Principles of Data Science for Lead Service Line Inventories and Replacement Programs

A white paper prepared for the Association of State Drinking Water Administrators (ASDWA)
by BlueConduit

September 2020

BlueConduit

asdwa
Association of State
Drinking Water Administrators

If you have questions, contact Ian Robinson at ian@blueconduit.com or Wendi Wilkes at wwilkes@asdwa.org

Introduction: Why data science is helpful to State Drinking Water Administrators and water systems

Substantial uncertainty still surrounds the nation's water systems regarding the number and locations of lead service lines (LSLs). This uncertainty stems primarily from inadequate records and data.

The accuracy of these records carries significant implications for utility operations, public health, regulatory compliance, and long-term asset management. Inaccurate or missing records make it difficult for utilities to understand the size of their LSL inventories, prepare budgets for replacement programs, address potential public health risks, prioritize high-likelihood homes for replacement, comply with state and federal regulations, control the cost of replacement programs, use this information to inform other capital projects, and communicate accurate information to their customers.

EPA's Proposed Lead and Copper Rule Revisions mandate that all US community water systems must develop an LSL inventory or demonstrate the absence of LSLs in their system within the first three years of final rule publication (the final rule publication is expected in the Fall of 2020). Each LSL inventory must be updated annually thereafter. Some states already have rules that go further than the existing federal guidance. Recent federal legislative efforts that specifically include funding for LSL replacement in addition to existing targeted funding and financing for full LSL removal make it even more imperative that utilities know the composition of their service lines (SLs). Such programs include the 2016 Water Infrastructure Improvements for the Nation Act grant programs, the Drinking Water State Revolving Fund (DWSRF), and the Water Infrastructure Finance and Innovation Act (WIFIA) program.

The kind of uncertainty that the LSL question presents is well-suited for data science methods that have evolved in recent years. There has been an increased adoption of predictive methods by utilities to guide their decision-making. With improved technology and innovative modeling approaches, there is greater ability to generate precise predictions for increasingly complex questions. Given the significant public health, regulatory, and financial implications of these decisions, it is essential that regulators and utilities be aware of and adhere to some fundamental statistical methods when using predictive methods to inform SL work.

The appropriate application of these methods supports both regulators and utilities in their work to ensure safe, clean, and affordable drinking water. Some of the principles in this white paper, such as good data management and transparency, should be considered as an important foundation for any utility, regardless of size or approach to building an LSL inventory. The principles contained in this document are relevant for all types of statistical analyses, from a simple linear regression to a machine learning model. Some of the more complex data science methods can be utilized for higher levels of sophistication and accuracy. Regardless of the utility size or statistical method used, these principles can inform a straightforward, data-driven approach.

Increasing the accuracy of LSL inventories yields actionable data that can be used to make replacement programs more efficient and reduce the risk of potential lead exposure. Having a clearer picture of how many LSLs might be in a water system and where they are concentrated is essential for accurate budgeting and management at the outset of and throughout replacement programs. Excavating every unconfirmed water SL would eliminate all uncertainty for inventories, but that kind of effort would cost too much money and time for utilities. Utilizing good data management and data science techniques can help water utilities build home-by-home predictions of SL material to guide decision-making for inventory and replacements programs.

Utilities can use multiple methods of SL identification, and they come with varying levels of accuracy. Visual verification of SL materials, through in-home inspection, curb box inspection, or potholing or hydrovacing on both sides of the line provide certainty of the pipe materials for the sections of SL that they reveal. Sequential sampling of an SL taken at the tap and analyzed for metals has also been used to gain insight about pipe composition but does not provide the certainty of visually inspecting each portion of the line. This list is not exhaustive and there may be other methods used to determine SL material. More information can be found on EPA's Lead Service Line Identification and Replacement Webinars. These methods are not the focus of this paper but data collected from these efforts can certainly be used in conjunction with the principles of this paper and to inform a statistical analysis.

Given that there will always be some uncertainty, these principles provide industry best practices to reduce that uncertainty through well-applied statistical methods, as well as steps to overcome practical issues in creating better SL materials inventories. The sections below detail the principles needed to use well-established, data-driven practices from statistics in order to estimate the SL material at every home in your system. These sections also recognize the realities that each water system is unique; their existing data management practices vary; their abilities to perform data analyses, or access and/or afford the services of experts who can, at each level of complexity, also vary. With this in mind, we put forward the following set of guiding principles:

1. Clean data management and organization;
2. Not accepting all historical records as truth;
3. Conducting a representative randomized sample of service lines;
4. Transparency in public outreach and reproducibility; and
5. Accuracy on held-out sample.

These principles can be used by regulators in guidance and/or rulemaking to encourage water systems to plan strategically, make data-driven decisions, set budgets and requests for funds, build capacity in some skill areas, communicate with the public and build trust, and, most importantly, continue to protect the health of all individuals in the system.

Throughout this paper, we will show how these methods were used by the city of Flint, Michigan and regulators in practice as part of Flint's LSL replacement program. In 2016, a team of researchers from the University of Michigan began working with Flint's SL replacement program. By that time, it was understood that the city's LSLs were the main source of lead in the drinking water but two key questions stood in the way of their progress: how many lead pipes are in the city and which homes have lead pipes? The researchers applied fundamental statistical methods to this problem. While the specific nature of the public health emergency in Flint is unlike the situation in other cities across the country, the inaccurate, outdated, and incomplete nature of their data is consistent with what other communities have faced in their LSL replacement programs. In this way, the Flint case study can provide insight for other water systems across the US. Other cities have also started using predictive models to guide their replacement programs. Their experiences and lessons will continue to inform how these methods are applied.

In 2019, the University of Michigan researchers who worked in Flint formed BlueConduit with the mission of supporting the large-scale removal of lead and other dangerous pipe materials from cities.

Principle 1: Clean data management and organization

The first principle in running any statistical analysis is to make sure that data is organized and consistent. Ensuring the data is organized means all of the information related to a point of service is associated with that point of service (i.e., a home's specific water SL). This refers to how the data is organized in the spreadsheet, database, or GIS system. Based on the number of data points and information collected, a system may be able to use Microsoft Excel and Access or may need to use a more sophisticated software to handle larger amounts of data. In practice, this means that each home/SL is a row in the spreadsheet and each column of the spreadsheet refers to different data points available about each SL (e.g., parcel information, age of home, water testing results, inspection records, historical records -- see Table 1 on page 5). Separate columns for the homeowner and utility segments are critical to keep track of what is known and unknown for each portion of the SL. There can also be a separate column for information regarding goosenecks, where applicable. Labeling should be consistent across the dataset and specify the pipe material (e.g., Copper, Lead, Galvanized, Brass, Plastic, Unknown, etc.).

An important planning step in setting up this database, or any database, is to make sure that it is designed to integrate any data points that are needed or desired. This level of detail is not only important for utility planning, but may be necessary for regulatory compliance. The proposed Lead and Copper Rule Revisions provides the following definition of an LSL:

> "Lead service line means a service line made of lead, which connects the water main to the building inlet. A lead service line may be owned by the water system, owned by the property owner, or both. For the purposes of this subpart, a galvanized service line is considered a lead service line if it ever was or is currently downstream of any lead service line or service line of unknown material. If the only lead piping serving the home or building is a lead gooseneck, pigtail, or connector, and it is not a galvanized service line that is considered an LSL the service line is not a lead service line."

Having a database or spreadsheet with intelligible data on each SL is a crucial baseline for establishing an LSL inventory no matter what level of analysis will ultimately be performed. Some utilities already have data management systems that do this; those that do not can generate spreadsheets with the information they have.

In some cases, utilities will have old physical records (e.g., notecards or maps) that have not been digitized. These can be an important piece of data in trying to predict or identify SL material and therefore must be digitized to be factored into any analysis. Digitizing paper records is a best practice for utility data management, whether it be for SL inventories or other records. New technologies, such as optical character readers, can digitize these old records, transforming images into spreadsheets. The outputs of these processes integrate directly into a utility's data management system.

In addition to being in a single spreadsheet or database, it is important that the data be consistent and able to be understood by current and future utility employees, in addition to external entities (e.g., regulators, consultants, or construction contractors). It must be clear what each of the columns refers to and what the labels in each column refer to. Data analyses often include a "data dictionary" or metadata explanation, which is used to define these column headers. Articulating what each of the columns means allows those using the spreadsheet across the utility and those using it at different times to be able to understand and replicate the analysis conducted. Defining the columns also ensures that each point of service is being assessed with the same criteria. For consistency, it is also important that utilities not simply "overwrite" data when work is completed, to ensure that a history of the record is maintained. Keeping a record of the pipe material and whether they were visually confirmed or from a historic record is helpful for any predictive analysis, as well as for tracking and communicating LSL replacement progress. This allows utilities to indicate to customers what pipes were made out of and when they did maintenance at an address, and also the assumed materials for pipes that have not been replaced yet.

Organizing data in a way that it can be searched, sorted or modeled is a necessary foundation to any analysis. Setting this up in a comprehensive way at the outset is important to making sure it is usable.

From a water administrator perspective, it can be helpful to include this principle in information regarding LSL replacement trainings, templates, and policies that administrators might develop. Better organized data makes it much easier for regulators to do their work, and for utilities to comply with regulatory needs.

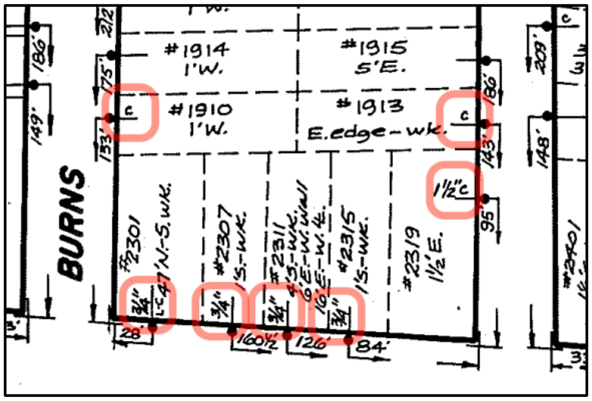
Under the proposed Lead and Copper Rule Revisions, water systems will have to submit an inventory of LSLs and SLs of unknown material to their state primacy agency and will then annually have to submit an updated inventory that reflects LSLs replaced and SLs of unknown material that have been evaluated. The proposed rule requires the states to maintain a record of all public water system LSL inventories and the annual updates. This information is necessary for the State to calculate its goal and mandatory LSLR rates and is a way to verify correct tap sample site selection tiering. The proposed rule also states that primacy agencies report the current number of LSLs at every water system to EPA. Additionally, the proposed rule includes, as a requirement for primacy, that “States would be required to provide a description of acceptable methods for verifying service line material under this proposal. Verification methods could include consultation of existing records or the physical examination of the service line.”

Attention water administrators:
Good data management is a crucial way to assess system progress and track regulatory compliance. Encouraging utilities to set up their LSL inventories in clear, accessible ways will improve efficiency and efficacy. Offering workshops and guidance on best practices in data management for utilities will help both utilities and administrators.

Attention water utilities:
Creating a data dictionary that clearly defines each of the columns in a dataset is an important step for any predictive model, including an SL materials dataset. It is important to include all data associated with an address in a data management system. It should be clear that each row should correspond to a service point.

Lessons from Flint:
At the beginning of the Flint pipe replacement project, the city had very few records of SL replacements, and all of the historical records were on index cards. The researchers digitized all of the records, which revealed the historic SL information records. Although these were not always accurate, the patterns in these records (and how accurate or inaccurate certain areas were) proved to be among the most useful pieces of data in the predictive model.

Figure 1: Digitization of Historical Records in Flint



The above image is an excerpt of a hard copy map in Flint. Records such as this were digitized by University of Michigan researchers using Optical Character Recognition technology. The digitization of records is crucial to incorporating this historical information into analyses.

Table 1: Excerpt from Flint Dataset Showing Historical and Verified Service Line Records

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
					1838914087	11 W MAIN ST	COPPER	COPPER		1952
COPPER	COPPER	11/13/18	Excavation	Firm 2	2870336116	31 WILSON AVE	COPPER	COPPER	03/15/86	1939
LEAD	GALVAN-IZED	5/25/18	Hydrovac	Firm 1	6187958482	13 E GROVER AVE	GALVAN -IZED	COPPER	01/11/52	1871
COPPER	LEAD	11/28/16	Excavation	Firm 3	5228472757	26 LAWN DR	Unknown	COPPER	06/15/29	1926
					3172003336	92 CLIFTON AVE	COPPER	COPPER		1955

The above table is an example of clean data management and organization (Addresses and Parcel IDs have been modified for privacy). Each SL has its own row and each column has a clearly defined title. Note that all of these addresses had the private SL as "copper" in historical records, yet three were verified as lead and one as galvanized. Blank cells indicate that there was no information available or no inspection yet conducted.

Principle 2: Not accepting all historical records as truth

There are many potential sources of data about SL information, and the types and accuracy of SL data will vary between water systems. Existing data about SL materials comes from different sources (e.g., water main repairs, water meter replacement programs, old construction records) and the accuracy and reliability of these records varies by record type and location. Experience in Flint and other cities is that historical records can be misleading (See Table 2 for a matrix that shows the patterns between the historical records and verified materials in Flint). Replacements may have been made over time without proper record keeping or records simply may be incomplete or incorrect. It is therefore crucial to establish how correct a water system's historical records are. Out of caution, it is important for water systems that historical labels not be considered as truth.

Many water systems will not know which types of records are correct and which are not. It is important to establish an understanding of how accurate those records are, while also noting that some types of records are going to be more accurate than others. In cases where there are recent or high confidence records indicating SL material, such as an SL whose material was confirmed as part of road construction in 2014, those types of records might be considered accurate. In other cases, where workslip index cards for buildings built in the 1950s-60s say "copper-?" water systems may regard these records as less reliable. The available and relevant documentation are often unique to each community based on the historical development patterns of its water distribution system. The process of learning just how accurate (or otherwise) a system's records are is a powerfully informative piece of this data-driven approach.

Attention water administrators:

This level of detail is not specifically called out in the proposed rule but could be included in guidance and training as a tactic for developing LSL inventories. A state may include it in a statewide template. When water utilities submit their LSL inventories, asking them to document the assumptions made in creating their inventories allows for increased accountability and transparency in their process. Having each utility submit using the same statewide or nationwide standard spreadsheet format electronically also facilitates future analyses like estimating overall scope and budgets. Utilities could submit a summary of the reliability of their historical records using a Historical Records Materials Confusion Matrix.

As outlined in Principle 1, historical records can be informative and should be included in the LSL inventory process. They should be preserved and not overwritten when replacements are made, as past existence of lead versus copper can be important data for SL predictions or future analyses. For example, if two SL records for the same location, from different points in time, are found, they should both be captured and present in the database to ensure a full history.

A clear way to understand the accuracy of historical records and to indicate that the water utility acknowledges the records' imperfections would be to report a "Historical Records Materials Confusion Matrix" (see the Table 2 below for an example). This is a table that simply counts the number of times historical material records say, for example, "copper," but the verified SL shows it is actually lead (or how often does the historical record say "galvanized," and the actual verified SL is copper). The percentage of times historical records aren't accurate (or are just incomplete), could all be summarized concisely in a single table for the utility and the state regulators.

Attention water utilities:

Historical records are important to maintain and can be an informative input into a predictive model, but treating them as the truth can lead to suboptimal decisions. Moreover, moving forward, utilities should develop processes to track SL materials in their daily operations.

Lessons from Flint:

Our team's experience in working in Flint and other cities is that historical records can be misleading. Table 2 below shows the results from the first 5,154 homes in Flint's SL replacement program (the sum of the total on the far right column in Table 2). It indicates the verified material of SLs in Flint based on historical records. The left column includes the different material labels that were found in the historical records. Each of the columns across the top refers to the materials that were found upon inspection. For example, when the historical records indicated "Copper," the replacement crews found a "Copper-Copper" connection 75% of the them (labeled A in Table 2). This means that historical copper labels turned out to have lead in some portion of the SL 25% of the time. Alternatively, 29% of the time that a historic record indicated "Lead," the SL was "Copper-Copper" (labeled B in Table 2).

Table 2: Confusion Matrix from Flint

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%
Galv/ Other	113	18	565	1286	81	31	2094
	5%	1%	27%	61%	4%	1%	100%
Lead	24	2	29	14	12	3	84
	29% (B)	2%	35%	17%	14%	4%	100%
Unknown	152	18	535	1169	118	42	2034
	7%	1%	26%	57%	6%	2%	100%

The above table shows the results from the first 5,154 homes in Flint that had their SL inspected/replaced as part of the replacement program. It indicates what the historical records said about each SL material (rows) and what was identified through physical inspection (columns). In the columns, the data is presented as "public side material - private side material" for the SL. The "Lesson from Flint" box above highlights how this kind of confusion matrix can be used.

Principle 3: Conducting a representative randomized sample of “unknown” SLs

Generating an estimate of the total number of LSLs in a system or the material at any given address will use information from previously verified SLs to estimate the materials at SLs of unknown material. The accepted best practice in statistics to be able to make these kinds of estimates is gathering verified SL data at a random set of homes where the SL material is unknown. Statistically, only such a representative set of verified service points will truly reflect the whole system.

The representative randomized sample is critical for understanding the entire system’s likely materials, which is especially useful for setting and requesting budgets. Combining the results of this inspection set with the characteristics of those addresses (e.g., age of home, neighborhood, water testing, etc.) allows a water system to calculate the probability of finding an LSL at other SLs with unknown materials. This is the definitive way of estimating or predicting the number of SLs with lead and any other type of materials of interest, as well as for understanding which areas are more likely to have LSLs.

As opposed to the representative set described above, a water system using a non-representative set of verified SLs may estimate that there are substantially fewer (or more) LSLs than there really are. The non-representative data comes about in many ways, such as only verifying SL materials where water mains have recently broken, where other road work has happened, or where new construction has just occurred. Since the proportion of LSLs found at these service points may differ from the remaining not-yet-verified service points with unknown materials, this data can be misleading. Using a non-representative set can skew the numbers and is not the best approach to estimate SL materials at any scale (i.e., across the whole system, area-by-area level within the system, or home-by-home). The non-representative data is still informative and will be used, but should not be used to extrapolate across the entire system.

A representative randomized set of SLs is one in which each SL in the water system that is categorized as “unknown” shall have an equal chance of being selected to be physically verified. Therefore, the overall characteristics of the representative set would have similar characteristics to other homes with unknown SL material (e.g., age of home, tax value, etc.). For example, if 30% of SLs of unknown material connect to homes built before 1940, then approximately 30% of the homes in the representative random set would be built before 1940.

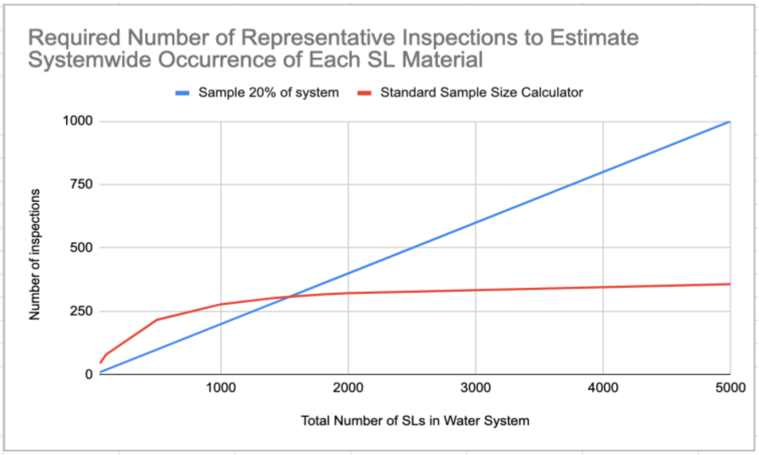
For each SL in the representative set, the materials of all SL portions should be verified via physical inspection (i.e., all segments of the SL and where the SL enters the home). In water systems where goosenecks or connectors have been used, verifying the material at the connector would also be necessary to characterize the full SL material. The resulting dataset of materials will reflect the entire water system. The proportion of LSLs in the sample should be, with high probability, within the margin of error of the true proportion of LSLs in the population. A sample size calculator can be used to calculate that number.

The graph below shows the output of a sample size calculator for the minimum required number of inspections based on the number of total “unknown” SLs in the water system.[1] Note that this is a statistical calculation to specifically assess the total number of LSLs in a system, which -- if using a randomized representative sample -- can be done with a relatively small number of inspections even for

[1] The sample size is based on having an estimate with a 5% margin of error and a 95% confidence interval for the proportion of LSLs within the population of SLs of unknown material.

larger systems. For water systems with more than 1,000 “unknown” SLs, the total sample size would not exceed 400 inspections.[2] For systems with fewer than 1,000 “unknown” SLs, the minimum required number of inspections may be more difficult to achieve, but asking for a certain percentage of their system to be verified could be an alternative (e.g., 20% of SLs for systems with fewer than 1,000 “unknown” SLs). Estimating the distribution of SLs across an area requires a slightly different calculation, but is still based on the principles of representativeness described above.

Figure 2: Modeling Showing the Necessary Number of Randomized Inspections Based on Size of the System



This graph shows the difference between the number of inspections based on a sample size calculator versus a flat 20 percent of the system. The appropriate proportion of SLs that need to be inspected for a statistically valid estimate of the number of LSLs flattens out once the system size is larger than about 1,500 SLs of unknown material. It might be challenging for smaller water systems to inspect at the recommended amount, but inspecting at a percentage of the SLs of unknown material (e.g., 20 percent) could provide valuable data to drive decision-making.

Estimating the number of LSLs through a representative set like the one described in this section is a crucial tool in planning replacement programs, especially for setting and requesting budgets. Such an understanding is also best for communicating to the public because it is easy to communicate that every SL of unknown material had an equal chance of being chosen for the sample inspections rather than inspections being construed as some sort of area-specific bias. Further, any model predictions of likelihood of lead are most accurate and reliable when they are based on data that come from a representative randomized inspection of a set of SLs.

Attention water administrators:
As water utilities request funding to replace their LSLs, the results of a representative sample can help develop accurate budgets for efficient replacement programs.

Attention water utilities:
Verified materials of a representative random sample of previously “unknown” SLs provide an accurate estimate of the concentration of LSLs — and is an essential input to more advanced modeling.

[2] Water systems seeking to use this process to estimate the distribution of LSLs between and across neighborhoods would use a modified sample size calculator that would take into account the water system’s geography and predicted proportion of lead. This is called a “multi-level sample size.”

Lessons from Flint:

Initially, Flint's historical records suggested 10-20 percent of the city's SLs contained lead. But after a representative set of inspections at 231 addresses, the resulting statistical analysis indicated that 52 percent of all parcels had lead in the inventory.[3] After over 25,000 excavations were completed, the estimated percentage of all Flint parcels with an LSL is now 51 percent. When narrowed down to include only active water accounts, a simulated representative sample estimates that 38.7 percent of active accounts would have an LSL. The SL replacement program has found LSLs at 37.2 percent of active water accounts. The estimates from the representative sample allowed the city to plan for, and request, the appropriate funding to remediate the problem.

Principle 4: Transparency

Whether a water system is using a statistical model to predict SL material or another approach, it is important that water utilities be transparent about their methods and results. Transparency about methods is important for communication with regulators and customers. Considering the public health and asset decisions that these inventories could support, being upfront about the steps used to develop a prediction is important. This also aligns with best practices for statistics and research, where organizations and researchers communicate their methods before they describe their results.

Public communication

From a public communication perspective, predictive modeling enables greater transparency by communicating the relative likelihood of having an LSL. Instead of categorizing pipes as just "known lead," "known non-lead," and "unknown," the use of a predictive model allows utilities to characterize the likelihood of any SL containing lead. The information about the likelihood can be communicated in a range of ways based on the needs of the utility, local authorities, or residents. This could take the form of a category based on a range of probabilities (e.g., "Likely LSL" as anything over 50%). The visual communication of this information is also important as it should be presented in an easy-to-use mapping format for the public. This kind of information can be included on maps as ways to foster trust and effectively communicate public health information.

This can reduce the potential panic that could occur when someone is told the material at their address is "unknown." By communicating a likelihood of lead based on the statistical methods addressed above, residents will be able to get a sense for their relative lead risk, along with information about how to take action. Some states already have enhanced transparency requirements: Ohio requires that each system submit a map of SLs and buildings likely to contain lead to the state Environmental Protection Agency;[4] Michigan requires each system to submit a Distribution System Materials Inventory;[5] Illinois and Wisconsin publish publicly accessible annual reporting data by their water systems; California requires community water systems to annually report SL inventory information and include a replacement

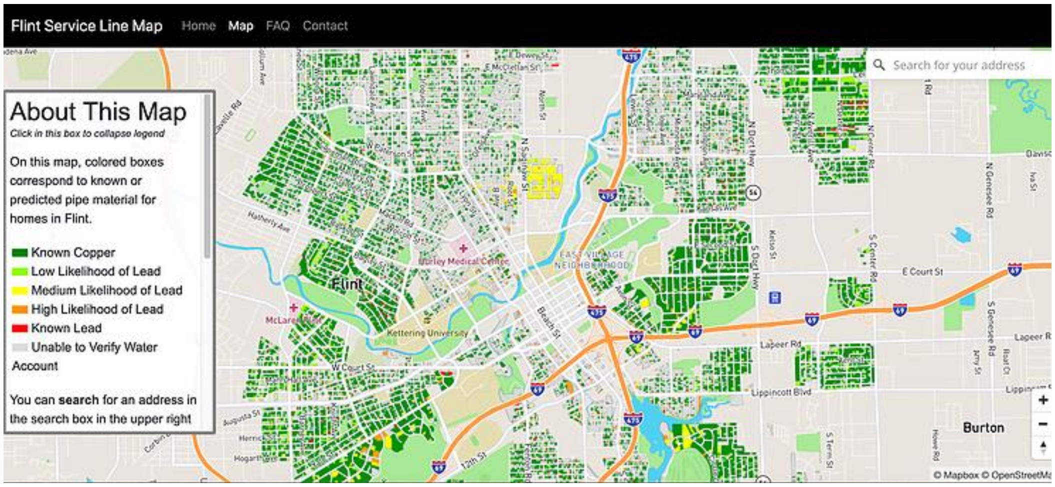
[3] Letter to DNR Creagh from City General McDaniel dated November 1, 2016 re: estimated number of service lines needing replacement. <https://www.michigan.gov/documents/flintwater/Letter_to_DNR_Creagh_from_Flint_McDaniel_dated_110116_545761_7.pdf>

[4] Ohio EPA, "Lead and Copper in Public Water Systems," https://epa.ohio.gov/ddagw/pws/leadandcopper#185385289-lead-service-lines-and-mapping
EDF, "State efforts to support LSL replacement," https://www.edf.org/health/state-efforts-support-lsl-replacement#ohio

[5] Michigan EGLE, "Michigan Service Line Materials Estimates Preliminary Distribution System Materials Inventories," https://www.michigan.gov/documents/egle/egle-dwehd-PDSMISummaryData_682673_7.pdf

schedule for any SL that contains lead or is categorized as "unknown".[6] Public knowledge of the material of residents' own SL materials and any health-protective actions needed are necessary for the success of SL inventory and replacement programs. The Lead Service Line Replacement Collaborative has done a lot of work around communication of LSLs, and they have many examples from communities across the country.

Figure 3: The Flint Service Line Map



Flint Service Line Map Home Map FAQ Contact

Search for your address

About This Map

Click in this box to collapse legend

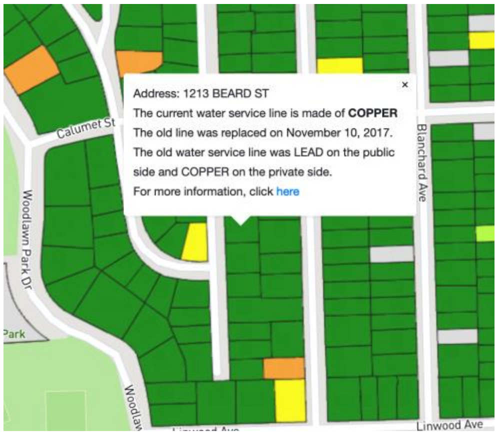
On this map, colored boxes correspond to known or predicted pipe material for homes in Flint.

- 🟩 Known Copper
- 🟩 Low Likelihood of Lead
- 🟨 Medium Likelihood of Lead
- 🟧 High Likelihood of Lead
- 🟥 Known Lead
- ⬜ Unable to Verify Water Account

You can **search** for an address in the search box in the upper right

The Flint Service Line Map is an interactive visual representation of the LSL replacement progress in Flint, Michigan. It is address-searchable and shows if/when an SL has been replaced as well as the known or estimated SL material. Figure 4 shows what a close-up view of the address-specific information provided.

Figure 4: A Close-Up Example of the Flint Service Line Map



Address: 1213 BEARD ST

The current water service line is made of **COPPER**

The old line was replaced on November 10, 2017.

The old water service line was LEAD on the public side and COPPER on the private side.

For more information, click here

This example shows the parcel-specific information provided by the Flint Service Line Map when a user clicks on or searches for a certain address. The Map also provides additional information regarding lead exposure risk and mitigation techniques.

[6] California Water Boards, "Lead Service Line Inventory Requirement for Public Water Systems," https://www.waterboards.ca.gov/drinking_water/certlic/drinkingwater/lead_service_line_inventory_pws.html

Reproducibility

From a regulatory perspective, it is important that water systems communicate the method that they used to calculate their predictions. Utilities can document and publicly disclose methods, assumptions, and data so that others can replicate the analysis and see how the results were obtained, creating opportunities for critical evaluation of the calculations, as well as disseminating methods and lessons learned. This level of transparency is important for regulators and public accountability about how water systems made determinations for their inventories.

Per the proposed Lead and Copper Rule Revisions, SLs listed as "unknown" in the initial inventory or the updated inventory must be counted as LSLs for purposes of calculating LSL replacement rates and for issuing targeted public education to consumers with an LSL or SL of unknown material. What is defined as an LSL and an unknown for inventories submitted to regulators is a policy decision with enormous potential impacts since replacement rates are based on the number of lead and "unknown" SLs. More details may be included in the final rule; if not, this will have to be solved through EPA guidance or at the state level. A statistical model does not replace the need for physical verifications, but can be used to inform SL inventory and replacement programs.

Attention water administrators:

An administrator could encourage or require the utility to submit a Public Communication and Community Engagement Plan to demonstrate their transparency. Possible criteria for an adequate plan could include a website featuring a map with SL inventory details for each address and a downloadable spreadsheet of underlying data. Another criteria could deal with summaries of the program, including overall counts of each type of SL material based on the inventory and an updating figure or table showing replacement progress over time. Administrators can integrate best practices in public health communication to better communicate uncertainty related to SL material. The results of the predictive model can support that communication. From a reproducibility perspective, the utility should be able to provide to the administrator (and potentially make publicly available online) the data and logic used to generate the analyses yielding the inventory, in a way that is clear and usable to other users.

Attention water utilities:

Communicating the SL inventory and the progress of replacements to the public is essential in maintaining trust and confidence of the community. But deeper engagement with the community around those issues will require continued transparency. Writing reproducible statistical analyses for SL inventories is important for future use and for the regulator and public to trust the inventory.

Lessons from Flint:

The researchers detailed their methods in reports to regulators and also published their work in peer-reviewed journals to ensure reproducibility and transparency. As the city of Flint worked to complete its pipe replacement program in 2020, BlueConduit collaborated with NRDC to release the Flint Service Line Map, an interactive tool that allows residents, policy makers, and advocacy groups to examine residential SL materials and predictions in Flint (see Figures 3 and 4 above). This map aims to empower residents with information about the status of the pipe replacement project at their address, what they can do to get their line confirmed and/or replaced, and steps they can take to reduce lead exposure.

Principle 5: Hold-out sample accuracy

When using a statistical model, it is important to continually evaluate model performance at every stage of model development and implementation. This is done through the use of a hold-out sample. A hold-out sample refers to withholding a random portion of a data set from an initial model and then using the withheld data to assess the statistical model's performance. This is a very common and important practice in statistical modeling and machine learning. It creates a high bar to measure a model's performance and provides evidence that the model's output could be trusted and used for decision-making. This principle can be applied with data from field work and in the building of the model.

Evaluating a model with true hold-out data from the field can be done simply by utilities. A utility would be able to make the following statement (with evidence): "As of May 31, 2020 based on the predictive model, we predicted about 65% of targeted SLs contained lead in some portion. And by August 31, 2020, after three months of inspections and replacements, we found 67% had lead." This would be evidence of accurate hold-out performance since the prediction closely matched what was then found in the field.

In addition to showing the performance using true hold-out data in the field, it is also best practice to use existing data as hold-out data while building the predictive model. Before a model is used to inform decisions, that model could be checked repeatedly by taking a subset of the existing data to be held out from the building of the model in order to use that data to see how well the model performs. This could mean that 25% of a dataset is not used in the construction of the model and using that subset to evaluate a model that is built on the remaining 75% of the data. Utilities could perform such checks of hold-out sample accuracy for different assumptions going into a model or for different predictive model techniques.

On such a hold-out sample, a utility could also show that its predictions are credible and well-calibrated. Demonstrating this involves the following steps: find all SLs that were predicted to have approximately 90% of lead by the model, and then for those SLs, calculate the percentage of them that truly did have lead. The calculated percentage from a well-calibrated model will be close to 90%.

Aside from making sure that model probabilities are well-calibrated, it is important to define the accuracy measures used to evaluate and monitor model performance. The key metric to be used for in-the-field true hold-out evaluation is "Hit Rate," the number of LSLs that were identified divided by the number of attempted replacements regardless of what was discovered. Hit rate can be computed for an entire region or broken down into a specific geography or time. It can also apply to any replacement project, whether or not it uses a predictive model.

Appendix 1 contains an in-depth discussion of different methods of analyzing model performance, including metrics to use when the presence of LSLs is extremely high (>90%) or extremely low (<10%) where hit rate is not the ideal metric.

Attention water administrators:

As utilities begin to rely more on predictive methods, they should be held to the standards typically used to evaluate those models, namely hold-out accuracy. Utilities using models could be required to submit predictions for all addresses with unknown SL materials on a particular date before beginning work in the field on those addresses. And then at a regular reporting interval (e.g., annually), they could be required to compare what they found to their initial predictions. Additionally, more technical reports could demonstrate hold-out accuracy as well.

Attention water utilities:

When using predictive models, the best way to ensure that they are working best for you -- and to convince others to trust them -- is to demonstrate accuracy and other performance measures on hold-out data of SLs not used to estimate the model.

Lessons from Flint:

The researchers provided extensive evidence that the predictive models were accurate and well calibrated by using hold-out data when creating their models and evaluating the model's performance in the field. They used hold-out accuracy to evaluate which of a set of possible predictive methods were most effective at predicting SL materials in Flint. To start they used straightforward statistical modeling approaches, such as regression-based models for classifying whether an SL had lead. While those methods yielded outputs that were easier to interpret, the accuracy on hold-out data was consistently poor. The use of machine learning methods proved to be more flexible in combining that same information into more accurate predictions of SL material. The house-by-house predicted likelihoods were calibrated to be close to the true percentage of leads. This was done using existing data collected, and then it was performed using true hold-out data from the field. The predictions, once demonstrated to be valid, have been used for prioritizing which neighborhoods, blocks, and specific SLs to service next. In the field, the researchers' predictions helped guide the LSL replacement program in 2016-2017. Of the 8,833 SLs dug up for replacement during that time, 6,228 contained lead or galvanized material. Some months, they achieved a 90% hit rate.

For a more detailed explanation of the specific data science methods used in Flint, please see this paper, which won an award for best applied data science paper at the 2018 KDD, or this paper from the Bloomberg Data for Good Exchange.

Conclusion

Having an accurate picture of the number and location of LSLs in a water system benefits all steps of the LSL inventory process, from budgeting to excavations. Following these guiding principles can help water systems plan for and efficiently execute their LSL replacement programs:

1. Clean data management and organization
2. Not accepting all historical records as truth
3. Conducting a representative randomized sample of service lines
4. Transparency in public outreach and reproducibility
5. Accuracy on held-out sample

Using the likelihood of an SL having lead, as predicted by a statistical model, is an important input to LSL inventory and replacement decisions, but is not the only criteria. Utilities weigh several factors when making these choices, including concerns about equity and logistical constraints. Having a statistically robust probability combined with other factors can improve the overall decision-making. LSL inventory and replacement projects can last for many years. Applying these principles can also prepare utilities for the multiple iterations or phases of these long-term replacement projects. The risk of not applying any or all of the previously outlined principles raises the potential of misinformed or biased decision-making, leading to potential delays or inefficiencies in mitigating lead exposure.

Moreover, water administrators can feel more confident in the accountability and transparency of a water system's process when submitting their LSL inventories, as these principles make clear the methods and assumptions of the analysis. They can also be used by regulators in training, guidance, and rulemaking to encourage water systems to use the most accurate and representative data to inform decisions, set budgets, plan strategically, and, most importantly, protect the health of all individuals in the system.

Some of these principles, such as good data management and transparency, could be considered as a necessary baseline for any inventory or replacement program, while some of the more complex statistical methods can be utilized for higher levels of sophistication and accuracy. Using a data-guided approach should be encouraging rather than daunting-- some systems may feel comfortable implementing this approach on their own, but there are also myriad organizations and companies ready to implement these principles on behalf of water systems across the country to help them more efficiently identify and replace LSLs.

If you have questions, contact Ian Robinson at ian@blueconduit.com or Wendi Wilkes at wwilkes@asdwa.org

Appendix 1

As mentioned above, these predictive models generate probabilities from 0 to 1 of the likelihood of an individual SL containing lead. They do not classify SLs as “lead” or “non-lead.” One consequence of this is that a model really generates a list of all SLs with unknown material and orders them in terms of likelihood of containing lead. What a water system chooses to do with that rank-ordered list of likelihoods of lead is a decision in the hands of the water system and not automatically determined by the model itself.

When a predictive model is used to help inform an SL replacement program, hit rate represents the combined success rate of both the model and the decision-maker using the model. It also captures the efficiency of the LSL replacement efforts, making it easy for a utility to calculate its cost per successful SL replacement.

Since statistical models generate a list of probabilities and the performance metrics require a classification, a cut-off must be chosen that divides the SLs into “most likely lead” and “most likely non-lead” categories. It’s important to underscore that the cut-off determination is only used to evaluate model performance and does not attempt to classify any individual SL as “lead” or “non-lead” for inventory purposes.

The standard measures of performance all can be visualized and computed from this 2x2 “confusion matrix.”

	Truly lead	Truly non-lead
Most likely label: “lead”	(A)	(B)
Most likely label: “non-lead”	(C)	(D)

The cutoff used to split between the “most likely” categories is usually the proportion of lead in the dataset (e.g. if 30% of SLs are estimates to be lead, anything with a probability above 30% would be considered “most likely lead”)

In the table, (A) is the number of cases where the model predicted the most likely label for an SL to be lead, and in fact, that SL is lead. These would be the count of “true positives.” (B) represents the number of “false positives” when the model’s label for an SL is lead, but the actual SL material is not lead. Similarly for (C), “false negatives,” the model labels the SL non-lead but the verification shows it is actually lead. And (D), “true negatives,” the model’s label for the SL is non-lead, and that SL is truly not lead.

Many different names are used to describe different percentages associated with these four counts, and we try to clarify those here. False positive rate is the percent of SLs given the label of lead but were found to not have any lead:

$$\frac{(B)}{(A)\ +\ (B)}$$

A model with a high false positive rate would have a lower hit rate and an increased average cost of LSL replacement to reflect the inefficiency in replacement attempts.

False negative rate represents the most important health challenge and the case to eliminate entirely in the long run because these are LSLs that are potentially being missed or non-prioritized for replacement:

$$\frac{(C)}{(C)\ +\ (D)}$$

This metric can be useful for model evaluation, but in isolation of the physical reality of the problem, does not reflect the nature of the decision-making based on the model over time. In practice, this corresponds to the proportion of all SLs in a community that were not targeted for replacement in that phase of the work (e.g., month, quarter, year), but, in fact, those SLs do have lead. As the program progresses, some of those SLs may be targeted for replacement. But at this time there are too many other SLs with higher likelihood of lead prioritized ahead. As discussed above, the use of the probabilities to help set a prioritization is distinct from deeming all SLs with low likelihood of lead to necessarily have no lead at all.

Finally, we note that even these metrics have limitations, especially if the water system has a fairly low (or fairly high) proportion of its service area containing LSLs. Some of the above measures of accuracy are not as helpful when the problem of finding LSLs resembles finding a "needle in a haystack." For example, if only 1% of all SLs are truly lead, then a simple model that predicts every SL is copper is "99% accurate" while not providing anything useful. In such a case, an additional pair of measures calculated from the confusion matrix, called "precision" and "recall," can be applied.

Precision is equivalent to true positive rate above, reflecting the ability to be correct when giving an SL the label of lead. It is expressed it as:

$$\frac{(A)}{(A)\ +\ (B)}$$

Recall is the model's ability to find LSLs among all of the true lead SLs in the data, including those in the "likely lead" and the "likely non-lead" groups.

$$\frac{(A)}{(A)\ +\ (C)}$$

This is a useful metric because even if only 1% of all SLs are LSLs, it evaluates how many of those true LSLs did the model correctly label as lead. Precision and Recall are scores between 0 and 1, with 0 indicating poor performance and 1 indicating perfect performance.

As we mention above, there are many reasons why it can be undesirable to choose a hard cutoff in likelihoods to apply it to all SLs. But choosing that cutoff can be avoided with a more advanced approach of evaluating the models. Those advanced methods are able to consider all possible cutoff points at once, evaluating the entire model-based rank ordering of all SLs. While beyond the scope of this white paper, such model evaluation methods are known as the Receiver-Operating-Characteristic Curve (for false positives and false negatives) and the Precision-Recall Curves. We leave these to the reader to investigate further.