

MassDEP Drinking Water Program

Statistical (Predictive) Modeling Guidance for Evaluating Unknown Service Lines

October 2023

Introduction: The Lead and Copper Rule Revisions (LCRR) Requirements

The 2021 EPA LCRR require public water systems (PWS) to develop a complete inventory of all service lines. This includes identifying the materials of both public and private portions of the service lines.

In this document, “known service line” is defined as a service line where the pipe material is categorized using records or other means. “Unknown service line” is defined as a service line of unknown material with no documented material history.

The statistical, or predictive, modeling approach provides a method to complete a service line inventory while, eliminating or prioritizing the need to inspect every unknown service line.

What is Statistical Modeling?

Statistical Modeling, sometimes called Predictive Modeling, is an identification method that uses machine learning to predict the material of an unknown service line based on the data set of identified service lines entered into the model.

If using a statistical model, MassDEP’s Drinking Water Program will expect the PWS’s model to demonstrate a minimum statistical confidence of 95 percent.

Prior to Using Statistical Methods: The Identification Process

1. Before using a statistical approach to identify unknown service lines, **the PWS must first use other MassDEP/EPA approved methodologies** (records review, including post 1986 construction, exclusion of larger pipe diameters, and optionally, customer data) to categorize service lines¹.
2. If the PWS still has unknown service lines in their inventories **after using other MassDEP/EPA approved methodologies**, a statistical approach may be used. **Please note: MassDEP considers Predictive Modeling as a last resort after other methods of identifying the materials of service lines has been exhausted.**
3. **If a PWS decides to use predictive modeling as part of their verification process, the PWS should ensure their selected product is using all the PWS available records to train the model.** Predictive models using borrowed records data (from other PWSs) to train the model will not be accepted by MassDEP.

¹ **Other methods include:**

Field Inspection by PWS: This is considered the most accurate verification method that uses a physical and visual inspection by a trained staff person. Typically, at the time of meter replacement, service line replacement, or special inspections such as pot holing and vacuum excavation.

Records Review: This verification method includes review of current or past PWS records including tap/tie cards, distribution system main replacement or leak detection or any projects where service line material may have been recorded by the PWS. Other potential sources of information in a community might include plumbing and building permits, or inspectional services records, or year of construction.

Customer Self-Identification: This verification method uses information collected from building occupants, and typically includes photos of the service line. The MassDEP crowdsourcing application or a similar software solution can be used to collect and verify the information.

MassDEP Drinking Water Program

Statistical (Predictive) Modeling Guidance for Evaluating Unknown Service Lines

October 2023

MassDEP may also require PWS that use predictive modeling to submit a long-term compliance plan using other methods to confirm identification for all service lines initially identified by predictive modeling.

Methodologies for identifying service line materials can be found in the EPA LCRR guidance at [Inventory Guidance_Final 080322.pdf \(epa.gov\)](#).

Note: If after using a statistical model more than 5% of the predicted service line materials are found to be a different material than predicted by the model, then the statistical method for determining unknowns must be re-evaluated. For further PWS specific guidance contact MassDEP Drinking Water Program at program.director-dwp@mass.gov, Subject: LCRR.

For PWSs interested in exploring the use of a statistical model, please be aware of the following information from the MassDEP LCRR Q&A located at <https://www.mass.gov/doc/frequently-asked-questions-about-the-lead-and-copper-rule-revisions-lcrr/download>:

MassDEP does not endorse third-party products and/or services including Predictive Modeling products, but we encourage consultants to educate their clients on the product being considered so that they can make an informed decision. PWSs considering Predictive Models, i.e., machine learning, for gathering service line information, required under the LCRR, need to ensure the product meets their goals for both the short and long term. MassDEP recommends that PWSs fully evaluate the options and ask all the necessary questions to make an informed decision prior to agreeing to any contract.

Some considerations when evaluating Lead Service Line Predictive Modeling products:

1. Will the product meet your objectives?
 - Provide Service Line Inventory acceptable for MassDEP reporting (See MassDEP Service Line Inventory (SLI) Workbook at <https://www.mass.gov/media/2480901>. Instructions can be found at <https://www.mass.gov/media/2480886/>)
 - Ability for improvement over time
 - Meet your confidence levels
 - Minimize resource inputs to alternatives (in-person verification)
 - Meet LCRR October 2024 reporting deadline
2. What are the biggest obstacles to getting this done?
 - Level of effort and resources to provide the data inputs, i.e., collecting and feeding data to the predictive model to achieve desired confidence level.
 - PWSs should be looking for a confidence level of 95% or greater with at least 20 - 25% field verification.
 - Responsibilities for data collection
3. Cost?
 - Upfront cost
 - Future maintenance costs
4. Has the model encountered barriers in the past?
 - Ask for references or examples from systems like yours
5. If the project doesn't succeed, what are the implications?

MassDEP Drinking Water Program
Statistical (Predictive) Modeling Guidance for Evaluating Unknown Service Lines
October 2023

- What are the guarantees to meet the 2024 deadline?
6. PWS must carefully evaluate all products.

How to Identify Service Lines Using Predictive Models

Step 1: Identify all service lines of unknown material.

Identify all water service lines that cannot be categorized using another approved methodology. Determine the total number of these unknown service lines.

Step 2: Select a statistical model that uses multiple verification methods to train the model.

The statistical model should train the model using 80% of the service lines with known materials and test the model with the remaining 20% of service lines (SLs) with known materials.

Please be aware that using only historical records like tie cards without proactively verifying their reliability, can lead to the model making inaccurate predictions from these records. Therefore, these records should be verified.

In addition, it is important that the model is used in a way that prevents biases. Biases might appear when specific home or neighborhood types show up too frequently or not at all in the data used for prediction. For instance, if a city's historical records are concentrated in one particular neighborhood, the model will probably perform well there but fall short elsewhere. Therefore, gathering representative data to feed the model is a crucial step in preventing bias (for example, data from across the city selected randomly rather than by convenience).

It is also possible to introduce biases when predicting service line materials by using only tie cards or only house age or building codes. However, when numerous inputs are provided to the model at once, such as tie cards, building age, customer self-identification, and construction codes, the model can predict service line materials with more accuracy.

See Appendix A to see how to randomly verify services lines using physical inspections.

Step 3: Predict the materials for all unknown service lines and record the predicted material.

Record the material identification method as “statistical” and the service line material as predicted by the model.

Step 4: Submit the service line inventory.

Submit the SLI and remember to select statistical analysis as the verification method for the service line materials that were predicted using the model.

Step 5: Verify the predicted materials.

Over time, during routine operations, verify the predicted materials, update the service line inventory and submit corrections as required by the LCRR. Depending on the number or percentage of predictions discovered

MassDEP Drinking Water Program
Statistical (Predictive) Modeling Guidance for Evaluating Unknown Service Lines
October 2023

to be inaccurate, we may require that all predictions revert to unknown status. Remember to re-run your predictive model with verified information to improve the accuracy for the remaining unknown service lines.

Step 6: Retain identification records.

- Create, compile, and retain documentation of all service line identification efforts.
- MassDEP/DWP may ask PWSs to produce or submit these records.

For any questions on this information please contact the MassDEP Drinking Water Program at program.director-dwp@mass.gov or 617-292-5770.

MassDEP Drinking Water Program
Statistical (Predictive) Modeling Guidance for Evaluating Unknown Service Lines
October 2023
Appendix A

Identify how many service lines must be physically inspected.

- PWSs with fewer than 1,500 unknown service lines must physically verify at least 20 percent of the total number of unknown lines.
- PWSs with more than 1,500 unknown service lines must physically verify enough lines to reach a minimum 95 percent confidence level.

See Table A to determine the number of service lines requiring verification. Table A uses a confidence level of 95 percent.

Randomly select service lines for physical inspection.

- From the list of unknown service lines identified in Step 1 of How to Identify Service Lines Using Predictive Models randomly select the number of service lines determined in Step 2 to be physically inspected.
- Selection must be uniformly random and not based on any specific criteria which can introduce bias. In other words, each unknown service line must have an equal chance of being chosen for verification.
- See Appendix B for an easy way to generate a uniformly random set of service lines for inspection.

Note: It may be tempting to introduce a “logic” to the site selection process, such as selecting within periods of construction or targeting portions of town. However, doing so can unintentionally bias the data set. Be certain to use a truly random selection method such as the one described in Appendix B.

Conduct a one-point (or more, if needed) physical inspection.

- At least one-point physical identification is required for **each** portion of the unknown service line. If the service line is jointly owned, each portion that is unknown (public and/or customer) must be inspected.
- Physical identification methods include excavation, in-home inspections, and other emerging methods and must be conducted or overseen by water system personnel.
- Record the actual material observed at each point.
- If inspecting near the meter, ensure the observed material is the actual service line and not part of the metering components.

Refer to EPA’s “Guidance for Developing and Maintaining a Service Line Inventory,” Chapter Five, for typical methods of service line identification. See [Inventory Guidance_Final 080322.pdf \(epa.gov\)](https://www.epa.gov/waterservices/inventory-guidance-final-080322.pdf).

If one or more of the original randomly selected sites cannot be physically inspected, the PWS must substitute it by randomly generating a new site using the process described in Appendix B.

Record results of the physical inspection process.

- The PWS should record the results of the physical inspection using their SLI database or the MassDEP SLI Excel workbook located at <https://www.mass.gov/media/2480901>. If using the MassDEP SLI workbook, in the dropdown list, enter the service line material observed at each point. The spreadsheet

MassDEP Drinking Water Program
Statistical (Predictive) Modeling Guidance for Evaluating Unknown Service Lines
October 2023

will automatically categorize the entire service line into one of the four required EPA categories. The four service line category types are: lead, non-lead, galvanized requiring replacement and unknown. The spreadsheet has further subclassification categories of material types that is recommended the water system utilize.

Table A. Minimum number of service lines requiring physical inspection.

Number of Unknown Service Lines*	Number to Physically Inspect
Fewer than 1,500	20% of unknown lines
1,500	306
1,600	310
1,700	314
1,800	317
1,900	320
2,000	322
2,200	327
2,400	331
2,600	335
2,800	338
3,000	341
3,500	346
4,000	351
4,500	354
5,000	357
6,000	361
7,000	364
8,000	367
9,000	368
10,000	370
15,000	375
20,000	377
30,000	379
40,000	381
60,000	382
90,000	383
225,000 or more	384
<i>Table taken from Oregon Health Authority: Statistical Guidance for Evaluating Unknown Service Lines.</i>	

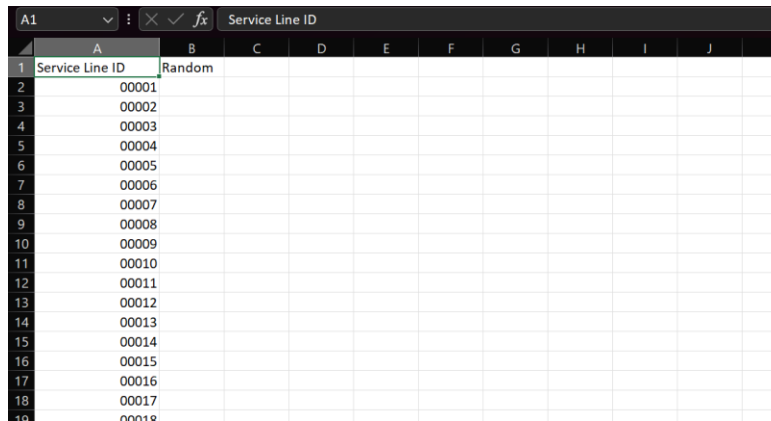
*For the purposes of this process, this number represents the number of service lines that cannot be categorized from records, installation date, diameter, previous physical inspection, or customer data. If the number of unknowns falls between two values on the chart, either interpolate or round up to the higher number.

MassDEP Drinking Water Program
Statistical (Predictive) Modeling Guidance for Evaluating Unknown Service Lines
October 2023

Appendix B
Generating a uniformly random set of service lines for inspection

You can use a spreadsheet (such as Microsoft Excel or Google Sheets) to generate a uniformly random set of locations of unknown service lines for inspection using the following Microsoft Excel steps (the same formulas and method work for Google Sheets):

1. In the first column of a spreadsheet, list every unique service line of unknown material. They can be listed by address, service line ID, or other identification method.

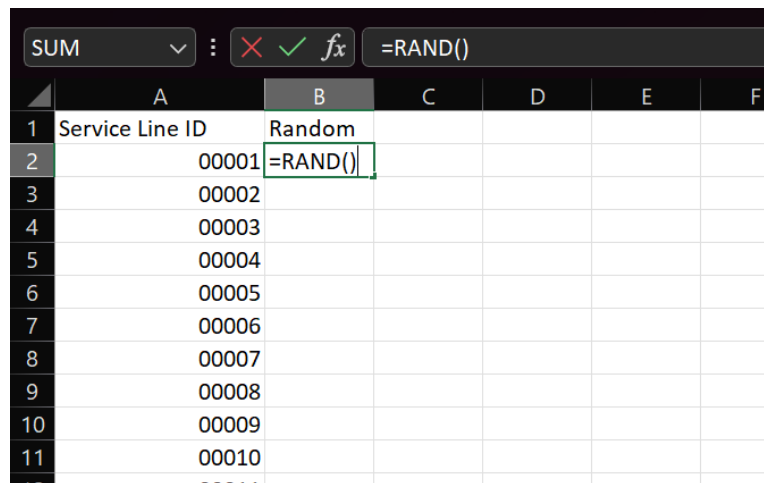


	A	B	C	D	E	F	G	H	I	J	K
1	Service Line ID	Random									
2	00001										
3	00002										
4	00003										
5	00004										
6	00005										
7	00006										
8	00007										
9	00008										
10	00009										
11	00010										
12	00011										
13	00012										
14	00013										
15	00014										
16	00015										
17	00016										
18	00017										
19	00018										

2. In the second column, generate uniformly random numbers, so that each service line is associated with a randomly generated number.

Follow these steps:

- a. Enter the formula =RAND() into the first cell of the second column next to the first service line location and press Enter. This generates a number between 0 and 1 for each service line.



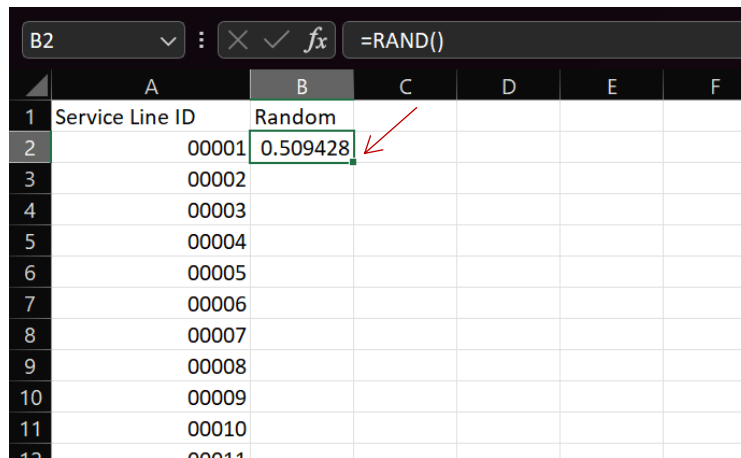
	A	B	C	D	E	F
1	Service Line ID	Random				
2	00001	=RAND()				
3	00002					
4	00003					
5	00004					
6	00005					
7	00006					
8	00007					
9	00008					
10	00009					
11	00010					
12	00011					

- b. Select the lower right corner of the first cell in the second column (the column with the random value) and double click the small square to copy the formula into the cells below it so that every service line location is assigned a random number.

MassDEP Drinking Water Program

Statistical (Predictive) Modeling Guidance for Evaluating Unknown Service Lines

October 2023

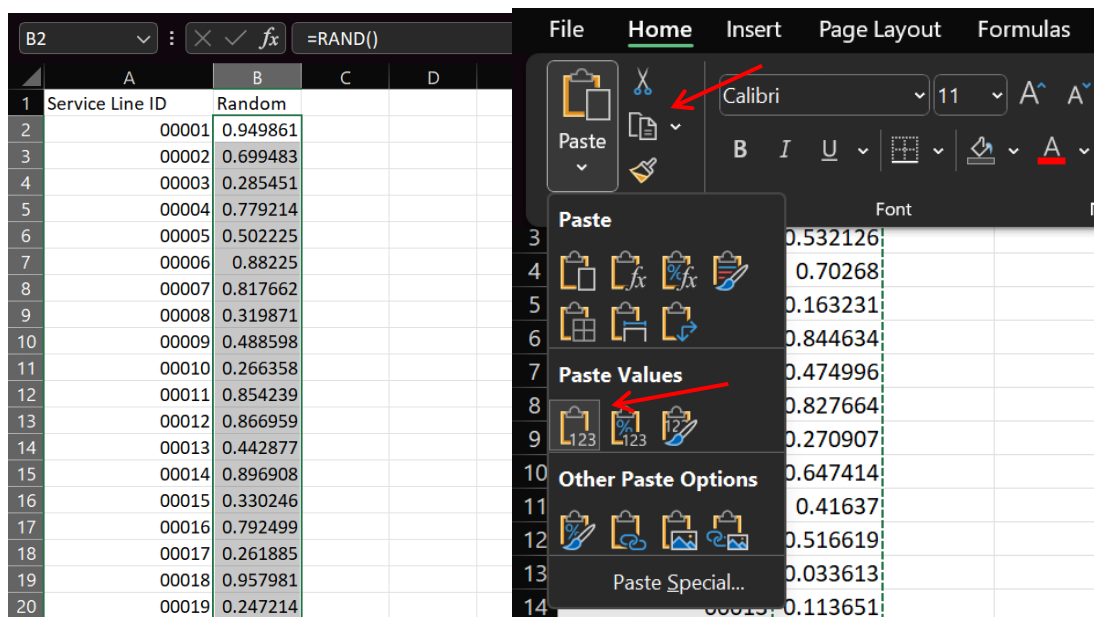


This screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F
1	Service Line ID	Random				
2	00001	0.509428				
3	00002					
4	00003					
5	00004					
6	00005					
7	00006					
8	00007					
9	00008					
10	00009					
11	00010					
12	00011					

The formula bar at the top shows the formula `=RAND()` being entered into cell B2. A red arrow points to the value 0.509428 in cell B2.

- c. With the entire second column still selected, select Copy and then the Paste Special option to Paste Values Only into that same column. This will overwrite the formula with the set of random numbers and ensure these random numbers remain static.



This screenshot shows the Excel spreadsheet after the Paste Special operation. The formula bar still shows `=RAND()`. The 'Paste' dropdown menu is open, and the 'Paste Values' option is selected. The random numbers are now static values in column B.

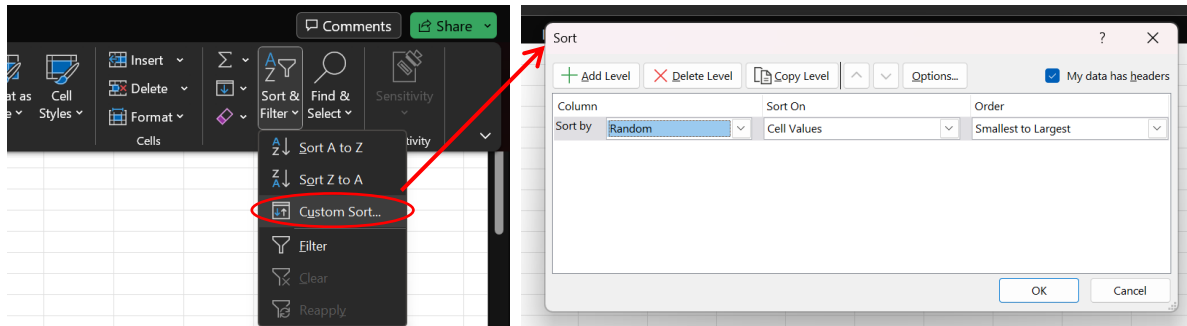
	A	B	C	D
1	Service Line ID	Random		
2	00001	0.949861		
3	00002	0.699483		
4	00003	0.285451		
5	00004	0.779214		
6	00005	0.502225		
7	00006	0.88225		
8	00007	0.817662		
9	00008	0.319871		
10	00009	0.488598		
11	00010	0.266358		
12	00011	0.854239		
13	00012	0.866959		
14	00013	0.442877		
15	00014	0.896908		
16	00015	0.330246		
17	00016	0.792499		
18	00017	0.261885		
19	00018	0.957981		
20	00019	0.247214		

- d. Use the Sort feature to list the randomly generated numbers from lowest to highest. If the Sort Warning appears, select Expand the Selection, then Sort.

MassDEP Drinking Water Program

Statistical (Predictive) Modeling Guidance for Evaluating Unknown Service Lines

October 2023



2. Select only the top N service lines, where N is the number requiring inspection. For example, if you need to inspect 20 service lines, select the first 20 service lines on the list. These are the 20 uniformly random service lines to be inspected.

	A	B	C	D	E
1	Service Line ID	Random			
2	00085	0.000922			
3	00071	0.004868			
4	00049	0.01286			
5	00027	0.018663			
6	00031	0.037531			
7	00036	0.061321			
8	00069	0.064076			
9	00029	0.066213			
10	00028	0.079171			
11	00021	0.098848			
12	00059	0.098886			
13	00054	0.10848			

See the brief video on-line tutorial at <https://www.youtube.com/watch?v=q8fU001P2II> for generating random samples on Microsoft Excel.