A Tale of Two States

Demographic Risk Factors for Drinking Water Contaminations in the U.S.

INTRODUCTION

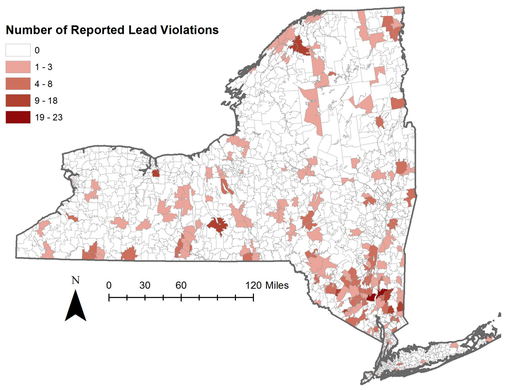
In April of 2014, the city of Flint Michigan switched its water supply from Detroit's system to the Flint River. In May of 2015, Flint residents began to complain of foul odor, taste and color of the water but it wasn't until February of the following year that city officials finally tested the water and found an astonishing 104 parts per billion in a sample of tap water (Lin et al. 2016). It took almost three years before drinking water contained lead levels within government standards. Exposure to any amount of lead during early childhood can lead to cognitive and behavioral changes that are irreversible so it is crucial that areas at risk are identified before any lead exposure occurs. The Flint water crisis brought to light the serious implications of lead contamination, resulting in an increased focus on the public health issue of water contamination, especially in urban areas. Since the first reports on Flint water quality were generated there have been a variety of analyses done for the area of Flint, but these analyses, which considered factors such as race, poverty and other socioeconomic factors, were limited to the Greater Flint area (Hanna-Attisha et al. 2016). In this project, I will be expanding this type of analysis to attempt to determine if specific socioeconomic factors are good indicators of areas at high risk for drinking water contamination. Previous researchers based contamination risk on age of housing and poverty and concluded that urban areas, especially old industrial cities, are at highest risk of lead contamination (Frostenson et al. 2016). Based on this I selected two states of interest, New York and Texas, to compare patterns of clustering.

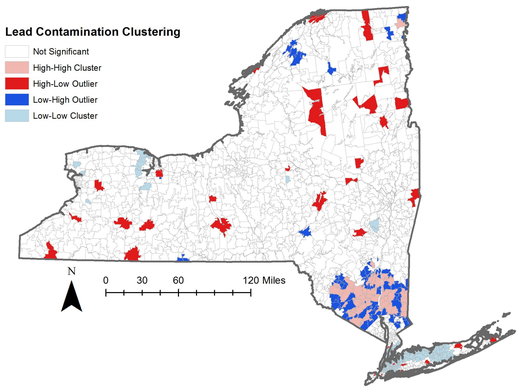
METHODOLOGY

Data was collected from the American Community Survey conducted in 2010 from the Tufts M Drive and data on drinking water violations was gathered with the help of Jessie Norriss. Following data collection, the reported lead contaminations were geocoded based on associated zip code and joined to zip code centroids and subsequently joined to demographic data from the American Community Survey collected in 2010. Chosen demographic variables of interest included income, housing value, age of housing, level of education and poverty. After extraneous data was removed, and variables of interest were normalized, spatial analysis was performed in ArcMap and GeoDa. Analyses were first conducted on the U.S. as a whole and then areas of interest were selected and further analyses were conducted on these two states. Spatial analysis methods included local and global Moran's I's as well as an Ordinary Least Squares (OLS) Regression for each state as well as the entire US.

CASE STUDY 1: NEW YORK

The state of New York was chosen as the first area of interest based on the Vox report which identified New York as an area with a high concentration of risk due to New York City's aging urban infrastructure.



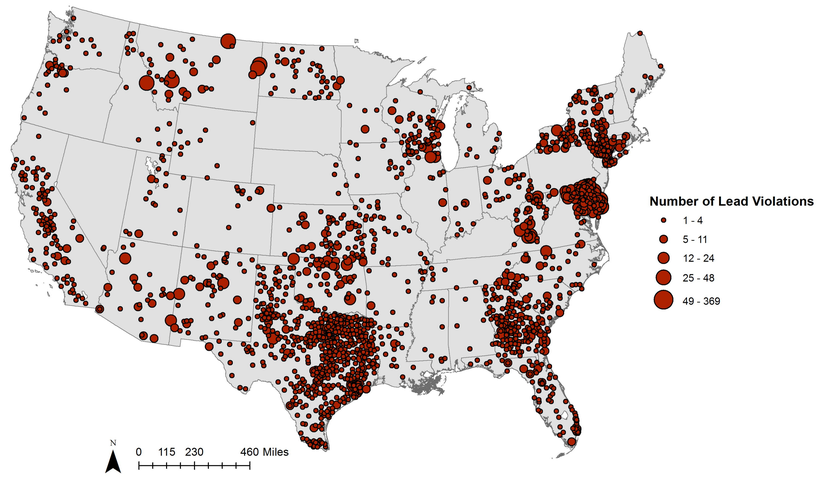


RESULTS

Cluster Analysis: The results of the Local Moran's I cluster analysis showed significant clustering in several zip codes in both New York and Texas. Unsurprisingly, there was an abundance of high-high clustering in the New York City area with many sections of central and upstate New York containing no significant clustering. This is consistent with findings from earlier reports linking urban areas to a greater number of lead violations. However, the Long Island area contained several low-low clusters even though it is one of the most densely populated areas in the state. In comparison, the state of Texas included several low-low clusters in urban areas near the cities of Dallas and San Antonio.

Regression Analysis: At the state level, results of the OLS regressions varied based on normalization of the chosen demographic variables and when variables were normalized the results were not significant at the 5% level. For example, in both Texas and New York the number of households below the poverty level showed a significant negative correlation with number of reported lead violations. However, once this variable was normalized by number of households in each zip code there was no statistical significance. At the country level, the results of the OLS regression did appear to be significant for all variables tested however some areas particularly the Midwest and Rocky Mountain area, had very few data points.

LEAD VIOLATIONS IN THE U.S.



Areas of Interest

U.S. REGRESSIONS

Dependent variable: Total number of reported lead violations

Independent Variable	Coefficient	Probability
Median Household Income	1.41×10^{-6}	0.00002*
Median Housing Value	2.14×10^{-7}	0.00003*
Median Year Built	7.35×10^{-5}	0.0517
Total Population	2.26×10^{-6}	0.00001*
Total Households	5.87×10^{-6}	0.00002*
Total Housing Units	8.087×10^{-6}	0.00000*

NEW YORK REGRESSIONS

Dependent variable: total number of lead violations in each zip code

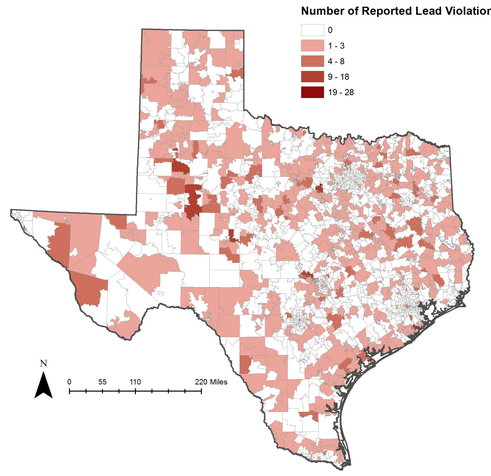
Independent Variable	Coefficient	Probability
Total Population	6.26×10^{-5}	0.00005**
High School Graduates*	6.86	0.184
College Graduates*	6.86	0.184
Renter Households	0.000163	0.0334**
Renter Households*	0.0436	0.977
Owner Households	0.000381	0.0000**
Owner Households*	-0.0448	0.976
Households below Poverty	0.000562	0.0163**
Households below Poverty*	-0.252	0.942
Number of houses built between 1970 and 1979	0.00140	0.00000**
Number of houses built between 1930 and 1939	0.000202	0.0614
Median Household Income	1.93×10^{-5}	0.0396**
Median Housing Value	2.46×10^{-6}	0.0938
Median Year Built of Housing Units	0.0502	0.0242**

A single asterisk indicates a normalized variable, for education demographics the number of graduates was normalized by population totals in each zip code and for household demographics the variables were normalized by number of households in each zip code.

DISCUSSION

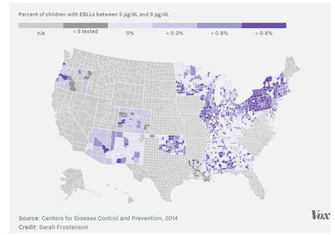
While there was significant clustering in reported lead violations in both Texas and New York the subsequent regressions resulted in mostly insignificant findings with normalized demographic variables. Significance was detected for the entire U.S. but all findings were limited by the amount of data available on drinking water contaminations. However, this is not to say that demographics play no role in areas at highest risk for drinking water contamination, as previous studies have demonstrated poverty and age of housing are two such factors with significant correlations. Part of the issue in this field of public health study results from the fact that government organizations such as the Centers for Disease Control (CDC) do not require states to submit lead exposure data, thus leading to huge gaps in data available as shown in Figure 1 (Frostenson, 2016a). In 2014 only half of U.S. counties reported lead poisoning data and within these reports some counties reported no confirmed cases of lead whereas some counties reported greater than 10% of blood tests reported lead in the bloodstream (Frostenson, 2016a). To truly determine the best predictors of drinking water contamination, more data will have to be gathered and analyzed to get a truly representative data set for the entire U.S.

CASE STUDY 2: TEXAS

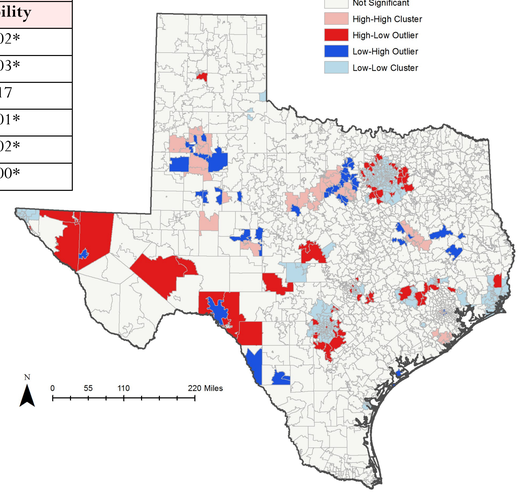


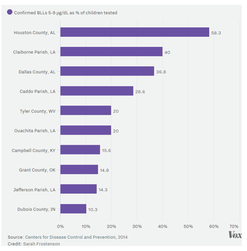
A second case study was conducted for the state of Texas, which was chosen based on the amount of available data on reported lead violations as well as to serve as a contrast to New York. While New York is expected to be home to a high number of lead violations due to New York city's aging infrastructure, the major cities in Texas were constructed more recently and thus are hypothesized to carry a lower risk of potential lead contamination.

Figure 1



Source: Frostenson, 2016a





Source: Frostenson, 2016a

TEXAS REGRESSIONS

Dependent variable: total number of lead violations in each zip code

Independent Variable	Coefficient	Probability
Total Population	-6.55×10^{-6}	0.00313**
High School Graduates*	0.208	0.741
College Graduates*	0.382	0.534
Renter Households	-3.44×10^{-5}	0.0124**
Renter Households*	-0.232	0.296
Owner Households	-2.12×10^{-5}	0.0385**
Owner Households*	0.280	0.203
Households below Poverty	-0.000103	0.00252**
Households below Poverty*	-0.0447	0.916
Number of houses built between 1970 and 1979	-6.62×10^{-5}	0.00850**
Number of houses built between 1930 and 1939	5.36×10^{-6}	0.947
Median Household Income	4.97×10^{-7}	0.793
Median Housing Value	4.61×10^{-7}	0.328
Median Year Built (Housing)	0.000331	0.218

A single asterisk indicates a normalized variable, for education demographics the number of graduates was normalized by population totals in each zip code and for household demographics the variables were normalized by number of households in each zip code.

Cartography by Valerie Willocq. GIS 102 Fall 2017.
Projection: NAD 1983 Texas Centric Mapping System Albers, NAD 1983 State Plane New York Central FIPS, North American Albers Equal Conic
Data Sources: Tufts M Drive, Lead data provided by Jessie Norriss

Literature Cited: Frostenson, Sarah. "America's Lead Poisoning Problem Isn't Just in Flint. It's Everywhere." *Vox.com*, 21 Jan. 2016.
Frostenson, Sarah, Kliff, Sarah. "Where Is the Lead Exposure Risk in Your Community?" *Vox.com*, 6 Apr. 2016
Hanna-Attish M., LaChance J. Sadler R.C., Schnepp A.C. "Elevated Blood Lead Levels in Children Associated with the Flint Drinking Water Crisis: A Spatial Ananlysis of Risk and Public Health Response. *American Journal of Public Health*. 2016: 283-290.
Lin, Jeremy C.F. et al. "Events That Led to Flint's Water Crisis." *The New York Times,* 21 Jan. 2016.

Tufts UNIVERSITY