How EPA Researchers Use Predictive Modeling for Lead Service Line Identification

United States Environmental Protection Agency

About the Project

This project aims to provide a scientifically grounded overview of predictive modeling methods, including outlining their capabilities, data requirements, and suitability for different community water systems.

EPA researchers are seeking to partner with communities, which will strengthen the project and help resulting methods to align with current needs and priorities. Insights into the unique needs of different communities will help researchers tailor solutions.

Interested in Partnership Opportunities?

- Helia Seifikar (seifikar.helia@epa.gov)
- Caleb Buahin (buahin.caleb@epa.gov)
- Emily Walpole (walpole.emily@epa.gov)
- Brian Dyson (dyson.brian@epa.gov)

Identifying Lead Service Lines (LSLs)

Service line inventories are often incomplete and may contain inaccuracies, making the mandated identification and replacement of lead service lines (LSLs) challenging and time-consuming.

Predictive modeling can be a complementary tool to expedite the identification process by using available data to identify where LSLs are most likely to be present. By identifying service lines with a higher probability of containing lead, communities can prioritize remediation efforts more efficiently.

What are predictive models?

Predictive and statistical models, such as machine learning decision tree and logistic regressions, attempt to identify patterns in data, such as housing age and community profile. These models then make predictions based on the data. Different methods use different strategies to achieve this goal.

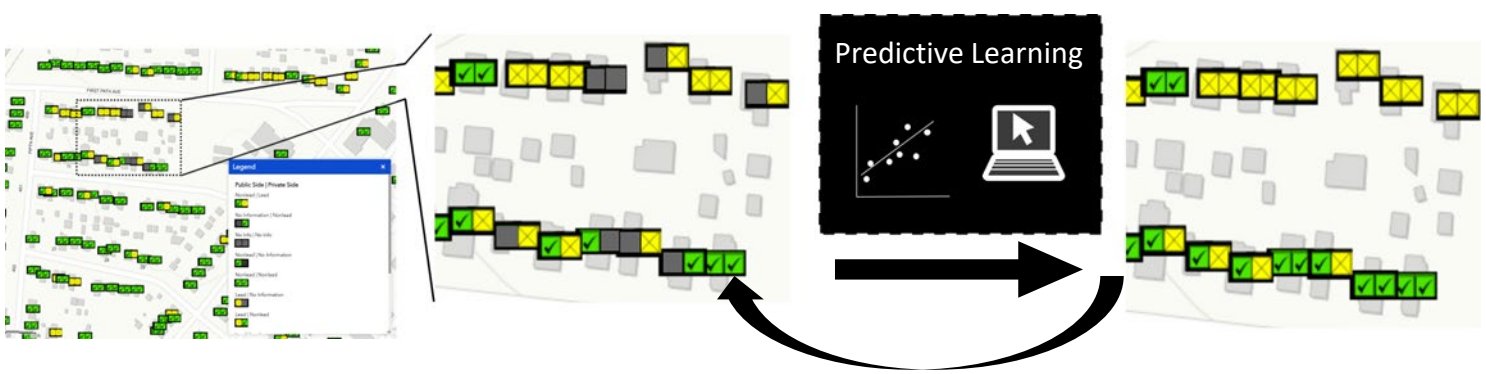
How can predictive models help with LSL identification and replacement?

Predictive models estimate where LSLs are most likely to be found, helping improve service line inventories and replacement efforts. For example, areas with a higher chance of having LSLs might use more thorough or intensive testing methods such as mechanical excavation, while areas with a lower chance might use simpler approaches such as visual inspection. Data used, likelihood thresholds, and model methods vary and should be evaluated to suit local needs.

The diagram below shows an incomplete inventory map, which is then updated with model-predicted outcomes. In the updated map, uncertain locations (gray) are replaced with low (yellow) and high (green) probabilities of lead occurrence. The updated map provides new data that can then be used as new input to run the models as part of an iterative process.

Map with Incomplete Inventory

Map with New Information Gained from the Outcome of Models



See Page 2 to learn about how predictive models are developed and evaluated.

Disclaimer: This document is for informational purposes only. EPA does not endorse one lead service line identification technique over another.

March 2025

How Predictive Models for LSL Identification Are Developed and Evaluated

United States Environmental Protection Agency

Input Data

Information — such as home age, property tax, income, geospatial data, prior testing results, and available LSL inventories — is put into the model.

Predictive Model Development

The models identify patterns in data to make predictions. Machine-learning models use algorithms for this purpose: in the **training phase** algorithms learn from a dataset and then adjust the model's internal parameters to minimize error in a **validation phase**. Lastly, **testing** the model on new, unseen data ensures it can effectively predict outcomes.

Evaluation

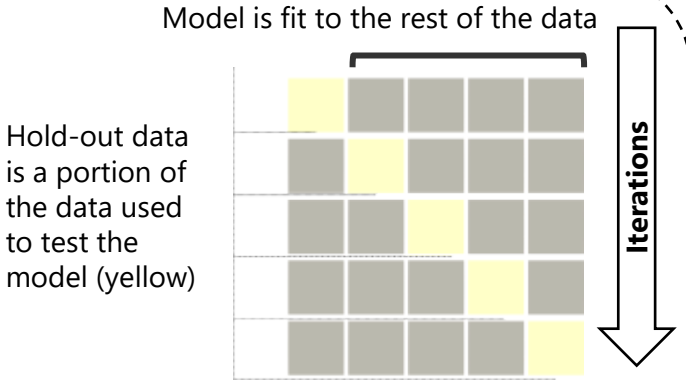
Models can be evaluated in several ways. Some common practices are **cross-validation** and using a **confusion matrix** (see examples below).

Predictive Model Utilization

Inspecting service lines with high probability of lead occurrence and updating inventories.

Updated inventories can be used as inputs in new model iterations.

Examples of Model Evaluation



Cross-validation
Data is partitioned in subsets. During each iteration, a hold-out dataset is used to assess how well the model performs on unseen data.

True positive How many lead service lines the model correctly identifies as lead.	**False negative** How many lead service lines the model misses.
False positive How many service lines the model incorrectly identifies as lead.	**True negative** How many non-lead service lines the model correctly identifies as non-lead.

Confusion Matrix
Metrics such as recall, precision, and accuracy are calculated to measure performance.

www.epa.gov/water-research/drinking-water-technical-assistance-support-bipartisan-infrastructure-law

Disclaimer: This document is for informational purposes only. EPA does not endorse one lead service line identification technique over another.

March 2025