Published in final edited form as:

Sci Total Environ. 2017 December 01; 599-600: 1552–1563. doi:10.1016/j.scitotenv.2017.05.094.

How Geostatistics can Help You Find Lead and Galvanized Water Service Lines: The Case of Flint, MI

Pierre Goovaerts

BioMedware Inc, Ann Arbor, MI 48106

Abstract

In the aftermath of Flint drinking water crisis, most US cities have been scrambling to locate all lead service lines (LSLs) in their water supply systems. This information, which is most often inaccurate or lacking, is critical to assess compliance with the Lead and Copper Rule and to plan the replacement of lead and galvanized service lines (GSLs) as currently under way in Flint. This paper presents the first geospatial approach to predict the likelihood that a home has a LSL or GSL based on neighboring field data (i.e., house inspection) and secondary information (i.e., construction year and city records). The methodology is applied to the City of Flint where 3254 homes have been inspected by the Michigan Department of Environmental Quality to identify service line material. GSLs and LSLs were mostly observed in houses built prior to 1934 and during World War II, respectively. City records led to the over-identification of LSLs, likely because old records were not updated as these lines were being replaced. Indicator semivariograms indicated that both types of service line are spatially clustered with a range of 1.4 km for LSLs and 2.8 km for GSLs. This spatial autocorrelation was integrated with secondary data using residual indicator kriging to predict the probability of finding each type of material at the tax parcel level. Cross-validation analysis using Receiver Operating Characteristic (ROC) Curves demonstrated the greater accuracy of the kriging model relative to the current approach targeting houses built in the forties; in particular as more field data become available. Anticipated rates of false positives and percentages of detection were computed for different sampling strategies. This approach is flexible enough to accommodate additional sources of information such as local code and regulatory changes, historical permit records, maintenance and operation records, or customer self-reporting.

Graphical abstract

Corresponding author phone: 517-900-9283; fax: 734-913-2201; goovaerts@biomedware.com. Corresponding author address: BioMedware, Inc. PO Box 1577, Ann Arbor, MI 48106 (USA).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

kriging; indicator semivariograms; Lead-and-Copper Rule; ROC curve; Lead service lines

Introduction

The drinking water contamination crisis in Flint, Michigan was a painful reminder of the threat posed by an aging infrastructure, in particular the occurrence of lead service lines (LSLs), to drinking water supply (Rosen *et al.*, 2017). Lead pipes started being installed on a large scale in the United States in the late 1800s, and by 1900 they were used in more than 70% of cities with populations greater than 30,000 (Troesken, 2006). The installation of LSLs slowed down after 1930 but was only prohibited nationwide with the passage of the Safe Drinking Water Act Amendments of 1986 (Rabin, 2008). In some cities, most of LSLs were actually put in from 1940 to 1945 (Wisely and Spangler, 2016).

According to a recent survey of US community water systems (Cornwell *et al.*, 2016), 6.1 millions of LSLs (either full or partial) are currently present nationwide, serving between 15 and 22 million people (7% of customers). A full LSL is defined as a lead line extending from the water main to the house. Partial LSL refers to the situation where a lead line is present on either the customer side (curb stop to the home) or the utility side (curb stop to the water main) of the service line but not both. Partial LSLs are important too as a culprit for high water lead levels in Flint was the destabilization of lead-bearing corrosion rust layers that accumulated over decades on a galvanized iron pipe downstream of a lead pipe (Pieper *et al.*, 2017). Galvanized steel pipes themselves can be a long-term source of lead as the “galvanized” surface zinc coating can contain up to 2% lead (Clark *et al.*, 2015). This material was actually the most widely used for most of the 20th century and an estimated 7.5% of households overall have steel or galvanized steel service lines (American Water Works Association, 1996).

In the aftermath of the water contamination crisis the city of Flint decided to replace all lead and galvanized service lines. Pipes at more than 700 homes have been replaced so far and recently a federal judge approved a deal to replace water lines at 18,000 homes by 2020 to

settle a lawsuit over lead-contaminated water (White, 2017). For 2017 alone, 6000 residential homes have been identified for service line replacement in 10 different zones, 600 per zone. (May, 2017). According to MDEQ plan, the selection of these zones was based on several factors, including the concentration of lead and galvanized service lines, active water accounts, density of children and elderly, and the water lead levels determined by testing.

This ambitious project has however been off to a slow start as the goal of replacing 1000 service lines in 2016 has still not been reached as of April 2017. The effort has been plagued by problems that include inaccurate records on the location of pipes and the type of material used in them (Ehrmann, 2017). For example, a number of homes that have been targeted for service line (SL) replacement turned out to have copper pipes that didn't need to be replaced. In the case a Flint, records consisted of 45,000 index cards and 240 parcel maps showing city blocks divided into little squares of property, with a code that indicated which type of SL was running into each house (Gold, 2016). Unfortunately this is not an isolated situation as many other communities still keep their water service connection records on manila-colored index cards, handwritten in pencil by installation crews decades ago (Wisely and Spangler, 2016).

In general, many cities lack a good estimate of how many lead pipes remain in their communities and where they are buried (Smith, 2016). This information is not only needed to plan the replacement of lead lines but it is critical to assess compliance with the Lead and Copper Rule (LCR) in the first place. Under this rule no more than 10% of first-draw 1-L water samples collected from high-risk homes can exceed the action level of 15 µg/L for lead and 1.3 mg/L for copper (LCR, US EPA, 1991, 2002, 2016a). The LCR includes a tiering system for prioritizing the selection of sample sites based on the likelihood of the sites to release elevated levels of lead; e.g., sites with LSL, lead pipes, or copper pipes with lead solder. Whenever possible all water samples should be collected from Tier 1 sampling sites which are Single Family Residences served by a LSL or withlead pipes or copper pipes withlead solder (US EPA, 2002). The delay in reporting highlevels of lead in Flint drinking water and the resulting extent of the ensuing public health crisis were partially caused by biased sampling; e.g., out of 324 sampling sites used to monitor lead in Flint over the years, only six could positively be determined in November 2015 as having a lead service line (Wisely and Spangler, 2016)

Lack of accurate records is forcing public water system and State agencies to rely on expensive survey by licensed plumbers to identify the location of lead and galvanized service lines. For example, the Michigan Department of Environmental Quality (MDEQ) conducted in 2016 an investigation of lead lines over 3000 homes across the City of Flint. Besides highlighting the limitations of city records (i.e., over-prediction bias for lead lines) this campaign early on revealed the larger frequency of occurrence of LSLs in houses built from the late 30s to the late 40s (Bryce Feighner, MDEQ, personal communication, March 2, 2017). Based on this information the sampling strategy was subsequently modified to target this particular housing segment. So far the geographical location of newly acquired on-site data has not been used during their combination with city records on SL material and construction years. One should however expect neighboring houses to have similar types of

service line, as they were likely built around the same time period and might have undergone similar upgrades

Geostatistical tools have been applied to the modeling of spatial distributions in many disciplines, such as environmental sciences (e.g., deposition of atmospheric pollutants, soil heavy metal concentration), ecology (characterization of population dynamics), and health (patterns of diseases and exposure to pollutants). These tools are increasingly coupled with GIS (geographical information system) capabilities for applications that characterize space-time structures (semivariogram analysis), spatially interpolate scattered measurements to create spatially exhaustive layers of information and assess the corresponding accuracy and precision (Goovaerts, 2011). Yet their application to drinking water distribution systems is far less common. Several authors (De Oliveira *et al.*, 2011; Zhao and Daly, 2015) used geospatial techniques and pipeline geographic information to detect clusters of pipe breaks (e.g., pipe body cracks or splits, joint failures, and hydrant valve failures) and prioritize pipes that should be selected for rehabilitation. Studzinski (2011) applied geostatistical prediction (kriging) to estimate water flow and water pressure at the nodes of a communal water distribution network. To the author's knowledge, these tools have never been used to model the spatial distribution of service line material.

This paper presents the first geospatial approach to predict the likelihood that a tax parcel has LSL or GSL based on neighboring field data (i.e., house inspection) and secondary information (i.e., construction year and city records). Cross-validation combined with ROC curves is introduced as a way to assess the accuracy of the predictions and provide decision-makers with anticipated rates of false positives and percentages of detection. The methodology is illustrated using data recently collected in Flint, MI.

2. Data Sources and Methods

2.1 Datasets

From February to early July 2016 a "Lead Line Investigation" (LLI) was conducted by MDEQ at over 3000 homes across the City of Flint. Plumbers classified the material of the service line coming into the home (i.e., customer-side service line) into six categories: lead, galvanized, copper, plastic, other, and unknown. Galvanized refers to iron pipe with a protective "galvanized" surface coating composed of zinc, lead, and cadmium. The "other" category includes material other than the four that are normally observed in a home (e.g., PEX or cast iron). The term "unknown" was used whenever the SL material could not be confirmed because, for example, the line was behind a wall or way back in a crawl space. Two data sources were combined in this paper: 1) Excel spreadsheet provided by MDEQ including the results of the inspection of 3132 houses identified by their address and tax parcel ID, and 2) data collected by MDEQ at 823 sites within the framework of the sentinel sampling program (Flint Safe Drinking Water Task Force, 2016) and posted online at <http://www.michigan.gov/flintwater>. After removal of data labeled as "unknown" (53 sites) and duplicates between the two datasets, the final database consists of 3254 data.

Sites were selected for house inspection according to the following criteria: i) spatial distribution to ensure coverage of all nine wards, ii) areas predicted to have high blood levels

based on Hanna-Attisha *et al.* (2016), and iii) environmental justice considerations, specifically lead paint indicators, minority population, and low income derived from EPA Environmental Justice Mapping and Screening Tool known as EJSCREEN (US EPA, 2016b). As the MDEQ campaign progressed and houses built from the late 30s to the late 40s turned out to be the main housing segment with LSLs, the sampling started targeting specifically houses built during that period (Bryce Feighner, personal communication, March 2, 2017), which explains the preferential sampling noticed in earlier studies (Goovaerts 2017a,b).

The following housing characteristics described in detail in Goovaerts (2017a) are available for all 56,039 tax parcels located within the City of Flint: i) the type of service line (lead, other, or unknown) retrieved from a digital map of Flint's lead water pipe (city records), and ii) the year the house was built. These two characteristics will serve as covariates to predict the occurrence of lead and galvanized service lines.

2.2 Hard and soft indicator coding of SL data

The analysis started with the coding of all n house inspection results (field data), georeferenced using their tax parcel's centroid geographical coordinates $\mathbf{u}_\alpha$, into two indicators of presence/absence of lead and galvanized SL:

$$i(\mathbf{u}_\alpha; LSL) = \begin{cases} 1 & \text{if SL is lead at } \mathbf{u}_\alpha \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$i(\mathbf{u}_\alpha; GSL) = \begin{cases} 1 & \text{if SL is galvanized at } \mathbf{u}_\alpha \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

These indicators are referred to as “hard data” as they correspond to precise measurements of SL composition. On the other hand, calibration of secondary data (built year, city records on SL material) provides imprecise or “soft” information about the presence of a particular type of SL material (Goovaerts, 1997). In this paper, soft data were computed as the frequencies of occurrence of LSL or GSL for 3×3 categories of built year (BY) and digital SL data (DD); see Section 3.1 for the definition of these categories. For example, soft indicators for LSL were calculated as:

$$j(\mathbf{u}_\alpha; LSL) = f(i(\mathbf{u}_\alpha; LSL) = 1 | BY(\mathbf{u}_\alpha) = y_k \text{ and } DD(\mathbf{u}_\alpha) = d_{k'}) \\ = \frac{1}{\sum_{\alpha=1}^n i_{BY}(\mathbf{u}_\alpha; y_k) \times i_{DD}(\mathbf{u}_\alpha; d_{k'})} \sum_{\alpha=1}^n i(\mathbf{u}_\alpha; LSL) \times i_{BY}(\mathbf{u}_\alpha; y_k) \times i_{DD}(\mathbf{u}_\alpha; d_{k'}) \quad (3)$$

where $i_{BY}(\mathbf{u}_\alpha; y_k) = 1$ if $BY(\mathbf{u}_\alpha) = y_k$ and zero otherwise, while $i_{DD}(\mathbf{u}_\alpha; d_{k'}) = 1$ if $DD(\mathbf{u}_\alpha) = d_{k'}$ and zero otherwise. In other words, at a house $\mathbf{u}_\alpha$ which was built during the period y_k (e.g., 1939–1949) and with a service line of composition $d_{k'}$ (e.g., lead) according to city records, the soft indicator for LSL $j(\mathbf{u}_\alpha; LSL)$ will be the average frequencies of lead service

lines found during inspection at all houses built during that time period and with that type of digital information. Unlike hard data which are binary (0,1) and only available at $n=3254$ sites, soft data range from zero to one and can be computed for all $N=56,039$ tax parcels.

2.3 Indicator semivariograms

The spatial pattern of SL data was first characterized using the indicator semivariogram which is traditionally computed as half the squared difference between indicator data separated by a distance h ; e.g., for LSL:

$$\hat{\gamma}_I(h; LSL) = \frac{1}{2N(h)} \sum_{a=1}^{N(h)} [i(\mathbf{u}_a; LSL) - i(\mathbf{u}_a + h; LSL)]^2 \quad (4)$$

where $N(h)$ is the number of sampled houses within a given class of distance, known as spatial lag. Since the spatial increment $[i(\mathbf{u}_a; LSL) - i(\mathbf{u}_a + h; LSL)]$ is non-zero only if one service line is lead (indicator=1) and the other one is of a different composition (indicator=0), the quantity $2\hat{\gamma}_I(h; LSL)$ measures the transition frequency between lead SLs and non-lead SLs as a function of h . In other words, the smaller $2\hat{\gamma}_I(h; LSL)$, the greater the spatial connectivity of LSLs. If LSLs are distributed randomly across the City, then we should expect the semivariogram to be flat (pure nugget effect) and close to the sill value $m_I(LSL) \times [1 - m_I(LSL)]$, where $m_I(LSL)$ is the mean of indicator data (Eq. 1) and represents the proportion of sampled houses with LSL.

To correct for the preferential sampling and clustering of LSLs and GSLs, the following rescaled semivariogram estimator (Pannatier, 1996; Goovaerts *et al.*, 2005) was also used:

$$\hat{\gamma}_I(h; LSL) = \frac{\sigma^2}{2N(h)} \sum_{a=1}^{N(h)} \frac{[i(\mathbf{u}_a; LSL) - i(\mathbf{u}_a + h; LSL)]^2}{\sigma^2(h)} \quad (5)$$

where σ^2 is the variance of the $2N(h)$ data used for estimation at lag h , and σ^2 is the variance of the entire dataset (n indicators). The rescaling allows one to account for large changes in variance from one lag to the next. This estimator can greatly attenuate erratic fluctuations in semivariograms and result in a more accurate quantification of the short-range variability (i.e., nugget effect); see Goovaerts *et al.* (2005) for an application to preferential sampling of water wells with high levels of arsenic.

A continuous function must be fitted to the experimental curve (Eqs. 4&5) so as to deduce semivariogram values for any possible lag h required by prediction algorithms (Section 2.4) and also to smooth out sample fluctuations. Semivariogram modeling was here conducted using weighted least-square regression as implemented in the freeware program AUTO-IK (Goovaerts, 2009).

2.4 Indicator kriging

If the semivariogram indicates the existence of spatial correlation between indicators, then field data can be used to predict the SL composition for houses that were not inspected by MDEQ. This prediction takes the form of the probability for a given SL material (e.g., lead) to be present at location $\mathbf{u}$ based in the $n(\mathbf{u})$ closest field data; $n(\mathbf{u})$ was here set to 16 based on sensitivity analysis using the AUC statistic (Section 2.6) as criterion. Mathematically, this probability is computed as the following linear combination of neighbouring indicator data; e.g., for LSL:

$$p_{IK}^*(\mathbf{u}; LSL) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha} \times i(\mathbf{u}_{\alpha}; LSL) \quad (6)$$

where the weights λ_{α} are solutions of a system of linear equations, known as ordinary indicator kriging (IK); see Journel (1983). Kriging weights are influenced by the proximity of the data to the location $\mathbf{u}$ (i.e., closest data receive more weight), the data configuration (e.g., spatially clustered data receive less weight since they provide redundant information), as well as the spatial pattern of indicator data as modeled using the semivariogram (Eqs. 4&5).

2.5 Indicator kriging using hard and soft data

The predictive model created by indicator kriging (Eq. 6) does not make full use of all the data available in that secondary information, such as built year and city records on SL material, is ignored. Residual kriging (RK) incorporates this missing information, which is captured by soft indicators (Eq. 3), using the following estimator:

$$p_{RK}^*(\mathbf{u}; LSL) = j(\mathbf{u}; LSL) + \sum_{\alpha=1}^{n(\mathbf{u})} \lambda'_{\alpha} \times r(\mathbf{u}_{\alpha}; LSL) \quad (7)$$

where $j(\mathbf{u}; LSL)$ is the probability estimated from calibration of secondary data, and $r(\mathbf{u}_{\alpha}; LSL) = [i(\mathbf{u}_{\alpha}; LSL) - j(\mathbf{u}_{\alpha}; LSL)]$ are indicator residuals representing the variability in SL composition that is not explained by secondary information. These residuals are interpolated to parcel $\mathbf{u}$ using a set of weights λ'_{α} which are the solution of a system of linear equations, known as a simple kriging with varying local means (Goovaerts and Journel, 1995; Grunwald *et al.*, 2006). As the parcel being predicted gets further away from any field data, the weights λ'_{α} assigned to these data decrease, resulting in the predicted value $p_{RK}^*(\mathbf{u}; LSL)$ getting closer to the soft datum $j(\mathbf{u}; LSL)$. In other words, if no house was inspected in the vicinity of location $\mathbf{u}$ the prediction of SL material relies primarily on the secondary data available for that parcel. The same happens if the spatial autocorrelation is weak (e.g., large nugget effect and/or short range of the residual semivariogram). As for indicator kriging (Section 2.4), the computation of weights λ'_{α} accounts for the spatial pattern of the residual data, which was modeled using the semivariogram (Eqs. 4&5) where hard indicators $i(\mathbf{u}_{\alpha}; LSL)$ are replaced by residuals $r(\mathbf{u}_{\alpha}; LSL)$. The number of data $n(\mathbf{u})$ was set to 8 based on

sensitivity analysis; compared to ordinary kriging fewer neighbors are needed as the soft datum collocated with the location being estimated, $\hat{J}(u; LSL)$, is used.

2.6 Validation analysis

The benefit of geostatistical prediction (Eqs. 6 and 7) over the straightforward approach adopted by the State to locate LSLs and GSLs (e.g., targeting houses built from the late 30s to the late 40s and without non-lead SL according to city records) was quantified using validation studies. The ability of the different interpolation techniques to detect problematic SLs was investigated over a range of sampling densities using the following procedure:

1. Select a random subset of n field data representing between 20% and 90% of the total number of houses inspected by MDEQ (prediction set). For each of the 8 sampling intensities (20%, 30%, ..., 80%, 90%), 100 different random subsets (k-fold cross-validation) were selected to account for sampling fluctuations. To ensure a uniform coverage of the City, a stratified random sampling was implemented whereby the same percentage of data was sampled in each of the nine wards.
2. For each random subset and sampling intensity, predict the probability of occurrence of LSL and GSL at the remaining N locations (validation set) using the following statistical approaches:
 - Calibration of secondary information (predictor = soft indicator data)
 - Ordinary indicator kriging which ignores secondary information (OK)
 - Residual kriging which combines both hard and soft data (RK)

For both types of kriging, the experimental semivariograms of hard indicators and residuals were computed on each subset and modelled using weighted least-square regression. The following two empirical approaches were also used as reference or baseline:

- Sample first houses where city records indicate the presence of LSL(GSL) or SL of unknown composition, then sample the other validation locations; within both groups, the selection or sampling was conducted randomly (method STRAT1)
 - Sample first houses built within a given time period (1938–1951 for LSL and prior to 1934 for GSL) and where city records indicate the presence of LSL(GSL) or SL of unknown composition, then sample the other validation locations; within both groups, the sampling was conducted randomly (method STRAT2)
3. Compare the predictions with the ground truth (i.e., hard indicators) and create a Receiver Operating Characteristics (ROC) curve which plots the probability of false positive versus the probability of detection (Swets, 1988; Fawcett, 2006; Goovaerts *et al.*, 2016). The accuracy of the prediction was quantified using the relative area under the ROC curve (AUC statistic), which ranges from 0 (worst case) to 1 (best case). The AUC 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).

3. Results and Discussion

3.1 Sampling design

Fig. 1A shows the location of all 3254 houses that were inspected by MDEQ. Houses with lead and galvanized SL are denoted by red dots and blue squares, respectively. Only 135 LSLs were found on-site (Table 1) and visually these appear to be strongly clustered in space. Many more GSLs were detected (593, Table 1) and they cover a much wider section of the city. These differences in spatial patterns are more apparent when aggregating the results at the block group level (Figs. 1C, D). Out of 128 block groups that were sampled, only 29 included houses with LSL and these represented between 1.5% and 36% of the total number of houses inspected in each block group. Conversely, galvanized lines were found in the majority (i.e., 94) of block groups, representing on average 24% of inspected lines.

It is noteworthy that the sampling density, hence the reliability of percentages displayed in Figs. 1C, D, varied greatly across the city (Fig. 1B). For example, the block group with 100% GSLs (Ward 1) had only one house inspected. The highest rate for LSLs (36% in one block group in Ward 6) is however computed from 75 field data representing 20% of the total number of tax parcels within that block group. As expected given the focus on locating LSLs, block groups that were the most densely sampled tend to be the ones with the largest proportions of LSL ($r=0.54$); compare Figs. 1B and C.

Results were also aggregated at the level of nine city wards whose limits are overlaid on all maps of Fig. 1. According to Fig. 1B and Table 2, wards 1, 3 and 5 were the least densely sampled (< 3% of tax parcels), although the largest percentages of lines identified as galvanized during the inspection were in wards 3 and 5. Ward 7 is the only ward with a sampling rate above 10% and it has the lowest poverty level (Table 2, last line). This sampling rate is even more impressive (24.1%) for two of the block groups marked by a star in Fig. 1B. Interestingly, only 5% and 11.3% of lines inspected in these two census units were classified as lead and galvanized, respectively. At the same time, these two block groups had a poverty level of 44.8%, much lower than the average for the City of Flint (66.3%) and Ward 7 (57.6%), which confirms the socio-economic bias mentioned in recent studies (Goovaerts, 2017a, b).

3.2 Calibration of secondary information

The contingency table of city records versus results of on-site inspections (Table 1) illustrates the benefits and limitations of using digital data to predict the presence of lead and galvanized service lines. For lead there were an overwhelmingly large number of false positives: LSLs were found at less than 12% of houses identified on the basis of city records (82/706). Most of these service lines were actually made of copper (567), while 51 lines were galvanized. This severe overestimation is most likely due to the failure to update city records as LSLs were being replaced with copper. This is confirmed by the fact that less than 6% of inspected LSLs (8/135) were incorrectly labelled as other-than-lead material by city

records: 61% (82/135) were correctly identified as lead, while the remaining 33% (54/135) were listed as unknown. Although city records do not provide any specific information on the presence of GSLs, most of them were classified as others (54%) or unknown (37%). Only 9% of GSLs were incorrectly labelled as LSLs in city records.

Construction year appears to be a more accurate predictor of the composition of service lines. Kernel smoothing introduced in Goovaerts (2017b) was used to explore how the relative frequency of LSLs, GSLs, and copper lines changes with built year. Every year between 1895 and 2010, frequencies of occurrence of the three types of SL material were computed from all houses built within an 11-year window (e.g., 1916–1926 for year 1921). Results based on less than 30 observations were discarded to avoid erratic fluctuations. Figure 2 shows the three frequency curves that were standardized by their global frequency (last line of Table 1) to facilitate visual comparison; value > 1 indicates that for this year of construction the percentage is larger than what was found on average over all inspected homes. Most GSLs (86%) were found in houses built before 1934, while the occurrence of LSLs spiked for houses built between 1938 and 1951: 89% of LSLs were found within that time period. The influence of construction year on copper lines is less drastic: the percentage of copper lines increased at the same time as the proportion of GSLs dropped, and then became the norm (97.40%) after 1951.

Table 3 illustrates the combined influence of city records and construction year on the percentages of LSLs and GSLs found during house inspection. Construction year was discretized into three categories based on the interpretation of Fig. 2. Regardless of the SL material reported in city records, lead and galvanized lines were absent from all houses built after 1951 and 1971, respectively. LSLs were rare for houses built before 1938 and wherever digital data indicate non-lead SLs (“Others” category). GSLs were found mostly in older homes (built prior to 1931), in particular where the SL material is marked as “unknown” in city records.

The spatial distribution of the two covariates is mapped at the parcel and block group levels in Fig. 3, while Table 2 provides summary statistics at the ward level. As fairly common in most US cities, increasingly older homes are found toward the center of the City. An exception is the cluster of post- 1985 constructions in Ward 5 (dark blue spot) corresponding to recent low-income housing developments and commercial buildings. Block group level maps (Figs. 3C–E) display spatial patterns for the two housing categories (pre-1934 houses, 1938–1951 built year) identified as the most likely to have GSLs and LSLs, respectively. Pre-1934 houses are confined to the central part of the City and are much less scattered than 1938–1951 houses that are found in small clusters in two thirds of the wards.

Unlike construction year, no clear spatial pattern emerges from the parcel-level map of SL material (Fig. 3B); this noise likely reflects the inaccuracy of city records. Aggregation to the block group level suggests the widespread presence of LSLs, except on the Eastern and Northern edges of the city (Fig. 3D). City records were however found to overestimate greatly the percentages of LSLs (Table 1). This high rate of false positives is also apparent when comparing Fig. 3D to Fig. 1C where MDEQ results are mapped using the same color scale. Another indicator of the inaccuracy of Fig. 3D is the lack of similarity with the 1938–

1951 built year map (Fig. 3E) despite the relationship revealed by MDEQ inspections (Fig. 2). The block group map of percentages of other-than-lead SL (Fig. 3F) shows a negative relationship with % LSLs, as expected given that these two categories represent 82% of the digital data.

3.3 Spatial pattern of SL composition

The visual description of maps of field inspection data (Figs. 1C, D) was followed by a quantitative analysis and modeling of scales of variability. Lead and galvanized SLs have sharply different semivariograms (Figs. 4A–B, green curve). The range of autocorrelation is much shorter for LSLs (300 m) compared to GSLs (2.8 km), which reflects the existence of small clusters of LSLs, while GSLs are distributed more broadly over the City. This pattern is coherent with the existence of small clusters of 1938–1951 houses identified on Fig. 3E. The GSL semivariogram also displays a large nugget effect (i.e., short-range variability), which should hamper any spatial prediction.

The impact of sampling configuration on results was investigated using the rescaled semivariogram estimator which accounts for large changes in variance from one lag to the next. This estimator attenuates erratic fluctuations in the two semivariograms (Figs. 4C–D, green curve), revealing a second longer range for LSLs (1.4 km) and eliminating the drop at the origin of the GSL semivariogram. Nevertheless, a large proportion of the total variability is still observed over a few hundred meters.

The residual semivariogram needed to combine hard and soft data using kriging (Eq. 7) is also displayed in Fig. 4 (red curve). In all cases, the residual semivariogram reaches a lower sill than the indicator semivariogram; the difference reflects the proportion of variance explained by the two covariates: year built and city records on SL composition. While the shape of both semivariograms remains fairly similar for LSLs, the contribution of the longer range (2.8 km) structure decreases sharply for the GSL residual semivariogram. This change indicates that this long-range structure reflected the spatial pattern of pre-1934 houses (Fig. 3C) which include most of GSLs.

3.4 Validation analysis: leave-one-out approach

The existence of spatial autocorrelation supported the application of geostatistics to predict the likelihood of finding LSLs and GSLs at unvisited sites. The benefit of these advanced techniques, in particular when combining hard and soft data, was first investigated using cross-validation and ROC curves to assess the accuracy of prediction. Figure 5 shows the ROC curves computed for both types of service lines and five interpolation techniques described in Section 2.6. In this instance, each field data was removed one at a time and re-estimated using the remaining 3253 data (leave-one-out crossvalidation).

Each ROC curve plots the probability of false positive (false alarm) versus the probability of detection. The most efficient algorithm is the one that allows the detection of a larger fraction of LSLs or GSLs at the expense of fewer false positives; that is the ROC curve should be as close as possible to the vertical axis. A quantitative measure of the classification accuracy is the relative area under the ROC curve (AUC), which represents the average frequency of detection (best if close to 1). For the detection of all LSLs or GSLs the

worst results (i.e., the smallest AUC) are obtained when the selection is based on city records (STRAT1), while residual kriging leads to the most accurate predictions, in particular for LSLs (Figs. 5A, B). Approaches that account for both construction year and city records (i.e., RK, Soft, and STRAT2) outperform kriging which uses only field data. Their ROC curves also exhibit an inflection point around a probability of detection of 0.8 when the probability of false alarms starts increasing at a fast rate. This increase reflects the challenge in locating LSLs and GSLs for categories of construction year and digital SL composition with low frequencies of occurrence of LSLs and GSLs (Table 3). Similarly, for the method STRAT1 the inflection point corresponds to LSLs located in the non-lead category according to city records (8 out of 135) or GSLs located in the lead category according to city records (51 out of 593). Because both STRAT1 and STRAT2 approaches select houses randomly within each category, the ROC curve is a linear segment within each category (results are average of 100 simulations to account for sampling fluctuations in the random selection of houses).

A second series of ROC curves was created specifically to assess the accuracy in identifying the 19 LSLs and 118 GSLs located in houses with low probabilities of including this type of SL material based on their construction year and city records. In this case, using soft data can actually result in a classification less accurate than a simple random selection (i.e., $AUC < 0.5$). This subset of “challenging” target SLs emphasizes the benefit of geostatistics over non-spatial selection. Using neighboring field data residual kriging was able to correct partially the underestimation of probability of occurrence of LSLs and GSLs by soft data. It is noteworthy that ordinary kriging becomes the most efficient algorithm to detect galvanized lines, although the AUC is still rather small (0.574).

3.5 Validation analysis: impact of sampling density

Leave-one-out cross-validation can lead to overly optimistic accuracy estimates, in particular when data are spatially clustered (Roberts *et al.*, 2017). The way the different interpolation techniques perform as the sampling density decreases was investigated using the k-fold cross-validation approach described in Section 2.6. Figure 6 shows for each prediction algorithm the value of the AUC statistic, averaged over 100 random subsets of increasing size (i.e., more field or hard data used for prediction). Both types of SL material exhibit contrasted behaviors caused by differences in spatial autocorrelation and strength of relationship with secondary information.

The small nugget effect of the LSL semivariogram model magnifies the benefit of adding hard data, in particular for ordinary kriging which ignores the secondary information available at the location of the prediction (Fig. 6A). As the sampling density increases, the accuracy of ordinary kriging gets closer to the straightforward approach of sampling first homes built between 1938 and 1951 and where city records indicate the presence of LSLs or SLs of unknown composition (STRAT2). Residual kriging gives the best prediction with soft data being a close second, in particular as sampling density decreases. The benefit of a geostatistical approach and increased sampling density is more pronounced for the prediction of 19 LSLs that are not located in the expected classes of construction year and digital SL composition (Fig. 6B). In that case, using the sole soft data is misleading and

ordinary kriging gives more accurate predictions for sampling density above 50%. Residual kriging is still the best predictor, illustrating the ability of the algorithm to correct for inaccurate soft data.

The impact of sampling density on prediction accuracy is negligible for GSLs (Fig. 6C), mainly because of the much larger nugget effect displayed by the GSL residual semivariogram (Fig. 4D). In that case, residual kriging assigns a very small weight to hard data, leading to estimates that are very close to soft data. Such reliance on secondary information is however hazardous for the detection of 118 GSLs that are not located in the expected classes of construction year and digital SL composition (Fig. 6D). Then, ordinary kriging becomes the best predictor, while residual kriging corrects to some extent the unduly influence of inaccurate soft data. Like for LSLs, the State approach of prioritizing houses based on construction year and city records yields the worst predictions.

3.6 Spatial prediction

Based on the validation analysis, residual kriging was selected to predict the probability of presence of LSLs and GSLs for each of the 56,039 tax parcels located within the City of Flint (Figs. 7A, B). Because kriging is an exact interpolator (i.e., field data are matched at housing inspection sites), these maps display a spatial pattern similar to the location map in Fig. 1A: LSLs are found in small pockets, while GSLs cover a much wider fraction of the City. This pattern is even more apparent after aggregating kriging estimates at the block group level (Figs. 7C, D). To visualize the impact of hard data (field inspections) versus soft data (calibration of built year and city records) on final results, soft indicators were also averaged at the block group level (Figs. 7E, F) and the location of inspected LSLs and GSLs is overlaid on all four maps.

Because the LSL semivariogram has a smaller nugget effect relative to GSLs, field data have generally a greater impact on the RK estimates: the probability inferred from soft information increases in the vicinity of LSLs found during house inspections; compare Figs. 7C to 7E. Higher probabilities in the Northern part of Ward 3 result mainly from the calibration of secondary information as only one LSL was found during field survey. Changes are more subtle for the GSL probability map, which agrees with the overlapping ROC curves and AUC statistics displayed in Figs. 5 and 6.

The expected number of tax parcels with LSL or GSL, which were not inspected by MDEQ, was computed as the sum of RK probability estimates at all unsampled locations. The results are 1075 LSLs and 12,255 GSLs. For comparison, city records indicated the existence of 4214 parcels with LSLs, which minus 706 parcels inspected by MDEQ (Table 1, gives 3508 remaining parcels with potential LSL. As expected, this number is much larger than the geostatistical estimate (1075) given the bias observed for city records. Note also that some tax parcels include multiple housing units and service lines, so the expected number of LSLs and GSLs is likely larger than the reported figures.

Probability maps (Figs. 7A, B) are best used to rank parcels according to their likelihood of having lead or galvanized service lines, allowing one to prioritize inspection and line replacement. Table 4 reports the numbers of parcels falling into different categories of

probability. For example, there was a zero probability of finding LSL at 27,397 parcels, while this number was 1638 for GSL. There is however some uncertainty attached to these predictions, which was assessed using the ROC curves created during the leave-one-out and k-fold validation analyses described in Sections 3.4 and 3.5. In particular, the 10 probability rules listed in Table 4 were applied to each of the 100 prediction sets generated for each of the nine sampling densities. The rates of true positives (i.e., % of LSLs or GSLs detected) and false positives (i.e., % parcels wrongly flagged as having LSL or GSL) were then averaged over all 100 prediction sets. Since these rates vary with the number of data available, the minimum and maximum values observed over the set of nine sampling densities are reported in Table 4.

Almost all LSLs and GSLs would be detected using the very conservative rule of inspecting all parcels with a non-zero probability of occurrence, but this naïve approach would be prohibitively expensive in terms of numbers of parcels inspected and rates of false positives. In contrast, inspecting all 3544 parcels with a probability of having a LSL above 0.15 should lead to the identification of between 81.9% and 87.4% of lead service lines, with a 10.2%–11.4% rate of false positives. Using lower probability thresholds (e.g., 0.05) would increase marginally the detection rate of LSLs (best scenario: 81.9% to 90.5%) at the expense of a 44% increase in the number of parcels to be inspected (5093 vs 3544), while the number of false positive could double (10.2% to 19.5%). A similar reasoning can be conducted for galvanized lines. In that case, the best strategy seems to be inspecting all 20,216 parcels with a probability of having a GSL above 0.25, which should lead to the identification of around 80% of galvanized service lines with a 16.2%–16.5% rate of false positives.

4. Conclusions

As most US cities are scrambling to locate lead and galvanized service lines in their water supply networks, there is a great need for techniques that can prioritize candidate sites for inspection and assess the accuracy of this classification. This paper presented the first geospatial approach to predict the likelihood that a home has lead or galvanized SL based on neighboring field data (i.e., house inspection) and secondary information (i.e., construction year and city records). This approach is flexible enough to accommodate additional types of information; in particular all data sources listed in Ohio EPA guidelines such as (Brush, 2017). Once this data is coded as a soft indicator of presence/absence of LSL or GSL, it can be easily incorporated in the indicator kriging approach and used to update the existing probabilistic model.

One key feature of the geostatistical approach is its ability to account for the geographical location of the data and their spatial pattern to make prediction at unsampled sites. Indeed, neighboring houses should be expected to have similar types of service line as they were likely built around the same time period and might have undergone similar upgrades to the water supply pipe network (e.g., replacement of lead service lines). In particular, kriging will assign less weight to clustered data which are often collected during preferential sampling surveys and carry redundant information.

Probabilities estimated by indicator kriging can range anywhere between 0 and 1, which allows the ranking and prioritization of all sites considered for inspection and line replacement. It also provides the justification to regulators that efforts are aimed at the highest ‘risk’ targets. Such mapping is thus much more informative than recent efforts (Hull & Associates, 2017) where tax parcels were simply assigned, based on their dates of construction, to the following three categories of likelihood that service lines include lead: i) high likelihood (built before 1940), ii) moderate likelihood (1940 to 1969), and iii) low likelihood (from 1970 to present). Cross-validation combined with ROC curves was also introduced as a way to assess the accuracy of the predictions and provide decision-makers with anticipated rates of false positives and percentages of detection. Cross-validation demonstrated the greater accuracy of the kriging model relative to the current approach targeting houses built in the forties; in particular as more field data become available.

The application of this methodology to Flint water public system provided important information on the spatial distribution of lead and galvanized service lines in the City. First, according to indicator semivariograms both types of SL are spatially clustered with a range of 1.4 km for LSLs and 2.8 km for GSLs. There was also substantial variability over a few hundred meters, which hampers any spatial interpolation in particular for GSLs. Second, construction year appears to be a good predictor of the SL material: galvanized lines were mostly found in pre-1934 houses, while the frequency of LSLs peaked for houses built around World War II. Third, city records led to the over-identification of LSLs, likely because old records were not updated as these lines were being replaced. Future research should investigate whether these findings can be generalized to other US cities.

Acknowledgments

The author is grateful to the Michigan Department of Environmental Quality for providing the service line inspection data and answering questions about their sampling strategy. This research was funded by grant R44 ES022113-02 from the National of Environmental Health Sciences. This views stated in this publication are those of the author and do not necessarily represent the official views of the NIEHS.

References

- American Water Works Association. Water://Stats 1996 Distribution Survey. AWWA; Denver, CO: 1996.
- Brush, M. [Accessed April 6, 2017] Here is how to tell if you have lead pipes in your home. 2017 Mar 24. Available at: <http://www.alleghenyfront.org/heres-how-to-tell-if-you-have-lead-pipes-in-your-home/>
- Clark BN, Masters SV, Edwards MA. Lead release to drinking water from galvanized steel pipe coatings. *Environ Eng Sci*. 2015; 32(8):713–721. DOI: 10.1089/ees.2015.0073
- Cornwell DA, Brown RA, Via SH. National survey of lead service line occurrence. *J Am Water Works Ass*. 2016; 108(4):E182–E191.
- de Oliveira DP, Neill DB, Garret JH, Soibelman L. Detection of Patterns in Water Distribution Pipe Breakage Using Spatial Scan Statistics for Point Events in a Physical Network. *J Comput Civ Eng*. 2011; 25(1):21–30.
- Ehrmann, C. Effort to replace pipes to Flint homes off to slow start. Associated Press; 2017 Mar 19. Available at: <http://www.freep.com/story/news/local/michigan/flint-watercrisis/2017/03/19/flint-lead-pipes-water/99385754/> [Accessed April 8, 2017]
- Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett*. 2006; 27:861–874.

- Flint Safe Drinking Water Task Force. Recommendations on MDEQ's Draft Sentinel Site Selection. Feb. 2016 Retrieved from https://www.epa.gov/sites/production/files/2016-02/documents/task_force_recommendations_on_sentinel_site_selection_2-16.pdf on August 20, 2016
- Gold, R. UM-Flint GIS Center Mapping Flint Water System's Lead Service Lines. Jan 28. 2016 Retrieved from <https://news.umflint.edu/2016/01/28/10668/>
- Goovaerts, P. Geostatistics for Natural Resources Evaluation. Oxford University Press; New York, NY: 1997.
- Goovaerts P. AUTO-IK: a 2D indicator kriging program for the automated non-parametric modeling of local uncertainty in earth sciences. Comput Geosci. 2009; 35:1255–1270. [PubMed: 20161335]
- Goovaerts, P. Fate and Transport: Geostatistics and Environmental Contaminants. In: Nriagu, JO., editor. Encyclopedia of Environmental Health. Vol. 2. Burlington: Elsevier; 2011. p. 701-714.
- Goovaerts P. The drinking water contamination crisis in Flint: Modeling temporal trends of lead level since returning to Detroit Water System. Sci of the Total Environ. 2017a; 581–582:66–79.
- Goovaerts P. Monitoring the aftermath of Flint drinking water contamination crisis: Another case of sampling bias? Sci of the Total Environ. 2017b; 590–591:139–153.
- Goovaerts P, Journel AG. Integrating soil map information in modelling the spatial variation of continuous soil properties. Eur J Soil Sci. 1995; 46(3):397–414.
- Goovaerts P, AvRuskin G, Meliker J, Slotnick M, Jacquez GM, Nriagu J. Geostatistical modeling of the spatial variability of arsenic in groundwater of Southeast Michigan. Water Resour Res. 2005; 41(7):W07013. doi:10.1029.
- Goovaerts P, Wobus C, Jones R, Rissing M. Geospatial estimation of the impact of Deepwater Horizon Oil Spill on plant oiling along the Louisiana shorelines. Journal of Environmental Management. 2016; 180(15):264–271. [PubMed: 27240202]
- Grunwald S, Goovaerts P, Bliss CM, Comerford NB, Lamsal S. Incorporation of auxiliary information in the geostatistical simulation of the spatial distribution of soil nitrate-nitrogen in a mixed-use watershed. Vadose Zone J. 2006; 5:391–404.
- Hanna-Attisha M, LaChance J, Sadler RC, Champney Schnepf A. Elevated Blood Lead Levels in Children Associated With the Flint Drinking Water Crisis: A Spatial Analysis of Risk and Public Health Response. Am J Public Health. 2016; 106:283–290. [PubMed: 26691115]
- Hull, &., Associates, Inc. [Accessed April 9, 2017] Water Distribution System Lead Mapping Submittal. 2017. Available at: <http://edocpub.epa.ohio.gov/publicportal/ViewDocument.aspx?docid=588050>
- Journel AG. Nonparametric estimation of spatial distributions. Math Geol. 1983; 15(3):445–468.
- May, J. [Accessed April 8, 2017] Where, how and when Flint plans to replace 6,000 water lines. 2017 Mar 17. Available at: http://www.mlive.com/news/flint/index.ssf/2017/03/where_how_and_when_flint_plans.html
- Ohio Environmental Protection Agency. [Accessed April 9, 2017] Guidelines for Lead Mapping in Distribution Systems. 2017. Available at: http://epa.ohio.gov/Portals/28/documents/pws/PWS-04-001_01-06-2017.pdf
- Pannatier, I. VARIOWIN: Software for Spatial Data Analysis in 2D. Springer-Verlage; New York, NY: 1996.
- Pieper KJ, Tang M, Edwards MA. Flint water crisis caused by interrupted corrosion control: Investigating “ground zero” home. Environ Sci Technol. 2017; 51(4):2007–2014. DOI: 10.1021/acs.est.6b04034 [PubMed: 28145123]
- Rabin R. The lead industry and lead water pipes “A MODEST CAMPAIGN”. Am J Public Health. 2008; 98:1584–1592. [PubMed: 18633098]
- Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Aroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017; doi: 10.1111/ecog.02881
- Rosen MB, Pokhrel LR, Weir MH. A discussion about public health, lead and Legionella pneumophila in drinking water supplies in the United States. Sci of the Total Environ. 2017; 590–591:843–852.

- Smith, L. [Accessed April 8, 2017] Some cities know where lead water service lines are, others have “very rough to no idea”. 2016 Apr 1. Available at: <http://michiganradio.org/post/some-cities-know-where-leadwater-service-lines-are-others-have-very-rough-no-idea>
- Studzinski, J. Application of kriging algorithms for solving some water nets management tasks. In: Pillmann, W., Schade, S., Smitts, P., editors. Innovations in Sharing Environmental Observations and Information, Part 1: Environmental Informatics; Proceedings of EnviroInfo Ispra; 2011; Aachen: Shaker Verlag; 2011. p. 493-488.
- Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988; 240:1285–1293. [PubMed: 3287615]
- Troesken, W. The Great Lead Water Pipe Disaster. MIT Press; Cambridge, MA: 2006.
- US Environmental Protection Agency,. Office of Water. Lead and Copper Rule 40 CFR Part 141 Subpart I. 1991; 1991 [Accessed April 8, 2017] Available at: <https://www.epa.gov/dwreginfo/lead-and-copper-rule>.
- US Environmental Protection Agency, Office of Water. [Accessed April 8, 2017] Lead and Copper Monitoring and Reporting Guidance for Public Water Systems. 2002. Available at: <https://www.epa.gov/dwreginfo/lead-and-copper-rule-compliance-help-public-water-systems>
- US Environmental Protection Agency, Office of Ground Water & Drinking Water. [Accessed April 8, 2017] Memorandum: Clarification of Recommended Tap Sampling Procedures for Purposes of the Lead and Copper Rule. 2016a. Available at: <https://www.epa.gov/dwreginfo/memo-clarifying-recommendedtap-sampling-procedures-lead-and-copper-rule>
- US Environmental Protection Agency, Office of Policy. [Accessed February 2, 2017] EJSCREEN Technical Documentation. 2016b. Available at: https://www.epa.gov/sites/production/files/2016-07/documents/ejscreen_technical_document_20160704_draft.pdf
- White, E. Michigan, Flint to replace 18, 000 lead-tainted water lines. Associated Press; 2017 Mar 27. Available at: <http://bigstory.ap.org/article/c85f78dc0f6042dbbcbb6236a6dfbac1/michiganflint-replace-18000-lead-tainted-water-lines> [Accessed April 8, 2017]
- Wisely, J., Spangler, T. [Accessed April 6, 2017] Where are the lead pipes? In many cities, we just don’t know. 2016 Feb 27. Available at: <http://www.freep.com/story/news/local/michigan/flint-watercrisis/2016/02/27/lead-water-lines-lurk-unknown-many-cities/80551724/>
- Zhao, CK., Daly, C. Where are the hot zones: prioritization with historical pipe break. In: Firat Sever, V., Osborn, L., editors. Pipelines 2015 - Recent Advances in Underground Pipeline Engineering and Construction - Proceedings of the Pipelines 2015 Conference; August 23–26, 2015; Baltimore, Maryland. 2015. p. 1602-1607.

HIGHLIGHTS

- Most US cities cannot locate all lead service lines in their water supply systems.
- 3254 homes have been inspected in Flint to identify service line (SL) material.
- Galvanized lines (GSLs) were widespread and mostly observed in pre-1934 houses.
- Lead lines (LSLs) were clustered and found in houses build during World War II.
- Probability of finding LSL and GSL can be mapped using kriging.

Fig. 1.

Spatial distribution of houses that were inspected by MDEQ (A, red dots indicate lead service lines, while blue square denote galvanized service lines). To facilitate visualization of spatial patterns, the following statistics have been computed at the block group level: (B) % tax parcels sampled (stars indicate two highly sampled block groups in Ward 7; see text for comments), (C) % LSLs among inspected lines, and (D) % GSLs among inspected lines. The boundaries of nine Flint wards are overlaid on all maps. For bottom maps the color scale is centered on the mean percentage of LSLs (4.15%) and GSLs (18.22%), respectively.

Fig. 2.

Relative proportions of service lines that are made of three main types of material as function of the year the house was built. All frequency distributions were smoothed using a kernel of size 11 years and results based on less than 30 observations were discarded. For comparison purposes each curve was rescaled by the global percentage of each type of material in the dataset; relative proportion > 1 indicates that for this year of construction the percentage is larger than what is observed on average over all inspected homes.

Fig. 3.

Maps of the two housing characteristics used as covariates to predict the likelihood of presence of LSLs and GSLs: (A) year built, and (B) city records on composition of service lines. To facilitate visualization of spatial patterns, the following statistics were computed at the block group level: % houses built prior to 1934 (C) or between 1938 and 1951 (E), % service lines made of lead (D) or other known material (F) according to city records.

Fig. 4.

Experimental semivariograms with the model fitted for indicators of presence/absence of two types of SL material identified during house inspection: lead and galvanized. Two estimators were used: traditional or Matheron's estimator (A, B) and rescaled estimator (C, D) to account for clustering of inspected sites. In all graphs, the lower curve (red line) corresponds to indicator residuals; i.e. spatial variability in SL material that is not explained by the two housing characteristics: year built and city records on SL composition.

Fig. 5.

Receiver operating characteristics curves obtained by leave-one-out cross-validation for five different predictors of the probability of presence of LSLs (A, B) and GSLs (C, D). Right column shows the ROC curves for the subset of service lines that are harder to locate on the basis of built year and city records. A quantitative measure of the classification accuracy is the relative area under the ROC curve (AUC), which represents the average frequency of detection (best if close to 1).

Fig. 6.

Impact of sampling density on the accuracy, as measured by the AUC statistic, of five different predictors of the probability of presence of lead (A, B) and galvanized (C, D) service lines. Results have been averaged over 100 subsets of the data created by ward-based stratified random sampling.

Fig. 7.

Maps of the probability of presence of lead (left column) and galvanized (right column) service lines computed by simple kriging with varying local means derived from built year and city records. Results are displayed at the level of tax parcels (A, B) and block groups (C, D) to facilitate visualization. Bottom maps show the block groups averages of soft indicators. Black dots on block group maps indicate the location of houses where MDEQ inspection found a lead or galvanized service line.

Table 1

Contingency table comparing for 3254 sampling sites the composition of service line recorded during the MDEQ inspection (on-site data) with the information retrieved from digitized city records (digital data).

	On-site data					Total	Percentage
	Lead	Galvanized	Copper	Plastic	Other		
Digital data							
Lead	82	51	567	3	3	706	21.7
Other	8	322	1604	6	8	1948	59.9
Unknown	45	220	327	1	7	600	18.4
Total	135	593	2498	10	18	3254	100
Percentage	4.15	18.2	76.8	0.30	0.55	100	

Table 2

Numbers of parcels and statistics on MDEQ inspection results and housing characteristics within each ward in Flint. Poverty level represents the percentage of the block group population living in households where the income is less than or equal to twice the federal “poverty level.”

Statistics	Flint Ward								
	1	2	3	4	5	6	7	8	9
Total number of parcels	6564	6682	8092	5941	7382	4996	5230	6058	5094
Number of inspections	154	348	211	266	218	427	615	532	483
% parcels inspected	2.34	5.21	2.61	4.48	2.95	8.55	11.8	8.78	9.48
% LSLs (field data)	0.0	3.45	2.37	1.50	0.46	6.79	4.88	3.38	7.45
% GSLs (field data)	7.79	11.5	28.4	9.40	28.4	12.7	19.0	22.2	21.7
% pre-1934 houses	13.1	27.5	56.6	35.6	82.4	50.2	39.0	35.1	36.3
% 1938–1951 houses	27.6	36.1	14.5	19.8	6.35	24.4	25.5	20.4	27.4
% LSLs (digital data)	3.40	8.25	8.85	5.44	10.8	9.45	6.31	7.35	7.03
Block group poverty level (%)	66.3	63.0	74.7	65.0	73.6	67.1	57.6	59.4	65.1

Table 3

Percentages of lead and galvanized service lines found during house inspection for three classes of construction year and the three types of SL material reported in city records. These results were used as soft indicator data in residual kriging.

Digital data	LSLs			GSLs		
	Construction year			Construction year		
	< 1938	1938–1951	> 1951	< 1934	1934–1971	> 1971
Lead	1.66	29.0	0.0	10.8	3.79	0.0
Other	0.65	0.45	0.0	46.7	4.20	0.0
Unknown	1.48	19.2	0.0	63.1	4.30	0.0

Table 4

Number of unsampled tax parcels where the probability of presence of LSL or GSL, as estimated by residual kriging in Fig. 7, equals or exceeds various probability thresholds. The corresponding rates of detection and false positives were computed by validation analysis (Section 3.5); the minimum and maximum average rates observed for sampling densities ranging from 20% to 100% are reported.

Probability rule	Lead SLs		
	Number of parcels	LSLs detected (%)	False positives (%)
$p_{RK}^* = 0$	27,397	2.4–5.5	42.0–46.1
$p_{RK}^* > 0$	25,388	94.5–97.6	53.9–58.0
$p_{RK}^* > 0.01$	10,611	91.2–92.6	28.0–34.2
$p_{RK}^* > 0.02$	6002	89.2–90.8	21.5–23.8
$p_{RK}^* > 0.05$	5093	87.7–90.5	16.8–19.5
$p_{RK}^* > 0.1$	4134	85.9–89.0	13.5–14.8
$p_{RK}^* > 0.15$	3544	81.9–87.4	10.2–11.4
$p_{RK}^* > 0.2$	1103	73.0–81.4	6.2–8.1
$p_{RK}^* > 0.4$	138	28.5–52.9	0.7–1.2
$p_{RK}^* > 0.6$	32	13.2–36.6	0.1–0.4
Probability rule	Galvanized SLs		
	Number of parcels	GSLs detected (%)	False positives (%)
$p_{RK}^* = 0$	1638	0.6–1.0	4.0–5.7
$p_{RK}^* > 0$	51,147	99.0–99.4	94.3–96.0
$p_{RK}^* > 0.03$	46,123	96.3–97.1	77.4–81.1
$p_{RK}^* > 0.05$	27,277	87.6–89.3	33.4–44.5
$p_{RK}^* > 0.1$	22,345	83.7–85.4	23.7–27.0
$p_{RK}^* > 0.25$	20,216	79.5–79.9	16.2–16.5
$p_{RK}^* > 0.45$	17,934	63.2–71.8	11.6–13.5
$p_{RK}^* > 0.5$	9648	40.2–43.2	5.6–6.6

Probability rule	Number of parcels	Lead SLs	
		LSLs detected (%)	False positives (%)
$p_{RK}^* > 0.6$	7037	24.3–34.8	3.2–4.5
$p_{RK}^* > 0.7$	302	5.0–6.6	0.5–0.8