Chapter 14
Flint Drinking Water Crisis: A First Attempt to Model Geostatistically the Space-Time Distribution of Water Lead Levels

Pierre Goovaerts

Abstract The drinking water contamination crisis in Flint, Michigan has attracted national attention since extreme levels of lead were recorded following a switch in water supply that resulted in water with high chloride and no corrosion inhibitor flowing through the aging Flint water distribution system. Since Flint returned to its original source of drinking water on October 16, 2015, the State has conducted eleven bi-weekly sampling rounds, resulting in the collection of 4,120 water samples at 819 "sentinel" sites. This chapter describes the first geostatistical analysis of these data and illustrates the multiple challenges associated with modeling the space-time distribution of water lead levels across the city. Issues include sampling bias and the large nugget effect and short range of spatial autocorrelation displayed by the semivariogram. Temporal trends were modeled using linear regression with service line material, house age, poverty level, and their interaction with census tracts as independent variables. Residuals were then interpolated using kriging with three types of non-separable space-time covariance models. Cross-validation demonstrated the limited benefit of accounting for secondary information in trend models and the poor quality of predictions at unsampled sites caused by substantial fluctuations over a few hundred meters. The main benefit is to fill gaps in sampled time series for which the generalized product-sum and sum-metric models outperformed the metric model that ignores the greater variation across space relative to time (zonal anisotropy). Future research should incorporate the large database assembled through voluntary sampling as close to 20,000 data, albeit collected under non-uniform conditions, are available at a much greater sampling density.

P. Goovaerts (✉)
BioMedware, Inc, 11487 Highland Hills Drive, Jerome, MI 49249, USA
e-mail: goovaerts@biomedware.com

B. S. Daya Sagar et al. (eds.), *Handbook of Mathematical Geosciences*,
https://doi.org/10.1007/978-3-319-78999-6_14

14.1 Introduction

The drinking water contamination crisis in Flint, Michigan has attracted national attention since extreme levels of lead were recorded in local water supplies and the percentage of children with elevated blood lead levels (BLL) increased in neighborhoods with the highest water lead levels (WLL). Problems started when the City of Flint, Michigan adopted the cost-saving decision of drawing and treating water from the Flint River instead of relying on the Detroit Water and Sewerage Department's system (DWSD) for its public water supply. A few months later, in December 2014, water samples showed elevated levels of trihaloethanes (THMs) a disinfection byproduct of chlorine, as well as high levels of lead and copper. A public health emergency was declared and residents were told to avoid drinking the water until it was tested or approved water filters were installed. In July 2015, public concerns were raised that lead and copper were being leached from corrosion (chlorine-induced) in the underground lead service lines and home plumbing fixtures as a result of not using corrosion control treatment (CCT). In August and September 2015, 16.6% of the 271 water samples collected by a Virginia Tech's team were found to exceed the EPA action level of 15 μg/L (ATSDR 2010). In September and October 2015, elevated childhood blood lead levels were confirmed and an emergency response was initiated (Hanna-Attisha et al. 2016), leading the city to switch back to the DWSD water supply on October 16, 2015.

Starting in February 2016, 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 for this smaller network including fewer than 200 sites. Overall these 11 sampling rounds resulted in the collection of 4,120 data at 819 different sites over a 40-week time period. This State-controlled monitoring program was supplemented by a voluntary or homeowner-driven sampling whereby concerned citizens received a testing kit and conducted sampling on their own (Goovaerts 2017a, b). Despite the larger size of this database (18,760 samples collected over 53 weeks at 10,341 sites), its heterogeneity and lack of systematic sampling across time prohibited its use in the present space-time analysis.

Except for a few graphs and location maps, the database assembled by the City of Flint and made available online has not undergone any rigorous statistical treatment by State employees and only a few studies have been published so far. 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 did not include a temporal component. A time trend analysis was conducted by Goovaerts (2017a) who used joinpoint regression to model time series of lead levels collected by the state-controlled 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. A space-time analysis of these data should however provide important information to identify residences where high levels of lead are expected. It would also support any assessment of past and current lead exposures among the population at risk, particularly pregnant women and children.

Geostatistical techniques have been routinely used to analyze and map the spatial variability of soil and sediment lead concentrations (Goovaerts et al. 1997; Cattle et al. 2002; Solt et al. 2015), yet their application to lead in drinking water is far less common and mainly concerns groundwater quality (Siddique et al. 2012). A recent study (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. To the author's knowledge, the present study is however the first application of geostatistics to lead in drinking water within a distribution system.

This chapter describes a new methodology to predict lead level in tap water, accounting for WLL measurements collected in neighboring houses, housing characteristics (e.g., age of the house or presence of lead pipes), and temporal trends (e.g., decline since return to pre-crisis source of drinking water). Linear regression was used to model temporal trends at sentinel sites, accounting for the composition of service line (SL), construction year, poverty level, and census tracts as covariates. Cross-validation analysis allowed one to assess the benefit of this approach and compare the results obtained using three different types of space-time covariance models. Both the cases of predicting unsampled times at monitored locations (i.e., filling gaps in time series) and making predictions at unsampled locations were investigated.

14.2 Materials and Methods

14.2.1 Datasets

4,150 WLL measurements recorded over the period 2/20/2016-11/20/2016 were downloaded from http://www.michigan.gov/flintwater (residential testing results).

Table 14.1 Datasets available for the space-time analysis: 4,120 water lead levels measured over 11 sampling rounds. Statistics include the number of data available, the sampling period, the percentage of WLL above 15 μg/L, the mean of logtransformed concentrations, and the composition of service line that was recorded for each sentinel site (three main categories besides plastic, unknown, and other)

Sampling round	Data (n)	Sampling period	%WLL > 15 μg/L	Mean Log_{10} (μg/L)	Composition of SL		
					Lead	Galvanized	Copper
Round S1	610	2/16/2016–2/29/2016	9.51	0.487	5.90	20.66	68.20
Round S2	606	2/24/2016–3/13/2016	8.42	0.465	8.91	19.97	67.00
Round S3	654	3/15/2016–3/24/2016	8.26	0.480	11.62	19.57	63.91
Round S4	644	3/29/2016–4/5/2016	7.14	0.457	13.66	17.39	64.29
Round S5	622	4/13/2016–4/15/2016	6.43	0.427	14.31	15.27	65.43
Round X1	170	5/23/2016–6/7/2016	7.06	0.604	45.88	9.41	44.71
Round X2	178	6/14/2016–6/30/2016	8.99	0.638	49.44	7.87	42.70
Round X3	167	7/19/2016–7/22/2016	6.59	0.557	46.11	8.38	45.51
Round X4	162	8/18/2016–8/22/2016	9.88	0.579	45.06	9.26	45.68
Round X5	158	9/19/2016–9/27/2016	6.33	0.522	45.57	9.49	44.94
Round X6	149	11/17/2016–11/23/2016	6.71	0.532	45.64	9.40	44.97

Data were then allocated to an individual tax parcel unit on the basis of their postal address. Data with incomplete address (two samples) or duplicates (e.g., samples taken from two different faucets on the same day in the same house) were discarded, leading to a total of 4,120 samples collected at 819 different sites; see Table 14.1. Because of their strongly positively skewed distribution (concentrations range from 0 to 5,986 μg/L) and large proportion of zero values (34.6%), data were transformed using the following formula $Log_{10}(z+1)$.

Sentinel sites were initially selected from a pool of 1,951 volunteer sites identified during door-to-door water distribution; in particular it included all 156 sites with lead or lead combination service lines according to City records. Other sites were added according to several criteria: (i) spatial distribution to ensure coverage of all nine City wards, (ii) measurements of high blood levels (Hanna-Attisha et al. 2016), and (iii) environmental justice considerations (e.g. presence of houses with lead-based paint, minority population, and lower socio-economic households). This

Table 14.2 Statistics computed for time series of different lengths: number of sentinel sites, percentage of WLL above 15 µg/L, the mean of logtransformed concentrations, and the composition of service line

Length	Sites	%WLL > 15 µg/L	Mean Log_{10} (µg/L)	Composition of SL		
				Lead	Galvanized	Copper
1	80	7.50	0.413	6.25	22.50	66.25
2	33	6.06	0.475	12.12	18.18	63.64
3	36	6.48	0.433	22.22	25.00	46.30
4	95	4.74	0.411	5.26	18.95	70.53
5	409	3.52	0.358	2.93	17.85	73.59
6	41	8.54	0.530	21.95	4.88	73.17
7	19	9.77	0.651	89.47	5.26	5.26
8	10	11.25	0.705	68.75	7.50	23.75
9	23	11.59	0.693	84.54	4.35	11.11
10	32	18.75	0.750	38.75	15.63	45.63
11	41	19.82	0.793	33.92	20.26	45.81

initial set evolved between sampling rounds as some residents stopped participating, while others asked to be included in the network (Goovaerts 2017b), which explains the fluctuation in the number of sampled sites during the first five rounds S1-S5: 607–621 (Table 14.1). Fewer sites (149–178) were then part of the "Extended Sentinel Site Program". Table 14.2 indicates that only 41 sites were sampled in all 11 rounds, while 80% of time series included five observations or less.

Each house selected to be part of the sentinel network was visited by a licensed plumber who 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, and therefore can be a long-term source of lead (Clark et al. 2015). 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.

City records were the only source of service line data available for the majority of 56,039 tax parcels which were not part of the sentinel sampling program. These records are however inaccurate and lead to the over-identification of lead SLs, likely because old records were not updated as these lines were being replaced (Goovaerts 2017c). The same author 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. This information was combined with field inspection data

and city records to predict by indicator kriging the likelihood that a home has lead or galvanized SL (Goovaerts 2017c).

Besides service lines, lead in drinking water mainly comes from lead-based solder and lead-containing plumbing fixtures (Lee et al. 1989; Cartier et al. 2011). Plumbing material is usually related to the installation year of a plumbing system, which can be approximated by the year of construction. For example, most faucets purchased prior to 1997 were made of brass or chrome-plated brass containing up to 8 percent lead (Rabin 2008). Construction year was retrieved from the 2016 Parcels GIS layer. The attribute "Year_built" was missing for 20,372 parcels and was estimated by ordinary kriging (Goovaerts 1997) with a mean absolute error of prediction of 6.43 years. Based on its relationship to water lead levels (Goovaerts 2017a), construction year was discretized into three classes: pre-1940, 1940–1959, and post-1959.

Poor workmanship as well as lack of regular maintenance can also lead to more corrosion and leaching, and the presence of lead particulates, such as disintegrating brass or detaching pieces of old solder (Wang et al. 2014). Socio-economic status was here assessed using 2015 ACS (American Community Survey) 5-year estimates of the percentage of the block group population living in households where the income is less than or equal to twice the federal "poverty level".

There are many other variables known to influence lead in drinking water. For example, longer water age (i.e., water travel time between the treatment plant and home plumbing system) can decrease the effectiveness of corrosion control; increasing leaching and water lead levels (US EPA 2002; Wang et al. 2014). This information was however unavailable for this study.

14.2.2 Space-Time Kriging and Covariance Models

Let $z(\mathbf{u}_\alpha;t)$ denote the water lead level recorded on time t at sentinel site α georeferenced by the geographical coordinates $\mathbf{u}_\alpha = (x_\alpha, y_\alpha)$ of the corresponding tax parcel centroid. Prediction of z-value at unsampled time t_0 and location u_0 was conducted using the following kriging estimator:

$$Z^*(u_0;t_0) = \sum_{t=t_0-\Delta t}^{t_0+\Delta t} \sum_{\alpha=1}^{n(t)} \lambda_{\alpha t} \times z(u_\alpha;t) \tag{14.1}$$

n(t) is the number of observations recorded at time t, within the time window $2\Delta t$, that were retained for estimation. The weights $\lambda_{\alpha t}$ are solution of the following space-time (ST) kriging system:

$$\begin{aligned}
&\sum_{t=t_0-\Delta t}^{t_0+\Delta t}\sum_{\alpha=1}^{n(t)}\lambda_{\alpha t}C\left(u_\alpha-u_\beta;t-t'\right)+\mu=C\left(u_0-u_\beta;t_0-t'\right) \quad \beta=1,\cdots,n\left(t'\right)\\
&\sum_{t=t_0-\Delta t}^{t_0+\Delta t}\sum_{\alpha=1}^{n(t)}\lambda_{\alpha t}=1 \qquad\qquad t'=t_0-\Delta t,\cdots,t_0+\Delta t
\end{aligned} \tag{14.2}$$

The parameter μ is a Lagrange multiplier accounting for the constraint on the weights. The term $C\left(u_\alpha-u_\beta;t-t'\right)$ is the ST covariance between any two observations recorded at locations u_α and u_β at times t and t', respectively. Euclidian distances were used here since most lead in drinking water comes from premise plumbing materials and service lines instead of being transported through water mains (Del Toral et al. 2013; EET Inc. 2015).

One challenge associated with the application of ST kriging is the choice of a ST covariance model within the ever growing class of models (Montero et al. 2015). The following three non-separable ST covariance models were compared in the present study:

- The generalized product-sum model (De Iaco et al. 2002):

$$C(h,\tau)=k_1C_s(h)+k_2C_t(\tau)+k_3C_s(h)C_t(\tau) \tag{14.3}$$

  where k_1, k_2, and k_3 are non-negative (strictly positive for k_3) coefficients estimated from the sills of the spatial, temporal, and spatio-temporal semivariograms (De Cesare et al. 2002).
- The metric model (Dimitrakopoulos and Luo 1994):

$$C(h,\tau)=C_{st}\left(\sqrt{\left(\frac{h}{a_s}\right)^2+\left(\frac{\tau}{a_t}\right)^2}\right) \tag{14.4}$$

  where a normalized space-time distance measure is created by rescaling the spatial and temporal lags, h and τ, by the ranges of the spatial and temporal semivariograms, a_s and a_t (case of geometric anisotropy).
- The sum-metric model (Heuvelink and Griffith 2010):

$$C(h,\tau)=C_s(h)+C_t(\tau)+C_{st}\left(\sqrt{\left(\frac{h}{a_s}\right)^2+\left(\frac{\tau}{a_t}\right)^2}\right) \tag{14.5}$$

  This model combines characteristics of the two previous models: (i) sum of spatial and temporal covariances allowing for the presence of zonal anisotropies (i.e., semivariogram sills are not the same in all directions), and (ii) a metric ST model for the residual variability (geometric anisotropy).

Two other classes of non-separable ST covariance models, Cressie-Huang model (Cressie and Huang 1999) and Gneiting models (Gneiting 2002), were not considered because: (1) the fitting of these models needs a complex iterative parameter optimization technique (De Iaco 2010), whereas the three selected models can be fitted using straightforward techniques similar to those already used for spatial-only and temporal-only semivariograms, and (2) recent studies (Guo et al. 2015) indicate that these two more complex models provide similar fits to experimental ST semivariograms and comparable prediction accuracy as the product-sum model, confirming previous findings (De Iaco 2010).

The main difficulty in the practical implementation of the product-sum and sum-metric models is the inference of the sill of the ST semivariogram model, $C_{st}(0)$, which is most often estimated visually from the 3D plot of the experimental ST semivariogram $\hat{\gamma}_{st}(h,\tau)$ (e.g., De Cesare et al. 2002; Heuvelink and Griffith 2010). In order to make the fitting procedure more user-friendly, the space-time sill $C_{st}(0)$ was here computed as the following weighted average of experimental space-time semivariogram values:

$$C_{st}(0) = \frac{1}{\sum_h \sum_\tau w_{h,\tau}} \sum_h \sum_\tau w_{h,\tau} \hat{\gamma}_{st}(h,\tau) \quad \text{if } \hat{\gamma}_{st}(h,\tau) \geq g_c \tag{14.6}$$

where the weight $w_{h,\tau}$ is the number of data pairs falling into the class of spatial and temporal lags (h,τ). Only the classes where the ST semivariogram values exceed a critical sill g_c, defined as the maximum of the spatial and temporal sills, were used.

14.2.3 Accounting for Secondary Information

Lead service lines are widely considered the main source of lead in drinking water (Lee et al. 1989; Clark et al. 2015). Another culprit is lead fixtures and pipes present within old houses (premises plumbing), and poverty can compound the problems through the lack of maintenance. Goovaerts (2017a) also found that temporal trends can vary greatly across the city. This secondary information was here incorporated in the definition of a stochastic trend model $M(u;t)$, leading to the following decomposition of the space-time random function (RF) (Kyriakidis and Journel 1999):

$$Z(u;t) = M(u;t) + R(u;t) \tag{14.7}$$

where $M(u;t)$ is a nonstationary spatiotemporal RF modeling the space-time distribution of the mean process, with $E[M(u;t)] = m(u;t)$ and $R(u;t)$ is a zero mean stationary spatiotemporal RF modeling space-time fluctuations around $M(u;t)$.

The trend component at each sentinel site $\mathbf{u}_\alpha$ was fitted using a linear model including six fixed factors: presence/absence of LSL, presence/absence of galvanized service line (GSL), time since first sample was collected (TIME), poverty

level (POV), house age (AGE), and census tract (CT). The model takes the following form:

$$\begin{aligned} M(u;t) = LSL(u) \times TIME + CT(u) \times TIME + LSL(u) \times CT(u) \\ + GSL(u) \times CT(u) + AGE(u) \times CT(u) \\ + POV(u) \times CT(u) \end{aligned} \tag{14.8}$$

This model naturally handles uneven spacing of repeated measurements within each time series, as well as their correlation which was modeled using a spherical variance-covariance structure. Once the trend model was fitted, regression residuals were interpolated using space-time simple kriging and the ST covariance models introduced in Sect. 14.2.2.

14.2.4 Cross-Validation

The accuracy of the predictive models created by the different approaches (e.g., three types of ST covariance models, univariate vs incorporation of secondary information) was assessed by cross-validation whereby each observation or time series (i.e., all data collected at the same site) was removed at a time and re-estimated using data collected at neighboring sentinel sites. The following performance criteria were then computed from n kriging estimates:

- the mean error (ME) of prediction as:

$$ME = \frac{1}{n} \sum_{t=1}^{T} \sum_{\alpha=1}^{n(t)} \left(z^*(u_a;t) - z(u_a;t) \right) \tag{14.9}$$

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

$$MAE = \frac{1}{n} \sum_{t=1}^{T} \sum_{\alpha=1}^{n(t)} \left| z^*(u_a;t) - z(u_a;t) \right| \tag{14.10}$$

- the mean square standardized residual (MSSR) as:

$$MSSR = \frac{1}{n} \sum_{t=1}^{T} \sum_{\alpha=1}^{n(t)} \frac{\left(z^*(u_a;t) - z(u_a;t) \right)^2}{\sigma_K^2(u_a;t)} \tag{14.11}$$

where $\sigma_K^2(u_a;t)$ is the kriging 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).

One application of the predictive models is to prioritize any further sampling or intervention by ranking 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 μg/L was assessed using Receiver Operating Characteristics (ROC) curves which plot the probability of false positive versus the probability of detection (Swets 1988; Fawcett 2006; Goovaerts et al. 2016). 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 (e.g., $z_c \geq 15$ μg/L) higher than a randomly chosen negative instance (e.g., $z_c < 15$ μg/L).

14.3 Results and Discussion

14.3.1 Spatial Distribution

Figure 14.1a shows the location of all 819 sentinel sites within the nine wards in the city of Flint. Site-specific statistics such as number of observations and average log concentrations recorded for each time series, as well as composition of service line (GSL vs. LSL), were aggregated at the census tract level for better visualization. Geographical clusters of sentinel sites can be distinguished in several census tracts (e.g. border of wards 2 and 6, wards 7 and 9) which tend to be the tracts with the largest WLLs (Fig. 14.1c) and percentages of sampled LSLs (Fig. 14.1d). There is also a clear spatial trend with fewer lead service lines (e.g., none in Ward 1) and shorter time series (Fig. 14.1b) sampled in the Northern part of the city. Ward 5 includes the oldest neighborhood where GSLs are prevalent (Fig. 14.1e), while LSLs appear as small clusters, in particular in wards 6, 7 and 9 (Goovaerts 2017c).

14.3.2 Temporal Trend Modeling

Temporal trends for the three major types of service line were visualized by aggregating observations within non-overlapping 14-day windows, which corresponds to the average time interval between sampling rounds during the first phase (Round S) of the sentinel monitoring program (Table 14.1). Except for LSLs water lead levels do not appear to have declined over the 40-week sampling period; actually they seem to have slightly increased for GSLs (Fig. 14.2a). These results are however a direct artifact of the sampling strategy whereby 80% of sentinel sites

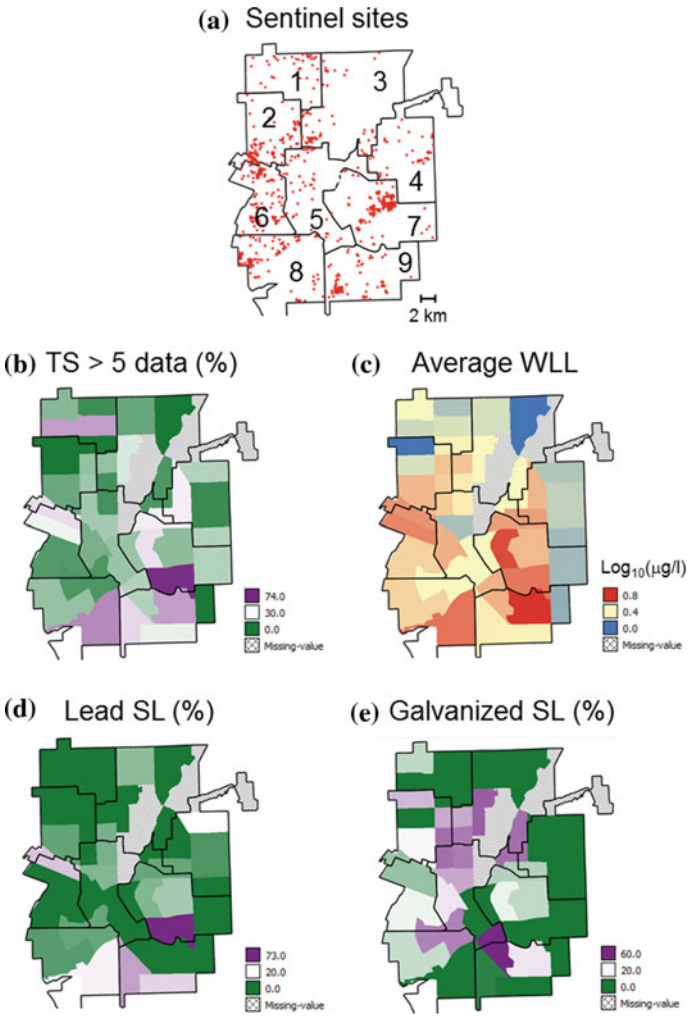


Fig. 14.1 **a** Location of sentinel sites in each of the nine wards, and several census tract-level statistics: **b** percentages of time series (TS) including more than five observations, **c** average water lead levels, **d** percentage of sites with lead service lines, **e** percentage of sites with galvanized service lines. Shaded polygons indicate census tracts that do not include any sentinel site (missing values)

were not sampled beyond week 16, while sampling continued at sites where the risk of exceeding the EPA action level of 15 µg/L was the greatest (Table 14.2).

After elimination of all sites where fewer than six observations were collected, the averaged time series display the expected decline (Fig. 14.2b). The impact is minimal for LSLs since most of these sites are considered at risk and were sampled during both the initial and extended sentinel sampling programs (Rounds S and X).

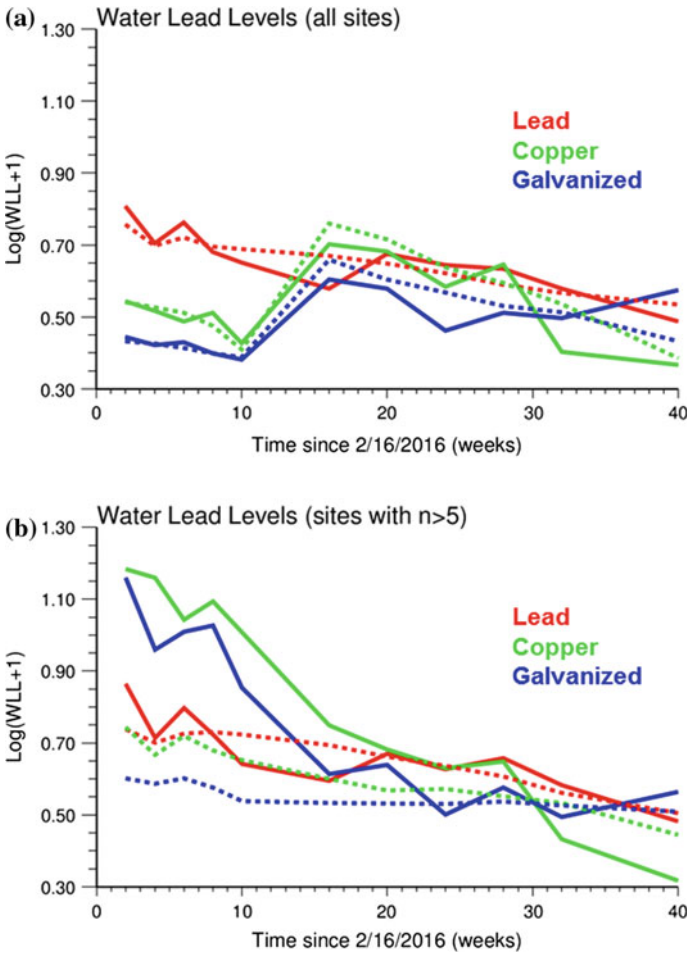


Fig. 14.2 Time series of observed (solid line) and predicted by regression (dashed line) water lead levels computed on average for the three major types of service line: lead, galvanized, and copper. Results (log transformed concentrations) are calculated from: **a** all sites, and **b** subset of sites where at least six observations were recorded

The selection bias is stronger for copper and galvanized lines, which explains the larger water lead levels recorded during the first 16 weeks relative to LSLs.

This sampling bias complicated greatly the modeling of temporal trends by regression. Indeed using all the data would underestimate the weekly rate of decline of water lead levels, whereas subsetting the dataset (e.g., using only time series including more than five data points as in Fig. 14.2b) will result in overestimating the concentrations at a majority of sites. In addition, the time series length cannot be used as covariate in the model to allow its application at unmonitored locations. Two modeling strategies were considered in this chapter. First, because of its

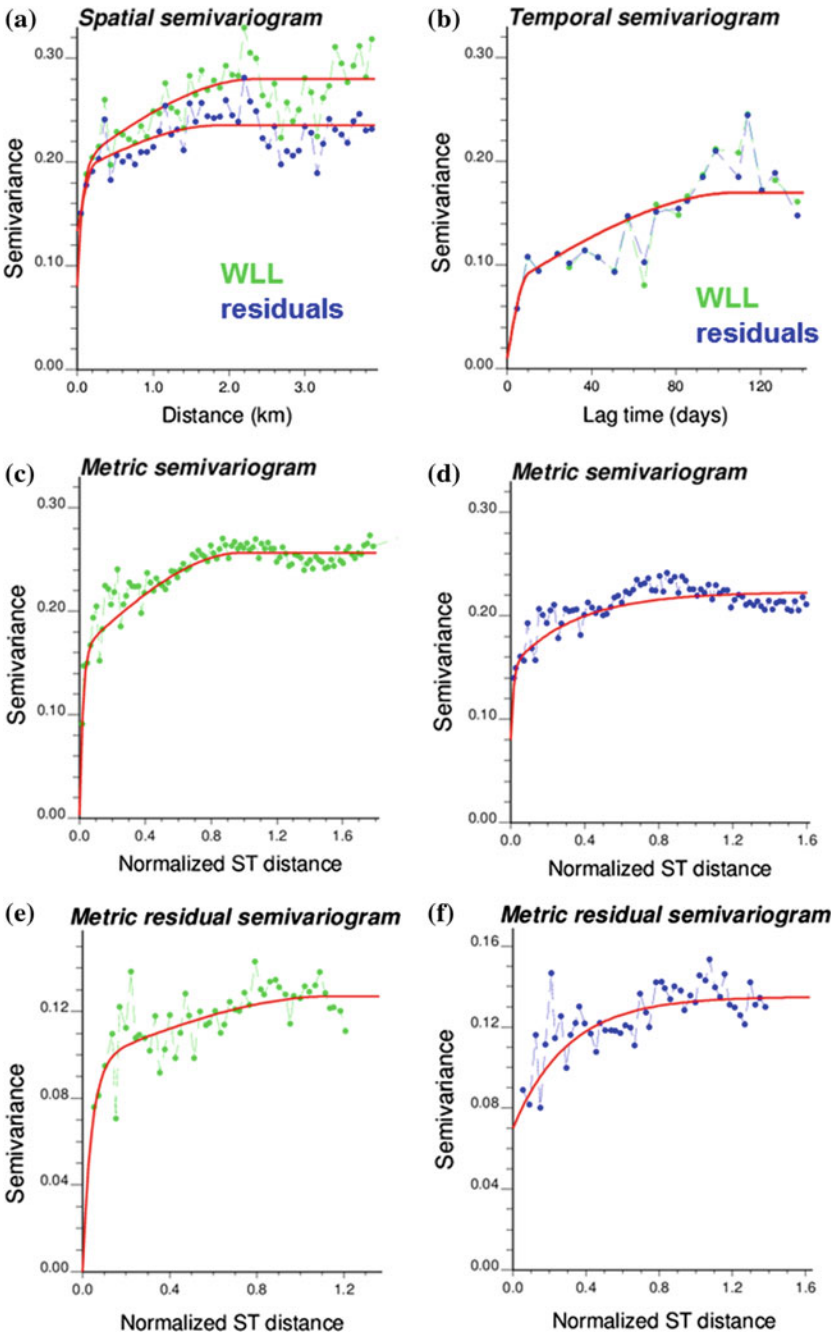
relationship with time series length (Fig. 14.1) census tract was used as covariate in the regression model (Eq. 14.8). The second more complicated approach was to allow the intercept to fluctuate among sentinel sites, even when located within the same tract; i.e., use a mixed model where the intercept is modeled as a random effect. The trade-off cost for this added flexibility was the need to estimate the intercept at unmonitored locations, which was accomplished using ordinary kriging. Despite providing a better fit than the first alternative, the mixed model did not lead to more accurate kriging estimates, hence only the first option is discussed hereafter.

All six interaction terms in the trend model (Eq. 14.8) were highly significant ($\alpha = 0.01$). The correlation between predicted and observed WLL is however rather weak ($r = 0.47$), which illustrates the challenge of predicting spatial and temporal variations in lead for drinking water (Bailey and Russell 1981; Del Toral et al. 2013). While the output of the regression model provides a reasonable fit to the SL-specific time series computed using all the data (Fig. 14.2a), it underestimates water lead levels for LSL and GSL when using only time series including more than five data points (Fig. 14.2b).

14.3.3 Variography

Semivariograms helped quantifying the scale and magnitude of the space-time variability displayed by the maps and time series of Figs. 14.1 and 14.2. The spatial semivariogram (Fig. 14.3a) shows three nested scales of spatial variability: (1) a long range (2.35 km) caused by the neighborhood effect since houses in the same neighborhood tend to be built at the same time (i.e., similar plumbing system) and have similar water age, (2) a short range (200 m) corresponding to variability between adjacent houses, and (3) a nugget effect or discontinuity at the origin which represents the variability among samples taken within the same tax parcel (i.e. different apartments and/or measurement error for samples taken within the same residence). The substantial short-range variability (71% of total sill) likely reflects the heterogeneity in housing conditions (e.g., renovated houses) 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). This interpretation is confirmed by the similar short-range variability displayed by the semivariogram of regression residuals (Fig. 14.3a, lower blue curve) since the regression model (Eq. 14.8) does not account for sampling characteristics. It is noteworthy that the longer range of 2.35 km is still fairly small relative to the size of the city (see legend of Fig. 14.1a), while the average separation distance between each sentinel site and the closest neighbor (293 m) exceeds the shortest range (200 m) that encapsulates 71% of the total spatial variability.

The temporal semivariogram (Fig. 14.3b) also displays three nested scales of variability although the longer range structure (110 days) represents here 53% of the total variability. Another difference with the spatial case is the overlap of


(a)
Spatial semivariogram
Semivariance
0.30
0.20
0.10
0.00
WLL
residuals
0.0
1.0
2.0
3.0
Distance (km)
(b)
Temporal semivariogram
Semivariance
0.30
0.20
0.10
0.00
WLL
residuals
0
40
80
120
Lag time (days)
(c)
Metric semivariogram
Semivariance
0.30
0.20
0.10
0.00
0.0
0.4
0.8
1.2
1.6
Normalized ST distance
(d)
Metric semivariogram
Semivariance
0.30
0.20
0.10
0.00
0.0
0.4
0.8
1.2
1.6
Normalized ST distance
(e)
Metric residual semivariogram
Semivariance
0.12
0.08
0.04
0.00
0.0
0.4
0.8
1.2
Normalized ST distance
(f)
Metric residual semivariogram
Semivariance
0.16
0.12
0.08
0.04
0.00
0.0
0.4
0.8
1.2
1.6
Normalized ST distance

◀**Fig. 14.3** Experimental semivariograms with the model fitted that were used to form the three types of ST covariance models (Eqs. 14.3–14.5) **a** spatial semivariogram (lower curve is for residuals), **b** temporal semivariogram, **c** metric semivariogram for WLLs, **d** metric semivariogram for regression residuals, **e** metric residual semivariogram (sum-metric model) for WLLs, **f** metric residual semivariogram for regression residuals

temporal semivariograms for WLLs and regression residuals, illustrating the inability of the trend model (Eq. 14.8) to capture purely temporal changes. This result is in agreement with the small magnitude of changes displayed by the time series of predicted values in Fig. 14.2 (dashed line). Comparison of the total sills of spatial and temporal semivariograms (Fig. 14.3a–b) indicates that the variability observed across space is greater than the temporal variability. Such zonal anisotropy is in conflict with the assumption underlying the metric ST covariance model (Eq. 14.4).

Figure 14.3c–d show the semivariograms computed using a normalized space-time distance (metric model). Because the spatial and temporal lags were rescaled using different constants for the WLL and residual semivariograms, these two curves are plotted separately. The vertical axis is however comparable and illustrates the smaller variability of residuals (i.e., lower sill for the semivariogram of Fig. 14.3d). Once again, both semivariograms display substantial short-range variability. The last two semivariograms (Fig. 14.3e–f) represent the metric space-time model that captures the residual variability in the sum-metric model (Eq. 14.5).

14.3.4 Cross-Validation Analysis

The semivariogram models of Fig. 14.3 were used to conduct a cross-validation analysis whereby one observation (LOO approach) or one time series (LTO approach) was removed at a time and re-estimated using data collected at neighboring sentinel sites. Based on a sensitivity analysis using ST ordinary kriging and MAE criterion, 48 observations with a maximum of three data points per site were retained for the estimation by univariate and residual ST kriging. Results obtained for predictions by the time trend model were also included as reference in Table 14.3.

The first three rows in Table 14.3 indicate that all algorithms give unbiased predictions (ME close to zero). As expected, the best prediction scores (i.e., lower MAE and higher AUC) are obtained when using data from the same time series (LOO approach) instead of relying solely on non-colocated data (LTO approach). Except for MSSR the product-sum model performs best, with the sum-metric model being a close second. The metric model underperforms the other two models because the combination of both spatial and temporal dimensions through a normalized space-time distance leads one to underestimate the correlation among observations of the same time series. In other words, the assumption underlying the

Table 14.3 Results of cross-validation analysis conducted by leaving one observation out (LOO) or one time series out (LTO) at a time. The four performance criteria described in Sect. 14.2.4 were computed for three types of space-time covariance models (generalized product-sum, metric, and sum-metric) and three space-time interpolation algorithms (ST ordinary kriging, trend model fitted by linear regression with and without interpolation by ST residual kriging)

Algorithm	Performance criteria					
	Product-sum model		Metric model		Sum-metric model	
	LOO	LTO	LOO	LTO	LOO	LTO
Mean error of prediction (ME)						
ST ordinary kriging	−0.001	0.009	0.003	0.007	−0.001	0.008
ST residual kriging	0.0	0.008	0.003	0.005	0.001	0.008
Trend model[a]	0.0					
Mean absolute error of prediction (MAE)						
ST ordinary kriging	0.257	0.375	0.336	0.384	0.263	0.378
ST residual kriging	0.251	0.337	0.318	0.346	0.254	0.343
Trend model[a]	0.331					
Mean square standardized residual (MSSR)						
ST ordinary kriging	1.326	0.954	1.026	1.208	1.190	1.111
ST residual kriging	1.119	0.912	0.957	1.086	1.015	1.086
Trend model[a]	74.9					
Area under the ROC curve for 15 μg/L (AUC)						
ST ordinary kriging	0.832	0.615	0.743	0.598	0.829	0.613
ST residual kriging	0.839	0.707	0.768	0.692	0.836	0.697
Trend model[a]	0.713					

[a]value for trend model is the same for all six combinations

metric model is incompatible with the zonal anisotropy detected on Fig. 14.3. Accounting for secondary information through residual kriging slightly improves the prediction relative to ST ordinary kriging; both kriging algorithms outperformed the trend model.

These results however apply only to the narrow situation where exposure to lead in drinking water is reconstructed at the sole sentinel sites. For prediction at sites where no data was collected, LTO results indicate that differences between ST covariance models are much smaller as purely temporal correlations are not used in the kriging system. Nevertheless, the product-sum model still performs best. The LTO approach also emphasizes the benefit of using trend models that account for secondary information (i.e., larger differences between residual kriging and ordinary kriging). Yet, prediction performances actually deteriorate when kriged residuals are added to the trend model: the sole trend model gives better prediction than residual kriging. It is however noteworthy that the trend model was not cross-validated, hence the observation being predicted was used to create the model.

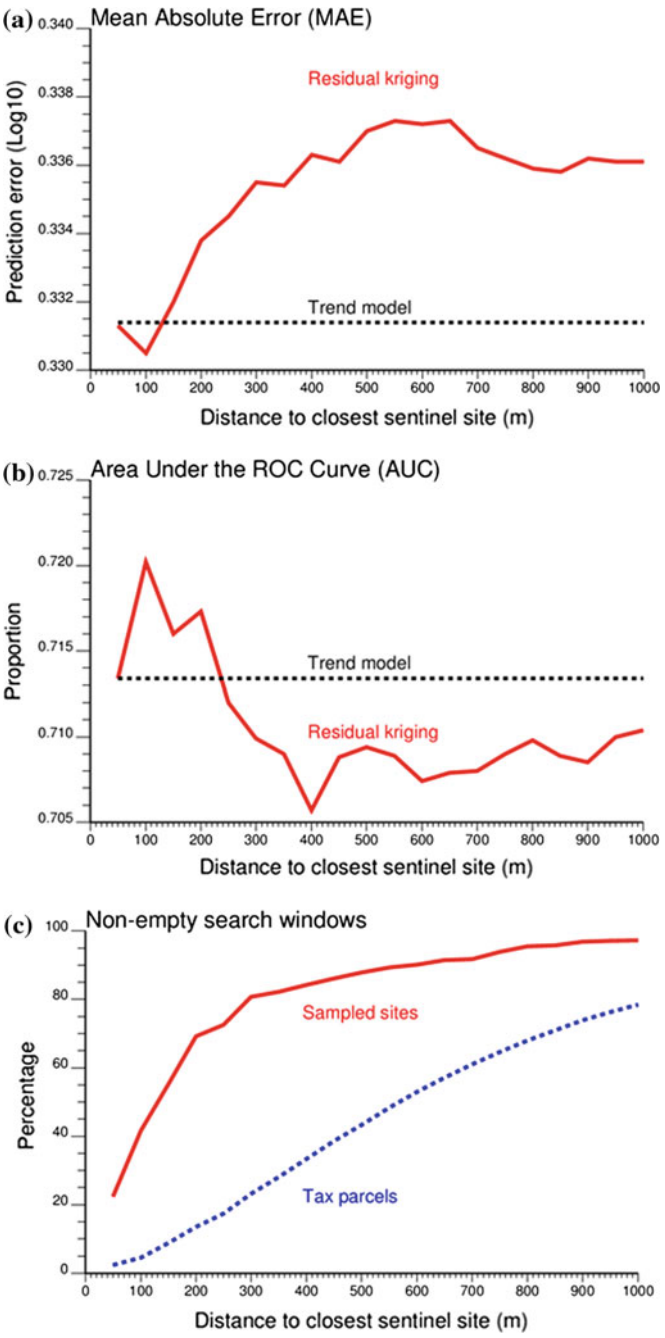


Fig. 14.4 Impact of the size of kriging search window on several statistics computed by the leave one time series out (LTO) approach: **a** mean absolute error of prediction, and **b** area under the ROC curve. Horizontal dashed lines represent the values obtained for the time trend model created by linear regression. **c** percentages of search windows that include at least one observation when centered on sampled sentinel sites or tax parcels

Because of the substantial short-scale spatial variability retaining increasingly distant data is expected to add more and more noise to the kriging estimate. This was investigated by changing the search strategy and selecting only sentinel sites located within a given distance of the site being predicted. If no data was located within the search radius, the kriged residual was zero and the residual kriging estimate was simply the value of the trend model. Figure 14.4 shows results of this sensitivity analysis conducted for the product-sum model over distances ranging from 50 m to 1 km. For the mean error of prediction the little benefit of residual kriging vanishes as soon as data beyond 100 m are used in the estimation (Fig. 14.4a), while this distance is 200 m for the area under the ROC curve (Fig. 14.4b). Figure 14.4c indicates that 42% of sentinel sites have another sentinel site within 100 m, while this percentage is only 4.6% for tax parcels (Fig. 14.4c). In other words, there is little benefit in applying geostatistics to model the space-time distribution of WLL over the 56,039 tax parcels in Flint using the data collected at sentinel sites.

14.4 Conclusions

This chapter presented the first application of space-time geostatistics to lead levels recorded in drinking water of a public distribution system. The methodology was illustrated using 4,120 water samples that were collected at 819 "sentinel" sites over a 40-week period in the city of Flint. Despite a sizable database assembled by the State of Michigan, the geostatistical analysis was hampered by a temporal sampling bias and the existence of substantial variability over a few hundred meters. Unlike other countries such as Canada or France, sampling is not conducted by a trained technician in the US. Instead, homeowners are expected to collect water samples after a minimum of 6 h. of stagnation (e.g., overnight stagnation) following specific instructions (US EPA 2016), which can cause substantial variability among households. 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.

In the present case-study, space-time kriging proved beneficial only in the situation where observations had been collected at the site being predicted; i.e., to fill the gaps in time series. The generalized product-sum and sum-metric space-time covariance models then outperformed the metric model that ignores the greater variation across space relative to time (zonal anisotropy). Sentinel sites represent however only 1.5% of tax parcels in the city of Flint. At unsampled sites the kriging prediction was no better than the temporal trend estimated by linear regression and it turned out to become less accurate if no data was collected within 100 meters. Although the regression model included site-specific characteristics, such as construction year and composition of service lines, it was unable to explain the

short-range variability, leaving 78% of the total variance unaccounted for ($R^2 = 22\%$).

In the future, several approaches will be investigated to tackle the impact of short-range variability on prediction. First, the data analyzed in this chapter represent less than 20% of the water samples available for the city of Flint. The majority of samples were collected by voluntary sampling whereby concerned citizens received a testing kit and conducted sampling on their own (Goovaerts 2017a, b). Despite the lack of periodic sampling in time and existence of temporal bias (e.g., houses with low lead levels were less likely to be tested again) the greater spatial coverage (i.e., more than 18% of tax parcels sampled) will reduce substantially the average distance between a tax parcel and the closest observation. However, spatial heterogeneity will likely still be present over short distances, leading one to question our ability to make prediction 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. As more US cities are facing similar drinking water crisis, reliable techniques for sampling and modeling spatial and temporal changes in water lead levels will be sorely needed.

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:1610.00580. https://arxiv.org/abs/1610.00580. Accessed 26 May 2017

Agency for Toxic Substances and Disease Registry (2010) Case studies in environmental medicine (CSEM) lead toxicity. Course: WB 1105 Original Date August 15. US Department of Health and Human Services, Public Health Service: Atlanta

Bailey RJ, Russell PF (1981) Predicting drinking water lead levels. Environ Technol Lett 2:57–66. https://doi.org/10.1080/09593338109384023

Cartier C, Laroche L, Deshommes E et al (2011) Investigating dissolved lead at the tap using various sampling protocols. J Am Water Works Assoc 103:55–67

Cattle JA, McBratney AB, Minasny B (2002) Kriging method evaluation for assessing the spatial distribution of urban soil lead contamination. J Environ Qual 31:1576–1588

Clark BN, Masters SV, Edwards MA (2015) Lead release to drinking water from galvanized steel pipe coatings. Environ Eng Sci 32:713–721. https://doi.org/10.1089/ees.2015.0073

Cressie N, Huang H-C (1999) Classes of nonseparable, spatio-temporal stationary covariance functions. J Am Stat Assoc 94:1330–1340

De Cesare L, Myers DE, Posa D (2002) FORTRAN programs for space–time modeling. Comput Geosc 28:205–212

De Iaco S (2010) Space–time correlation analysis: a comparative study. J Appl Stat 37:1027–1041

De Iaco S, Myers DE, Posa D (2002) Nonseparable space-time covariance models: some parametric families. Math Geol 34:23–42
Del Toral MA, Porter A, Schock MR (2013) Detection and evaluation of elevated lead release from service lines: a field study. Environ Sci Technol 47:9300–9307
Dimitrakopoulos R, Luo X (1994) Spatiotemporal modeling: covariances and ordinary kriging systems. In: Dimitrakopoulos R (ed) Geostatistics for the next century. Kluwer, Dordrecht, pp 88–93
EET Inc (2015) Evaluation of lead sampling strategies. WRF, Final grant report 4569. http://www.waterrf.org/PublicReportLibrary/4569.pdf. Accessed 26 May 2017
Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27:861–874
Flint Safe Drinking Water Task Force Recommendations on MDEQ's Draft Sentinel Site Selection (2016). https://www.epa.gov/sites/production/files/2016-02/documents/task_force_recommendations_on_sentinel_site_selection_2-16.pdf. Accessed 26 May 2017
Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New-York
Goovaerts P (2017a) The drinking water contamination crisis in Flint: modeling temporal trends of lead level since returning to Detroit water system. Sci Total Environ 581–582:66–79
Goovaerts P (2017b) Monitoring the aftermath of Flint drinking water contamination crisis: another case of sampling bias? Sci 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 Total Environ 599–600:1552–1563
Goovaerts P, Webster R, Dubois J-P (1997) Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. Environ Ecol Stat 4:31–48
Goovaerts P, Wobus C, Jones R et al (2016) Geospatial estimation of the impact of deepwater horizon oil spill on plant oiling along the Louisiana shorelines. J Environ Manag 180:264–271
Gneiting T (2002) Nonseparable, stationary covariance functions for space-time data. J Am Stat Assoc 97:590–600
Guo L, Lei L, Zeng Z et al (2015) Evaluation of spatio-temporal variogram models for mapping Xco2 using satellite observations: a case study in China. IEEE J Sel Topics Appl Earth Observ Remote Sens 8:376–385
Hanna-Attisha M, LaChance J, Sadler RC et al (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
Heuvelink GBM, Griffith DA (2010) Space–time geostatistics for geography: a case study of radiation monitoring across parts of Germany. Geogr Anal 42:161–179
Kyriakidis PC, Journel AG (1999) Geostatistical space-time models: a review. Math Geol 31:651–684
Lee RG, William CB, David WC (1989) Lead at the tap: sources and control. J Am Water Works Ass 81:52–62
Montero JM, Fernandez-Aviles G, Mateu J (2015) Spatial and spatio-temporal geostatistical modeling and kriging. Wiley, New York
Rabin R (2008) The lead industry and lead water pipes "A Modest Campaign". Am J Public Health 98:1584–1592
Siddique A, Zaigham NA, Mohiuddin S et al (2012) Risk zone mapping of lead pollution in urban groundwater. J Basic Appl Sci 8:91–96
Solt MJ, Deocampo DM, Norris M (2015) Spatial distribution of lead in Sacramento, California, USA. Int J Environ Res Public Health 12:3174–3187
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
Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240:1285–1293
US Environmental Protection Agency, Office of Ground Water & Drinking Water (2002) Effect of water age on distribution system water quality. https://www.epa.gov/sites/production/

files/2015-09/documents/2007_05_18_disinfection_tcr_whitepaper_tcr_waterdistribution.pdf. Accessed 26 May 2017

US Environmental Protection Agency, Office of Ground Water & Drinking Water (2016) Memorandum: clarification of recommended tap sampling procedures for purposes of the lead and copper rule. https://www.epa.gov/sites/production/files/2016-02/documents/epa_lcr_sampling_memorandum_dated_february_29_2016_508.pdf. Accessed 26 May 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