Flint Water Data Analysis Using Linear Mixed Models

STT [redacted] Project

[redacted]

Introduction

On April 25, 2014 the City of Flint switched its water supply from Detroit's system to the Flint river as a cost-saving measure for the financially challenged city. This act led to the city's water lead contamination due to the high corrosiveness for lead of the Flint river. What is now known as the Flint water crisis reportedly caused, among many other issues, developmental problems in a group of Flint children. On October 16, 2015 the City of Flint recognised the crisis and reconnected back to Detroit's system.

This project aims to analyse the residential water testing samples collected between September 3, 2015 and March 15, 2016 and a sentinel sampling data set collected between February 15, 2016 and February 26, 2016. The sentinel samples correspond to the households chosen by the Environmental Protecting Agency (EPA) and the Michigan Department of Environmental Quality (MDEQ). In particular, the sites were selected as those with higher lead concentration in water due to their lead service lines or documented high lead levels in blood of the kids living in those homes.

First, based on the residential water testing samples, we examine the effect of reconnecting to the Detroit's system on the lead concentration in the water and try to asses the effectiveness of this measure. Second, based on the same data set as above, the difference in the lead contamination among several ZIP code areas in Flint is considered. Finally, we compare the conclusions drawn from the sentinel data set with the residential water testing results in order to determine the necessity of this secondary sampling method for monitoring of the lead contamination in Flint.

Results

The residential water testing samples revealed a significant decrease in the lead concentration after switching back to Detroit's treated water supply. Moreover, we have found evidence for difference in the lead contamination of water among Flint's ZIP code areas 48503, 48504, 48505, 48506, and 48507 with the highest lead concentration for the ZIP code 48503 and the lowest for 48507. Our analyses also showed that the conclusions drawn from the sentinel data set depart from the results based on the residential water samples. Therefore, both sampling methods should be considered for monitoring of the lead contamination in the City of Flint.

Methods

Model Building

For the sake of building a model for the residential water testing samples, the data has been first cleaned from observations with missing values which were mostly due to omitted ZIP code information. Then, an uneven sampling frequency of observations among different ZIP code areas was considered. Therefore, we focus in the subsequent analyses only on 48503 - 48507 ZIP code areas with relatively even sampling frequencies. Finally, we removed the extremes cases for the lead concentration in the water.

In order to answer the questions proposed in the introduction, we consider the cube root of *water lead concentration* in parts per billion (ppb) as our response variable as to compensate for the skewness of its distribution. For the modelling purpose, the *ZIP code area* and an indicator variable for the *measurements collected after October 16, 2015* were determined to be fixed effects in our model; they are the main factors under questioning. However, the lead concentration in water was collected repeatedly for many of the households. This introduces dependency into the data set, and therefore, a random effect for *repeated measurements* needs to be placed into the model. It is also likely that households in geographical vicinity have similar service lines and plumbing conditions, we expressed this fact by adding a random effect for *street on which were the houses build.*

Overall, a **linear mixed model** was fitted to the residential sample data using unrestricted ML method, where normality assumption of the error terms and random effects is reasonable because of the central limit theorem. The unrestricted ML method was chosen to ensure meaningful inference.

So as to asses the equivalence of both sampling procedures, the sentinel data set was preprocessed in alignment with the household results. Moreover, 9 sentinel sites (out of 552) with repeated measurements were removed to simplify the interpretation of the analyses. Linear mixed models with fixed effect of *ZIP code area* and random effect of *site's street address* were fitted for the sentinel data and the household measurements collected between February 15, 2016 and February 26, 2016. Again, the unrestricted ML method was chosen to ensure meaningful inference.

Final Models and Statistical Analysis

The final linear mixed model for the residential water testing samples is of the following form:

$$(Y_{ijk})^{\frac{1}{3}} = \mu + \beta^T X_{ijk} + \gamma_i + u_j + \epsilon_{ijk}, \tag{1}$$

where Y_{ijk} is the lead concentration in the water of the k^{th} measurement of the i^{th} household. $X_{ijk} = (X_{ijk}^{(1)}, X_{ijk}^{(2)}, X_{ijk}^{(3)}, X_{ijk}^{(4)}, X_{ijk}^{(5)})^T$ are the covariates for the fixed effects, $X_{ijk}^{(1)} = 1$ if the k^{th} measurement of the i^{th} household was taken after October 16, 2015, and

$$X_{ijk}^{(2)} = \begin{cases} 1 & \text{if the } i^{th} \text{ household belongs to 48504 ZIP code} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ijk}^{(3)} = \begin{cases} 1 & \text{if the } i^{th} \text{ household belongs to 48505 ZIP code} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ijk}^{(4)} = \begin{cases} 1 & \text{if the } i^{th} \text{ household belongs to 48506 ZIP code} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ijk}^{(5)} = \begin{cases} 1 & \text{if the } i^{th} \text{ household belongs to 48507 ZIP code} \\ 0 & \text{otherwise.} \end{cases}$$

γ_i are IID random intercept for the i^{th} household, γ_i is assumed to follow $N(0, \sigma_\gamma^2)$. u_j are IID random intercept for the j^{th} street, u_j is assumed to follow $N(0, \sigma_u^2)$, and ϵ_{ijk} are IID random error with $\epsilon_{ijk} \sim N(0, \sigma^2)$. Moreover, both random effects and the random error are independent of each other.

The unrestricted ML method yields the following estimates of the parameters and their standard errors:

Parameter	Estimate	Standard Error
μ	1.347	0.064
β_1	-0.411	0.061
β_2	-0.091	0.039
β_3	-0.130	0.039
β_4	-0.281	0.042
β_5	-0.135	0.041
σ_γ^2	0.314	
σ_u^2	0.102	
σ^2	0.573	

Table 1: Parameter estimates and their standard errors

First, the effect of reconnecting to the Detroit's system on the lead concentration in the water was examined by testing the null hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. We performed the LRT test for two nested models that differ only in their fixed effects. The resulting P-value of the test statistics is < 0.001 and the approximate 95% confidence interval for β_1 is $(-0.531, -0.292)$. Since the P-values generated by the LRT for the fixed effects tend to be too small, and therefore overestimate the importance of some effects, a parametric bootstrap procedure was also performed resulting in P-value = 0. Both of the procedures gives strong evidence that **the lead concentration in the water in the City of Flint was significantly reduced after October 16, 2015** and shows the effectiveness of reconnecting back to Detroit's water supply. Nevertheless, the cube root transformation of the response variable prevent us from meaningful interpretation of the size of this effect.

Second question under consideration is the difference in the lead contamination among several ZIP code areas in Flint. A similar procedure as above was adopted to test the null hypothesis $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ against H_1 : At least one β_i $,i = 2, 3, 4, 5$ is different from 0. The resulting P-value of the LRT test statistics for two nested models is < 0.001 and the parametric bootstrap P-value = 0. It shows a significant effect of ZIP code area on the lead concentration in the water.

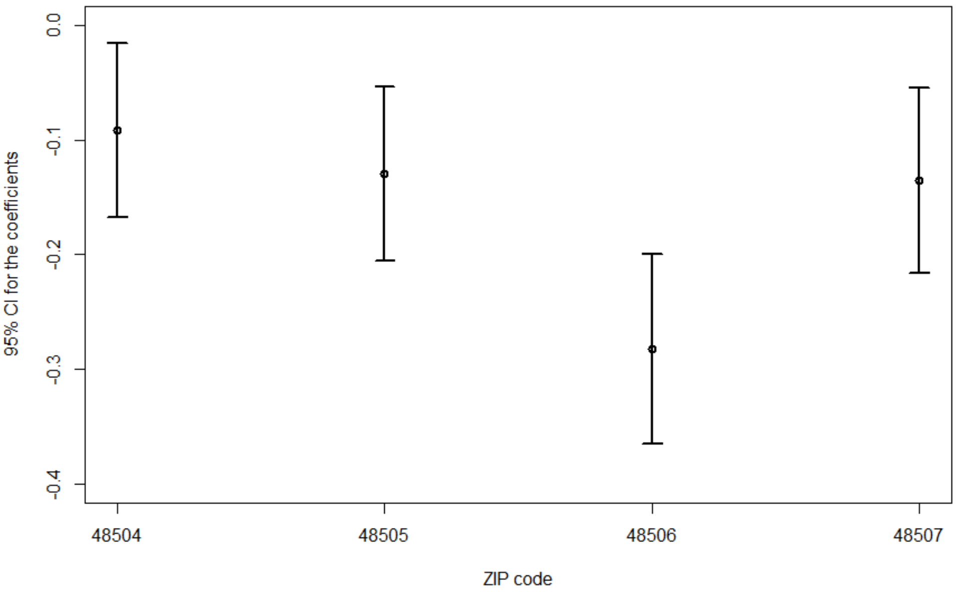
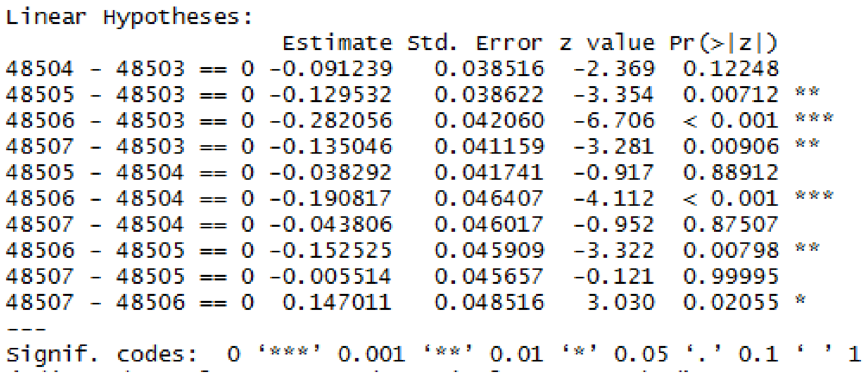


Figure 1: Approximate 95% confidence intervals for the parameters $\beta_2, \beta_3, \beta_4$ and β_5

The next step si to examine the difference in the lead contamination among the ZIP code areas. Figure 1 displays the approximate 95% confidence intervals for the area coefficients and suggest that ZIP codes 48504 and 48506 significantly differ in the lead concentration in the water. However, to prevent the inflation of experiment-wise Type I error, Tukey's multiple comparison method was performed with the following results:



```
Linear Hypotheses:
                      Estimate Std. Error z value Pr(>|z|)    
48504 - 48503 == 0 -0.091239   0.038516  -2.369  0.12248    
48505 - 48503 == 0 -0.129532   0.038622  -3.354  0.00712 ** 
48506 - 48503 == 0 -0.282056   0.042060  -6.706  < 0.001 ***
48507 - 48503 == 0 -0.135046   0.041159  -3.281  0.00906 ** 
48505 - 48504 == 0 -0.038292   0.041741  -0.917  0.88912    
48506 - 48504 == 0 -0.190817   0.046407  -4.112  < 0.001 ***
48507 - 48504 == 0 -0.043806   0.046017  -0.952  0.87507    
48506 - 48505 == 0 -0.152525   0.045909  -3.322  0.00798 ** 
48507 - 48505 == 0 -0.005514   0.045657  -0.121  0.99995    
48507 - 48506 == 0  0.147011   0.048516   3.030  0.02055 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2: The results of Tukey's multiple comparison procedure

We can see that the **ZIP code areas 48505, 48506, and 48507 have significantly lower lead concentration in water than 48503**. In addition, the residential data gives strong evidence that the **lead contamination in water for the ZIP code area 48506 is the smallest and the highest for 48503 among all the other ZIP code areas** considered for the City of Flint.

Finally, we compare the sentinel data set with the residential water testing results using the linear mixed model of the form:

$$(Y_{ij})^{\frac{1}{3}} = \mu + \beta^T X_{ij} + u_j + \epsilon_{ij}, \tag{2}$$

where Y_{ij} is the lead concentration in the water of t the i^{th} household. $X_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)}, X_{ij}^{(3)}, X_{ij}^{(4)})^T$ are the covariates for the fixed effects

$$X_{ij}^{(1)} = \begin{cases} 1 & \text{if the } i^{th} \text{ household belongs to 48504 ZIP code} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ij}^{(2)} = \begin{cases} 1 & \text{if the } i^{th} \text{ household belongs to 48505 ZIP code} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ij}^{(3)} = \begin{cases} 1 & \text{if the } i^{th} \text{ household belongs to 48506 ZIP code} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ij}^{(4)} = \begin{cases} 1 & \text{if the } i^{th} \text{ household belongs to 48507 ZIP code} \\ 0 & \text{otherwise.} \end{cases}$$

u_j are IID random intercept for the j^{th} street, u_j is assumed to follow $N(0, \sigma_u^2)$, and ϵ_{ij} are IID random error with $\epsilon_{ij} \sim N(0, \sigma^2)$. Moreover, both random effect and the random error are independent of each other.

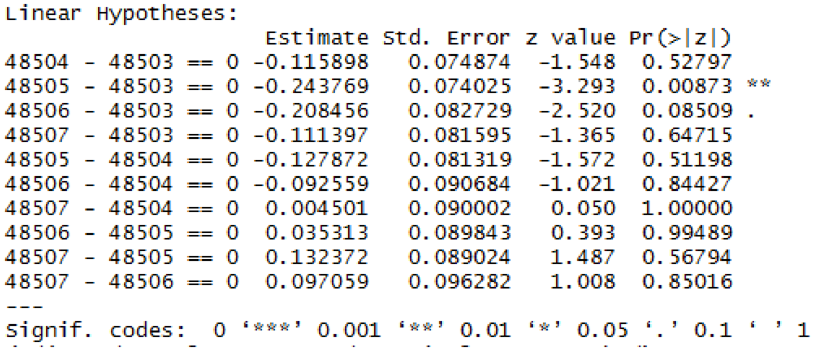
The unrestricted ML method yields the following estimates of the parameters and their standard errors:

		Residential Data		Sentinel Data
Parameter	**Estimate**	**Standard Error**	**Estimate**	**Standard Error**
μ	1.021	0.047	1.043	0.075
β_1	-0.116	0.075	0.128	0.127
β_2	-0.244	0.074	-0.140	0.127
β_3	-0.208	0.083	0.015	0.171
β_4	-0.111	0.082	-0.395	0.181
σ_u^2	0.147		0.100	
σ^2	0.863		0.854	

Table 2: Parameter estimates and their standard errors

We start by considering the significance of the fixed effect. Both LRT and parametric bootstrap procedures were used to test the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ against H_1 : At least one β_i ,$i = 1, 2, 3, 4$ is different from 0. For both data sets, the LRT statistics P-value was < 0.001 and the parametric bootstrap P-value $= 0$. Therefore, there is a significant effect of ZIP code area on the lead concentration in the water for the sentinel measurements and the residential data as well.

As in the previous case, Tukey's multiple comparison method was performed in order to examine the difference in the lead contamination among the ZIP code areas. Here, the Tukey's method didn't find any significant difference between ZIP code areas for the sentinel data set, whereas the residential water testing samples yield the following result:



```
Linear Hypotheses:
                     Estimate Std. Error z value Pr(>|z|)
48504 - 48503 == 0 -0.115898   0.074874  -1.548  0.52797
48505 - 48503 == 0 -0.243769   0.074025  -3.293  0.00873 **
48506 - 48503 == 0 -0.208456   0.082729  -2.520  0.08509 .
48507 - 48503 == 0 -0.111397   0.081595  -1.365  0.64715
48505 - 48504 == 0 -0.127872   0.081319  -1.572  0.51198
48506 - 48504 == 0 -0.092559   0.090684  -1.021  0.84427
48507 - 48504 == 0  0.004501   0.090002   0.050  1.00000
48506 - 48505 == 0  0.035313   0.089843   0.393  0.99489
48507 - 48505 == 0  0.132372   0.089024   1.487  0.56794
48507 - 48506 == 0  0.097059   0.096282   1.008  0.85016
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3: The results of Tukey's multiple comparison procedure for the residential water testing samples

We can see that there is evidence for different lead concentration in water between at least two ZIP code areas (48505 and 48503). Overall, both models show significant effect of ZIP code area on water lead concentration in the City of Flint. Nevertheless, the conclusions drawn from the model for residential water samples vary from those based on the sentinel data, and therefore, **we cannot state that both sampling methods are equivalent for monitoring of the lead contamination in Flint.**

Model Diagnostics

The appropriateness of linear mixed models above relies on two main assumptions. They are the form of the fitted function and the distribution of random effects and residuals. Figure 4 shows residual plot and normal Q-Q plot for the residential data model (1).

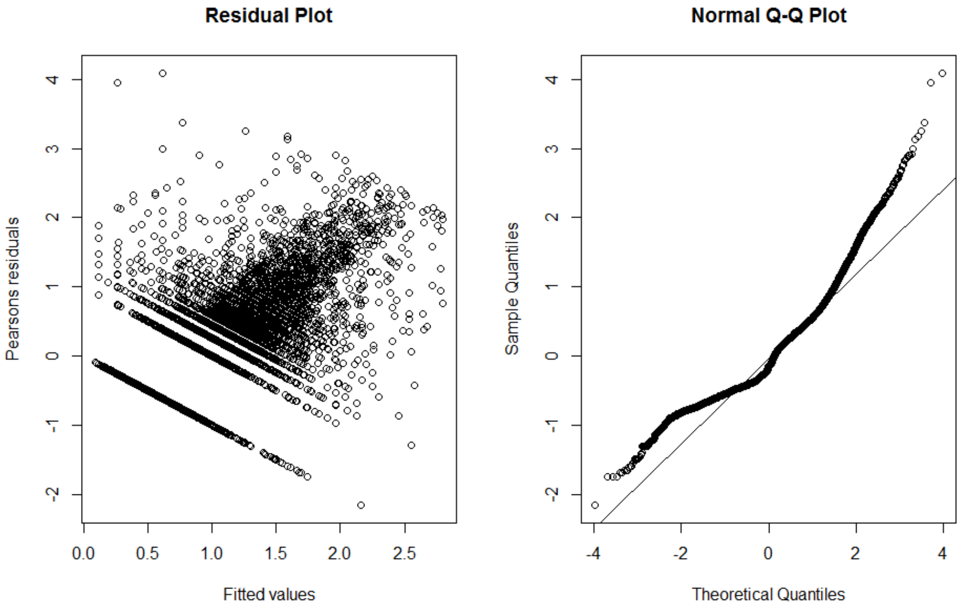


Figure 4: Residual plot and normal Q-Q plot for model (1)

The normal Q-Q plot shows slight right-skewness of the residuals, however, linear models are quite robust when it comes to violating the normality assumption, and therefore we consider this small violation acceptable. The appropriateness of the linearity of the fitted function can be evaluated based on the residual plot (Pearson's residuals). We don't see any systematic tendency in the plot, which means that the linear form of the fitted function is reasonable and the homoscedasticity of the error terms appropriate.

Similarly, Figure 5 shows residual plots and normal Q-Q plots for the model (2).

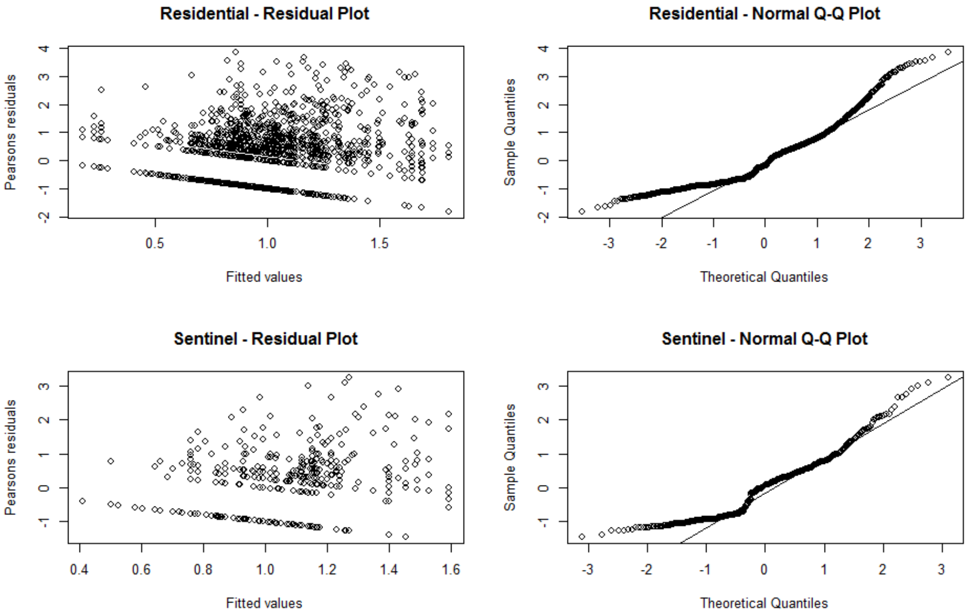


Figure 5: Residual plots and normal Q-Q plots for model (2)

Using the same criteria as above, the normality of the error terms is again slightly violated for both data sets, however, the robustness of linear mixed models allows some departure from normality. Equivalently, we don't see any systematic tendency in the residual plots, which means that the linear form of the fitted functions is reasonable and the homoscedasticity of the error terms again appropriate.

Discussion

Even though the analyses above reveals some aspects about both the residential water testing and the sentinel data set, we have mainly focused on the former. Sentinel data, however, contain information about the type of service line in each of the households. An analysis of significance of the service line effect on the lead concentration would be useful to determine the scale on which it will be most effective to contain the Flint water crisis. Moreover, only the compound symmetric covariance structure was used for the linear mixed models. Better fit of the models might be achieved, if we considered other covariance structures as well.