Guidance for Utilizing Predictive Modeling to Identify Lead Service Lines for Inventory Development

Author and Affiliation:

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

New Jersey Department of Environmental Protection *Division of Science and Research*

February 9, 2024 [1]

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

[1] The definition of the "three-point verification method" on page 5 was updated on 2/9/2024. All other text in this version remains consistent with the original 11/15/2022 predictive modeling guidance document.

Please cite as: Lester, L.A. 2024. Guidance for utilizing predictive modeling to identify lead service lines for inventory development. New Jersey Department of Environmental Protection. Trenton, NJ. 9 pages. Available at https://dep.nj.gov/wp-content/uploads/dsr/lsl-predictive-modeling-guidance.pdf.

Introduction

Many water systems in New Jersey are currently required to develop a service line inventory to identify lead service lines within their system (See: N.J.S.A. 58:12A-42, 40 C.F.R 141.84 (a)(6)). Water systems may develop this inventory using a number of state approved methods for identifying lead service lines, including predictive modeling. For some water suppliers, predictive modeling may be a useful tool for obtaining a more accurate estimate of the number and the locations of LSLs throughout their system. However, predictive modeling of LSLs is not required for all water suppliers. If a water supplier elects to utilize predictive modeling in creating its LSL inventory, this guidance should be closely followed, and may be amended and supplemented in the future. The main goal of this guidance document is to provide more detailed information on the best statistical practices for utilizing predictive modeling to determine the likelihood of a property containing an LSL. Simultaneously, the New Jersey Department of Environmental Protection (Department), Division of Science and Research has released a rationale in support of the utilization of predictive modeling as an effective and useful method for locating drinking water service lines made from lead. The rationale synthesized the available literature on predictive modeling of LSL locations and determined that predictive modeling, in conjunction with other resources, may be an appropriate tool for some water suppliers.

Predictive modeling may be especially helpful in areas where historical records are known to be inaccurate or incomplete. Prior to the promulgation of the Federal Lead and Copper Rule (LCR) in 1991, some municipalities rarely maintained accurate records of water service line material (Blackhurst et al. 2019). Predictive modeling may also be useful in areas where many buildings were built prior to the lead ban. The use of lead pipes was prohibited in the United States beginning in 1986 (New Jersey banned the use in 1987), but lead pipes were used less frequently starting in the 1940s (Calabrese 1989). Although the usage of LSLs decreased starting in the 1940s, galvanized service lines continued to be commonly installed in properties until 1960 (NJDEP 2022).

Although some water suppliers may find predictive modeling helpful to prioritize properties for LSL replacement and to develop their inventory, predictive modeling of LSL locations may be unnecessary for other water suppliers. For instance, the historical records may be accurate enough to develop an inventory directly from records in some areas. In other cases, the majority of buildings may have been constructed more recently than the lead ban, and thus LSLs may not be a concern in these areas. Furthermore, predictive modeling may be more useful for large systems with many service lines and/or regions with many service lines of unknown material. However, the following guidance, especially the sections on data management and transparency, may still be useful for suppliers who choose not to use predictive modeling as a tool.

Water suppliers do not need to physically verify every service line in order to develop or use a predictive model. Instead, the water suppliers are required to inspect a statistically sound subset of their service area[1]. However, even when predictive modeling is utilized, the Department may require additional excavations, if deemed necessary, to identify the service line material (see N.J.S.A. 58:12A-42(f)(2)).

[1] Further details regarding best practices for performing a predictive model are provided below in Sec. 3: Verify service line material for a randomly selected sample of service lines.

To ensure that the predictive modeling results are accurate and interpreted appropriately, water suppliers should adhere to the following fundamental statistical and research principles when using predictive modeling to estimate the service line material at each property in their system:

1. Develop a data management plan.
2. Evaluate historical records.
3. Verify service line material for a randomly selected sample of service lines.
4. Demonstrate that the predictive model is accurate for the supply region.
5. Ensure transparency by submitting model results and explaining how results were utilized.

The Department may request documentation supporting determinations made by a water supplier, including but not limited to, a written predictive modeling report that fully describes (1) how the predictive model was developed, evaluated, and assessed for accuracy, (2) the results of the model (i.e., likelihood of each property having an LSL), and (3) how the results were utilized to inform the prioritization of properties for service line replacement and/or to develop an inventory. Records pertaining to the development of the model and inventory should be maintained/retained pursuant to 40 C.F.R. 141.91. The Department strongly recommends that this predictive modeling report be submitted along with the annual inventory required pursuant to N.J.S.A. 58:12A-42(f), and the predictive modeling report may be subject to the Department's approval in the future. Furthermore, the predictive modeling report may be updated and resubmitted to the Department on an annual basis as new data are incorporated into the model.

No information contained in this guidance obviates any legal requirements under the Lead and Copper Rule or the Safe Drinking Water Act rules. Also, please note that although several companies are referenced in the following guidance, the Department does not officially endorse any of these entities.

1. Develop a data management plan

In addition to submitting a service line inventory report to the Department, the water supplier should collect and maintain all historical data associated with each property's service line for their service area in an organized way. The data should be stored in a spreadsheet or database computer program (e.g., Microsoft Excel or Access), or a more sophisticated database software (e.g., PostgreSQL, dBASE, etc.) if necessary due to a large amount of data. Within each spreadsheet, each property should have its own row, and each column should contain the different data available regarding the service line material at that property (Table 1).

For each property, separate columns should exist for the utility side of the service line and the property owner side of the service line, as applicable, to ensure that both sides of the service line are considered for material type. The database must be designed to contain all types of data that may be needed or desired to be known. Any old physical records, such as notecards or maps, should be digitized and included in the database. If more than one record is found for the same location, both should be included in the database (i.e., no data should be discarded). At minimum, the database should contain all available historical information regarding both the public and private sides of the service line. Finally, a "data dictionary" should be included in the database which thoroughly defines and explains the column headers and what information is contained in each column. The Department may request that the database and the "data dictionary" be included in the predictive modeling report.

Table 1. This sample database (BlueConduit 2020) shows an example of how to potentially organize historical records and service line records that were verified by physical examination of the line.

Verification Details					Location Details		Historical Details			
Verified Public Service Line Material	Verified Private Service Line Material	Date Verified	Method	Contractor	Parcel ID	Address	Historical Public Service Line Material	Historical Private Service Line Material	Date of Historical Records	Year Built
COPPER	COPPER	12/6/18	Excavation	Firm 1	4489186533	60 KALAMAZOO AVE	COPPER	COPPER	12/01/56	1951
GALVAN-IZED	LEAD	10/25/17	Excavation	Firm 3	5006830967	34 OAK ST	GALVAN -IZED	COPPER		1935
GALVAN-IZED	LEAD	6/20/18	Excavation	Firm 2	9362055119	31 CATHERINE AVE	UN-KNOWN	COPPER		1927

2. Evaluate historical records

Some historical records that indicate the service line material may be inaccurate, and thus water suppliers are responsible for assessing whether historical records are accurate in their service area. Many potential sources of service line data may exist, including water main repair records, water meter replacement records, old construction records, etc. However, the accuracy, accessibility, and reliability of these records may vary by record type and location. Replacements may have been completed over time that were not properly recorded or historical records may be inaccurate or incomplete. Some types of historical records, such as recent records, may be more accurate than others. For example, if the service line material was confirmed in 2018 during road construction, this record is likely more accurate than a 1956 record handwritten on a notecard that states the material was “copper?”.

Water suppliers should track what materials were found as service lines are inspected and replaced, so that they can compare these results with the historical records. The Department suggests that water suppliers submit an evaluation of their historical records in their predictive modeling report along with their annual inventory submission required pursuant to N.J.S.A. 58:12A-42(f). In order to evaluate historical records, the water suppliers may develop a “Historical Records Materials Confusion Matrix”. This confusion matrix would report whether or not the historical records were found to be accurate following physical verification. The matrix would also calculate the percentage of times that the historical records were accurate. An example confusion matrix is displayed below (Table 2). The top highlighted blue row shows the number of properties with a historical record stating that the service line was copper that were verified to have each service line material (e.g., Copper-Copper, Copper-Galvanized, etc.). The bottom highlighted blue row displays the percentage of service lines that were assigned to each verified service line material (e.g., 1,115 properties were assigned to Copper-Copper out of a possible 1,489; i.e., 1,115/1,489*100 = 75% accuracy).

Table 2. This sample "Historical Records Materials Confusion Matrix" (BlueConduit 2020) reports whether historical records were found to be accurate. For example, the cells highlighted in blue under the "Copper-Copper" header show the number of service lines (1,115) and the percentage of total service lines (75%) that were correctly identified as copper.

Verified SL Materials (Public-side Material - Private-side Material)							
Historical records	Copper- Copper	Copper-Galvanized	Lead-Copper	Lead- Galvanized	Lead- Lead	Other Safe Materials (e.g., plastic)	Totals for historical records by label
Copper	1115	10	258	84	13	9	1489
	75% (A)	1%	17%	6%	1%	1%	100%
Copper/ Lead	109	20	816	91	15	25	1076
	10%	2%	76%	8%	1%	2%	100%

3. Verify service line material for a randomly selected sample of service lines

Water suppliers will be responsible for verifying the service line material for a randomly selected sample of service lines in their respective service regions when developing a predictive model. Although service line material data may exist from previous construction projects, these data may be misleading to use as input to a predictive model because it may not be representative of the whole system (e.g., the data may be biased because it was only collected in one section of the supply region). If a sample is not representative of the whole system, the predictive model may produce incorrect results. Instead, the randomized sampling approach should be utilized to randomly select a subset of service lines from the water supply region. According to the randomized sampling technique, each building serviced by the system should have an equal chance of being selected as a sample.

In order to select random samples, the water supplier should start by deciding what area of interest to study which will likely be the whole region that they service. Then, an appropriate sample size must be determined for the dataset. A large enough sample must be collected in order to ensure that the resulting dataset accurately reflects the entire water system. In order to determine the appropriate number of samples necessary, the following information may be needed: (1) number of service lines in the system, (2) best estimate of the number of LSLs in the system, (3) acceptable amount of error (also commonly referred to as the confidence interval, e.g., ± 5%), and (4) desired size of confidence level (e.g., 95%). An online sample size calculator may be useful to determine the appropriate sample size for the water system (e.g., https://www.surveysystem.com/sscalc.htm).

After the appropriate sample size has been determined, buildings should be chosen randomly from all that are available in the region of interest. The random number method could be utilized to randomly select the sample. This method involves assigning each building in the area of interest a random number[2], and then selecting the first records up to the pre-determined sample size as the random

[2] Random numbers can be generated using the random number function (RAND) in Microsoft Excel.

sample (i.e., if the appropriate sample size is 2,000, then buildings 1 through 2,000 would be selected for the random sample).

Once the proper sample size is determined, the water suppliers should verify the service line material at each property for the randomly selected sample. Visual inspections will be necessary to confirm material at the sampling locations. The randomized sample can then be utilized to create a predictive model to estimate the probability of finding an LSL at locations where the service line material is unknown or questionable because of inaccurate historical records. For each service line of unknown or questionable material, the materials of all service line portions should be physically verified (i.e., all segments of the service line and where the line enters the building). If goosenecks or connectors have been used, the materials of these should also be verified. In particular, the material must be confirmed using a three-point verification method, which means visually confirming (1) the pipe entering the curb box (main to curb stop), (2) the pipe leaving the curb box (curb stop to building), and (3) the pipe entering the building (interior portion, connected to the premise plumbing). Statistical models do not replace the need for physical verifications, including possible excavation (see N.J.S.A. 58:12A-42(f)(2)), but can be used to inform service line inventory and replacement programs by prioritizing locations with high likelihood of LSLs for replacement.

Once the dataset has been collected, an appropriate predictive model can be selected and performed. In previous studies, machine learning models (e.g., Xgboost, random forest, etc.) tended to perform best (Abernethy et al. 2016, 2018, Blackhurst et al. 2019, CO DPHE 2019, Kontos et al. 2019). Geospatial models were attempted in Flint, MI (Goovaerts 2017), but did not perform as well as machine learning models. Once a system decides to use predictive modeling, the water supplier will be expected to continue training and improving the model throughout the replacement program. The methods utilized to verify service line material for the random sample may be described in the predictive modeling report provided to the Department, and all data records should be maintained by the water supplier.

4. Demonstrate that the predictive model is accurate for the supply region

If a water supplier decides to utilize predictive modeling to determine the likelihood of LSLs at properties for development of the LSL inventories, they should assess the accuracy of their model continually throughout all stages of model development and usage. In particular, the hold-out sample method[3] should be utilized to evaluate model performance from development to implementation and improvement. By assessing accuracy of the model, this will ensure that the model predictions are accurate and thus can be used for decision-making purposes. Without an accuracy assessment, there is no way to know whether the results are meaningful.

The accuracy of the predictive model may be assessed by (1) comparing the model results from the data used to train the model to the results from using the hold-out data to test the accuracy of the predicted outcomes and (2) comparing the model results to data collected in the field. For the first phase of accuracy assessment, a portion of the initial dataset (e.g., typically 20-30% of samples) could be utilized as a hold-out group and thus not used to train the model during development. The remaining percentage of the initial samples (e.g., 70-80% of samples) would be utilized to develop and train the model. Following model development, the accuracy of the model results could then be tested using the hold-out sample to see whether the model can accurately determine the service line material in the

[3] The hold-out sample method refers to the process of setting aside a portion of the original dataset to utilize after the model is developed to test model accuracy (i.e., the hold-out sample is not used to train the initial model). A hold-out sample is commonly referred to as “test” data.

hold-out samples with known service line material. For the second phase of model accuracy assessment, the water supplier would evaluate the model results by comparing the model results to new data collected in the field. For example, the water supplier could perform inspections and replacements for a few months following model development and then calculate the percentage of properties that had LSLs during inspections. If the percent of properties with LSLs during excavations (e.g., 75% of houses had LSLs) were similar to the prediction from the model (73% of properties were expected to have LSLs), it could be assumed that the model was accurate.

For further information, it may also be helpful to refer to the many previous predictive models that have utilized similar techniques to assess model accuracy (Chojnacki et al. 2017, Goovaerts 2017, Abernethy et al. 2018, Blackhurst et al. 2019). Finally, the Department may request that the water supplier submit both types of accuracy assessments (i.e., hold-out and field assessments) in their predictive modeling report annually along with the inventory as new data are incorporated into the model as inspections and replacements continue.

5. Ensure transparency by submitting results and explaining how results were utilized

Water suppliers should be transparent when utilizing predictive modeling to determine potential LSL locations for inventories. In many cases, water suppliers may need assistance from statistical consultants with expertise in predictive modeling techniques. A basic summary of the predictive modeling methods and assumptions, data, and results should be provided by the supplier to the Department in their predictive modeling report. The public may appreciate access to maps showing locations of potential LSLs. Many communities have released maps showing the potential likelihood of LSLs at properties including Newark, NJ (CDM Smith 2021), Pittsburgh, PA (Pittsburgh Water & Sewer Authority 2021), and Flint, MI (Webb et al. 2021).

The results from predictive models may be helpful as water suppliers prioritize locations for LSL replacements. The model results will show the likelihood of an LSL at each property in the service area where the service line material is unknown and/or the historical records are questionable. Water suppliers could potentially use these likelihoods to target the properties most likely to have LSLs and ensure that replacements happen as soon as possible in these areas, while leaving properties with lower likelihoods of LSLs for later in the replacement program.

If water suppliers choose to use predictive modeling results to inform their inventory, they should demonstrate to the Department how they determined which service lines were listed as lead in their inventory in the predictive modeling report. Two options may be considered by water suppliers for selecting which service lines to include as lead in their inventory including: (1) the inflection point[4] in the distribution of lead likelihoods produced by the predictive model may be demonstrated where the locations over this inflection point are considered likely to have LSLs or (2) any property with a likelihood greater than a certain threshold should be considered a likely LSL.

The inflection point approach may be especially useful in regions where there are small numbers of properties with likelihoods near 50%. For example, Figure 1 displays a scenario where most homes (*n* = 58,000) have a low likelihood of having LSLs (i.e., 0-10% likelihood of lead). Moreover, few properties in this case have likelihoods near 50% where there is a close to equal chance of having lead or non-lead

[4] An inflection point is the place in the histogram where the bars change direction (see red line in Figure 1).

material (i.e., the material remains unknown). In this example, the inflection point occurs at 70% likelihood of lead (red line) because the number of properties for each likelihood bin starts to increase again. If the material remains unknown at a large number of properties (i.e., many properties have likelihoods of lead near 50%), the inflection point approach should not be utilized. In addition, if the likelihoods do not continue to increase from the inflection point to the maximum likelihood (e.g., the >70-80% bin has a larger number of houses than the >90-100% bin), then the inflection point approach should be deemed insufficient.

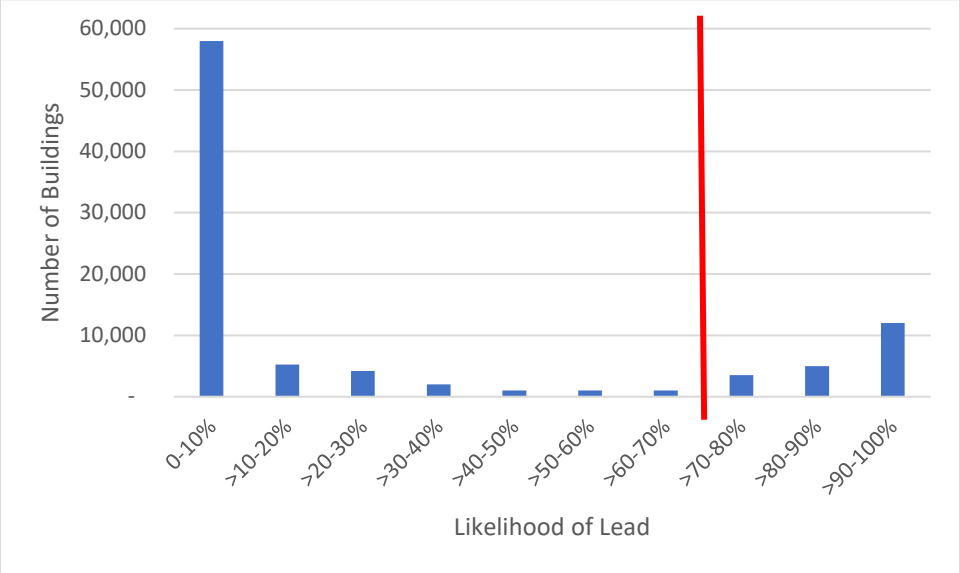

Figure 1. Identifying the Inflection Point in the Likelihood of Lead Dataset. To determine which properties to count as LSLs in the inventory, the inflection point should be demonstrated in the data where locations over this point are considered likely to have LSLs. In this example, the inflection point approach would suggest that any property with a likelihood of lead greater than 70% (red line) should be included in the inventory as an LSL.

If the inflection point approach is inappropriate, a threshold may be selected above which the water supplier considers service lines as lead for inventory purposes. The decision of where to set the threshold for determining which properties to identify as served by a LSL in the inventory must be thoroughly explained and defended in the predictive modeling report, and the Department retains the authority to reject the threshold. As previously mentioned, the Department retains the authority to request physical verification, including possible excavation of additional service lines. Even when there is a low likelihood of a property having an LSL (e.g., 90% chance that service line is non-lead), there remains a small chance that the property will have an LSL (e.g., 10% chance of LSL). Water suppliers are encouraged to contact the Department prior to submitting the predictive modeling report with any questions, particularly those regarding the usage of predictive modeling results to inform the inventory.

Conclusion

Predictive modeling may be a useful tool for some water systems, particularly in regions with inaccurate historical records and/or properties built prior to the lead ban (Calabrese 1989). The likelihood of a service line being lead is an important component to the LSL inventory and replacement decisions, but it is not the only criteria. Other components such as equity, logistical constraints, and high-risk populations

(i.e., elderly, pregnant people, or children) are also important to consider. For additional information regarding the usage of predictive modeling to inform service line inventories, see Chapter 5.5 in EPA's "Guidance for Developing and Maintaining a Service Line Inventory" (EPA 2022). In conclusion, the Department encourages water suppliers to utilize the most accurate data and predictive modeling, if deemed useful, to inform decision-making, to plan strategically, and to protect the health of all individuals in their systems.

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 University of Pittsburgh. Pittsburgh, PA.

BlueConduit. 2020. Principles of data science for lead service line inventories and replacement programs. Pages 1–17 White paper prepared for the Association of State Drinking Water Administrators (ASDWA).

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

CDM Smith. 2021. Lead service line replacement program. https://www.newarkleadserviceline.com/check-your-address.

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.

EPA. 2022. Guidance for developing and maintaining a service line inventory. Pages 1–164 Office of Water. Washington, DC.

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.

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.

NJDEP. 2022. For residents receiving notice of a lead service line. https://www.nj.gov/dep/lead/notices.html.

Pittsburgh Water & Sewer Authority. 2021. Lead map. https://lead.pgh2o.com/your-water-service-line/planned-water-service-line-replacement-map/.

Webb, J., S. Woods, J. Abernethy, and E. Schwartz. 2021. Flint water service line materials map. https://www.flintpipemap.org/.