‘Pb-Predict’

Using Machine Learning to Locate Lead Plumbing in a Large Public Water System

by

Raanan Sharohn Gurewitsch

Bachelor of Philosophy, University Honors College, 2019

Submitted to the Undergraduate Faculty of

University Honors College

of the requirements for the degree of

Bachelor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This thesis was presented

by

Raanan Sharohn Gurewitsch

It was defended on

April 5, 2019

and approved by

Dr. Pierre Goovaerts, PhD, BioMedware

Dr. Saumyadipta Pyne, PhD, Graduate School of Public Health

Michael Blackhurst, PhD, University Center for Urban and Social Research

Thesis Advisor: Dr. Hassan Karimi, School of Computing and Information

‘Pb Predict’: Using Machine Learning to Locate Lead Plumbing in a Large Public Water System

Raanan Sharohn Gurewitsch, BPhil

University of Pittsburgh, 2019

Struggling to respond to elevated lead levels in residential tap water, cities like Flint, MI and Pittsburgh, PA are undergoing large-scale efforts to remove the lead pipes that bring water service to their customers. However, limited geographic data on plumbing materials throughout housing stocks represents a logistical challenge for local authorities to locate and replace lead service lines. This study tests whether available geographic data on housing conditions and plumbing materials can effectively inform risk assessment and thus, expedite replacement programs and help prevent exposure to lead. To do so, we train and compare multiple types of machine learning classification algorithms to predict the presence or absence of lead service lines at properties in Pittsburgh. The results show that the probability of having a lead service line increases for houses built before 1930 and demonstrate the significance of parcel age, spatial proximity and other housing characteristics as predictive features for locating lead in water hazards. Accurate targeting of high-risk housing units may inform the strategy of decision-makers working to ensure that residents of aging American homes have safe drinking water. Therefore, the results are mapped to simulate the prevalence of lead service lines throughout the City of Pittsburgh and a framework for other cities is discussed.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

To this day, public water systems throughout the United States remain riddled with lead plumbing, an issue that gained significant attention throughout the country following the crisis in Flint, Michigan. The threat of harmful exposure to this invisible, tasteless neurotoxin, primarily to pregnant women and young children, is a serious public health concern with wide-ranging socioeconomic implications. Without appropriate corrosion control measures, the drinking water in a public system will leach lead from pipes, fixtures and solder, endangering those who consume it. According to the Centers for Disease Control and Prevention (CDC), blood lead levels under 10 µg/dL, which were previously considered safe, have been associated with behavioral issues, poor academic performance and learning disabilities in children (Advisory Committee on Childhood Lead Poisoning Prevention, 2012). Lead in drinking water, though less common a source of poisoning than lead-based paint or contaminated soil or dust, has been known as a cause of elevated blood lead levels in children since the 1980's (Shannon & Greaf, 1989; Cosgrove et al., 1989). Lead pipes, which were most commonly installed at American residences from the late 1800s until the 1930s, deliver water to homes in over 70% of cities with populations over 30,000 people (Troesken, 2008). However, even after widespread installation of LSLs became less common in the 1930's, the practice continued in several major US cities including Philadelphia, PA; Milwaukee, WI; Boston, MA; and Chicago, IL (Rabin, 2008).

The Safe Drinking Water Act's Lead and Copper Rule (LCR) banned the material from use in public water systems in 1986 and enacted a federal mandate for local water authorities to conduct periodic testing and remediation efforts. This rule constitutes noncompliance as either a failure to implement testing procedures or when at least 10% of homes that are tested have lead

concentrations above an “action level” of 15 parts per billion (U.S. Environmental Protection Agency, 2016). In 2016, a Natural Resources Defense Council report showed that over 5,300 community water systems throughout the US, serving over 18 million people, were in violation of the LCR, while other surveys have estimated that 6.1 million lead service lines (LSL) are still in use around the country (Olson & Fedinick, 2016, Cornwell, Brown & Via, 2016). In addition to compliance sampling practices that systematically underestimate high lead levels and potential human exposure, as suggested by Del Toral, Porter, & Schock (2013), lack of enforcement of 88.8% of LCR violations demonstrates a substantial need for more effective responses to what is a widespread environmental health and justice problem (Katner et al, 2016; U.S. Environmental Protection Agency, 2016).

Several American cities have undertaken largescale remediation efforts with respect to LSL replacements in the interest of public health. Service lines are pipes that carry tap water from the water main to a residential unit through a curb stop and can be fully or partially made of lead. Partial LSLs have a non-lead portion either from the main to the curb stop or from the curb stop to the home. Occurrence of partial LSLs have not only been linked with higher water lead levels (WLLs) than full LSLs, but also with higher prevalence of elevated blood lead levels among children who reside at those homes (Trueman, Camara, & Gagnon, 2016; Dore, Deshommes, Laroche, Nour, & Prevost, 2019). The first major city in the United States to implement a complete LSL replacement program was Madison, WI, where more than 8,000 LSLs were removed from residential units between 2000 and 2012 at a cost of approximately $15.5 million. To gather data on private side plumbing material, the City of Madison commissioned a compulsory survey to thousands of property owners after holding several community meetings where consumers were informed of how to locate their servile lines and test for lead (City of Madison, 2016). As similar

initiatives are carried out in different cities around the United States, further investigation into lead's prevalence throughout public water systems will help protect at-risk families and children from harmful exposure.

1.1 Background

Pittsburgh, Pennsylvania's lead problem emerged in 2016 after the Pittsburgh Water and Sewer Authority's (PWSA) failed its LCR compliance testing. The PWSA's immediate response included statements that LSLs remained at an estimated 25% of homes and that homes built before 1986, which represent 94% of Pittsburgh's housing stock, were "more likely to have lead pipes" (Pittsburgh Water and Sewer Authority, 2016) (Data.gov, 2017). Of the 100 homes sampled during the PWSA's 2016 compliance testing, more than 10% of homes sampled tested above the federal action level, triggering mandatory public education as well as certain SL replacement measures (U.S. Environmental Protection Agency, 2016). Although studies have demonstrated inherent variability and limited reliability of standardized sampling for lead testing (Masters, Parks, Atassi, & Edwards, 2016), the PWSA's results still reflect an LCR violation and a public health hazard. Regardless of the difficulty and uncertain effectiveness of voluntary tap water testing, the PWSA must still replace 7% of remaining LSLs annually until the public system is lead-free, as well as other public engagement efforts (U.S. Environmental Protection Agency, 2016). In 2017, the PWSA and City of Pittsburgh stepped up their community response, allocating $1 million to the Safe Water Program for free voluntary water testing and discounted water filters, and halting the practice of partial LSL replacements, which Trueman et al. (2016) and Del Toral et al. (2013) have shown to increase lead contamination. However, a lack of available data remains the most

significant logistical challenge in identifying LSL locations where replacements and additional remediation efforts should take place (Blackhurst, 2017).

The challenge of locating Pittsburgh's remaining LSLs presents an opportunity to test computational approaches to achieving lead-free water systems. To date, there is a small body of scientific literature focused on modeling the distribution of lead hazards in a large public water system such as Pittsburgh. With each focusing on the city of Flint, similar analyses have mapped the city-wide prevalence of LSLs using geostatistical prediction and machine learning classification (Goovaerts, 2017; Abernethy, Schwartz, Chojnacki, Webb, & Farahi, 2018). Other studies have employed similar methodology to model the geographic distribution of WLLs in Flint, as well as deterministic factors for plumbosolvency in Raleigh, North Carolina (Goovaerts, 2018; Wang, Devine, Zhang, & Waldroup, 2014; Abernethy et al., 2016). This study approaches city-wide risk assessment of LSL with respect to Pittsburgh, a much larger city than Flint, as a binary classification problem. A thorough analysis of publicly available housing data from the City of Pittsburgh enabled the inclusion of several address-level predictors in a suite of seven machine learning classification algorithms. The models were trained to predict the presence or absence of full and or partial LSLs and then compared to assess predictive accuracy. Building upon previous studies of Flint, Michigan's city-wide LSL distribution, this approach incorporates a variety of data types as predictive features into a suite of classification models. The results demonstrate the relative importance as well as the limitations of housing conditions and spatial data in locating lead hazards. For policy makers and health advocates in Pittsburgh, as well as other cities with known LSL stocks, this analysis provides insight that could improve the efficiency of future remediation efforts.

2.0 Methodology

2.1 Data Sources

In July 2018, the PWSA released an online map of over 43,000 homes within its service area with information on service line material (Clift, 2018). To date, employees at the PWSA have digitized more than 120,000 paper records and conducted over 5,300 curb box inspections (CBI) to create this LSL inventory (Pittsburgh Water and Sewer Authority, 2018). The CBI field data consist of predominantly unknown or partially unknown values, as seen in Figure 1, due to the inability of PWSA field engineers to locate curb boxes or identify SL material at many of the addresses. In fact, only 29% of the PWSA's inspections had entirely conclusive results. The digitized historical records, however, show a more complete picture of lead's prevalence throughout Pittsburgh. Each observation in this data set was re-coded into a binary indicator of 'Lead' or 'Non-Lead' and georeferenced using the centroid of each observation's tax parcel (Table 1). When projected geographically, the data show high levels of clustering with respect to these newly labeled observations (Figure 2).

Table 1: Recoding of PWSA Data

'Lead'	**Unknown (Null)**	**'Non-Lead'**
Lead / Lead	Unknown / Unknown	Non-Lead / Non-Lead
Lead / Non-Lead	Unknown / Non-Lead	
Non-Lead / Lead	Non-Lead / Unknown	
Lead / Galvanized	Unable to Locate	
Non-Lead / Galvanized	No Data	
Lead / No Data	Non-Lead / No Data	

No Data / Lead
No Data / Galvanized
Lead / Unknown
Unknown / Lead

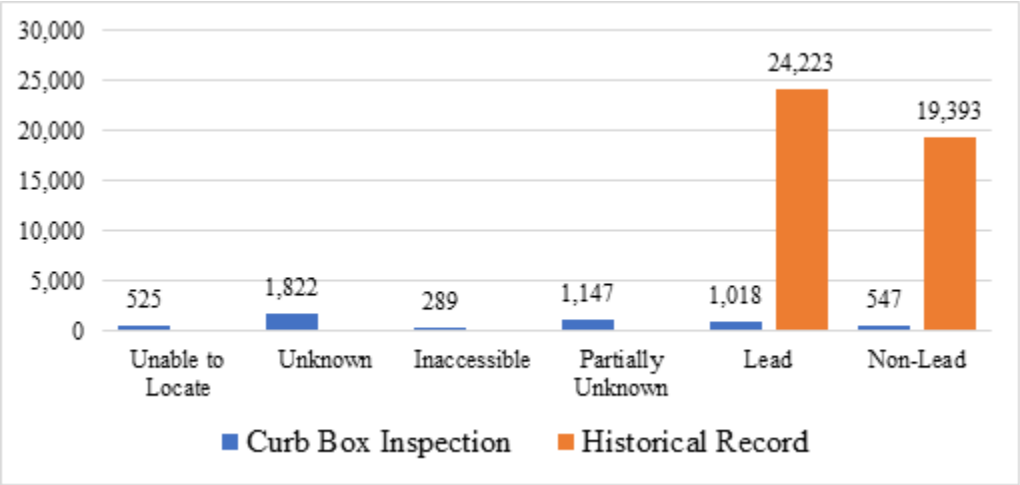


Figure 1: Distribution of Existing Lead Service Line Data

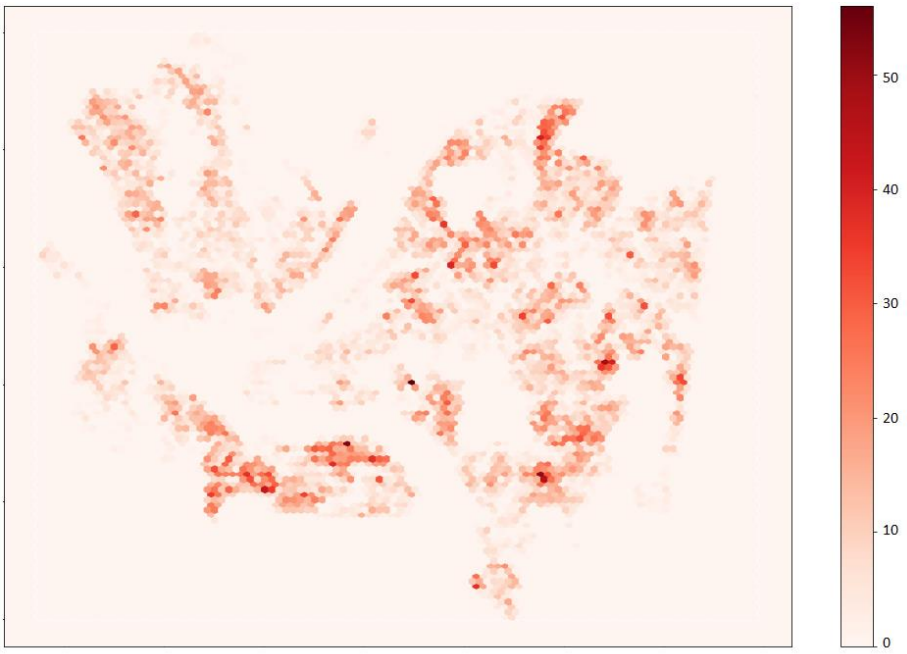


Figure 2: Spatial Density of Known Lead Observations

Data on Pittsburgh's housing stock was acquired from the Western Pennsylvania Regional Data Center. The Allegheny County Property Assessments data set contains property tax

information on residential building characteristics such as property value, condition and the year built (Snyder, 2018). These two sources were merged in order to build a data set containing each address in the PWSA service area with corresponding values for location, SL material, age of house and other characteristics. Determining the latitude-longitude point coordinates of each address facilitated a spatial analysis of the data. An initial geographic analysis of this data shows only a small number of neighborhoods with an average year built later than 1946, forty years prior to the national lead ban (Figure 3).

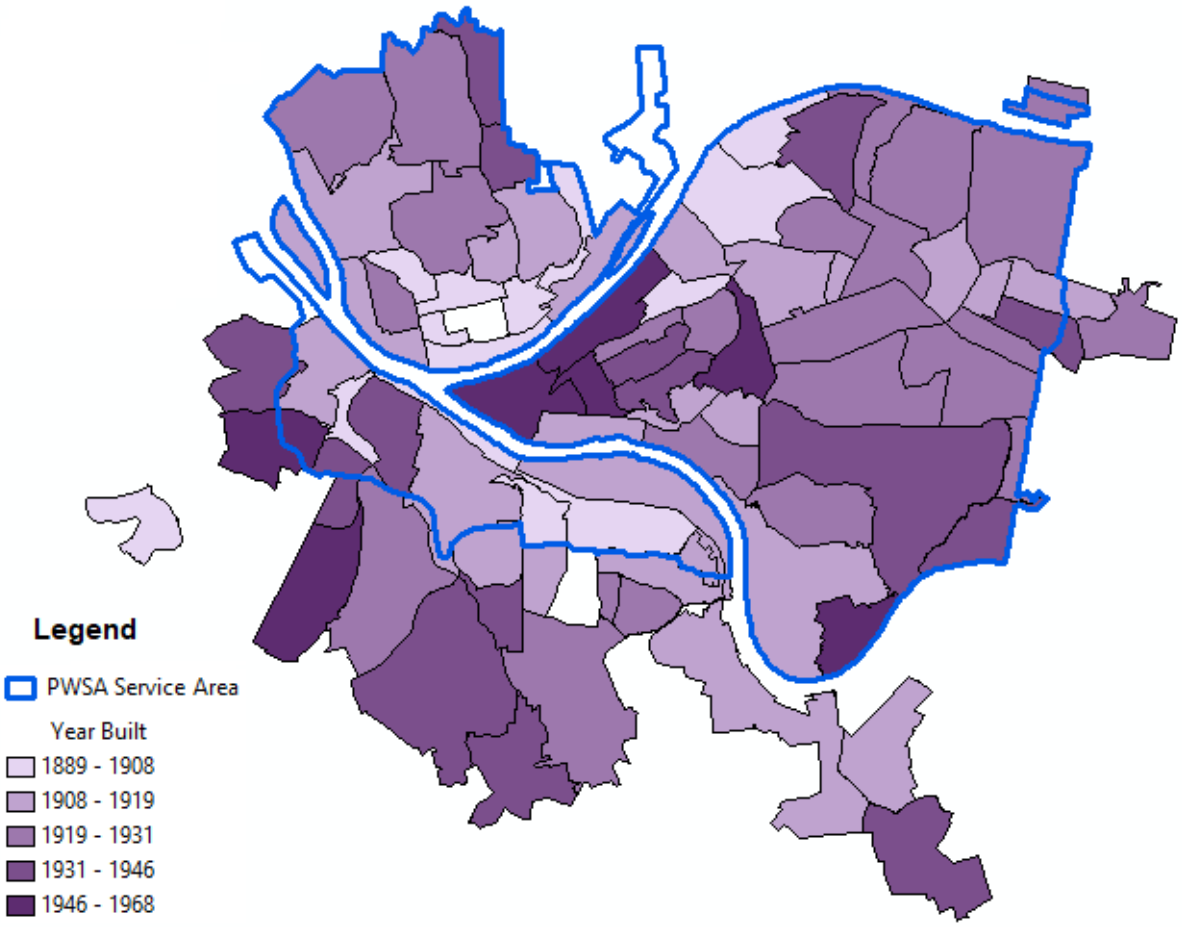


Figure 3: Average Year Built by Pittsburgh Neighborhood

2.2 Selecting and Eliminating Features

The historical legacy of lead in the United States would suggest that housing age as an important predictor of lead throughout Pittsburgh's housing stock. Of the 43,616 LSL records, 36,115 addresses have data for year built, according to PWSA historical records. At those houses in the data set built before the 1930s, majority have a partial or full LSL. However, patterns in the Allegheny County property assessments data set demonstrate the potential for inaccurate or biased reporting. The higher frequency among the first years of each decade (Figure 4) skew the distribution of housing age in Pittsburgh and suggest that data collectors may have estimated the year built of those houses lacking definitive records. To account for potential bias in these reported values, different coding formats for housing age were considered in the final training set of 37,532 addresses. These consisted of age (current year minus year built) as well as 20-year and 40-year age groups starting at year 1900 (e.g., pre-1900, 1900-1919, etc.).

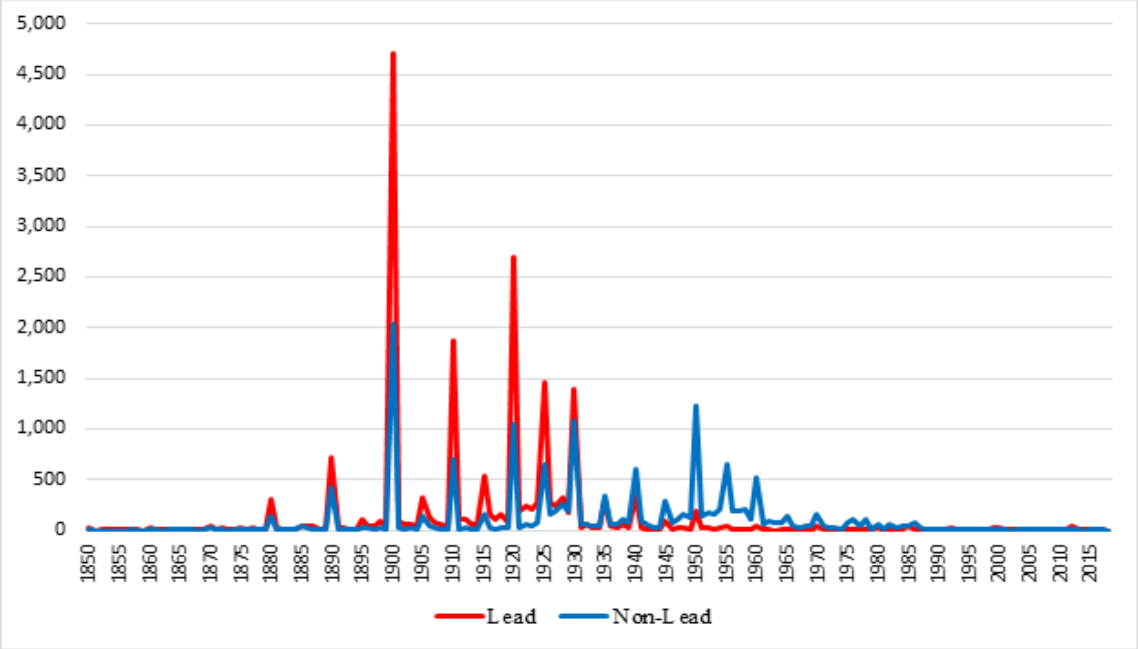


Figure 4: Lead and Non-Lead Observations by Year Built (Historical Records)

Due to the geographic nature of the data, we can evaluate the relationship between events as a spatial pattern of lead and non-lead observations. To do so, each address in the dataset was georeferenced with the coordinates of each corresponding tax parcel's centroid. When modeling the distribution of point events over space, it is important to describe the patterns made by the observations of each class within the study region (O'Sullivan & Unwin, 2014). In this example, latitude/longitude data were used to determine the impact of proximity on the presence or absence of lead. After splitting the observations in the data set into lead and non-lead point patterns, we determined the Euclidean distance between every address, *i* and the nearest event with service line material m_i, denoted by,

$$\boldsymbol{m_i} = \begin{cases} \boldsymbol{LSL}, & \boldsymbol{nearest\ neighbor\ with\ LSL} \\ \boldsymbol{NLSL}, & \boldsymbol{nearest\ neighbor\ with\ NLSL} \end{cases}$$

Equation A

$$d_i^m = \sqrt{(X^m - X_i)^2 + (Y^m - Y_i)^2}$$

Equation B

$$D_i = d_i^{LSL} - d_i^{NLSL}$$

Equation C

where X^m and Y^m represent the latitude-longitude coordinates of the nearest neighbor events in the lead (LSL) and non-lead (NLSL) point pattern, respectively. The difference, D_i, between the two distance measures, d_i^{LSL} and d_i^{LSL}, represents the relative proximity of each observation to another unit with an LSL. In other words, D_i is a measure of whether an address is closer to an event in the LSL point pattern or one in the NLSL one. In Figure 6, it is evident that on average, points in the LSL pattern are closer to each other than to NLSL events, due to the distribution's negative skew. For units with LSLs, the median relative proximity measurement is –0.000257º compared to

0.000119º for non-lead observations, suggesting that spatial proximity would improve our ability to distinguish between the two classes.

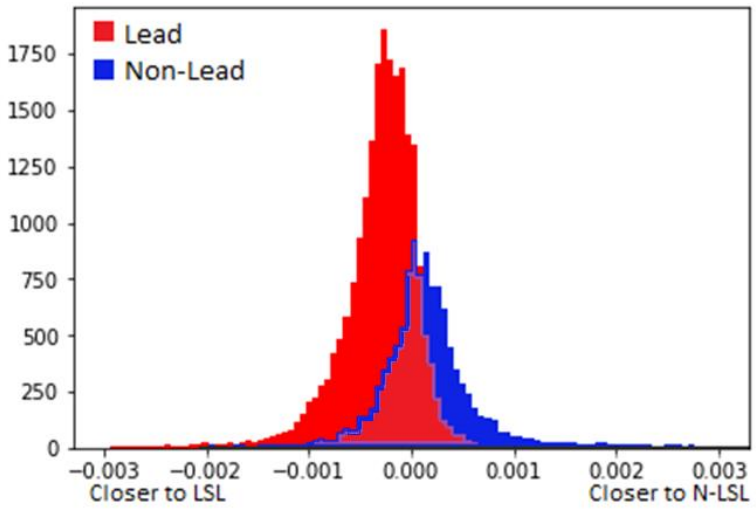


Figure 5: Frequency Distribution of Relative Proximity at LSL and Non-LSL Locations

To further examine the geospatial differences between the two target classes, we consider the number of lead observations in the immediate vicinity of each home. In effect, we extend a circle of radius r from centroid of each tax parcel and count the number of lead observations within the resulting buffer. This approach is similar to that of Blackhurst (2018a), who considered the number of LSL locations among up to twenty neighboring houses on the same street when estimating total number of remaining LSLs in the PWSA's service area. Instead, the total number of events was counted within a distance from each housing unit equal to one-half percent of the height of the overall study area. By doing so, we build on the insight of Blackhurst (2018a), who demonstrated that the likelihood of having a non-LSL at a given property decreases steadily as the number of houses with LSLs in the immediate vicinity grows (Figure 6). Introducing this new spatial feature allows us to examine whether there is more separation created by one approach or another. We first plot the relative proximity and then the number of neighboring LSLs against year built to see the extent to which the two target classes are separate or overlapping.

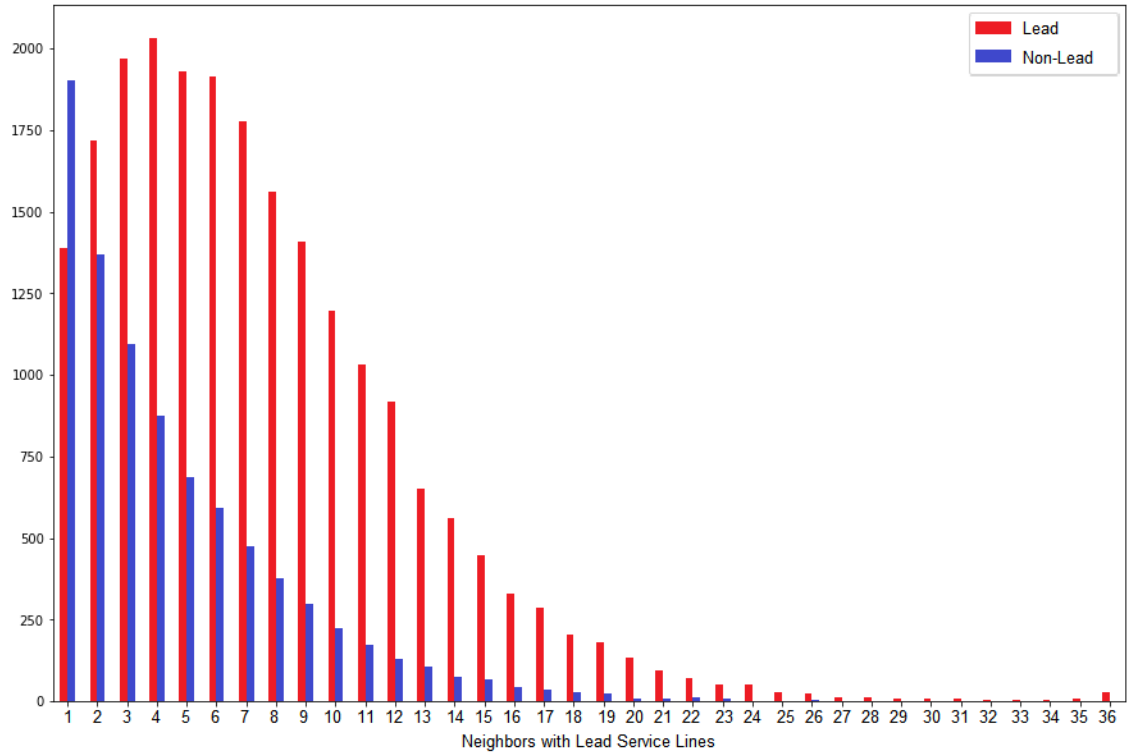


Figure 6: Nearest Neighbors with LSLs at Lead and Non-Lead Observations

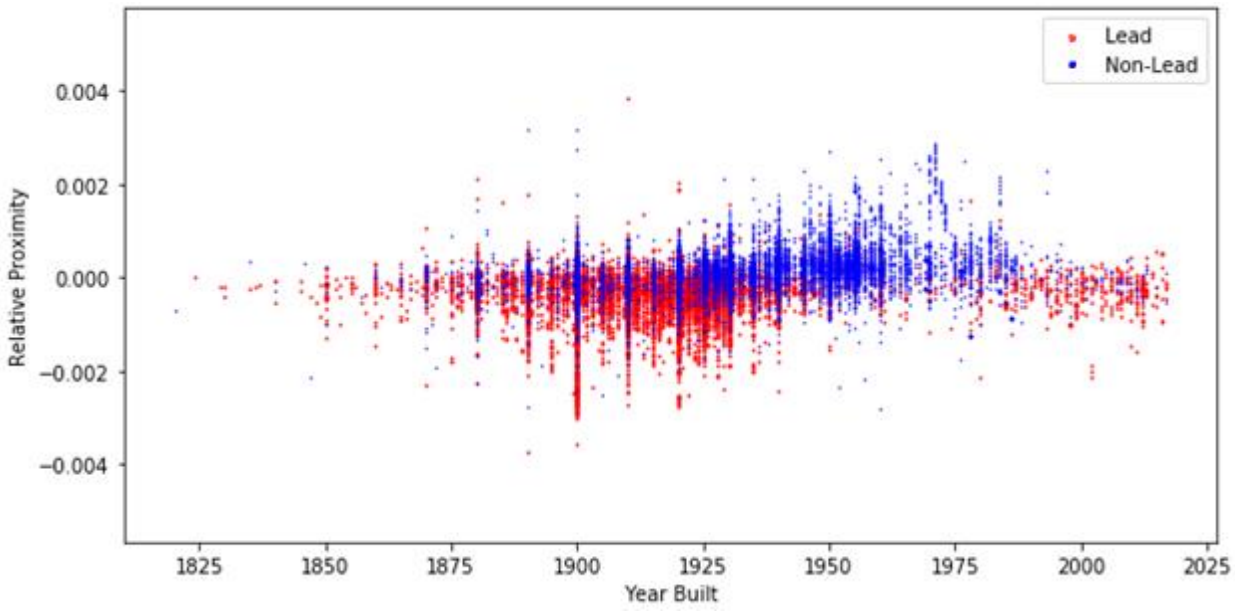


Figure 7: Scatterplot of Relative Proximity and Year Built

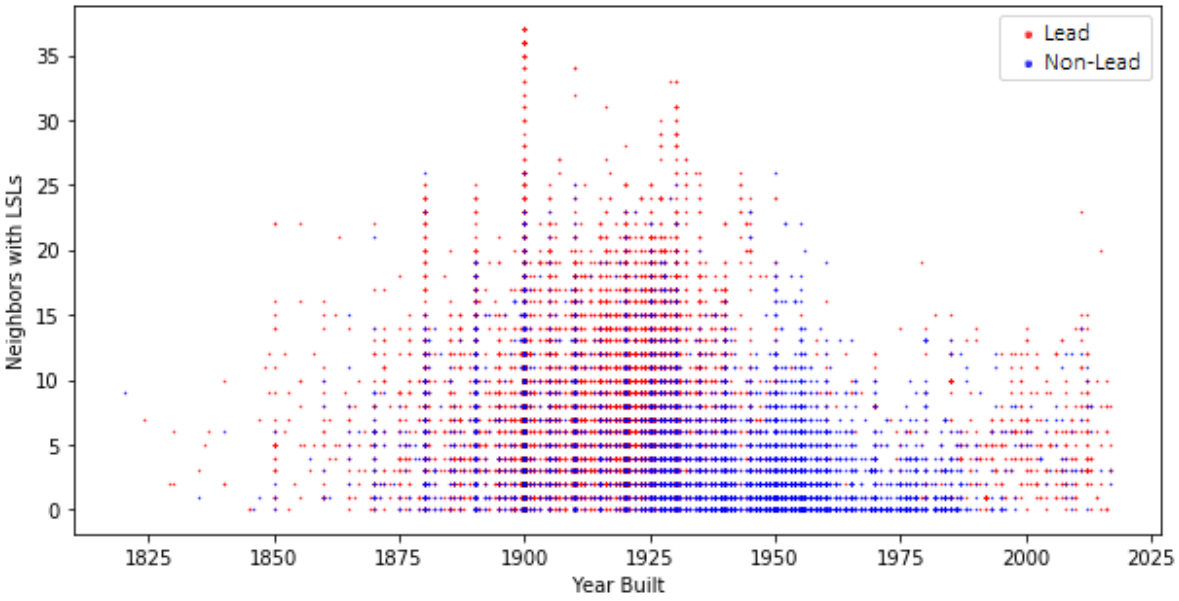


Figure 8: Nearest Neighbors with LSLs and Year Built

After exploring age and location as predictors, we performed a recursive feature elimination analysis (RFE), which produces a model with those attributes that contribute most to predicting the target class (e.g., lead or non-lead), on the relevant fields in the merged dataset. The benefit of RFE is the removal of superfluous attributes that could either marginally or negatively impact model performance, as well as an empirical ranking of each remaining features in the data set. The RFE process uses a logistic regression model as an external estimator with which to generate coefficients for increasingly smaller and smaller subsets of features. After irrelevant features are eliminated, the remaining three are those that best predict the response variable (Sci-kit Learn Developers, 2018). In this application, using RFE helps demonstrate the added value of publicly available housing data in predicting the presence of lead plumbing by distinguishing the most determinative characteristics. In Table 2, we report the name and description of each field in the property assessments data set included in the RFE process. The results show that the 40-year age grouping, physical condition and number of rooms were the three supported features. While the number of nearby lead observations ranked immediately after these fields, the relative

proximity measurement was second to last in the order, despite having more visible separation between lead and non-lead classes in the initial analysis (Figure 7). The assessed value of each property was also included in the feature selection analysis but performed poorly, ranking third to last. A higher ranking of property value would have suggested that the risk for LSLs increases for homes with lower value. Blackhurst (2018b) evaluates this relationship by assessing the impact of LSL presence on property values in Pittsburgh, concluding that the presence of the hazard resulted in an average price reduction of $9,700. The 'StatisticalModel' used to predict LSL locations in Flint included property values as a key predictor as well; however, while property values are an important indicator of lead in Flint, its impact may ultimately be context dependent (Abernethy et al., 2018).

Table 2: Recursive Feature Elimination Results

Feature	***Support***	***Rank***	***Description***
'AGE GROUP 40'	True	1	40-year age grouping (e.g., 1900-1939)
'CONDITION'	True	1	Physical condition of the building
'TOTALROOMS'	True	1	Number of stories in the structure
'COUNTPB'	False	2	Number of lead observations within standard distance from the unit
'BEDROOMS'	False	3	Number of bedrooms
'AGE GROUP 20'	False	4	20-year age grouping (e.g., 1900-1919)
'AGE' (normalized)	False	5	Current year minus year built
'GRADE'	False	6	Quality of construction
'CDU'	False	7	Composite rating of condition, desirability and utility of the property
'STYLEDESC'	False	8	Building style (e.g., townhouse)
'HOOD'	False	9	Pittsburgh Neighborhood
'USEDESC'	False	10	Primary use of the parcel (e.g., Two family home)

'OWNERDESC'	False	11	Owner type (e.g., individual, corporation, etc.)
'LOCALBUILDING' (normalized)	False	12	Locally assessed building value
'LOCALTOTAL' (normalized)	False	13	Locally assessed property value
'DIF' (normalized)	False	14	Proximity difference
'LOTAREA' (normalized)	False	15	Total square footage of land

2.3 Supervised Learning Approach

This study matches the PWSA's historical records and inspections results with county property assessments data to train seven binary classification models. The output is an indicator of whether or not the hazard exists at a house with unknown LSL data. Multiple types of machine learning algorithm—specifically logistic regression, support vector machine (SVM), k-nearest neighbors (k-NN), decision tree and random forest—were trained and ten-fold cross validated to determine the approach with the highest relative accuracy. In addition, four separate sets of features were trained and compared to help decide on an optimal approach. The first three sets consist of the age of the housing unit, and then the age and the two geospatial statistics, relative proximity and number of nearby LSLs. These first three sets are meant to provide a baseline predictive accuracy level with which to compare higher-dimensional models that include housing characteristics data. With respect to the scalability and functionality of this approach, determining a baseline level of accuracy with only historical and geographic information is advantageous because detailed housing characteristics may not be available in certain cities. Given the national scope of the lead water contamination issue, a more scalable baseline approach allows for potential adaption to more locations with aging water infrastructure.

Table 3: Machine Learning Models and Configurations

Algorithm	*Model-Specific Configurations*
Logistic Regression	n/a
Support Vector Machine	RBF kernel
Support Vector Machine	Linear kernel
K-Nearest Neighbors	k = 3
K-Nearest Neighbors	k = 5
K-Nearest Neighbors	k = 7
Decision Tree	n/a
Random Forest	1,000 decision trees

The k-NN algorithm is a non-linear model trained with k = 3, 5 and 7 to test which cluster size yielded the lowest error. The kernel size, k, indicates the number of closest observations in the test data which determine the prediction class by taking a "vote" (James, Hastie, & Witten, 2017). Non-linear models such as k-NN or decision tree and random forest are expected to perform better than linear ones where data are not linearly separable. The SVM algorithm, in contrast to non-linear models, employs a so-called "kernel trick," which constructs a maximum margin separator in multi-dimensional space in order to better differentiate between data that are not easily separable in the original input space. In this application, the SVM is the hyperplane separator at the largest possible distance between 'Lead' and 'Non-Lead' observations, mapped in *n* dimensions, where *n* is the number of features in the support vector. While SVMs are an effective tool for classification in high-dimensional spaces and are memory efficient, they tend to perform worse (i.e., longer training time, lower accuracy) with large, noisy data sets in which target classes are overlapping and more difficult to separate (Russel & Norvig, 2016). This approach determines the relative advantage of linear and non-linear, parametric and non-parametric and simple and ensemble models. Comparing several classification algorithms allows us to account for different relationships between the various features that were ultimately supported after RFE.

For each of the four feature sets in this analysis, we choose the determine algorithm by computing the following two cross-validation statistics: 1) cross-validated accuracy score, and 2) area under ROC curve (AUC). The cross-validated accuracy score measures the expected accuracy of out-of-sample predictions by removing one-by-one each observation from the training set and re-classifying using the existing model (James et al., 2017). A Receiver Operating Characteristics (ROC) curve plots the probability of false positive versus the probability of detection (Swets, 1988; Fawcett, 2006; Goovaerts et al., 2016). The relative area under the ROC curve (AUC statistic), which ranges from 0 (worst case) to 1 (best case), is equivalent to the probability that the classifier will rank a randomly chosen positive instance (e.g., presence of LSL) higher than a randomly chosen negative instance (e.g., absence of LSL). Therefore, the cross-validated accuracy score is treated as a measure of the model's expected performance and the AUC statistic as a measure of its ability to distinguish between the presence or absence of an LSL. Once the models with the highest accuracy were determined, predictions were made at addresses where CBIs had successfully gathered ground truth data to further evaluate the approach's effectiveness.

Table 4: Feature Set Components

	Features	**Units in Training Set**
Set 1	Age	34,769
Set 2	Age	34,769
	D_i	
Set 3	Age	34,769
	Nearby LSLs	
Set 4	Age Group (40-years)	34,769
	Nearby LSLs	
	Condition	
	Total Rooms	

3.0 Results and Discussion

Table 5: Classification Performance Scores

	Set 1		***Set 2***		***Set 3***		***Set 4***	
Model	*CV Score*	*AUC*	*CV Score*	*AUC*	*CV Score*	*AUC*	*CV Score*	*AUC*
Logistic Regression	0.7409	0.6845	**0.7998**	**0.7458**	0.7750	0.7174	0.7169	0.7011
SVM RBF	0.4390	0.7085	0.6465	0.5322	0.4432	0.7311	0.7765	0.7221
SVM Linear	**0.7455**	**0.6826**	--	--	**0.7825**	**0.7213**	0.7178	0.7005
k-NN, k = 3	0.3656	0.6988	0.7271	0.7299	0.4099	0.7097	0.7285	0.7126
k-NN, k = 5	0.3571	0.7057	0.7533	0.7367	0.4214	0.7161	0.7421	0.7181
k-NN, k = 7	0.4084	0.7094	0.7661	0.7400	0.4500	0.7158	0.7515	0.7234
Decision Tree	0.4396	0.7079	0.4025	0.7208	0.4303	0.7265	0.7630	0.7297
Random Forest	0.4396	0.7081	0.4026	0.7114	0.4535	0.7269	**0.7800**	**0.7334**

3.1 Comparing Model Performance

After preprocessing, training and resampling the data from each of the four sets of features, the performance statistics were reported to help choose the best model for informing our LSL risk assessment (Table 5). In Set 1, the SVM linear kernel model's cross-validated accuracy score (CV

score) of 74.55 percent was the highest, indicating a strong link between an address' documented age and its water service line material. Due to the simplicity of predicting on housing age alone, the results of the logistic regression model in Set 1 are promising with respect to the approach's baseline accuracy and geographic scalability, as discussed in the previous section. However, predicting only by the age of a household neglects the numerous other factors as well as the potential advantages of more robust, multivariate approaches. With Set 2 predicting on age and relative proximity, we achieved the highest maximum performance score, 79.99 percent, with logistic regression. The entire set of performance statistics for Set 2 demonstrate that on average, the relative proximity increases cross-validated accuracy by 20.8 percent when added to the model along with housing age. Using our second spatial measurement, the number of nearby LSL observations, we see a 2.8 percent increase in the average cross-validated accuracy score compared to the univariate set. However, for both Sets 2 and 3, the linear parametric models (logistic regression and SVM linear kernel) perform at least 74 percent accuracy compared to a maximum of nearly 45% among the remaining algorithms.

By considering housing characteristics however, we only see a marginal increase in overall cross-validated accuracy for the feature set. The highest cross-validated score in Set 3, which consisted of the house's 40-year age group, number of nearby LSLs, condition and number of rooms, was the random forest algorithm at 78.00 percent. While the logistic regression model from Set 2 had the highest individual score, Set 3's predictions were approximately 29.5 percent more accurate on average. The results of ten-fold cross-validation in this machine learning analysis demonstrate a relatively high level of expected accuracy for out-of-sample units (i.e., houses with none or unknown LSL data). However, to decide the optimal model for predicting the unknown LSL locations, we must also consider the discrimination ability of each model, represented by the

AUC score. In the previous section, we explore the separation and overlap that exists when age and spatial features are plotted together. To visualize the separation created by Set 4, we plot each address in three dimensions using age, nearest neighbors, condition and total rooms.

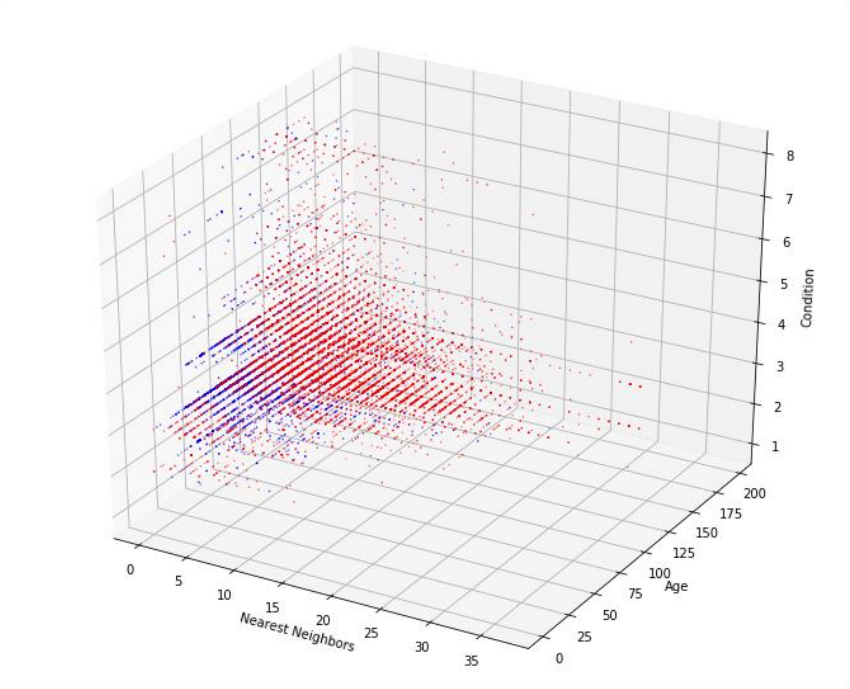


Figure 9: Nearby LSL Locations, Housing Age and Condition

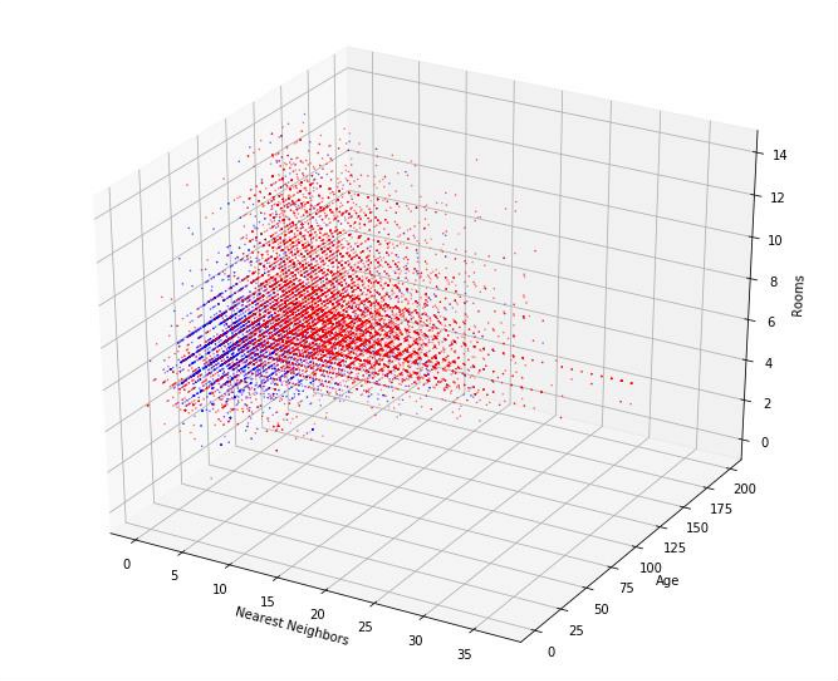


Figure 10: Nearby LSL Locations, Housing Age and Total Rooms

3.2 Choosing a Final Model

Table 6: Classification Results

	Model	True Positive	False Positive	True Negative	False Negative	Sensitivity (Recall)
SET 1	Logistic Regression	4412	1337	968	237	**95%**
	k-NN (k = 3)	4277	1204	1101	372	92%
	k-NN (k = 5)	4325	1196	1109	324	93%
	k-NN (k = 7)	4343	1188	1117	306	93%
	Decision Tree	4368	1207	1098	281	94%
	Random Forest	4361	1203	1102	288	94%
	SVM (RBF Kernel)	4361	1201	1104	288	94%
	SVM (Linear Kernel)	4423	1351	954	226	**95%**
SET 2	Logistic Regression	4167	958	1363	441	90%
	k-NN (k = 3)	3993	944	1377	615	87%
	k-NN (k = 5)	4087	960	1361	521	89%
	k-NN (k = 7)	4150	976	1345	458	90%

	Decision Tree	3913	946	1375	695	85%
	Random Forest	3745	905	1416	863	81%
	SVM (RBF Kernel)	4490	2112	209	118	**97%**
	SVM (Linear Kernel)	--	--	--	--	--
SET 3	Logistic Regression	4240	1100	1205	409	91%
	k-NN (k = 3)	3949	991	1314	700	85%
	k-NN (k = 5)	4254	1113	1192	395	92%
	k-NN (k = 7)	4302	1138	1167	347	**93%**
	Decision Tree	4331	1103	1202	318	**93%**
	Random Forest	4310	1091	1214	339	**93%**
	SVM (RBF Kernel)	4313	1073	1232	336	**93%**
	SVM (Linear Kernel)	4319	1121	1184	330	**93%**
SET 4	Logistic Regression	4147	1140	1176	489	89%
	k-NN (k = 3)	4127	1077	1239	509	89%
	k-NN (k = 5)	4178	1077	1239	458	90%
	k-NN (k = 7)	4197	1062	1254	439	91%
	Decision Tree	4332	1100	1216	304	93%
	Random Forest	4298	1066	1250	338	93%
	SVM (RBF Kernel)	4419	1179	1137	217	**95%**
	SVM (Linear Kernel)	3907	1023	1293	729	84%

In this specific problem, a false negative prediction means that the predicted material of a service line is safe at a particular house when in fact, there is an LSL that could potentially harm its residents. Therefore, a basic understanding of expected accuracy from cross-validation is not sufficient for determining the best approach. Table 6 reports the specific classification results of each model that was trained in this study. These scores allow for a more robust performance analysis focused on minimizing the number of false negative outcomes in implementation. We compare the sensitivity of each model, which specifically tells us how effective the approach is at detecting the presence of a LSL by dividing the number of true positive predictions by the total number of actual positives in a twenty percent subset of the data. Overall, we see an average recall above 90 percent, meaning that this approach is particularly effective with respect to predicting the presence of an LSL.

Ultimately, the models which perform the best while also making the fewest false negative predictions are the logistic regression model in Set 1 and the SVM RBF kernel model in Set 4, which both have a sensitivity of 95 percent. The highest recall rate was 97 percent, produced by the SVM RBF kernel model in Set 2; however, this model predicted LSL presence at all but five percent of the units in the sample. Thus, despite a marginally lower cross-validated accuracy than the logistic regression model in Set 2, we choose to implement the SVM RBF kernel model from Set 4, which predicts on the age group, total neighboring LSLs, physical condition and total rooms. The resulting predictions for LSL locations are visualized in Figure 11, which provides a detailed look at the predicted aggregate prevalence in each neighborhood. We see that in 68% of the Pittsburgh's neighborhoods, the final model predicted the presence of an LSL at over half of the addresses. The average neighborhood prevalence of LSLs was just over 60 percent, further highlighting the scope of the problem faced by Pittsburgh residents and the PWSA.

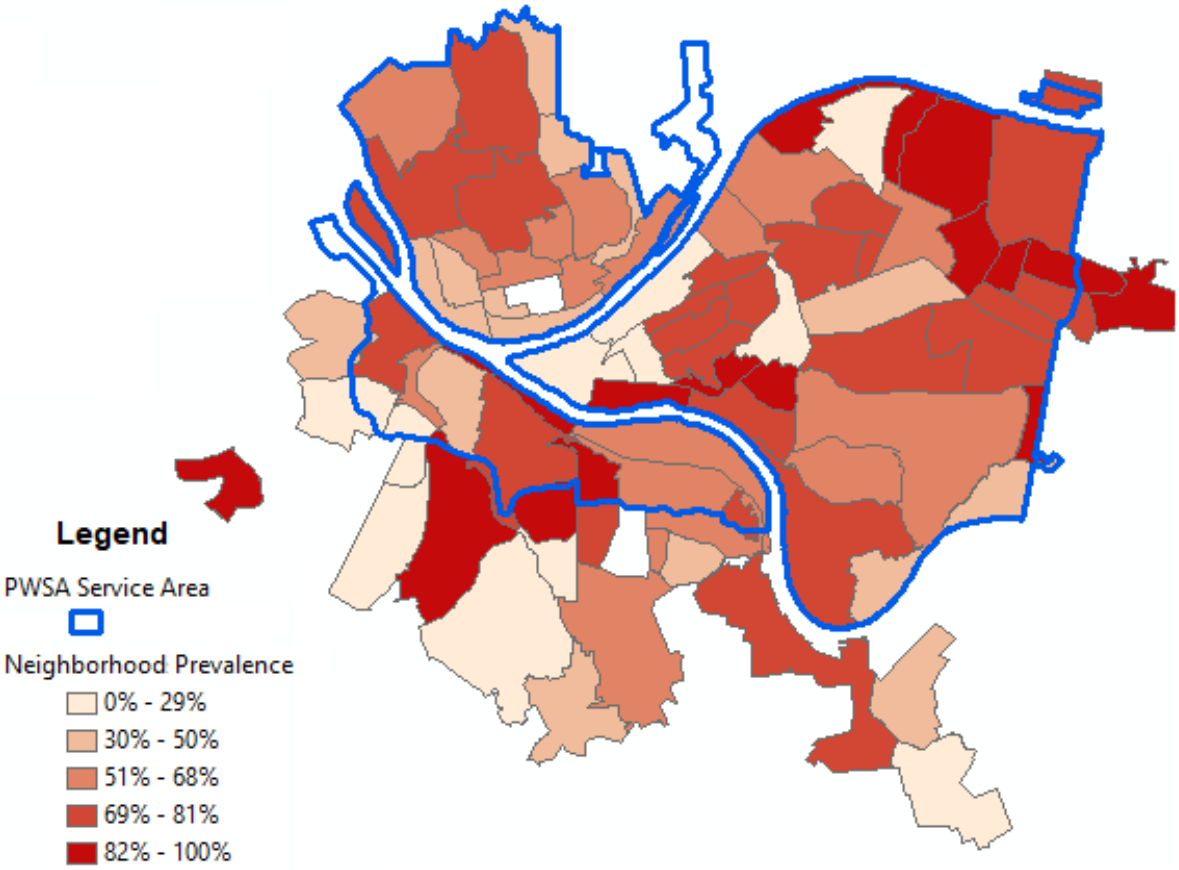
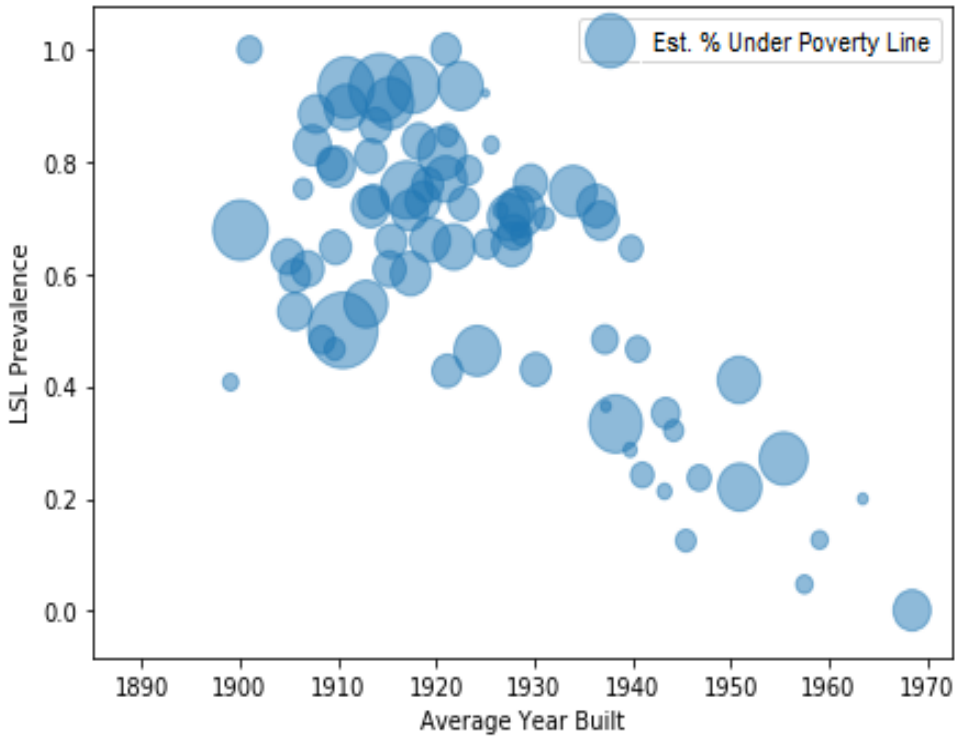


Figure 11: Predicted Neighborhood Prevalence of LSLs (Final Model)

3.3 Environmental Justice

A serious concern for decision makers in the PWSA, local government and within communities is the impact of lead in water on disadvantaged communities. As seen in the bubble plot in Figure 9, many of the neighborhoods with a simulated LSL prevalence of 50% or higher also have high rates of poverty, which is represented by the size of each bubble. This implies that lead water contamination may disproportionately affect communities with lesser means and

therefore fewer resources. The implications of this are important in part because it reflects the financial means of particular neighborhoods to remediate exposure, either by paying for private-side LSL replacements or repeatedly purchasing lead-certified water filters, which can become expensive over time. Lead poisoning in particular causes cognitive deficiencies that have been shown, at least in part, to undermine long term economic performance and contribute to the perpetuation of low socioeconomic status (Clay, Troesken, & Haines, 2014). Environmental justice is a crucial factor in the reality of the lead problem in Pittsburgh, as it is in Flint and other cities in the United States (Olson & Fedenick, 2016; Katner et al., 2016; Goovaerts, 2018).



4.0 Future Work and Conclusions

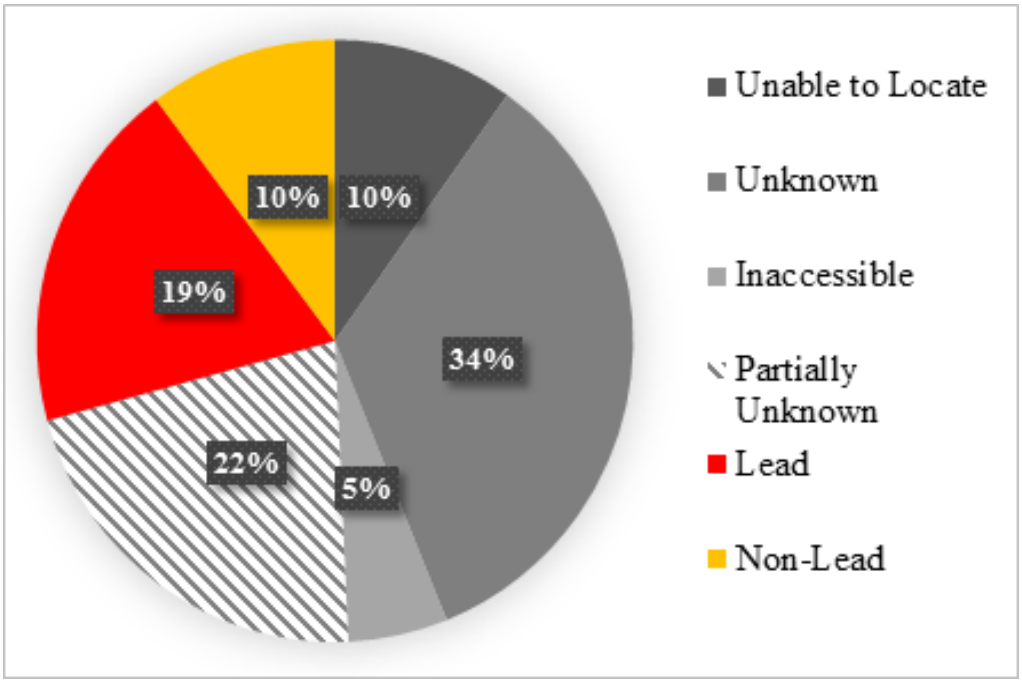


Figure 12: Breakdown of PWSA Curb Box Inspection Results (July 2018)

4.1 Comparing Similar Approaches

Despite practical and computational limitations, the results of this analysis nonetheless demonstrate a promising application of machine learning to modeling the distribution of lead hazards throughout a water service area. As additional address-level data become available (e.g., water contamination, SL materials and unaccounted-for SL replacements), the inclusion of more domain-specific features that are indicative of lead contamination could potentially improve the effectiveness of this approach. For example, further analysis of spatial point patterns could potentially improve the performance of Models 2 and 3 by including density-based geostatistical

features. In Abernethy et al. (2018), the classification accuracy of the 'StatisticalModel,' which predicted 'safe' or 'unsafe' service line materials in Flint, MI, improved over time as more data on LSL locations became available, reaching an accuracy level in a holdout set of 91.8 percent. This approach is also unique in that the problem is then extended from the identification of hazard locations to the decision-making for inspections and replacement. To emulate this approach in Pittsburgh, further cooperation with the PWSA, particularly with respect to the sharing of sensitive customer data, is necessary. The only existent analysis in Pittsburgh is that of Blackhurst (2018a) in which the number of young children, the cost of replacements and the income levels in each neighborhood are considered. Additionally, an approach which separates field data from historical records could strengthen the analysis by treating CBIs and historical records as primary and secondary data, respectively. This methodology resembles that of Goovaerts (2018), which matches address-level field data from inspections throughout the City of Flint with aggregated statistics based on historical records. Other potential extrapolations from this analysis include matching simulated LSL locations to data from annual blood testing. Such a study could help estimate the impact of LSLs on water contamination and childhood lead poisoning and as Potash et al. (2015) argues, help implement successful interventions that prevent rather than remediate harmful exposure.

4.2 Conclusions

The approach taken in this study builds off a strong theoretical background of geostatistical and machine learning applications for environmental risk assessment. Combining existing data on housing age and location, we can expect a baseline accuracy of nearly 80 percent in our prediction

of the presence or absence of an LSL at addresses without SL material records. This baseline was determined by training and testing three simple sets of features, which account for the age of the house and its age as well as its proximity to other properties where LSLs are known to exist. We also see that using a machine learning approach is very effective at locating LSL locations specifically, predicting correctly over 90 percent of the time, on average, when the actual material is lead. Given the age distribution of the PWSA's existing plumbing records, it is likely that a considerable number of homes built between 1900 and 1930 once had LSLs that have since been replaced, making it more difficult to differentiate between housing units with LSLs and Non-LSLs. Despite this challenge, however, this approach is particularly promising due to the nature of the problem at hand. In our results, we see that false negatives, which are the worst possible outcome, are significantly less common among our predictive models. While false positive predictions could result in unnecessary inspections or excavations by the PWSA, a more precautionary approach will ultimately be more effective at preventing exposure in the short term while additional analyses are carried out.

Leveraging Pittsburgh's publicly available housing data offers insights into the process by which city officials, regulators and non-government decision makers can identify high risk constituents and act accordingly to protect public health. The results demonstrate the extent to which housing age contributes to predicting lead hazard locations throughout Pittsburgh's housing stock without the consideration of other potentially valid predictors such as, tap water test results or aggregated poverty levels and average property values. Of the 5,348 reported curb box inspections as of July 2018 (Figure 10), only 29 percent yielded conclusive results, leaving the remaining 71 percent of homes either with incomplete or no information regarding the presence of a lead hazard. Moving forward, policy makers should consider the cost-effectiveness of conducting

further curb box inspections without specifically targeting houses identified as high-risk. As cities across the country embrace data-driven approaches to tackling public health emergencies, thorough analyses like the one presented in this study, as well as those conducted in Flint, MI can offer valuable insight into individualized risk-assessment and the formation of best practices. By applying machine learning to the process of locating and replacing LSLs, we can prevent lead poisoning through water effectively, efficiently and equitably.

Bibliography

Abernethy, J., Farahi, A., Schwartz, E., Stroud, J., Anderson, C., Nguyen, L., . . . Webb, J. (2016, September 25). Flint Water Crisis: Data-Driven Risk Assessment Via Residential Water Testing. *Bloomberg Data for Good Exchange Conference*.

Abernethy, J., Chojnacki, A., Farahi, A., Schwartz, E., Webb, J. (2018). "ActiveRemediation." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, doi:10.1145/3219819.3219896

Blackhurst, M., Center for Social & Urban Research, University of Pittsburgh. (2018a). Predicting Lead Water Service Line Locations for Improving Mitigation Strategies /. Retrieved from https://ucsur.pitt.edu/environ/Lead_CE.html.

Blackhurst, M., Center for Social & Urban Research, University of Pittsburgh. (2018b). Do lead water laterals affect property values? A Case Study of Pittsburgh, PA, 2018 /. Retrieved from https://ucsur.pitt.edu/lead_water_laterals_2018.php.

Blackhurst, M. (2017, June). "Regulatory Gaps May Increase Risks of Lead in Drinking Water with Service Line Replacements." *Pittsburgh Economic Quarterly, University of Pittsburgh Center for Social and Urban Research.* pp. 4-6.

City of Madison. (n.d.). Information for Water Utilities on Lead Service Line Replacement. Retrieved from http://www.cityofmadison.com/water/water-quality/lead-service-replacement-program/information-for-utilities-on-lead-service

Clay, K., Troesken, W., & Haines, M. (2014). Lead, Mortality, and Productivity. *Review of Economics and Statistics, 93*(3). doi:10.3386/w16480

Cosgrove, E., Brown, M. J., Madigan, P., McNulty, P., Okonski, L., Schmidt, J., (1989). Childhood lead poisoning: case study traces source to drinking water. J. Environ. Health 52 (1), 346–349.

Data.gov. (2017, June 03). PGH SNAP Census Data. Retrieved from https://catalog.data.gov/dataset/pgh-snap

Del Toral, M. A., Porter, A., and Schock, S. (2013). "Detection and Evaluation of Elevated Lead Release from Service Lines: A Field Study." *Environmental Science & Technology* 47.16: 9300-307.

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters, 27*(8), 861-874. doi:10.1016/j.patrec.2005.10.010

Goovaerts, P. (2017). How geostatistics can help you find lead and galvanized water service lines: The case of Flint, MI. *Science of The Total Environment, 599-600,* 1552-1563. doi:10.1016/j.scitotenv.2017.05.094

Goovaerts, P. (2019). Geostatistical prediction of water lead levels in Flint, Michigan: A multivariate approach. *Science of The Total Environment, 647,* 1294-1304. doi:10.1016/j.scitotenv.2018.07.459

"Important Information About Lead in your Drinking Water." (2016). *Pittsburgh Water and Sewer Authority.*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). "Support Vector Machines." In James, G., Witten, D., Hastie, T., & Tibshirani, R. (Eds.), *An Introduction to Statistical Learning: With applications in R* (337-356). New York: Springer.

Katner, A., Pieper, K. J., Lambrinidou, Y., Brown, K., Hu, C., Mielke, H. W., & Edwards, M. A. (2016). Weaknesses in Federal Drinking Water Regulations and Public Health Policies that Impede Lead Poisoning Prevention and Environmental Justice. *Environmental Justice, 9*(4), 109-117. doi:10.1089/env.2016.0012

Lead Map. (2018). Pittsburgh Water and Sewer Authority. Retrieved July, 2018, from http://lead.pgh2o.com/your-water-service-line/planned-water-service-line-replacement-map/

Masters, S., Parks, J., Atassi, A., & Edwards, M. A. (2016). Inherent variability in lead and copper collected during standardized sampling. *Environmental Monitoring and Assessment, 188*(3). doi:10.1007/s10661-016-5182-x

Olson, E. D., & Fedinick, K. P.. (2016). "What's in Your Water: Flint and Beyond." *Natural Resources Defense Council*.

O'Sullivan, D., & Unwin, D. (2014). "Point Pattern Analysis." In O'Sullivan, D., & Unwin, D. (Eds.), *Geographic Information Analysis* (121-130) Hoboken: Wiley.

Potash, E., Ghani, R., Brew, J., Loewi, A., Majumdar, S., Reece, A., . . . Mansour, R. (2015). Predictive Modeling for Public Health. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 15*. doi:10.1145/2783258.2788629

Rabin, R. (2008). The Lead Industry and Lead Water Pipes "A MODEST CAMPAIGN". *American Journal of Public Health, 98*(9), 1584-1592. doi:10.2105/ajph.2007.113555

Russell, S. J., & Norvig, P. (2016). "Learning from Examples." In Russell, S. J., & Norvig, P. (Eds.), *Artificial intelligence: A Modern Approach* (695-737). Upper Saddle River: Pearson.

Scikit-Learn Developers. (2018). Sklearn.feature_selection.RFE¶. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

Shaddick, G., & Zidek, J. V. (2016). *Spatio-temporal methods in environmental epidemiology*. Boca Raton, Fla: CRC Press.

Shannon, M., & Graef, J. W. (1989). Lead Intoxication. *Clinical Pediatrics, 28*(8), 380-382. doi:10.1177/000992288902800810

Snyder, M. B. (2018, August 21). Allegheny County Property Assessments. *Western Pennsylvania Regional Data Center*. Retrieved from https://data.wprdc.org/dataset/property-assessments.

Swets, J. (1988). Measuring the Accuracy of Diagnostic Systems. *Science, 240*(4857), 1285-1293. doi:10.1126/science.3287615

Troesken, W. *"The Great Lead Water Pipe Disaster."* MIT Press, (2008). Print.

TribuneReview. (2018, March 20). PWSA plans to check 15K homes for lead lines this year. Retrieved from https://triblive.com/local/allegheny/13442955-74/pwsa-plans-to-check-15000-homes-for-lead-lines-this-year.

Trueman, Benjamin F., Eliman Camara, and Graham A. Gagnon. (2016). "Evaluating the Effects of Full and Partial Lead Service Line Replacement on Lead Levels in Drinking Water." *Environmental Science & Technology* 50.14: 7389-396.

U.S. Environmental Protection Agency, Office of Water. (2016). "Lead and Copper Rule Revisions White Paper." Environmental Protection Agency.

Wang, Z. M., Devine, H. A., Zhang, W., & Waldroup, K. (2014). Using a GIS and GIS-Assisted Water Quality Model to Analyze the Deterministic Factors for Lead and Copper Corrosion in Drinking Water Distribution Systems. *Journal of Environmental Engineering, 140*(9). doi:10.1061/(asce)ee.1943-7870.0000816