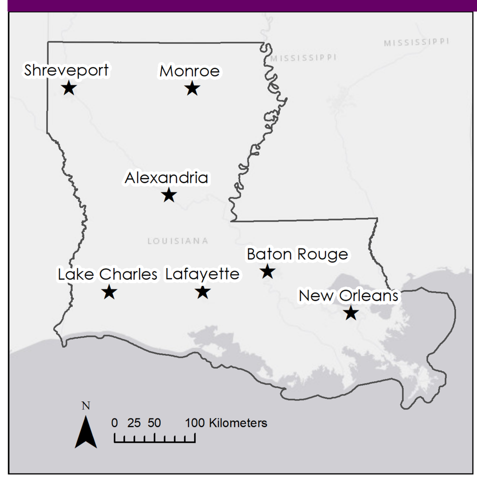
The MAUP Problem: How Much Does Spatial Unit Matter?

Mapping Industrial Pollution in EJ Communities to examine the MAUP Problem as a source of spatial error

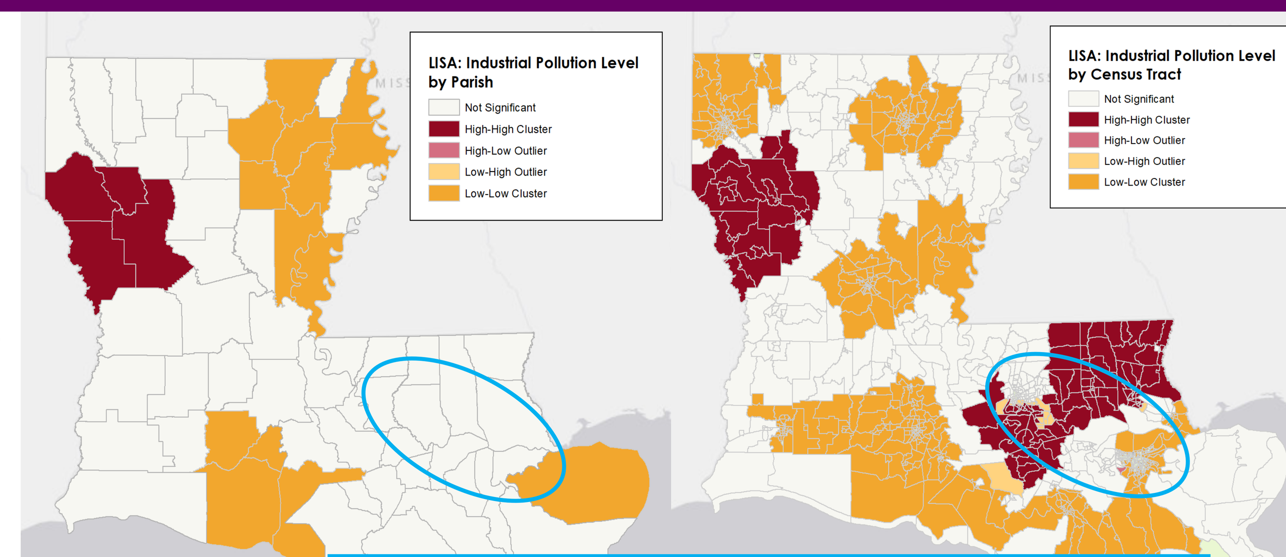
Introduction: The MAUP Problem & Environmental Justice



The MAUP problem, or the "Modifiable Areal Unit Problem" refers to the error you get when you aggregate data points into an artificial spatial unit. Depending on how you draw the lines across the data, points will aggregate into the polygons in different ways, creating potential for spatial error.

Environmental Justice communities are communities disproportionately affected by industrial communities. The presence of environmental pollutants, often from industrial plants like Coke processing facilities or chemical plants, sickens the surrounding population, causing disproportionately high mortality rates. Environmental theories suggest that systematic oppression makes these communities more likely to be low income communities and communities of color.

Research Question: How does the MAUP problem affect how we identify environmental justice communities and evaluate the impact of industrial pollution?



Blue ovals identify a region referred to as "Cancer Alley", where industrial pollution contributes to extremely high cancer incidence rates. The cluster appears at the Tract level, but not the Parish level.

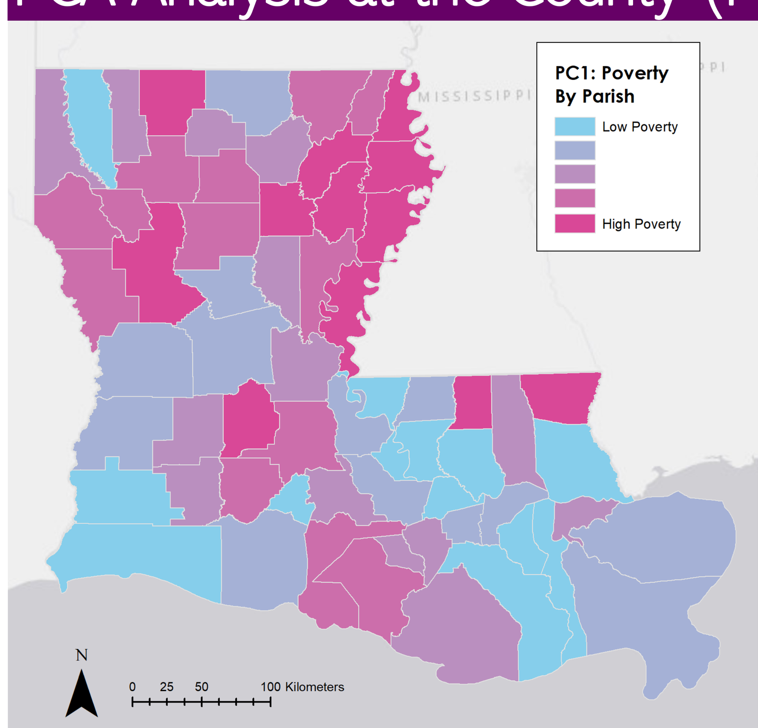
Data and Methods

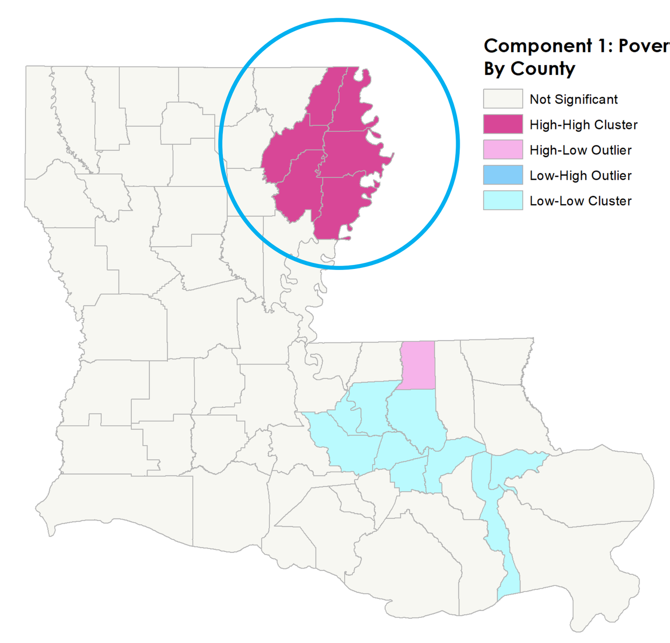
The most important part of my project consists of spatial unit: I conducted every part of my analysis at two different spatial units, Parish (Louisiana's County equivalent) and Census Tract.

First, I interpolated US EPA Toxic Release Inventory (TRI) data using the Kriging Method and then ArcMap's Zonal Statistics in order to create an industrial pollution index for Louisiana. Next, I conducted a Cluster Analysis and a Principal Components Analysis of various socioeconomic indicators using RStudio. For this analysis, I selected features from the ACS 2013-2018 5-Year Survey, which I processed in Excel. I joined this data to TIGER shapefiles for Parish and Census Tract. I mapped the results of my PCA in ArcMap and calculated Global Moran's I to determine spatial autocorrelation, and used univariate Local Moran's I's to identify clustering amongst the components.

Finally, I conducted regressions between mean industrial release per spatial unit and the PCA components. I chose the regression method for each spatial unit based on the spatial dependence diagnostics: for Parish, I conducted an Ordinary Least Squares regression, while for Census Tract, I conducted a Spatially Lagged Regression.

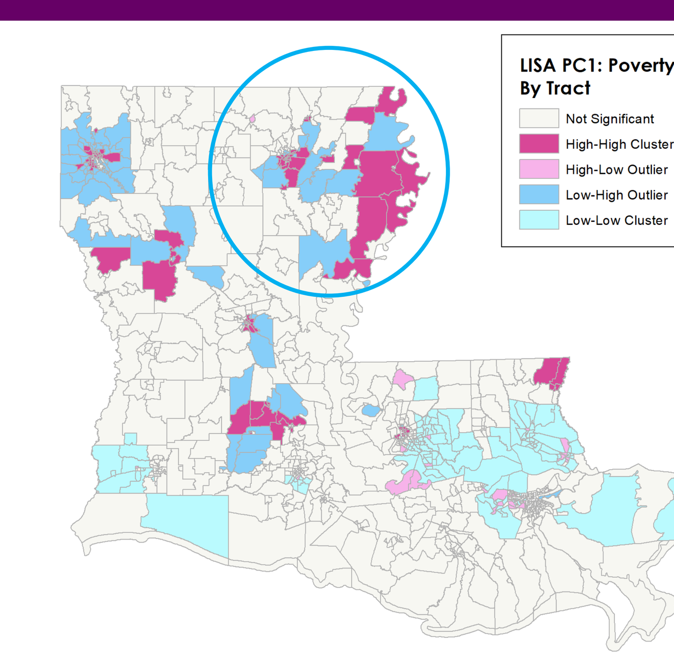
PCA Analysis at the County (Parish) Level and Tract Level

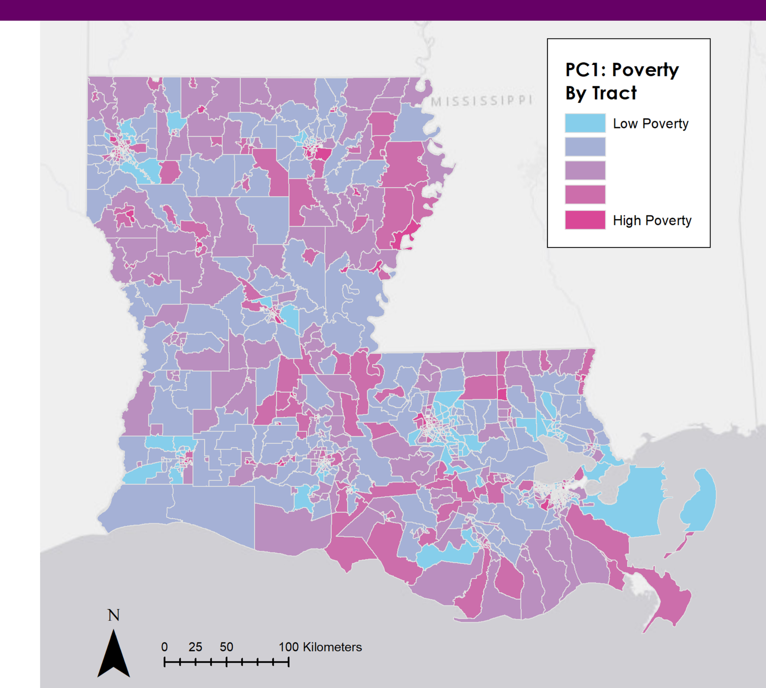




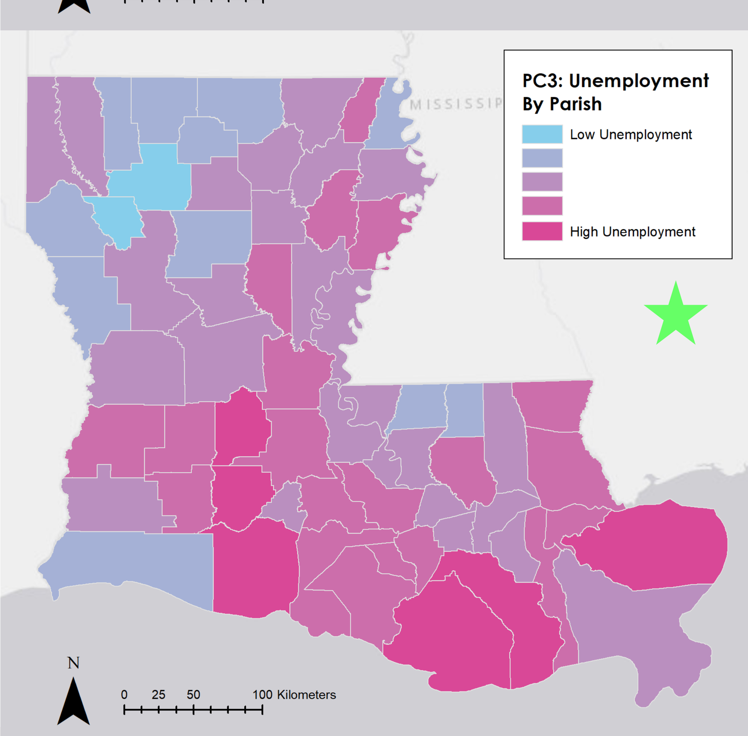
PC1: an example of the of the MAUP Problem

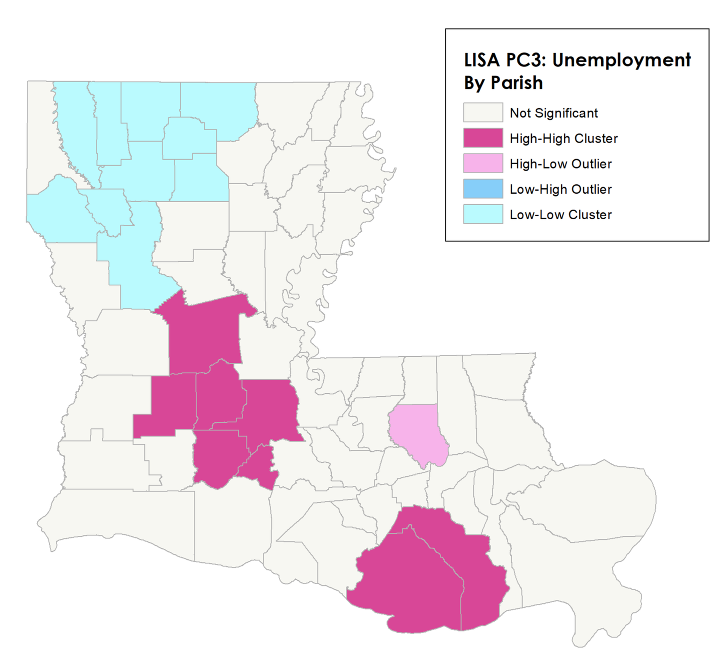
Counties show the area as one, uniform unit of high poverty, while tracts show clusters of high income. Same data, same analysis, two vastly different stories.

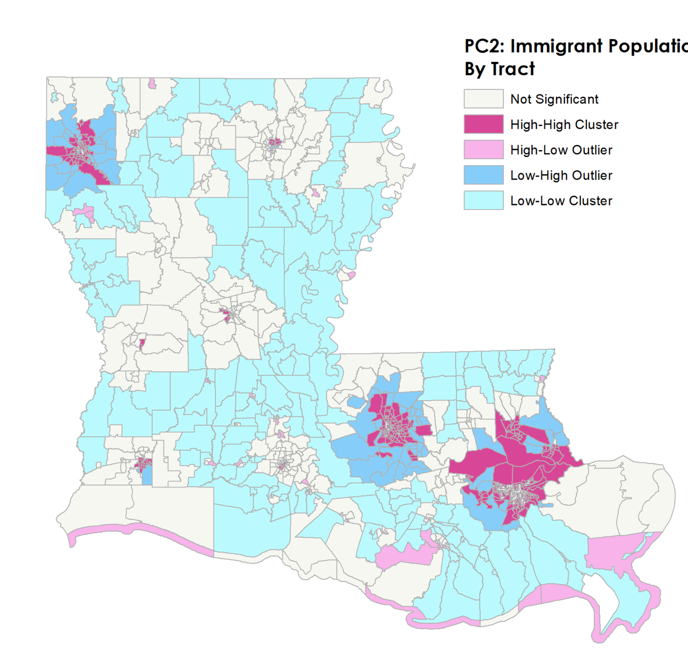


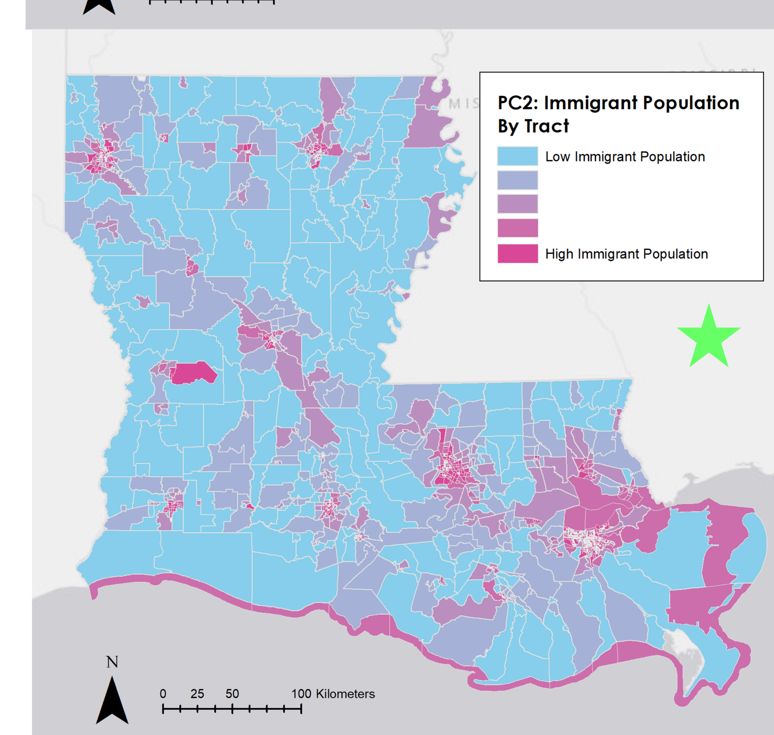


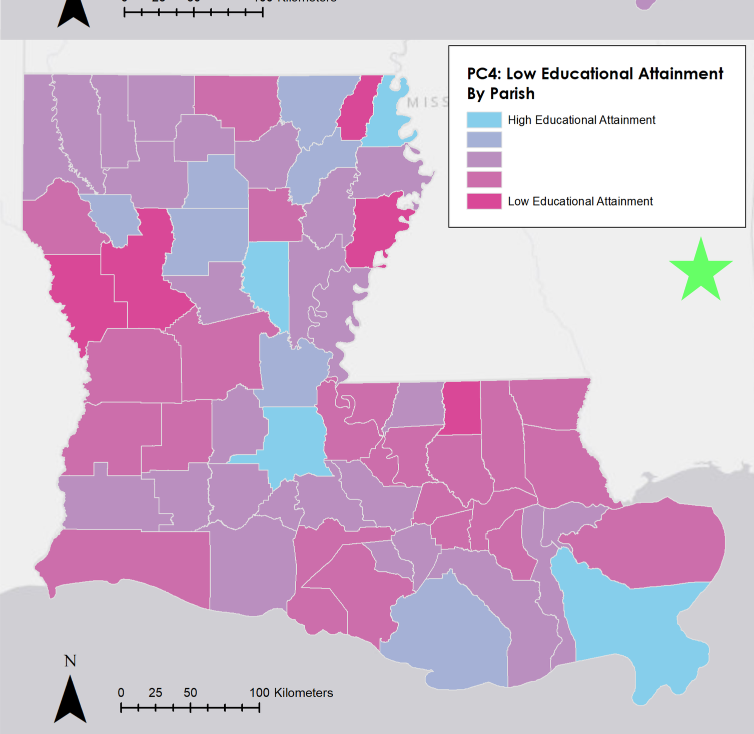
★ = statistically significant component

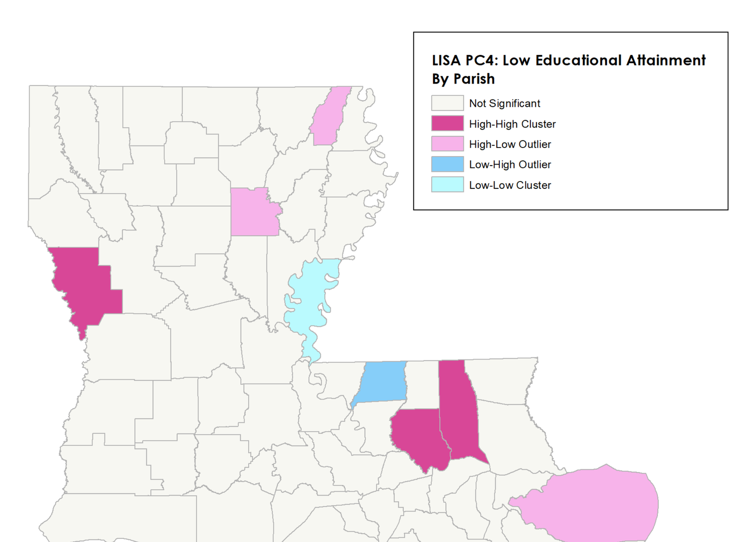


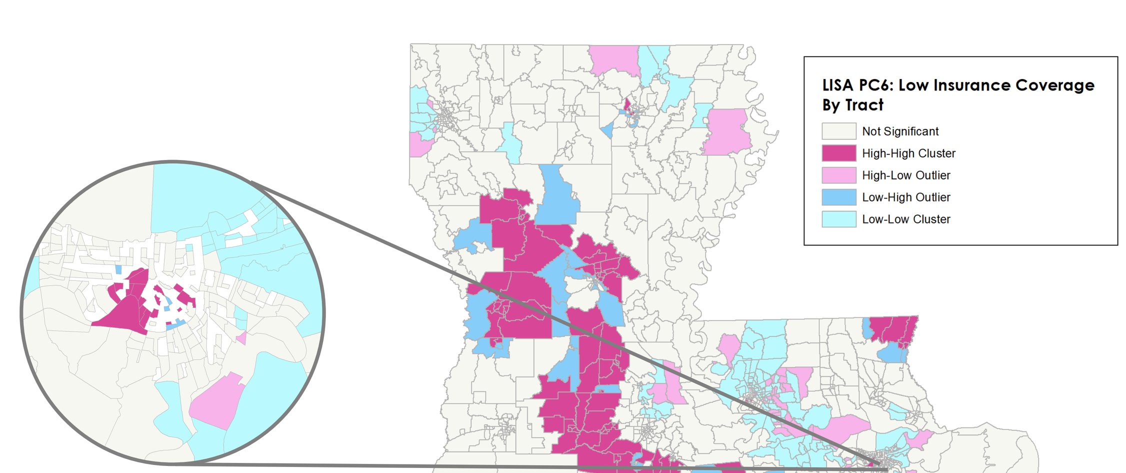




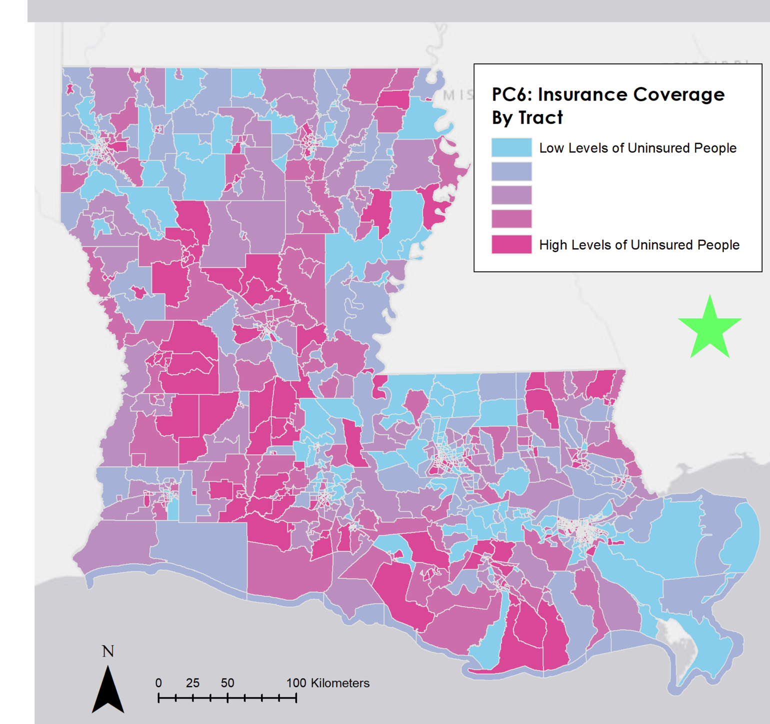








Central New Orleans is a cluster of low health insurance coverage



Statistical Analysis: Do these components help to explain Industrial Pollution Patterns in Louisiana?

PCA Results , OLS Regression Results Between Industrial Pollution & Principal Components

Component	Description	Explained Variance	Coefficient	P-Value	Significant?
PC1: Parish	Poverty Index	45.5%	-887.0	0.85	No
PC2: Parish	Immigrant Population	20.4%	-6571.5	0.22	No
PC3: Parish	Unemployment	10.9%	-25266.1	0.003	Yes
PC4: Parish	Low Education	8.77%	15825.2	0.07	Yes*

PCA Results , Spatially Lagged Regression Results Between Industrial Pollution & Principal Components

Component	Description	Explained Variance	Coefficient	P-Value	Significant?
PC1: Tract	Poverty	44.9%	-4.26	0.99	No
PC2: Tract	Immigrant Population: Low Income Renters	15.6%	-789.9	0.07	Yes
PC3: Tract	Immigrant Population: High Income	10.8%	1201.7	0.09	No
PC4: Tract	Unemployment	8.23%	-49.5	0.95	No
PC5: Tract	Low Education, Home Owners	7.04%	-206.5	0.81	No
PC6: Tract	Insurance Coverage	5.66%	-2477.6	0.01	Yes

Discussion, Limitations & Conclusions

The MAUP problem is clearly an issue when examining the impacts of industrial pollution, as the results I obtained at the Census Tract level vs. the Parish level were quite different. It's clear through looking at both the LISA Cluster Maps and the PCA that using Parish spatial unit gives large summaries of the data, while using the Census Tract spatial unit reveals more nuances. PC1 at both Tract and Parish level indicates that poverty is the main factor tying the PCA variables together, with PC1 explaining about 45% of the variance respectively. LISA analysis of PC1 indicates that poverty in Louisiana is mostly clustered in the Northeast near the Mississippi border, with smaller pockets in Shreveport and Lafayette, not in Cancer Alley as one would assume. However, according to my regression results, poverty is not the best indicator of industrial pollution: at both the tract and parish level, the correlations between PC1 and industrial pollution level were not significant. Furthermore, past PC1, the rest of the variance is explained differently depending on the spatial unit. At the Tract level, 6 components represent the same amount of variance as 4 components at the Parish level, indicating that data at the tract level captures more variation.

In terms of regression results, the components overall are not good indicators of environmental pollution. At the parish level, OLS regression shows a negative correlation between Unemployment and pollution, meaning that as pollution increases, unemployment decreases by a factor of –25,266.1. It also shows a positive correlation between Low Educational attainment and pollution, meaning that pollution increases, the population of undereducated people also increases by a factor of 15,825.2. Both of these correlations make sense: the first indicates that people who live near industrial pollution are not likely to be unemployed, which makes sense as these people would tend to be employed by these polluting industries. Furthermore, industrial jobs tend to require unskilled labor, meaning they tend to attract people with lower educational attainment. t the Census Tract level, Spatially Lagged Regression shows a negative correlation between Low Income immigrant population and pollution: as pollution increases, the low-income immigrant population (PC2) decreases by a factor of –789.9. It also shows a negative correlation between No Insurance Coverage and pollution: as pollution increases, the population of people without insurance decreases by –2477.6. The second correlation makes sense, as people living near industrial pollution tend to be employed, they must also have insurance coverage. The first correlation, however, is confusing as environmental justice theories postulate that industrial pollution has a disparate affect on low income people of color. This could indicate that industrial pollution affects homeowners more than renters: if people living near industrial pollution are likely to be employed, it could mean they are more likely to own homes. However, this could also indicate the presence of spatial error as a result of the MAUP problem. It could also represent error in our understanding of pollution data: we used total onsite releases as an indicator of industrial pollution, and assumed that communities nearest industrial facilities were most affected by industrial pollution. It could be that the level of release is a bad indicator, and that the toxicity of the release should be measured instead, or that industrial pollution travels and has greater impact on communities that aren't necessarily the closest, depending on how the pollutants are disposed of.

Given these potential sources of error, if repeating this analysis, I would examine the correlations between cancer incidence rate and these principal components. This could help determine who is most impacted by industrial pollution.

Cartographer: Kenia French
UEP 236: Spatial Statistics, Sumeeta Srinivasan

Data Sources: US EPA, ACS 5-Year 2013-2018
NAD 1983 UTM Zone 15N, Transverse Mercator
Spring 2020

Tufts