Published in final edited form as:

Sci Total Environ. 2019 January 10; 647: 1294–1304. doi:10.1016/j.scitotenv.2018.07.459.

Geostatistical Prediction of Water Lead levels in Flint, Michigan: a Multivariate Approach

Pierre Goovaerts

BioMedware Inc, Ann Arbor, MI 48106

Abstract

Despite several environmental crises, little research has been conducted on citywide geospatial modeling of water lead levels (WLL) in public distribution systems. This paper presents the first application of multivariate geostatistics to lead in drinking water within a distribution system, specifically in Flint, Michigan. One of the key features of the Flint data is their collection through two different sampling initiatives: (i) voluntary or homeowner-driven sampling whereby concerned citizens decided to acquire a testing kit and conduct sampling on their own (10,717 sites), and (ii) State-administered sampling where data were collected bi-weekly at 809 selected sites after training of residents by technical teams (sentinel sites). These two datasets were first averaged over the 41-week sampling period and each tax parcel to attenuate sampling fluctuations and create a set of 420 tax parcels sampled by both protocols. Both variables displayed a correlation of 0.62 while their direct and cross-semivariograms showed substantial nugget effect and a long range of 7.5 km. WLLs recorded at sentinel sites and deemed more reliable by city officials were then interpolated using cokriging to account for the more densely sampled voluntary data and information on service line composition (lead, other, or unknown) available for each of 51,045 residential tax parcels. Cross-validation demonstrated the greater prediction accuracy of the multivariate geostatistical approach relative to kriging and inverse square distance weighting interpolation using only sentinel data. This general procedure is applicable to other cities with aging infrastructure where lead in drinking water is a concern.

Graphical abstract

Corresponding author address: BioMedware, Inc. 167 Little Lake Drive, Ann Arbor, MI 48106 (USA), phone: 517-900-9283; fax: 734-913-2201; goovaerts@biomedware.com.

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

cokriging; inverse square distance weighting; cross-validation; voluntary sampling; cross-variogram

Introduction

According to a recent poll (McCarthy, 2017) more than half of Americans say they have concerns about the quality of their water as more people become educated about specific contaminants and take action in their homes. These worries have been fueled by the repeated occurrence of drinking water contamination crises that have plagued water distribution systems of cities like Washington, DC in 2001 (Edwards and Dudi, 2004), Greenville and Durham, NC in 2006 (Renner, 2009), and more recently Flint, MI (Hanna-Attisha *et al.*, 2016). In all cases the culprit was a change in water chemistry (e.g., change in disinfectant, switch in water supply, lack of corrosion control treatment) that caused lead to leach from underground service lines connecting each house to the water main as well as from indoor plumbing. A recent survey of US community water systems (Cornwell *et al.*, 2016) indicates that 6.1 millions of lead service lines (LSL) are currently present nationwide, serving between 15 and 22 million people (7% of customers). Galvanized steel pipes can also be a long-term source of lead as the “galvanized” surface zinc coating can contain up to 2% lead (Clark *et al.*, 2015). An estimated 7.5% of households overall have steel or galvanized steel service lines (American Water Works Association, 1996). Another source of lead is home plumbing which includes metal pipes, brass faucet fixtures and solder at pipe connections installed prior to 1986 (Cartier *et al.*, 2011).

Public water systems are responsible for monitoring lead and copper levels in their distribution system to assess compliance with the Lead and Copper Rule (LCR). 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 $\mu\text{g/L}$ for lead and 1.3 mg/L for copper (LCR, US EPA, 1991). High-risk homes are defined as sites where elevated levels of lead are likely to be found based on the presence of LSLs, lead pipes, or copper pipes soldered with lead. If non-compliance is observed the system must undertake a number of additional actions to control

corrosion. If the action level for lead is exceeded, the system must also inform the public about steps they should take to protect their health and may have to replace lead service lines under their control.

The delay in reporting high levels of lead in Flint drinking water was 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). Lead poisoning in Flint was exposed when water was sampled by an independent research team from Virginia Tech and 16.6% of the 271 water samples were found to be in excess of 15 µg/L of lead (FlintWaterStudy.org, 2015). This was the start of a vast sampling campaign to assess the extent of the lead contamination and restore public confidence in local officials and agencies by posting all results online (<http://www.michigan.gov/flintwater>).

Starting in February 2016, water samples were collected bi-weekly at more than 600 sentinel sites chosen by the EPA and MDEQ (Michigan Department of Environmental Quality) across the city to determine the general health of the distribution system and to track changes in lead concentrations over time (Flint Safe Drinking Water Task Force, 2016). After five rounds of sentinel sampling, a new sentinel program called “Extended Sentinel Site Program” started in June 2016, targeting specifically sites with high WLL during previous rounds or located in the highest-risk areas. Six additional sampling rounds were conducted in 2016 for this smaller network of less than 200 sites. Overall, these 11 sampling rounds resulted in the collection of 4,120 datapoints at 819 different sites over a 40-week time period (Goovaerts, 2018). Sentinel teams including a member of DEQ, a licensed plumber and a community member visited the houses selected to be part of the sentinel network. A plumbing inspection was conducted to verify the service line material entering the home and residents were shown how to draw samples of their water in a scientifically accurate manner (Goovaerts, 2017a).

This State-administered monitoring program was supplemented by a voluntary homeowner-driven sampling whereby concerned citizens received a free testing kit with written instructions to conduct sampling on their own (Goovaerts, 2017a,b). Despite the larger size and wider temporal extent of this database (18,760 samples collected over 53 weeks at 10,341 sites), officials have largely dismissed these data on the ground that there is a lack of uniformity and control on sampling procedure as homeowners did not undergo training (Bryce Feighner, personal communication, February 2, 2017). Albeit potentially less accurate, the spatial density of the voluntary sampling network can help capturing complex geographical patterns related, for example, to the time it takes for the water to travel from the treatment plant to the home plumbing system (water age). Indeed, longer times can decrease the effectiveness of corrosion control, increasing the dissolution of the layer coating the inside of the pipes (passivation layer) resulting in leaching and higher water lead levels (Wang *et al.*, 2014; Zahran *et al.*, 2017). In addition, Goovaerts (2017b) showed that the sentinel monitoring program was biased as sentinel sites with LSL were mostly built between 1935 and 1950 and located in parts of the city that were less poor. On the other hand, voluntary sampling turned out to reproduce fairly well the main characteristics (i.e. presence of lead service lines, construction year) of Flint housing stock.

Housing characteristics, in particular the presence of lead SL, represent another source of information on water lead level in a house. Although city records on the type of service lines are not always accurate (Goovaerts, 2017c), this information is spatially exhaustive as it is available for all 56,039 tax parcel units located within the City of Flint. It is noteworthy that a tax parcel can include multiple sampling sites, such as multiple rental units within an apartment complex or low-income housing developments.

The database assembled by the City of Flint and made available online has yet to undergo any rigorous statistical analysis beyond the production of location maps and computation of summary statistics. Using a data-driven approach, Abernethy *et al.* (2016) developed an ensemble of predictive models (e.g., random forest, logistic regression, linear discriminant analysis) to assess the risk of lead contamination in individual homes and neighborhoods in Flint. They trained these models using a wide range of data sources, including residential water tests, historical records, and city infrastructure data. Their analysis, however, ignored the spatial correlation among data and their origin (i.e., sentinel vs voluntary sampling networks). A time trend analysis was conducted by Goovaerts (2017a) who used joinpoint regression to model time series of water lead levels collected by the sentinel and voluntary sampling programs. This analysis carried out at the city and ward levels still ignored the spatial correlation among data and did not provide any tax parcel-based prediction. Goovaerts (2018) also performed a space-time analysis of these data, yet it was limited to sentinel data.

In general, little research has been devoted to citywide geospatial modeling of water lead levels in public distribution systems. Besides the aforementioned studies of Flint data, Wang *et al.* (2014) applied geographic information systems (GIS) and a hydraulic model of distribution systems to test the influences of pipe material, pipe age, water age, and other water quality parameters on lead/copper leaching in Raleigh (NC). In Symanski *et al.* (2004), mixed effect models were used to assess spatial fluctuations, temporal variability, and errors due to sampling and analysis for levels of disinfection by-products in water samples collected in households within the same distribution system.

The present study is the first application of multivariate geostatistics to the modeling of water lead levels in a distribution network. Since lead in drinking water mostly comes from premise plumbing materials and service lines instead of being transported through water mains (Del Toral *et al.*, 2013), traditional techniques based on Euclidean distances are appropriate. Cokriging is introduced as a way to combine sentinel data (primary variable) and densely sampled voluntary WLL data (auxiliary variable) in spatial interpolation, while accounting for their respective pattern of spatial variability and sampling procedure. Another auxiliary variable is composition of service lines which is available for each tax parcel unit where the prediction is conducted. The benefit of this multivariate approach over univariate kriging and inverse square distance weighting interpolation (IDW) is assessed using cross-validation.

2. Material and Methods

2.1. Flint data

24,515 WLL measurements recorded over the period 1/1/2016–10/13/2016 were downloaded from <http://www.michigan.gov/flintwater> (residential testing results). Records with an address that was incomplete or located outside the city of Flint were discarded, leading to a total of 22,672 observations: 18,701 collected by voluntary sampling and 3,971 recorded at sentinel sites. Data were then allocated to an individual tax parcel unit on the basis of their postal address, and only residential tax parcels were retained for the subsequent analysis (e.g., commercial and industrial parcels were excluded). A detailed description of sampling procedures and analysis of temporal variability can be found in Goovaerts (2017a,b,c).

For both voluntary and sentinel databases, observations were aggregated over the 41-week sampling period within each tax parcel, which is the smallest geographical unit for which housing characteristics are available. Time-averaging allowed filtering random fluctuations between sampling events which are caused, for example, by the random release of particulate lead following fluctuations in water pressure and vibrations in indoor pipes. The final dataset includes WLL values for 809 sentinel sites and 10,717 residences that were sampled voluntarily by homeowners. More than half of the sentinel sites (420) are located on tax parcels that were also visited as part of the voluntary sampling program, allowing the calibration of the relationship between both datasets.

A digital map of Flint's lead water pipe was obtained from the GIS center at the University of Michigan, Flint. The shapefile includes, for each of the 51,045 residential tax parcel units, the type of service line (SL) classified into three categories: lead (7.3%), other (69.2%), and unknown (23.5%). According to Goovaerts (2017c) city records result in the over-identification of LSLs, likely because old records were not updated as these lead lines were being replaced over time. Thus, data on SL composition were coded as a binary variable indicating whether the service line was non-lead (i.e., other) or lead/unknown. In other words, the two least accurate categories were merged. For prediction, these indicators were then replaced by the corresponding mean WLL observed at sentinel sites.

2.2. Multivariate prediction

Let $z_1(\mathbf{u}_0)$ be the water lead level for a tax parcel georeferenced by its centroid's geographical coordinates $\mathbf{u}_0$. The estimated value $z_1^*(\mathbf{u}_0)$ is computed as the following linear combination of n_1 closest sentinel WLL data $z_1(\mathbf{u}_\alpha)$, n_2 closest voluntary WLL data, $z_2(\mathbf{u}_\beta)$, and the sentinel mean WLL conditional to the type of service line (SL=other or lead/unknown) listed in city records for $\mathbf{u}_0$:

$$z_1^*(\mathbf{u}_0) = \sum_{\alpha=1}^{n_1} \lambda_{1\alpha} \times z_1(\mathbf{u}_\alpha) + \sum_{\beta=1}^{n_2} \lambda_{2\beta} \times z_2(\mathbf{u}_\beta) + \lambda_{30} \times z_3(\mathbf{u}_0) \quad (1)$$

with $z_3(\mathbf{u}_0) = m_1 |SL(\mathbf{u}_0)|$. The weights λ are solution of a system of linear equations known as rescaled or standardized ordinary cokriging (Isaaks and Srivastava, 1989, p. 409–416; Deutsch and Journel, 1992, p. 70), which is a variant of ordinary cokriging where the following single unbiasedness constraint is imposed on the kriging weights:

$$\sum_{\alpha=1}^{n_1} \lambda_{1\alpha} + \sum_{\beta=1}^{n_2} \lambda_{2\beta} + \lambda_{30} = 1 \quad (2)$$

Removing the traditional constraint that all secondary data weights sum to zero attenuates two common shortcomings of traditional ordinary cokriging (Goovaerts, 1998): (1) some of the secondary data weights are negative, thereby increasing the risk of getting unacceptable estimates such as negative WLLs, and (2) most of the weights assigned to secondary data tend to be small, thus reducing the influence of the secondary information. Under the single constraint (2) the unbiasedness of the ordinary cokriging estimator is typically ensured by rescaling the secondary variables so that their means are equal to that of the primary variable. This step was not required here as all three variables represent water lead levels and have the same measurements units.

Solving the cokriging system requires a model for the semivariogram of each variable and the cross-semivariogram between any two variables. The matrix of semivariograms $\Gamma(\mathbf{h})$ was modeled using the following linear model of coregionalization:

$$\begin{pmatrix} \gamma_{11}(\mathbf{h}) & \gamma_{12}(\mathbf{h}) & \gamma_{13}(\mathbf{h}) \\ \gamma_{21}(\mathbf{h}) & \gamma_{22}(\mathbf{h}) & \gamma_{23}(\mathbf{h}) \\ \gamma_{31}(\mathbf{h}) & \gamma_{32}(\mathbf{h}) & \gamma_{33}(\mathbf{h}) \end{pmatrix} = \sum_{l=1}^L \begin{pmatrix} b_{11}^l & b_{12}^l & b_{13}^l \\ b_{21}^l & b_{22}^l & b_{23}^l \\ b_{31}^l & b_{32}^l & b_{33}^l \end{pmatrix} g_l(\mathbf{h}) = \sum_{l=1}^L \mathbf{B}_l g_l(\mathbf{h}) \quad (3)$$

where L is the number of basic semivariogram models (e.g., nugget effect and two exponential models here), and the corresponding matrices of sills $\mathbf{B}_l$ are all positive semi-definite. The coregionalization matrices $\mathbf{B}_l$ were estimated using the iterative procedure developed by Goulard and Voltz (1992).

2.3. Cross-validation

The accuracy of the predictive models was assessed by cross-validation whereby each observation at sentinel site was removed one at a time and re-estimated using data collected at neighboring sentinel and voluntary sites, as well as the co-located SL-based mean WLL. The following performance criteria were then computed from n kriging estimates:

- the mean error (ME) of prediction as:

$$ME = \frac{1}{n} \sum_{a=1}^n (z_1^*(\mathbf{u}_a) - z_1(\mathbf{u}_a)) \quad (4)$$

- the mean absolute error (MAE) of prediction as:

$$MAE = \frac{1}{n} \sum_{\alpha=1}^n |z_1^*(\mathbf{u}_a) - z_1(\mathbf{u}_a)| \quad (5)$$

- the mean square standardized residual (MSSR) as:

$$MSSR = \sum_{\alpha=1}^n \frac{(z_1^*(\mathbf{u}_a) - z_1(\mathbf{u}_a))^2}{\sigma_{CK}^2(\mathbf{u}_a)} \quad (6)$$

where $\sigma_{CK}^2(\mathbf{u}_a)$ is the cokriging variance.

A mean error close to zero indicates a lack of bias, while the mean absolute error should be as small as possible. If the actual estimation error is equal, on average, to the error predicted by the model, the MSSR statistic should be about one (Wackernagel, 1998, p. 91). A fourth statistic is the linear correlation coefficient ρ between observed and estimated WLLs.

One application of the predictive models is to prioritize any further sampling or intervention through a ranking of tax parcels from highly hazardous to less hazardous on the basis of kriging estimates. The ability of this ranking to identify successfully sites where WLL is greater or equal to the EPA action level of 15 $\mu\text{g/L}$ was assessed using Receiver Operating Characteristics (ROC) curves which plot the probability of false positive versus the probability of detection (Fawcett, 2006). The accuracy of the classification 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 (i.e., $z_c \geq 15 \mu\text{g/L}$) higher than a randomly chosen negative instance (i.e., $z_c < 15 \mu\text{g/L}$).

3. Results and Discussion

3.1 Data configuration and summary statistics

Figure 1 (left column) shows the location of all primary and secondary data within the nine wards in the City of Flint: 809 sentinel sites (A), 10,717 parcels sampled voluntarily (C), and 51,045 residential tax parcels with city records on type of service line (E). For visualization purposes, these parcel-level data were aggregated within each of the 40 census tracts (Figure 1, right column). Because of their strongly positively skewed distribution (concentrations range from 0 to 5,986 $\mu\text{g/L}$) and large proportion of zero values (41.6%), all WLL data (i.e., WLLs aggregated over the 41-week sampling period within each tax parcel) were first transformed using the following formula $\text{Log}_{10}(z+1)$.

Location maps (Figs. 1A,C) illustrate the spatial clustering of sentinel sites, in particular along the boundary that Ward 2 shares with wards 3 and 6 and in the center of Ward 7. On the other hand, much fewer sentinel sites relative to non-sentinel sites are located in wards 1, 3, 4 and 5. Such sampling bias, which was studied in details in Goovaerts (2017b),

emphasizes the benefit of incorporating data collected by the voluntary network since it captures Flint housing stock more accurately. This preferential sampling also explains the discrepancies between the census-tract level maps of WLL (Figs. 1B,D); CT sentinel values display a wider range because averages are based on fewer observations. Both maps however indicate the presence of lower WLLs in the Eastern and Northern parts of the city where most service lines are not made of lead (Fig. 1F). Sampling gaps correspond to non-residential parcels (e.g., commercial and industrial parcels, Figure 2) that were excluded from the analysis.

The relationship between the primary and secondary variables is illustrated using the scatterplots in Fig. 3. Water lead levels recorded by both sampling initiatives at 420 tax parcels (green dots in Fig. 3A) display a linear correlation of 0.62 (Fig. 3B), which validates the visual comparison of Figs. 1B and 1D. The greater density of dots under the 45° line highlights the tendency for values measured at voluntary sites to be larger than their sentinel counterparts. The relationship between WLLs and presence/absence of lead service lines was first assessed by comparing the mean WLL for two categories of tax parcels: lead or unknown SL vs other-than-lead SL. The respective means (units = $\log_{10}(\mu\text{g/L})$) are 0.650 vs 0.486 for the sentinel program and 0.571 vs 0.383 for the voluntary sampling networks. As expected, lower lead levels are found in houses that are not serviced by lead lines. Another way to visualize this relationship for both datasets is to compare the values after aggregation to the census tract level. The relationship between WLL and SL composition is especially strong ($r=-0.62$) for the voluntary sampling network (Fig. 3D); the weaker relationship for the sentinel sampling program ($r=-0.25$) could be partially caused by the lower accuracies of CT values computed from fewer sentinel sites (small number problem).

The outlier in Fig. 3D corresponds to the most central census tract in the City (blue polygon in Fig. 1D) where the 68 recorded water lead levels are much lower than what would be expected given the large percentage of lead SL (i.e., low percentage of other-than-lead SL). The culprit is the location of 67 out of the 68 samples in a public housing complex built post-1970, so likely without lead service lines. Yet, because many public records on SL composition have not been updated (Goovaerts, 2017c), half of these parcels are wrongly shown as having lead SL.

3.2 Coregionalization analysis

Figure 4 shows the experimental direct and cross semivariograms with the linear model of coregionalization fitted using a FORTRAN implementation of the algorithm described in Goovaerts (1997, p. 443–446). This model consists of three basic structures: a nugget effect and two exponential models with ranges of 1.5 and 7.5 km, respectively. The corresponding sills are listed with the nugget effects in Table 1. All three direct semivariograms display very large nugget effect. The substantial short-range variability for WLLs likely reflects the heterogeneity in housing conditions (e.g., renovated houses with new meter and pipes located next to houses with original plumbing) as well as the lack of uniformity of sampling conducted by homeowners since even with simple instructions it is difficult to ensure strict adherence to any sampling protocol (Del Toral *et al.*, 2013). On-site training of homeowners by scientists is expected to attenuate sampling variability, which would explain the smaller

nugget effect for sentinel sites (Fig. 4A) compared to the voluntary sampling network (Fig. 4B). The nugget effect for service line composition could reflect the inaccuracies of city records as these consisted of 45,000 index cards and 240 parcel maps showing city blocks divided into little squares of property, with a code indicating which type of SL was running into each house (Goovaerts, 2017c).

The spatial structure of WLLs is related to the “neighborhood effect” since houses in the same neighborhood tend to have been built at the same time (i.e., similar plumbing system) and have similar water age. Goovaerts (2017b) found that construction year was a good predictor of service line material: galvanized lines were mostly found in pre-1934 houses, while the frequency of lead service lines (LSLs) peaked for houses built around World War II. Thus, the neighborhood effect could also explain the spatial structure of the semivariogram for the presence/absence of other-than lead service lines (Fig. 4C).

Attributing the large nugget effect of the direct semivariograms to sampling variability for WLL and data inaccuracies for SL is confirmed by the small nugget effect displayed by the cross-semivariograms, in particular for the pair voluntary WLL vs SL (Fig. 4F). Indeed, an interesting property of the cross-semivariogram is that its nugget effect is due only to the covariance between the microscale variations of the two variables (Goovaerts, 1997). If measurement errors greatly contribute to the nugget effect observed on the direct semivariograms, then a small nugget component on the cross-semivariogram is expected since these measurement errors are likely independent. Conversely, a large nugget effect on the cross-semivariogram would suggest that the nugget effect on the direct semivariograms is mainly due to microscale variations common to the two variables. The interpretation of the nugget effect on the two other cross-semivariograms is hampered by the smaller size of the sentinel dataset; i.e., only 420 tax parcels were jointly sampled by the voluntary and sentinel monitoring networks.

The lack of spatial structure for the cross-semivariogram between sentinel WLLs and SL (Fig. 4E) is in agreement with the lack of clear relationship between CT-level aggregates of these two variables (Fig. 3C). Despite being based on fewer data pairs (420 vs 809), the cross-semivariogram between sentinel and voluntary WLLs displays a clearer structure (Fig. 4D). Last, the best structured cross-semivariogram is for the two most densely sampled variables, voluntary WLL and parcel SL, which are also strongly correlated (Fig. 4F). Note that data on SL composition were transformed into WLL before performing cokriging (recall Section 2.1), which explains why the scatterplots display a negative relationship (i.e., % other-than-lead SL vs WLL), while the relationship is positive on the cross-semivariograms (SL-based WLL vs WLL).

3.3 Cross-validation

The coregionalization model of Fig. 4 was used to conduct a cross-validation analysis whereby each observation at sentinel site was removed at a time and re-estimated using data collected at neighboring sentinel and voluntary sites, as well as the co-located SL-based mean WLL. To assess the benefit of secondary information, a similar analysis was performed in a univariate setting where only sentinel sites were used for estimation with

ordinary kriging and the direct semivariogram of sentinel WLL data (Fig. 4A). Calculations were performed using Gslib software (Deutsch and Journel, 1992).

The impact of the number of neighbors (primary and secondary data) on results was investigated using a sensitivity analysis. In the univariate case, kriging was applied with $n_1=5$ to 100 sentinel sites and in each case all five performance criteria described in Section 2.3 were computed. A principal component analysis was then performed to summarize the redundant information provided by these five statistics; e.g., MAE, ME and ρ are strongly correlated. Note that two statistics had to be transformed so that their value could directly be interpreted as deviation from the optimum of 0 for $|\text{ME}|$ and 1 for MSSR: $|\text{ME}|$ and $|1-\text{MSSR}|$. The first component (PC1) is positively correlated with ρ and ROC AUC, while the contributions of MAE and ME are negative. Its value increases with the number of neighbors before levelling off around $n_1=59$, which is the optimum (Fig. 5E). Fluctuations are however of small magnitude in absolute terms. For example, Figs. 5A&C indicate that MAE decreases from 0.39 to 0.37, while the ROC AUC statistic increases from 0.595 to 0.630. A similar cross-validation was conducted for IDW and results (Figs. 5A&C; dashed line) illustrate the benefit of a geostatistical approach (i.e., lower MAE and larger AUC) although differences are rather small, likely because of the large nugget effect of the semi-variogram. The MSSR statistic could not be computed given that IDW does not provide a prediction error variance; hence no PCA was performed for IDW.

Similar plots were created for cokriging where the same three variables (MAE, AUC, PC1) are plotted as a function of the number of voluntary WLL data ($n_2=2$ to 100) for four different numbers of sentinel sites $n_1=9, 19, 39$, and 59 (optimum for OK). As expected the magnitude of prediction errors decreases with the number of secondary data (Fig. 4B). The decrease in AUC (Fig. 4D) might seem counter-intuitive; yet these fluctuations are of small magnitude and the impact of the number of sentinel sites, as measured by differences between curves, is negligible as well. However, these plots emphasize the benefit of incorporating secondary information: the peak MAE decreases from 0.40 (OK) to 0.32 (CK), while the AUC statistic rises from 0.63 (OK) to 0.76 (CK). A principal component analysis was conducted for $n_1=59$ and the first component is plotted in Fig 5F. Unlike for ordinary kriging, PC1 is negatively correlated to ROC AUC, hence the maximum value is not optimum. The number of voluntary WLL data was set to $n_2=28$, since it leads to the smallest MAE (Fig. 5B) while corresponding to a peak in the ROC AUC (Fig. 5D) and the beginning of a plateau for the first principal component (Fig. 5F).

Results in Fig. 5 are based on 809 sentinel sites, including 420 tax parcels that were also visited by the voluntary sampling program (co-located data, COL dataset). One should expect the cokriging estimates to be more accurate at these 420 sites where secondary WLL data are available compared to the 389 sites that were sampled only by the sentinel program (SEN dataset). This is illustrated in Table 2 where performance statistics were computed for the two datasets separately. Results are also listed for ordinary kriging and IDW to emphasize differences between the two datasets even for univariate prediction. For the COL dataset, all interpolation algorithms lead to underestimation (i.e., negative ME) and larger prediction errors (i.e., greater MAE) compared to the SEN dataset. The three other statistics (ρ , MSSR, AUC) indicate better predictions for the COL dataset. The culprit for these

discrepancies is the tendency for both sampling initiatives to target the most hazardous sites, resulting in higher water lead levels and proportions of data above 15 µg/L for the COL dataset (14.3% observations > 15 µg/L) versus the SEN dataset (5.14%). Higher WLLs tend to be under-estimated with larger prediction errors because of kriging smoothing effect. On the other hand, these higher values are easier to detect (i.e., larger ROC AUC) and their estimates are more strongly correlated with the observed values (Table 2). Regardless of the dataset the correlation is however very weak ($\rho=0.18\text{--}0.24$) and the benefit of OK vs IDW is small because of the large nugget effect displayed by the direct semivariogram of sentinel WLL data (Fig. 4A).

Except for the MSSR statistic in the COL dataset, cokriging systematically leads to better prediction performances than univariate predictions. As expected the benefit is the largest for parcels with WLL data collected by voluntary sampling (COL dataset). The improvement is much smaller where only secondary information on service lines is available (SEN dataset), which is caused by the inaccuracies of the SL data and the substantial short-range variability displayed by both sentinel and voluntary WLL data (Figs. 4A,B).

3.4 Mapping water lead levels

Based on cross-validation results, water lead levels were predicted for each of the 51,045 residential tax parcels in Flint using cokriging with WLL data collected at the closest $n_1=59$ sentinel sites and $n_2=28$ voluntary sites, as well as the co-located SL-based mean WLL (Fig. 6A). Cokriging was conducted using the public-domain software SGEMS (Remy *et al*, 2009). The map of cokriging estimates displays a pattern similar to the one observed for the voluntary WLL data, in particular after aggregation at the census tract level; compare Fig. 6B to Fig. 1D. Larger WLLs are predicted in the central and southwest parts of the city which includes the largest percentages of lead service lines (Fig. 6D). This relationship is illustrated in the scatterplot of Fig. 6E; the correlation coefficient is 0.77. For intervention, it is also useful to map the percentage of cokriging estimates above the action level of 15 µg/L (Fig. 6C). This map provides a clear picture of census tracts that should be targeted for further sampling, which might not necessarily be the ones with the largest percentage of lead services lines as illustrated by the scatterplot of Fig. 6F. This result confirms earlier findings that home lead service lines may not be the largest contributor of lead in Flint water, and lead contamination may be caused by interior plumbing (Dolan, 2016a; Goovaerts, 2017a,b).

4. Conclusions

This paper presented the first application of multivariate geostatistics to the modeling of water lead levels in a distribution network. The methodology was illustrated using a large database of observations collected in Flint, MI through two different sampling initiatives: a State-administered program that monitored lead levels at 809 sentinel sites and a voluntary homeowner-driven approach that sampled 20% of the 51,045 residential tax parcel units in the city. Unlike the sentinel program, the voluntary sampling captured fairly well the main characteristics of Flint housing stock (i.e., presence of lead service lines, construction year); see Goovaerts (2017b). Yet, the lack of supervised training of homeowners led authorities to question the way water samples were collected on a voluntary basis and the reliability of test

results. Cokriging allowed the incorporation of both datasets in the predictive model without the need for merging them and assuming they have the same spatial variability, nor requiring the availability of both variables at all locations (isotopic or equally-sampled case). It also provided a flexible approach to account for a third exhaustively sampled data layer in the form of old city records on service line composition. Even in presence of moderate correlation (i.e., 0.4–0.6), the much larger sampling density of secondary data relative to primary data (i.e., the number of parcels that were sampled voluntarily is one order of magnitude larger than the number of sentinel sites) is expected to increase the benefit of cokriging over any univariate interpolation method (Goovaerts, 1997).

Despite a sizable database assembled by the State of Michigan, the geostatistical analysis was hampered by the existence of substantial variability over a few hundred meters. In many countries such as Canada or France water quality sampling is conducted by a trained technician. On the other hand, in the US water samples are collected directly by homeowners, which can create variability among households even after the on-site training that took place at sentinel sites. Other sources of fluctuation include heterogeneity in the plumbing system (e.g., renovation, installation of a new meter), location of sampled faucets (e.g., bathroom vs kitchen), or water temperature (e.g., lead solubility increases with water temperature), to name a few. Variability in Flint water samples was also partly due to insoluble lead being released throughout the water system, effectively creating a ‘Russian Roulette’ problem with respect to the temporal variability of lead exposure (Dolan, 2016b). This random variability was partly filtered by aggregating the data over the 41-week sampling period within each tax parcel. In the future, such within-parcel variability could be incorporated as a local measure of data reliability and accounted for using kriging with non systematic measurement errors (Chilès and Delfiner, 1999).

The present study was the first to describe and quantify the spatial relationship and correlation between water lead levels recorded by the two sampling programs. Given the aforementioned sources of sampling heterogeneity and the fact that tax parcels can include multiple housing units, the correlation of 0.61 between the two datasets was unexpected and improved substantially the prediction of sentinel WLLs for parcels that were also sampled on a voluntary basis. For the other parcels cokriging improved little over kriging and IDW given the substantial nugget effect for all direct semivariograms and the outdated city records on service line composition. The mediocre prediction performances (e.g., correlation coefficient between observed and predicted values is 0.2) are a concern since more than 80% of tax parcels in Flint were not sampled at all.

In the future, several approaches will be investigated to improve the preliminary multivariate analysis described here. First, indicator transform should provide a better, yet more demanding, alternative to lognormal transform of water lead levels to account for the very large proportion of zero values (41.6%). This non-parametric approach would also provide a straightforward approach to map the probability of exceeding the action level of 15 µg/L at unsampled locations, which is critical to prioritize interventions. Given the amount of spatial heterogeneity recorded over short distances one should question the pertinence of making predictions at the tax parcel level. More appropriate spatial supports for prediction could be census block groups which are statistical divisions of census tracts and are generally defined

to contain between 600 and 3,000 people. The city of Flint includes 132 block groups and 40 census tracts. Such spatial aggregation or upscaling would be a way to filter between-household fluctuations which appears to be mainly noise. A stochastic simulation approach should thus be implemented to derive statistic such as probability of exceeding specific thresholds at these different spatial scales (Goovaerts and Glass, 2014).

Acknowledgements

This research was funded by grant R44 ES022113-02 from the National Institute of Environmental Health Sciences. The views stated in this publication are those of the author and do not necessarily represent the official views of the NIEHS.

References

- Abernethy J, Anderson C, Dai C et al. (2016). Flint water crisis: Data-driven risk assessment via residential water testing. arXiv preprint arXiv:1610.00580. Available at <https://arxiv.org/abs/1610.00580>. Accessed November 5, 2017.
- American Water Works Association. (1996) Water://Stats 1996 Distribution Survey AWWA, Denver, CO.
- Cartier C, Laroche L, Deshommes E, Nour S, Richard G, Edwards M, Prévost M (2011). Investigating dissolved lead at the tap using various sampling protocols. J. Am. Water Works Ass 103, 55–67.
- Chilès JP, Delfiner P (1999). Geostatistics: Modeling Spatial Uncertainty John Wiley and Sons, New York, 695 p.
- Clark BN, Masters SV, Edwards MA (2015). Lead release to drinking water from galvanized steel pipe coatings. Environ. Eng. Sci 32(8), 713–721. 10.1089/ees.2015.0073.
- Cornwell DA, Brown RA, Via SH (2016). National survey of lead service line occurrence. J. Am. Water Works Ass 108(4), E182–E191.
- Del Toral MA, Porter A, Schock MR, 2013 Detection and evaluation of elevated lead release from service lines: A field study. Environ. Sci. Technol 47(16), 9300–9307. [PubMed: 23879429]
- Deutsch CV, Journel AG (1992). GSLIB: Geostatistical Software Library and user's guide. Oxford Univ. Press, New York, 340 p.
- Dolan M (2016a, 9 8). Study: Flint lead contamination goes beyond service pipes. Detroit Free Press Available at <http://www.freep.com/story/news/local/michigan/flint-water-crisis/2016/09/08/study-flint-lead-contamination-goes-beyond-service-pipes/89994636/> Accessed November 1, 2017.
- Dolan M (2016b, 4 12). Researcher: Flint water 'like Russian roulette'. Detroit Free Press Available at <https://www.freep.com/story/news/local/michigan/flint-water-crisis/2016/04/12/researcher-improved-flint-water-still-isnt-safe-enough/82928284/> Accessed June 25, 2018.
- Edwards M, Dudi A (2004). Role of chlorine and chloramine in corrosion of lead-bearing plumbing materials. J. Am. Water Works Ass 96(10), 69–81.
- FlintWaterStudy.org (2015). Lead Results from Tap Water Sampling in Flint, MI during the Flint Water Crisis. Available at <http://flintwaterstudy.org/2015/12/complete-dataset-lead-results-in-tap-water-for-271-flint-samples/> Accessed August 28, 2016.
- Flint Safe Drinking Water Task Force Recommendations on MDEQ's Draft Sentinel Site Selection. (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.
- Fawcett T (2006). An introduction to ROC analysis. Pattern Recognit. Lett 27, 861–874
- Goovaerts P (1997). Geostatistics for Natural Resources Evaluation Oxford Univ. Press, New-York, 483 p.
- Goovaerts P (1998). Ordinary cokriging revisited. Math. Geol 30(1), 21–42.
- Goovaerts P (2017a). The drinking water contamination crisis in Flint: Modeling temporal trends of lead level since returning to Detroit Water System. Sci. of the Total Environ 581–582, 66–79.
- Goovaerts P (2017b). Monitoring the aftermath of Flint drinking water contamination crisis: Another case of sampling bias? Sci. of the Total Environ 590–591, 139–153.

- Goovaerts P (2017c). How geostatistics can help you find lead and galvanized water service lines: The case of Flint, MI. *Sci. of the Total Environ* 599–600, 1552–1563.
- Goovaerts P (2018). Flint drinking water crisis: a first attempt to model geostatistically the space-time distribution of water lead levels. In: Sagar BSD, Cheng Q, Agterberg F (Eds.), *Springer Handbook of Mathematical Geosciences: Fifty Years of IAMG*, Chapter 14.
- Goovaerts P, Glass G (2014). Geostatistical modeling of the spatial distribution of surface soil arsenic around a smelter. *J. Jpn. Soc. Soil Phys* 128, 5–10.
- Goulard M, Voltz M (1992). Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Math. Geol* 24(3), 269–286.
- Hanna-Attisha M, LaChance J, Sadler RC, Champney Schnepf A (2016). 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* 106, 283–290. [PubMed: 26691115]
- Isaaks E, Srivastava R (1989). *An introduction to applied geostatistics* Oxford Univ. Press, New York, 561 p.
- McCarthy J (2017). In US, water pollution worries highest since 2001 Gallup News, March 31, 2017. <http://news.gallup.com/poll/207536/water-pollution-worries-highest-2001.aspx>. Accessed October 19, 2017.
- Remy N, Boucher A, Wu J (2009). *Applied Geostatistics with SGEMS: A User's Guide*, Cambridge University Press, New York, NY, 312 p.
- Renner R (2009). Out of plumb: when water treatment causes lead contamination. *Environ. Health Perspect* 117(12), A542–7. [PubMed: 20049189]
- Symanski E, Savitz DA, Singer PC (2004). Assessing spatial fluctuations, temporal variability, and measurement error in estimated levels of disinfection by-products in tap water: implications for exposure assessment. *Occup. Environ. Med* 61, 65–72. [PubMed: 14691275]
- US Environmental Protection Agency, Office of Water (1991). Lead and Copper Rule 40 CFR Part 141 Subpart I. 1991 <https://www.epa.gov/dwreginfo/lead-and-copper-rule>. Accessed November 5, 2017.
- Wackernagel H (1998). *Multivariate geostatistics*, 2nd completely revised edition Springer, Berlin
- Wang Z, Devine H, Zhang W et al. (2014). Using a GIS and GIS-assisted water quality model to analyze the deterministic factors for Lead and Copper corrosion in drinking water distribution systems. *J. Environ. Eng* 140, A4014004
- Wisely J, Spangler T (2016, 2 27). Where are the lead pipes? In many cities, we just don't know <http://www.freep.com/story/news/local/michigan/flint-water-crisis/2016/02/27/lead-water-lines-lurk-unknown-many-cities/80551724/> Accessed April 6, 2017.
- Zahran S, McElmurry SP, Sadler RC (2017). Four phases of the Flint Water Crisis: Evidence from blood lead levels in children. *Environ. Res* 157, 160–172. [PubMed: 28570960]

HIGHLIGHTS

- Flint water lead level was monitored by the State of Michigan at 809 sentinel sites.
- 10,717 residences were sampled voluntarily for lead levels by homeowners.
- Cokriging allowed incorporating both sentinel and voluntary data in prediction.
- Substantial short-range variation hampered spatial prediction of water lead levels.
- Interior plumbing might contribute more to lead in Flint water than service lines.

Fig. 1.

Location of residential tax parcel data and results of their aggregation at the CT level: (A,B) water lead levels recorded at sentinel sites, (C,D) water lead levels recorded at sites visited during the voluntary homeowner-driven sampling, (E) city records on composition of service lines for all residential tax parcels, and (F) percentage of all residential tax parcels with other-than-lead service lines. The boundaries of nine Flint wards are overlaid on all maps. Shaded polygons indicate census tracts without any sentinel site (missing values), while white areas in location maps correspond to non-residential tax parcels (Figure 2) that were excluded from the analysis. Maps in right column were created using a diverging color scheme (i.e., 3 break values); locations with values in between the breaks receive a blend of the two break colors.

Parcel classification

Fig. 2.

Parcel classification use and possible ROW (right of way). The focus of geostatistical modeling is on residential tax parcels since they were the target of sentinel and voluntary sampling programs.

Fig. 3.

(A) Location of tax parcels visited only by the sentinel sampling program (×) and by both sampling programs (green dots). Scatterplots illustrate the relationship between: (B) water lead levels recorded by the sentinel and voluntary sampling networks for 420 tax parcels (solid line depicts perfect agreement, correlation =1), (C,D) sentinel and voluntary water lead levels averaged at the CT level and the corresponding percentages of other-than-lead service lines. Star (*) in panel D indicates an outlier; see text for explanations.

Fig. 4.

Experimental direct and cross semivariograms for the primary and two secondary variables, with the linear model of coregionalization fitted; see Table 1 for parameters.

Fig. 5.

Impact of the number of primary and secondary data on the results of a cross-validation analysis conducted for univariate (left column) and multivariate interpolation (right column). Performance statistics include: (A,B) the mean absolute error of prediction (MAE), (C,D) the ROC AUC, and (E,F) the first principal component of a PCA conducted on the correlation matrix of five performance statistics (higher scores indicate better predictions).

Fig. 6.

(A) Water lead levels (log₁₀ scale) estimated by cokriging for each of the 51,045 residential tax parcels in Flint, and results of their aggregation at the CT level: (B) average WLL, (C) percentage of estimates above the action level of 15 µg/L, (D) percentage of tax parcels with lead service lines. Scatterplots illustrate the relationship between CT percentages of tax parcels with lead service lines and: (E) average WLL estimates, and (F) percentages of cokriging estimates above the action level of 15 µg/L. All maps were created using a diverging color scheme (i.e., 3 break values); locations with values in between the breaks receive a blend of the two break colors.

Table 1.

Parameters of the linear model of coregionalization (i.e., nugget effect and sills for both exponential models) fitted to all six direct and cross-semivariograms displayed in Figure 4.

Semivariograms	Basic models		
	Nugget effect	Exp(range=1.5km)	Exp(range=7.5km)
Pb (sentinel)	$1.806 \cdot 10^{-1}$	$6.527 \cdot 10^{-2}$	$5.03 \cdot 10^{-2}$
Pb (voluntary)	$3.635 \cdot 10^{-1}$	$3.042 \cdot 10^{-2}$	$6.949 \cdot 10^{-2}$
Pb (SL parcels)	$4.536 \cdot 10^{-3}$	$1.527 \cdot 10^{-3}$	$8.565 \cdot 10^{-4}$
Pb (sentinel vs voluntary)	$1.294 \cdot 10^{-1}$	$4.452 \cdot 10^{-2}$	$5.912 \cdot 10^{-2}$
Pb (parcel vs sentinel)	$4.667 \cdot 10^{-3}$	$5.366 \cdot 10^{-4}$	$7.157 \cdot 10^{-4}$
Pb (parcel vs voluntary)	$5.599 \cdot 10^{-4}$	$6.667 \cdot 10^{-4}$	$9.114 \cdot 10^{-4}$

Table 2.

Results of a cross-validation analysis conducted by removing one sentinel data at a time and re-estimating this value using inverse square distance weighting interpolation (IDW), ordinary kriging (OK), and cokriging (CK). The five performance criteria described in Section 2.3 were computed for two sets of sentinel sites: 420 sites visited by both sentinel and voluntary sampling programs (Co-located, COL), and 389 sites that were sampled only by the sentinel program (Sentinel, SEN).

Performance criteria	Datasets					
	Sentinel (n=389)			Co-located (n=420)		
	IDW	OK	CK	IDW	OK	CK
ME	0.074	0.085	0.058	-0.091	-0.079	-0.027
MAE	0.342	0.332	0.322	0.411	0.405	0.308
MSSR	NA *	0.695	0.775	NA *	1.046	1.113
ROC AUC for 15 µg/L	0.593	0.582	0.618	0.635	0.65	0.807
p observed vs estimated WLLs	0.195	0.178	0.207	0.24	0.211	0.634

* MSSR cannot be computed for IDW as it is a deterministic procedure