DS4A Project Report - Team 31

The City-Dweller at the Crossroads: The Intersection of Public Health Risks, Income Disparity, and Seismic Hazard in the Bay Area

Bin Lin, Xiran Liu, Natasha Toghramadjian, Athena Zhang
Mentors: Geetha Gopakumar, Maria Fernanda Osorio Moreno

October 2020

1 Introduction

1.1 Background

Urban environments are the crossroads where hazards to human health and well-being play out across diverse demographic swaths.

In the Bay Area of California, public health hazards, socioeconomic disparities, and seismic activity intersect–posing multi-layered risks to the more than 7 million residents of the metro area. These risks play out differently in each community. To effectively respond to ongoing public health risks and natural disasters, it is necessary to develop individualized assessments of the overlapping risks facing specific geographies.

We thus seek to understand the interplay of these diverse hazards on a highly granular scale. We leverage several robust, independent data sets to investigate relationships between multitude of public health risks, socioeconomic disparities, and community characteristics at the census tract scale for the entire Bay Area. We further perform a focus study on the most densely-populated and seismically active sector, San Francisco proper, integrating seismic hazard data for an added dimension of risk analysis.

1.2 Motivation and Problem Statement

"People in real life are simultaneously exposed to multiple contaminants from multiple sources and also have multiple stressors based on their health status as well as living conditions. Thus, the resulting cumulative health risk is also often influenced by non-chemical factors such as socioeconomic and health status of the people living in a community."

- California Environmental Screening Report, January 2017

As outlined by the 2017 California Environmental Health Screening report, the cumulative health of a community and its individuals is influenced by distinct stressors that can mutually exacerbate one another and heighten health risks.

The Bay Area, comprised of nine counties (Alameda, Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, and Sonoma), is home to over 7 million people and 101 cities. Highly urbanized and seismically active, the Bay Area faces dangerous levels of pollution, stark disparities in socioeconomic conditions, and frequent earthquake occurrences. It is thus a natural laboratory for investigating the relations between these diverse threats to public well-being, particularly the question of which environmental hazards are highly correlated in census tracts marked by majority high- or low-income.

1.3 Objectives

With the above motivation in mind, we intend to identify how relationships between public health risks and socioeconomic disadvantages vary between communities of different income levels.

The key goals of this study are:

- To investigate the relationships between diverse public heath hazards, socioeconomic indicators, seismic hazards, and income levels on regional and localized scales.
- To identify which disadvantages are most correlated in low-income census tracts, and if this differs from correlated risks in middle- and high-

income census tracts.

- To build a model that can inform regional and local policy-makers with actionable insight.

2 Data Analysis and Computation

2.1 Datasets and Data Processing

2.1.1 Description of the Datasets

The following data sets were utilized in this study:

- **California Communities Environmental Health Screening, CES3.0:** CES3.0 identifies Californian communities by census tract that are disproportionately burdened by, and vulnerable to, multiple sources of pollution and socioeconomic disadvantages. The project was initiated in 2012 and most recently updated in 2018 (https://oehha.ca.gov/calenviroscreen).

  "CalEnviroScreen uses a science-based method for evaluating multiple pollution sources in a community while accounting for a community's vulnerability to pollution's adverse effects." (CES3.0 Report, 2017).

- **UC San Francisco Health Atlas:** Health Atlas is an interactive tool that collates public data on neighborhood-level physical and social characteristics that influence health. We extracted Census tract-level crude prevalence rate of adults with arthritis, high blood pressure, current asthma, coronary heart disease, chronic obstructive pulmonary disease (COPD), diabetes, high cholesterol, chronic kidney disease, poor mental health, obesity, poor physical health, and stroke. (https://healthatlas.ucsf.edu).

- **Social Explorer Database:** Social Explorer is a database and visualization tool that provides hundreds of thousands of data indicators across demography, economy, health, education, religion, crime and more. (socialexplorer.com). We leverage census tract-scale income and rent burden data for the ACS 2018 5-year estimates for our analysis:

  - Proportion of households in each census tract that makes specified levels of income–both discrete brackets (e.g. Percentage of

households making $25,000 to $49,999) and cumulative values (e.g. Percentage of households making Less than $75,000).

- **Seismic Hazard Zone Maps and Evaluation Reports**: California state government reports and maps of probabilistic seismic peak ground acceleration for soft rock conditions, from the California Department of Conservation and California Geological Survey (conservation.ca.gov). Based on geologic and tectonic features, these values exist on a geologic timescale and do not change for decades.

2.1.2 Data cleaning, filtering, joining, etc.

- CES3.0 and Health Atlas, which covers all of California, was filtered to include only census tracts within the nine aforementioned Bay Area counties, for a total of 1,581 census tracts.

- Social Explorer data was filtered for the same region, reformatted, and then joined to the CES3.0 spreadsheet, matching each census tract.

- Seismic ground acceleration maps were first georegistered in QGIS. Registered maps were then loaded into the modeling software GOCAD, in UTM coordinates. The centerpoint latitude-longitude coordinates of every Bay Area census tract were converted to UTM, and then logged as unique points in GOCAD, overlaying the seismic peak ground acceleration (PGA) maps. Probabilistic PGA values were then extracted for every census tract in San Francisco proper (232 census tracts in San Francisco and San Mateo counties), and joined to the CES3.0, Health Atlas, and Social Explorer data set. The seismic PGA data coverage represents the most densely populated and urbanized sector of the Bay Area, where the majority of active faults are.

2.2 Exploratory Data Analysis

2.2.1 Missing Data

Reasons for missing data in CES3.0 included 1) no monitoring or reporting conducted or 2) no population reported within that census tract. For example, ambient air quality measures were not available for all California census tracts because many places do not have an air monitor close enough to reliably estimate air quality in that community. In the Bay Area, 1 out of 1581

census tracts had zero population reported by the U.S. Census so all population characteristic indicators assigned "NA" for this tract. In addition, census tracts with highly unreliable estimates for the educational attainment, linguistic isolation, poverty, and rent-adjusted income indicators were assigned "NA."

	Missingness
Drinking Water	2
Traffic	12
Low Birth Weight	25
Education	14
Linguistic Isolation	37
Poverty	12
Unemployment	24
Housing Burden	23

Table 1: Count of missing data in CES3.0.

2.2.2 Demographic Information

We first looked at the demographic information of the population we are working with. The population size, age and race or ethnicity are summarized in Fig. 1 and 2.

The distribution of age is similar among most areas, with the majority of the population within the group of 11-64, and a small portion of child and elderly. The distribution of race or ethnicity varies more between difference census tract areas. There are very few Native American population in this dataset. We can observe from the fraction of different colors in Fig. 2c that in some areas, there is a larger Hispanic population (blue), while in some others, Asian population (purple) forms the majority of the total population.

2.2.3 Pollution Burden Variables

There are 24 pollution burden variables, as descried in Section 2.1.1. We visualized the value distribution of each of these variables on the map of Bay Area (provided in Fig. 15 in Appendix.A) and examined their summary

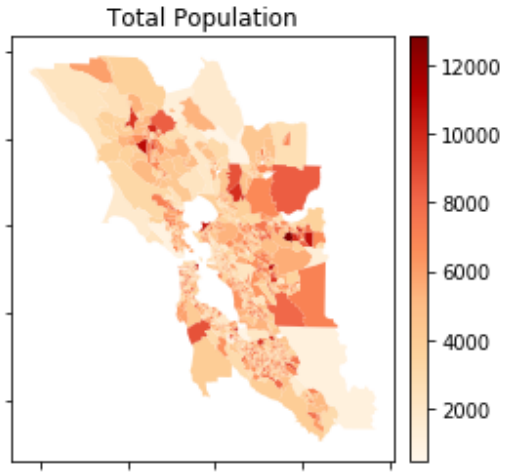
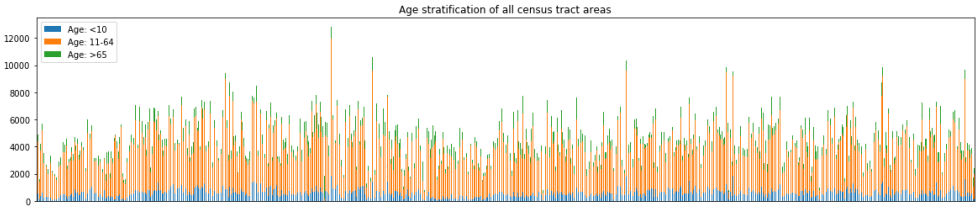


Figure 1: Population size of each census tract area

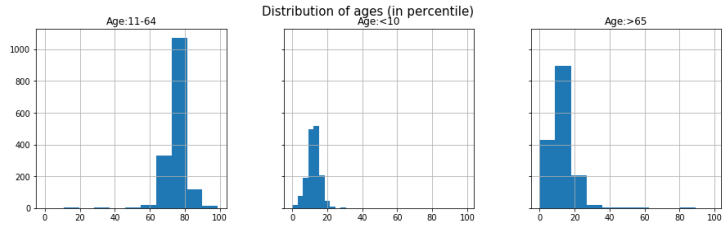
statistics (described in Table. 6 in Appendix.A). The distribution of these variables is shown in Fig. 3 and the corresponding pairwise scatter plot is shown in Fig. 16 in Appendix.A.

2.2.4 Population Characteristic Variables

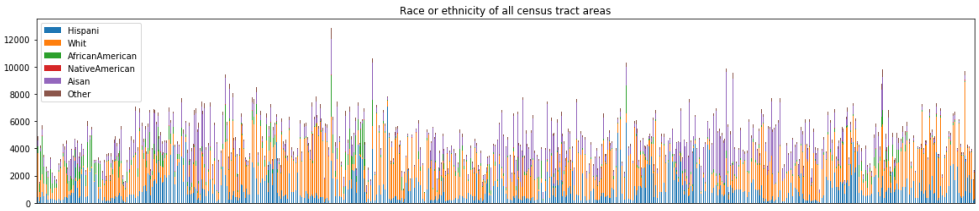
There are 15 population characteristic variables, as described in Section 2.1.1. We visualized the value distribution of each of these variables on the map of Bay Area (provided in Fig. 17 in Appendix.B) and examined their summary statistics (described in Table. 7 in Appendix.B). The distribution of these variables is shown in Fig. 4 and the corresponding pairwise scatter plot is shown in Fig. 18 in Appendix.B.



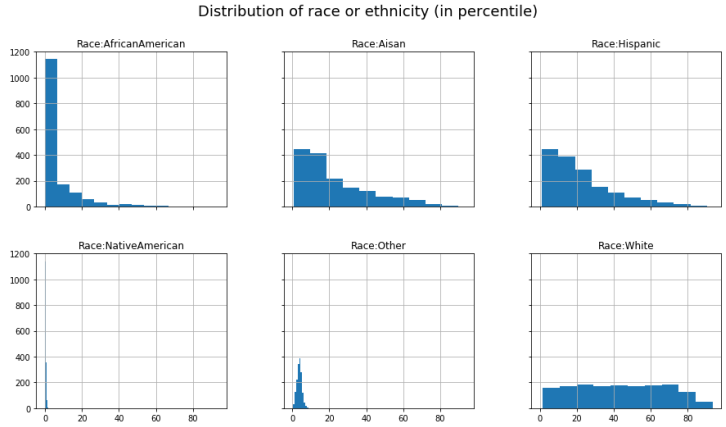
(a) Number of people in each of the three age groups for all areas.



(b) Percentile distribution of three age groups.



(c) Number of people in each of the six race or ethnicity groups for all areas.



(d) Percentile distribution of six race groups.

Figure 2: Age and Race (or ethnicity) stratification of the Bay Area population: Age divided into three groups: main population (11-64), children (<10), and elderly (>65); Race or ethnicity divided into six groups: Hispanic White, African American, Native American, Asian American, and Other.

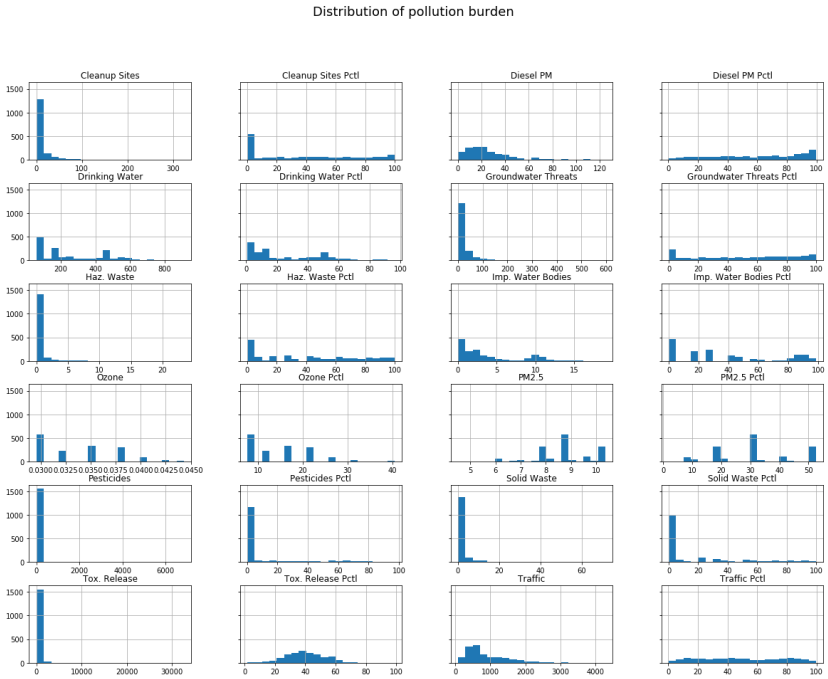


Figure 3: Distribution of pollution burden variables.

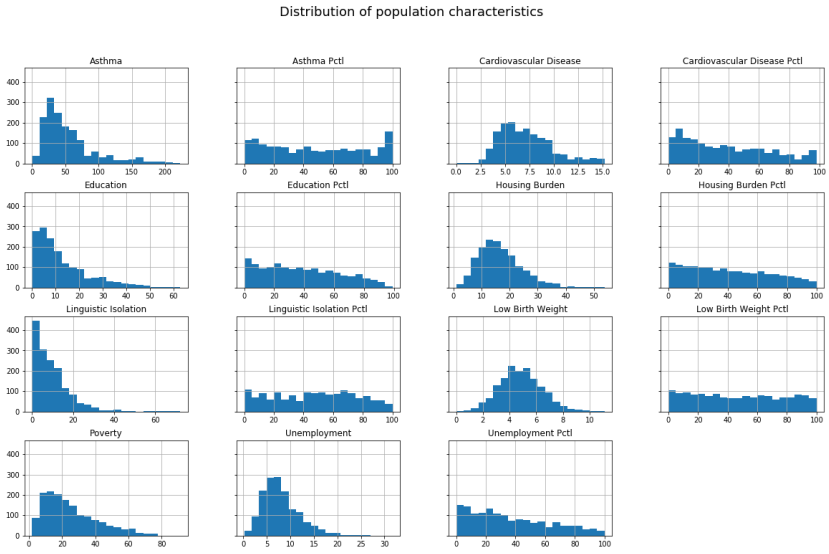


Figure 4: Distribution of population characteristics variables.

2.2.5 Income

Looking at the percentages of population with income below each of the four thresholds shown in Fig. 5, the northern Bay Area has higher proportion of population with low income than the southern Bay Area.

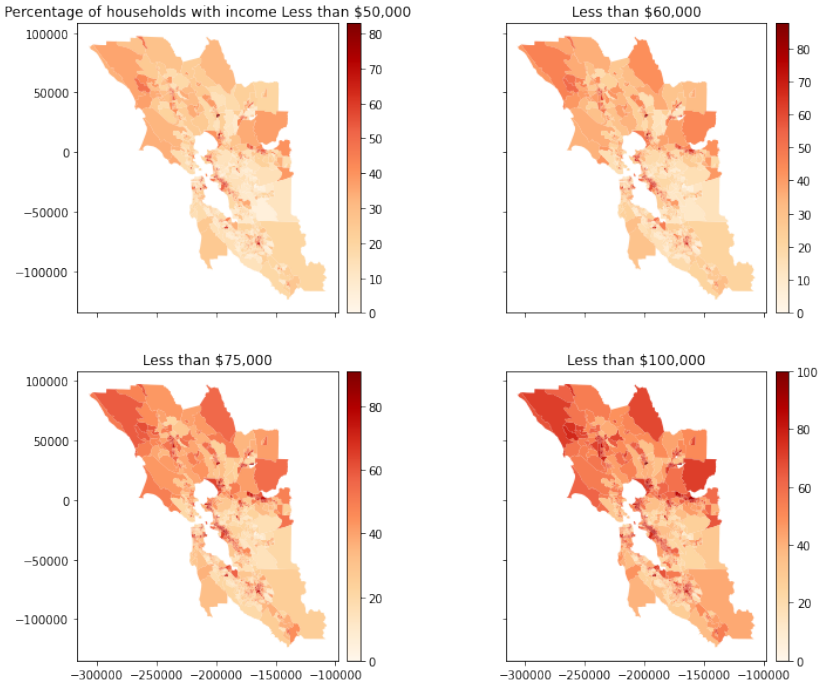


Figure 5: Percentages of population with income below thresholds: $50,000, $60,000, $75,000, and $100,000.

2.2.6 Distribution of Health Outcomes

Stratified by income above and below $75,000, we compare the distributions of crude prevalence rate of adults with arthritis, high blood pressure, current asthma, coronary heart disease, chronic obstructive pulmonary disease (COPD), diabetes, high cholesterol, chronic kidney disease, poor mental health, obesity, poor physical health, and stroke (Fig. 6). While the average prevalence for all conditions may be similar between the income groups, there is less variability in the higher income group. This could be due to better reporting and access to healthcare.

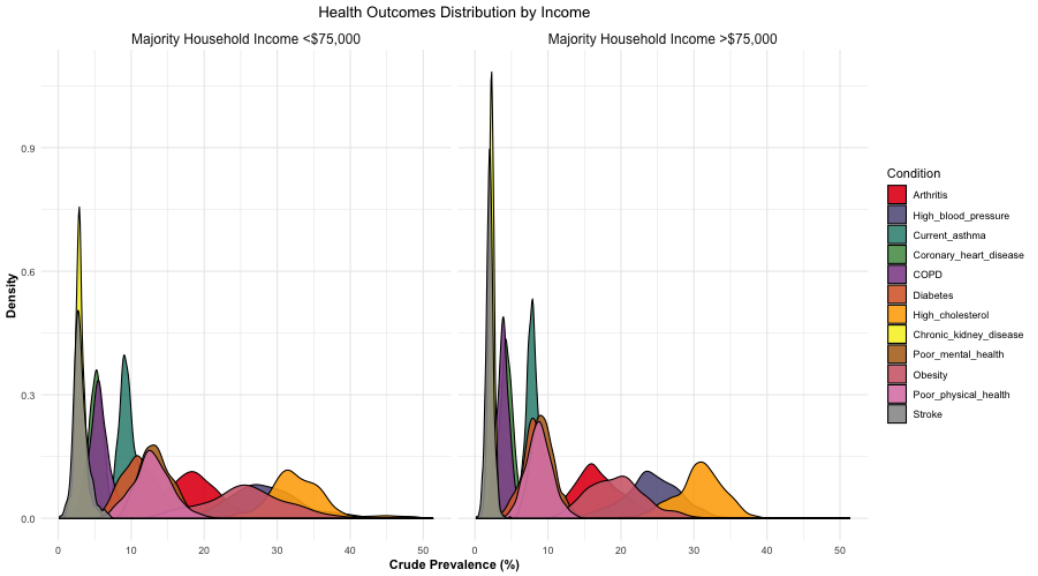


Figure 6: Distribution of adults with health conditions stratified by income.

2.2.7 Seismic Data

The peak ground acceleration (PGA, maximum acceleration the ground will experience at a given location during an earthquake, expressed as a fraction of gravitational acceleration g) of which there is a 10% chance of being exceeded in the next 50 years, for every census tract in the San Francisco area, is shown in Fig. 7. PGA is a physical measure used directly in the seismically-safe engineering of civil structures, such as residential high-rise buildings and large bridges. The first-order pattern of seismic hazard levels is a clear systematic spatial distribution, based on local surface geology and proximity to active faults. In terms of exploratory data analysis, the spatial characteristics of seismic hazard are straightforward: gradual increase in expected PGA from northeast to southwest. However, while this distribution is not physically related to any of the socioeconomic or pollutant variables investigated, high seismic hazard may overlap with high risk levels in other variables and thus merits integrated spatial analysis.

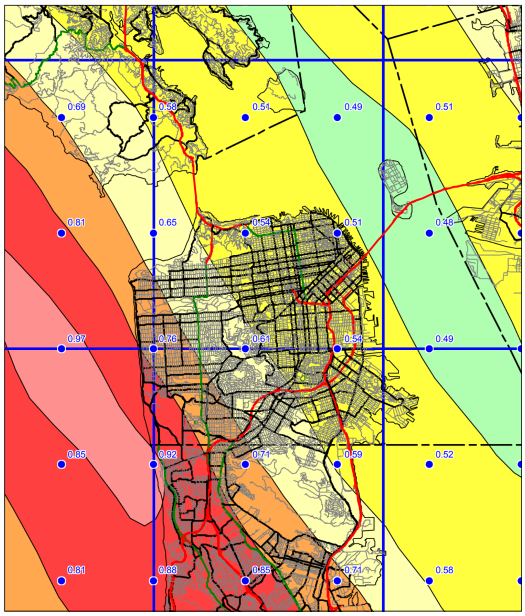


Figure 7: Peak ground acceleration (PGA) with 10% probability of exceedance in 50 years in San Francisco (U.S. Geological Survey).

2.3 Statistical Analysis

2.3.1 Spatial Autocorrelation Analysis

If the spatial distribution of the variables was random, then we should not see any clustering of similar values on the map. However, visualizations (in Fig. 15 and 17) suggest that spatial autocorrelation may exist for some of the variables, therefore we analyzed the relationship between spatial similarity and attribute similarity.

In spatial autocorrelation analysis, we used queen contiguity weight (neighbors that share either a corner or an edge) as the spatial weight $w_{i,j}$ to define spatial similarity between two areas i and j, i.e., to indicate if the two are neighbors. The attribute similarity is measured by spatial lag, which is defined as $x_i^{lag} = \sum_j w_{i,j} x_j$ for area i and attribute x.

We tested for the global autocorrelation statistics. By rejecting the null of complete spatial randomness, with realizations generated using random spatial permutations of the observed attribute values, we can conclude that

the patterns of some variables we looked, including poverty, unemployment, ozone, income percentage of less than $75,000, pollution burden, and seismic PGA, are all spatially-nonrandom, and show different levels of spatial correlation.

A local Moran's statistic I [1] is designed to be a local indicator of spatial association or LISA (local indicators of spatial association) statistic, which gives an indication of the extent of significant spatial clustering of similar values around that observation and has sum proportional to a global indicator of spatial association. A Moran Scatterplot [2] shows the relationship between the value of the chosen variable and the average value of its neighbors for the same variable. We computed the Moran's statistics (Table 2) and generated the Moran scatterplots (Fig. 8), with statistically significant points highlighted in dark red and mean values of the variables and the lagged variables given by the dashed lines.

Variable	Moran statistics I	p-value
Poverty	0.60	<0.001
Unemployment	0.42	<0.001
Ozone	0.98	<0.001
Income% Less than $75,000	0.59	<0.001
Pollution Burden	0.63	<0.001
SEISMIC PGA	0.94	<0.001

Table 2: Moran statistics I for selected variables.

We then distinguished each of the four specific types of local spatial association reflected in the four quadrants of the Moran scatterplot and marked the hotspots and the coldspots on the map, as shown in Fig. 9.

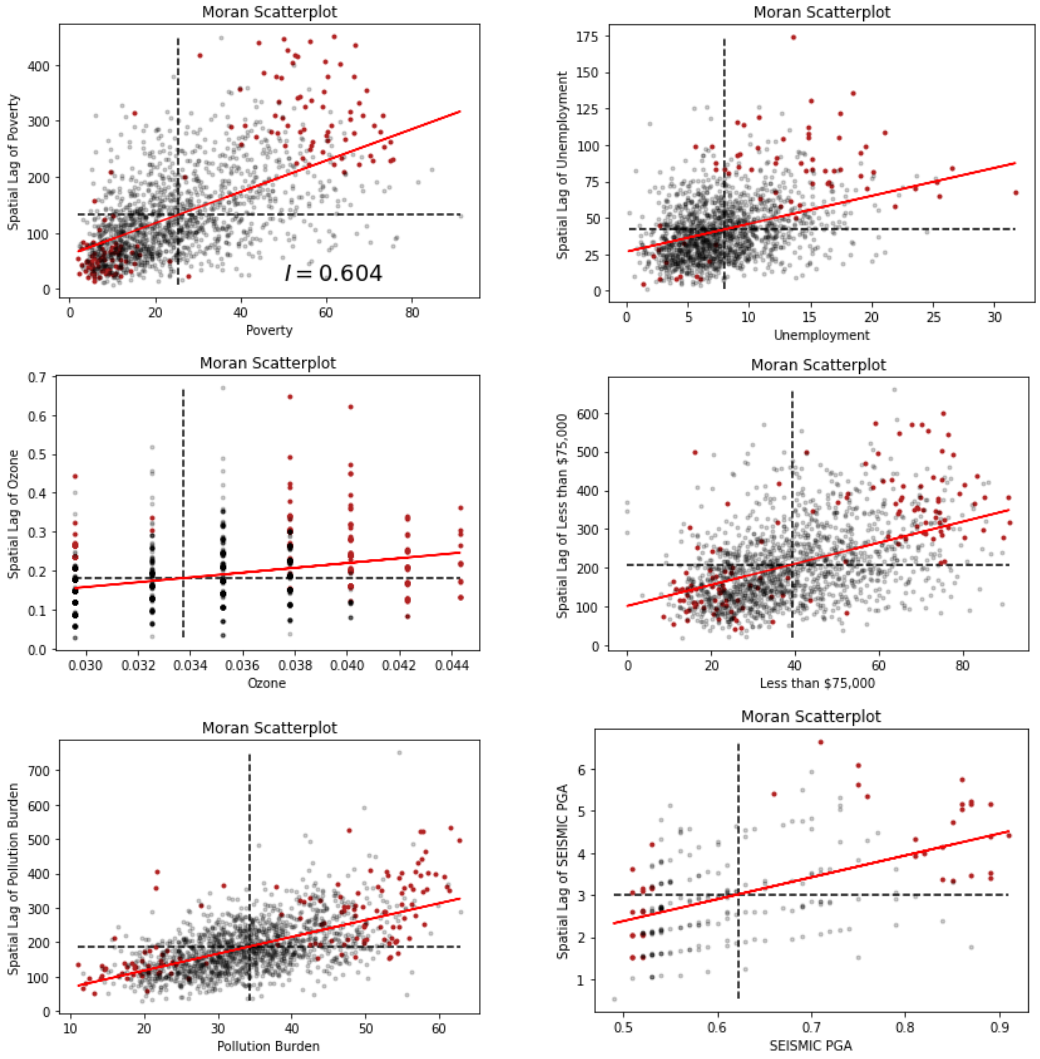


Figure 8: Moran scatterplot for variables: poverty, unemployment, ozone, income percentage of less than $75,000, pollution burden and seismic PGA. Red lines are the best fit using global I as slopes. Statistically significant points are marked in dark red. Dashed lines are the mean values of the variable and the lagged variable.

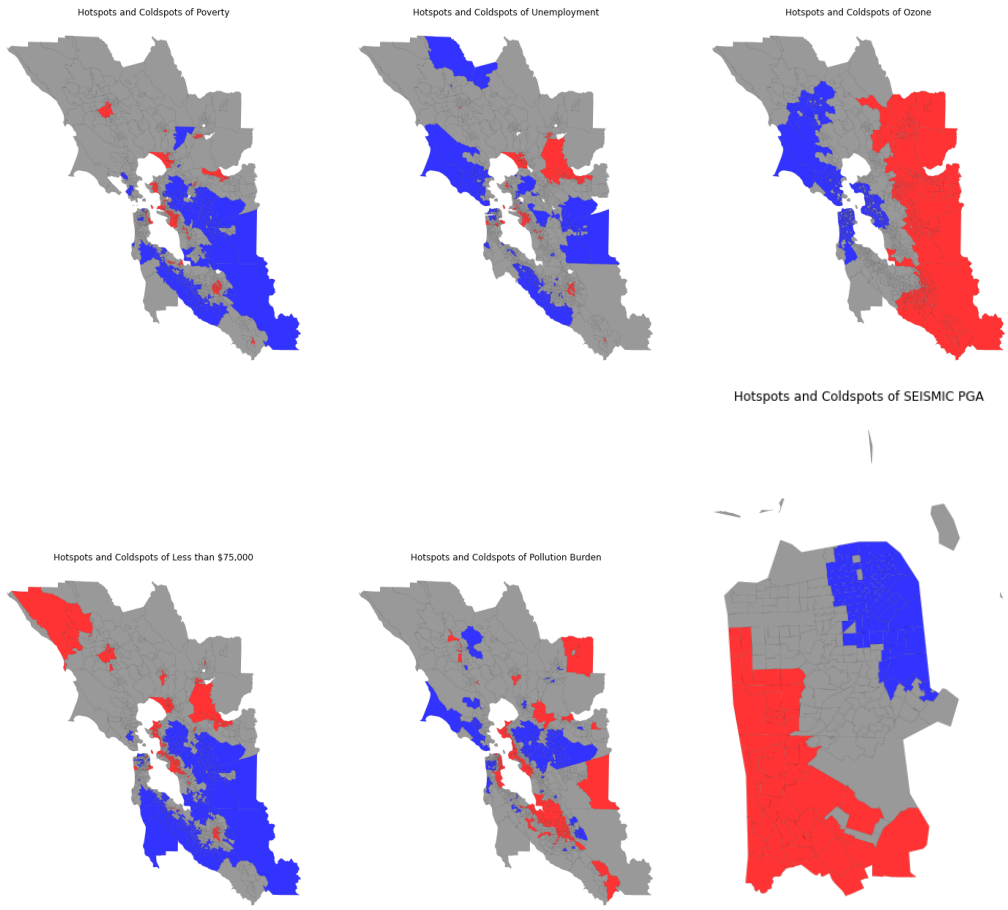


Figure 9: The hotspots and coldspots from the Moran scatterplots for variables: poverty, unemployment, ozone, income percentage of less than $75,000, pollution burden and seismic PGA.

2.3.2 Spatial Cross-Correlation Analysis

Chen [3] proposed a method for spatial cross-correlation analysis, which can reveal the importance of the part of correlation between two variables played by geographical distances or spatial relationships, and can therefore complement the spatial autocorrelation analysis on single variables. In the paper, he proposed the use of spatial cross-correlation index (SCI), which is the coefficient of spatial cross-correlation, and its goodness of fit, i.e., the coefficients of determination, in evaluating the spatial cross-correlation between two variables. Dual cross-correlation scatterplots were used to visualize the correlation relationships, and together with a scatterplot of local SCIs of the two variables, were used as another visual aid for categorization of each spatial data point.

Each of the dual scatterplots put the spatial points into four types according to the quadrants of a Cartesian coordinate system, with the first quadrant representing the high-high (H-H) type, the second quadrant representing the low-high (L-H) type, the third quadrant representing the low-low (L-L) type, and the fourth quadrant representing the high-low (H-L) type. H-H type means that an element and its neighbors are both at the higher level, and L-H type means an element is at a lower level with neighbors at a higher level, and so on and so forth. The scatterplot of local SCIS distrbutes rearrange and further support the clustering result from the dual cross-correlation scatterplots.

Using the method in paper, we examined the pairwise spatial cross-correlations between all variables, including population characteristic variables, pollution burden variables and the percentages of income below thresholds. A heatmap of all the pairwise coefficients of determination is shown in Fig. 10. We can see that for some pairs of variables, one can explain about a decent amount of the spatial change of the level of another. For example, as shown in Fig. 11, PM2.5 can explain about 22.4% of the spatial change of the level of asthma, which does not influence the spatial change of level of PM2.5.

As shown in Fig. 12, diesel particulate matter can explain about 29.8% of the spatial change of the level of impaired water bodies, and impaired water bodies can explain about 18.4% of the spatial change of the level of diesel particulate matter. Some relationships are expected. For example, the traf-

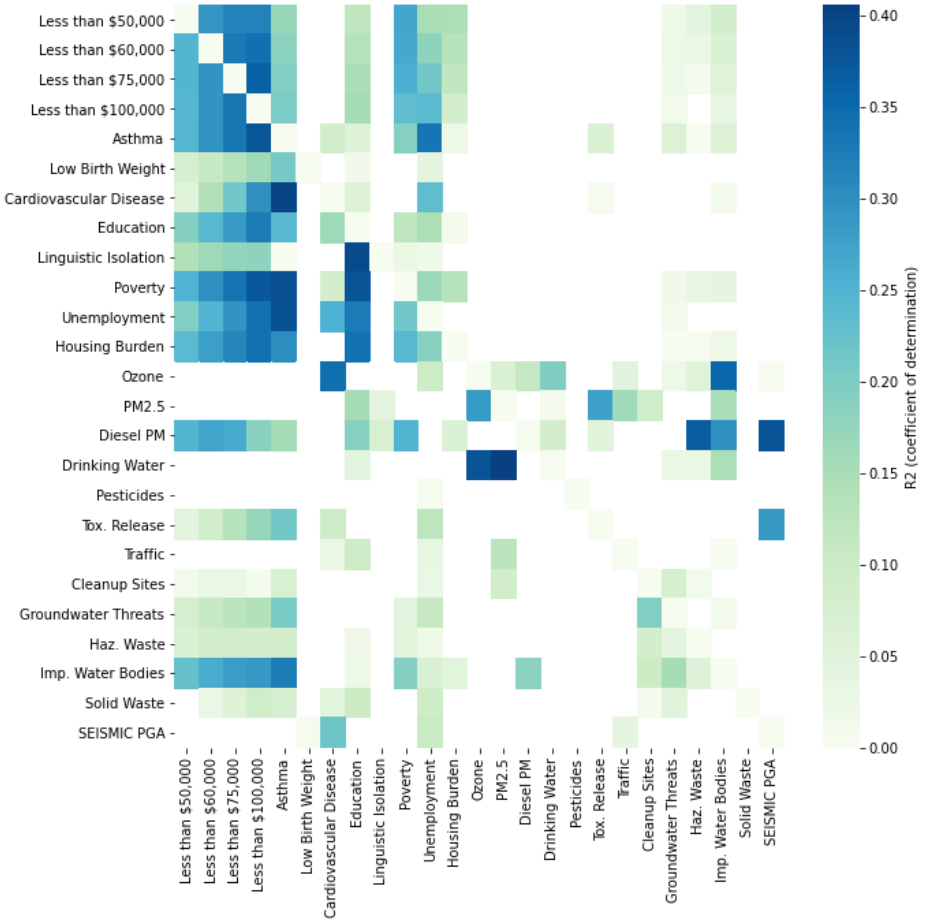


Figure 10: The heatmap of coefficients of determination for pairwise spatial cross-correlation analysis. (Blank entries correspond to pairs with no meaningful spatial cross-correlations.

fic level explains about 12.7% of the spatial change in PM2.5 level, and the unemployment level explains about 32.6% of the spatial change in education level. Theses analysis also reveal some interesting relationships, for example, that the relatively large influence (42.1% explained) that seismic PGA has on the spatial change of level of toxic release.

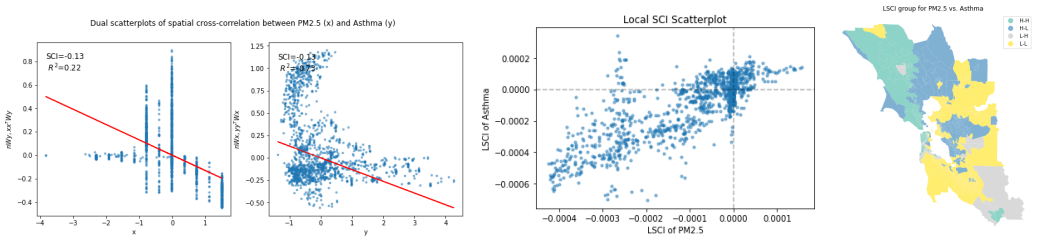


Figure 11: Scatterplots for cross-correlation analysis between PM2.5 and asthma, and the distribution of four LSCI type groups on the map. Lines of fit are colored in red.

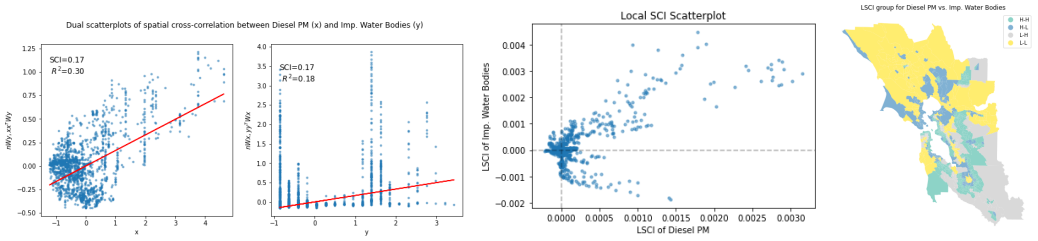


Figure 12: Scatterplots for cross-correlation analysis between diesel particulate matter and impaired water bodies, and the distribution of four LSCI type groups on the map. Lines of fit are colored in red.

2.3.3 Associations between Disease Burden and Socioeconomic-Environmental Indices

Finally, we wish to examine how environmental exposures (12 variables) and socioeconomic characteristics (5 variables) interplay and affect community health. Although CES3.0 reports three specific health outcomes, we decided to use a separate dataset independent of CES3.0. As a naive proxy for Disease Burden, the crude prevalence estimates of the 12 conditions from Health Atlas were combined using principal component analysis. The first principal component (hereafter Disease Burden) explains 73.8% of the variance and is positively associated with all the conditions. Therefore, the higher the Disease Burden index, the more a particular Census tract is affected by various conditions.

Prior to modelling, predictors were preprocessed by centering and scaling, missing data were imputed using K-nearest neighbor imputation, and pre-

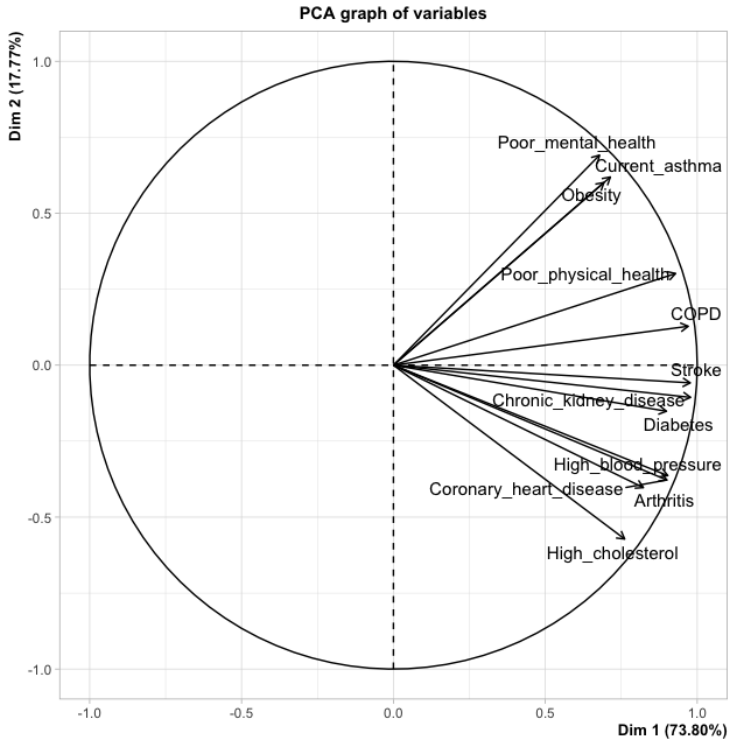


Figure 13: Results of PCA of the 12 conditions from Health Atlas with variability explained by the components.

dictors with near-zero variance removed. From the assessment of spatial autocorrelation and cross-correlation above, we need to incorporate spatial dependency: the impact of neighboring tracts affect each other. Ignoring this dependency would give biased results and underestimated standard errors.

Spatial data have two properties that add an additional layer of complexity to modelling: 1) spatial autocorrelation of variables and 2) heterogeneous behavior of predictors where different areas are influenced by the predictors differently. Spatial lag model (SLM) is a global model that isolates the spatial lag and treats it as another predictor. Geographically weighted regression, on the other hand, is able to address both properties of spatial data by essentially estimating a model for every Census tract and calibrates the model estimates by using neighboring tracts with closer tracts having more influence than farther tracts [4][5]. We explore both models and examine their

differences.

Spatial Lag Model (Global): The results of the SLM are shown in Table 3. Since the predictors were centered and scaled previously, we interpret the coefficients based on direction and magnitude. Interestingly, 10 of 12 environmental predictors showed a negative correlation with Disease Burden with Diesel PM being the most significant. In contrast, 4 of the 5 socioeconomic predictors were positively associated with Disease Burden with Poverty being the strongest predictor overall. The SLM results showed positive spatial autocorrelation of Disease Burden ($\rho = 0.53$, p-value $<.001$).

	Coefficient	Std. Error	P-value
(Intercept)	0.006	0.049	0.896
Ozone	-0.038	0.061	0.534
PM2.5	-0.150	0.063	0.017
Diesel PM	-0.255	0.063	$<.001$
Drinking Water	-0.146	0.056	0.009
Pesticides	-0.073	0.051	0.150
Toxic Release	-0.043	0.050	0.397
Traffic	-0.049	0.051	0.331
Cleanup Sites	-0.151	0.070	0.031
Groundwater Threats	-0.119	0.064	0.064
Hazardous Waste	-0.061	0.058	0.293
Impaired Water Bodies	0.191	0.056	0.001
Solid Waste	0.035	0.057	0.539
Education	0.318	0.099	0.001
Linguistic Isolation	0.077	0.081	0.342
Unemployment	0.289	0.066	$<.001$
Housing Burden	-0.442	0.079	$<.001$
Poverty	0.967	0.115	$<.001$

Table 3: Results of global spatial lag analysis of Disease Burden.

Geographically Weighted Regression (Local): The results of GWR are shown in Table 4. By allowing the predictors to vary over space, we are able to see how the predictors are associated with Disease Burden differently in different tracts and that the coefficients from the global model are between

the first and third quartiles of the GWR model. Most notable, Poverty and Education (that is, percent of population over 25 with less than a high school education) were the two predictors positively associated with Disease Burden consistently over all tracts.

The coefficients of three representative predictors are shown in Figure 14, illustrating the usefulness of local models in policy-making. The tracts rendered using the darkest colors indicate where the coefficient is the largest.

- Poverty is more associated with Disease Burden in San Francisco and southern Solano County and less so in San Mateo around the areas of Palo Alto.
- Impaired water bodies is potentially an important indicator in the South Bay, particularly in the areas around San Jose and less so in the Northeastern areas.
- Disease Burden and Unemployment are more associated in the North compared to the South.

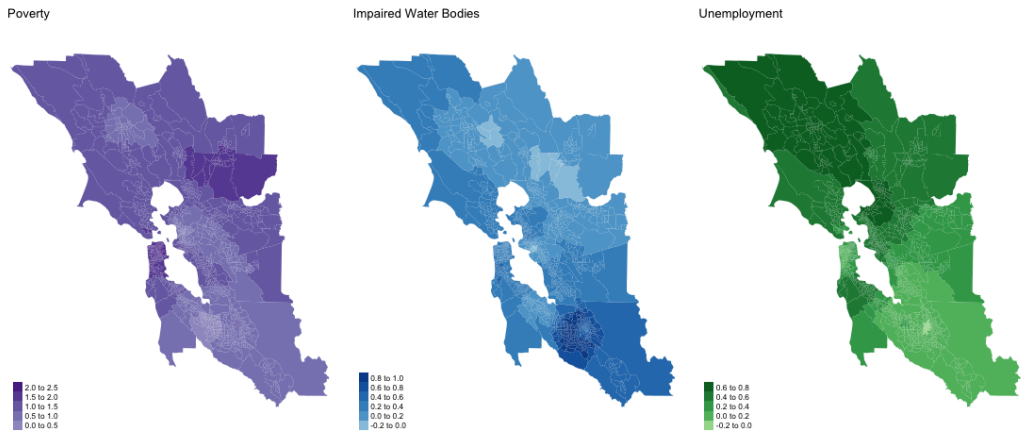


Figure 14: Three predictors showing how GWR allows for coefficients to vary across space and different influences on Disease Burden in different tracts.

Model Comparison: The root mean squared errors (RMSE) and Akaike information criterion (AIC) of the global and local models are shown in

	Minimum	1st Quartile	Median	3rd Quartile	Maximum
(Intercept)	-2.679	-0.245	0.154	0.444	2.426
Ozone	-1.605	-0.321	0.228	0.946	2.124
PM2.5	-3.161	-0.819	-0.404	-0.051	1.397
Diesel PM	-1.250	-0.481	-0.293	-0.188	0.145
Drinking Water	-1.137	-0.294	-0.093	0.049	1.008
Pesticides	-1.190	-0.115	0.067	0.480	3.043
Toxic Release	-4.922	-0.080	0.285	1.234	3.401
Traffic	-0.396	-0.062	-0.005	0.028	0.123
Cleanup Sites	-0.680	-0.518	-0.176	0.133	0.579
Groundwater Threats	-0.784	-0.137	-0.052	0.028	0.188
Hazardous Waste	-0.942	-0.262	0.008	0.139	0.423
Impaired Water Bodies	-0.101	0.123	0.261	0.385	0.968
Solid Waste	-0.178	0.010	0.105	0.185	0.356
Education	0.067	0.665	0.955	1.195	1.816
Linguistic Isolation	-1.463	-0.557	-0.070	0.405	0.892
Unemployment	-0.021	0.147	0.325	0.522	0.700
Housing Burden	-0.910	-0.554	-0.405	-0.227	0.073
Poverty	0.103	0.842	1.043	1.351	2.295

Table 4: Results of GWR analysis of Disease Burden showing the range of coefficient values for Census tracts in the Bay Area.

Table 5, both suggesting GWR provided a better fit, though both models can provide useful information depending on the level of policy implementation.

Model	RMSE	AIC
SLM	1.933	6710.6
GWR	1.853	6266.2

Table 5: Model fit measures for the two models.

3 Discussions and Future Work

3.1 Conclusions

The distributions of many of the environmental and socioeconomic factors are spatially non-random. Not only are the pollution burdens in different areas highly correlated to their spatial location, but population characteristics also show the tendency of having similar levels in census tracts that are geographically closer.

Moreover, these two sets of factors influence each other through both the direct relationship, which is the correlation between variables, and the indirect relationship, which is the spatial cross-correlation between them. The indirect interplay of different pairs of environmental and socioeconomic variables though space allows us to observe several interesting patterns. The influence that one factor has on the other in spatial change provides insight into how much these factors could interact and co-evolve given different geographical background. These indirect relationships could inform policy-making on a highly granular scale.

The spatial cross-correlation of variables also provides us with a unified framework for categorizing the different areas by the pairwise behavior of their environmental and socioeconomic factors. For each pair of variables, the distribution of the local spatial cross-correlation indices estimates how similar or different the levels of a variable and its neighbors are. Based on these statistics, areas can be categorized into different groups sharing similar relationships within the groups, and showing differing relationships across the groups. Further analysis can be done on the inherent characteristics of these areas that are causing these patterns, which will provide guidance on how policy-making should take into account the neighboring districts' influences, and promote the implementation of similar strategies in areas within the same group for addressing the environmental problems, preparing for natural disasters, lowering health risks, reducing economic inequality, etc., which will likely be more effective due to the closeness of areas revealed by these indirect patterns, and more efficient in terms of drawing experience from neighbors.

For Census Tracts in the Bay Area, disease burden was more associated

with income and socioeconomic status than environmental pollution exposure. While the negative effects of environmental hazards cannot be understated, we can conjecture that better-off communities are better equipped to deal with pollution. Some examples of this are households owning air and water filters. The result of our GWR model may be useful for local community leaders to know what to address specifically in order to reduce disease burden in their jurisdiction. Both the global model and local model showed that areas with high percentage of poverty and education less than high school was positively associated with disease burden, stressing more educational outreach, poverty relief programs, or access to affordable healthcare.

3.2 Limitations and Future Work

- Using CalEnviroScreen, a publicly-available environmental justice dataset, we cannot establish causality or make inferences on an individual's disease burden. There is also the issue of underreporting of health outcomes with the Health Atlas. People of low SES often do not seek routine care or have access to quality healthcare. These issues cannot be captured with the data sources.

- Creating a Disease Burden using PCA based on crude prevalence of select chronic conditions is a naive method of capturing health disparity. It does not consider other dimensions of disease burden such as healthy behavior, admissions, mortality, non-chronic conditions and cancer, and creating a valid index requires expert epidemiological knowledge.

- Pollution and seismic hazard only scratch the surface of environmental hazards that currently threaten the Bay Area, as climate change increases occurrence of drought, wildfires, flooding, and storms.

- The complexity of the questions addressed in this study merits both more data, and data over time; these should be incorporated in any future studies.

Acknowledgement

We would like to thank our mentors, Dr. Geetha Gopakumar and Maria Fernanda, for providing us unwavering support and guidance during the project.

References

[1] L. Anselin, "Local indicators of spatial association-lisa," *Geographical analysis*, vol. 27, no. 2, pp. 93–115, 1995.

[2] ——, "The moran scatterplot as an esda tool to assess local instability in spatial," *Spatial Analytical*, vol. 4, p. 111, 1996.

[3] Y. Chen, "A new methodology of spatial cross-correlation analysis," *PloS one*, vol. 10, no. 5, p. e0126158, 2015.

[4] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically weighted regression the analysis of spatially varying relationships*. Wiley, 2010.

[5] I. Gollini, B. Lu, M. Charlton, C. Brunsdon, and P. Harris, "GWmodel: An R package for exploring spatial heterogeneity using geographically weighted models," *Journal of Statistical Software*, vol. 63, no. 17, pp. 1–50, 2015. [Online]. Available: http://www.jstatsoft.org/v63/i17/

A Visualization and Summary Statistics of Pollution Burden Variables

	count	mean	std	min	25%	50%	75%	max
Ozone	1581.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ozone Pctl	1581.0	14.7	7.2	7.6	7.6	11.1	22.3	40.5
PM2.5	1581.0	8.7	1.1	4.5	7.9	8.7	9.5	10.4
PM2.5 Pctl	1581.0	30.8	13.7	1.9	17.8	30.7	40.9	52.6
Diesel PM	1581.0	26.1	21.3	0.1	12.1	21.0	33.7	124.7
Diesel PM Pctl	1581.0	60.4	29.6	0.3	35.2	64.7	89.0	99.8
Drinking Water	1579.0	277.9	203.3	58.5	83.5	186.9	479.2	911.5
Drinking Water Pctl	1579.0	27.4	23.9	0.7	7.3	14.1	51.0	96.3
Pesticides	1581.0	26.2	285.0	0.0	0.0	0.0	0.0	6855.8
Pesticides Pctl	1581.0	11.4	23.0	0.0	0.0	0.0	6.4	97.2
Tox. Release	1581.0	421.3	1004.0	0.0	153.7	234.0	442.5	32668.4
Tox. Release Pctl	1581.0	40.4	13.0	1.7	31.2	39.5	48.9	99.0
Traffic	1569.0	875.1	606.2	73.4	448.3	657.8	1185.2	4291.7
Traffic Pctl	1569.0	49.2	27.4	0.5	25.7	46.8	74.8	99.7
Cleanup Sites	1581.0	11.5	23.6	0.0	0.0	3.0	12.0	323.8
Cleanup Sites Pctl	1581.0	37.4	35.0	0.0	0.0	32.7	69.4	100.0
Groundwater Threats	1581.0	24.7	41.4	0.0	3.0	12.0	27.9	597.0
Groundwater Threats Pctl	1581.0	50.4	33.0	0.0	21.9	54.3	79.8	100.0
Haz. Waste	1581.0	0.5	1.6	0.0	0.0	0.1	0.3	23.4
Haz. Waste Pctl	1581.0	38.4	33.8	0.0	0.0	31.7	67.9	100.0
Imp. Water Bodies	1581.0	3.8	4.4	0.0	0.0	2.0	7.0	19.0
Imp. Water Bodies Pctl	1581.0	36.5	33.8	0.0	0.0	29.3	71.6	98.6
Solid Waste	1581.0	1.6	4.3	0.0	0.0	0.0	1.0	71.5
Solid Waste Pctl	1581.0	18.7	29.6	0.0	0.0	0.0	32.8	100.0

Table 6: Summary statistics of pollution burden variables.

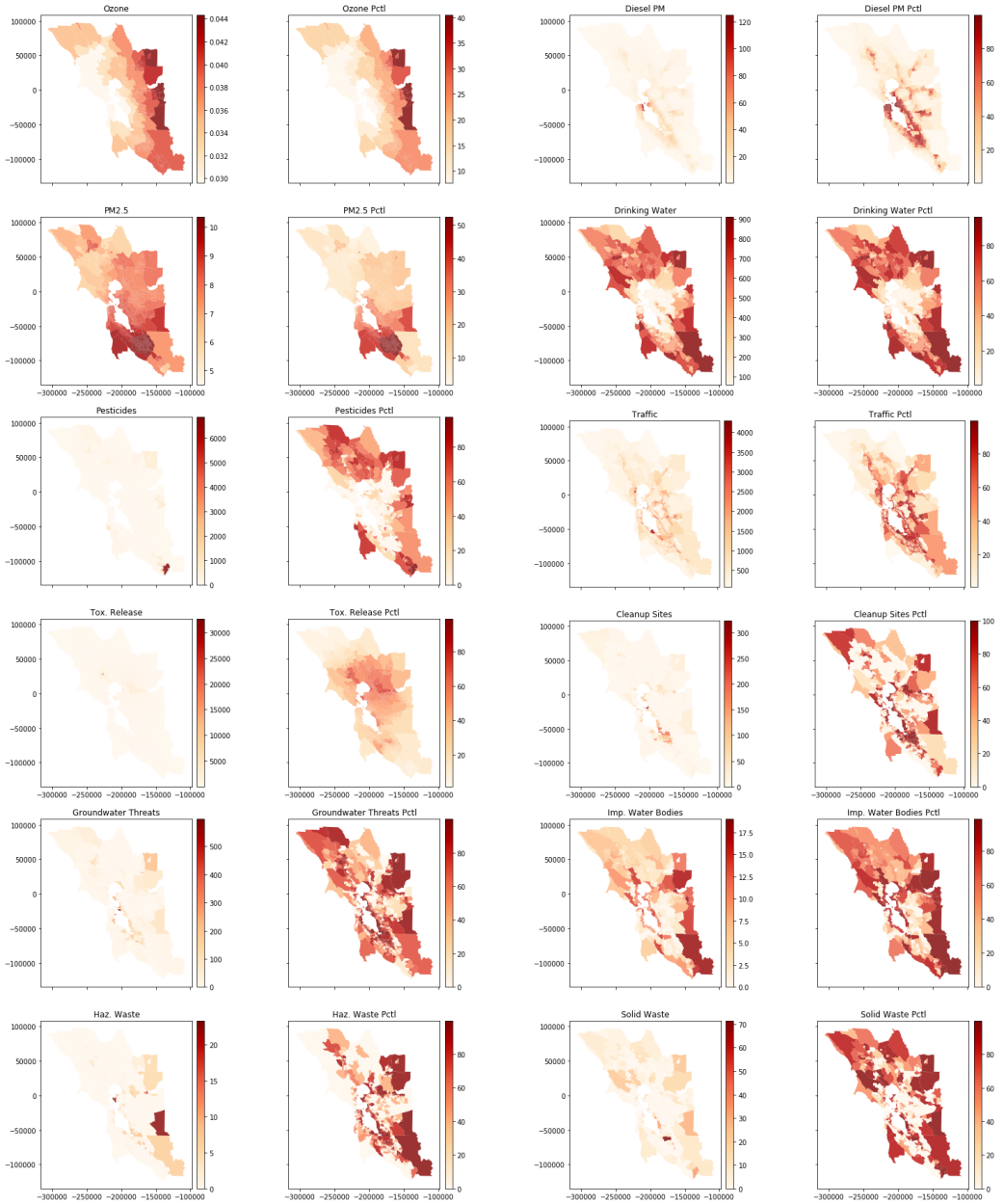


Figure 15: Pollution burden variables for each census tract area, with values color-coded on map.

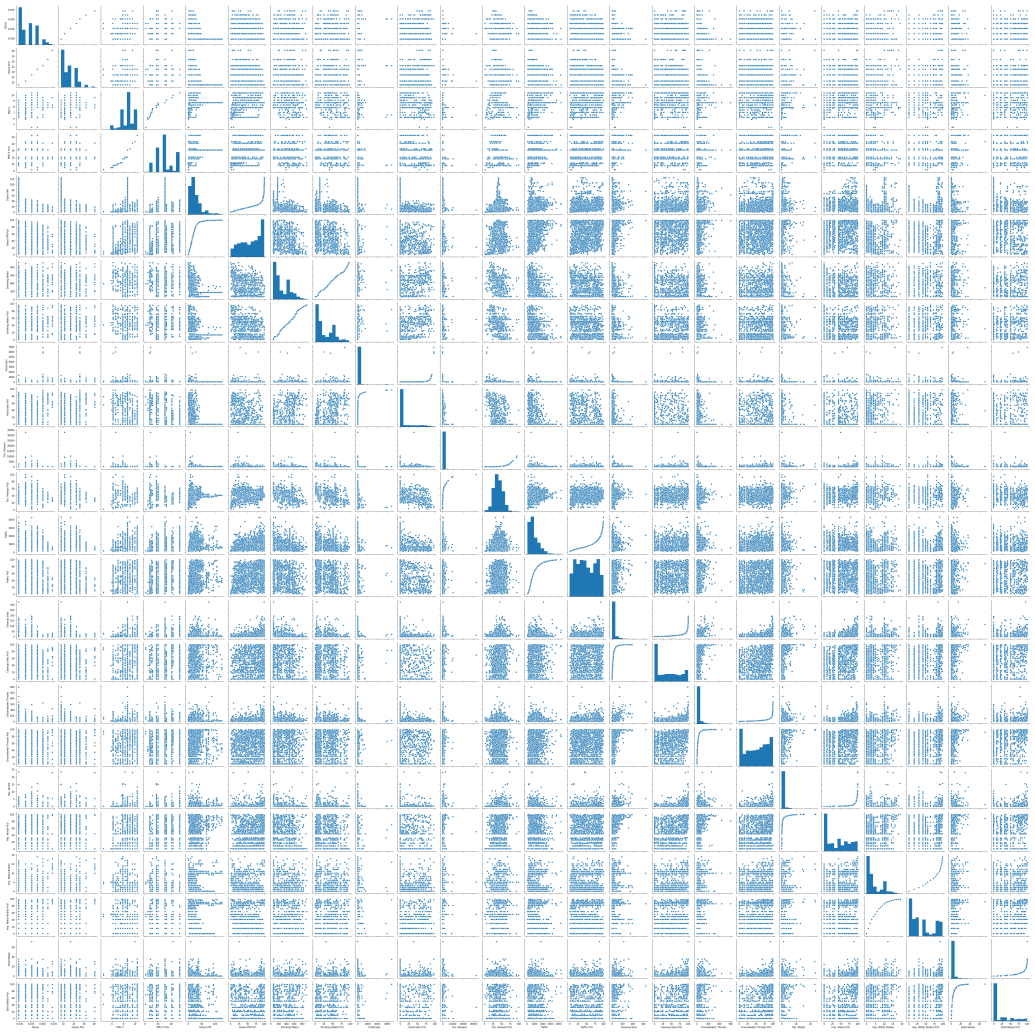


Figure 16: Pairwise scatter plot of pollution burden variables.

B Visualization and Summary Statistics of of Population Characteristics Variables

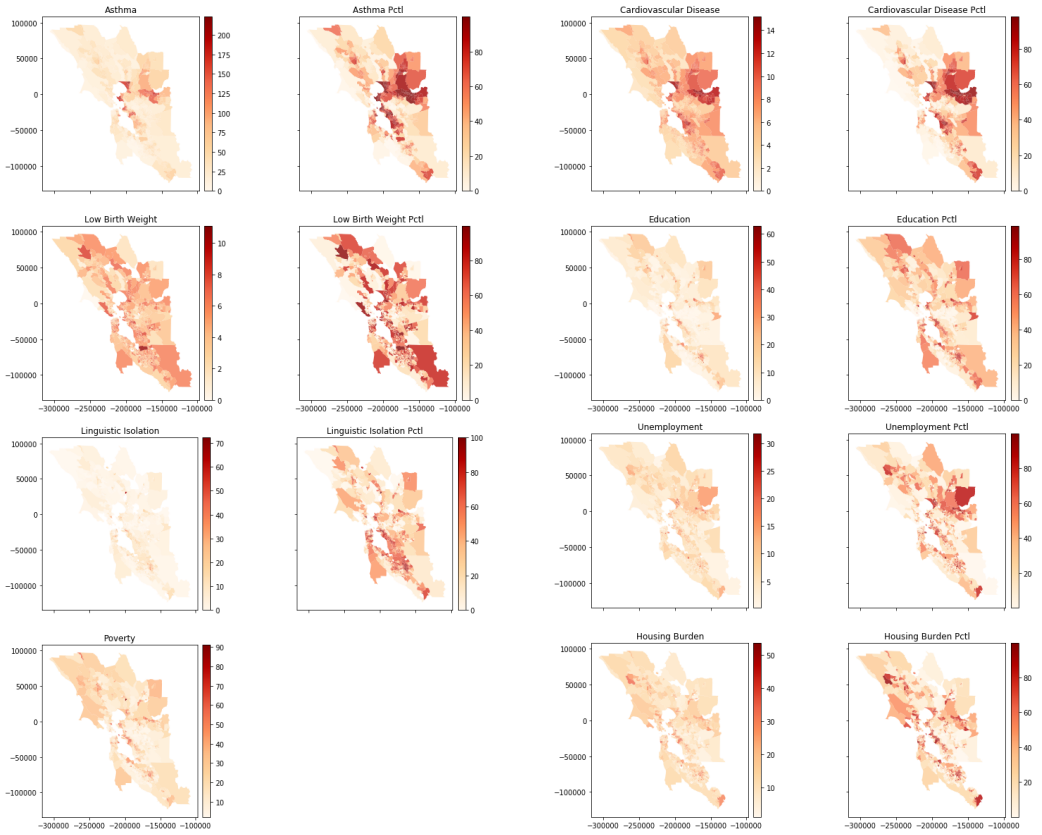


Figure 17: Population characteristics variables for each census tract area, with values color-coded on map.

	count	mean	std	min	25%	50%	75%	max
Asthma	1581.0	54.4	39.7	0.0	26.2	42.2	67.3	223.3
Asthma Pctl	1581.0	48.0	31.8	0.0	18.1	45.4	76.3	100.0
Low Birth Weight	1556.0	4.8	1.6	0.0	3.8	4.8	5.8	11.1
Low Birth Weight Pctl	1556.0	47.5	29.7	0.0	20.9	46.5	73.2	99.9
Cardiovascular Disease	1581.0	7.1	2.7	0.0	5.1	6.7	8.7	15.2
Cardiovascular Disease Pctl	1581.0	38.2	27.8	0.0	13.5	33.4	59.6	98.0
Education	1567.0	12.6	11.1	0.0	4.3	9.1	17.6	63.0
Education Pctl	1567.0	38.3	25.5	0.0	16.2	36.3	57.5	99.0
Linguistic Isolation	1544.0	9.5	8.7	0.0	3.1	7.5	13.2	72.3
Linguistic Isolation Pctl	1544.0	47.4	27.7	0.0	23.3	48.8	70.0	100.0
Poverty	1569.0	25.5	16.3	2.0	12.6	21.5	35.3	91.3
Unemployment	1557.0	8.0	4.0	0.2	5.2	7.3	10.0	31.7
Unemployment Pctl	1557.0	36.9	26.6	0.0	14.2	32.1	56.1	99.8
Housing Burden	1558.0	16.7	7.6	1.0	11.2	15.8	21.2	53.8
Housing Burden Pctl	1558.0	41.4	27.4	0.0	17.6	39.1	63.8	99.9

Table 7: Summary statistics of population characteristics variables.

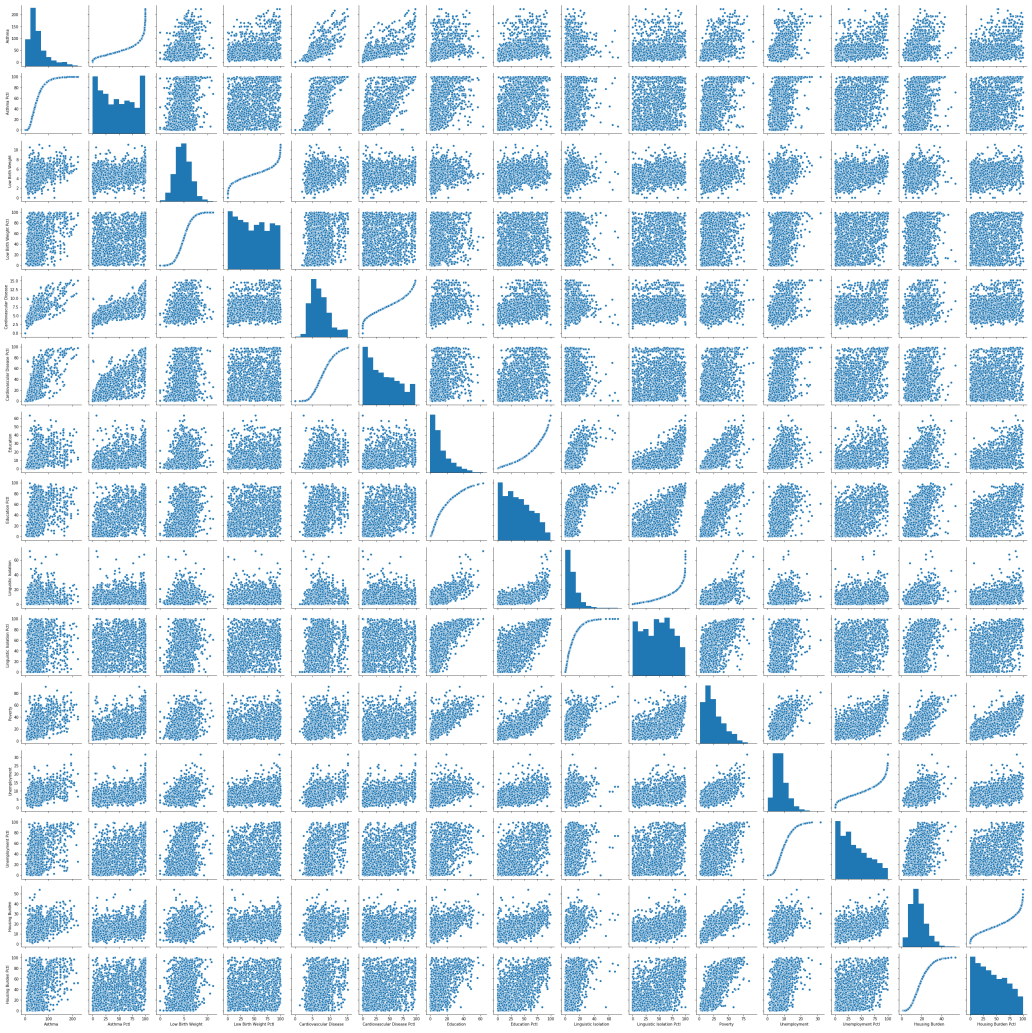


Figure 18: Pairwise scatter plot of population characteristics variables.