially interesting insights. A similar analysis was performed for all the 9 variables for which correlation higher than 0.9. We also removed geographic factors such as county name, latitude and longitude since we were more interested on the influence of only environmental and socioeconomic factors. This analysis helped us to further clean our data and prepare it for the multivariable regression model.<sup>25</sup>

### 3.4 Confirmatory Data Analysis

After finishing the data cleaning and exploratory data analysis, we created a multivariable regression model. We used 47 variables that represented environmental and socioeconomic factors. The main objective was to identify the most statistically significant variables that would explain the health metric from the 500 Cities dataset. We ran a model with all 47 variables and found that the model had an  $R^2$  value of 0.803. Table 2 shows the 13 variables we identified with a p-value lower than 0.05. That means that these variables were the most significant in explaining the CDC health index.

	Estimate	Pr(> t )	
(Intercept)	6.798	3.00E-08	***
Population Count	4.44E-04	< 2e-16	***
Per Capita Income	-2.92E-05	7.06E-09	***
% People Under Poverty Line	1.18E-01	1.07E-06	***
% People >25 with no high school diploma	1.33E-01	3.00E-06	***
Solid Waste	-5.27E-02	6.43E-05	***
Tox. Release Count	5.31E-04	0.000474	***
Pesticides Count	-5.15E-03	0.000581	***
People in institutionalized groups	-6.40E-02	0.00521	**
Disability	9.79E-02	0.01789	*
PM2.5	-1.53E-01	0.024878	*
AGE17	6.60E-02	0.02666	*
People below poverty level	-1.05E-03	0.033655	*
Diesel PM Count	8.00E-03	0.034945	*

**Table 2: Most significant variables to explain CDC Health Indicator**

<sup>25</sup> The R code for EDA and CDA is available in Github: <https://github.com/rubi1rdz/A9-Final-Project>

As a next step, we ran another model with only the identified 13 variables and achieved a performance of 0.742  $R^2$ . This means that we can fairly well explain the CDC Health Indicator with only 13 out of the 47 variables. Therefore, we decided to use these 13 variables to create our model. Table 3 shows the results from this analysis.

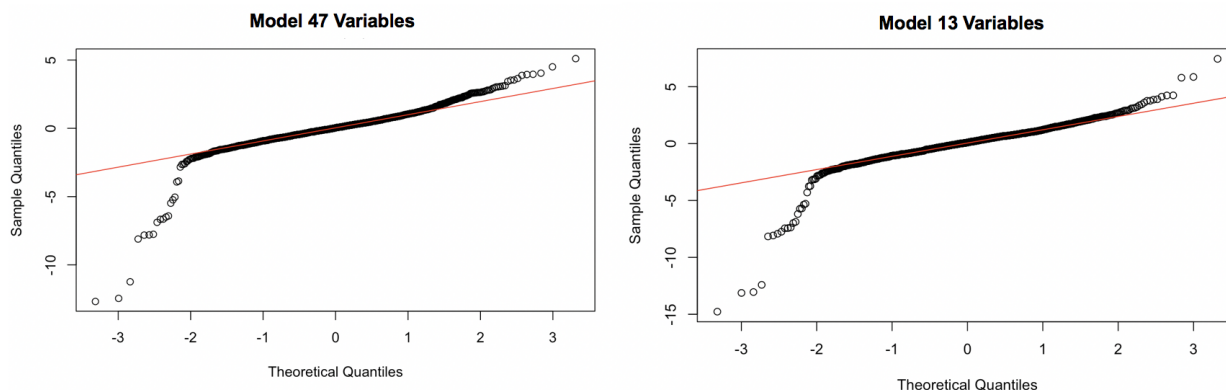
.	Estimate	Pr(> t )	
(Intercept)	7.32E+00	<2.00E-16	***
PopCountCDC	2.36E-04	1.39E-10	***
PM2.5CES	-3.69E-01	8.88E-11	***
DieselPMCountCES	1.33E-03	0.66033	
PesticidesCountCES	-4.97E-03	0.002329	**
Tox.ReleaseCountCES	5.19E-04	0.000858	***
SolidWasteCES	-4.43E-02	0.000194	***
PerCapitaIncome	-1.96E-05	2.56E-06	***
PeopBelowPoverty	-1.38E-03	2.43E-07	***
PercPeopBelowPoverty	1.36E-01	<2.00E-16	***
PeopNoDiploma	1.16E-01	<2.00E-16	***
Age17	7.87E-02	1.56E-13	***
PeopleWithDisability	1.89E-01	<2.00E-16	***
PeoplInstGroups	-6.42E-02	2.52E-13	***

**Table 3: Results from the second iteration yielded a performance of 0.742**

The final equation for our indicator is given by:

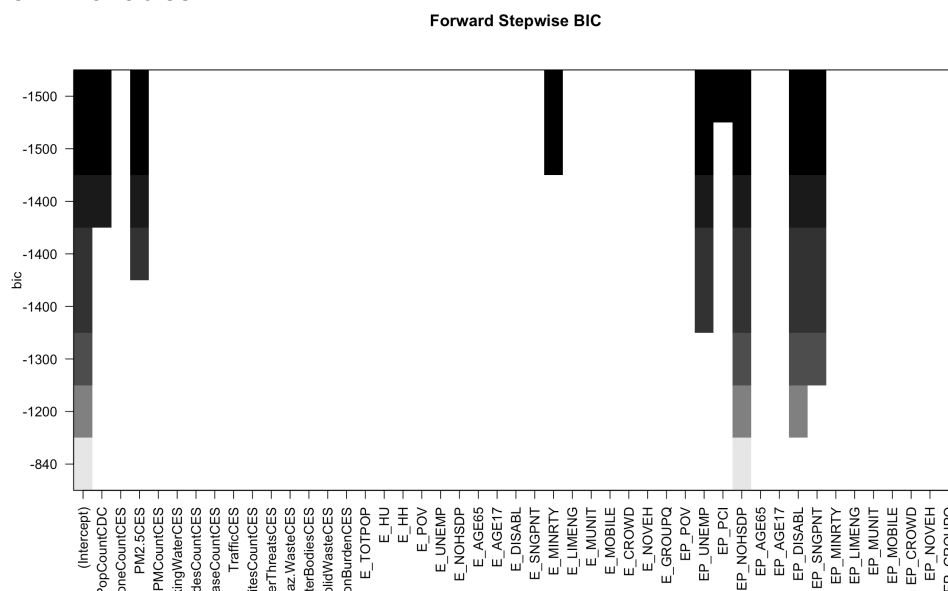
$$\text{NewIndex} = 7.32 + 2.36\text{E-}04 \cdot \text{PopulationCount} - 1.96\text{E-}05 \cdot \text{PerCapitaIncome} + 1.36\text{E-}01 \cdot \text{PercPeopBelowPoverty} + 1.16\text{E-}01 \cdot \text{PeopNoDiploma} - 4.43\text{E-}02 \cdot \text{SolidWaste} + 5.19\text{E-}04 \cdot \text{Tox.ReleaseCount} - 4.97\text{E-}03 \cdot \text{Pesticides Count} - 6.42\text{E-}02 \cdot \text{PeoplInstGroups} + 1.89\text{E-}01 \cdot \text{PeopleWithDisability} - 3.69\text{E-}01 \cdot \text{PM2.5} + 7.87\text{E-}02 \cdot \text{Age17} - 1.38\text{E-}03 \cdot \text{PeopBelowPoverty} + 1.33\text{E-}03 \cdot \text{DieselPMCount}$$

As a note, both models were validated by proving the normality of the residuals as shown in Figure 4.



**Figure 4: Normality of the residuals is proved for the first model with all 47 variables (left) and for the one we used as basis for our index (right).**

In addition to Multivariable Linear Regression, we considered using a Machine Learning model for Regression known as Forward and Backward Stepwise Bayesian Information Criterion (BIC) as a method to develop our indicator. Figure 5 shows the results for Forward Stepwise BIC which included the 47 variables.

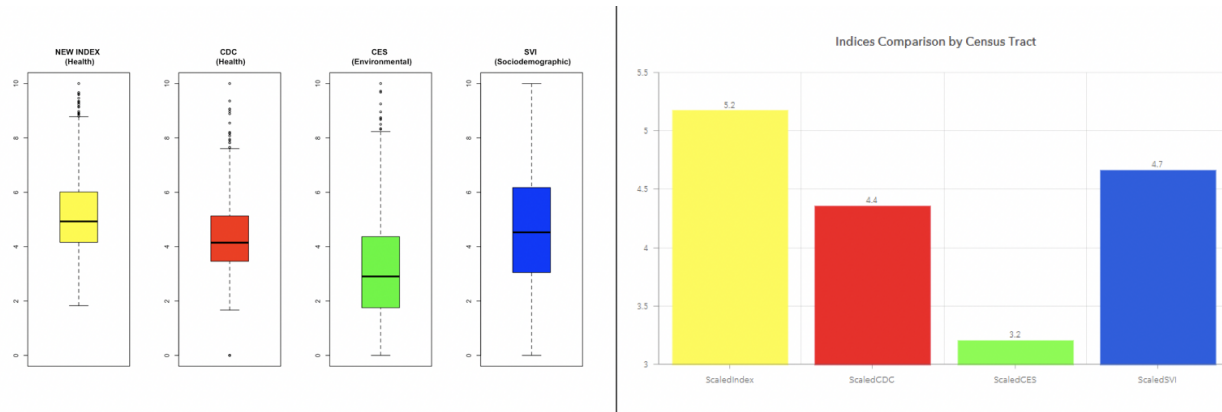


**Figure 5: Results from the Forward BIC model**

The black columns shows the variables that better explain the behavior of the CDC Health Index. If we had used the Forward BIC model to define the indicator, socioeconomic factors would be far more important than environmental factors in explaining health outcomes. PM 2.5 is the only environmental factor that the model would have included. Since we are interested in understanding the importance of several environmental factors in addition to socioeconomic factors, we chose to use multivariable linear regression instead of Forward BIC as our methodology.



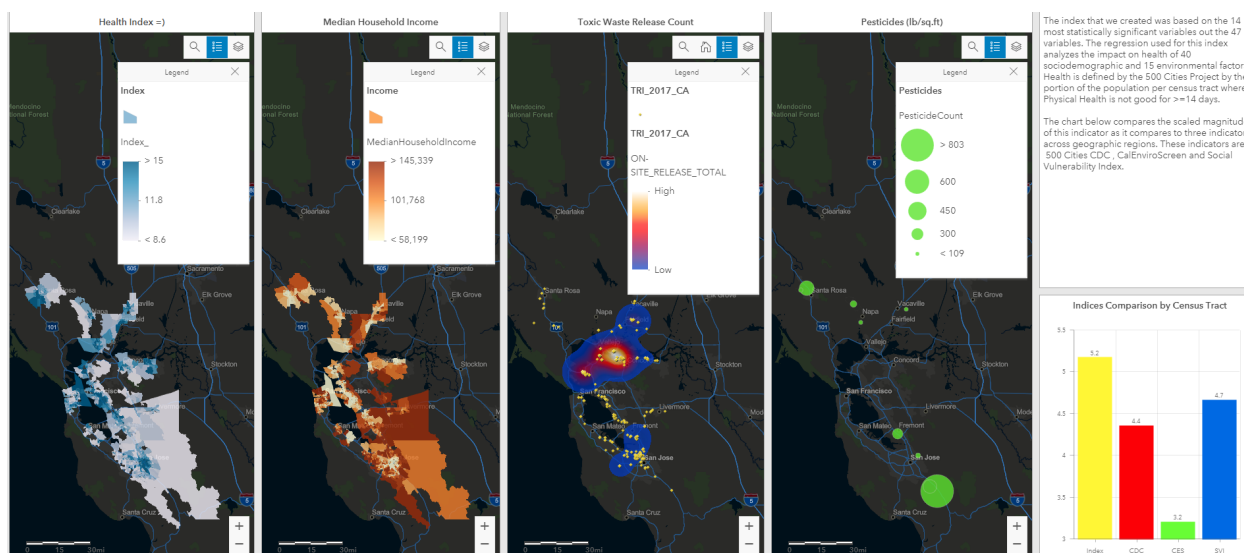
Finally, in order to understand how our index compared to the three indices that were used to inform our analysis, we created boxplots to analyze the distribution across the Bay Area, as shown in Figure 6. These boxplots gave us a reference point when we analyzed the scaled indices in our ArcGIS dashboard. The right side of Figure 6 shows an example from the dashboard of the scaled indices in a single census tract.



**Figure 6: Indices distribution across the Bay Area (left) and by Census Tract (right)**

### 3.5 Spatial Analysis

From our statistical analysis, we found that income, pesticide concentration, and toxic waste releases were three main contributors to our new index. Therefore, we proceeded to explore the distribution of these variables further. We created a dashboard to compare our original health index to the distributions of toxic waste releases, pesticides and median household income as shown by Figure 7. On the bottom right of the dashboard, we displayed a comparison of the indices.



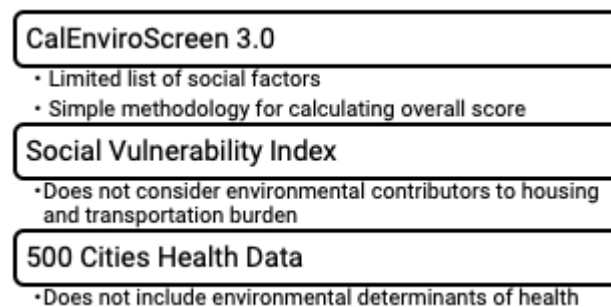
**Figure 7: Dashboard with spatial distribution of health, median income, pesticides and toxic waste release count.**



## 4. Results

### 4.1 General results of Analysis

Figure 8 summarizes the limitations of each of the indicators. The new health index benefits from the rich data from each of the three contributing indicators. It overcomes some of the limitations of each metric by consolidating data across social, environmental, and health outcomes.



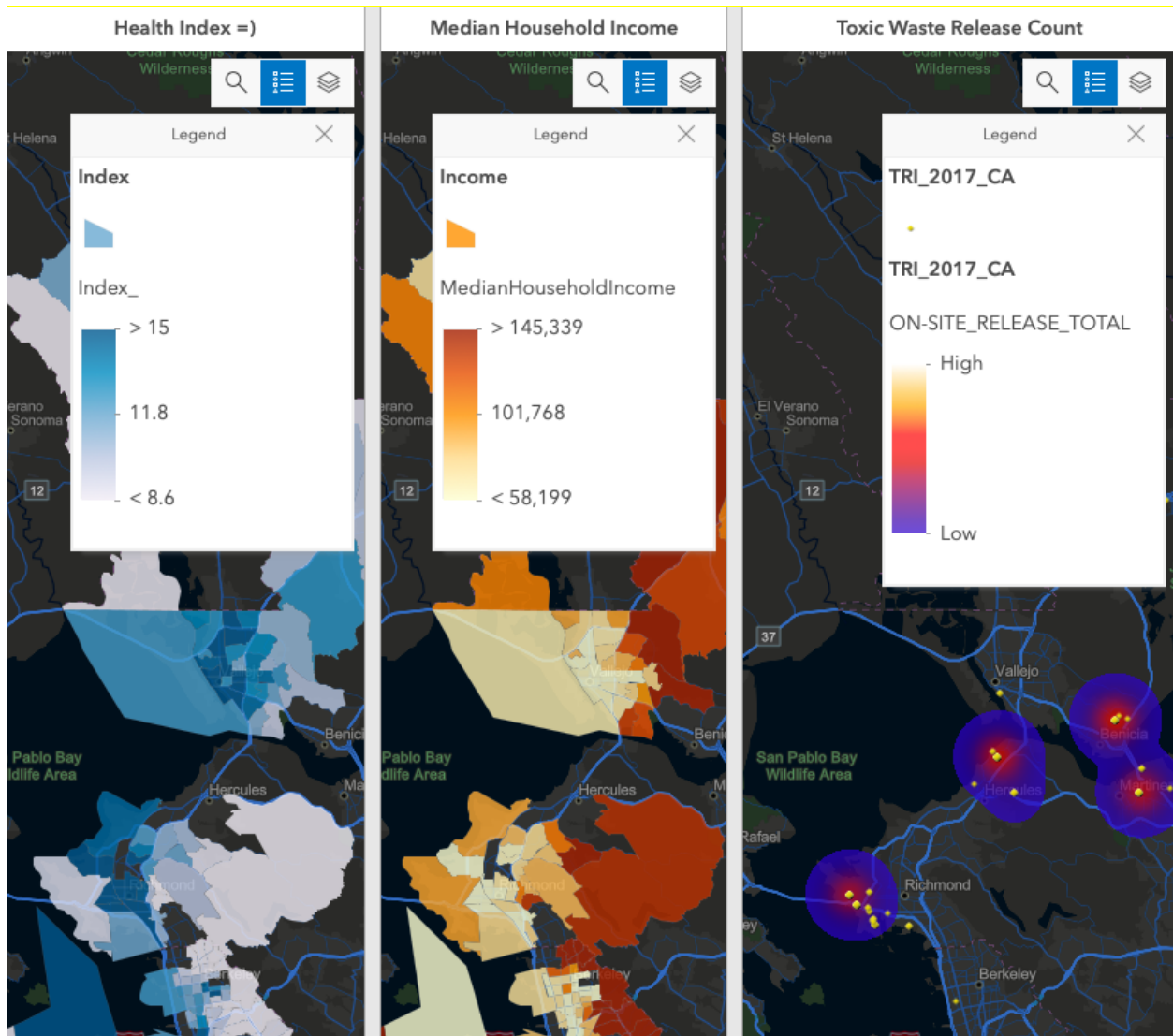
**Figure 8: Limitations of each of the underlying datasets**

CalEnviroScreen 3.0 includes 27 environmental factors but provides limited data about health outcomes. This metric considers asthma, low birth weight, and cardiovascular health. The Social Vulnerability Index contains 32 socioeconomic descriptors including housing and transportation, but the index does not quantify any of the environmental factors that may influence these social outcomes. It also does not consider health outcomes. The 500 Cities dataset includes information about various health conditions but does not examine environmental or social contributors to health. Therefore, our study adds value by overcoming some of the limitations of each of the individual indicators and providing a more comprehensive indicator that draws from the extensive amount of data contained within each source.

We found that in general, areas with higher median income corresponded with areas of a lower health index. This confirmed our hypothesis because it is likely that people with higher incomes have access to healthcare and healthier environments which may contribute to fewer days of poor physical health. While toxic waste sites and releases did correspond to a higher incidence of poor health, the trend was not as clear across the geography. For instance, near San Jose, there are several facilities documented Toxic Release Inventory, however, this does not necessarily correspond to areas of higher health burden. One potential explanation is that median household income has a greater influence than exposure to Toxic Releases. This is supported by the finding that income had a p value of 7.06E-09 whereas toxic release count had a p value of 4.74E-04. While these observations are intended to provide insight, they do not imply causation or predict outcomes.

## 4.2 Analysis of Richmond, San Jose, and Morgan Hill, California

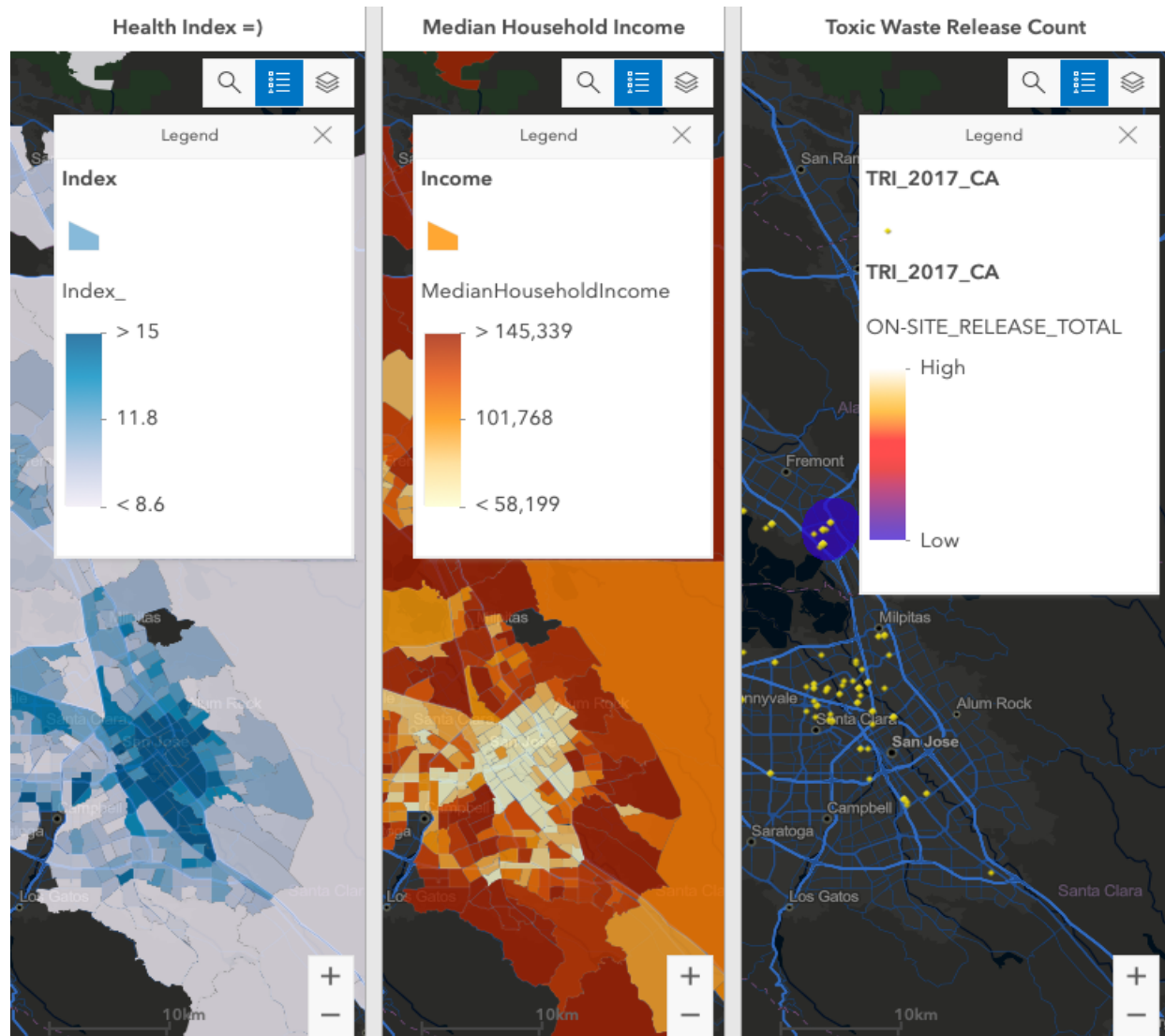
After reviewing literature about activism and health burdens in various communities, we became interested in how exposure to toxic releases corresponds to income and health.<sup>26</sup> As discussed in the literature review, in the East Bay region, particularly the Richmond area, toxic releases have contributed to many hospitalizations for respiratory infections as well as chronic health conditions. As Figure 9 shows, low median income corresponds with areas of higher health burden and proximity to toxic waste sites.



**Figure 9: Spatial representation of health, income, and toxic waste in Richmond, California**

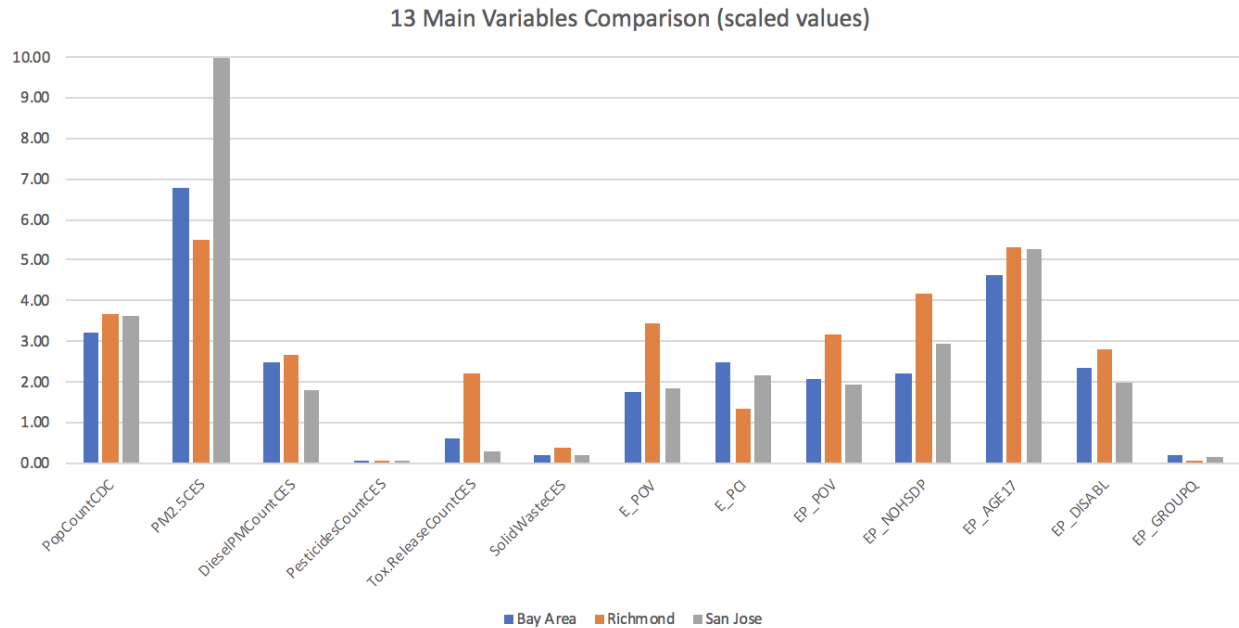
<sup>26</sup> "Major Accidents at Chemical/Refinery Plants in Contra Costa County," Contra Costa Health Services, Available at: <https://cchealth.org/hazmat/accident-history.php>

We were also interested in studying the spatial representation of health, income, and toxic wastes sites in San Jose, California. In San Jose, there are a number of toxic release sites, but the health indicator seems to correlate more closely with median income as shown by Figure 10.



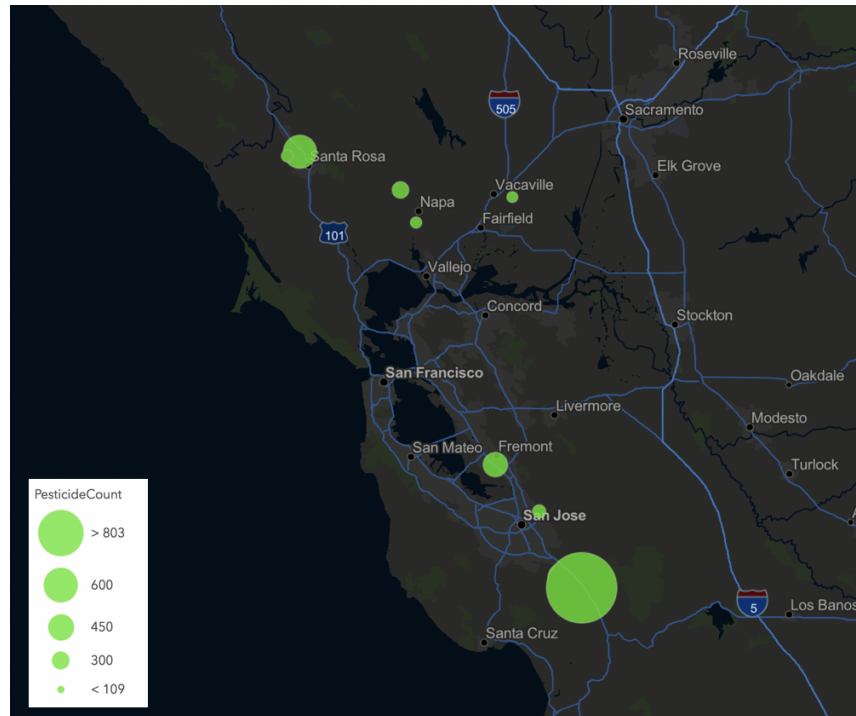
**Figure 10: Spatial representation of health, income, and toxic waste in San Jose, California**

As expected, social indicators (per capita income, percent of the population in poverty, level of disability, and number of young people under 18) indicated that on average San Jose has a higher standard of living than Richmond. However, as Figure 11 shows, while the relative impact of toxic release count for Richmond was higher, San Jose experienced a higher amount of PM 2.5 pollution. This could be partially a result of San Jose's large geographic spread and demographic diversity. However, this comparison is illustrative because it shows how these urban areas are impacted by different environmental and social factors, but the overall trends are consistent with what we would expect.



**Figure 11: Comparison of the 13 factors contributing to the health index across Richmond and San Jose with the Bay Area average**

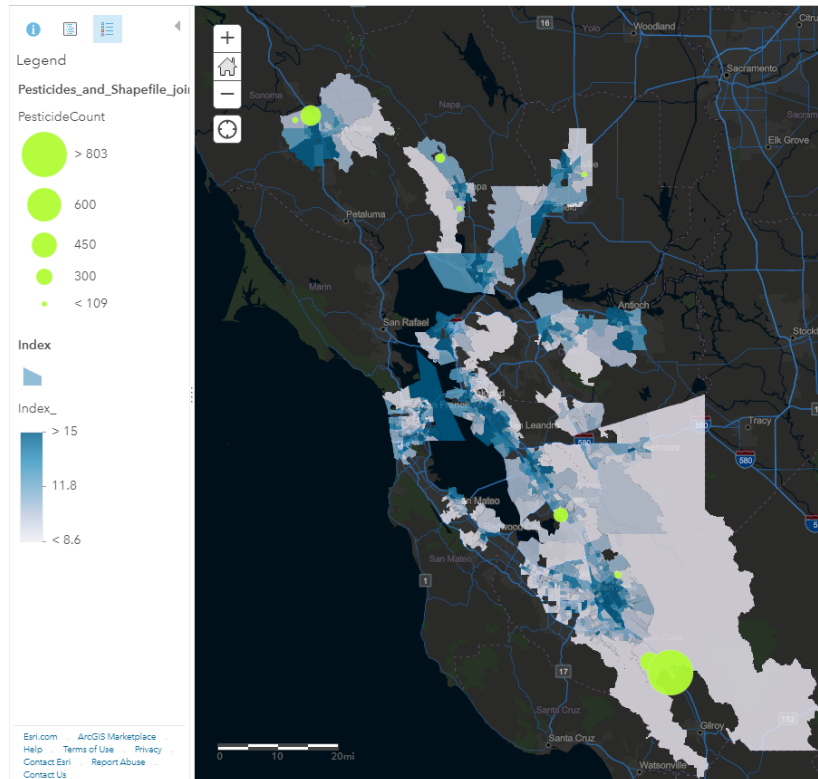
We also decided to look further into the concentration of pesticides. It was surprising to us that out of the 47 variables, pesticide concentration made it to the top 13 in statistical significance when we were predominantly considering urban or suburban areas. As we analyzed this variable further, we realized that out of the 1106 census tracts included in the analysis, 901 had a value of zero for pesticide count while the rest vary from 0 to 803 lb/sq. ft. As shown in Figure 12, we mapped the census tracts that had a concentration of higher than 100 lb/sq.ft. across the Bay Area.



**Figure 12. Pesticides concentration in lb/sq.ft. The map displays the top 10 areas in the Bay with concentrations higher than 100 lb/sqft.**

Based on our experience living in the Bay Area, we recognized that the regions shown in the map correspond to suburban and agricultural areas. For example, Morgan Hill, the highest concentration in the Bay, has a large amount of agricultural land. Therefore, it would be interesting to conduct further research and collect data related to land use in order to find better correlations that could explain our insights.

We next sought to further explore the contributions of pesticide concentration to our index. According to our model, pesticide concentration was one of the top 13 variables that explained the value of the CDC health indicator. Our hypothesis was that higher concentrations of pesticides would correspond with higher amounts of health burden (and therefore a higher score on our health index). We mapped both the index and pesticide concentration as shown by Figure 13.



**Figure 13: There is no apparent direct impact on the index due to high concentrations of pesticides**

Figure 13 shows an interesting result. High levels of pesticide concentration do not necessarily contribute to high scores on our index even for census tracts where pesticide concentration is statistically significant. As we go back to the p-values of the models, we can see that Pesticides ranks 11th out of 13 variables in terms of significance. Although it is a significant factor, other factors likely have a more significant impact on explaining health.

This leads us to one of our key takeaways from the analysis. When creating a linear model we realized that is very important to always be careful in interpreting the significance of each variable. For instance, we may have been misguided in our conclusions if we only considered the results from our regression analysis and the distribution shown by Figure 12. We would expect that a higher proportion of people living in areas of higher pesticide concentration, such as Morgan Hill, might experience greater health burdens. However, because we dug a bit deeper into the p-values and created Figure 13, we realized that our initial hypothesis was incorrect. We are missing information about the spread of the area, proximity of households to the source of pollution, and type of land use that could help us better explain the good levels of health regardless of the high concentrations of pesticides. These type of data could be included in future research to better understand the impact of pesticides in the region and their correlation with health burden.

## 5. Conclusions and Future Work

### 5.1 Conclusions and Limitations

Our analysis does not imply causation or predict outcomes, but is helpful in finding some environmental and social factors that contribute to health outcomes. Our preliminary findings support the body of literature that factors such as exposure to toxics and income affect health impacts. The literature review highlighted some limitations of each of the indicators. In general, it seemed that social and demographic factors had a higher correlation to health than the environmental factors that we analyzed. These findings also highlight the need for public education about each of these various factors. For instance, we anticipated that air quality indicators such as PM 2.5 and Ozone would have a greater impact than proximity to solid waste sites in urban and suburban areas. By focusing on smaller geographies and census tracts, we were able to see that the overall trend may not be representative of the impact of some of the indicators on specific census tracts that was less significant in the overall analysis.

Furthermore, since each of these indicators come from various data sources with different methodologies, we found that it was important to deeply understand how each of these metrics were derived and to acknowledge that there are many gaps in the data. For instance, when we analyzed exposure to PM 2.5 and ozone, we realized that the data that was collected and interpreted by the CalEnviroScreen 3.0 was averaged across several census tracts in some cases which may have masked some underlying trends. While our model did attempt to provide a more comprehensive analysis of the data from each of the three indicators we used as input data, our model has likely inherited some of the shortcomings of data collection and analysis. This exploratory effort shed light on gaps in the data, and we urge government agencies and foundations to invest in collecting robust environmental and social data and conducting more longitudinal studies that would make this type of analysis more meaningful and generalizable. This study could be a starting place in identifying where to collect more information and how policy makers can allocate funding to addressing data gaps and employing effective public health and environmental interventions.

### 5.2 Future Work

Because environmental justice is a movement rooted in increasing equity and access to decision making, an important next step would be to critically evaluate how this analysis could be made more accessible and useful for the needs of various stakeholders including activists, policymakers, industrial facilities, and healthcare providers. Some future steps could include a series of listening sessions and engagement with grassroots organizations such as the California Environmental Justice Alliance and the Greenlining Institute could be helpful in shaping future

iterations of this analysis. Another extension could be to analyze the impacts of legislation such as SB 535 on environmental and social determinants of health. Particularly since this legislation was heavily impacted by CalEnviroScreen 3.0 and its underlying datasets, it would be interesting to see the impact of these metrics on actual health incomes in the long term. One critique of both the CalEnviroScreen 3.0 and the 500 Cities Health Indicators is that both do not adequately capture the environmental and social realities of rural populations. While this is a data gap that needs to be filled by the research community, we could try to mine additional datasets in creative ways to try to fill in some of the geographic gaps in the analysis.

Future work could include a predictive analytics component to expand upon our exploratory observations and findings. For instance we could train a Multilinear Regression, MARS or NeuroNets to predict the values for the rest of the census tracts on the Bay Area which were not included by the 500 Cities Project. However, before doing so, we would like to do further analysis by different scales of geographies across the Bay to see if the variables that are more significant in our overall analysis are as significant in a more local level. We would consider conducting this analysis at a more granular level by block group if we could find this data. This is mainly because a predictive model assumes that the data that would be used to predict the new outcome comes from the same distribution as the data that was used to train the model. Therefore, we need first to make sure this assumption holds.

Finally, targeting our model to a specific audience and geography can make our indicator an even more powerful tool for policymakers and stakeholders. For instance, policymakers can use a refined version of this tool to assess progress of existing environmental health interventions and propose new policies and programs. Activists and community organizers can use this information to ground their advocacy, produce educational materials, and advance their goals in stakeholder meetings, regulatory comments and in other political action. Healthcare providers and local government agencies can better understand the environmental and social factors that affect local health and design targeted interventions to address these issues. While this tool is still in its infancy, it could eventually be developed into an instrument that expands upon and overcomes some of the limitations of existing social and environmental metrics. This can help drive the action needed to foster more equitable and healthy communities.