MassDEP Drinking Water Program

Statistical and Predictive Modeling Guidance for Evaluating Unknown Service Lines

Updated July 2025

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

The 2021 EPA Lead and Copper Rule Revisions (LCRR) require~~d~~s public water systems (PWS) to develop a complete inventory of all service lines. This include~~d~~s identifying the materials of both public and private portions of the service lines.

~~In this document, a "known service line" is defined as a service line where the pipe material is categorized using records or other means. An "unknown service line" is defined as a service line of unknown material with no documented material history.~~The 2024 Lead and Copper Rule Improvements (LCRI) requires PWS to build on their initial Service Line Inventory (SLI) with the creation of a Baseline SLI. Additional information such as including connector and street address data is required in the Baseline SLI. The LCRI also includes further requirements to identify all unknown service lines and replace all Lead and Galvanized Requiring Replacement (GRR) Service Lines.

The statistical and predictive modeling approach(es) provides ~~a~~ method~~s~~ to complete and update an ~~service line~~ inventory while eliminating or prioritizing the need to inspect every lead status unknown service line.

What ~~are~~ is Statistical and Predictive Modeling?

Statistical Modeling is an identification method that uses the composition of known service lines to predict the material of unknown service lines in a service area. To do this with a statistically significant result, it is normally performed with a randomly selected group of service lines. Predictive modeling is a ~~form~~ version of statistical modeling, or often a further step after statistical modeling, that uses machine learning to predict the material of unknown service lines based on the previously selected group of known service lines.

MassDEP DWP Approval is Now REQUIRED Prior to PWS Utilizing Statistical Analysis or Predictive Modeling for their SLI

As of July 3, 2025, prior written approval from MassDEP DWP is required for the use of statistical analysis or a predictive model.

Note the following basic requirements PWS will need to meet to create a statistical analysis/predictive model

PWS must contact MassDEP DWP at program.director-dwp@mass.gov, subject: Requesting Statistical Analysis/Predictive Modeling SLI Approval. PWS should include any relevant information regarding their PWS, including information about their distribution system, whether they are requesting to use a

statistical analysis and/or predictive model, if lead or galvanized service lines have been found previously, and so on. From there, MassDEP DWP will work with PWS to schedule a meeting to discuss the proposal if necessary, and provide either approval or rejection of the proposal, and possibly next steps if further work or justification is needed to approve the request.
PWS should note the following guidelines regarding the use of a statistical analysis prior to considering using it as a verification method. However, PWS should note that no matter what conditions are met, PWS still need approval from MassDEP DWP before beginning the process of creating a statistical analysis or predictive model.

Limitations of a Statistical Analysis or Predictive Model

PWS should consider the following information when determining whether a statistical analysis or predictive model is right for your PWS:

~~PWS should consider:~~

- ~~Data used to train predictive or statistical models must belong to the PWS using the model (i.e. **PWS must use their own records for training, testing, and using a model, and cannot "borrow" data from another system at any point in this process**).~~
- ~~MassDEP reserves the right to reject a statistical or predictive model as a verification method if the required submitted documentation does not demonstrate an unbiased or representative model of the system.~~
- ~~MassDEP may require PWS to submit a plan to perform an agreed upon amount of field inspections on service lines identified using a statistical/predictive model after October 16th, 2024.~~
- ~~Both predictive and statistical models will require a confidence level of **95% or greater.**~~

Prior to Using Statistical/Predictive 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** (such as records review, including post 1986 construction, exclusion of larger pipe diameters, and optionally, customer data) to categorize service lines[1].~~
2. ~~If the PWS still has unknown service lines in their inventories **after using other**~~

[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 the 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/EPA approved methodologies, a statistical/predictive approach may be used. **Please note: MassDEP considers predictive modeling as a last resort after other methods of identifying the materials of service lines have 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 verified and accurate PWS records to train the model.** Predictive models using borrowed data (from other PWSs) to train the model will not be accepted by MassDEP.

MassDEP may also require PWS that use predictive or statistical modeling to submit a long-term compliance plan using other methods to confirm identification for all service lines initially identified by statistical or predictive modeling. Methodologies for identifying service line materials can be found in in the EPA LCRR guidance at Inventory Guidance_Final 080322.pdf (epa.gov).

For PWSs interested in exploring the use of a statistical/predictive 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.

Considerations

Some considerations when evaluating Service Line Predictive Modeling products:

1. ***Will the product meet the following objectives to:***
   - Provide a 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 confidence levels (of 95% or greater)
   - Minimize resource inputs to alternatives (in-person verification)
   - Meet LCRR October 16th, 2024,your reporting deadline
   -
2. ***What can be the 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 and MassDEP

strongly recommends PWS verify 20 - 25% of the predicted service lines through field inspections.

- ~~Data~~
  - ~~Does your PWS have the capacity to handle the data output of a predictive model?~~
  - ~~What format will the data be presented?~~
  - ~~Whose responsibility is it to put this model into the MassDEP accepted format?~~
-

- Responsibilities for data collection
- *Cost~~?~~*
  - Upfront cost
  - ~~Future maintenance costs~~
  -

3. ***Has the model encountered barriers in the past?***
   - ~~Ask for references or examples from systems like yours~~
   - ~~Follow up with provided references for their experience in their own words.~~
   -
4. ~~***If the project doesn't succeed, what are the implications?***~~
   - ~~What are the guarantees to meet the 2024 deadline?~~
4.
5. ***PWS must carefully evaluate all products.***

For PWSs interested in exploring the use of a statistical/predictive model, please be aware of the above 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.

~~5.~~

Cost Limitations

Statistical Analysis and Predictive Modeling can both have high upfront costs and maintenance costs, depending on the composition of your PWS (for example, the size of system, age, variety of pipe materials used) and the work planned. PWS are encouraged to consider cost now as well as long term with all other pros and cons to determine if either (or both) types of analysis are right for your PWS.

Statistical Analysis:

Statistical Analysis as a method alone can be a very cost-effective verification method, because the PWS can verify a small number of service lines with potholing and in person inspections and then use these results to predict that the rest of the PWS's unknown service lines are non-lead, instead of verifying the material of every single service line. **However, if a single lead service line is found, statistical analysis can no longer be used.** PWS at this point may need to pivot to a predictive model, or continue to physically verify all service lines, which can cost more than the PWS may have originally expected when selecting statistical analysis alone.

If the PWS has any galvanized service lines on the private side, this method may result in increased cost, as PWS with galvanized service lines may not be approved to use the verification method of statistical analysis, such as if the PWS did not meet the GRR acceptance limit and process.

Predictive Modeling:

Predictive Modeling, similar to statistical analysis (as it uses the analysis as a first step to the model), can be cost effective to PWS, as it requires less physical inspection and digging to

determine all service line material. **However, the cost will increase for the following reasons:**

- If not enough lead service lines are found, creating an accurate predictive model may not be possible, and the PWS will not be approved to use a statistical analysis if any lead service lines are found. At this stage, a PWS would be in "limbo", unable to use either method of verification.
- Predictive models are living models, and updated overtime. Consider what the cost is for maintenance, and how this will affect your PWS overtime. Are costs secured by your PWS with contracts?
- Will your PWS require additional models, such as a model to determine GRR service lines?
- If your PWS finds 5% (or 1% if using a statistical analysis) of service lines were classified incorrectly, all service lines verified by this method will need to be reclassified as UNK-LG.

Future Non-Lead Validations:
All PWS should remember that the Lead and Copper Rule Improvements (LCRI) will require PWS to validate (confirm) that their non-lead service lines are non-lead by creating a pool of applicable non-lead service lines, selecting a random group of these non-lead service lines, and conducting 2-point physical inspections to confirm the material. Service lines verified by statistical analysis/predictive modeling will be included in this validation pool and may be required to be inspected in the future.

Statistical Analysis Limitations

If a PWS has Lead Service Lines:
PWS which have any known lead service lines will not be approved to perform a statistical analysis, unless there are certain conditions met which MassDEP DWP approves of. This may include scenarios such as:

- If a certain section of the PWS used lead, however there are clear records stating that lead was only used in that specific area of the PWS. This may have happened if a PWS expanded into another area where the previous owner/PWS installed lead in that area.

PWS must note that if any lead is found during the initial investigations to create a statistical analysis, the statistical analysis cannot be used to determine that unknown service lines are not lead or galvanized (UNK-NOLG). If a PWS creating a statistical analysis finds lead at any point during their investigations to create an analysis, they must stop immediately.
Should a PWS find lead during initial investigations, they must do one of the following:

1. Continue performing investigations and turn to a predictive model, or
2. Begin to plan for another method to determine service line materials, such as records review, field inspections, customer self-identifications, or other approved methods of verification which are not Statistical Analysis.

If a PWS has Galvanized Service Lines:

PWS should not create a statistical analysis on the private side of their service lines if there is a likelihood that the PWS may have multiple Galvanized Requiring Replacement (GRR) service lines. This means that if a PWS has, or expects, Galvanized service lines on the private side, and they have lead, unknown, or non-lead service lines where it is unknown if they were ever lead, on the public side,

they should consider if a statistical analysis is in their best interests and cost effective, see the new GRR acceptance limit and process below for more information. PWS in this scenario may not be approved to use this verification method, or not have the method accepted in their service line inventory, should they not meet the GRR acceptance limit and process.

PWS should note that if any galvanized service lines are found on the private side during a statistical analysis, this may impact the classification of all service lines being verified by this method.
In the interest of protecting public health and to assist PWS to identify all service lines which are lead or contain lead, MassDEP has developed the following **Galvanized Requiring Replacement (GRR) acceptance limit and process** for PWS that have used statistical analysis as their verification method in accordance with MassDEP/DWP Statistical Analysis requirements in their Service Line Inventory (SLI): **A maximum of 2.5% of all service lines verified by statistical analysis, which must be less than or equal to 25 service lines, that could be GRR if the material is discovered to be a galvanized material is acceptable.** PWS which meet this limit must also provide for MassDEP/DWP's approval a Non-Lead Validation Compliance Plan, which describes the PWS's approach to finding possible GRR service lines during non-lead validations, which are required under the Lead and Copper Rule Improvements (LCRI).

PWS with galvanized service lines will have a higher chance of approval to use a statistical analysis on the private side, if they do the following:

- Perform a statistical analysis on the public side of the service lines and determine that lead is not installed on the public side of the service lines and never was.
- The PWS is aware due to other reasons that lead is not on the public side and never was installed on the public side.
- The PWS, based on evidence in their records, can document that there is a very small number of galvanized service lines in the private side, and are prepared to conduct a large number of field inspections to meet the GRR acceptance limit and process.
- PWS should note that if there are records which indicate galvanized was used only during certain times or in certain areas, this information may be taken into consideration when MassDEP is considering approving your PWSs use of statistical analysis.
- **PWS must note that if PWS cannot meet the GRR acceptance limit and process, the statistical analysis cannot be used to determine that unknown service lines are not lead or galvanized (UNK-NOLG).** PWS must reclassify all service lines classified as non-lead due to statistical analysis, which could be GRR, as UNK-LG. If the PWS is able to conduct more inspections to meet the GRR acceptance limit and process, the PWS can resubmit their SLI and required documentation for approval later on.

Note: MassDEP DWP reserves the right to reject any proposals for statistical analysis due to concern with misclassifying possible GRR service lines as non-lead.

If PWS is Using a Water Main/Block Level or Neighborhood Wide Level Analysis:

If the PWS is using a smaller scale analysis/model, MassDEP DWP may allow the PWS to use smaller scale analysis/models to determine the composition of certain areas of the PWS, even if the PWS has lead or galvanized service lines in their service area. This may be allowed if and when the PWS:

- only installed lead or galvanized in certain areas of the PWS, or

- did so only during certain time periods,

and the PWS can support a belief that certain areas of a PWS have no lead or galvanized service lines. In these cases, MassDEP DWP will review all provided information and may allow PWS to conduct an analysis/model. However, PWS should note that if they are approved to use an analysis/model, they must still follow all requirements discussed in this guidance and may have to revert selected service lines to UNK-LG if certain conditions and acceptance limits/processes are not met. For this reasoning, PWS should proceed with caution when evaluating all verification methods.

~~Choosing a Model and Sample to Train the Model~~

~~*Level of Model Identification*~~

~~Models are dependent on the type of data entered into the model to train it. Models can vary in the area they predict/represent, dependent on the data entered and the bounds of the model.~~

~~There are Three Types of Statistical/Predictive Models:~~

- ~~System Wide Level~~
- ~~Neighborhood Wide Level~~
- ~~Water Main/block level~~

~~Systems are reminded to discuss with their contractor what the best model may be for their service area, and which has the most representative results. Systems with lead congregated in certain areas of the service area/town may benefit from a neighborhood level approach, to focus on areas with a higher likelihood of lead, while others may prefer a system wide level.~~

~~*Sample Groups/Investigation Pools*~~

~~**Investigation Pool:** This term refers to the service lines chosen randomly that must be identified,~~

~~i.e. a sample group. Identification may include verified methods such as field inspections, PWS records, operator knowledge, and other approved methods. PWS should note that field inspections are the preferred and recommended verification method for this process. If your PWS is unsure on the validity of a source of information, such as tie cards or certain PWS records, PWS should exclude them from their verification methods used in this process.~~

~~Statistical/Predictive Models may also have varying types of sample groups/investigation pools used to train the model. There are two current options for PWS to create their investigation pools which MassDEP will accept for statistical and predictive models.~~

~~**1.** A pool of randomly chosen known and unknown service lines chosen from your **entire service area.**~~

- ~~This sample/investigation pool must meet the numbers provided in Appendix A, Table A.~~

~~**2.** A pool of randomly chosen unknown service lines chosen from your entire inventory of **unknown service lines.**~~

- ~~This sample/investigation pool must meet the numbers provided in Appendix A, Table A.~~

~~***Note:***~~

- ~~All predictive models should be trained using an 80/20 model, meaning 20% of the~~

known service lines should be held out to test the predictive model while training. The recommendation is to test a predictive model many times and choose the best version moving forward, then improving with further identifications of service lines.

- **PWS must use a random method to find the service lines included in the initial sampling/investigation pool.** See Appendix B for one method of doing so.
- **Sample Groups/Investigation Pools do NOT need to be 100% verified by field inspections.** Only **20%** of the investigation pool of service lines must have their verification method confirmed by field inspection. This field inspection must have been conducted within the last 10 years.

General Statistical/Predictive Model Verification Method Requirements Performing a Statistical Model

Statistical Models, or statistical analysis are a way PWS may use statistically significant means to predict the composition of their service line inventory. This method tends to be used by PWS that expect to have no lead or galvanized requiring replacement (GRR) service lines in their service area. By randomly selecting a statistically significant number of service lines to

identify, PWS may extrapolate that number to the rest of the service area. This method does not use machine learning but may be a step for PWS who will use machine learning after finding out if their system does or does not contain lead.

PWS must be aware:

- PWS can only use statistical modeling as a verification method if there are **no GRR or lead service lines discovered during the initial investigation.** If any lead or GRR service lines are found, systems must use another method to find the location of all expected lead and GRR service lines, whether that be predictive modeling or another approved method.
- **PWS must use a random method to find the service lines included in the initial sampling/investigation pool.** See Appendix B for one method of doing so.
- PWS may use their entire service area, known and unknown service lines, to pull from for their investigation pool. However, this pool of service lines must meet the required numbers in Appendix A, Table A to be **statistically significant.**
- If PWS find a total of **1% or more** service lines that are lead or GRR in years following the submission of their SLI, PWS must revert all service line materials back to an unknown status.[2]

[2] Statistical Models and Predictive Models have a different accepted rate of inaccuracy following the submission of the SLI. Statistical Model predictions must be reverted to unknown more than 1% of the service lines are inaccurate (lead), while predictive model prediction must be reverted back to unknown status if more than 5% of service line classifications are inaccurate.

- Since PWS may have their submission rejected if 1% or more of the service lines in their SLI are discovered to be lead or GRR, PWS may prefer to err on the side of caution and identify more service lines before using this method and create a more statistically significant model (ex: 99% confidence level). PWS should discuss this concern with their contractors when considering this process.

Training your Predictive Model

Training your Model

PWS models should first be trained by using a pool of service lines where all materials are known. PWS should use 80% of the service lines with known materials to train the model and test the model by having the model predict the material of the remaining 20%[3] of service lines (SLs) that PWS have already identified. See Appendix A, Table A for the required number of service lines that must be included in your initial investigation pool.

Using Historical Records

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, if there is concern over the accuracy of records, records included in the model should be verified through other methods, **such as field inspections or operator knowledge.**

Preventing Biases in the Model

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 may perform well there but fall short elsewhere.

It is also possible to introduce biases when predicting service line materials by using only tie cards, building age, or building codes.

How to Prevent Biases

Neighborhood Bias

- Gather representative data to feed the model

[3] The 20% Testing Data is not included as part of the investigation pool of random samples, see Appendix A for more information.

- ~~Service line data of all expected materials.~~
  - ~~Service line data from multiple regions of the PWS service area. Tie Card Bias~~

- ~~Provide numerous inputs into the model.~~
- ~~Tie cards.~~
- ~~Building age~~
- ~~Customer self-identification~~
- ~~Construction codes~~

~~PWS are encouraged to discuss all concerns with their contractor and continue to train their model with multiple iterations to strengthen the results.~~

General Statistical/Predictive Model Verification Method Requirements ~~Summary~~
Please note that if any additional requirements are included in the rest of this document not included here, PWS must still follow those additional requirements. Please read this document carefully and ensure that all requirements and guidance are followed.
Predictive Model Requirements Only
~~Predictive Model Requirements Only~~ • PWS must use their own records for training, testing, and using a model, and cannot "borrow" data from another system at any point when using a statistical/predictive model. • All predictive models should ~~train the model~~be trained using an 80/20 ~~testing pattern~~model, meaning 20% of the known service lines should be held out to test the predictive model while training. The recommendation is to test a predictive model many times and choose the best version moving forward, then improve it with further identifications of service lines. • PWS must use MassDEP's defined thresholds below to define the material of their service line, or stricter thresholds.   • PWS using a predictive model will likely receive a percentage for each service line that will provide a likelihood of lead per service line. Service Lines with an 80% or higher likelihood of lead must be classified as "lead" in the SLI. Service lines with a 15% or lower likelihood of lead may be classified as "Unknown, definitely does not contain lead or galvanized" (UNK-NOLG). Service lines with a likelihood of lead between 15.01% and 79.99% must be categorized as "Unknown, may contain lead and/or galvanized" (UNK-LG).
Statistical Model Requirements Only

- PWS cannot use statistical modeling (without predictive modeling) as a verification method if the PWS has known lead service lines. PWS may not be approved to use statistical analysis if the PWS is aware of/expects multiple galvanized service lines to be found on the private side which could be GRR service lines.
- PWS must meet with MassDEP DWP and receive approval for their statistical analysis prior to beginning work.
- PWS will likely only be able to use statistical modeling as a verification method if there are **NO lead service lines discovered during the initial investigation.** If any lead service lines are found, systems must use another method to find the location of all expected lead service lines, whether that be predictive modeling or another approved method, or will be required to revert any predicted service lines back to UNK-LG.

  In the interest of protecting public health and to assist PWS to identify all service lines which are lead or contain lead, MassDEP has developed the following **Galvanized Requiring Replacement (GRR) acceptance limit and process** for PWS that have used statistical analysis as their verification method in accordance with MassDEP/DWP Statistical Analysis requirements in their Service Line Inventory (SLI): **A maximum of 2.5% of all service lines verified by statistical analysis, which must be less than or equal to 25 service lines, that could be GRR if the material is discovered to be a galvanized material is acceptable.** PWSs that do not meet this limit must reclassify all non-lead services that could be GRR under their analysis/model as UNK-LG for MassDEP/DWP approval.

Statistical and Predictive Model Requirements

- All Statistical analysis and predictive modeling must be approved by MassDEP DWP prior to work beginning. If PWS do not have this verification method approved by MassDEP DWP, they cannot use this method.
- 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 lines before using a statistical/predictive model.
- PWS should ensure their model is using all verified and accurate PWS records to train the model.
- PWS must use a random method to find the service lines included in the initial sampling/investigation pool[4]. See Appendix B for one method of doing so.
- Only 20% of the investigation pool of service lines must be verified by field inspection. This field inspection must have been conducted within the last 10 years.
- Models will require a confidence level of 95%.
- All PWS must provide a disclaimer with all public facing SLI related materials, including public notices and the public inventory, which follows the language stated in the ***Public Material Requirements*** section below.
- If more than *5% of predicted service line predictions* and **1% of statistical model predictions** are discovered to be inaccurate, all predicted service lines must revert to unknown status.
- MassDEP may also require PWS that use predictive or statistical modeling to submit a long-term compliance plan using other methods to confirm identification for all service lines initially identified by statistical or predictive modeling.

[4] **Investigation Pool:** This term refers to the service lines chosen randomly that must be identified, i.e. a sample group.

- MassDEP may ask PWSs to produce or submit identification records at any point. PWS should create, compile, and retain documentation of all service line identification efforts.
- PWS must ensure that all service lines verified by statistical analysis or predictive modeling are listed as verified by method "A", in the SLI (this option is available in the Verification Method column of the SLI Workbook). All service lines which were part of the initial investigations must be noted in the comments section.
- PWS are required to submit a statistical analysis/predictive modeling report, from the PWS and Contractor (if used), which details:
    - For Statistical Models:
        - a map of the investigation pool of service lines which were used in the model and
        - the statistical analysis used to develop the conclusions of the model.
    - For Predictive Models:
        - how the predictive model was created,
        - the service lines used to train the model,
        - how these service lines were identified to be used in the training set, and
        - information on the training results and confidence interval.

~~**Statistical and Predictive Model Requirements**~~

~~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 lines before using a statistical/predictive model.~~

~~PWS should ensure their model is using all verified and accurate PWS records to train the model.~~

~~PWS must use a random method to find the service lines included in the initial sampling/investigation pool. This investigation pool must meet the amounts provided in Appendix A, Table A.~~

~~Only 20% of the investigation pool of service lines must have their verification method be by field inspection. This field inspection must have been conducted within the last 10 years.~~

~~Models will require a confidence level of 95%.~~

~~PWS are required to submit a report with their service line inventory which includes details listed above regarding their statistical/predictive model.~~

~~All PWS must provide a disclaimer with all public facing SLI related materials, including public notices and the public inventory, which follows the language stated above.~~

~~If more than *5% of predictive model service line predictions* and 1% of statistical model predictions are discovered to be inaccurate, all predicted service lines must revert to unknown status.~~

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

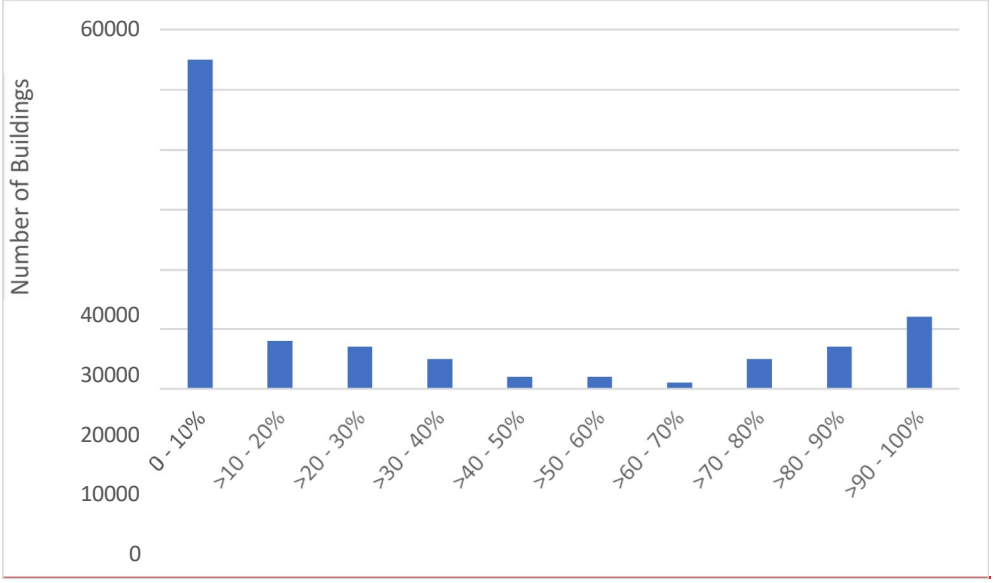
~~MassDEP may ask PWSs to produce or submit identification records at any point. PWS should create, compile, and retain documentation of all service line identification efforts.~~

~~**Record**ng Predictive Modeling as your Verification Method in the Service Line Inventory CSV File~~

When data entering the service lines predicted by your predictive model in your service line inventory, record the verification method as "statistical analysis" and the service line material as predicted by the model in the comments.

The predictive model will calculate a percentage for each service line, with the higher the percentage meaning a higher likelihood that a service line is lead.

See the graph below for an example of this.



Service Lines with an **80% or higher likelihood of lead** may be classified as "lead" in the SLI. Service lines with a **15% or lower likelihood of lead** may be classified as "Unknown, definitely does not contain lead or galvanized" (UNK-NOLG). Service lines with a likelihood of lead between 16% and 79% must be categorized as "Unknown, may contain lead and/or galvanized" (UNK-LG).

Submitting your Inventory

PWS must submit the SLI by the October 16th, 2024 deadline and remember to select statistical analysis as the verification method for the service line materials that were predicted using the model in the SLI certification form[5].

[5] Service Line Inventory (SLI) Certification Forms will be distributed to PWS after they have submitted their SLI, and the SLI has been validated by MassDEP staff.

~~Along with the SLI CSV File and SLI Certification Form, PWS are required to submit a report, from the PWS and Contractor (if used), which details:~~

~~For Statistical Models:~~

- ~~a map of the investigation pool of service lines which were used in the model and~~
- ~~the statistical analysis used to come to the conclusion of the model.~~

~~For Predictive Models:~~

- ~~how the predictive model was created,~~
- ~~a map of the investigation pool of service lines used to train the model,~~
- ~~how these service lines were identified for inclusion in the training set, and~~
- ~~information on the training results and confidence interval.~~

Public Material Requirements

- All PWS must provide a disclaimer with their public inventory that states: **"This Service Line Inventory was created with the use of Statistical/Predictive Modeling to predict and identify the material of unknown service lines."**
- All PWS must provide a disclaimer with all LCRR Lead Service Line Notices that states **"Your home is served by a lead service line confirmed through the use of Predictive Modeling"**; the letters provided to consumers must also ~~include the~~list the ~~percentage~~ % likelihood of lead presented by the model. See below.
- ~~All PWS must provide a disclaimer with all LCRR Unknown Service Line Notices that states: **"Through use of predictive modeling, your service line is has over a [percentage] .**~~
- **likelihood~~y~~ of being lead."**. The percentages provided in these letters may be exact percentage found for each service line, or within the ranges of the likelihood of lead provided below:

- ~~**20%-30%**~~
- ~~**31%-40%**~~
- ~~**41%-50%**~~
- ~~**51%-60%**~~
- ~~**61%-70%**~~
- ~~**71%-80%**~~
  - 15.01%-19.99%
  - 20%-30.99%
  - 31%-40.99%
  - 41%-50.99%
  - 51%-60.99%
  - 61%-70.99%
  - 71%-79.99%

Verifying ~~the~~ Predicted Service Lines

Over time, during routine operations, PWS must verify the predicted materials, update the service line inventory, and submit corrections as required by the LCRR/~~LCRI~~. If more than 5%[~~6~~7] of service line predictions made by the predictive model[8] are discovered to be inaccurate, MassDEP DWP may require that all predicted service lines revert to unknown status. When required, PWS must re-run its predictive model with new verified information to improve the accuracy of the model as service lines are identified.

MassDEP DWP may require PWS to create a plan to identify all service lines in their inventory that were previously identified using a statistical or predictive model within a time frame determined by MassDEP DWP.

Retaining Identification Records.

MassDEP may ask PWSs to produce or submit identification records at any point. PWS should create, compile, and retain documentation of all service line identification efforts.

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

~~**Statistical/Predictive Model Verification Method Requirements Summary**~~

~~**Predictive Model Requirements Only**~~

- ~~PWS must use their own records for training, testing, and using a model, and cannot "borrow" data from another system at any point when using a statistical/predictive model.~~
- ~~All predictive models should train the model using an 80/20 testing pattern.~~
- ~~PWS must use MassDEP's defined thresholds to define the material of their service line, or stricter thresholds.~~

[6] Because Predictive Models usually classify service lines only as lead or non-lead, if a service line predicted to be lead is discovered to be GRR, this is not counted towards this inaccurate total. This number should, however, be noted in future reports to MassDEP for reference.

[~~7~~] ~~Because Predictive Models usually classify service lines only as lead or non-lead, if a service line predicted to be lead is discovered as GRR, this is not counted towards this inaccurate total. This number should, however, be noted in future reports to MassDEP for reference.~~

[~~8~~] ~~Statistical Models and Predictive Models have a different accepted rate of inaccuracy following the submission of the SLI. Statistical Model predictions must be reverted to unknown more than 1% of the service lines are inaccurate (lead), while predictive model prediction must be reverted back to unknown status if more than 5% of service line classifications are inaccurate.~~

~~**Statistical and Predictive Model Requirements**~~

- ~~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 lines before using a statistical/predictive model.~~
- ~~PWS should ensure their model is using all verified and accurate PWS records to train the model.~~
- ~~PWS must use a random method to find the service lines included in the initial sampling/investigation pool. This investigation pool must meet the amounts provided in Appendix A, Table A.~~
- ~~Only 20% of the investigation pool of service lines must have their verification method be by field inspection. This field inspection must have been conducted within the last 10 years.~~
- ~~Models will require a confidence level of 95%.~~
- ~~PWS are required to submit a report with their service line inventory which includes details listed above regarding their statistical/predictive model.~~
- ~~All PWS must provide a disclaimer with all public facing SLI related materials, including public notices and the public inventory, which follows the language stated above.~~
- ~~If more than *5% of predictive model service line predictions* and **1% of statistical model predictions** are discovered to be inaccurate, all predicted service lines must revert to unknown status.~~
- ~~MassDEP may also require PWS that use predictive or statistical modeling to submit a long-term compliance plan using other methods to confirm the identification of all service lines initially identified by statistical or predictive modeling.~~
- ~~MassDEP may ask PWSs to produce or submit identification records at any point. PWS should create, compile, and retain documentation of all service line identification efforts.~~

Questions and Answers (Q&A) about MassDEP DWP Statistical Analysis and Predictive Modeling Requirements

Q. Does MassDEP DWP allow PWS to do different levels of models for statistical analysis and predictive modeling?

A. Yes, MassDEP DWP allows the ~~There are Three Types of Statistical/Predictive Models:~~following levels of models:

- System Wide Level
- Neighborhood Wide Level
- Water Main/block level

Systems are reminded to discuss with their contractor/persons completing the analysis/model what the best model may be for their service area, and which has the most representative results. Systems with lead congregated in certain areas of the service area/town may benefit from a neighborhood level approach, to focus on areas with a higher likelihood of lead, while others may prefer a system wide level.

PWS should discuss the planned procedure for their model with MassDEP DWP, should they plan to create an analysis or model, when meeting with MassDEP DWP to discuss analysis/model approval.

Q. How should PWS account for possible biases in their predictive model?

A. 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 neighborhood, the model may perform well there but fall short elsewhere.

It is also possible to introduce biases when predicting service line materials by using only tie cards, or only housing age, or building codes.

PWS should plan to address biases by doing the following:

To avoid Neighborhood Bias:

- Gather representative data to feed the model
  - Service line data of all expected materials.
  - Service line data from multiple regions of the PWS service area.

To avoid Tie Card Bias:

- Provide numerous inputs into the model.
- Tie cards.
- Building age
- Customer self-identification
- Construction codes

PWS are encouraged to discuss all concerns with their contractor and continue to train their model with multiple iterations to strengthen the results.

Q. Can a PWS use their entire service area to pull from their investigation pool?

A. Yes, PWS may use their entire service area, known and unknown service lines, to pull from for their investigation pool. This can then allow PWS to use service lines where a service line material is already known, instead of requiring immediate field inspections. However, this pool of service lines must meet the required numbers in Appendix A, Table A to be **statistically significant.**

Q. How is the GRR acceptance limit calculated, i.e., what is the GRR acceptance process?

A. This acceptance limit is generated for PWS by doing the following:

1. Accounting for the total number of service lines inspected/included in your pool of randomly chosen service lines, which was used to create your statistical model **(Total # of Service Lines Identified to Create your Statistical Model).**

2. We then review the information included in your report and SLI to calculate the total number of galvanized service lines found during inspections. If this number is not provided in the report, it is calculated based on the number of Field Inspected service lines included in the SLI, and the total number of Galvanized service lines discovered with this verification method **(Total # of Private Galvanized Service Lines Found During Inspections)** *(If this information is unclear, MassDEP DWP can reach out for further clarification).*
3. We then divide the total number of galvanized service lines found by the total number of inspections**: (Total # of Private Galvanized Service Lines Found During Inspections) / (Total # of Service Lines Identified to Create your Statistical Model) = GRR %. GRR %** is your PWSs percentage (%) of the Highest Number of Estimated GRRs, it must be equal to or lower than the acceptance limit of 2.5% to be accepted.
4. We then use that percentage and multiple it by the total number of service lines in your SLI which could be GRR if a galvanized service line was found on the private side **(Total Number of Service Lines which Could Be GRR).** Meaning, the total number of service lines which have been classified as NON-LEAD due to statistical analysis, which are: UNK-LG on the public side, were lead previously, or it is unknown if they were ever lead previously. **GRR % * (Total Number of Service Lines which Could Be GRR) = GRR #. GRR #** is the highest estimated number of GRR service lines we could expect in your PWS. This number must be 25 or lower to be accepted by the GRR acceptance limit and process.
5. We then compare these numbers (**GRR % and GRR #)** to the acceptance **limit (2.5% & 25).** If either of the numbers are above the acceptance limit, the analysis cannot be accepted at this time.

Equations:

$$\frac{\textbf{Total \# of Private Galvanized Service Lines Found During Inspections}}{\textbf{Total \# of Service Lines Identified to Create your Statistical Model}} = \boldsymbol{GRR}\ \%$$

$$\boldsymbol{GRR}\ \% * \textbf{ Total Number of Service Lines which Could Be GRR } = \textbf{ GRR \#}$$

(GRR % and GRR #) is then compared to the acceptance **limit (2.5% & 25)**

For example:

- The system has **500** service line inspections for their statistical analysis**. (Total # of Service Lines Identified to Create your Statistical Model).)**
- The system found **15** galvanized service lines during these inspections. **(Total # of Private Galvanized Service Lines Found During Inspections)**
- Total service lines that are predicted through statistical analysis and could be GRR if the private service line was galvanized is **2,000** service lines. **(Total Number of Service Lines which Could Be GRR)**
- 15 / 500 = 0.03 or **3% *(3% = GRR %)***
- 2,000 * 0.03 = **60 *(60 = GRR #)***
- This systems number and percentage of possible estimated GRR service lines is **60 & 3%,** compared to the GRR acceptance limit of **25 and 2.5%,** this PWSs analysis cannot be accepted.

Appendix A: Creating an Investigation Pool of Service Lines

To use a statistical model, PWS must have a predetermined amount of verified service lines in their service area. See the requirements below:

- PWSs with fewer than 1,500 unknown service lines must have an investigation pool of at least 20 percent of their total number of service lines, which may include known and unknown service lines.
- PWSs with more than 1,500 unknown service lines must have an investigation pool with enough lines to reach a minimum 95 percent confidence level. This investigation pool may include known and unknown service lines that must all be identified before continuing with a statistical model. See *Table A* to determine the number of service lines required. Table A uses a confidence level of 95 percent.

Selecting the Service Lines to Include in an Inspection Pool

Randomly select service lines for physical inspection.

- Compile a list of all service lines (known and unknown) in your PWS service area.
- Your selection must be uniformly random and not based on any specific criteria which can introduce bias. In other words, each 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.

Verify All Unknowns in the Investigation Pool

PWS may use other verification methods for this method, however, PWS must use methods they believe are valid, and use records that are likely to be accurate. MassDEP recommends PWS use field inspections whenever possible, as it is the most accurate verification method.

Field Inspection Reminders

- When performing field inspections, 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.[9]
- 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.

Record results of the physical inspection process.

- The PWS should record the results for their investigation pool using their own SLI database or the MassDEP SLI Excel Workbook see https://www.mass.gov/lists/lead-copper- forms-and-templates#lead-&-copper-rule-revisions-(lcrr)- for more information. If using the MassDEP SLI workbook, in the dropdown list, enter the service line material observed at each point. The spreadsheet will automatically categorize the entire service line into one of the many categories that align with the required EPA categories; lead, non-lead, galvanized requiring replacement and unknown.

[9] 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).

Table A. Minimum number of service lines requiring verification.

This table refers to a PWS creating a random sample of service lines from their entire service area. If a PWS is only using a model for a section of the service area or will be creating a model by only finding a sample/investigation pool from their unknowns, use that number in the lefthand column instead of your total service lines.

Service Lines in Service Area	Number of Required Service Lines to be Verified *	Minimum number of known service lines[10] required to test your predictive model during the training process. (20% Testing Pool)**
Fewer than 1,500	20% of service lines	5% of service lines
1,500	306	75 - 80
1,600	310	
1,700	314	
1,800	317	
1,900	320	80 - 85
2,000	322	
2,200	327	
2,400	331	
2,600	335	
2,800	338	85 - 90
3,000	341	
3,500	346	
4,000	351	
4,500	354	
5,000	357	
6,000	361	90 - 95
7,000	364	
8,000	367	
9,000	368	
10,000	370	
15,000	375	
20,000	377	
30,000	379	
40,000	381	95 - 100
60,000	382	
90,000	383	
225,000 or more	384	

Table adapted from Oregon Health Authority: Statistical Guidance for Evaluating Unknown Service Lines.

*The number of service lines that must be physically inspected is based on the required number to meet a 95% confidence interval. MassDEP recommends that PWS inspect/verify as many service lines as possible, meeting this number and going beyond, to improve the accuracy of your statistical and/or predictive model.

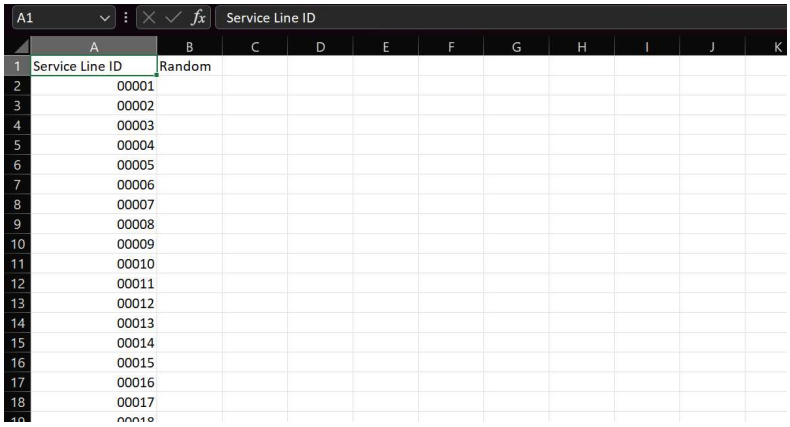
** This Column refers to the number of service lines that must be identified, and used to test the model. This is the 20% of the 80/20 model required by MassDEP. The 20% is a number of service lines **separate** from the number of service lines that must be identified in the investigation pool.

[10] The service lines used to test the model do not need to be chosen randomly, unlike the service lines in the investigation pool ~~must be~~.

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 service lines for verification using the following Microsoft Excel steps (the same formulas and method work for Google Sheets):

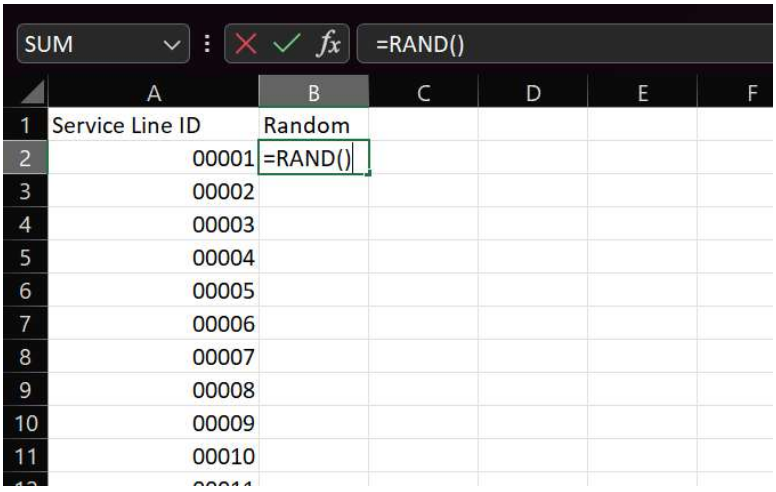
1. In the first column of a spreadsheet, list every unique service line. They can be listed by address, service line ID, or other identification method.



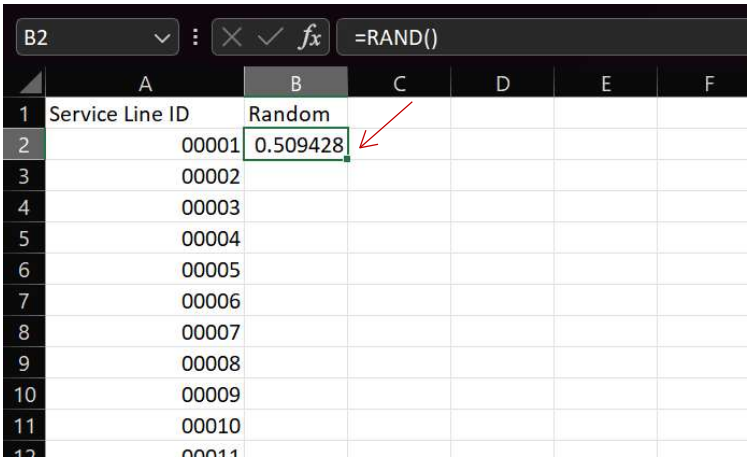
2. In the second column, generate uniformly random numbers, so that each service line is associated with a randomly generated number.

Follow these steps:

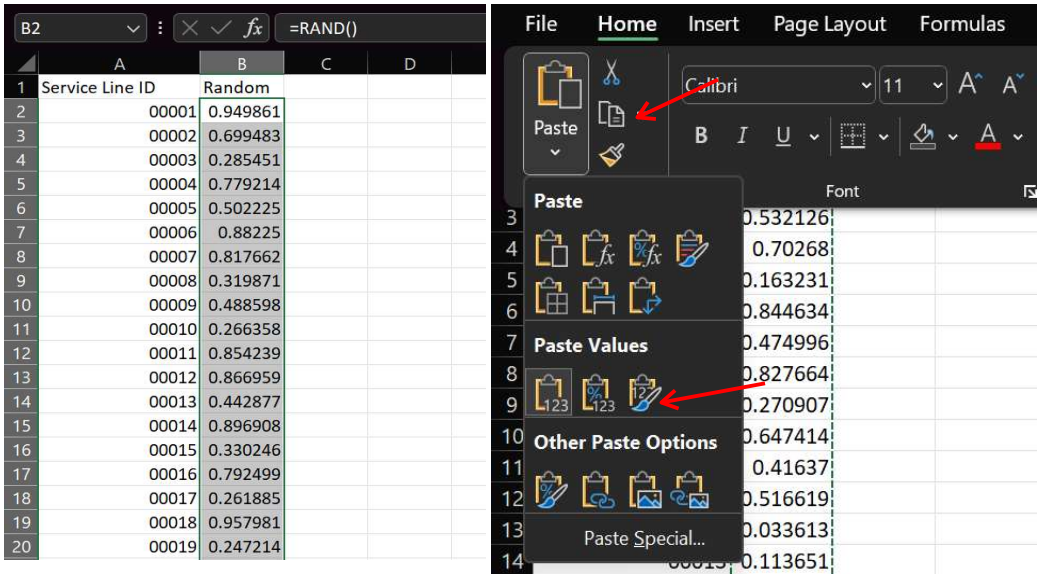
a. Enter the formula =RAND() into the first cell of the second column next to the first service line location and press Enter. This generates a number between 0 and 1 for each service line.



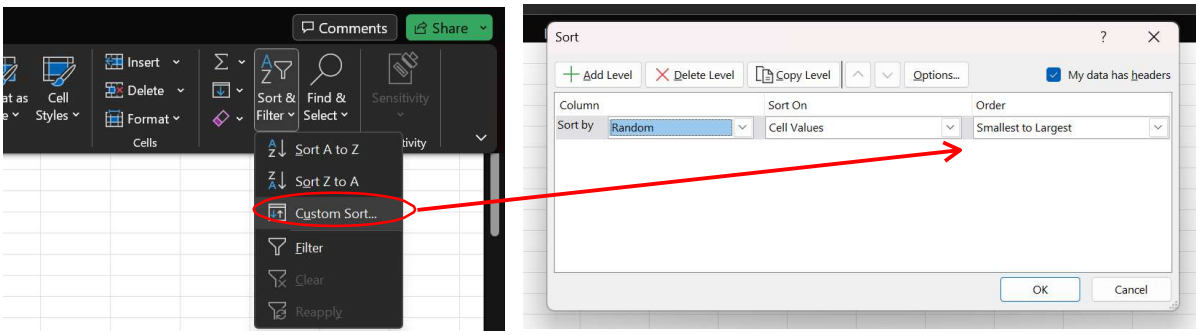
b. Select the lower right corner of the first cell in the second column (the column with the random value) and double click the small square to copy the formula into the cells below it so that every service line location is assigned a random number.



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.



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.



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 ~~identified~~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=q8fU001P2lI for generating random samples on Microsoft Excel.