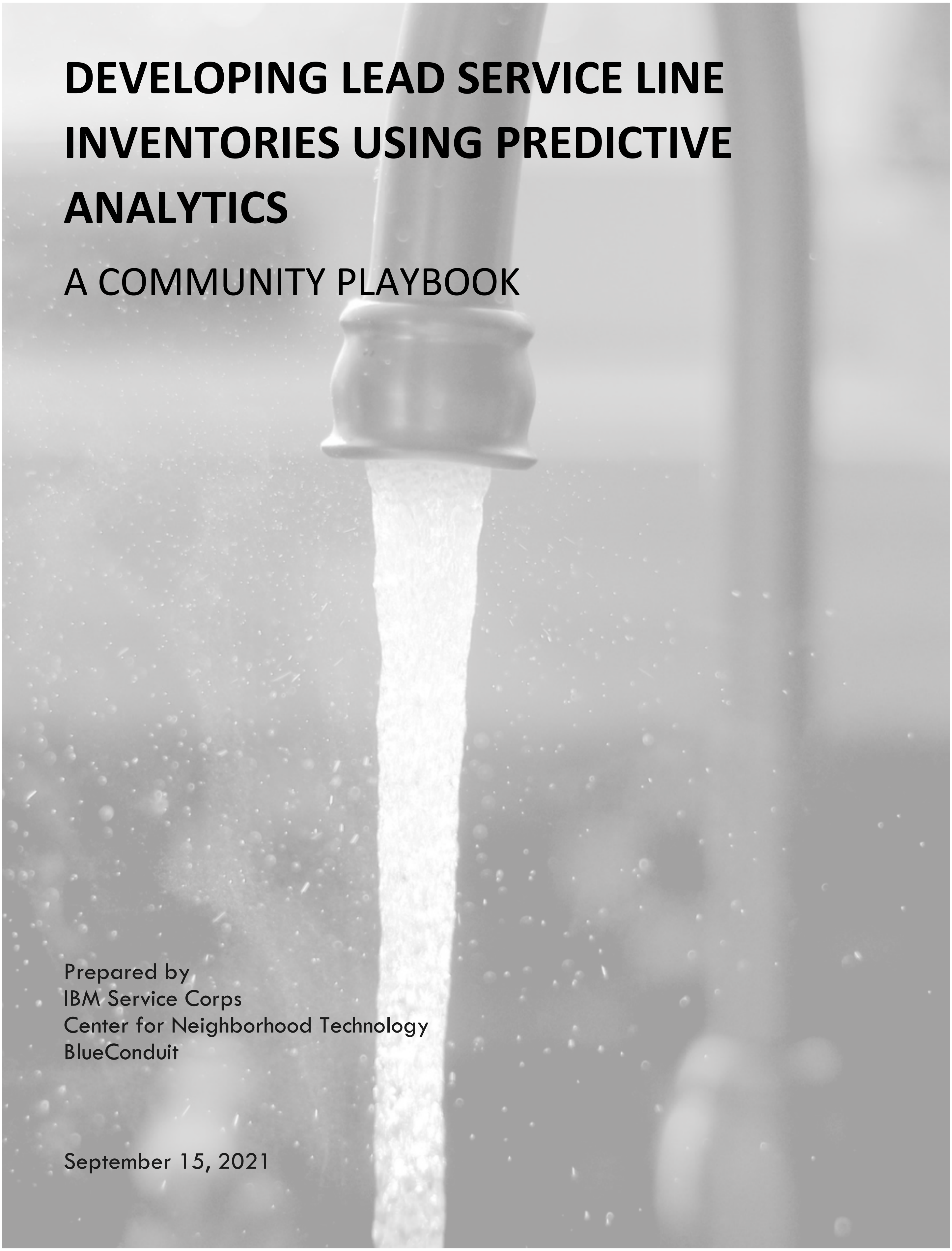


DEVELOPING LEAD SERVICE LINE INVENTORIES USING PREDICTIVE ANALYTICS

A COMMUNITY PLAYBOOK

Prepared by
IBM Service Corps
Center for Neighborhood Technology
BlueConduit

September 15, 2021

ACKNOWLEDGEMENTS

VILLAGE OF HAZEL CREST

Dante Sawyer, Village Manger
Charles Dryer, Director of Public Works
Elliot Eldridge, Director of Inspectional Services
Joe Shupryt, Sewer and Water Division, Public Works Department
Ed Melby, Sewer and Water Division, Public Works Department

VILLAGE OF FLOSSMOOR

John Brunke, Public Works Director
Dan Milanovich, Assistant Public Works Director
Scott Bugner, Building and Zoning Administrator, Building Department

SOUTH SUBURBAN MAYORS AND MANAGERS ASSOCIATION

Sara Kobetis, Director of Planning
Freddy Segura, GIS Analyst
Kai Funahashi, GIS Analyst

OTHER PARTNERS

Justin Keller, Manager, Metropolitan Planning Council
Cheryl Watson, Blacks in Green

ABOUT CENTER FOR NEIGHBORHOOD TECHNOLOGY

The Center for Neighborhood Technology (CNT) is a leader in promoting more livable and sustainable communities. CNT's mission is to make cities work for everyone. CNT works at the intersection of environmental sustainability, social equity, and technology — with particular attention on creating efficient and affordable solutions for low-income communities and communities of color.

ABOUT IBM SERVICE CORPS

For more than ten years, IBM's Service Corps has given employees the opportunity to use their professional skills to help people and communities tackle complex issues. Small squads of IBMers partner for several weeks with nonprofit, government, educational and civic leaders to address high-priority issues in education, sustainability, health, and economic development. Six IBM staff from the Chicago region supported CNT during the development of this playbook and water service line inventories for the Villages of Hazel Crest and Flossmoor

ABOUT BLUECONDUIT

BlueConduit is a water infrastructure analytics consulting company that uses data and machine learning to help cities do service line inventories and removal. BlueConduit pioneered this use of predictive modeling to help the City of Flint, Michigan, efficiently replace its pipes. These methods saved the city tens of millions of dollars and accelerated the removal of dangerous infrastructure. BlueConduit played an advisory role on this project, providing insights and support to the CNT and IBM teams during the data collection, cleaning, and analysis phases of the project.

CONTENTS

About **3**

Primary Steps **6**

Step 0: Understanding the Environment 7

Step 1: Data Identification 8

Step 2: Data Gathering 9

Step 3: Data Ingestion 10

Step 4: Analyze and Data Visualization 11

Step 5: Use and Maintain 13

Step 6: Leverage the Data 13

Tips and Tricks **14**

Initial Questions for the Community 14

data Collection and organization 14

Optical Character Recognition (OCR) 16

Visualization Examples 16

Key Takeaways **17**

Case Study: Developing a Lead Service Line Inventory Using Predictive Analytics for the Village of Hazel Crest **18**

Appendix A: Sample set of Parcel Level Variables **22**

BACKGROUND

Lead service lines are pervasive throughout US water supply infrastructure due to the material's popularity as a plumbing material for centuries. In light of recent public health crises and, at this point, a nearly universal acknowledgement that there is no safe level of lead exposure, communities of all sizes are working to understand the prevalence of the issue and level of investment needed to get the lead out of their systems. Many mid-sized and larger cities, like Newark, NJ, and Cincinnati, OH, are implementing lead service line replacement programs, dedicating resources and using novel approaches to tackle this huge and urgent local issue. However, getting started may be a daunting task for communities of any size, but especially for those that are smaller, older, and capacity constrained. The first step in developing a replacement strategy is to develop a baseline or an inventory. This can prove challenging when local data is on lead service line locations is unreliable, outdated, or missing. And it can be expensive to go house to house to verify pipe material, without having a sense of where to prioritize verification.

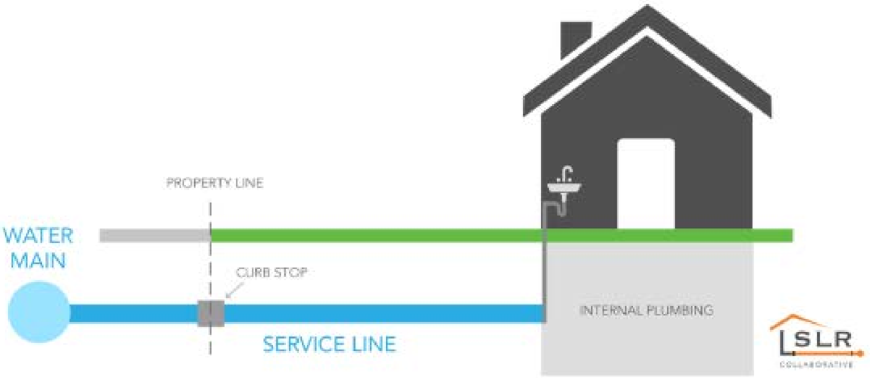
Predictive models and data-driven methodologies can be applied to more accurately predict the location of lead services lines, thus saving communities money by improving field verification and removal efficiency. This Community Playbook (Playbook) provides a framework for collecting data, applying a version of the lead service lines predictive analytics developed by BlueConduit, and provides public usage ideas for small to medium communities around developing lead pipe service line inventories.

The Playbook is based on the body of work completed by the Center for Neighborhood Technology (CNT), BlueConduit, and IBM, between February 2020 and August 2021, in partnership with two Chicago suburbs, the Villages of Hazel Crest and Flossmoor. The team worked with each community to collect municipally owned data (e.g., building permits), data available from the Cook County Assessor's office (e.g., housing age, value, and type), and IL EPA service line testing data to develop a probabilistic service line inventory, based on key housing characteristics, to help each municipality get a preliminary and predictive sense of the extent of lead service lines throughout their community.

While the Playbook is informed by lessons learned while working with small communities (5-15k residents), the collection of insights (which questions to ask, which data to gather, different data analysis methods) will prove helpful to communities of any size as they set out to apply predictive tools in the development of service line inventories.

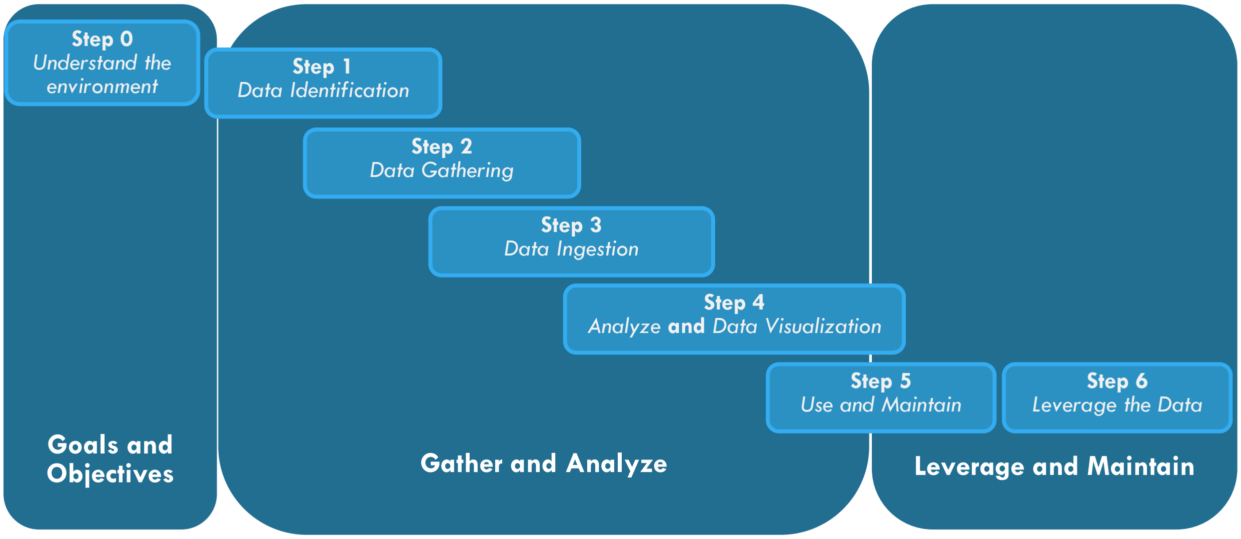
DEFINITIONS

Term	Definition
PIN	Property Identification Number – a unique identifier typically used by assessors to identify a property location
OCR	Optical Character Recognition – ability to scan and identify text and handwriting from a document and store it digitally
Models	Analytic method of defining how to interpret data
Predictive Analytics	Use of algorithms, models, and data – in this case used to predict where lead may be prevalent in a town or by address
Water Supply Service Line	The water supply service line runs from the water main to curb, or property line, and then from the curb to home. Many municipalities have records for the mains and to the curb, but the private service line material to the home may or may not be documented. Service lines may be lead, copper, galvanized metal, PVC, or some combination

	 PROPERTY LINE WATER MAIN CURB STOP INTERNAL PLUMBING SERVICE LINE SLR COLLABORATIVE
Ground Truth	Information from direct observation and not from inference. For water service lines, it is seeing or testing the materials the pipes are made of at the site, as opposed to inferring the materials based on the age of the construction or other data.

PRIMARY STEPS

There are three main stages to developing an inventory using predictive analytics: Goals and Objectives (the stage to get everyone on the project on the same page about expected outcomes and project goals); Gather and Analyze (the stage during which data are named, found, cleaned, and analyzed); and Leverage and Maintain (the stage during which maps are created and are actively used to support field testing and replacement). The following section outlines the questions and considerations a community should consider at each stage of the inventory development process, from Step 0 – Understanding the Environment through Step 6 – Leverage the Data. The team reflected on the questions asked/options considered during its work with the Villages of Hazel Crest and Flossmoor to inform the following considerations.



Step 0
Understand the environment

Step 1
Data Identification

Step 2
Data Gathering

Step 3
Data Ingestion

Step 4
Analyze **and** *Data Visualization*

Step 5
Use and Maintain

Step 6
Leverage the Data

Goals and Objectives

Gather and Analyze

Leverage and Maintain

STEP 0: UNDERSTANDING THE ENVIRONMENT

1. What are the needs of the community for building knowledge of lead water service lines?
    - Identify whether there are current public health concerns actually or potentially connected the community's water supply. Define the urgency/timeframe.
    - Define the scope of the analysis – water mains, service lines to houses/buildings, public and/or private ownership, lead only, or other materials subject to replacement
        - Are you dealing with residential property only, or small businesses, like daycares? Schools? Other public buildings?
    - Goals of the project – planning for replacement schedules, public knowledge/community engagement, community trust building, fulfillment of state EPA reporting requirements, anything else?
    - Is there a source of Ground Truth or a plan to verify the predictions and update the records as lines are confirmed or replaced?
2. Size of community
    - Does predictive analysis make sense for the size and age of your community?
        - If the number of homes with lead service lines are easily tracked manually or understood without inference, you may not need to build the system.
        - If the community has been totally built after 1990, you probably do not have many or any lead pipes
        - If you are expecting growth, adding mapping capabilities may help in many more ways than lead line identification. Are there other activities that might benefit from data visualization efforts (i.e., other infrastructure investment projects, built and natural asset mapping)? Identifying multiple uses for a mapping tool might open up new streams of funding to support the effort.
    - Can you team with surrounding towns? Regional groups? Is there a council of governments (COG) or a local metropolitan planning organization (MPO) in your area that could help coordinate partnerships or resource sharing?
3. Identify which groups are responsible for water services
    - Water line resources – city engineers, private water line companies, city or area water bureaus, building inspectors, utilities
    - Who has knowledge (and access) of existing data sources?
    - Who has knowledge of the data – origins, accuracy, updates?
    - Have there been meter replacement projects or water testing projects that may have identified lead service lines?
    - Have any lead service line identification projects been run and what were the results?
    - How is the town/village reporting lead service lines to the EPA or state EPA? What processes have been used to gather the data today, and how accurate is the data?
4. Age of homes/subdivisions/areas
    - Older homes/areas will tend to have more lead.
    - Use assessor's information to identify homes older than the non-lead requirements of the town or federal law.
    - Anything older than the 1980's should be higher priority
5. Who will maintain the digital information and where? Do they have the data analytics and mapping skills to maintain the data and run the analysis?
    - Is there a regional shared resource that could be used? Again, check with the local COG or MPO to see if they can offer support
6. Is there a Go-No Go decision to make?
    - Understand the minimum set of data requirements to apply a predictive analytics model (e.g., the BlueConduit model). Can these minimum requirements be met? Are there other variables that might be predictive in your community?

- Create backup plan if you cannot collect sufficient data to support the development of a predictive analytics informed inventory (e.g., it could be as simple as prioritizing testing and replacement based on ages of homes; or a community-wide survey with field-testing follow ups)
- Is there a requirement and budget to perform non-invasive or house to house checks?
    - For example – what are the cost factors associated with digging up every pipe?
    - Is basic info such as age of homes/ known problem areas enough to provide you with a starting point and subset/sampling for digging?
- Need to build the most accurate picture with the most efficient use of resources, with public safety as top of mind.

STEP 1: DATA IDENTIFICATION

1. Based on the goals, identify the data needed
    - Leverage the list of data elements from existing predictive analytics models (e.g., BlueConduit's Model[1])
    - Identify what additional data may be predictive in your community
    - Highlight which data is most predictive, i.e., critical to have and most important- these are the highest priority items to find.
    - Put the checklist in a priority order based on what has been seen in other communities. Example: Building Permits are a really good source of data
    - Leave open spaces for other data that the community may have and want to add
    - Indicate what was used for Ground Truth
    - Age of the homes is generally a good datapoint, but be sure to understand what the "age" is referring to – is it the initial build? Is it renovation/rebuild?

2. Using the checklist, identify what data you have available. See Appendix A for a list of the variables the team considered during its work with Flossmoor and Hazel Crest. Understanding how utilities are tracked in the town is a great place to start. Here are some suggestions on sources. Each community may store data in different places and different formats.
    - County records - County assessor's office / Website / Other government agencies like Fire or County recorder of deeds.
    - Is there an aggregated source that is free? Inexpensive? If not, where else to look?
    - EPA documents such as inventory of service lines, water quality - FOIA requests
    - Paper documents and forms such as building permits - local city office, county offices
    - Local building / Construction codes (hint: include history, this will tell you when the laws changed to require non-lead pipes)
    - Building records / inspection records
    - Water supply side vendors (house to curb)
    - Builders, plumbers (curb to house)
    - Any work already done to identify lead pipe issues - other ways in use (Example: HydroPak is a non-invasive method to identify pipe materials)
    - Water meter replacement records
    - Demographics – assessors' information, city building and subdivision records

[1] Abernathy, et al. (Sept 2016). Flint Water Crisis: Data-Driven Risk Assessment Via Residential Water Testing. University of Michigan – Ann Arbor. https://michigandatascienceteam.github.io/assets/files/flint-water-crisis.pdf

- Multi-tenant buildings vs single family – single water feed into building for multiple addresses
- Business/Commercial property water feeds
- Do you want to use data on women and children to prioritize where to tackle first? (If your town is looking to prioritize which areas in which to replace lead pipes, you may want to think about demographics)
- Keep in mind that other data may be useful even if it's not on this list.
- Be thinking of how to get to Ground Truth

Interacting with Municipal Departments - Interviews/Surveys

Over the course of an inventory development project there are many different interactions between the team creating the inventory (could be an intra-municipal department) and other departments throughout the municipality. One recommended communication method is interviews with the village staff that is responsible for utilities and infrastructure improvements.

Who should you interview?

- Village decision makers
- Public works, buildings department employees
- Public health officials
- Schools

When should interviews be conducted?

- Can you meet prior to any data collection?
- Can you set up multiple meetings with the interviewees after initial interviews to validate understandings and close gaps?
- Should you meet with different / smaller groups to allow people to speak freely?

Discussion points during interviews

- What output or tools are you looking for?
- How do you hope to use the output?
- What benefits are you looking for related to the output?
- Who is your audience?
- What will the data be used for?
- What criteria must be met prior to publishing data?
- Have you considered long term use?
- What ideas do you have for data maintenance?
- Do you have a vision for presenting the data?

STEP 2: DATA GATHERING

1. Go through your records and match what you have to the checklist. Evaluate source docs for:
   - Quality
   - Completeness
   - Digital versus paper

- Gaps - what is missing and how to find it and complete it, if possible
- Updates – were lines replaced when homes rebuilt/upgraded? Mains replaced?
- Assess dates/eras – did forms change over time? Did laws change over time?
- Agree on sampling or full data collection - this will be based on desired uses and budget for scanning/analysis, accuracy – would a subset of homes on a block be representative of the full block?

2. How do you determine if a dataset is useful?
    - Do the sources have information about pipes, plumbing, date of construction or other fields in the checklist?
    - Are they already digital?
    - Are there forms used that can be OCR ready?
    - How old are the documents and how good is the quality (legibility, completeness, accuracy, etc.)?
    - Ideas of other data that may help
    - Do you have a Ground Truth or any plans to get it?
    - Is there a single identifier per address? Sometimes as communities grow and change, street names and sometimes addresses change. Can you identify those situations on your records?
3. Gather the data sources
    - Digital
    - Paper
    - Interviews
        - City inspectors, city engineers, water utility teams, water main construction companies, plumbers, builders
        - Ride-alongs, anecdotes, questions on changes to an area, construction history can all be used to identify high priority areas
    - Fill in the gaps manually
    - Ground Truth: how can this be established if not known?
    - Cost/Benefit analysis – is the cost of obtaining and maintaining the data worth the effort required?
4. Are there any privacy concerns or rules that you need to take into account?

STEP 3: DATA INGESTION

1. Create the digital repository – database, spreadsheets, mapping systems
2. Data cleansing/normalization as needed
3. Methods to data ingestion
    - Load existing datasets
    - Use Optical Character Recognition (OCR) on paper documents to convert handwritten / scanned text to analyzable values and load data
        - Do the source documents have a set form? Note that OCR works best with repeatable forms.
        - Do the source documents have handwriting and typed information? How free-form is the information?
        - Refer to the OCR tips and tricks section starting on page 16.
    - Manual updates if needed
4. Data organization
    - Databases
    - Key fields – identify by address? Assessors' property id?
    - Build the database on the biggest available dataset
    - Examples of key information and potential sources

    - Assessor's data (PIN)
    - Hydrants (latitude, longitude; nearest neighbor - hydrant to any given parcel)
    - Meter replacement information (address; PIN)
    - Water main replacement (houses along street with replacement; assign by PIN)
    - LCR samples (PIN)
    - Some building permits (PIN)
        - large format drawings
        - permit forms – note that forms may have changed over time, group them together for better ingestion
        - wild card – unknown what these will contain
    - ACS (block group)
    - Zoning maps - assign zoning to each parcel
    - Community surveys (PIN)
- See Table 1 for a sample database structure and management (each property should have a unique, non-address ID)

5. Consider data limitations
    - Different datasets will have different levels of reliability, depending on how, when, and by whom it was collected
    - "Verified" data are always subject to interpretation – verified by whom? What has changed since data were verified?
    - Data are helpful, even if not 100% accurate. Developing the initial database is a way to organize data, regardless of how (in)accurate or (in)complete the datasets are. It's a necessary starting point toward developing a fuller picture of what a community has and still needs to gather

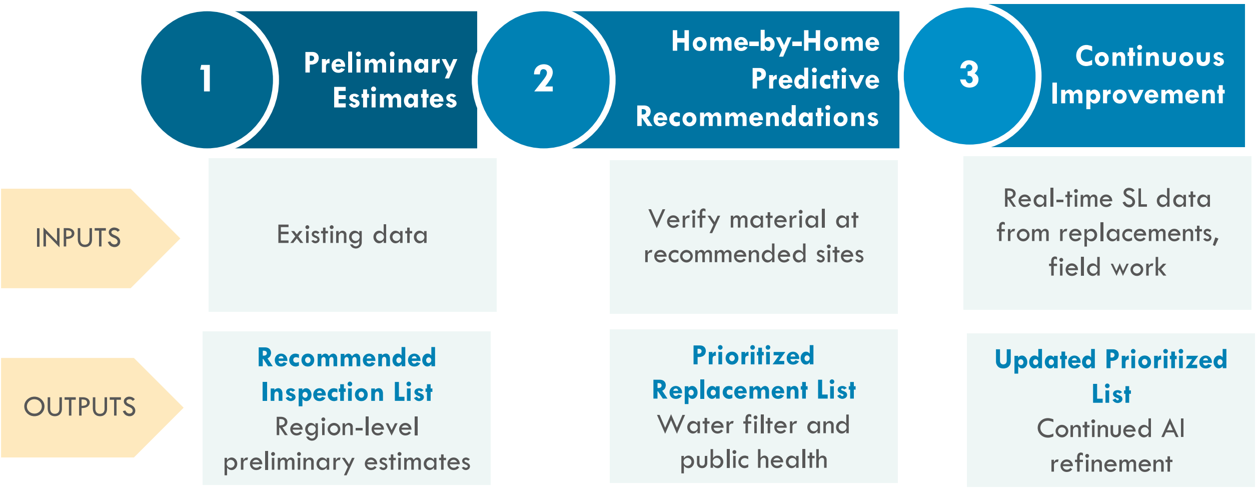
Table 1 Sample Property Data and Organization

Service Line Verified (SL)* (Private)	SL Verified (Public)	Parcel ID	Address	Year Built	Latitude	Longitude	Home value	SL Historical Record (Private)	SL Historical Record (Public)	Water Test Results (ppb)
Galvanized	Copper	401446	6201 MAPLE	1925	43.00802	-83.6365	7800	Copper	Copper	0
?	?	401215	1234 MAIN	1957	43.07282	-83.7219	19300	Copper	-	2
?	?	410545	567 WALNUT	1921	43.02797	-83.7171	10900	Copper	Lead	8
Copper	Copper	401438	1486 OAK	-	43.02454	-83.6655	4000	Copper	Copper	-
Lead	Lead	462645	2014 ELM	1970	43.05868	-83.725	17700	-	-	-

STEP 4: ANALYZE AND DATA VISUALIZATION

1. Data analysis
    - Apply models to the data; leverage opensource algorithms and approach to start with, modify as needed for better results
    - Confirm accuracy of the results by leveraging your Ground Truth information
    - Update training models as needed
2. Blue Conduit employs a three-phase approach (see Figure 1):
    - Preliminary estimates (i.e., what are the existing data and what is the recommended inspection list)

- Home-by-Home Predictive Recommendations
    - Get ground truth at recommended sites and update prioritized replacement list
- Continuous Improvement
    - Get real time service line data from replacements/field work and continue to update prioritization list as you feed new ground truth back into the model



	1 Preliminary Estimates	2 Home-by-Home Predictive Recommendations	3 Continuous Improvement
INPUTS	Existing data	Verify material at recommended sites	Real-time SL data from replacements, field work
OUTPUTS	**Recommended Inspection List** Region-level preliminary estimates	**Prioritized Replacement List** Water filter and public health	**Updated Prioritized List** Continued AI refinement

Figure 1 Blue Conduit Approach

3. Data visualizations
    - Mapping outputs are the most efficient for this type of data. Please refer to the tips and tricks for examples from other communities
    - Before designing the visualization/s, several questions should be answered:
        - Use/Purpose:
            - What is the goal of your visualization?
            - How will it be used?
        - Audience:
            - Who is the audience for the outputs? Why?
            - How will they use the map? Will they be able to update/edit certain features of the map? Or will it be view only?
            - Are there multiple audiences? If so, do there need to be multiple interfaces?
            - Privacy concerns? Do you show the parcels using a property identification number or address? How do you balance privacy and transparency?
        - Data:
            - What data is required to build the feature/s that you like?
            - If you can't obtain all the data points, will the visualization still be useful or do you need to modify the design?
            - Are there other mapping overlays for other information that may be important for the audience?
            - What boundaries will be shown? Village/Town? Neighborhood/subdivision?
        - Several communities have existing visualizations. Of these, what do you like, dislike, want to replicate or avoid?

STEP 5: USE AND MAINTAIN

1. Feedback from the audiences on the effectiveness of the information
2. Ground truth – update the data as you collect ground truth
3. Public access to data – refer to the sample visualizations
4. Continue updating the model based on what is most effective/needed in your community
5. Confirm that the usage and value meets your original goals (public health; modify outputs as needed
6. Create a process for new data collection
    - Consider how the will municipality collect and incorporate data from new construction, subdivisions, rebuilds, main replacements
7. Consider how to organize your database and data so that it is actionable and can be used for other potential purposes (best data management practices)
    - Create a process for automated reports to comply with local, state and federal reporting requirements
8. Develop a representative recommended inspection set based on the inputs in the model (refer back to Figure 1)
    - The representative set should be representative of the whole system – should encompass all service line types, ages of homes, parcels throughout community, property value
    - Inspections and verification to augment existing datasets
    - Getting the biggest data gain for the effort
9. Active learning
    - Feed the ground truth data back into the model to train it based on the new information. This simply means updating the parcel level data in the model based on field testing. For example, if a parcel in the model was predicted to be lead and the ground truth investigation shows that it is lead, update the model to say that the parcel is now confirmed lead. This new data improves the accuracy of the model's predictions, further increasing efficiency in its outputs. The predictions will continue to increase in accuracy as the ground truth dataset grows

STEP 6: LEVERAGE THE DATA

1. Now that you have this data, are there other things you can do with it?
    - Did you gather other data that can be used for other uses?
    - Are you getting other requests from public/private entities?
2. Are there other community projects that the maps/data can support?
3. How will you use the data and mapped inventory to improve public communication and transparency?
    - Consider creating a dedicated website
    - Ensure that there are interim public health measures available to residents (i.e., if someone knows their property has a lead service line, but it won't be replaced for a year, ensure they have access to filters and water use instructions to keep themselves and their family safe
    - Consider partnering with other communities on communications strategies
4. Highlight what your community will be able to do with the tool or how it will benefit them
    - E.g., System-wide predictions
        - Budget estimate based on number of service lines predicted to be lead
        - Apply for grants and other funding
        - Public health communication prioritization

TIPS AND TRICKS

There are several aspects to account for when developing a water service line inventory, some of which may seem obvious (who is the audient and what is the purpose) and others more subjective and case specific (what are the particular data useful in your community). The following section outlines some questions to consider when beginning a service line inventory process.

INITIAL QUESTIONS FOR THE COMMUNITY

1. Initial sample questions
    - Who is the audience (town/community vs. Individual homeowner) and what are they expecting to use the tool/data for?
    - What are the uses for the information?
    - What output or tools do you expect?
    - How do you hope to use the output?
    - What are the benefits of each use/tool?
    - Based on your conversations with CNT so far, what are you expecting to receive?
    - How many homes must have known data before publishing online?
    - What format would data need to be in for the next phase of digital representation?
    - Will this base map be used for other information?
2. Example use cases for town
    - Tracking lead service line replacement
    - Reporting data to state EPAs and/or US EPA
    - Other infrastructure improvement purposes (streets, mains…), etc.
    - Homeowner education (buyers/sellers/real estate)
    - Broader mapping efforts (natural and built assets)
3. Phased Approach needed? Start with most pressing needs and build upon it over time
    - For town use only, what information would be useful?
    - EPA: What information/reports are needed?
    - External users: homeowners, contractors, realtors

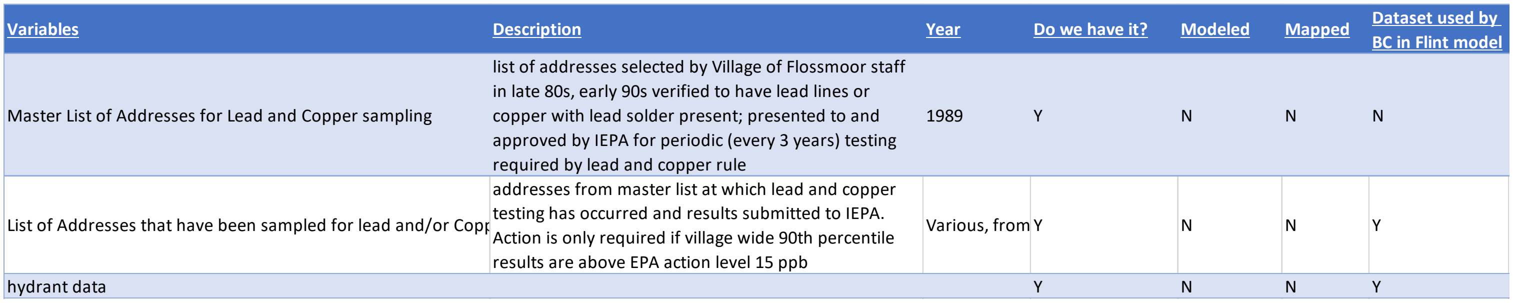
DATA COLLECTION AND ORGANIZATION

1. Data needs and tracking
    - Demographics (optional): General Population Characteristics – Total Population, Number of housing units, Household Income, Race and Ethnicity, Age Cohorts, Education, Languages Spoken at home
        - May be used to identify areas in the town that may be more prone to having lead service lines. Also useful to see if there is any correlation between presence of lead service lines and particular socioeconomic variable
    - Housing Types: Housing Age (by decade), Housing types
        - Older homes tend to have more lead service lines, especially those built before lead service lines were banned in
    - Water Supply System information – what do you know about regional or state regulations/requirements for the water supply system (water loss audits, service line inventories, etc.)? Are you already collecting data that might fulfill these requirements? If not, what and how might you begin collecting to support the development of an inventory?

Like many states, the Illinois EPA (IEPA) requires all communities to submit an inventory of their service lines. Following is a list of data points Illinois requires for the inventory. Knowing these requirements may help communities plan for what to collect as they build out at a database / inventory.

- Water Service Name
- Water Service Type
- Owner Type
- Service Line Ownership (public, private, combination)
- Primary Water Source
- Population Served Count
- AWWA Category
- Established Post 1986
- County Illinois LSL 2018 Date
- Total Service Connections – IL LSL 2018
- Wholesale Connections – IL LSL 2018
- Retail Connections – IL LSL 2018
- Lead Services IL LSL 2018
- Lead Services percent
- Unknown Services IL LSL 2018
- Unknown Services percent
- Copper-Lead Solder Services IL LSL 2018
- Copper-Lead Solder Services percent
- Galvanized Services IL LSL 2018
- Galvanized Services percent
- Unknown Not Lead Services IL LSL 2018
- Copper No Lead Solder Services percent
- Cast Ductile Iron Transite Services IL LSL 2018
- Plastic Services IL LSL 2018

2. Data Inventory Tracker: Start with the list of variables/data needed and identify if you have it and if the model needs it, this will help define the gaps and the model to be used for analysis. It is also used to track what has been included in your outputs to date. Example:



Variables	Description	Year	Do we have it?	Modeled	Mapped	Dataset used by BC in Flint model
Master List of Addresses for Lead and Copper sampling	list of addresses selected by Village of Flossmoor staff in late 80s, early 90s verified to have lead lines or copper with lead solder present; presented to and approved by IEPA for periodic (every 3 years) testing required by lead and copper rule	1989	Y	N	N	N
List of Addresses that have been sampled for lead and/or Copp	addresses from master list at which lead and copper testing has occurred and results submitted to IEPA. Action is only required if village wide 90th percentile results are above EPA action level 15 ppb	Various, from	Y	N	N	Y
hydrant data			Y	N	N	Y

OPTICAL CHARACTER RECOGNITION (OCR)

There are many options to select for OCR, some open source, some paid services. Each starts with a scanned or digital document. Then the OCR software, typically using a pre-defined template, interprets the documents and stores the data in some form of output (spreadsheet, database, json, etc.) that can then be used by the model/algorithms to analyze and map. Here is a subset of potential OCR tools that can be used. This is not an exhaustive list, but the ones we tested during this assessment.

1. Template Matching (Open CV)
   - Works well when the data to be extracted across images (e.g. map data) is consistent but the structure of each image varies
2. Proprietary OCR tool (Example: Vidado)
   - Ideal for structured forms
   - Must know what information you want to extract and where it resides on the form
   - Works on typed text as well as handwriting
3. Free OCR tools
   - Tesseract – Open sourced OCR tool
     - Works well with typeset characters in varying locations but not handwritten characters
     - Good for extracting typed text in unstructured documents
4. Topic Modeling
   - When a community contains too many forms to sift through manually it difficult to identify relevant forms to apply any of the above methods.
   - Instead we can look at all the forms associated with a home and try to group homes based on the similarity of forms attached to each home.
   - Topic modeling allows us to represent each home as a distribution of topics (which is itself a distribution over words) with respect to the words that occur across all documents for a given home. Such a distribution may be used as features in building the predictive model

VISUALIZATION EXAMPLES

City/State	Site
Flint, MI	https://www.flintpipemap.org/map
Indiana	https://www.edf.org/health/mapping-lead-pipes-water-utility
Barrington, IL	https://www.barrington-il.gov/government/departments/public_works/programs___services/lead_copper_water_service_information.php
Washington, DC	https://www.dcwater.com/leadmap
Wheaton, IL	https://www.wheaton.il.us/253/Lead-in-Drinking-Water
Evanston, IL	https://evanston.maps.arcgis.com/apps/webappviewer/index.html?id=4da43f5f440d46cc8c330c69666e99fe
Moline, IL	https://www.moline.il.us/1067/Lead-and-Service-Line-Information
Naperville, IL	http://naperville.maps.arcgis.com/apps/webappviewer/index.html?id=342c782b2d2e46b68102e6f4f3a38833
Rockford, IL	https://rockfordil.maps.arcgis.com/apps/webappviewer/index.html?id=cd512abcf49445e686ea8e090865b5f0
Reference	IMPJ_121-139_Municipal_Strategies_Lead_Service_Replacement.pdf, pages 14-15

KEY TAKEAWAYS

The above steps and strategies, while not exhaustive, are meant to be a guiding set of questions and considerations for communities embarking on a lead service line inventory and replacement plan journey. Even those communities that do not use predictive analytics to inform a preliminary inventory may still find many of these recommendations useful in getting started. Above all, communities should remember:

1. Goals and limits should be set at the outset of a project
   Define the goals of the program, and specifically goals for data usage early. This will dictate the data requirements, the output required and the audience for the analysis. Use a data inventory spreadsheet to track what information is needed, what data you have, what additional data you need to find, and the level of effort to find it. It's also critical to understand the limits of data collection – some data simply won't be available. This need not mark the end of the project, but will inform how you proceed, specifically what resources you will need to find and track data, and other continued support needs.
2. Data Accessibility and quality are key
   Finding data sources takes some investigation and understanding the quality and relevance of the sources takes patience. Many communities will have only paper records, and they may be in different formats and different quality. Understand your budget for scanning, OCR and weight the value of manual versus automated action (size and relevancy of the paper documents)

3. Decisions should be made quickly and with flexibility and agility
   Build in several checkpoints throughout the project to iterate on the worthiness of time spent doing something. In the early stages, reflect on community involvement and support – are you taking an equitable and authentic approach to community engagement? Are there issues being raised by community members that you need to respond to? During the data analysis stages, consider the data quality and depth of scan (are certain methods worth the effort). When do you shift from manual to automated: size of effort, cost of solutions? Can you obtain enough useful data to make the modeling process worthwhile and effective? Do you have a process to define Ground Truth?

Taking these aspects into consideration early and often throughout the process can support a thoughtful, adaptable, and thorough approach to developing a service line inventory using predictive analytics.

CASE STUDY: DEVELOPING A LEAD SERVICE LINE INVENTORY USING PREDICTIVE ANALYTICS FOR THE VILLAGE OF HAZEL CREST[2]

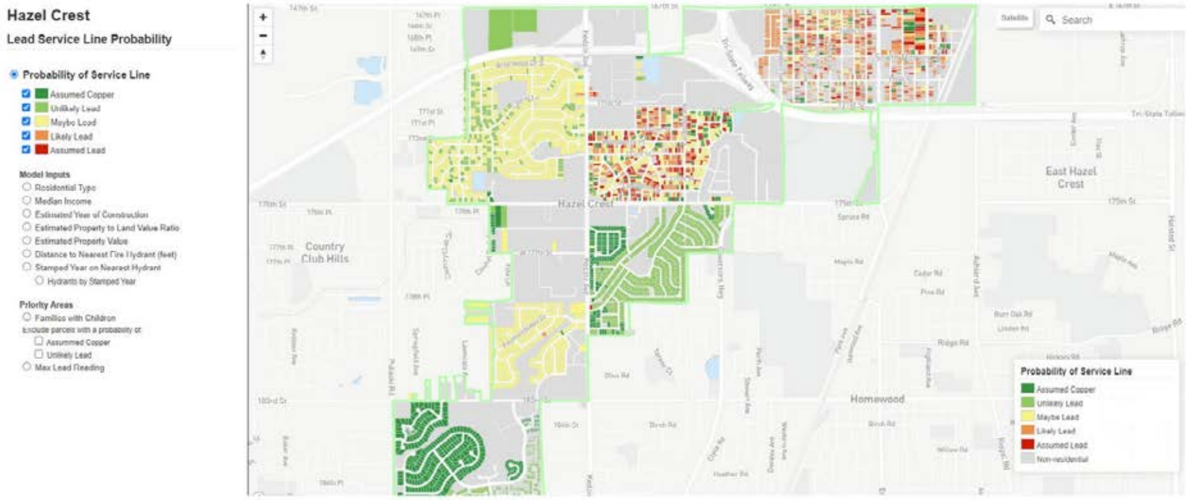
INTRODUCTION

The Center for Neighborhood Technology (CNT) worked closely with the Village of Hazel Crest, MPC, IBM Service Corps, and BlueConduit to develop a probability-driven inventory of lead service line locations throughout the municipality. The team collected and analyzed data from the Cook County Assessor, IEPA, and the Village of Hazel Crest to design and build a predictive model using a multiple linear regression analysis that determined whether the presence of a lead service line is likely or unlikely at each residential parcel. CNT created an interactive map that presents the probability for each parcel. Fig. 2 shows a screenshot of this interactive map. Checkboxes allow the user to visualize the model inputs or selectively turn on/off specific inputs. The map also includes replacement priority areas, e.g., parcels in block groups with a high percentage of families with children.

METHODOLOGY

Developing a probability-driven lead service line inventory requires an initial evaluation of available data, a designation of parcels as dependent variables, or "labels" (i.e., parcels where there is service line material data), the assignment of independent variables, and application of multiple linear regression analysis to develop probabilities for the rest of the parcels for which there is no service line material data.

Fig 2 Screenshot of Hazel Crest Lead Service Line Probability interactive map



Hazel Crest

Lead Service Line Probability

● Probability of Service Line
- ☑ 🟩 Assumed Copper
- ☑ 🟩 Unlikely Lead
- ☑ 🟨 Maybe Lead
- ☑ 🟧 Likely Lead
- ☑ 🟥 Assumed Lead

Model Inputs
- ○ Residential Type
- ○ Median Income
- ○ Estimated Year of Construction
- ○ Estimated Property to Land Value Ratio
- ○ Estimated Property Value
- ○ Distance to Nearest Fire Hydrant (feet)
- ○ Stamped Year on Nearest Hydrant
  - ○ Hydrants by Stamped Year

Priority Areas
- ○ Families with Children

Exclude parcels with a probability of
- □ Assummed Copper
- □ Unlikely Lead
- ○ Max Lead Reading

DEPENDENT VARIABLES

In 2017, the Village of Hazel Crest mailed a paper survey to every residence in Hazel Crest asking residents to verify their water service line material. To guide residents' decision-making, the survey included a picture of what copper versus lead service lines look like. The village received 452 responses, which represented a response rate of approximately 10 percent. Of these, 153 respondents identified their line as copper, 286 respondents identified their line as lead, and 12 respondents were unable to confirm the material make-up of their service line.

[2] Excerpt from a December 2020 report by Metropolitan Planning Council and CNT, "Village of Hazel Crest Lead Service Line Inventory and Replacement Plan" https://cnt.org/sites/default/files/publications/Hazel_Crest_Lead_Service_Line_Replacement_Plan.pdf

In addition to the community administered survey, Hazel Crest's Public Works Department has a list of 85 high-risk sample sites from which they sample 30 homes every three years. This is based on IEPA's administration of the Lead and Copper Rule, which requires water utilities to identify properties that have a higher risk for elevated levels of lead and/or copper and submit the list for approval by IEPA. Upon approval, the utility is required to sample these homes every three years for lead and copper levels. In the model, parcels with confirmed presence of lead or copper — based on the community administered survey results and the IEPA Approved Sample Sites — were set as the dependent variables, or "labels." Age of construction was also used as a label. Properties built after 1991 are assumed to have non-lead service lines because the Lead and Copper Rule was passed in 1991.

As labels, these data were used to train the model based on the relationship between the label and independent variables of property characteristics, such as property age and property value. Once a relationship between the label and independent variables was established, a probability of presence of lead was assigned to other parcels, based on similar or divergent characteristics. In general, the more labels in a model, the more accurate the initial assigned probabilities.

INDEPENDENT VARIABLES

For this model, independent variables were property characteristics that might ultimately predict the presence of lead service lines, depending on its predictive relationship with the dependent variables. The independent variables used in this model included residence type (i.e., number of stories), median income of the census block group, the parcel's estimated year of construction, estimated property-to-land-value ratio, estimated property value, and the stamped year on the nearest fire hydrant. Table 2 lists the dependent and independent data used in the multiple linear regression model.

Table 2 Variables used in predictive model

Variable	**Data source**	**Independent or dependent variable (Label)**	**Data scale (parcel, block group, etc.)**
Community-administered survey	Village of Hazel Crest	Dependent	Parcel
Village of Hazel Crest selected and IEPA approved Lead and Copper Rule sample sites	IEPA	Dependent	Parcel
Estimated year of construction	Cook County Assessor's Database	Dependent and Independent	Parcel
Estimated property value	Cook County Assessor's Database	Independent	Parcel
Estimated property to land value ratio	Cook County Assessor's Database	Independent	Parcel
Residence type: number of stories	Cook County Assessor's Database	Independent	Parcel
Fire hydrant stamp year (year of manufacture)	Village of Hazel Crest	Independent	X, Y coordinates

RESULTS

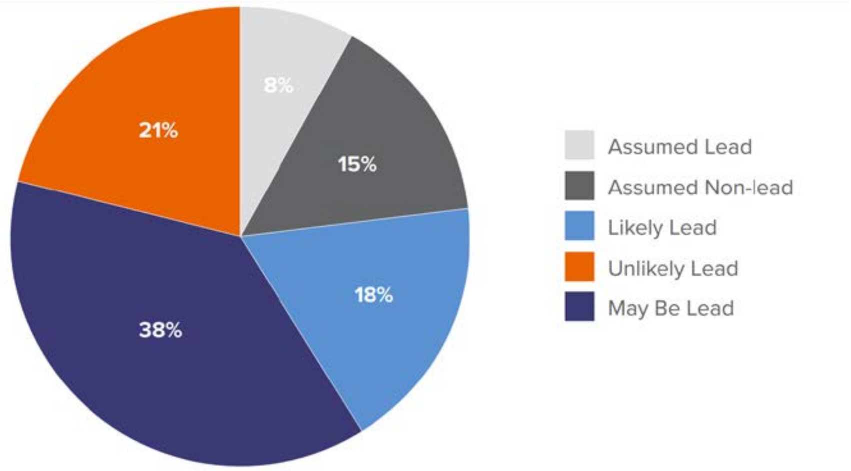
Based on community survey data, age of construction, and data from IEPA, 15 percent of Hazel Crest parcels are assumed to have non-lead service lines, and 8 percent of parcels are assumed to have lead service lines. The probability, or likelihood, of the presence of a lead service line in the remaining 77 percent of parcels was determined using a multiple linear regression. This regression analysis found that 56 percent of remaining parcels "maybe" or "likely" have a lead service line (see Fig. 3).

The majority of parcels in the “assumed lead” and “likely lead” categories are in Hazel Crest Proper and the Pott Hills/English Valley and Twin Creeks subdivisions. The Pacesetter/Stonebridge/Carriage Hills and Village West/Dynasty Lakes subdivisions are newer, and, therefore, service lines in those locations are classified as “assumed non-lead” or “unlikely lead.”

Based on the model outputs, of the 4,221 residential parcels in the Village of Hazel Crest:

- 633 parcels (approx. 15 percent) are assumed to have non-lead service lines, either because they were built after 1991 (i.e., after the Lead and Copper Rule banned the use of lead in plumbing) or because the village-administered community survey visually confirmed the presence of a copper service line
- 347 parcels (approx. 8 percent) are assumed to have lead service lines, either because the addresses are on the IEPA’s Approved Sample Sites list or because the village administered community survey visually confirmed the presence of a lead service line
- 758 parcels (approx. 18 percent) are likely lead, meaning they are predicted to have a high likelihood (66 percent to 99.99 percent probability) of having a lead service line
- 1,595 parcels (approx. 38 percent) may be lead, meaning they are predicted to have a medium likelihood (33 percent to 65.99 percent probability) of having a lead service line; and
- 888 parcels (approx. 21 percent) parcels are unlikely lead, meaning they are predicted to have a low likelihood (.001 percent to 32.99 percent probability) of having a lead service line.

Fig 3 Service line materials based on regression analysis



UPDATING THE INVENTORY

These initial results can help guide the village in preliminary decision-making on a lead service line replacement effort. However, owing to the small number of “labels” — or visually confirmed service lines made of lead, copper, or another material — there are limitations to the predictiveness of the model. As the village begins field testing and replacing service lines in high priority areas, those data can be input back into the map database to increase the number of known service line materials. This is important for two reasons:

1. The village can record and visualize its replacement progress in real time, supporting its goals of compliance and transparency; and
2. As the village collects more “labels,” these data points can be used to train and update the model used to create the initial probabilities, producing more accurate probabilities for the service lines yet to be tested/replaced.

BlueConduit, one of the project partners, originated the use of predictive analytics with machine learning to predict lead service line locations through their work in Flint, Michigan. They followed a similar methodology as the one outlined above but also applied machine learning techniques to make their preliminary predictive model smarter. The basic principle behind machine learning is to feed the initial model a random representative sample of "ground truth," or field data, from the community, which would then correct or strengthen the preliminary predictions of where lead service lines are likely or unlikely to occur throughout the community.

Even without a random, representative sample, the Village of Hazel Crest can strengthen the predictiveness of the model by inputting data from field testing and lead service line replacement efforts.

In 2021, CNT collaborated with the South Suburban Mayors and Managers Association (SSMMA) to develop a map and database maintenance plan to ensure that when Hazel Crest begins collecting field data, either by on-site testing or through community surveys, there will be a straightforward process by which to update the map and train the model. SSMMA currently hosts the interactive map and is primed to execute updates as the Village begins field testing.

INVENTORY LIMITATIONS

As discussed, one of the principal limitations in developing this inventory was a small number of "labels," or known service line materials. Fewer labels decrease the accuracy of predicted service line materials because there aren't as many known data correlations to inform the probabilities. This limitation can be overcome as the Village starts field testing and increasing the number of labels in the database.

Another notable limitation is the lack of data on the service line materials on either side of the curb stop, i.e., from the curb stop to the water main or to the water meter inside a residence. Regarding the inventory labels, the community survey results provide data for the residential side of the service line, the IEPA-approved list of addresses for sampling likely provide data for the entire service line, and homes built after 1991 should be lead free. While these distinctions have not been included in the map or database due to the level of uncertainty, a next step for the inventory database manager would be to add this distinction to the database and assign a designation based on available information. For instance, some of the community survey results for copper came from homes built well before 1991. This might indicate a home rehabilitation project, where the residential half of a lead service line was replaced with another material, but the portion from the curb stop to the water main is still lead. Again, this level of granularity can further support decision-making and efficient use of funds.

APPENDIX A: SAMPLE SET OF PARCEL LEVEL VARIABLES

Parcel Level Dataset Columns				
Column Name	**Description**	**Is it available?**	**Does CNT have it?**	**Source**
Lead (ppb)	Lead level in the submitted sample (ppb)	For some addresses	For some addresses	IL EPA and Village of Hazel Crest Community Survey Results, and Village of Flossmoor observed materials during construction
PID	Unique parcel ID	Y	Y	Assessor
Property Zip Code	Property zip code	Y	Y	Assessor
Owner Type	Owner type: including residential, commercial, and industrial	Y	Y	Assessor
Homestead	Homestead is a person's or family's residence, which comprises the land, house, and outbuildings, and in most states is exempt from forced sale for collection of debt	Maybe	N	
Homestead Percent	0-100	Maybe	N	
Home SEV	SEV is State Equalized Value. That's what the government thinks your home is worth.	Maybe	Y	Assessor
Land Value	Land value	Y	Y	Assessor
Land Improvements Value	Value of improvements on the parcel	Maybe	N	
Residential Building Value	Residential building value (only for residential buildings)	Y	Y	Assessor
Commercial Building Value	Commercial building value (only for commercial buildings)	N/A	N/A	
Building Storeys	Number of storeys	Y	Y	Assessor
Parcel Acres	Parcel acres	Y	Y	Assessor
Use Type	Residential, commercial, or industrial use	Y	Y	Assessor
Prop Class	Whether a parcel is agricultural, industrial, residential, or commercial property	Y	Y	Assessor
Old Prop class	Previous Prop Class	Maybe	N	
Year Built	Year which the building was built	Y	Y	Assessor
Original building materials	As build materials for original structure	Maybe	For some	Assessor/Large format permits
USPS Vacancy	Vacancy status of property according to USPS records	Y	N	USPS?

Assessor Vacancy	Vacancy status of property according to assessor data	Y	Y	Assessor
Zoning	City of Flint zoning assignment	Y	Maybe	Village, SSMMA
Future Land use	Planned use for land in the future	Maybe	N	Village
DRAFT Zone	Future assigned zoning	N/A	N/A	
Housing Condition 2012	Building condition according to the city record in 2012 (only for residential properties)	Maybe	Maybe	Assessor
Housing Condition 2014	Building condition according to the city record in 2014 (only for residential properties)	Maybe	Maybe	Assessor
Housing Condition 2013	Building condition according to the city record in 2013 (only for residential properties)	Maybe	Maybe	Assessor
Commercial Condition 2013	Building condition according to the city record in 2013 (only for commercial properties)	N/A	N/A	
Rental	Rental Residential Building or not	Y	Y	Assessor
Residential Building Style	Style of Residential Building			
Latitude, Longitude	Latitude and Longitude	Y	Y	Assessor
Hydrant Type	Type of closet hydrant to the property	Yes for Flossmoor; maybe for Hazel Crest	Yes for Flossmoor	Municipality
Ward	A ward is an optional division of a city or town for administrative and representative purposes, especially for purposes of an election	Y	N	Municipality
PRECINCT	Voting location the parcel belongs to	Y	N	Municipality
CENTRACT	A census tract/area is a geographic region defined for the purpose of taking a census. Numbers in the column are the population sizes (number of people).	Y	Y	Assessor
CENBLOCK	A census block is the smallest geographic unit used by the United States Census Bureau	Y	N	Census
SL_Type	Service line connection type	Maybe	N	Municipality
SL_Type2	Second service line connection type, if more than one connection for one parcel	Maybe	N	Municipality
SL_Lead	Lead/No lead connection	For some	For some	Municipality (historical/current records; observed)