GETTING THE LEAD OUT: DATA SCIENCE AND WATER SERVICE LINES IN FLINT

Jared Webb
University of Michigan
Ann Arbor, MI
jaredaw@umich.edu

Jacob Abernethy
Geogia Institute of Technology
Atlanta, GA
prof@gatech.edu

Eric Schwartz
University of Michigan
Ann Arbor, MI
ericmsch@umich.edu

February 12, 2020

ABSTRACT

We give a brief outline of the Flint Water Crisis and our previous work making lead service line discovery more efficient. We also give a case study, comparing the efficacy of our data driven models to a non-data driven approach run by an independent engineering firm in 2018. The data show that using a machine learning approach significantly improves the efficiency of the lead service line replacement program in Flint.

***K*eywords** Water Infrastructure · Machine Learning · Flint Water Crisis

1 Introduction

The story of the Flint Water Crisis is related to the general economic decline in Rust Belt cities across the American Midwest. Flint, once one of the most prosperous cities in the United States, has seen its population cut in half since its peak in the 1960s. In the same time period, the number of automotive jobs has decreased from more than 80,000 to less than 10,000. As the city's tax base and economy shrunk, so did tax revenues. After the 2008 financial crisis, the city was forced to declare bankruptcy.

In the face of bankruptcy, a state-appointed emergency manager took control of city government in 2011 and began to address the city's "structural debt" [1] by cutting services and implementing cost-saving measures. In particular, the city was forced to switch its drinking water from the Detroit Municipal System to its back-up source, the Flint River. After the switch, the specific chemical properties of the new water were overlooked or improperly treated by water officials. Most importantly, the new water was more acidic than water from the previous source and stripped away protective mineral buildup on pipes, eventually exposing raw metal. When the acidic water flowed through lead pipes, the metal leached into the drinking water supply. Though there were numerous complaints by the citizenry about off-color and foul-smelling water, these were dismissed as anecdotal. The water remained incorrectly treated for more than 2 years, until a doctor working in Flint noticed elevated levels of lead in children's blood [2] [3].

News of the crisis became an international media story, and during this period of intense scrutiny, both State and Federal government allocated funds for recovery. Attention soon focused on aging lead service lines (see Figure 2) that connect water mains to residences, which have been identified in the scientific literature as a source of lead contamination in drinking water [4].

The city and state jointly appointed a team to coordinate using allocated funds to replace the lead service lines. This team, called the Flint Fast Action and Sustainability Program (or FAST Start), had a primary objective of removing as much hazardous infrastructure as possible, up to funding levels. This would be a relatively straightforward task except for one thing: no one knew where the dangerous pipes were.

In the 1980's, the federal government instituted a regulation governing drinking water in the United States called the Lead and Copper Rule. The Environmental Protection Agency was tasked with enforcing the regulation. Among other things, it mandated a ceiling on lead levels in drinking water and that cities maintain an inventory of their lead

Figure 1: Schematic describing a residential service line. The private portion of the service line from the house to the curb box is owned by the resident. The public portion from the curb box to the water main is owned by the city (Courtesy Michigan Department of Environmental Quality). The image on the right shows a service line replacement in Flint.

infrastructure. Flint failed to maintain such records [5], and so the data indicating the location of dangerous pipes lay buried underground.

This made it costly to discover the material of even a single pipe, and this information bottleneck stood to use up valuable resources in exploration that could be used replacing dangerous pipes. Our team began working with FAST Start in 2016 to provide technical, statistical, and algorithmic support in order to improve information availability and help guide decision making. We built machine learning models to predict which neighborhoods were most likely to have lead pipes and recorded consistently high rates of correct predictions during pipe excavations through 2017.

In 2018, the city replaced the FAST Start team with an international engineering firm. The new firm disregarded previous model-based recommendations in favor of their own prioritization program. The number of excavated lead service lines dropped precipitously, sinking below 20 percent, resulting in projections by the city that more than 20 million dollars would be spent to dig up and bury lines that did not require replacement [6] [7].

Following a lawsuit by the Natural Resources Defense Council (NRDC) [8], the city was obligated to return to the data-driven modelling approach to prioritize service line replacement. Once this was re-implemented, correct identification rates returned to their previous levels, greatly improving the per-replacement costs for the city.

In this paper, we briefly outline our previous published work on lead service line prediction and then compare outcomes between the model and non-model approaches. The measurable improvements in dangerous service line discovery gives strong evidence that data-driven approaches provide significant return on investment.

2 Previous Work

We have written extensively about our previous work applying machine learning to the Flint recovery effort (See [9] and [10]). We provide a brief outline here for context.

2.1 Available Data

Our models are built primarily on parcel information provided by the city. These records include tax information, such as the state estimated value of the property and its buildings. They also include physical attributes of the properties, such as the acreage of the plots and the style of residential buildings.

While initially no historical service line records were known, the city eventually discovered water-logged, hand-drawn maps that included this data. These were digitized by Dr. Martin Kauffman and his students at the GIS center of the University of Michigan - Flint [11]. These records were initially viewed as a source of ground truth for existing service lines, even though information pertaining to a large number of homes was missing. However, as excavations began, it became apparent that many of the records were wrong almost as often as they were right. Instead of discarding this data, we incorporated it into our models.

Another important data source was residential water testing data. When the scale of the crisis became apparent, the Michigan Department of Environmental Quality made home testing kits available to city residents. The results of each test were published at michigan.gov/flintwater.

We also incorporated other data from various sources, such as the American Community Survey and the Michigan DEQ. A full descriptions of all of our data can be found [9].

2.2 Model Development

We developed a machine learning model called ActiveRemediation to predict the locations of dangerous service lines and prioritize regions for exploration to reduce uncertainty. ActiveRemediation consists of an exploration step that uses active learning to prioritize locations for label discovery in order to reduce uncertainty and a training step that uses available data to train an XGBoost [12] model to predict remaining unknown labels with all available data.

Since contractors were assigned to replace groups of homes, rather than single plots, care had to be taken to choose a cross-validation strategy to prevent over-fitting our models. If the training and test sets contain next homes within the same neighborhoods, the model could easily make accurate predictions. To prevent this, we made sure that neighborhoods in the training sets had no homes in the test set and vice versa.

Using our techniques in simulation, we conducted experiments that show that several million dollars could be saved by the city avoiding unnecessary excavations (i.e., digging up a copper pipe that doesn't need to be replaced). This research generalizes well beyond Flint, as many municipalities grapple with the presence of antiquated and potentially dangerous water infrastructure.

3 A Comparison of Two Approaches

In this section we will compare the hit rates and efficiency costs our model (2016-2017, 2019) and the approach of the engineering firm the city hired to run excavations in 2018. We define the hit rate to be the number of service lines that contained dangerous material divided by the total number of excavations. For example, if 100 excavations take place and 75 pipes contain lead, the hit rate is 0.75. If the hit rate for a given approach is r, then the total number of excavations to remove N lead service lines is $\frac{N}{r}$.

In Flint, contractors are paid \$5,000 to replace a service line, and \$3,000 for an excavation that does not yield a replacement (for example, if they dig and find a copper pipe). Using these numbers, we can calculate the cost C of replacing N lead service lines as a function of our hit rate:

$$C(r|N) = N \cdot 5000 + N(\frac{1}{r} - 1) \cdot 3000$$

where r is the hit rate and $N(\frac{1}{r} - 1)$ gives the number of excavations that does not yield replacement. This function is an inverse function in r, and so costs blow up as the hit rate approaches 0 and decrease only marginally as the hit rate approaches 1. See Figure 4.

3.1 Results in 2018

Excavation information in Flint is publicly available as part of compliance with the NRDC lawsuit [8]. In response to an NRDC request, in late 2018 we gathered the public data on all excavations that took place that year to analyze the hired firm's effectiveness at replacing dangerous lead service lines. The hit rate for each month in 2018 was consistently below 0.2, and 0.15 for the whole year.

To highlight the disparity between the predictions available at the beginning of 2018 and the action taken by the firm, in Table 1 we examine voting precincts in Flint that saw the most service line replacement attempts in 2018. Each of these precincts were among the least likely to have lead service lines according to our models, yet saw hundreds of excavations with very little payoff. Compare this to Table 2, which gives the same breakdown for those precincts predicted to have the highest number of lead service lines.

3.2 Results in 2019

After the city returned to our data-driven approach, we have been tracking hit rates over time. Our results show that using the model has significantly improved dangerous service line discovery, nearly returning to pre-2018 levels. See Figure 2 and Figure 3.

Voting Precinct	Number of Excavations 2018	Expected Hit Rate	Actual Hit Rate
8	417	0.14	0.00
10	346	0.20	0.03
48	318	0.35	0.35
22	267	0.05	0.00
9	217	0.23	0.05

Table 1: Number of excavations and hit rates in the most excavated voting precincts of 2018. The expected hit rate is the result given by our models trained on data obtained before 2018. These precincts were among the least likely to contain lead service lines according to our models, and yet they were the most excavated.

Voting Precinct	Number of Excavations 2018	Expected Hit Rate	Actual Hit Rate
31	0	0.93	N/A
27	6	0.91	1.0
40	0	0.89	N/A
39	10	0.85	0.87
28	58	0.84	1.0

Table 2: Number of excavations and hit rates in the voting precincts with highest predicted concentrations of lead service lines. The expected hit rate is the result given by our models trained on data obtained before 2018. Even though the engineering firm was given these numbers, these precincts were largely ignored during the 2018 excavation program.

Recall that the cost to replace N dangerous service lines increases dramatically as the hit rate decreases. In Figure 4, we plot the cost function for 100 service line replacements, as well as corresponding costs of the 2018 and 2019 replacement programs. We also include the estimated costs of random guessing. The hit rate in 2018 was low enough that marginal improvements would have resulted in significant improvements in the per replacement costs.

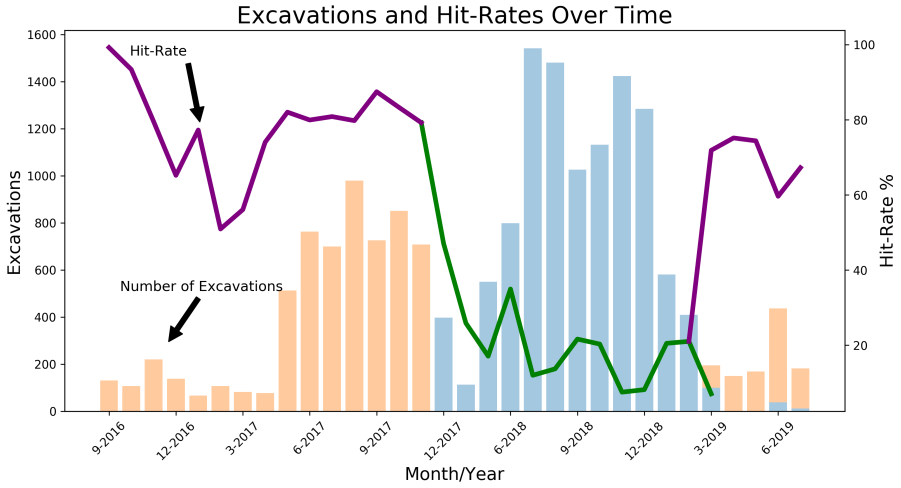

Figure 2: Hit rate for found lead service lines over time. The valley (in green/blue) in 2018 is when the city-hired engineering firm was running the replacement program. Once the city returned to the model based approach, hit rates increased dramatically.

4 Conclusions

This case study demonstrates the serious financial risk that governments or other stakeholders take when data are ignored. Conversely, policy makers stand to dramatically increase efficiency and reduce uncertainty in a crisis if they can formulate it as a data science problem. Our work has been able to dramatically reduce the cost per replacement in Flint, and we hope that other cities facing similar problems can find ways to integrate similar approaches.

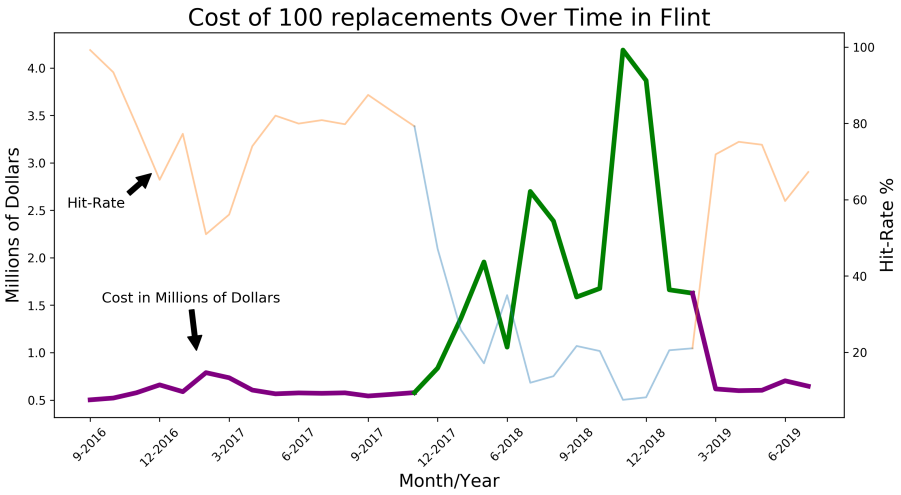

Figure 3: The cost to replace 100 service lines in Flint since 2016. As the hit rate decreased, the cost dramatically increases (See Figure 4). When the city returned to using our model in 2019, average costs quickly decreased to their previous levels.

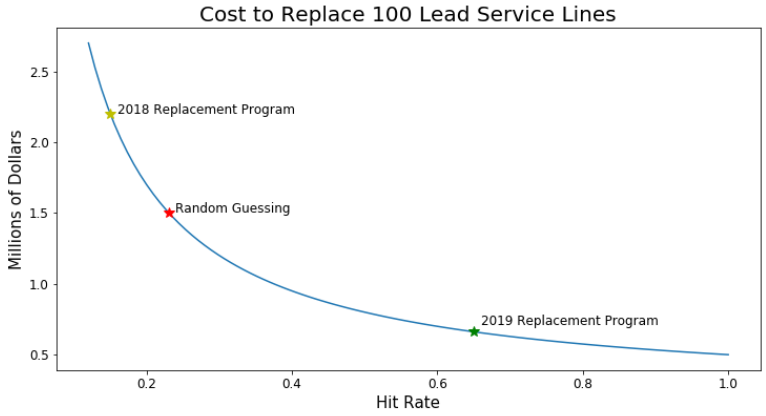

Figure 4: Cost function for replacing 100 lead service lines in Flint. Note that as the hit rate approaches 0 the function blows up. Stars indicate the historical costs in 2018 and 2019, as well as the projected cost from random guessing. Low hit rates like those in 2018 yield very expensive per replacement costs for the city. On the other hand, as the hit rate approaches 1 the cost function yields rapidly diminishing returns.

References

[1] Kristin Longley. Emergency manager michael brown appointed to lead flint through second state takeover. https://www.mlive.com/news/flint/index.ssf/2011/11/emergency_manager_michael_brow.html, November 2011. (Accessed July 8, 2018.

[2] Mona Hanna-Attisha, Jenny LaChance, Richard Casey Sadler, and Allison Champney Schnepp. Elevated blood lead levels in children associated with the flint drinking water crisis: a spatial analysis of risk and public health response. *American journal of public health*, 106(2):283–290, 2016.

[3] Michael Torrice. How lead ended up in flint's tap water. *Chem. Eng. News*, 94(7):26–29, 2016.

[4] Anne Sandvig, P Kwan, G Kirmeyer, B Maynard, D Mast, R Rhodes Trussell, S Trussell, A Cantor, and A Prescott. *Contribution of service line and plumbing fixtures to lead and copper rule compliance issues*. Environmental Protection Agency, Water Environment Research Foundation, 2008.

[5] Ron Fonger. Documents show flint filed false reports about testing for lead in water. https://www.mlive.com/news/flint/index.ssf/2015/11/documents_show_city_filed_fals.html, November 2015. (Accessed July 8, 2018.

[6] Ron Fonger. Flint agrees to return to data-driven approach to find remaining lead service lines. https://www.mlive.com/news/flint/2019/02/flint-agrees-to-return-to-data-driven-approach-to-find-remaining-lead-service-lines.html, 2019. (Accessed May 2019).

[7] Alexis Madrigal. https://www.theatlantic.com/technology/archive/2019/01/how-machine-learning-found-flints-lead-pipes/578692/, 2019. (Accessed May 2019).

[8] Flint case documents. https://www.nrdc.org/resources/flint-case-documents, 2019. (Accessed May 2019).

[9] Alex Chojnacki, Chengyu Dai, Arya Farahi, Guangsha Shi, Jared Webb, Daniel T. Zhang, Jacob Abernethy, and Eric Schwartz. A data science approach to understanding residential water contamination in flint. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1407–1416, New York, NY, USA, 2017. ACM.

[10] Jacob Abernethy, Alex Chojnacki, Arya Farahi, Eric Schwartz, and Jared Webb. Activeremediation: The search for lead pipes in flint, michigan. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, New York, NY, USA, 2018. ACM.

[11] Ron Fonger. Flint data on lead water lines stored on 45,000 index cards. 2015. (Accessed Feb, 16, 2017).

[12] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.