Can 120 Seconds Make a Difference?: A Statistical Analysis of the Flint Water Crisis

Samuel Isken, Xinyi Wang, Zhishan Li, Alison Cronander
Michigan State University

Abstract

The Flint Water Crisis has had, and continues to have, a lasting impact on generations of residents in the city of Flint as well as the socioeconomic level of the city as a whole. More background will be provided throughout the paper, but in essence, the term “The Flint Water Crisis” refers to the situation currently occurring in Flint, Michigan where residents of the city are without clean drinking water due to the presence of toxic chemicals in the water system. This paper focuses on a statement made often by city officials and tests its accuracy. The common statement made by the city was that turning on the water and allowing it to run for a set amount of time would allow a clean flow of water. We will test the accuracy of this statement by examining a dataset looking at the levels of a specific indicator chemical (which presence indicates unclean / unsafe water conditions) at 0 seconds after initiation, 45 seconds after initiation and 120 seconds after initiation (initiation refers to the action of turning on access point to water supply, for example faucet, shower, etc.). Our analysis is designed by using a regression to examine the relationship between the variable “time’ (t) and the level of chemical presence (pb). We model our regression using an exponential decay function and examined the results in order to present our conclusion. Our paper is organized by looking first at background information regarding “The Flint Water Crisis” as well as our data and modeling strategy. Second, we will graphically examine our data in order to explore it. Finally, we will present the results of our regression, discuss our methods and apply our results to current policy.

Keywords: Flint Water Crisis, Exponential Decay Function, Ward Code, Environmental Protection Agency, Chloride, Linear Regression, Non-Linear Regression, Non-Linear Least Squares Method, Indicator Variable

Imagine waking up one day and not knowing where you would get your water for the day. You would likely wonder how you would go about your day lacking the ability to complete the most basic tasks such as shower, cook and drink. This was, and still is, a reality for the people of Flint, Michigan and is a massive public health crisis. Numerous statements have been made by government officials in order to ensure citizens that the water they are consuming is safe. However, many of these statements did not take into account scientific analysis.

Our paper will seek to analyze a common sentiment made under the tenure of Governor Rick Snyder. The idea was that citizens of Flint could let their water run for up to two minutes and the lead chemicals would "fall out" ensuring the water was safe to drink. In our work to conclude whether this statement is accurate we will touch on the following topics:

1. General Background on the Flint Water Crisis, Root Causes and Public Backlash
2. Data Set and Strategy, Graphical Analysis
3. Mathematical Analysis & Regression
4. Results and Concluding Remarks Regarding Policy Implications

The Flint Water Crisis began in 2014 when the city of Flint switched their water source from Lake Huron and the Detroit River to the Flint River. The city had been in a financial emergency and made the switch with the hope of saving $5 million, however the switch caused many unforeseen issues. The Flint River water caused the pipes to corrode due to the high levels of chloride.

People of Flint began noticing their water was coming out of the tap orange, but the Michigan Department of Environmental Quality (MDEQ) insisted that it was okay. When samples were taken from homes, many biases came into play. For example, before collecting the testing samples, the preparation instructions given were meant to lower the levels of lead and the samples were collected in bottles which could only take in water at a low-pressure, which could have lead to underestimated results.

The state has now issued 43 criminal charges against 13 officials. The residents of Flint also filed a class action lawsuit against Governor Snyder and others that were involved the corruption that lead to Flint's negative trajectory.

We examined a data set containing water samples from 270 households from eight different zip codes and ten different wards. There was initially 271 samples, however, there existed a massive outlier that appeared to be the result of human error. Under advisement we removed this outlier as it did not fit the trends we were seeing in the data and was essentially scientifically impossible. The graphics titled "Pie Chart of Ward Code Quantity" and "Pie Chart of Zip Code Quantity" provide information on the composition of data with regards to zip code and ward code. (All graphics are located in our appendix attached to the report.)

Ward denotes a variable dependent on location similar to zip code. Each household sample contained three unique data points denoting the lead level in water drawn after waiting a specific amount of time (0 seconds, 45 seconds, and 120 seconds).

The result of the graph titled “Correlogram of Flint Lead Data (All Variables)” shows us that there does exist correlation at a relatively high level (.6) between the variable relationships "First & Third" as well as "Second & Third". We would expect these to not be equivalent if the time the water runs truly was reducing the lead content. Though, the correlation is not enough so while this was a good starting point much more analysis was clearly needed.

From the graphs titled “Pie Chart of Ward Code Quantity” and “Pie Chart of Zip Code Quantity” we learned that some zip codes only contained 1-2 samples, the rest contained roughly the same amount of samples which were much higher numbers.

After examining our data from this high level We decided to do this by examining whether there appeared to exist a relationship between geographical area (Zip and Ward Codes) and lead level. We produced the two graphics title “Lead Level Comparison Between Zip Codes (First)” and “Lead Level Comparison Between Ward Codes (First)” in order to explain this. After examining the initial draw lead count against "Zip Code" and "Ward Code" we stopped. It became very clear there was no clear relationship/correlation between area and lead content. It was true that some areas contained higher levels, however there is not clear relationship within the data. It became clear this issue does not discriminate by area and all of Flint was effected.

Upon this discovery we had to re-think our methodology and manipulate our data. We now knew that zip code / ward code / location was not going to play a large role in our study when concerning the relationship between time and lead level in water. We needed to find a way to rework the data into a more useful format in order to examine the relationships between "First", "Second" and "Third" and look at how lead content was changing over time. The column "t" represents time quantified and is dependent on the column "Time". "Time" can be considered and indicator variable while "t" is the time column we will be using for analysis and computation. We have attached the data (in the RMD file) so the reader can examine the rework for themselves.

Next, upon recommendation per prior information provided we decided to use an exponential decay function to model our data using a regression. The exponential decay function is displayed below to remind readers of the makeup of the function.

$$f(x, \theta_1, \theta_2) = \theta_1 e^{-\theta_2 x}$$

Initially we used the NLS (Non-linear Least Squares Method) and the LM (Linear Model) method. We used the standard form of the exponential decay function and estimated our parameters to be:

$$(\theta_1 = 2,\ \theta_2 = -0.01)$$

We chose these parameters after running a regression on the natural log of the data. We used our results in order to estimate the parameters of the NLS model. After running these

regressions we were able to make some conclusion on our data. The main result we examined and used for the following is the “plot(all_zip_lm)” results.

1. Based on the results from the plot functions which is “QQ-Plot”, the results of our residuals are considered normal. (Graph#2 Normal Q-Q)
2. Based on the Residual vs. Fitted plot we determined our residuals do not possess heteroscedasticity, thus our variances remain equal throughout our range of values of variable that predicts it. (Graph #1 Residuals vs. Fitted)
3. We examine Cook's distance (which is used to estimate the influence of a single data point) to determine our X variables do not have undue influence on our model's result.

These three results confirm that our model is accurately showing a relationship between lead content in water and time. The accuracy of our Linear Model implies (by mathematical computation) the Non-Linear Least Squares Method using the exponential decay function also accurately models our data. To examine how much influence time has on lead content we used both our R-Squared value and our slope.

Our R-Squared Value was small (roughly .128). This means only 12.8% of the variability in lead levels can be attributed to a change in time. Our slope was also only roughly equal to -.01.

Let us now examine the policy implications of our results. In order to do this, we first had to convert the data form parts per billion to parts per million and then calculate the outliers from each the first, second, and third draw. We calculate the mean of each draw and created a plot titled “How Many Houses are Above Average?” that showed the lead in each house with respect to the average lead level line in Flint. This gave us an easy way to visualize how many houses in have a lead level that is above the average lead level in Flint.

The EPA says “Lead and copper are regulated by a treatment technique that requires systems to control the corrosiveness of their water. If more than 10% of tap water samples exceed the action level, water systems must take additional steps. For copper, the action level is 1.3 mg/L, and for lead is 0.015 mg/L.” We used these guidelines to write a code that looped through the data for the first, second and third draws, counted how many were above the limit of 0.015 ppm and then turned it into a percentage of houses needing action.

Our percentage for the first draw was 16.67% of samples, 6.30% for the second draw, and 4.44% for the third draw.

We concluded the following:

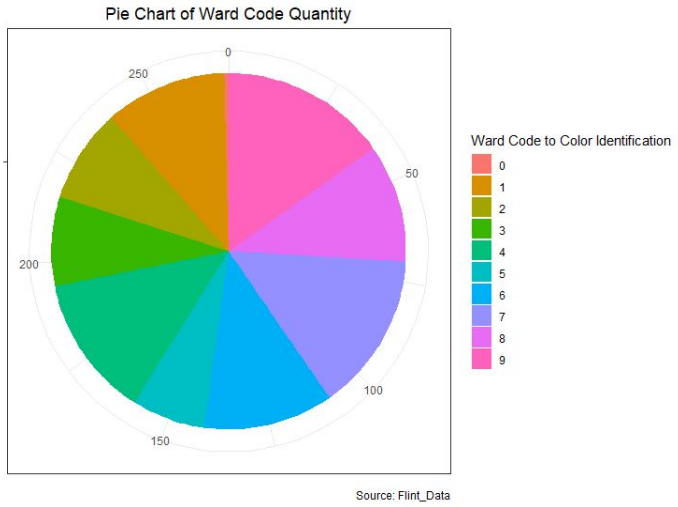
There did appear to be a negative relationship between time and lead content that can be accurately modeled through the exponential decay function. However, when applied to the situation the aggregate slope only has a value of -0.01 and the relationship can only account for 12.8% of variability. This means that in actuality the amount of lead is barely decreasing as time passed increases (essentially having no effect when considering the size and scope of the problem) and the content will still have extremely negative health implications.

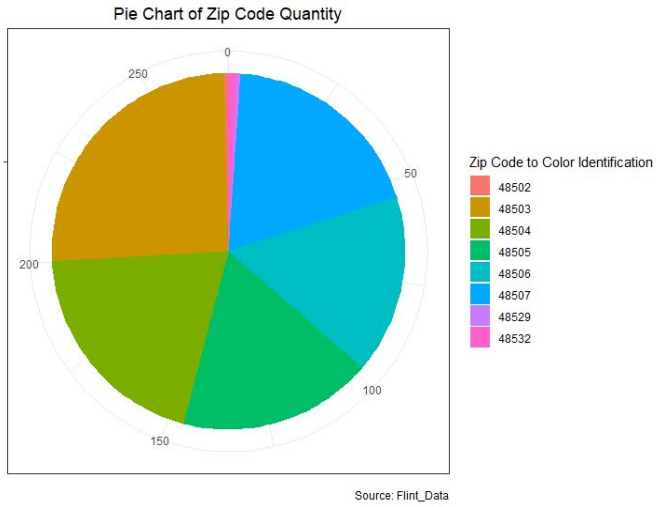
Based on policy standards, Flint should be doing more in regards to their water systems since more than 10% of water samples exceeded 15 ppm. The water is too corrosive according to the EPA and more needs to be done to help Flint fix the problem.

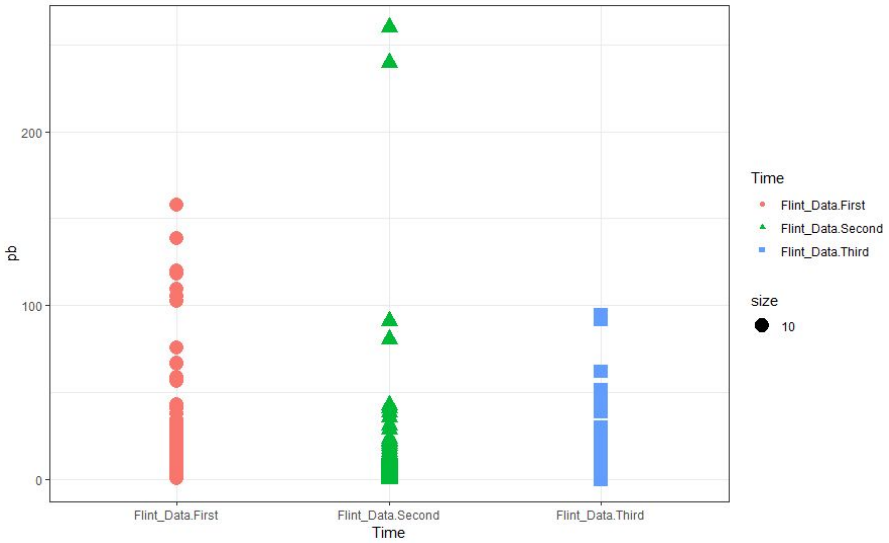
Below we have chosen to also include code for individual regressions for each zip code. Because we determined earlier that the zip code had little effect on lead level we decided not to evaluate these but to include them if the reader wishes to see how the result of the zip code mirror the results of the aggregate model. These results and models would be useful for a future study regarding the impact of the Flint Water Crisis on individual areas of Flint.

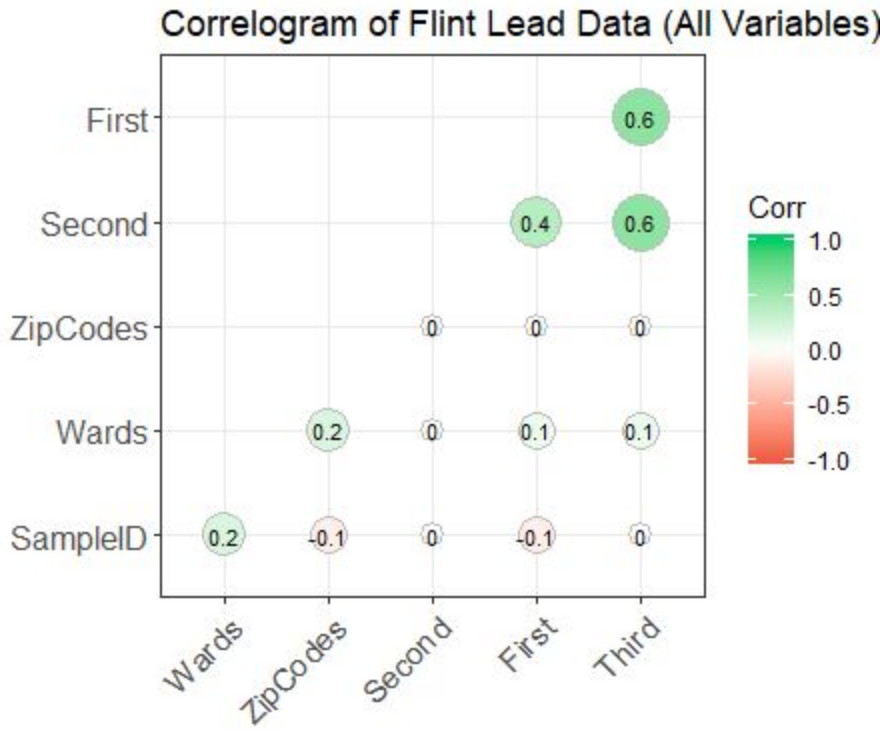
In conclusion, the exponential decay function accurately models our data and provides useful visuals for analysis of the crisis. However, in the context of policy the small values for slope and R-Squared lead to little change in lead levels as time increases. This means we proved the local government was providing false information to the citizens of Flint. To conclude our research we expanded on this and showed that by EPA standards Flint should be taking immediate action to resolve this crisis. We urge policy makers to take this analysis into account and increase research and action regarding the Flint Water Crisis.

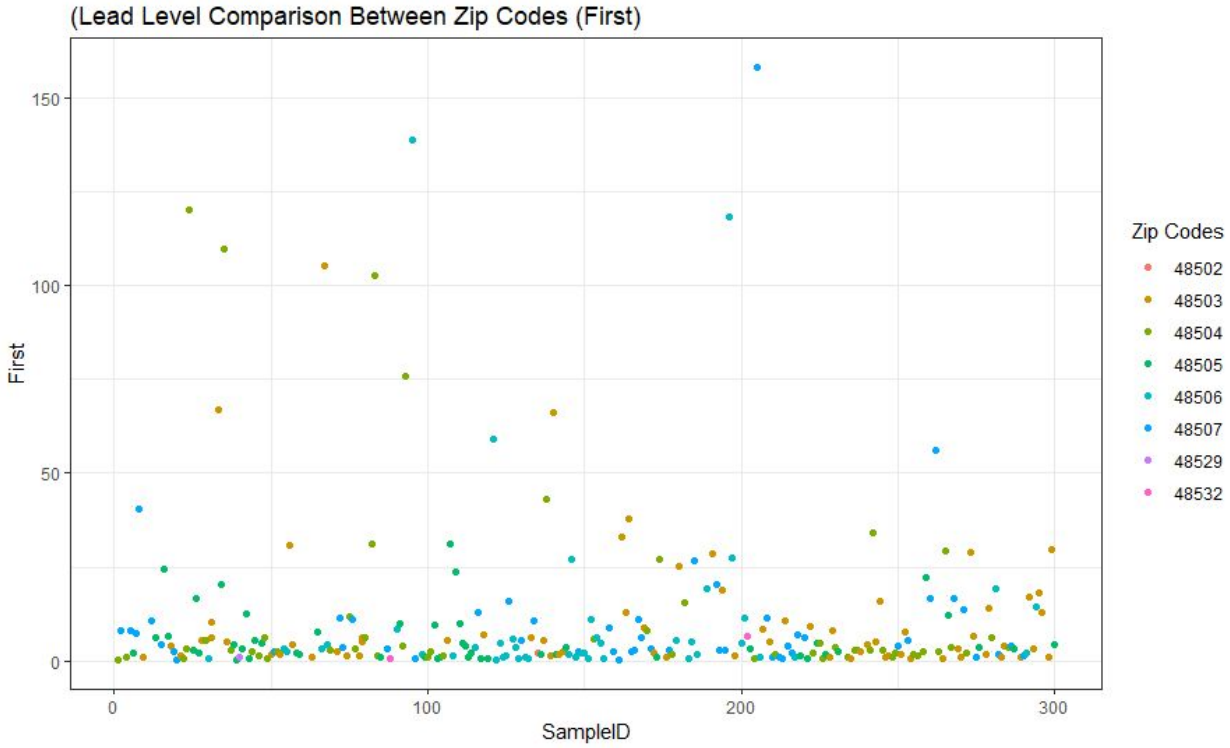
Appendix

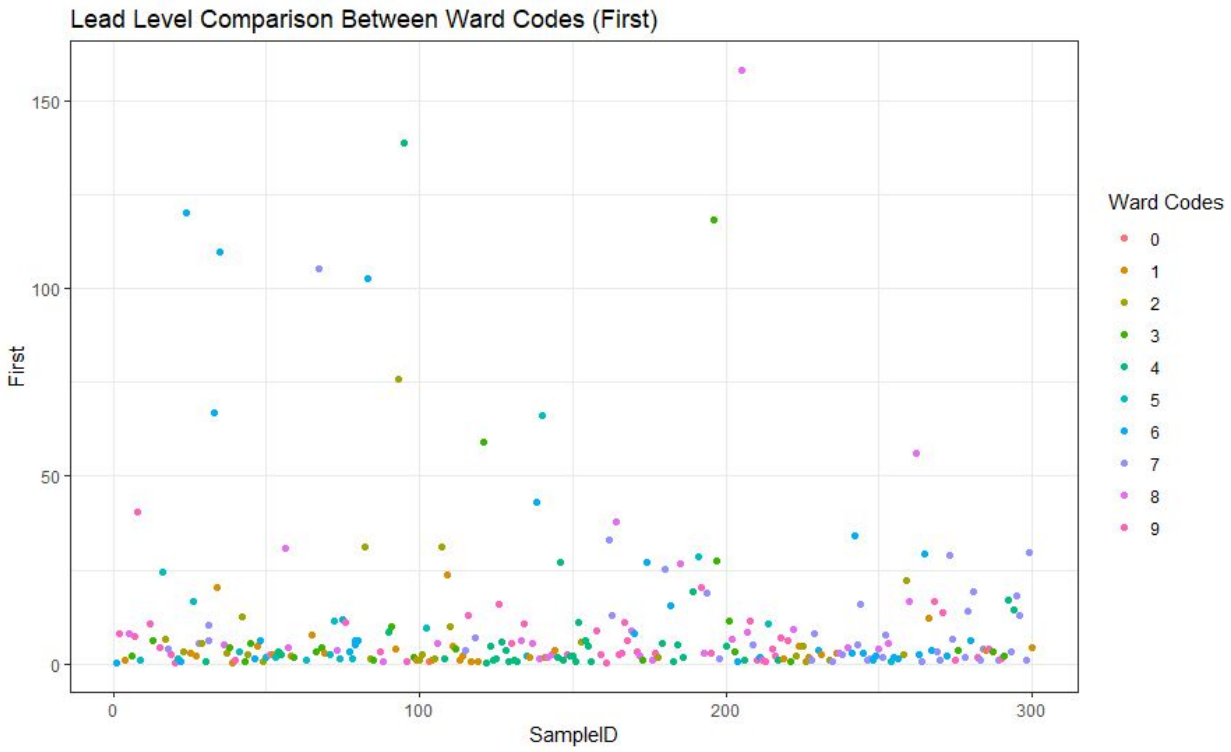

Pie Chart of Ward Code Quantity
0
50
100
150
200
250
Ward Code to Color Identification
0
1
2
3
4
5
6
7
8
9
Source: Flint_Data

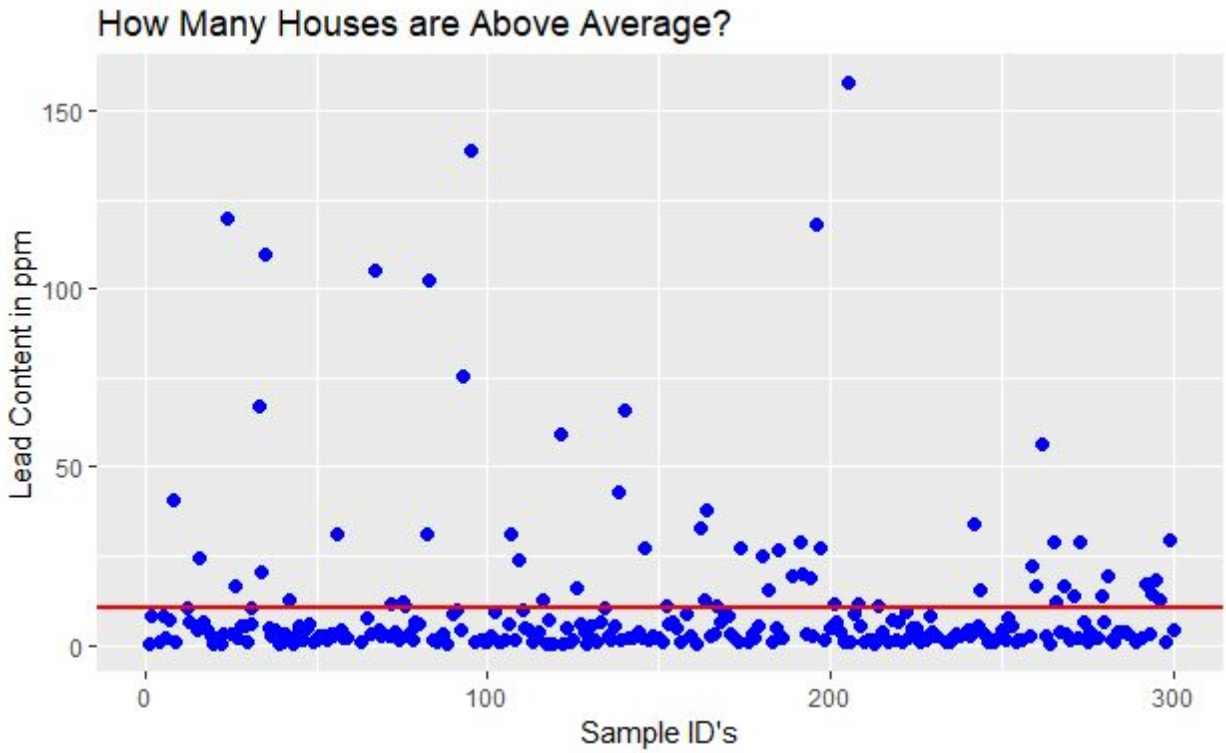

Pie Chart of Zip Code Quantity
0
50
100
150
200
250
Zip Code to Color Identification
48502
48503
48504
48505
48506
48507
48529
48532
Source: Flint_Data

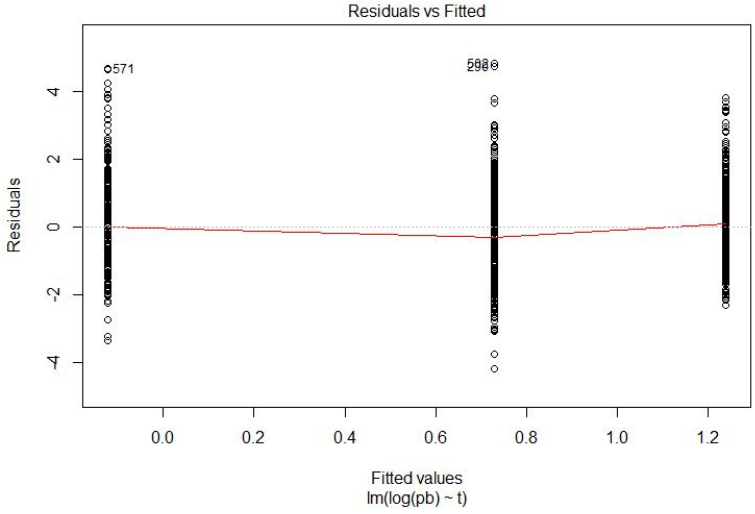

pb
0
100
200
Flint_Data.First
Flint_Data.Second
Flint_Data.Third
Time
Time
Flint_Data.First
Flint_Data.Second
Flint_Data.Third
size
10

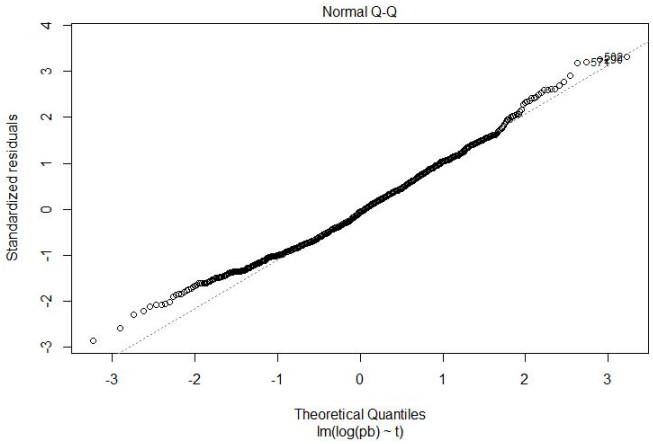

Correlogram of Flint Lead Data (All Variables)
First
Second
ZipCodes
Wards
SampleID
0.6
0.4
0.6
0
0
0
0.2
0
0.1
0.1
0.2
-0.1
0
-0.1
0
Wards
ZipCodes
Second
First
Third
Corr
1.0
0.5
0.0
-0.5
-1.0

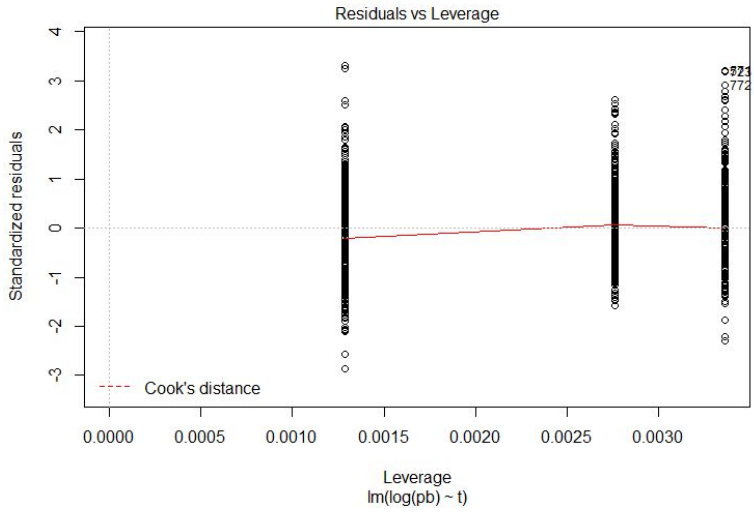

(Lead Level Comparison Between Zip Codes (First)
First
150
100
50
0
0
100
200
300
SampleID
Zip Codes
48502
48503
48504
48505
48506
48507
48529
48532


Lead Level Comparison Between Ward Codes (First)
First
SampleID
Ward Codes
0
1
2
3
4
5
6
7
8
9
150
100
50
0
0
100
200
300


How Many Houses are Above Average?
Lead Content in ppm
Sample ID's
150
100
50
0
0
100
200
300


Residuals vs Fitted
Residuals
Fitted values
lm(log(pb) ~ t)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(pb) ~ t)


Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(log(pb) ~ t)

Bibliography

Clark, Anna. "'Nothing to Worry about. The Water Is Fine': How Flint Poisoned Its People." *The Guardian*, Guardian News and Media, 3 July 2018, www.theguardian.com/news/2018/jul/03/nothing-to-worry-about-the-water-is-fine-how-flint-michigan-poisoned-its-people.

Langkjaer-Bain, Robert. "The Murky Tale of Flint's Deceptive Water Data." Significance, vol. 14, no. 2, 2017, pp. 16–21., doi:10.1111/j.1740-9713.2017.01016.x.

"National Primary Drinking Water Regulations." *EPA*, Environmental Protection Agency, 22 Mar. 2018, www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations#seven.