Rationale for Utilizing Predictive Modeling to Identify Lead Service Lines

Author and Affiliation:

Lori A. Lester, Ph.D., Chief | Bureau of Risk Analysis

New Jersey Department of Environmental Protection

Division of Science and Research

November 15, 2022

State of New Jersey
Phil Murphy, Governor

Department of Environmental Protection
Shawn M. LaTourette, Commissioner

Division of Science & Research
Nicholas A. Procopio, Ph.D., Director

Visit the DSR website:
https://dep.nj.gov/dsr

Please cite as: Lester, L.A. 2022. Rationale for utilizing predictive modeling to identify lead service lines. New Jersey Department of Environmental Protection. Trenton, NJ. 8 pages. Available at https://dep.nj.gov/wp-content/uploads/dsr/lsl-predictive-modeling-rationale.pdf.

Charge

The Bureau of Safe Drinking Water requested that the Division of Science and Research develop a rationale to explain whether predictive modeling is as effective as other methods to locate drinking water service lines made of lead. The location of lead service lines can be determined by different methods, including screening records (such as municipal codes, plumbing codes, and construction specifications), visual examination of plumbing, water quality sampling, excavation of water service line, and predictive modeling (Hensley et al. 2021). However, each of these methods provide differing levels of accuracy for predictions and differing levels of costs. In most cases, predictive models can both improve the accuracy of locating lead service lines and reduce the costs associated with replacing lead service lines by excavating fewer unnecessary (i.e., non-lead) service lines.

Rationale

Although the use of lead pipes in the United States was not formally banned until 1986, lead pipes were used less frequently starting in the 1940s (Calabrese 1989). Prior to the Environmental Protection Agency's (EPA) promulgation of the Lead and Copper Rule (LCR) in 1991, municipalities rarely maintained accurate records of water service line material (Blackhurst et al. 2019). To comply with the LCR, water systems that exceeded the lead action level under certain circumstances were required to annually replace 7% of the lead service lines that it owned until the water system met the action level. In many cases, water systems could not rely on existing historical records to locate and replace lead service lines for compliance with the 7% replacement rule.

The EPA's revised LCR, which became effective on December 16, 2021, requires that utilities submit lead service line inventories and make that information available to the public, which may entail further investigation and evaluation by water systems using the list of sources outlined in the revised LCR (US EPA 2021). On July 22, 2021 in New Jersey, the Governor signed legislation into law for mandatory lead service line replacement, "NJ Bill A5343/S3398", P.L.2021, Ch.183, effective immediately upon signature. Under this law, the first deadline for community water systems to submit information on service line inventories occurred in September of 2021. Initial service line inventories were required in January of 2022, with updated inventories required in July of 2022 and annually in July thereafter. Further, the law requires that public water systems also make their most recent lead service line inventories publicly available.

Thus, there is an urgent need to develop and validate efficient methods to identify the locations of lead service lines for many reasons including (1) the human health consequences associated with lead exposure from lead service lines, (2) the public's desire to reduce lead exposure, and (3) the costs associated with accurately identifying lead service line locations for replacement. When deciding how to determine the location of lead service lines, there are many considerations, including costs, staffing requirements, and time (Hensley et al. 2021). Predictive modeling can help reduce the uncertainty associated with locating lead service lines in a cost-efficient manner.

Different Types of Predictive Models

There are two main categories of predictive models that are currently being utilized to locate lead service lines: (1) geospatial models and (2) machine learning models. Geospatial models use Geographic Information Systems (GIS) to analyze spatial relationships between known lead service line locations and other predictive environmental parameters (i.e., construction year and lead concentration in the water) to predict the likelihood of a lead service line at locations with unknown service line types. In response to the drinking water crisis in Flint, Michigan, both geospatial (Goovaerts 2017) and machine learning models (Abernethy et al. 2018) were developed in an attempt to locate potential lead services lines. Overall, the

machine learning models were found to be more accurate than the geospatial model (Abernethy et al. 2018).

Machine learning models expand on geospatial models by incorporating computer algorithms that improve over time as new data (e.g., historical records, water samples, city records, etc.) are incorporated into the models (Mitchell 1997, Hensley et al. 2021). In regard to locating lead service lines, machine learning models use a subset of the sample data, referred to as the "training data", to predict the locations of lead service lines at houses with an unknown service line type (Mitchell 1997). The other subset of the sample data is referred to as the "testing data", and this subset is used to evaluate how accurate the model is at predicting lead service line locations. Machine learning modeling techniques have been applied in several water systems nationwide (e.g., Flint, Michigan; Pittsburgh, Pennsylvania; Denver, Colorado) to improve the accuracy of lead service line inventories (Abernethy et al. 2016, 2018, Blackhurst et al. 2019, CO DPHE 2019, Kontos et al. 2019). The trend of municipalities and water suppliers utilizing predictive models to locate lead service lines will likely continue. For example, the City of Toledo, Ohio was recently awarded an EPA grant to create a machine learning model to predict lead service line locations (Goldstein 2020, Smith 2020).

Predictive Models are More Accurate than Historical Records

Utilizing machine learning techniques to model the locations of lead service lines is often more accurate than using historical data alone. In September 2016, the city of Flint, Michigan selected 171 homes for service line replacement (Abernethy et al. 2018). The homes were selected for replacement based on the presence of high lead levels in water or high-risk individuals (i.e., pregnant women, children under six years of age, or elderly people). During this study, which was referred to as Phase One, the rate of lead pipes discovered in houses during excavations was 96% (165/171 houses) whereas the city records suggested only 40% (68/171) of those homes would contain lead (Abernethy et al. 2018). The overall accuracy of the machine learning predictive model performed in Flint was 91.6%, suggesting that the model would be better at predicting the presence of lead than the historical records alone. In 2018, the city of Flint chose to contract with a different engineering firm who stopped utilizing the machine learning model results to prioritize locations for replacement (Fussell 2021). Instead of targeting residences with high likelihood of lead service lines, Flint's mayor demanded that the new firm excavate every house on randomly selected blocks across each of the city's wards (Madrigal 2019). The accuracy of locating lead service lines decreased by more than 15% when the new firm stopped using the predictive modeling results (Fussell 2021).

Similar trends where models more accurately predicted lead service line locations than historical records alone have been found in other areas. In Pittsburgh, many different types of machine learning models were tested to predict the probability of locating lead service lines (Gurewitsch 2019). In Gurewitsch's study, approximately 80% of lead service lines were correctly identified using the best fit model[1]. Although the overall predictive accuracy of machine learning models is generally much higher than that of using historical data alone, the development of machine learning models is not always easy. Oftentimes, the initial model needs to be improved and refined as new data become available. For example, in another Pittsburgh study, the initial machine learning model was only 73% accurate when predicting the presence of lead service lines, whereas the historical data was 63% accurate (Blackhurst et al. 2019). In addition, the model predicted that many residences had an equal chance of having lead or non-lead service lines. Thus,

[1] The best fit model was selected based on the relative area under a Receiver Operating Characteristics (ROC) curve plot. An ROC curve displays the probability of false positive versus the probability of detection. The area under the curve can range from 0 to 1 where values closer to 1 suggest better fitting models.

the predicted model results did not offer the clarity needed to be able to distinguish between lead and non-lead service lines. Due to the fairly low accuracy rate and the complications with differentiation capability of this prediction model, this initial model was not recommended to be used to estimate lead service line inventories in Pittsburgh. Researchers are currently retraining the model using 5,000 new customer samples, aiming to improve the predictive outcome.

Geospatial models have also been utilized to predict locations of lead service lines. However, the geospatial models tend to perform less well than the machine learning models because the geospatial models do not include a learning component that improves model accuracy over time as new data is acquired. In Flint, Michigan, geospatial methods were used to predict the likelihood of a home having lead and galvanized service lines (Goovaerts 2017). In particular, kriging models were generated using neighboring field data (i.e., locations with known service line type) and secondary information (e.g., construction year, city records, etc.). A kriging model is a geostatistical method to predict a probability (in this case, the likelihood of lead or galvanized service line) in unknown locations based on a given set of measurements. In Flint, the accuracy of the geospatial models for predicting lead service lines was approximately 72% (Goovaerts 2017), which was about 20% lower than the 91.6% accuracy of the best fit machine learning model (Abernethy et al. 2018).

Cost Effectiveness of Predictive Modeling

Furthermore, prioritizing locations for excavations based on the results from predictive modeling of lead service line locations is often more cost effective than depending solely on historical records. However, the costs may vary for water suppliers based on factors such as the size of the service area and the age of the town. Replacing lead service lines has a high financial impact because much of the necessary information is buried underground. It costs thousands of dollars to mechanically excavate a water service line pipe in a resident's yard to determine what material the line is made of (Abernethy et al. 2018). Considering that historical records are often inaccurate for determining service line material at individual houses, there are major cost implications because many pipes end up being excavated that do not need to be replaced. Another option is vacuum excavation, where a small hole is dug to inspect the service line (Hensley et al 2021). Vacuum excavation is more cost-effective, with costs being as low as $77 to $400 per inspection point (Hensley et al 2021, Abernethy et al 2018). Unfortunately, some lead service lines may be missed during vacuum excavations because only a small portion of the line is visually inspected. To lower costs, it is essential to be able to accurately determine where lead service lines are most likely to be found, and the results of predictive models may reduce the number of unnecessary excavations (i.e., excavations of non-lead service lines).

One of the first major cities in the United States to implement and complete a full lead service line replacement was Madison, Wisconsin (Madison Water Utility 2012, Gurewitsch 2019). The city of Madison stopped using lead for service lines in the 1920s and maintained accurate historical records, so predictive modeling was not necessary to determine the locations of lead service lines. From 2000 to 2012, more than 8,000 lead service lines were removed with a cost of approximately $15.5 million (Madison Water Utility 2012). The cost of major lead service line removal projects is commonly even higher than in Madison. Currently in Denver, Colorado, there are 64,000 suspected lead service lines, and the fiscal impact of Denver Water's lead service line removal program is expected to be between $190 and $272 million over the next fifteen years (CO DPHE 2019).

In Pittsburgh, predictive modeling is being utilized to reduce the number of unnecessary excavations (Blackhurst et al. 2019). The predictive model is expected to improve accuracy to 90%, which would greatly reduce the number of unnecessary excavations. By the end of 2019, 9,080 locations were excavated in Pittsburgh. Approximately 4,000 of these 9,080 excavations had non-lead service lines (44%). If the 9,080

excavations had been performed by following the likelihood rankings from the model, only approximately 910 service lines of the excavated service lines would be non-lead and the remaining 3,090 would have been lead service lines. In this area, the estimated cost is $1,300 per excavation. Thus, Pittsburgh spent $5.2 million excavating non-lead locations ($1,300 per excavation X 4,000 residences) originally, and they would have saved $4 million (3,090 excavations of lead service lines that were originally non-lead X $1,300) had the excavations been at the locations were predicted most likely to be lead. Similarly, in Flint, Michigan, the estimated cost to replace a single home's water service line was around $5,000 (Chojnacki et al. 2017). The predictive modeling conducted in Flint yielded approximately $10 million in cost savings due to the reduction in unnecessary excavations of service lines (Abernethy et al. 2018).

The potential cost of developing models or even paying for a consulting firm to develop and run the predictive models would likely be significantly less than the cost attributed to the trial and error associated with accurately locating lead service lines. In 2021, BlueConduit informed the Department that it was charging between $40,000-$55,000 for the first year to develop a predictive model and to test the model's accuracy (i.e., perform cross validation), and then between $15,000-20,000 per update after the first year to update the model with new data (pers. comm. with Andy Rosenblatt, 2021).

Model Criteria

Although there are many advantages to the predictive modeling of lead service line locations, the model must meet certain criteria to ensure that the results are statistically accurate. A poorly designed model is more likely to make incorrect predictions of lead service line locations. First, the baseline dataset should be generated using a randomized sampling approach where each sample has an equal probability of being selected. A randomized sampling approach ensures that the sample is an unbiased representation of the whole population of interest (i.e., area where lead service line locations will be predicted). Second, the sample size of known lead service line locations must be large enough to have the statistical power[2] necessary to accurately predict where lead service lines may be located. For example, 8,100 training data points were utilized in the Pittsburgh machine learning model (Blackhurst et al. 2019) and 15,447 data points were utilized in the machine learning model in Flint (Abernethy et al. 2016). Third, an appropriate type of model must be selected. In many of the cases mentioned thus far, the researchers utilized machine learning techniques to predict lead service line locations, but the types of machine learning models varied: in Pittsburgh the Recursive Feature Elimination with Random Forest model was performed (Blackhurst et al. 2019); in Flint the XGBoost model was utilized (Chojnacki et al. 2017); and in Denver the Random Forest model was used (Kontos et al. 2019). In most studies, multiple types of models were performed, and the model with the best fit was selected. Although machine learning models (Abernethy et al. 2018) tend to be more accurate than geospatial models (Goovaerts 2017), geospatial models may be appropriate for determining the location of lead service lines if the model results (i.e., estimated lead service line locations) are demonstrated to be accurate. Finally, the accuracy of the models (both geospatial and machine learning) must be validated using a model cross-validation technique[3] (Chojnacki et al. 2017, Goovaerts 2017, Blackhurst et al. 2019).

Due to the nature of statistical prediction, predictive modeling has its own limitations and uncertainties including: 1) the prediction accuracy and robustness largely depend on the sample size of training data sets (i.e., known lead service lines); 2) the accuracy of the input data directly affects the prediction outcome;

[2] Statistical power is the probability of detecting an effect if there is a true effect present to be detected. Power ranges from 0 to 1, with values closer to 1 being more powerful.

[3] Cross-validation is a way to assess whether a model is robust by holding out a portion of the original data (i.e., hold-out sample) to use after the model is generated to test whether the model results are correct.

and 3) technical statistical support may be needed by the utilities. Therefore, these factors must be considered in advance before utilizing predictive modeling.

Other Uses of Predictive Modeling

Predictive modeling is a commonly employed approach to solving many environmental problems. In Flint, Michigan, geospatial modeling was utilized to predict lead levels in samples of drinking water (Goovaerts 2019), and the location of lead service lines (Goovaerts 2017). However, cross-validation of this model suggested that the model performed poorly in predicting the presence of lead in drinking water at locations where the service line material was unknown (Goovaerts 2018). Machine learning models have also been implemented in Pittsburgh to predict locations with high tap water lead concentrations (Hajiseyedjavadi et al. 2020). Overall, these machine learning models were able to accurately predict high tap water lead concentrations 71.6% of the time.

In addition to utilizing predictive models for locating lead service lines and lead water levels, previous research has utilized predictive modeling successfully to predict the risk of fires in residences (76% accuracy, Lau et al. 2015) and pipe failures in water systems (~60% accuracy depending on model type; Li et al. 2014). Predictive modeling is a highly effective scientific technique that is commonly used to solve a plethora of environmental problems.

Potential Caveats to the Usage of Predictive Modeling

Some residents in communities impacted by lead water crises (especially Flint, Michigan) have reported distrust with the predictive modeling of lead service line locations (Fussell 2021). The city of Flint prioritized households for lead service line replacement according to the likelihood of a lead service line being present in that home according to the results of the machine learning model. Residents reported frustration when the model led the city to inspect their neighbor's house, but not their own. Information about vulnerable populations (i.e., children, pregnant women, and elderly) was then included as inputs in the predictive models to allow for prioritization of houses for replacement (Abernethy et al. 2018, Hensley et al. 2021). In Toledo, Ohio, officials are hoping to lessen or avoid this problem of residents mistrusting model results by expanding community outreach and involvement (Goldstein 2020, Smith 2020). The current plan in Toledo is to include the knowledge of homeowners (e.g., the number of children under 6 in a residence, the presence of pregnant women, the presence of senior citizens, etc.) to complement the modeling efforts to prioritize service lines for replacement based on likelihood of lead being present in an attempt to lessen community distrust.

Another potential limitation of modeling the location of lead service lines is that there is some uncertainty associated with the results, and thus the material of some service lines will remain unknown in the inventory. According to BlueConduit (pers. comm. with Eric Schwartz and Jacob Abernethy, 2021), the distribution of lead likelihoods produced by the machine learning models are commonly bimodal, where the majority of service lines are classified as either very likely to be lead (e.g., > 90%) or very unlikely to be lead (e.g., < 10%). Therefore, model outputs are beneficial for prioritizing houses for replacement projects, by allowing municipalities to start with houses that are very likely to have lead service lines. However, some locations will be assigned likelihoods that are closer to 50% where it is unknown whether the service line is lead. At these locations, the service line material will likely remain unknown until visual inspection. In order to determine which locations to include in the inventory as lead service lines, the water suppliers would either need (1) to demonstrate the inflection point[4] in the distribution of lead likelihoods produced

[4] An inflection point is the point in a curve where the curve changes direction.

by the predictive model where the locations are very likely to have lead service lines or (2) to select and defend a threshold where any property with a likelihood of lead higher than the threshold should be considered a lead service line (see guidance document for more information). Water suppliers would also be required to continue to test and improve the predictive model as new data are collected.

Conclusion

In conclusion, using predictive modeling to determine potential lead service line locations is an effective and scientifically-sound approach. Although predictive modeling has its own limitations and uncertainties, in the majority of the studies presented above, the uncertainty associated with other techniques (e.g., historical records, excavation, etc.) tended to be higher than the uncertainty associated with predictive modeling results. Furthermore, by replacing the lead service lines based on the more accurately predicted locations from models, the resources needed for lead service line replacement projects would be greatly reduced. Therefore, predictive modeling is an accurate and cost-effective approach to identify lead service line locations.

References

Abernethy, J., C. Anderson, C. Dai, A. Farahi, L. Nguyen, A. Rauh, E. Schwartz, W. Shen, G. Shi, J. Stroud, X. Tan, J. Webb, and S. Yang. 2016. Flint water crisis: Data-driven risk assessment via residential water testing. Pages 1–8 Bloomberg Data for Good Exchange Conference. New York City, NY.

Abernethy, J., A. Chojnacki, A. Farahi, E. Schwartz, and J. Webb. 2018. Active Remediation: The search for lead pipes in Flint, Michigan. Pages 5–14 Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery.

Blackhurst, M., H. Karimi, and S. Hajiseyedjavadi. 2019. Predicting lead water service lines in the Pittsburgh water and sewer authority service area. Pages 1–28. Pittsburgh, PA.

Calabrese, E. J. 1989. Safe drinking water act. CRC Press. Pages 1-240. Boca Raton, FL.

Chojnacki, A., C. Dai, A. Farahi, G. Shi, J. Webb, D. T. Zhang, J. Abernethy, and E. Schwartz. 2017. A data science approach to understanding residential water contamination in Flint. Pages 1407–1416 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA.

CO DPHE. 2019. Watershed & wastewater stakeholders summary report. Pages 1–71 Colorado Department of Public Health and Environment. Denver, CO.

Fussell, S. 2021. An algorithm is helping a community detect lead pipes. https://www.wired.com/story/algorithm-helping-community-detect-lead-pipes/.

Goldstein, P. 2020. AI will help Toledo, Ohio, find and replace lead pipes. https://statetechmagazine.com/article/2020/11/ai-will-help-toledo-ohio-find-and-replace-lead-pipes.

Goovaerts, P. 2017. How geostatistics can help you find lead and galvanized water service lines: The case of Flint, MI. Science of the Total Environment 599–600:1552.

Goovaerts, P. 2018. Flint drinking water crisis: A first attempt to model geostatistically the space-time distribution of water lead levels. Pages 255–275 *in* B. S. Daya Sagar, Q. Cheng, and F. Agterberg, editors. Handbook of Mathematical Geosciences. Springer, Cham, Switzerland.

Goovaerts, P. 2019. Geostatistical prediction of water lead levels in Flint, Michigan: A multivariate approach. Science of The Total Environment 647:1294–1304.

Gurewitsch, R. S. 2019. 'Pb-predict': Using machine learning to locate lead plumbing in a large public water system. University of Pittsburgh, Pittsburgh, PA.

Hajiseyedjavadi, S., M. Blackhurst, and H. A. Karimi. 2020. A machine learning approach to identify houses with high lead tap water concentrations. Proceedings of the AAAI Conference on Artificial Intelligence 34:13300–13305.

Hensley, K., V. Bosscher, S. Triantafyllidou, and D. A. Lytle. 2021. Lead service line identification: A review of strategies and approaches. AWWA Water Science 3:e1226.

Kontos, C., C. Pawlowski, M. Harris, and E. McIlwee. 2019. Appendix III.B.3 Predictive model and prioritization. Pages 1–14 Denver Water. Denver, CO.

Lau, C. K., K. K. Lai, Y. P. Lee, and J. Du. 2015. Fire risk assessment with scoring system, using the support vector machine approach. Fire Safety Journal 78:188–195.

Li, Z., B. Zhang, Y. Wang, F. Chen, R. Taib, V. Whiffin, and Y. Wang. 2014. Water pipe condition assessment: A hierarchical beta process approach for sparse incident data. Machine Learning 95:11–26.

Madison Water Utility. 2012. Information for utilities on lead service replacement. https://www.cityofmadison.com/water/water-quality/water-quality-testing/lead-copper-in-water/information-for-utilities-on-lead-service.

Madrigal, A. C. 2019. How a Feel-Good AI Story Went Wrong in Flint. The Atlantic:1–12.

Mitchell, T. 1997. Machine learning. Pages 1–414. McGraw Hill.

Smith, A. 2020. City of Toledo receives EPA grant to utilize artificial intelligence to identify lead water lines. https://freshwaterfuture.org/uncategorized/city-of-toledo-receives-epa-grant-to-utilize- artificial-intelligence-to-identify-lead-water-lines/.

US EPA. 2021. National Primary Drinking Water Regulations: Lead and Copper Rule Revisions. Page 86 Fed. Reg. 4198.