Water quality in Chicago, IL: Predicting lead hazards

using housing assessment and socioeconomic variables

CAPP 30254: Machine Learning for Public Policy

Valeria Balza (vbalza)

Tarren Peterson (tarren)

James Midkiff (jmidkiff)

June 4, 2021

1. **Executive Summary**

The city of Chicago has more lead service water lines than any other city in the United States, with nearly 400,000 lead service lines connecting family homes and small apartment buildings to water mains.[[1]](#footnote-1) Corrosion of these lines may expose residents to toxic lead in drinking water, which even in small concentrations can result in developmental problems in children and contribute to health complications in adults. The magnitude of this problem was recently acknowledged by the local government when Chicago Mayor Lori Lightfoot announced the first phase of the Lead Service Line Replacement Program in September 2020.[[2]](#footnote-2) The program, which is estimated to cost 8.5 billion USD and take multiple decades to complete, must prioritize those communities that observe the highest risk of lead exposure.

We drew on a range of housing assessment, demographic, and socioeconomic variables to train classification algorithms—namely, logistic regression, random forests, and linear support vector classification models—that predict high lead exposure at the census block group level in the city of Chicago. Across all our models, we used recall as the main evaluation metric, as our objective given the context of this project was to minimize false negatives. [**Insert blurb on the best models and feature importance**]. Ultimately, our analysis reveals machine learning algorithms can provide insight into which communities may face the highest risk for lead exposure through lead service lines and may be used to inform the city’s strategy as it launches the Lead Service Line Replacement Program in Summer 2021.

1. **Background: Water Quality in Chicago**

~~What is the policy problem you are addressing and why is it important? Provide citations if relevant to support your claims of relevance (e.g. government requests for comments, white papers by policy organizations, etc.). What is your solution to the problem? What will you be using machine learning for? Who is the audience for your analysis? What kinds of actions could be taken based on~~

~~your results, and who would be equipped to take those actions?~~

While the U.S. Environmental Protection Agency (EPA) establishes a 15 parts per billion (ppb) action level—requiring water systems to undertake corrosion control measures and to inform the public when 10% of water samples show lead concentrations exceeding 15 ppb—the organization recognizes lead can be harmful to humans even at low exposure levels. In children, low lead exposure has been linked to damage to the central and peripheral nervous system, learning disabilities, shorter stature, impaired hearing, and impaired formation and function of blood cells.[[3]](#footnote-3) In adults, lead exposure can result in reduced growth of the fetus and/or premature birth among pregnant women as well as contribute to cardiovascular complications, decreased kidney function and reproductive problems in both men and women.[[4]](#footnote-4)

In the city of Chicago, a study conducted in 2018 found lead in the tap water of nearly 70 percent of 2,797 homes tested between 2016 and 2017.[[5]](#footnote-5) This same study reported that 30 percent of the samples had lead concentrations of higher than 5 ppb—the maximum level allowed in bottled water by the U.S. Food and Drug Administration. The gravity of this problem was recently acknowledged by the local government when Chicago Mayor Lori Lightfoot announced the first phase of the Lead Service Line (LSL) Replacement Program in September 2020. The program, which aims to replace the city’s nearly 400,000 LSLs, is estimated to cost 8.5 billion USD and take multiple decades to complete.[[6]](#footnote-6)

Local government officials have already recognized the importance of adopting an equity-driven approach, prioritizing low-income residents who (i) own and reside in their home; (ii) have a household income below 80% of the area median income (72,800 USD for a family of four); and (iii) have consistent lead concentrations above 15 ppb in their water, as tested by the Department of Water Management.[[7]](#footnote-7) However, some experts have already criticized this approach for having stringent requirements, noting the current prioritization framework may overlook low-income renters who are traditionally more exposed to lead because their housing is in poor condition.[[8]](#footnote-8) Given the cost and scope of the LSL Replacement Program, it is imperative that Chicago’s city officials adopt a data-driven approach that prioritizes vulnerable communities and allocates resources accordingly.

To identify communities that may be at higher risk for lead exposure in their drinking water, we analyzed water quality data from the Department of Water Management showcasing lead concentrations of residential water samples across Chicago. We complemented our analysis by including socioeconomic and demographic indicators from the American Community Survey’s Five-Year Estimates and aggregate housing variables derived from the Cook County Assessor’s Residential Property Characteristics dataset. We trained classification algorithms on the final dataset combining socioeconomic, demographic and housing assessment indicators to predict high lead exposure. We then used our models to develop a risk profile for the city of Chicago at the census block group level.

We hope our analysis can be leveraged by Chicago city officials—namely those in the Department of Water Management and the Mayor’s Office. Ultimately, we hope the LSL Replacement Program adopts an equity- and data-driven prioritization framework as it launches in the upcoming months. We also hope our analysis helps inform Chicago city residents concerned about lead in their drinking water, as there are control measures that can reduce the risk of lead exposure as the city works to replace the 400,000 LSLs in the coming decades.

1. **Data**

~~Describe the data that you used. Be sure to specify what a row represents in your final dataset (your unit of analysis), and your outcome label (what you’re predicting). Include some descriptive statistics and/or visualizations of your data.~~

The data used in our analysis is all publicly available and can be accessed online via the relevant government organizations. Our primary analysis was conducted at the census block group level for the city of Chicago, with each row representing a block group including the aggregate features described below and an outcome variable classified as 1 if any house in that block group exceeded the lead concentration threshold. We developed two different thresholds (tested separately): a high threshold of 15 ppb and a medium threshold of 5 ppb. The high threshold matches the action level established by the E.P.A., while the medium threshold of 5 ppb matches a lower standard that the U.S. Food and Drug Administration has established as the limit for bottled water. Other organizations also advocate for reducing the EPA lead action level from 15 ppb to 5 ppb to match the FDA’s standard.[[9]](#footnote-9)

*Water Quality Data*

Water quality data featuring lead concentrations (ppb) in water were obtained from the Chicago Department of Water Management (DWM).[[10]](#footnote-10) The dataset features the results of water samples conducted across Chicago residences between 2016 and early 2021. Lead tests are initiated when a resident requests a free testing kit from the DWM. The homeowner gathers the water samples according to the directions provided; if the samples were correctly gathered, the department analyzes the results, reports them to the homeowner, and adds them to its dataset, which at the time of analysis contained over 24,000 observations.

Each observation in the dataset contains three sample readings: (i) immediately after first turning on the tap; (ii) two-three minutes after turning on the tap; and (iii) five minutes after turning on the tap. For our analysis, we considered the maximum of these three readings as the final reading for the corresponding observation. Further, if any water sample returns fewer than 1.0 ppb, the DWM replaces that value with “<1.0”. We replaced any values of “<1.0” with 1.0.

Observations in the original dataset are partly obfuscated, in that the last two digits of a homeowner’s address are replaced with “XX” for anonymization reasons. We imputed these values with “00” such that “13XX E Hyde Park Blvd”, for example, became “1300 E Hyde Park Blvd”, yielding block-level addresses. We then geocoded these addresses to retrieve a sample’s location within a city block. We used this data to construct two outcome variables that we tested separately: (i) *threshold (high)*, coded as 1 if the maximum sample reading exceeded 15 ppb and 0 otherwise; and (ii) *threshold (medium)*, coded as 1 if the maximum sample reading exceeded 5 ppb and 0 otherwise.

*American Community Survey Five-Year Estimates (2019)*

To build our set of features, we first drew on the American Community Survey (ACS) Five-Year Estimates (2019), which contain socioeconomic, demographic, and housing variables at the census block group level. We anticipated demographic features to have predictive power because of Chicago’s redlining history discriminating against Black/African American and other minority communities— making it more likely for these populations to live in lower quality housing.[[11]](#footnote-11)

Indeed, a study by the Metropolitan Planning Council found Black- and Hispanic-majority communities in Illinois to be more likely to live in residences with LSLs relative to White-majority communities.[[12]](#footnote-12) Demographic and socioeconomic variables in the feature set include (i) total population; (ii) median income; (iii) White population (percentage of total); (iv) Black/African American population (percentage of total); and (v) Non-White population (percentage of total, to account for Hispanic/Latino and other non-Black minority groups).

Further, since Chicago required the use of lead service lines until 1986, when the practice was banned by the federal government, individuals living in single-family or two-flat homes built before that year have a higher likelihood of being connected to a LSL (unless it was replaced during a renovation).[[13]](#footnote-13) As such, we also included ACS housing variables in our analysis to capture differences in living conditions across census block groups. These include (i) average household size; (ii) number of occupied housing units; (iii) median gross rent; and (iv) number of owner-occupied housing units.

*Cook County Assessor's Residential Property Characteristics*

Finally, we complemented the ACS feature set with data from the 2020 Cook County Assessor’s Residential Property Characteristics dataset.[[14]](#footnote-14) Filtering for properties in the city of Chicago, we used aggregate indicators at the census block group level for (i) mean/median land value; (ii) mean/median building value; (iii) mean/median land size in square feet; (iv) mean/median building size in square feet; (v) mean/median building age; and a series of one-hot encoded binary variables for property type, wall material, roof material, repair condition, and renovation status of the property.

Of the 1,995,108 total residences in the Cook County assessment dataset, we focus on the 728,543 that occur within Chicago’s city boundaries. The following table shows summary statistics for the assessment variables we have utilized in our models

**Assessment Variables**

| **Column** | **Non-Null Values** | **Null Values** | **Mean** | **Std. Dev** | **Median** | **Min** | **Max** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Prior Tax Year Market Value Estimate (Land) | 728543 | 0 | 45602 | 54501 | 34320 | 0 | 3040940 |
| Prior Tax Year Market Value Estimate (Building) | 728543 | 0 | 225590 | 253792 | 164960 | 0 | 13358380 |
| Land Square Feet | 728543 | 0 | 17032 | 37369 | 4495 | 0 | 7753680 |
| Building Square Feet | 438342 | 290201 | 1875 | 1234 | 1489 | 0 | 19992 |
| Age | 728543 | 0 | 71 | 39 | 70 | 1 | 205 |
| Property Class | 728543 | 0 | -- | -- | -- | -- | -- |
| Wall Material | 438329 | 290214 | -- | -- | -- | -- | -- |
| Roof Material | 438329 | 290214 | -- | -- | -- | -- | -- |
| Repair Condition | 438329 | 290214 | -- | -- | -- | -- | -- |
| Renovation | 1243 | 727300 | -- | -- | -- | -- | -- |

The Cook County assessment data includes information on smaller residential properties that have a property class in the 200-series.[[15]](#footnote-15) The dataset does not include properties categorized as multifamily housing, which will have a property class in the 300-series. In other words, no apartment buildings with 7 or more units will appear in our dataset.

**Property Class Counts**

|  |  |  |  |
| --- | --- | --- | --- |
| **Property Class** | **Description** | **Count** | **Proportion** |
| 202 | One-story Residence, any age, up to 999 square feet | 56905 | 7.81% |
| 203 | One-story Residence, any age, 1,000 to 1,800 square feet | 135822 | 18.64% |
| 205 | Two-or-more story residence, over 62 years of age up to 2,200 square feet | 36312 | 4.98% |
| 211 | Apartment building with 2 to 6 units, any age | 125638 | 17.25% |
| 299 | Residential condominium | 277880 | 38.14% |
| All Others | -- | 95986 | 13.17% |
| **Total** | **--** | **728543** | **100.00%** |

See Appendix 1 for figures that provide the full distribution of residential property classes and property age.

*Descriptive Statistics*

The described datasets were cleaned, wrangled and merged into one final dataset containing aggregate features and outcome labels at the census block group level. The total features and outcome variables are listed in Section 4.

[**Valeria to insert descriptive statistics/graphs.**]

1. **Machine Learning and Details of Solution**

~~What type of machine learning problem is this? Are you developing a classification or regression technique? Clearly articulate the learning that your resulting models enable. What types of models did you apply? Justify your choice of models. Your considerations could include the nature of your dataset (e.g. types and nature of features, size of the data), the requirements for model training or testing (e.g. real-time classification), or any other considerations you might have. What features did you use to train your model? Did you use feature engineering to expand your feature set? If so, please justify.~~

Drawing on the described set of socioeconomic, demographic, and housing features, we trained several classification models to predict whether or not a census block group will observe high lead exposure levels (depending on the threshold used, either (high) 15 ppb or (medium) 5 ppb). Logistic regression, random forest, and linear support vector classification models were all fit to the training data and then evaluated using testing data. Similar modeling approaches have been adopted by other scholars predicting the likelihood of lead exposure in children, including one that draws on home inspections and property value assessment data to predict lead level from blood tests in children for Chicago from 1993-2013.[[16]](#footnote-16) As described in Section 3, our final dataset includes the following features at the census block group level:

Total population; median income; White population (percentage of total); Black/African American population (percentage of total); Non-White population (percentage of total); average household size; number of occupied housing units; median gross rent; and number of owner-occupied housing units; mean and median land value; mean and median property value; mean and median land size in square feet; mean and median property size in square feet; mean and median property age; and a series of one-hot encoded binary variables for property type, wall material, roof material, repair condition, and renovation status of the property.

For the assessment data, we selected these variables on the basis of features that seemed most relevant to predicting the presence of LSLs (i.e. not the size of the garage), were not directly constructed from other features (such as the size of the lot, squared), were not prone to overfitting the data (such as the deed number), and that had the most potential for policy recommendations (i.e. not the basement finish). In addition, Cook County describes to a limited extent the dataset features on its website and says that some are not suitable for analytical purposes, such as the construction quality or site desirability.

We then one hot encoded the assessment data for two reasons. First, many of the features are categorical and have no intrinsic hierarchy; for example, ‘wall material’ takes on the values of 1, 2, 3, or 4, corresponding to wood, masonry, wood & masonry, or stucco materials comprising a residence’s external walls. Second, many of these categorical features are missing a significant number of values for which there is no obvious replacement value. After aggregating these variables at the block group level, we converted them to proportions of the total number of residences within that block group. For any numerical features missing values, we imputed with the feature median. And while not necessary for the random forest model, we normalized all non-target variables to a mean of 0 and standard deviation of 1.

Our models were trained separately on the two binary targets: threshold (high), classified as 1 if any house in the respective census block group exceeded 15 ppb and 0 otherwise; and threshold (medium), classified as 1 if any house in the respective census block group exceeded 5 ppb and 0 otherwise. [**Repeat/insert justification for these separate thresholds**]. [**Insert modeling approach/description**]. [**Do we need to discuss data imbalance and method to mitigate this issue?**] During the model training stage, we used **10**-fold cross-validation to avoid overfitting. We then evaluated our models against the testing set, maximizing the recall metric in order to minimize false negatives. We evaluated models based on the maximum average recall score achieved during k-fold cross validation, as a false negative (incorrectly labeling a block group as having a low risk of lead exposure) likely has a higher societal cost than a false positive. Since the City of Chicago aims to replace all LSLs in the long term, a false positive was deemed less of a concern than a false negative.

1. **Evaluation and Results**

Describe and interpret your results. Include tables and plots where relevant to summarize your models (e.g. precision-recall curves). Describe and justify the evaluation metrics that you chose. Describe the feature importances of your best model(s).

We used three different metrics to evaluate the performance of all our models: accuracy, precision and recall. As noted above, our models were evaluated based on the highest recall achieved—as our objective was to minimize false negatives. [**Insert justification of accuracy/precision metrics inclusion**]. Table X summarizes the results of our analysis. [**Insert analysis of the table, description of the best model, and any potential tradeoffs across models**].

[**Insert table summarizing model results**].

The features with the highest predictive power for our models comprised [**insert discussion of feature importance and any variation across models**].

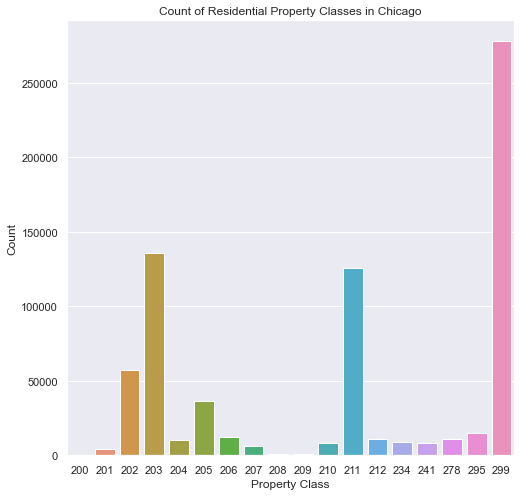
[Insert any additional evaluation/performance insights].

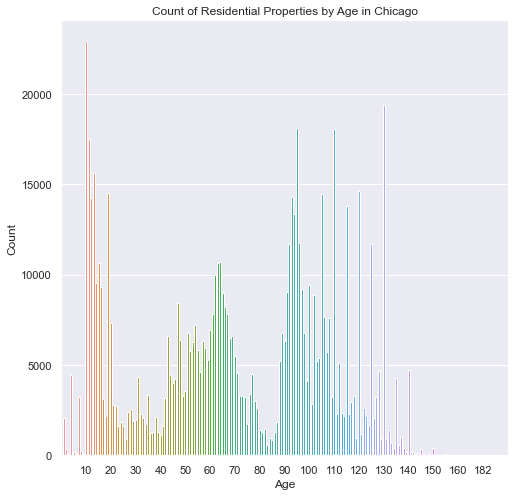
1. **Discussion**

[**Insert conclusion/policy implications/ethical considerations using this model/etc**].

1. **Appendix**

*Appendix 1: Additional Assessment Data Statistics*





1. Paul Caine, “Chicago Has More Lead Service Pipes Than Any Other US City, Illinois the Most of Any State,” *WTTW Chicago*, March, 24, 2021, <https://news.wttw.com/2021/03/24/chicago-has-more-lead-service-pipes-any-other-us-city-illinois-most-any-state>. [↑](#footnote-ref-1)
2. Mayor’s Press Office, “Mayor Lightfoot Launches Equity-Focused Lead Service Line Replacement Program,” Office of the Mayor, September, 10, 2021, <https://www.chicago.gov/city/en/depts/mayor/press_room/press_releases/2020/september/EquityFocusedLeadServiceReplacement.html>. [↑](#footnote-ref-2)
3. United States Environmental Protection Agency, “Basic Information about Lead in Drinking Water”, <https://www.epa.gov/ground-water-and-drinking-water/basic-information-about-lead-drinking-water#health> [↑](#footnote-ref-3)
4. Ibid. [↑](#footnote-ref-4)
5. Michael Hawthorne, “Brain-damaging lead found in tap water in hundreds of homes tested across Chicago, results show,” *Chicago Tribune*, April, 12, 2018, <https://www.chicagotribune.com/investigations/ct-chicago-water-lead-contamination-20180411-htmlstory.html>. [↑](#footnote-ref-5)
6. Monica Eng, “What We Do — And Don't — Know About Chicago's Lead Water Problem,” *NPR*, September 21, 2020, <https://www.npr.org/local/309/2020/09/21/915248028/what-we-do-and-don-t-know-about-chicago-s-lead-water-problem>. [↑](#footnote-ref-6)
7. Mayor’s Press Office, “Mayor Lightfoot Launches Equity-Focused Lead Service Line Replacement Program,” Office of the Mayor, September, 10, 2021, <https://www.chicago.gov/city/en/depts/mayor/press_room/press_releases/2020/september/EquityFocusedLeadServiceReplacement.html>. [↑](#footnote-ref-7)
8. Monica Eng, “Despite A Promise, Chicago Has Made No Progress On Removal Of Lead Pipes,” *WBEZ Chicago*, April, 18, 2021, <https://www.wbez.org/stories/despite-a-promise-chicago-has-made-no-progress-on-removal-of-lead-pipes/02644e18-4cd5-4e7e-b595-3ebc111c62a6>. [↑](#footnote-ref-8)
9. Kristi Pullen Fedinick, “Millions Served by Water Systems Detecting Lead,” *Natural Resources Defense Council,* May 13, 2021, <https://www.nrdc.org/resources/millions-served-water-systems-detecting-lead>. [↑](#footnote-ref-9)
10. Data from Chicago Department of Water Management’s Water Quality Study. Data was retrieved from <https://www.chicagowaterquality.org/home#results>. [↑](#footnote-ref-10)
11. Natalie Moore, “New Redlining Maps Show Chicago Housing Discrimination,” *WBEZ Chicago*, October, 28, 2016, <https://www.wbez.org/stories/new-redlining-maps-show-chicago-housing-discrimination/37c0dce7-0562-474a-8e1c-50948219ecbb>. [↑](#footnote-ref-11)
12. Justin Williams, “Data Points: the environmental injustice of lead lines in Illinois,” Metropolitan Planning Council, November, 10, 2020, <https://www.metroplanning.org/news/9960/Data-Points-the-environmental-injustice-of-lead-lines-in-Illinois>. [↑](#footnote-ref-12)
13. Lead-Safe Chicago, Lead Service Line Replacement, <https://www.leadsafechicago.org/lead-service-line-replacement>. [↑](#footnote-ref-13)
14. Data retrieved from Cook County Assessor's Office, Cook County Assessor's Residential Property Characteristics. Data last updated on November, 27, 2020. Data retrieved via <https://datacatalog.cookcountyil.gov/Property-Taxation/Cook-County-Assessor-s-Residential-Property-Charac/bcnq-qi2z>. [↑](#footnote-ref-14)
15. See the *Definitions for the Codes for Classification of Real Property* available at: <https://prodassets.cookcountyassessor.com/s3fs-public/form_documents/classcode.pdf>. [↑](#footnote-ref-15)
16. Eric Potash et al., “Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning”, KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August, 2015, <https://dl.acm.org/doi/10.1145/2783258.2788629>. [↑](#footnote-ref-16)