A Machine Learning Approach to Identity Houses with High Lead Tap Water Concentrations

Seyedsaeed Hajiseyedjavadi[1], **Michael Blackhurst**[2], **Hassan A. Karimi**[1]

{ [1,] School of Computing and Information, [2] Center for Social and Urban Research}, University of Pittsburgh, USA
{seh138, mfb30, hkarimi}@pitt.edu

Abstract

Over a century separates initial lead service lateral installations from the federal regulation of lead in drinking water. As such, municipalities often do not have adequate information describing installations of lead plumbing. Municipalities thus face challenges such as reducing exposure to lead in drinking water, spreading scarce resources for gathering information, adopting short-term protection measures (e.g., providing filters), and developing longer-term prevention strategies (e.g., replacing lead laterals). Given the spatial and temporal patterns to properties, machine learning is seen as a useful tool to reduce uncertainty in decision making by authorities when addressing lead in water. The Pittsburgh Water and Sewer Authority (PWSA) is currently addressing these challenges in Pittsburgh and this paper describes the development and application of a model predicting high tap water concentrations (> 15 ppb) for PWSA customers. The model was developed using spatial cross validation to support PWSA's interest in applying predictions in areas without training data. The model's AUROC is 71.6% and primarily relies on publicly available property tax assessment data and indicators of lateral material collected by PWSA as they meet regulatory requirements.

Introduction

Exposure to lead can cause serious health effects, particularly in children. Lead in paint, soil, and water are the primary sources of exposure risks (Gould 2009).

By 1900, lead was a predominant material used for residential water laterals in U.S. cities owing to its ease of installation and resistance to corrosion. The water later–or water service pipe–connects a building's interior plumbing to the main water system. As awareness of lead's negative public health effects grew, installations of lead laterals were eventually phased out. In 1986, federal regulations were revised to ban the use of lead plumbing (US EPA 2013)

However, many homes are still served by lead water laterals, which are the primary source of lead in drinking water in the U.S. As a result, federal guidelines require public water municipalities to monitor for lead concentration in their customers' taps. When lead tap water concentrations are elevated and subsequent treatment changes cannot control lead levels, municipalities are mandated to replace lead laterals until consecutive samples fall below federally accepted levels (US EPA 2015)

Municipalities can struggle to initiative effective replacement strategies due to missing information describing initial lead lateral installations and subsequent replacements. In the short-run, municipalities often provide customers with point-of-use protection (bottled water, water filters) and initiate lateral material inventory efforts, such as excavating laterals or inspecting them through curb boxes (Enking 2019; Ingber 2019)

Recent events in Pittsburgh, PA, demonstrate these challenges. The Pittsburgh Water and Sewer Authority (PWSA) provides drinking water to nearly all of the City of Pittsburgh, including approximately 70,000 residential customers. In 2016, tap water samples collected by the PWSA exceeded federal and state action levels for lead. PWSA thus entered a replacement mandate in June 2016 and is still operating under this mandate. In addition to replacements, PWSA recently implemented the use of orthophosphate as a corrosion control, and currently provides free water filters to affected residents (Shoemaker 2018) and has started building an inventory of water lateral materials through digitizing historical records, inspecting curb boxes, and excavations.

The inventory of service line materials is critical in providing a cost-effective replacement program. Excavating a service line and determining that the materials are not lead (or galvanized iron which is treated the same as lead at PWSA) is costly and diverts resources from the important work of actually replacing lead service lines. Therefore, having an accurate inventory is an important first-step in any lead service line replacement program. New approaches are needed to accurately predict the locations and numbers of lead service lines in a community.

Public concern and understanding about lead in drinking water grew from the Flint, MI, water crisis, which started in 2014. In coordination with state agencies, local municipal water providers in Flint switched water supplies from a corrosion-controlled source to a non-corrosion controlled source without making appropriate treatment changes (Olson et al. 2017). Since corrosion control limits the leaching of lead from lead laterals, these changes caused tap water levels to spike.

The historical nature of lead lateral installations, expected property development patterns, and influence of socioeconomics on replacements suggest that predictions of lead exposure may be feasible with appropriate training data.

After Flint's water crisis, a team of researchers from the university of Michigan addressed the problem of predicting houses with elevated lead level in their drinking water supply (Abernethy et al. 2016; Chojnacki et al. 2017). Using a wide range of data sources as their model features, they employed different classifiers for the prediction. In their work, they used traditional K-fold cross validation to estimate the performance of their models in face of real-world test set. Despite providing valuable predictions, we presume that their reported performances could be overestimated for samples located outside of areas for which training data were available. Essentially, spatial data require particular methods for cross validation to avoid non-independence associated with spatial proximity of the samples. Models that are solely relying on conventional cross validation may select less than ideal hyperparameters and underperform when predicting in areas unrepresentative of the training data.

In this work, we develop and apply models that predict houses with tap water concentration of lead that exceed 15 parts per billion (ppb), which is the federal action level for lead. To achieve highly accurate prediction, we meticulously interpolated missing data, balanced the data set, and pruned weak predictors. We also utilize spatial cross validation to find ideal hyperparameters as well as a more reliable technique to estimate the performance of our model. Eventually, we provide a ranking list of the most important features contributed in prediction of the response variable.

These models can be useful in balancing resources allocated for collecting better information on lateral material, and replacing lead laterals.

Data Sources and Limitations

Historical or observed lateral material, lateral installation and inspection dates, customer-provided tap water concentrations, and administrative customer data were provided by PWSA[1]. These data were merged with property assessment data–which describes household, use, and lot characteristics—collected by Allegheny County, PA.

Lateral Installation History

The other source of data is digitized historical data describing service laterals. This includes location of available data, description of laterals by their diameter, inspection results (date and materials), installed material and date, and a notes field. However, as these data has not perfectly maintained, many records are missing and there are inaccuracies.

Curb Box Inspection

In order of increasing expected accuracy, indicators of lateral material include historical records, curb box inspections (CBI), and materials recorded when laterals are replaced. Historical data reflect material indicators on paper maps prepared when laterals were first installed. CBI, which we refer to as "camera inspections" herein, reflects judgments made from photographs taken of laterals through curb boxes, which contain the main shut-off valve located near the street curb. Cross-referencing the materials indicated from CBI's and excavations, PWSA found a 97% true positive rate and 72% true negative rate.

The CBI program was conducted at about 17,500 locations in 2017 and 2018, with results frequently released to the public via an online lead service line map. The CBI data include binary indicators of lead or non-lead. The historical and excavation data specify the non-lead material (e.g., copper). We mapped the material indicators to binary values of 1 or 0 corresponding to "lead" or "non-lead" values. However, despite being a valuable source of information, a considerable portion of CBI data source is incomplete. In most cases, CBI inspectors could not locate the curb box, the curb box was located in an area that could not be accessed, or the service line material could not be identified during inspection. As a result, the total number of curb boxes inspected with detected line material is only 5,600 out of about 70,000 PWSA's customers.

Water Quality Test

Public water municipalities are required to collect tap water lead concentrations in compliance with federal regulations. Like many water municipalities, PWSA provides free certified lead testing for any customers electing to provide a sample, leading to significantly more samples than required for compliance. The raw dataset includes

[1] Datasets are provided to the University of Pittsburgh through a non-disclosure agreement signed with the Pittsburgh Water and Sewer Authority.

7,409 samples covering 5,634 homes. For homes with multiple samples, we used the maximum value.

Property Assessment

Allegheny county property assessments include 86 fields describing information used for the purposes of taxing real estimation. Example data includes property use (e.g., single family residential), age, floor space (for residential properties only), assessed value, building quality, architecture style, and lot size.

After merging this dataset with the customers dataset, the portion of the missing values of property assessment fields ranges between less than 1% to about 7%.

Sample Preparation

In this section, we present a detailed description of the steps we took to prepare a complete and relatively balanced dataset containing the most predictive features.

Handling Imbalanced Labels

One of the issues of our dataset is the imbalanced ratio of samples representing both classes which reduce stability and performance of a classifier. To be exact, the portion of water samples with the lead level below the federal action level of 15 ppb is 91.38%, while only 8.62% of samples show values of 15 ppb and above. In such an imbalanced sampling, a classifier is inclined to represent the majority class (below 15 ppb). To address this issue, we employed the SMOTE technique as a state-of-the-art approach to artificially increase the number of samples representing the minority classes (Chawla et al. 2002). This technique artificially generates samples of the minority class in the proximity of the existing feature space so that the ratio of majority and minority classes becomes less imbalanced. Note that to keep the evaluation set in accordance to real-world test set, we only synthetically balance the training set.

Handling Missing Values

Like many real-world problems a substantial portion of our data have missing values. With incomplete dataset the usability of predictive model would be very limited to those customers that have non-missing values for all the selected features. This means that without imputing the missing values, our model is only applicable to less than 8% of the customers.

Table 1 indicates a positive spatial autocorrelation in the spatial structure of the CBI data as measured by Moran's I statistics with queen contiguity-based spatial weights (Moran 1950). Based upon these results, we used the Inverse Distance Weighting (IDW) interpolation method to spatially estimate the missing values from the known ones (Myers 1994). We specifically interpolated the missing values of two important features: CBI (both public and private as independent observations) and the year the property was built. There are other features with less than 2% missing values. We replaced these with the mode or the median of the corresponding non-missing values.

Table 1: Moran's I calculated by Monte Carlo simulations over a thousand permutations for three independent variables

Feature	Moran's I	P-value
CBI (public side)	0.124	< 0.001
CBI (private side)	0.122	< 0.001
Origin year of the construction	0.667	< 0.001

Figure 1 shows the estimation of the missing values of CBI data for the public lateral, where values closer to 0 or 1 indicate higher or lower probability, respectively, of lead in the material of the public side of the service line.

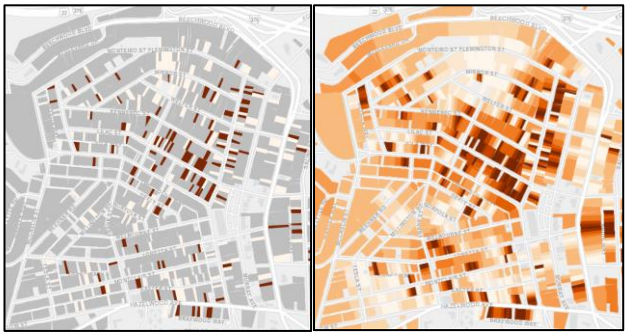

Figure 1: Example results of interpolating missing curb box inspections using the IDW technique with a power of 2. Training data are on the left and interpolations are on the right. Results are for the public portion of water laterals. Darker and lighter color indicates a higher or lower probability of lead, respectively.

Feature Selection

Starting with more than 150 features, we used Recursive Feature Elimination (RFE) (Guyon et al. 2002) as an effective feature selection technique, where at each step recursively remove features with the weakest predictive power, until no feature is left. The importance of the predictors is calculated at each iteration so that eventually, RFE selects and returns a subset of features with the highest predictive power.

In our work, we paired RFE with random forest (Tin Kam Ho 1995) to reduce our feature space from more than 150 to 16 features by pruning the weakest predictors.

Prediction

We developed a set of predictive models to classify PWSA's customers based on the observed lead level in their drinking water. The ultimate goal was to train effective classifiers to predict houses having drinking water levels equal or above the federal action level of 15 ppb. Observations were labeled as either equal or greater than 15 ppb (positive class) or less than 15 ppb (negative class).

Our training sample consisted of 429 positive and 4,552 negative samples for training the models to predict tap water levels for more than 60,000 customers.

Spatial Cross Validation

To evaluate model performance, we used the well-known statistical resampling procedure known as K-fold cross validation. In this procedure, the training set is split *randomly* into K subsets (i.e., folds) in such a way that a fold is put aside for the validation set and the model is learned using the remaining K-1 folds. The process is repeated K times so that every fold is used once as the validation set.

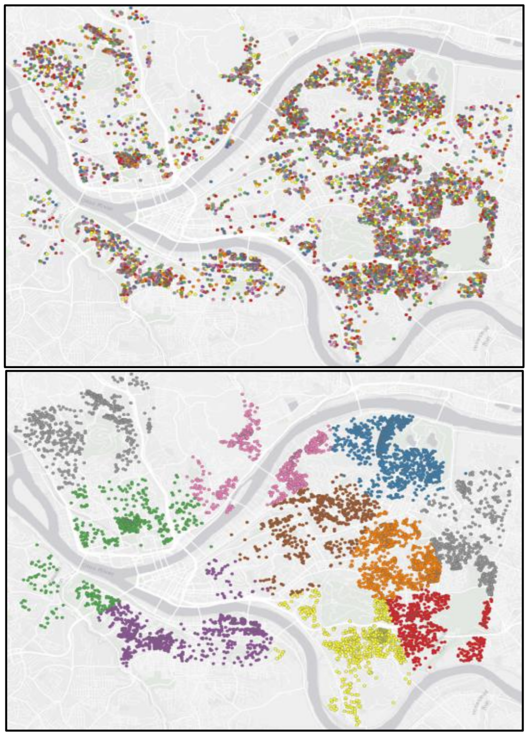

Figure 2: The difference between non-spatial (top) and spatial (down) cross validations. Points with the same color is to be considered as the validation fold for each iteration of the validation.

While cross validation works well for non-spatial data, it is shown that it leads to overestimation of the performance of the predictive model and inappropriate model selection for spatial data (Roberts et al. 2017; Brenning 2012). In traditional cross validation, due to the randomness of the partitioning procedure, it is highly likely that spatially close samples be divided into test and train subsets. Since spatially proximate observations are nearer each other, we can infer that in contrast to the primary objective of out-of-sample validation, the test and training samples are not statistically independent (Miller 2004)

To produce more robust and realistic estimation of the predictive performance of the models, we employed spatial cross validation. In spatial cross validation, samples are partitioned based on their geographical coordinates so that the dependency of samples become minimum.

We used K-means clustering to split the training set into K clusters and similar to K-fold cross validation, at each iteration one cluster was used for validation and the rest for training the predictive model. In Figure 2 both spatial and non-spatial cross validations are illustrated.

Predictive Models

To compare and select the best model with the highest predictive power, we deployed a set of state-of-the-art classifiers. Our goal was to explore a wide range of classifiers to identify the finest classifier with the highest predictive power for the problem.

For each observation, we reduced the set of features to 14 as it is described in feature selection section. All the deployed models, except the logistic regression, require turning hyperparameters. The best hyperparameters for each model were chosen by the highest AUROC (area under the receiver operating characteristics) score that we obtained via grid search. The AUROC was measured through validating each model by the spatial cross validation explained in the previous section. Table 2 shows a summary of the best settings for each model.

Table 2: Deployed classifiers with the best hyperparameters tuned via grid search.

Classifier	Hyperparameters
Artificial Neural Network	#Hidden Layer=1, #Neurons= 7
K-Nearest Neighborhood	K=187
GBM	Depth = 8, #Trees=17, Shrinkage = 0.1
SVM	Radial Kernel, Sigma=0.005, C=0.25
Random Forrest	#Randomly Selected Predictors=11
Logistic Regression	-

Results

In this section, we present the performance of the deployed models. We tested and evaluated the models by both non-spatial K-fold cross validation and K-means spatial validation and set the parameter *K* for both validations as 10. Table 3 shows the AUROC acquired for both validations.

As it can be seen in Table 3, among the classifiers we deployed, GBM (Friedman 2001) outperforms the other models with the AUROC of 0.716 for spatial validation and 0.744 for non-spatial validation.

Table 3: AUROCs for the deployed classifiers for both spatial and non-spatial cross validations.

Classifier	Spatial CV	Non-spatial CV
Logistic Regression	0.648	0.679
K-Nearest Neighborhood	0.652	0.687
Artificial Neural Network	0.651	0.711
SVM	0.667	0.730
Random Forests	0.714	0.733
GBM	**0.716**	**0.744**

It is worth mentioning tha the performances of all classifiers are lower when we evaluate them via spatial cross validation. This is due the fact that in spatial cross validation, a test set is far from, and thus independent of, the training set. For some classifiers like SVM and artificial neural networks the differences between AUROC of the two validation techniques are relatively higher than the other ones.

Predictors Importance

Figure 3 shows the top 10 features that the best model (i.e., GBM) identified as the most effective factors for the classification. These features and their corresponding importance scores are measured based on the procedure described in (Breiman 2001).

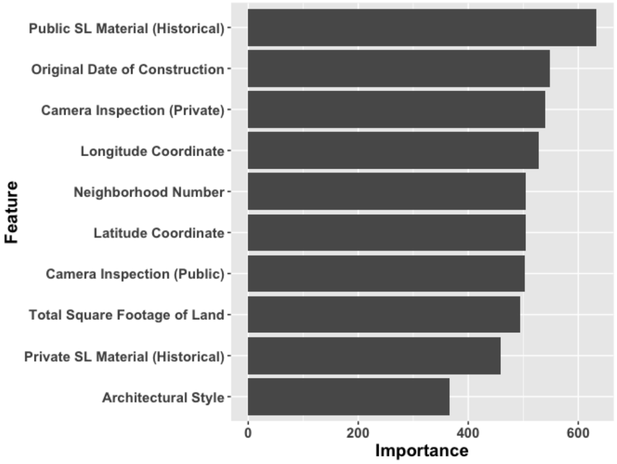

Figure 3: Top predictive features and their corresponding importance

As it can be seen in Figure 3, despite being outdated and unreliable, historical data of the material used in the public/private side of the service line are among the top important predictors for our model. In addition, camera inspections have a significant role in providing high accurate predictions as well. The other category of important predictors are the geographical features. The geographical coordinates and the neighborhoods of the places the samples were taken from show essential effects on the predictions. This illustrates how the high level of lead in drinking water could be concentrated in specific urban areas. In addition, property features, such as year the building was built and the architecture style of the building, are deemed important.

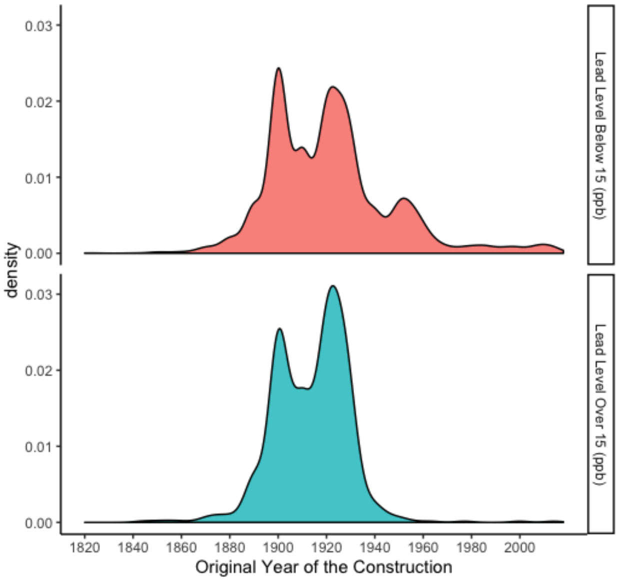

Figure 4: Distribution of the building year of the constructions over time for samples above and below 15 (ppb) lead level in tap water

In Figure 4, the distributions of the year that buildings were originally constructed with both positive and negative classes are illustrated. Similar to the findings of (Abernethy et al. 2016) we observe a strong relationship between high level of lead in tap water and the buildings built between 1910 to 1930. This could be due to the fact that these old lead lines laid in these places started to be corroded and as a result, high lead tap water concentration were observed in samples collected from these places.

Conclusion

To identify houses with tap water concentration of lead that exceed 15 (ppb), we deployed a set of classifiers trained on features collected from information that is either publicly available or collected in response to meeting regulatory requirements for lead in drinking water. We applied a series of data processing techniques to provide a more reliable and useful form of data for our predictive models. For evaluating our models, we used spatial cross validation as a reliable technique to estimate the performance of our model for data points located outside of the vicinity of the training samples.

We provide a list of features that are strong predictors of our best classifier. We found that geographical location, building characteristics, and indicators of lateral materials are among the top features.

As more and more lead service lines are detected and replaced by PWSA, we get a better access to a reliable source of information of the materials used in the service lines. For the next step, we plan to develop and deploy predictive models to exploit this valuable data source to identify houses with lead service lines to reduce the number of wrong excavations.

Acknowledgments

The authors would like to thank the Pittsburgh Water and Sewer Authority for their support of this work.

References

Abernethy, Jacob; Anderson, Cyrus; Dai, Chengyu; Farahi, Arya; Nguyen, Linh; Rauh, Adam; Schwartz, Eric; et al. 2016. Flint Water Crisis: Data-Driven Risk Assessment Via Residential Water Testing. https://arxiv.org/abs/1610.00580v1.

Breiman, Leo. 2001. Random Forests. Machine Learning 45(1): 5–32. https://doi.org/10.1023/A:1010933404324.

Brenning, A. 2012. Spatial Cross-Validation and Bootstrap for the Assessment of Prediction Rules in Remote Sensing: The R Package Sperrorest. In 2012 IEEE International Geoscience and Remote Sensing Symposium, 5372–75. Munich, Germany. https://doi.org/10.1109/IGARSS.2012.6352393.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research 16(June): 321–57. https://doi.org/10.1613/jair.953.

Chojnacki, Alex, Dai, Chengyu; Farahi, Arya; Shi, Guangsha; Webb, Jared; Zhang, Daniel T.; Abernethy, Jacob; and Schwartz, Eric. 2017. A Data Science Approach to Understanding Residential Water Contamination in Flint. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1407–16. New York: ACM. https://doi.org/10.1145/3097983.3098078.

Enking, Molly. 2019. From Flint to Newark to Pittsburgh: Why Do American Cities Fail to Protect Our Water? Rolling Stone (blog). https://www.rollingstone.com/politics/politics-news/flint-newark-pittsburgh-lead-in-drinking-water-873584/.

Friedman, Jerome H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics 29(5): 1189–1232.

Gould, Elise. 2009. Childhood Lead Poisoning: Conservative Estimates of the Social and Economic Benefits of Lead Hazard Control. Environmental Health Perspectives 117(7): 1162–67. https://doi.org/10.1289/ehp.0800408.

Guyon, Isabelle; Weston, Jason; Barnhill, Stephen; and Vapnik, Vladimir. 2002. Gene Selection for Cancer Classification Using Support Vector Machines. Machine Learning 46(1): 389–422. https://doi.org/10.1023/A:1012487302797.

Ingber, Sasha. 2019. Newark's Drinking Water Problem: Lead and Unreliable Filters. NPR.Org. https://www.npr.org/2019/08/13/750806632/newarks-drinking-water-problem-lead-and-unreliable-filters.

Miller, Harvey J. 2004. Tobler's First Law and Spatial Analysis. Annals of the Association of American Geographers 94(2): 284–89. https://doi.org/10.1111/j.1467-8306.2004.09402005.x.

Moran, P. A. P. 1950. Notes on Continuous Stochastic Phenomena. Biometrika 37(1/2): 17–23. https://doi.org/10.2307/2332142.

Myers, Donald E. 1994. Spatial Interpolation: An Overview. Geoderma 62(1): 17–28. https://doi.org/10.1016/0016-7061(94)90025-6.

Olson, Terese M.; Wax, Madeleine; Yonts, James; Heidecorn, Keith; Haig, Sarah-Jane; Yeoman, David; Hayes, Zachary; Raskin, Lutgarde; and Ellis, Brian R. 2017. Forensic Estimates of Lead Release from Lead Service Lines during the Water Crisis in Flint, Michigan. Environmental Science & Technology Letters 4(9): 356–61. https://doi.org/10.1021/acs.estlett.7b00226.

Roberts, David R.; Bahn, Volker; Ciuti, Simone; Boyce, Mark S.; Elith, Jane; Guillera-Arroita, Gurutzeta; Hauenstein, Severin; et al. 2017. Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. Ecography 40(8): 913–29. https://doi.org/10.1111/ecog.02881.

Shoemaker, Dale. 2018. Test Results Show Lead Levels Falling in Pittsburgh amid Questions over PWSA's Future. PublicSource | News for a Better Pittsburgh. https://www.publicsource.org/test-results-show-lead-levels-falling-in-pittsburgh-amid-questions-over-pwsas-future/.

Ho, Tin Kam. 1995. Random Decision Forests. In Proceedings of 3rd International Conference on Document Analysis and Recognition, 278–82. Montreal, Quebec, Canada. https://doi.org/10.1109/ICDAR.1995.598994.

US EPA, OA. 2013. Summary of the Safe Drinking Water Act. Overviews and Factsheets. US EPA. https://www.epa.gov/laws-regulations/summary-safe-drinking-water-act.

US EPA, OW. 2015. Lead and Copper Rule. Policies and Guidance. US EPA. https://www.epa.gov/dwreginfo/lead-and-copper-rule.