Lead and Copper Rule: Classification of Unknown Service Line Materials

Document no: 1
Version: 1

City of Newport: Department of Utilities (Newport Water Division)
Rob Schultz

Lead Service Replacement Program
January 23, 2025

Lead and Copper Rule: Classification of Unknown Service Line Materials

Client name:	City of Newport: Department of Utilities (Newport Water Division)		
Project name:	Lead Service Replacement Program		
Client reference:	Rob Schultz	**Project no:**	22-012-4
Document no:	1	**Project manager:**	Ed Roworth
Version:	1	**Prepared by:**	Joshua Registe, Tugba Akgun, Meredith Sullivan, Ed Roworth
Date:	January 23, 2025	**File name:**	
Document status:	Draft		

Document history and status

Version	Date	Description	Author	Checked	Reviewed	Approved
1.0	1/17/2025	Draft Results	Ed Roworth		Peter Von Zweck	Ed Roworth

Distribution of copies

Version	Issue approved	Date issued	Issued to	Comments
1.0		1/17/2025	Rob Schultz	

Jacobs Engineering Group Inc.

120 St. James Avenue
5th Floor
Boston, MA 02116
United States

T +1.617.242.9222
F +1.617.242.9824
www.jacobs.com

Contents

Appendices

Tables

Figures

Acronyms and abbreviations

CFR	Code of Federal Regulations
EPA	U.S. Environmental Protection Agency
Jacobs	Jacobs Engineering Group Inc.
LCR	Lead and Copper Rule
LCRR	Lead and Copper Rule Revisions
LSL	Lead Service Line
NWD	Newport Water Division
NA	Not Applicable
QA/QC	Quality Assurance/Quality Control
RIDOH	Rhode Island Department of Health
GIS	Geographic Information System
ML	Machine learning
GRR	Galvanized Requiring Replacement
LPPA	RI Lead Poisoning Prevention Act
LCRI	EPA Lead Copper Rule Improvements

Definitions

Geographical Information System. A conceptualized framework that provides the ability to capture and analyze spatial and geographic data.

Lead and Copper Rule. A subsection of the CFR that provides rules and regulations related to lead and copper levels in drinking water.

Lead Service Line. A water service line or portion that connects the water main to the building inlet. 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. Note: For the purposes of the 40 CFR §141.86(a) only, a galvanized service line is not considered a lead service line.

Project Team The City of Newport Water Division and Jacobs Engineering Group staff members.

Water System. Serves at least fifteen service connections used by year-round residents of the area served by the system or regularly serves at least twenty-five year-round residents.

1. Executive Summary

This document outlines the Newport Water Divisions efforts to comply with the U.S. Environmental Protection Agency's (EPA) **Lead and Copper Rule Improvements (LCRI)** and Rhode Island's **Lead Poisoning Prevention Act (LPPA)**. These regulations aim to ensure that public drinking water is free from dangerous levels of lead and other harmful materials. The Newport Water Division (NWD) is tasked with identifying public and private service lines and replacing lead service lines within its water distribution system to meet these critical public health goals.

1. Purpose of This Document

The primary goal of this report is to explain how statistical modeling, specifically a machine learning technique called **Random Forest Classification**, is being used to:

- **Identify Unknown Service Line Materials**:
    - Many unknown water service lines (6070)—the pipes that connect water mains to homes or buildings—have unknown materials on the City (public side) or Private-side of the curbstop. Determining whether these pipes are made of copper, lead, or other materials is essential for prioritizing replacements.
- **Streamline Compliance:**
    - Predictive modeling helps estimate which service lines are likely made of copper, reducing the need for costly and disruptive physical inspections. By focusing on the most at-risk lines, this approach saves time and resources while ensuring compliance with EPA and state requirements.
- **Prioritize Public Health:**
    - By accurately predicting where copper service lines are located, NWD can prioritize the investigation and removal of known or suspected lead, reducing potential exposure to lead in drinking water faster.

2. How It Works

NWD is leveraging historical data and modern machine learning techniques to identify service line materials. Key information used in the analysis includes:

- The year the service line, water main, or building was installed.
- Known material types on either the city or private side of the service line.
- Geographic data and inspection records.

Using this data, the **Random Forest Classifier** can predict the likelihood of service lines being made of copper, lead, or other materials, however; NWD is solely focusing on the data outputs classifying copper material. This method allows NWD to classify previously unknown lines with a high degree of accuracy (>96% accuracy), focusing efforts on areas most likely to contain lead.

3. Benefits of Statistical Modeling

- **Cost Savings**:
    - Modeling reduces the need for extensive physical inspections of copper service lines saving the city an estimated **$3.57 million**.
- **Time Efficiency:**

  - This approach accelerates compliance efforts by as much as 3 years.

- **Improved Public Health**:
  - By prioritizing high-risk areas, the city ensures safer drinking water for residents more quickly.

2. Statistical Modeling and Application for Service Line Inventories

Statistical analysis and predictive modeling are increasingly being used by water systems to estimate the likelihood of lead service line (LSL) occurrence without the need for physically excavating every buried service line connection. This approach, acknowledged by the US Environmental Protection Agency (EPA) as an effective service line material identification method, leverages historical data such as existing records, building age, geographic information, and other distribution system characteristics to predict the likelihood presence of LSLs, reducing the time and cost of compliance efforts. The EPA recommends leveraging predictive modeling to estimate the probability that a service line is lead, prioritize areas for service line investigations, and to expedite lead service line replacements (LSLR) (USEPA, 2022).

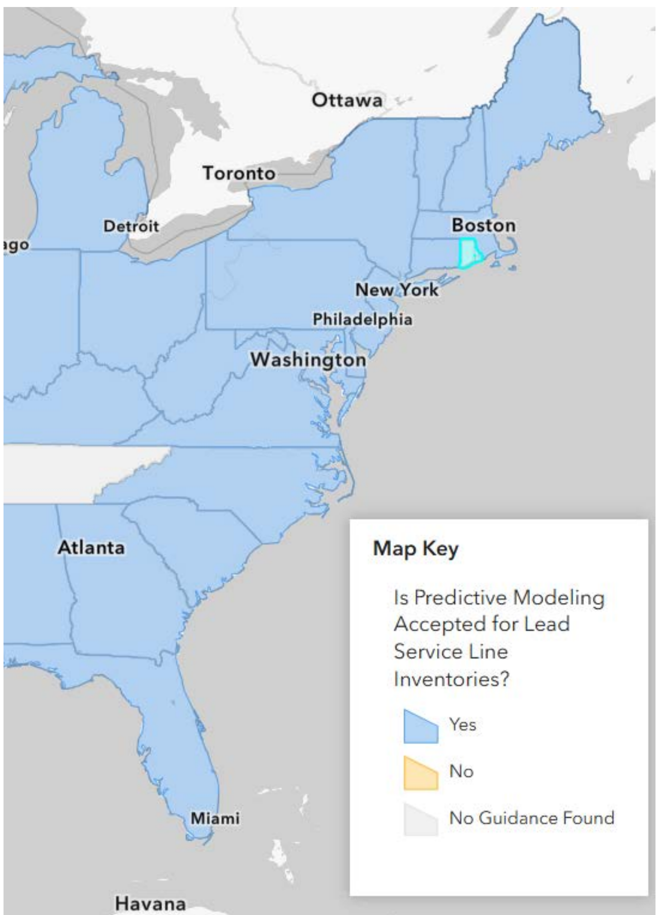


Figure 1. Map of States Accepting Predictive Modeling as a Method. *Source:[https://blueconduit.com/post/state -guidance-and-predictive-modeling](https://blueconduit.com/post/state-guidance-and-predictive-modeling)*

While the EPA recognizes this as a viable strategy, the decision to adopt it is left to individual states. Many state regulators across the country have approved or issued guidance on utilizing statistical analysis and predictive modeling for lead detection, streamlining compliance with the Lead and Copper Rule Revisions (LCRR). **Figure 1** shows neighboring states such as Connecticut, Massachusetts, and other States that have embraced this method and issued guidance on its use. These states typically require systems to submit documentation demonstrating that their predictive model is representative of the system. Additionally, they request a plan to conduct an agreed-upon number of field inspections to validate the model's accuracy and establish a minimum level of confidence, often set at 95% or higher (the industry standard in scientific research). This ensures the reliability of the predictive model while balancing compliance costs. However, Rhode Island is an outlier in the Northeast and mid-Atlantic regions, as it has not approved the use of statistical analysis or predictive modeling, despite the detection of LSLs within the state's initial inventory submissions.

In the absence of state-specific guidance, this report will outline the development and application of predictive modeling to meet the core principles and best practices from regulatory agencies that have issued guidance. Additionally, our approach will remain adaptable, allowing for adjustments should state-specific guidance from the RIDOH emerge in the future.

2.1 Lead Service Line Inventory

The Newport Water Division (NWD) Lead Service Line Inventory aims to address the U.S. EPA's Lead and Copper Rule Revision (LCRR) and the State of Rhode Island's recently amended Lead Poisoning Prevention Act (LPPA), to identify and remove all lead services throughout NWD's system within 10 years. In accordance with the EPA's LCRR, NWD prepared and submitted to RIDOH's 120Water Portal a service line material inventory for all services within its water system. This inventory (**Figure 2**) is based on historical book records review of documented service line installations and replacements, Asset Management field verifications, Meter maintenance, and resident reports. This approach ensured that all potential lead and galvanized requiring replacement (GRR) service lines have been identified and prioritized for identification and replacement through the program.

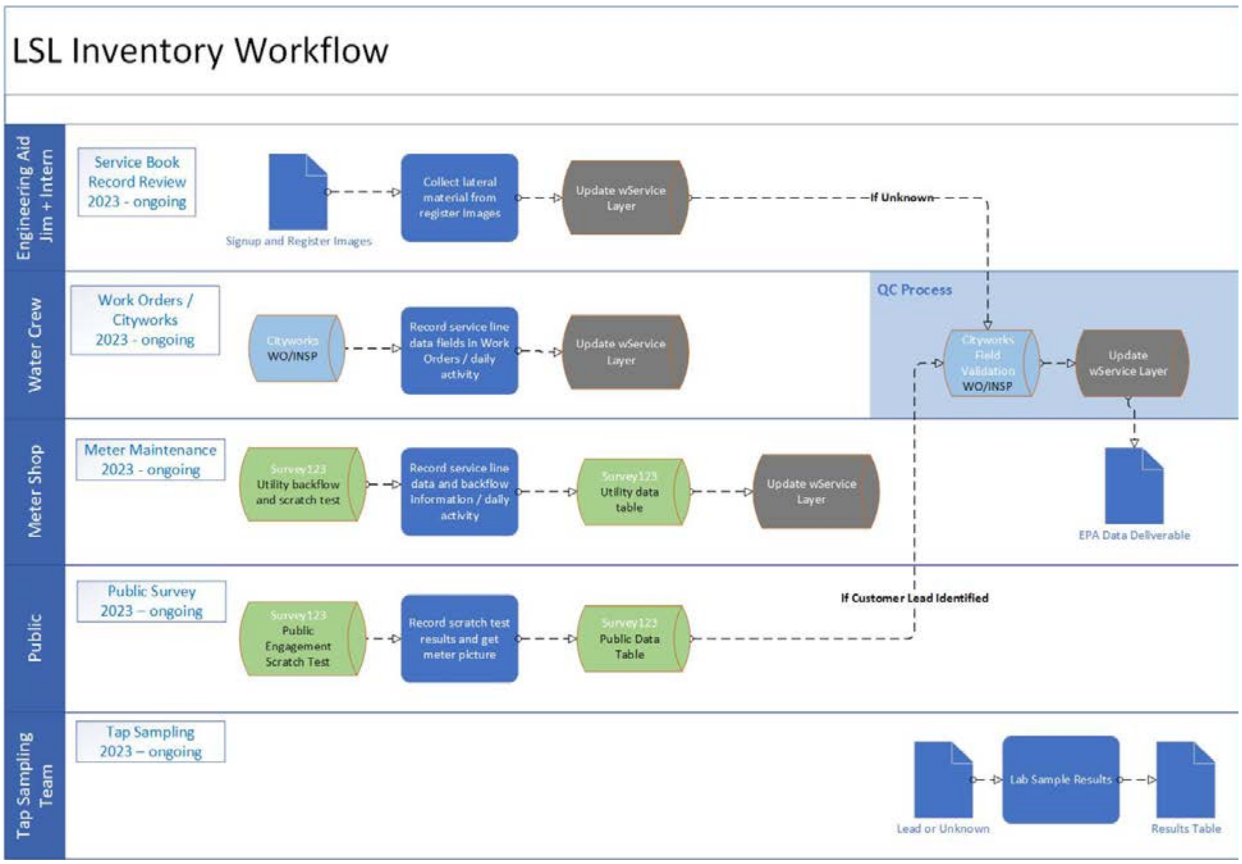
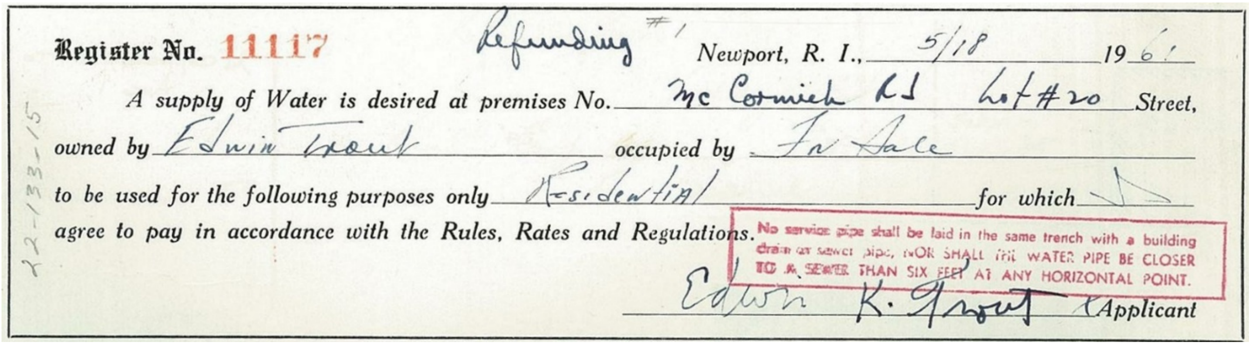


Figure 2. NWD's service line inventory data sources.

As part of the effort to be compliant with the LCCR and LPPA inventory submittal deadline, over 33,000 historical records were reviewed. The original water works started in 1876, and was formally incorporated as the Newport Water Works Company five years later in 1881. The company was succeeded by the Newport Water Corporation in 1929, and since 1936, the City of Newport has owned and operated the system. When a documentation system was eventually established, for over 100 years a paper-based water system service was utilized, detailing the original customers water service installation date (**Figure 3**), and a corresponding field drawing showing dimensions, diameter, and service line materials (**Figure 4**). Service line material would always be shown on the NWD side of the curb stop or property line. However, most drawings did not indicate the material on the private customer side as it's not NWD property. With these and other records, installation dates were gathered for a majority of the services in the system.



Register No. 11117 Refunding #1 Newport, R. I., 5/18 1961

A supply of Water is desired at premises No. Mc Cormick Rd Lot #20 Street,

owned by Edwin Trout occupied by For Sale

to be used for the following purposes only Residential for which

agree to pay in accordance with the Rules, Rates and Regulations.

No service pipe shall be laid in the same trench with a building drain or sewer pipe, NOR SHALL THE WATER PIPE BE CLOSER TO A SEWER THAN SIX FEET AT ANY HORIZONTAL POINT.

Edwin K. Trout Applicant

Figure 3. Original customer sign up for water service.

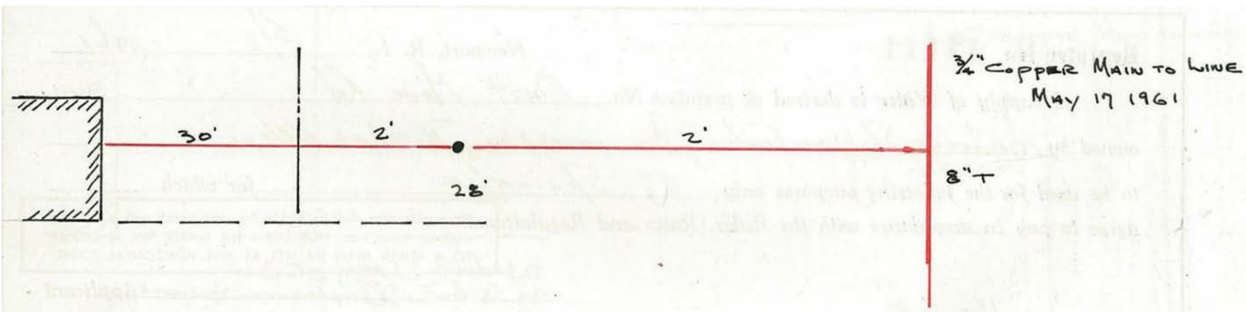


Figure 4. Corresponding drawing for a typical water service.

2.2 Inventory Results

Table 1 presents the outcomes of the service line material inventory, conducted on a comprehensive basis for the entire service line. Classification is irrespective of ownership and in compliance with the regulations established by RIDOH and EPA LCRR. The analysis reveals a substantial percentage of service lines that lack known material composition on one side of the service line.

RIDOH Classification	
Not Lead	5925
Lead	2963
Unknown	6070
Initial Totals	14958

Table 1. Inventory results with RIDOH classification

Material Type	NWD Count	Private Count
Asbestos Cement	77	76
Cast Iron	84	83
Cement Lined Galvanized	445	1237
Copper	9603	5427
Copper or Plastic	29	545
Ductile Iron	162	179
Galvanized	11	199
Lead	1377	392
Other	47	71
Plastic	208	453
TubLoy	5	23
Unknown	2833	6189
Wrought Iron	29	36

Table 2. Actual NWD and Private service line materials.

Table 2 presents the outcomes of the service line material inventory with the actual materials present. The analysis reveals a substantial percentage of service lines on the customers' private side that lack known material composition as would be expected from an ownership perspective.

2.3 Additional Data

NWD Utilizes the Esri GIS to maintain all water system asset data (**Figure 5**). GIS layers in the system include the most recent parcel data and the year the pressure main was installed or replaced. For each service location, the year the structure was built was added from the Newport, Portsmouth, or Middletown, RI parcel data. Similarly, the year the pressure main was installed was added for each service location. The year the structure was built is a good data point for the private side service installation year (**Figure 3**). The year the pressure main was installed or replaced is a good secondary data point for the NWD side lateral date. This additional data along with the original install or signup dates, lateral material, and pipe diameters is used with the machine learning algorithm explained in the next section.

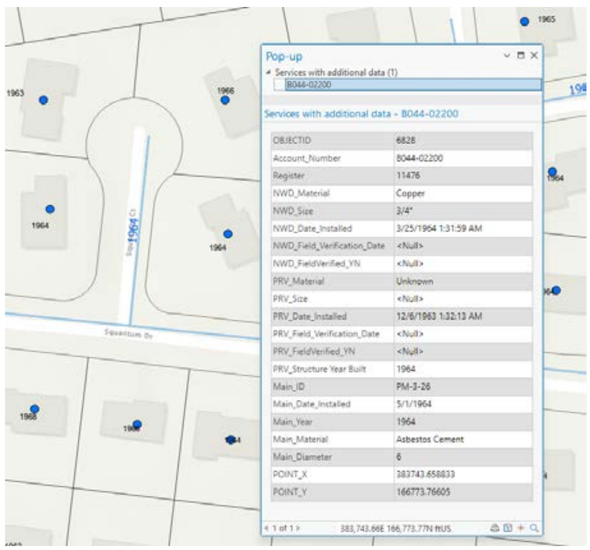


Figure 5. Additional GIS data included for each service

3. Machine learning

3.1 Random Forest Explained

The Random Forest Classifier is a powerful ensemble machine learning algorithm used for classification tasks, including predicting categorical outcomes. It combines multiple decision trees to improve accuracy, reduce overfitting, and provide robust predictions.

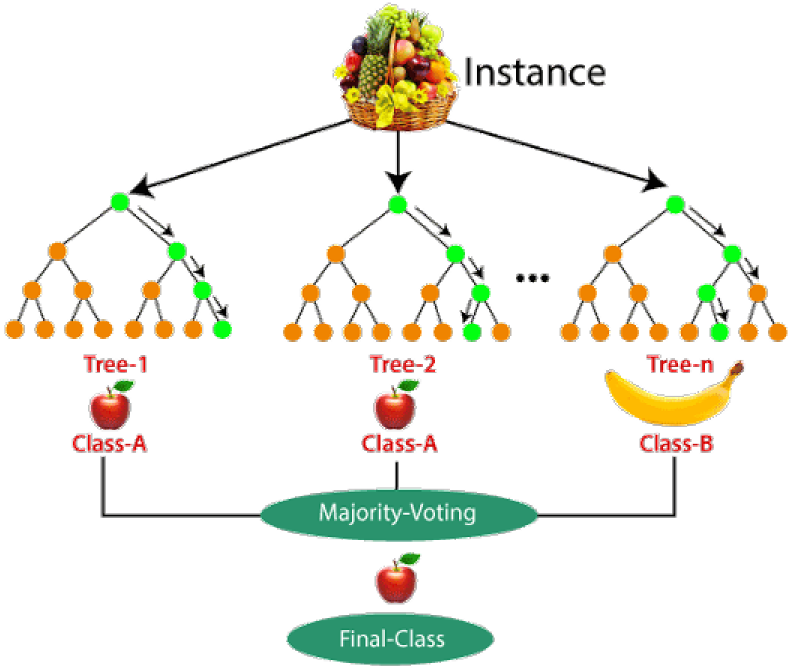


Figure 6. Simplified representation of Random Forest Classifier.

Figure 6 provides a simplified visual representation of a Random Forest Classifier and how it works. Here's a breakdown:

1. Input Data (Instance):
   - The "basket of fruits" represents the input data instance that needs classification (e.g., to determine whether it's an apple or a banana).
2. Decision Trees:
   - Multiple decision trees (Tree-1, Tree-2, ..., Tree-n) are built using different subsets of the data and data rules (e.g., the data shows the fruit is yellow, it could be X fruits).
   - Each tree independently makes a prediction. For example:
     - Tree-1 predicts Class A (Apple).
     - Tree-2 predicts Class A (Apple).
     - Tree-n predicts Class B (Banana).
3. Majority Voting:
   - After all trees make their predictions, the model performs majority voting to decide the final class.
   - In this case, most trees predict Class A (Apple).
4. Final Prediction:
   - Based on the majority vote, the Random Forest classifier assigns the final class (here, Apple).

Key Takeaway:
The strength of Random Forest comes from the diversity and aggregation of multiple trees, which makes it robust, reduces overfitting, and improves accuracy. Each tree gives its "opinion," and the "majority wins"!

3.2 Random Forest used to Predict Unknowns

1. Input GIS Data (Features): the dataset included the following key features:
   - Year the lateral was installed (customer and city sides).
   - Lateral material (e.g., copper, plastic, galvanized, etc.).
   - Main material (e.g., ductile iron, PVC, etc.).
   - Year the structure was built.
   - Year the pressure main was installed.

   These features describe each water service lateral and provide clues about whether it might be made of copper.

2. Prediction Goal:
   The goal is to predict the likelihood that a water service lateral (customer or city side) is made of copper, based on its characteristics.

3. How Random Forest Works:
   - The algorithm takes the dataset and learns patterns from the provided features.
   - It builds multiple decision trees, where each tree focuses on different relationships in the data. This occurs randomly with a technique known as bootstrap resampling for statistical robustness. Through this randomized approach you eliminate the potential bias of a single decision tree and alternatively:
     - One tree might focus on the year of installation (e.g., copper was more common in certain decades).
     - Another tree might focus on pipe material (e.g., copper laterals might be associated with ductile iron mains).
     - A third tree might consider the structure’s age (e.g., older structures might be less likely to have copper).

4. Majority Voting:
   - Each tree predicts whether a service lateral is likely made of copper or not.
   - The Random Forest aggregates the predictions using majority voting, and the final output is the probability of copper.

5. Example Use Case: a service lateral with the following data:
   - Installed in 1970 on both customer and city sides.
   - The pressure main was installed in 1965.
   - The structure was built in 1972.

- The pipe material is ductile iron.

The model might predict:

- A high likelihood of copper because copper was a common material for laterals during that time period and often paired with ductile iron mains.

6. Feature Importance: Random Forest will also show you which features are most important for predicting copper. For example:
    - The year of installation might have the strongest influence.
    - The pipe material might be the next most important factor.

Why Random Forest is a Good Fit for This Prediction:

- Handles Complex Patterns: Random Forest can learn the relationships between the features (e.g., how copper use varies with installation year and pipe material).
- Feature Importance: It can help you understand which factors drive the prediction.
- Robust Predictions: By aggregating the results of multiple decision trees, the model is less prone to overfitting or bias.

4. Results of Newport Statistical Modeling

4.1 Results

Random Forest Classification was able to predict which unknown materials classify as copper utilizing the data input into the model. Our objective was to predict and classify unknowns and assign these services to copper with at least a 95% confidence interval. While NWD's intent is to use the model to classify copper service lines, probabilities of lead and galvanized materials are noted as well. Separate predictions were made for the NWD side and the Customer side with separate model runs for each. Data was split into training and model dataset randomly.

Results for the NWD side of the service line are shown in **Table 3**. The model will predict a material for all unknowns but we are only interested in the Copper result. <u>The results shows that 2998 NWD services can be reclassified to Copper with a > than 96% probability.</u> There was also a 99.9% probability to classify 245 services as lead. Galvanized and other material types were also classified but did not meet our 95% confidence threshold.

Results for the Private side of the service line are shown in **Table 4**. The model will predict a material for all unknowns but again we are only interested in the Copper result. <u>The results show that 6097 Private customer services can be reclassified to Copper with a > than 96% probability.</u> There was also a 99.9% probability to classify 359 services as lead. Galvanized and other material types were also classified but did not meet our 95% confidence threshold.

NWD

Predicted Class	Count
Copper	2998
Lead	245
Asbestos Cement	48
Galvanized	27
Ductile Iron	10
Plastic	1
Grand Total	**3329**

Probability Copper	Count
0.96429	**2997**
0	245
0.096	48
0.020	27
0.059	10
0.223	1
1	**1**
Grand Total	**3329**

Probability Galv	Count
0.007	2997
0.000	257
0.021	48
0.852	27
Grand Total	**3329**

Probability Lead	Count
0.005	2997
1	**245**
0	60
0.076	27
Grand Total	**3329**

Table 3. NWD side results of Random Forest Classification with Probability

Private

Predicted class	Count
Copper	6097
Lead	359
Galvanized	192
Asbestos Cement	47
Plastic	8
Grand Total	**6703**

Probability Copper	Count
0.96429	6091
0.000	359
0.020	192
0.096	47
0.22	8
1.00	6
Grand Total	**6703**

Probability Gal	Count
0.007	6091
0	373
0.852	192
0.021	47
Grand Total	**6703**

Probability Lead	Count
0.005	6091
1	359
0.075	192
0	61
Grand Total	**6703**

Table 4. Customer side results of Random Forest Classification with Probability

To further reduce the chance for error in assigning copper to unknowns, only residential properties were considered for changing the material type on the NWD or Customer side. QA/QC reduces the total number of services to be considered for change to copper to **5588** as shown in the map in **Appendix A4**.

ML models have the ability to extract which features drive the predictions strongest and **Figure 7** provides a depiction of a model trained to predict whether the type of material being used at a given site. Features associated with different service line material types such as size, date installed, location, as well as private side material (if known) are strong drivers of these predictions.

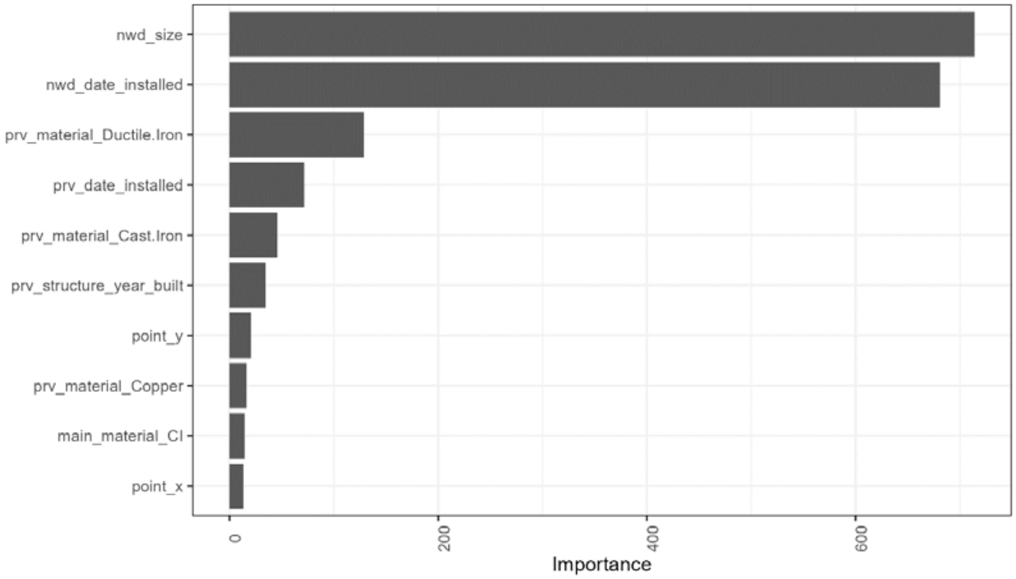


Figure 7 - Importance of Features when predicting Material Type on NWD Side

On the private side, the features of importance are similarly ordered, however the date installed on the private side is more influential to the quality of the private side size as expected and shown in **Figure 8**.

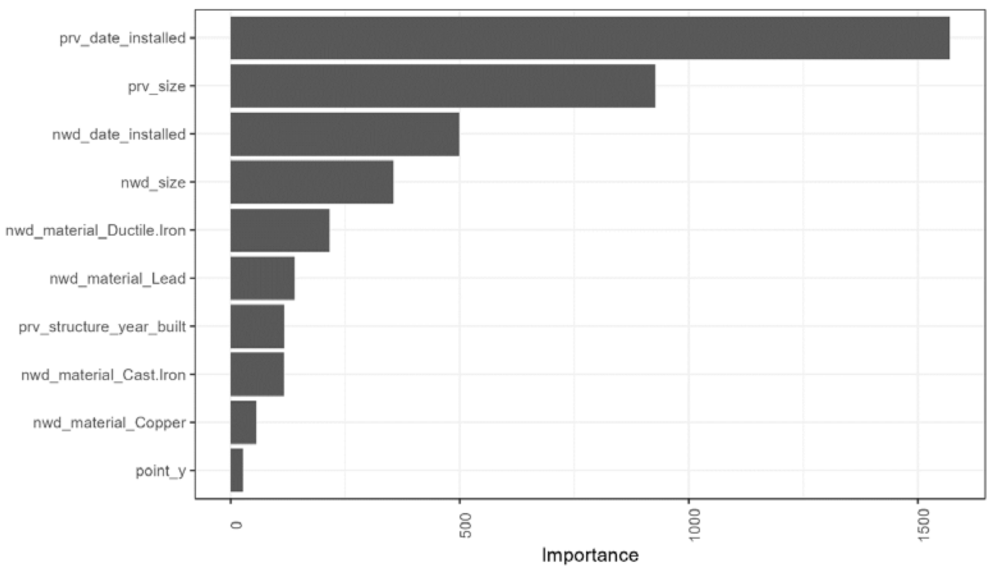


Figure 8 - Importance of Features when predicting Material Type on Private Side

4.2 Model Validation Next Steps

Field verification of model results for predicted copper will involve a statistically significant number of inspections. Assuming a 95% confidence level and a 5% margin of error this would be 385 inspections. On the NWD side this will involve visual identification utilizing hydrovac excavation. On the private customer side visual identification and scratch test will be performed at the structure interior lateral entry point typically in the basement. Since the submittal of the initial inventory on 10/16/2024 over 800 private customer inspections have been performed by NWD field staff.

5. Financial and Other Implications of Modeling

Implementation of statistical modeling will have a significant impact on the cost of the City's overall LSLR program as well as an impact on schedule. By utilizing the statistical modeling approach, the City would be able to eliminate a significant number of service lines from its "Unknown" category, thereby reducing the number of service line investigations needed to eliminate unknows from the inventory, and ultimately eliminate lead services from the system.

As noted in the sections above, investigation efforts required to eliminate the unknown services from the City's current inventory are time consuming, require significant customer coordination and cooperation, and can be disruptive. Internal investigations within a customer's property requires coordination and access to their property to conduct private-side investigations. For unknown service lines on the private side which cannot be easily accessed as well as all public side services, hydro-vacuum excavation is required. These excavations, especially those within the City's Right-of-Way can be disruptive to City traffic and require restoration to return the surface to its original condition.

By implementing statistical modeling to eliminate a portion of these unknowns from the inventory, the City will be able to focus more of its available funding on replacing known lead services as opposed to investigating service line materials, a significant number of which are anticipated to be non-lead. **Table 5** below provides an estimated savings that the City would expect to see through implementing the statistical analysis methodology.

Table 5. Estimated Investigation Costs Savings

Estimated Number of Private Unknowns Eliminated through Statistical Modeling	Cost for Private Side Unknown Investigations	Estimated Number of Public Unknowns Eliminated through Statistical Modeling	Cost per Public Side Unknown Investigations (HydroVac)	Total Estimated Savings
5,588	$100	2,512	$1200	$3,573,200

In addition to the anticipated cost savings this approach would provide, there would also be a measurable time savings recognized by the City for these efforts. The City currently has approximately 6,189 unknowns on the private side and 2,833 unknowns on the public side. Without implementing this statistical modeling approach, each of these locations would need to be further investigated to fully understand the number of lead service lines exist within the City's system. Furthermore, completing these investigations in advance of the November 2027 LCRI Compliance Date is crucial, as any remaining unknowns in the baseline inventory will contribute to the required number of annual LSLRs in order to meet the 10-year removal schedule.

Based on the City's hydrovac investigation program, it is estimated to take approximately 2 hours to complete a hydrovac excavation (not including advanced coordination efforts). Given the remaining number of public side unknowns, it is estimated that it would take the City 2.8 years to complete these investigations, based on a full-time crew completing at least 4 locations per work day. Given the complexity of these investigative efforts and the required coordination, it would be unlikely that crews could maintain this pace for an extended period of multiple years, therefore, it is expected that it would take longer to successfully complete all hydrovac activities. In addition to these public side investigations, there are also 6,189 private-side unknowns requiring internal investigations. Based on the City's experience with these internal investigations to date, it takes approximately 1 hour to complete these inspections. Given this timeframe, it is estimated that the City would need an additional 3.1 years to complete these internal investigations. It is also important to note the significant amount of staff time this would require. The City would need to dedicate several full-time employees to complete each of these investigation activities, which is not feasible given the size of the City's staff and the other activities the Newport Water Division staff must complete to maintain functionality of the water system. All of these factors contribute to the fact that it would be infeasible for the City to complete these investigations by the November 2027 Compliance Date, which would have significant consequences on the City's replacement schedule.

References

ASTM International. 2017. *ASTM D1193-99e1: Standard Specification for Reagent Water.* https://www.astm.org/d1193-99e01.html.

Code of Federal Regulations. 1975. *§ 141.23(k)(1) Inorganic Chemical Sampling and Analytical Requirement.* https://www.ecfr.gov/current/title-40/chapter-I/subchapter-D/part-141/subpart-C/section-141.23.

Code of Federal Regulations. 1991. *Subpart I - Control of Lead and Copper.* https://www.ecfr.gov/cgi-bin/text-idx?SID=9c5415b2fe8eb76878a169c14454171f&mc=true&node=sp40.25.141.i&rgn=div6 (accessed January 10, 2023).

U.S. Environmental Protection Agency. 2021a. *National Primary Drinking Water Regulations: Lead and Copper Rule Revisions.* 40 CFR Parts 141 and 142. Posted on Regulations.gov. https://www.regulations.gov/document?D=EPA-HQ-OW-2017-0300-1550 (accessed January 10, 2023).

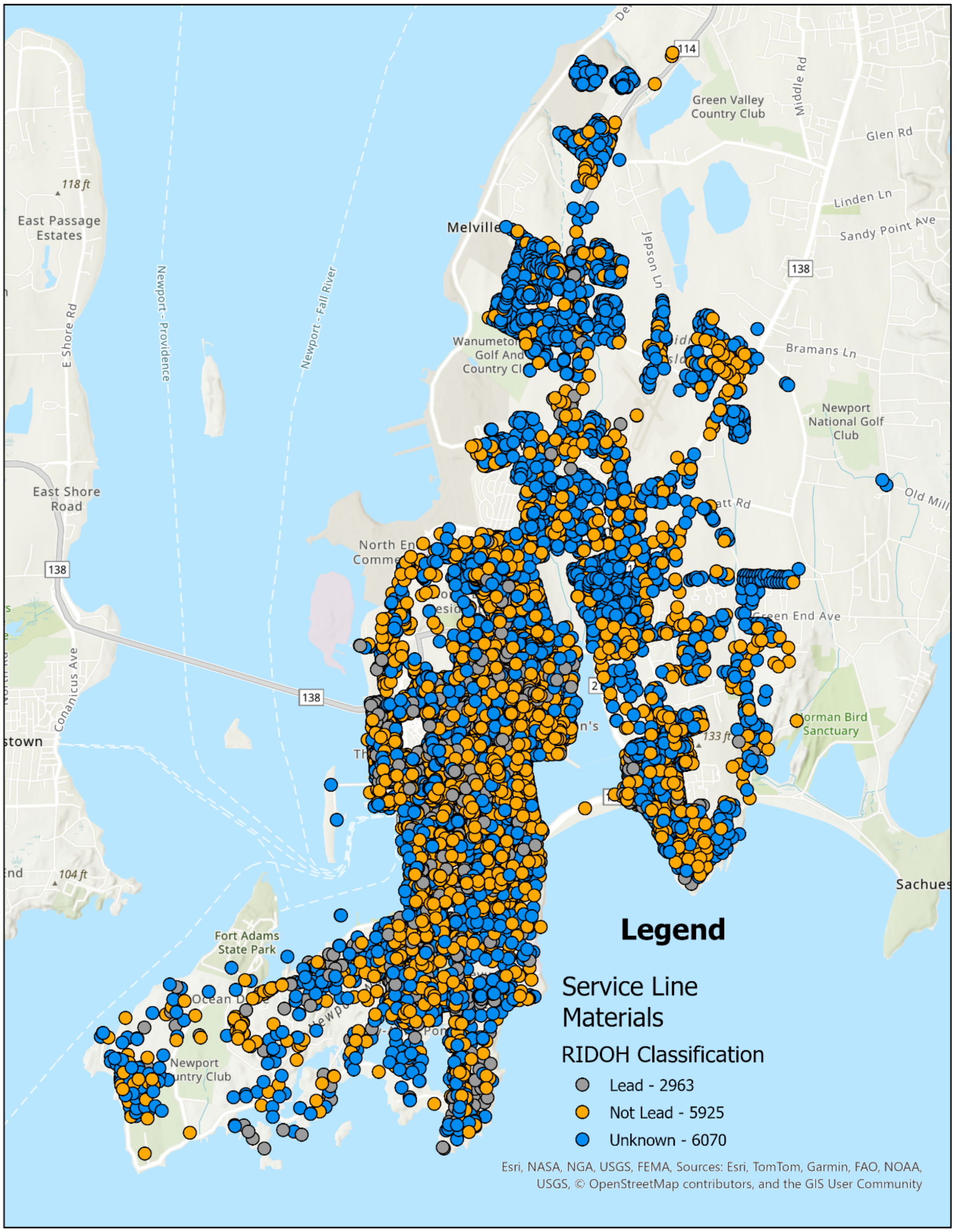
U.S. Environmental Protection Agency. 2021b. *Quality Assurance Project Plan Development Tool.* https://www.epa.gov/quality/quality-assurance-project-plan-development-tool#tab-1 (accessed January 10, 2023).

U.S. Environmental Agency. 2022. *Guidance for Developing and Maintaining a Service Line Inventory*. https://www.epa.gov/system/files/documents/2022-08/Inventory%20Guidance_August%202022_508%20compliant.pdf

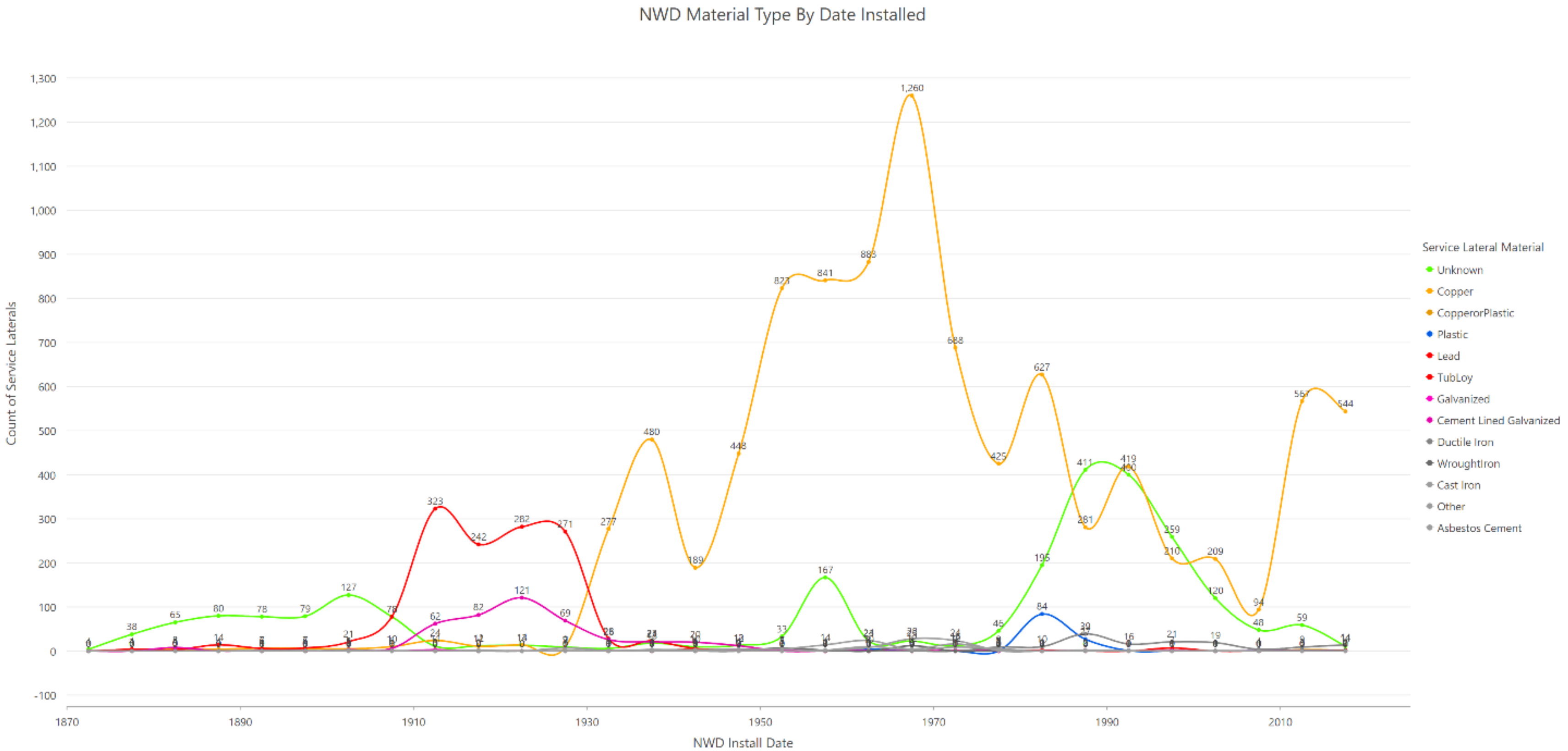
"A Brief Review of Random Forests for Water Scientists and Practitioners"
Authors: Tyralis, H., Papacharalampous, G., & Langousis, A.
Journal: Water, 2019.
Summary: This paper reviews the use of Random Forest (RF) algorithms in water resources, highlighting their potential in solving diverse practical problems in the water sector.

Appendix A.

A.1 Inventory submitted to RIDOH 10/15/2024

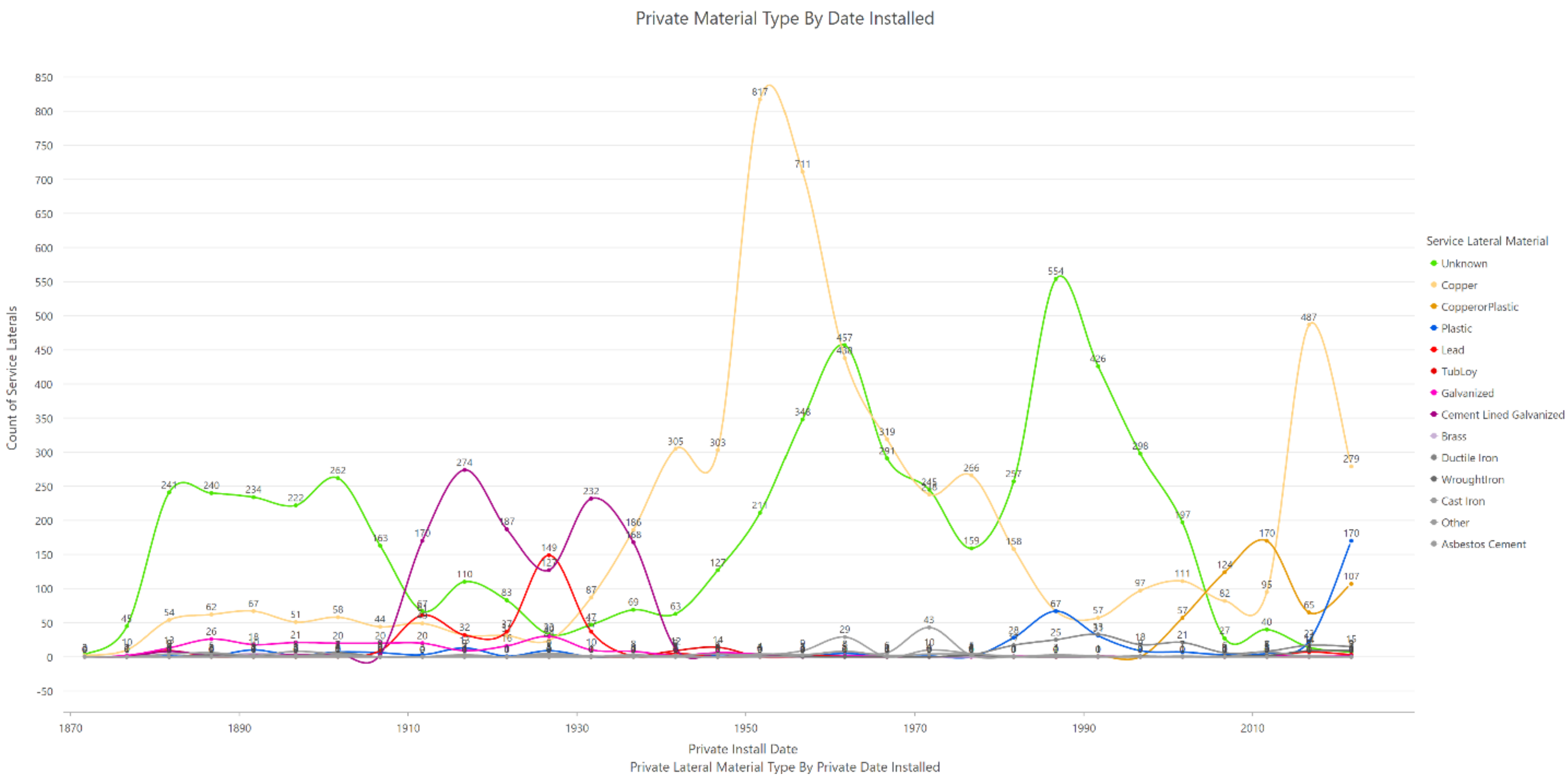


A.2 NWD Materials by installation date



NWD Lateral Material Type By NWD Date Installed

A.3 Private Materials by installation date



A.4 NWD Predicted Material Classification