

# The Geographic Distribution of Soil Lead Concentration: Description and Concerns

*Daniel A. Griffith*

*Abstract: Pollution of inhabited geographic landscapes is often a public health concern. Ascertaining the degree and scope of such pollution is a difficult task and frequently involves analysis of soil samples. This article describes ways to statistically analyze the geographic distribution of soil samples to better understand polluted landscapes, in part to determine whether more samples are needed; special reference is made to lead. The soundness of this description rests on the identically distributed assumption of statistical analysis as well as attributes of soil samples supporting evaluations of constant variance; methods are outlined for assessing both. Empirical findings are summarized regarding the distribution of soil lead concentration in the environment extracted from three types of geographic landscapes: an urban area, superfund sites, and a flood plain. These regions are described in terms of the statistical frequency distribution of soil lead concentration and the nature and degree of spatial autocorrelation latent in the geographic distribution of soil lead concentration. The importance of knowing such results is demonstrated by examining three concerns: 1) why a spatial analysis of soil samples is worth undertaking; 2) the cost-effectiveness of spatial sampling; and 3) the ability to predict soil lead contamination at unsampled locations based on data from sampled locations.*

## Introduction

Lead (Pb; a heavy, soft, malleable, bluish-gray metal) is a ubiquitous element that is found in rocks, soil, plants, animals, and human beings; it naturally occurs in quite low levels. It is also one of the toxic heavy metals that have been geographically concentrated or whose elevated levels have been made ubiquitous in the inhabited environment because of human activities. Three principal sources of this pollution are the widespread use of lead-based paints, lead emissions in gasoline in earlier years, and lead waste from mining/commercial/manufacturing processes. Community health issues associated with this pollution, including childhood lead poisoning, are subjects of research across the nation (e.g., Spake and Couzin 1999). This research increasingly involves geographic information systems, with recent spatial analyses of pediatric lead poisoning appearing for Syracuse, New York (Griffith et al. 1998a) and Jefferson County, Kentucky (Reissman et al. 2001), and for the states of New York (Raucci 1999) and Indiana (McGarigle 2000).

Current policies aimed at reducing lead exposure are based on the assumption that the greatest lead hazard comes from lead-based paints. Mielke (1999) argues that dust is another form of lead pollution that poses a threat to the health of children. Soil, which contains tiny particles of lead, functions as a giant reservoir of lead dust in the inhabited environment. Accordingly, children face their greatest risk for exposure in yards around their houses and, to a lesser extent, in open public spaces in which they play. Mielke further contends that only an accurate and complete appreciation of the distribution of lead in the environment can help shape policies that

more effectively protect the health of children. To this end, this article addresses the following general question:

How can the statistical and geographic distributions of soil lead concentration in the inhabited environment be described?

More specifically, what information do spatial statistics give us regarding georeferenced soil lead concentration measurements, and what are the implications gleaned from empirical analysis of selected landscapes for understanding soil lead concentrations in other landscapes? Answers to these questions are based on summarized findings regarding the distribution of soil lead concentration in various types of geographic landscapes: an urban area (Syracuse, NY), two superfund sites, and a flood plain. A spatial analysis methodology is described and employed to analyze four specific sites, with the goal being to establish expectations about non-geographic and geographic variability when similar environmental conditions prevail at other locations.

## Data Analysis Requirements

The desired data descriptions require assessment of soil lead concentrations in terms of their landscape-wide means, their variances, their frequency distributions, and latent levels of spatial autocorrelation in their geographic distributions. To complete these assessments, proper data collection dictates that a number of protocol features are satisfied. To assess soil lead contamination across a given landscape, georeferenced soil samples are needed. To assess the identically distributed assumption of statistical analysis, attributes of soil samples supporting evaluations of constant variance are needed. Additionally, to assess the appropriateness of sample size, it is necessary to establish adequate geo-

graphic coverage of a sampling network. Failure to fulfill these requirements severely compromises data descriptions.

The statistical analytic techniques employed here are based on the normal curve theory. A bell-shaped curve is symmetric and has the data concentrated near the mean, resulting in few unusually high or low values. It is the foundation upon which much constant variance testing, model parameter estimation, and spatial autocorrelation inferences are built. Development of the geographic sample-size concept and the map hole plugger discussed in this article is based on this. In other words, the statistical frequency distribution dictates the degree to which numerical results reported in this article are meaningful.

The methodology outlined here emphasizes each of the aforementioned data collection protocol features. Four landscapes were selected because their soil samples are georeferenced; a fifth site was a superfund site; it was dismissed because most of its soil samples lack locational tags. One assessment of constant geographic variance can be achieved by comparing variation in different regions of a landscape. A convenient regionalization scheme is to partition a landscape into the four quadrants of the plane; another is to make comparisons through the use of attribute features of sample locations. Finally, adequate geographic coverage can be established in two ways: deviation of a given sampling network from a hexagonal grid can be quantified (see Stehman and Overton 1996), and an effective sample size can be computed based on the nature and degree of latent spatial autocorrelation (Griffith and Zhang 1999).

The interested reader may wish to consult Cressie (1991) and Griffith and Layne (1999) for more comprehensive discussions of the geostatistical and spatial autoregressive tools employed for the analysis summarized in this article. More conventional tools for evaluating statistical model assumptions can be reviewed in sections of the *Encyclopedia of Statistical Science*. Additionally, Sen and Srivastava (1990) furnish a readable treatment of Box-Cox power transformations.

## Background: Samples From Selected Geographic Landscapes

Four geographic landscapes are explored in this study for illustrative purposes. Evidence is sought from them to address the question of what expectations can be obtained from these landscapes regarding the variability of soil lead contamination in other places.

### Site 1

Griffith et al. (1998a) analyzed the spatial distribution of pediatric blood-lead levels in Syracuse, NY. That article presented the first choropleth-map generalization of the geographic distribution of lead concentration in surficial soil across the city, much of which was deposited by gasoline emissions. However, the map is incomplete, being constructed using 112 samples covering 50 of the city's 61 census tracts; 139 samples were collected, but only 112 have locations within the city. Geometric means of samples were calculated for each census tract.

Johnson together with Bretsch (1998) subsequently expanded the soil sampling project, augmenting the number of samples by 167 and nearly completing coverage of the city. The measures used here were obtained with a 2-mm sieve for soils 0 to 10 cm in depth, with 162 falling into the city itself, which covers an area of roughly 25 square miles.

### Site 2

A total of 100 soil samples were collected from a roughly 1-square mile portion of the flood plain of the Geul River valley, located in the south of The Netherlands. This region is perilously polluted by heavy metals in stream sediments resulting from historic metal mining and deposited by flooding (Heuvelink 1999). Assay results for the soil samples collected from this region constitute part of the data for a pediatric lead ingestion study. These soil sample locations were determined by the play habits of children residing in campground sites for at least 3 days. Composite soil samples from the upper 5 cm of approximately 100 g were taken in duplicate, air-dried, processed through a 2-mm sieve, and powdered (van Wijnen et al. 1990, Leenaers 1991).

### Site 3

In all, 277 surface (0-2") soil samples were collected in and around an abandoned lead smelting facility superfund site located in Murray, Utah. This area was polluted by airborne emissions and placement of waste slag from the smelting process. Three of these samples failed to have a geocode recorded, 173 samples were collected from the superfund site itself, and 101 samples were collected from the adjacent community. The composite study area covers roughly a 0.5 square mile area. Thirty-eight soil samples can be grouped into 17 clusters on the basis of their common georeference coordinates; in these samples, the juxtaposed assay results were pooled for composite measures. The result is 253 locations for which lead concentrations were measured. A single spot within the site was intensively sampled.

### Site 4

A total of 236 surface soil samples were collected for a skeet-and-trap shooting range superfund site housed on a roughly 0.1 square mile area. The shooting positions were located along the southern boundary of the site. The measurements of lead concentration were taken in the top 6 inches of soil. A single spot within the site was intensively sampled. The sampling network used throughout most of the site reflects a square grid pattern.

## Step I: The Statistical Frequency Distribution Description of Soil Lead Concentration

Many environmental measures conform to a log-normal distribution (Gilbert 1987, Millard and Neerchal 2001) or empirically a frequency distribution where changing each data value to its natural logarithmic counterpart yields a set of values that conforms to a normal distribution. This frequency distribution tends

**Table 1** Shapiro-Wilk normality assessment statistics for soil lead samples (null hypothesis probabilities in parentheses). The null hypothesis value is 1

	<b>pb</b>	<b>LN(pb)</b>	<b>LN(pb + <math>\delta</math>)</b>	<b>Residuals</b>
City of Syracuse	0.621 (<0.0001)	0.947 (<0.0001)	0.969 (0.0008)	0.982 (0.0350)
Geul River flood plain	0.917 (<0.0001)	0.957 (0.0027)	0.959 (0.0033)	0.980 (0.1233)
Smelter superfund site	0.523 (<0.0001)	0.976 (0.0003)	0.990 (0.0819)	0.990 (0.0819)
Smelter superfund site: nonresidential	0.602 (<0.0001)	0.975 (0.0032)	0.987 (0.1047)	*****
Skeet/trap shooting range superfund site	0.228 (<0.0001)	0.917 (<0.0001)	0.968 (<0.0001)	0.995 (0.5858)

**Table 2** Null hypothesis probabilities of homogeneity of variance assessments for log-transformed soil lead concentration measures from selected geographic landscapes. For each statistic, the null hypothesis value is 0

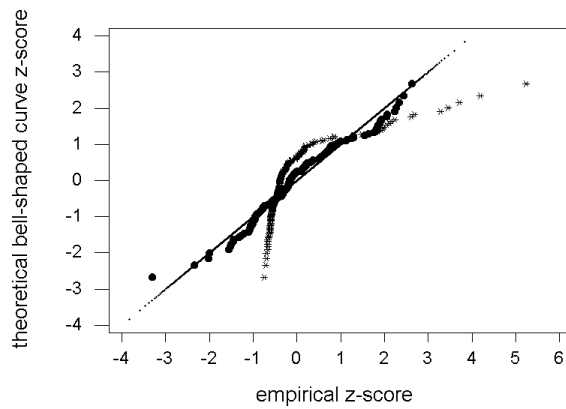
	<b>Bartlett's Statistic</b>	<b>Levene's Statistic</b>
City of Syracuse: location types	0.032	0.018
City of Syracuse: quadrants of the plane	0.194	0.345
Geul River: sides of river	0.644	0.823
Geul River: quadrants of the plane	0.066	0.123
Smelter superfund site: residential/nonresidential	0.000	0.000
Smelter superfund site: quadrants of the plane	0.000	0.000
Skeet/trap range superfund site: distance from gallery	0.000	0.000
Skeet/trap range superfund site: quadrants of the plane	0.000	0.000

to describe pollution measures well because they are bounded below at zero and are often strongly positively skewed. Because a heavy metal such as lead occurs naturally in all soils, its lower bound may differ from zero, requiring a threshold parameter to be included in the log-normal distribution specification. Pollution is deposited in a geographic landscape by point source human activities, such as leaded gasoline emissions dispersing from cars moving along roads. Relatively small amounts are deposited in most locations, while increasingly larger amounts are deposited in fewer and fewer locations. One of the critical properties of lead is that it does not typically migrate from where it is deposited and it does not decay or biodegrade into something else; rather, it adheres to fine clay and organic matter particles, accumulating in the upper few inches of undisturbed soil. If the process depositing lead pollution is repetitive, then, with some stochastic fluctuation, each layer of pollution has the same geographic distribution resulting in new deposit amounts being proportional to existing deposit amounts at each location. Thus, the cumulative effect of many layers of small deposits is multiplicative, resulting in the log-normal distribution. This type of depositing process is the expected outcome of gasoline exhaust emissions, periodic river flooding, smelter air pollution, or even skeet shooting waste. Even if pollution were repeatedly deposited at random, following a uniform distribution, the results would be approximately log-normally distributed. The first scenario is consistent with the presence of positive spatial autocorrelation; the second scenario is consistent with the absence of spatial autocorrelation. In either case, the repetitiveness of the depositing mechanism would result in some type of patterned variance.

The functional form of the logarithmic transformation utilized here includes a translation parameter,  $\delta$ , rendering  $LN(\text{lead concentration} + \delta)$ , where the translation term may be a function

of the minimum lead concentration measure (see Griffith et al. 1998b). This translation term is a mean response for a non-linear model specification. Of note is that soil sample locations are rarely randomly selected. Regardless of whether or not they are, formal tests for normality and homogeneity of variance are based on random sampling from the lead concentration distribution rather than the set of soil sample locations. Accordingly, these attribute values are what statistical tests of the hypotheses for normality and constant variance relate to; statistical significance is established using a model-based inference framework.

For the Syracuse geographic landscape, the minimum soil lead concentration measure of 0.4 parts per million (ppm) is aberrantly low,  $\delta = 3$ , and the log-transformed version of the 167 measures conforms much more closely to a normal distribution (see Table 1 and Figure 1). Pooling the second sample with the first sample decreases  $\delta$  to 2. Increasing the complexity of the transformation to account for this single deviant fails to dramatically improve conformity with normality. Meanwhile, two assessments of constant variance are possible with these data. Bretsch (1998) recorded the location type for soil samples using the following categories: streetside soil ( $n = 74$ ), park soil ( $n = 30$ ), playground soil ( $n = 17$ ), house lot soil ( $n = 30$ ), and vacant lot soil ( $n = 17$ ). In addition, the geographic landscape can be arbitrarily divided into four regions by centering the soil sample geocodes; the geocodes are then grouped together according to their locations in each of the four standard quadrants of a plane. This procedure results in groups of 35, 45, 44, and 43 soil sample locations. The Bartlett and the Levene statistics used to evaluate constant variance based on these two different schemes appear in Table 2. The Bartlett statistic requires a normal frequency distribution; the Levene statistic is relatively insensitive to departures from a normal distribution, and furnishes a reliability gauge for the corresponding Bartlett statistic.



**Figure 1:** Quantile Plot for Syracuse Soil Samples: Perfect Normality (*dot*), Raw Data (*asterisk*), and Log-Transformed Data, Including  $\delta$  (*solid circle*).

While variability of log-lead concentration within these five groups has overlapping 95% confidence intervals, the confidence interval for vacant lots is substantially wider; the result is both a Bartlett and a Levene test statistic for homogeneity of variance that is significant. More generally, while variability is very similar for log-lead concentration in park, playground, and house lot soil, it is modestly less in streetside soil and markedly greater in vacant lot soil. Hence, the vacant lot soil samples are the primary source for a potentially statistically significant difference in log-lead concentration variances across the five categories. In contrast, variability of log-lead concentration within the four quadrants of the plane fails to exhibit a significant difference, with the four 95% confidence intervals having considerable overlap.

For the Geul River flood plain geographic landscape,  $\delta = 19$ , the log-transformed version of the 100 measures conforms much more closely to a normal distribution (see Table 1), with a plot very similar to that appearing in Figure 1. Again, two assessments of constant variance are possible with the data. On the one hand, no apparent difference exists between the means or variances of the log-lead concentration measures obtained for the east ( $n = 46$ ) and the west ( $n = 54$ ) banks of this river. On the other hand, after employing a  $45^\circ$  rotation of the geocoding axes in order to obtain a more uniform distribution of soil sample locations by quadrant of the plane—resulting in group sizes of 24, 23, 27, and 26—a modest difference in variances is detectable. As before, the Bartlett and the Levene statistics used to evaluate constant variance based on these two different schemes appear in Table 2. Meanwhile, the 95% confidence intervals for variability of log-lead concentration within the four quadrants of the plane overlap. Both the accompanying Bartlett and Levene test statistics suggest that the variability in log-lead concentration is roughly homogeneous across this region. A Bonferroni adjustment for simultaneous testing reinforces this conclusion.

For the smelter superfund site geographic landscape, both the minimum (i.e., 37 ppm) and the maximum (i.e., 33,000 ppm) soil lead concentration measures are aberrant ones,  $\delta = -32$ , and the log-transformed version of the 277 measures conforms much more closely to a normal distribution (see Table 1), again with a

plot very similar to that appearing in Figure 1. As before, increasing the complexity of the transformation to account for these two extreme deviants fails to dramatically improve conformity with normality. However, restricting attention to the superfund site itself does yield a somewhat closer correspondence with a bell-shaped curve, in part because the minimum outlier value of 37 ppm is located in the residential portion of this landscape.

Inspection of Table 2 confirms a substantial variance difference between residential and non-residential regions. As one might expect, the superfund site exhibits a markedly greater level of variability. Variability again can be analyzed in terms of the four quadrants of the plane, which respectively contain 70, 65, 76, and 63 soil sample locations. Considering these groups in counter-clockwise order, log-lead concentration in the fourth quadrant displays substantially less variability than do the measures in the remaining three quadrants. This finding is attributable to this set of locations mostly being in the residential area of the region. The second quadrant is a mixture of superfund site and residential sample locations, which suppresses the variability displayed by the log-lead concentration measures obtained for it. Restricting attention to the superfund site itself does not render a more favorable assessment. The group sizes are currently 41, 19, 50, and 63. While all four of the 95% confidence intervals overlap, the log-lead concentration variability displayed by the second quadrant continues to deviate markedly from that displayed in the remaining three quadrants.

The skeet-and-trap shooting range superfund site also renders a frequency distribution of lead concentration that is approximately log-normally distributed, with  $\delta = 12$ . The log-transformed version of the 236 measures conforms much more closely to a bell-shaped curve (see Table 1); however, even though substantial improvement is attained, the transformed values continue to deviate from normality. Comparing the log-lead concentration variability across the superfund site once more reveals nonconstant variance (see Table 2). Analysis based on the four quadrants of the plane involves subregions containing 49, 60, 68, and 49 soil sample locations. The average log-lead concentration is roughly the same in the upper left- and lower right-hand quadrants, and is markedly greater than that for the upper right- and lower left-hand quadrants, whose average log-lead levels are approximately the same. The variance for these log-lead concentration measures is approximately the same in the top two quadrants, is marginally greater than that displayed in the lower right-hand quadrant, and is noticeably greater than that displayed in the lower left-hand quadrant. Inspection of variability from the southern border of the site, where the shooting gallery was located, also suggests the presence of nonconstant variance.

In conclusion, an analyst should expect that the statistical distribution of soil lead concentration measures for geographic landscapes might be best described with a log-normal distribution coupled with a landscape-specific translation parameter. The analyst should also expect that these measures will display heterogeneous variability across heavily polluted (e.g., superfund) sites and more or less display homogeneous variability across other

types of geographic landscapes, whether this variability is based on geographic subregions or attribute-based location categories. Of note is that, while the log-transformation fails to completely equalize variance in the two superfund site cases, diagnostics suggest that it is the most suitable transformation to use. Exploratory work to date with these transformed data has failed to reveal a good weighting scheme to employ in order to compensate for any persisting nonconstant variance. Additionally, both the Bartlett and the Levene test statistics warrant inspection here, since the Bartlett test statistic is both more sensitive to deviations from constant variance and far less robust against deviations from normality than the Levene test statistic.

## Step 2: Spatial Autocorrelation and The Geographic Distribution of Soil Lead Contamination

Two geographic distribution features of soil samples merit investigation. Spatial autocorrelation allows predictions of soil lead concentrations at unsampled locations. It also enables an effective sample size (the number of equivalent independent values) to be computed for data whose frequency distribution emulates a bell-shape curve. This second feature can be better understood through the calculation of an additional statistic pertaining to the spacing of soil samples. The average first nearest-neighbor distance supplies such a statistic. If the soil sample locations are randomly positioned, this statistic takes on a value of 1; if the locations are uniformly positioned, this statistic takes on a value of approximately 2.15.

A map for the Syracuse geographic landscape appears in Figure 2. The top map shows the census tracts for the city, together with their centroids and the most recent 167 soil sample locations. The bottom map shows a Thiessen polygon surface partitioning constructed with the soil sample locations, upon which the census tract centroids have been superimposed. The Moran Coefficient—a spatial autocorrelation index that is similar to a product moment correlation coefficient—based on this tessellation is 0.16663, which is both significant and indicates that a weak tendency exists for similar values of log-lead concentration measures to be in nearby sample locations. A pure simultaneous spatial autoregressive (SAR) model— $Y = \rho WY + \epsilon$ , for georeferenced variable  $Y$ , spatial weights matrix  $W$ , and random error  $\epsilon$ —quantifies the nature and degree of spatial autocorrelation as  $\hat{\rho} = 0.35196$  (on a scale of 0 to 1), which also indicates the

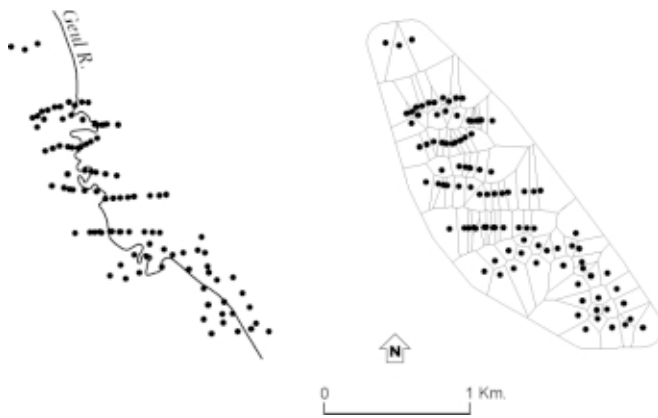


**Figure 2:** City of Syracuse, New York, with Soil Sample Locations (cross) and Tract Centroids (solid circle). *Top:* Census Tracts. *Bottom:* Thiessen Polygon Surface Partitioning for the Soil Sample Locations.

presence of a weak tendency for similar values to cluster in the city. The most appealing semivariogram description is furnished by a Bessel function [see Figure 6; this function can be estimated following Griffith and Layne (1999) or with ESRI's *Geostatistical Analyst*]; parameter estimates include 0.274 for the nugget, 2.490 for the slope parameter, and 0.146 for the range parameter (based on standardized distance), with the relative error sum of squares being 0.228. Hence, the geographic distribution of soil lead concentration across Syracuse may be described in a manner that supports spatial interpolation of the surface.

The effective sample size is 43.1% of  $n$ ; approximately 12% of the variance in log-lead concentration is accounted for by nearby values of log-lead. The distribution of the 162 soil sample points within the city is essentially random, raising the possibility of poor geographic coverage by the sampling network. In fact,

Landscape	$\hat{\rho}$	$n$	Effective Sample Size	Spatial Autoregressive % Variance Accounted	First Nearest-Neighbor Statistic
City of Syracuse	0.37111	167	71.9	11.5	1.02868
Geul River flood plain	0.79251	100	12.0	54.9	0.82032
Smelter superfund site	0.53603	253	68.2	23.8	0.06208
Skeet/trap shooting range superfund site	0.76404	236	25.3	53.2	1.61596

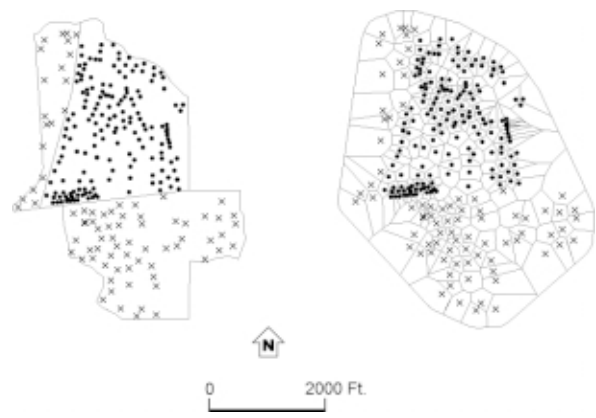


**Figure 3:** Geul River, The Netherlands, with Soil Sample Locations (solid circle). *Left:* Soil Samples Spanning the River. *Right:* Thiessen Polygon Surface Partitioning for the Soil Sample Locations.

four census tracts (CTs) (i.e., CTs 15, 28, 38, and 42) are without soil samples. The semivariogram model renders the following respective  $LN(Pb + 3)$  predicted values for the centroids of these four census tracts, when kriging weights are restricted to being positive: 4.99744, 4.02726, 4.04000, and 4.81637. These results suggest that only a modest amount of new information will be gained from additional soil samples and that any supplemental sample locations need to be judiciously selected.

A map for the Geul River flood plain appears in Figure 3. The left-hand map shows the river, flood plain, and 100 soil sample locations. The right-hand map shows a Thiessen polygon surface partitioning constructed with the soil sample locations. The Moran Coefficient based on this tessellation is 0.42492, which is significant and also indicates that a moderate tendency exists for similar values of log-lead concentration measures to be in nearby sample locations. A pure SAR model quantifies the nature and degree of spatial autocorrelation in this case to be  $\hat{\rho} = 0.79251$ , which confirms the presence of a moderate tendency for similar values to cluster in this flood plain. Again, the most appealing semivariogram description was obtained with a Bessel function (see Figure 6); parameter estimates include 0.061 for the nugget, 0.550 for the slope parameter, and 0.358 for the effective range (based on standardized distance), with the relative error sum of squares being 0.250. Hence, as was found for the Syracuse case, the geographic distribution of soil lead concentration across the Geul River flood plain may be described in a manner that supports spatial interpolation of the surface.

The effective sample size is a mere 12% of  $n$ ; approximately 55% of the variance in log-lead concentration is accounted for by nearby values of log-lead. The distribution of the 100 soil sample points within the flood plain is random, with a noticeable tendency toward clustering, raising the possibility of poor geographic coverage by the sampling network. This finding is partly an artifact of the use of transects for sampling. These results suggest that considerable redundant information is contained in the soil samples already collected. Additional information can be gained from supplementary soil samples if they are collected



**Figure 4:** Murray, Utah, Smelter Superfund Site, with Soil Sample Locations in the Site Itself (solid circle) and in Nearby Residential Neighborhoods (cross). *Left:* Soil Samples Distributed across the Superfund Site and Adjacent Neighborhoods to its West and South. *Right:* Thiessen Polygon Surface Partitioning for the Soil Sample Locations.

in a non-transect fashion from those sections of the flood plain in which few samples already have been collected. Conspicuous undersampled subregions include the northeast and the northern parts of the landscape.

Finally, 253 soil samples were collected for the smelter superfund site and 236 were collected for the skeet-and-trap shooting range superfund site. A map depicting this first case appears in Figure 4. The left-hand side of the figure presents a map showing the distribution of soil samples with both the site itself and the nearby residential community. The right-hand side of the figure presents a Thiessen polygon surface partitioning for the soil samples. The Moran Coefficient based on this tessellation is 0.26588, which is both significant and indicates that a weak-to-moderate tendency exists for similar values of log-lead concentration measures to be in nearby sample locations. A pure SAR model quantifies the nature and degree of spatial autocorrelation as  $\hat{\rho} = 0.53603$ , which also indicates the presence of a weak-to-moderate tendency for similar values to cluster in this superfund site. Again, the most appealing semivariogram description was obtained with a Bessel function (see Figure 6); parameter estimates include 0.071 for the nugget, 2.660 for the slope parameter, and 0.132 for the range parameter (based on standardized distance), with the relative error sum of squares being 0.234. Of note is that, while the mean log-lead concentration levels for the superfund and residential areas appear to be statistically significantly different (see Table 2), adjusting for the difference of means yields a Moran Coefficient of 0.23963, which deviates little from that for the unadjusted data. Hence, as for the two preceding cases, the geographic distribution of soil lead concentration across the superfund site may be described in a manner that supports spatial interpolation of the surface.

A map portraying a Thiessen polygon surface partitioning for the skeet-and-trap shooting range superfund site appears in Figure 5. In this case the log-transformed measures yield a Moran

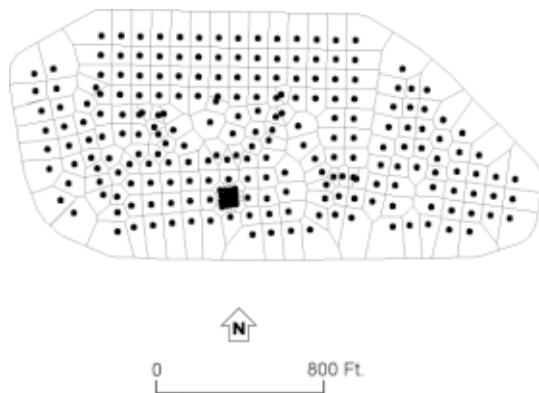
Coefficient value of 0.48873, which is both significant and indicates that a moderate tendency exists for similar values of log-lead concentration measures to be in nearby sample locations. A pure SAR model quantifies the nature and degree of spatial autocorrelation as  $\hat{\rho} = 0.76404$ , which also indicates the presence of a moderate, pronounced tendency for similar values to cluster across this superfund site. As with the preceding three empirical examples, the most appealing semivariogram description was obtained with a Bessel function (see Figure 6); parameter estimates include 0.894 for the nugget, 3.510 for the slope parameter, and 0.138 for the range parameter (based on standardized distance), with the relative error sum of squares being 0.071. Once more, the geographic distribution of soil lead concentration may be described in a manner that supports spatial interpolation of the surface.

The effective sample sizes for these two superfund sites, respectively, are 27.0% and 10.7% of their corresponding  $n$  values; their respective percentages of the variance in log-lead concentration (accounted for by nearby values of log-lead) are approximately 24 and 53. The distribution of soil sample points across each of the two superfund sites is quite different: the smelter site has a strong tendency for its sample locations to cluster, whereas

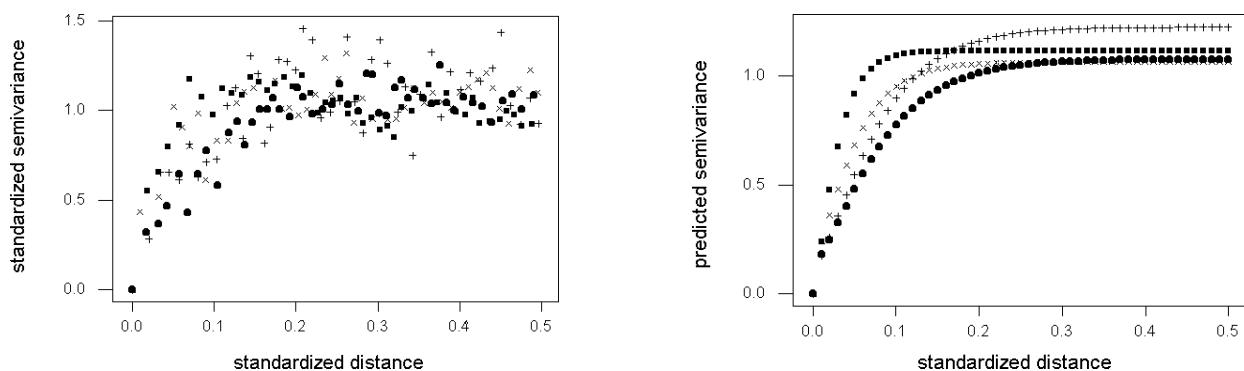
the skeet-and-trap shooting range site has a strong tendency for its sample locations to be uniformly spaced. These results are due in part to: 1) a very intensively sampled section of the smelter superfund site coupled with several sparsely sampled subregions, and 2) the visibly noticeable square grid network component used in the skeet-and-trap shooting range superfund site (see Figure 5). Interestingly, considerable redundant information is contained in the soil samples already collected for the shooting range superfund site, even though its soil sample locations are more uniformly spaced, and less redundant information is contained in the samples already collected for the smelter superfund site. This outcome is partly due to the sampling intensity per unit area in the shooting range site being nearly six times that in the smelter site. Thus, little additional information can be gained from supplementary soil samples for the first site, while judiciously selected supplemental soil sample locations could yield considerable new information for the second site.

Therefore, the geographic distribution of soil lead concentration measures for these geographic landscapes may be described as containing weak-to-moderate positive spatial autocorrelation and a spatial dependency structure that may be described with a Bessel function semivariogram model. For comparative purposes, the log-transformed lead concentration measures were converted to  $z$ -scores for each of the four landscapes, with the resulting semivariogram plots appearing in Figure 6. A comparison of these plots reveals that: 1) the City of Syracuse and the Geul River flood plain display considerably more variability with increasing distance than do the superfund sites, 2) the ascending rank-order levels of autocorrelation should be for the smelter superfund site, followed by the City of Syracuse, and then roughly a tie for the top rank by the Geul River flood plain and the skeet-and-trap shooting range superfund site. The effective standardized distance ranges for these cases fall into the interval (0.08, 0.23).

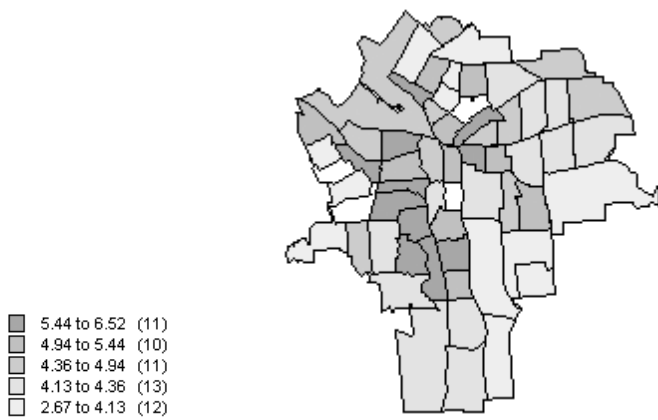
Non-zero semivariogram nuggets may be attributed to measurement error, specification error, or some other unaccounted for source of error. The nugget values reported here may relate to a geographic scale effect and/or its interaction with sampling intensity (a form of resolution), since the landscapes range in size from 0.1 to 25 square miles while their intensities range in magnitude from 7 to 2360 samples per square mile. Unfortunately,



**Figure 5:** Skeet-and-trap Shooting Range Superfund Site, with Soil Sample Locations (solid circle) and a Thiessen Polygon Surface Partitioning for the Soil Sample Locations.



**Figure 6:** Soil Lead Concentration Semivariogram Plots: City of Syracuse (cross), Geul River Flood Plain (plus), Smelter Superfund Site (solid square), and Skeet-and-trap Shooting Range Superfund Site (solid circle). *Left:* Empirical Semivariogram Plots. *Right:* Predicted Semivariogram Plots.



**Figure 7:** Geographic Distribution of Soil Lead Concentration in the City of Syracuse, New York, Aggregated by Census Tract.

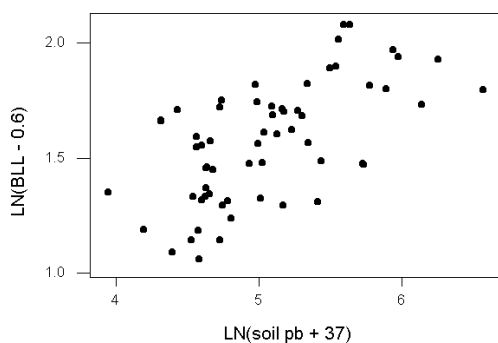
with only four landscapes, it is not possible to determine whether or not this is the case; figures for the City of Syracuse are inconsistent with a possible trend portrayed by the remaining three landscapes.

## Exploring Concerns

Three concerns merit attention here, all of which pertain to what has been learned from the data analyses presented in this article that is transferable to other places. The first concern focuses on why a spatial analysis of soil samples is worth undertaking. The second concern relates to cost-effectiveness of spatial sampling. The third concern centers on the ability to predict soil lead contamination at unsampled locations based on data from sampled locations.

### Pediatric Lead Poisoning: A Public Health Concern

Mielke (1999) argues that an accurate and complete appreciation of the distribution of lead in the environment is needed. Findings reported in this article seek to establish a description of this distribution, in terms of both statistical frequency and geographic variability. The utility of this type of description may be demonstrated by employing it in observational public health studies, such as those addressing pediatric lead poisoning. For example, the combined two sets of Syracuse soil lead concentration measures were used to construct a choropleth map of soil lead



**Figure 8:** Scatter Plot of Relationship between Soil Lead and Pediatric Blood-lead Levels in Syracuse, New York, by Census Tract.

levels by census tract (see Figure 7). This map is a marked improvement over the one presented in Griffith et al. (1998a), which is based only on the first soil sample set. Soil sample values within a census tract can be aggregated by summarizing them with a single geometric mean. The geometric mean is the preferred measure of central tendency when soil lead concentrations conform to a log-normal distribution, and actually represents an arithmetic mean of the log-transformed values. The combined two soil sample sets yield 57 aggregate soil concentration geometric mean values that closely conform to a log-normal distribution (the Shapiro-Wilk test statistic null hypothesis probability is 0.495), with  $\delta = 0$ , display constant variance across the four quadrants of the plane and across sample sizes (the respective null hypothesis test statistic probabilities are: for Bartlett, 0.895 and 0.422; for Levene, 0.616 and 0.643), and exhibit weak positive spatial autocorrelation (Moran Coefficient = 0.17050;  $\hat{\rho} = 0.33962$ ). Covariation between these measures and mean pediatric blood-lead levels is portrayed in Figure 8; spatial autocorrelation latent in the soil lead values allows interpolated measures to be calculated for the four tracts lacking soil samples. A conspicuous positive relationship is displayed between these two georeferenced ecological variables. Figure 8 portrays this relationship, displaying a positively sloping underlying (invisible) trend line from which the scatter of points considerably deviates.

The spatial statistical model relating pediatric blood-lead levels and soil lead concentration may be written as follows:

$$\overline{LN(pb_{bll} - 0.6)} = 0.58948 \sum_{j=1}^n w_{ij} \overline{LN(pb_{soilj} - 0.6)} - 0.93676(1 - 0.58948) + 0.20251LN(pb_{soil} + 37) + e, \text{ adj-}R^2 = 0.566,$$

where  $Pb_{bll}$  denotes the blood-lead level of an individual child (in micrograms/deciliter),  $Pb_{soil}$  denotes the geometric mean of lead concentration in soil samples (in ppm),  $w_{ij}$  denotes the geographic weight for census tracts  $i$  and  $j$  ( $0 \leq w_{ij} \leq 1$ ;  $\sum_{j=1}^n w_{ij} = 1$ ),  $LN$  denotes the natural logarithm (base  $e = 2.7182818$ ), and  $e$  is an error term.

Diagnostics associated with this equation are as follows: normality of residuals: Shapiro-Wilk statistic = 0.985 ( $p = 0.661$ ); attribute homogeneity of variance: lack of any conspicuous patterns in the  $\hat{y}$ -versus- $e$  plot; homogeneity of spatial variance null hypothesis probabilities: 0.460 for Bartlett, 0.730 for Levene; and spatial autocorrelation: Moran Coefficient =  $-0.04785$  ( $z_{MC} \approx 0.5$ ),  $GR = 1.01429$ .

These diagnostics imply that this equation lacks specification error due to assumption violations. However, the percentage of variance accounted for suggests that covariates may be missing from the equation.

Therefore, children living in census tracts with higher soil lead concentrations appear to be at higher risk of becoming lead poisoned than do children living in census tracts with lower soil lead concentrations. The spatial statistical analysis and methodology outlined in this article make such an assessment of an important public health problem possible.

## Representative Maps: Spatial Sampling Concerns

Numerous features of sampling warrant discussion here, namely intensity, spacing, coverage, information content, and precision. As mentioned earlier, the best spatial sampling network is based on a hexagonal grid, and as such has a first nearest neighbor statistic close to 2.15. Deviations from this grid are often necessitated by factors such as cost and the feasibility of collecting a sample. Both the City of Syracuse and the Geul River flood plain samples are dramatically impacted by these considerations. In the case of Syracuse, mostly public locations have been sampled because of permission and access constraints. In the case of the Geul River, transects were used to help reduce costs and because campgrounds were of interest. Some transect sampling appears to have occurred in the smelter superfund site as well (see Figure 4).

Meanwhile, as the intensity of sampling increases, spacing will decrease, resulting in an increase in the degree of spatial autocorrelation for sample values. In turn, incremental information content decreases. Both of the superfund site samples are impacted by this consideration. When hot spot subregions were detected in these sites, the subregions were intensively sampled in order to confirm their elevated pollution status. Both sites have one subregion that has been intensively sampled. While the replicate information is redundant, it also is confirmatory, illustrating one of the values of securing duplicate information during sampling.

In other words, the effective sample size is helpful—as is the first nearest neighbor statistic—when assessing coverage of a landscape, especially in cases where supplemental sampling is expected to occur. Knowledge of the effective sample size informs a scientist about where new samples should be taken, in an attempt to obtain as representative a geographic coverage as possible. If intensity is great enough, little beyond confirming previous sample results will be gained by securing new samples. If subregions are more intensively sampled, little will be gained by securing new samples from them; rather, more new information can be gained by securing new samples from the least intensively sampled subregions. By doing so, the sampling grid would undergo a modifi-

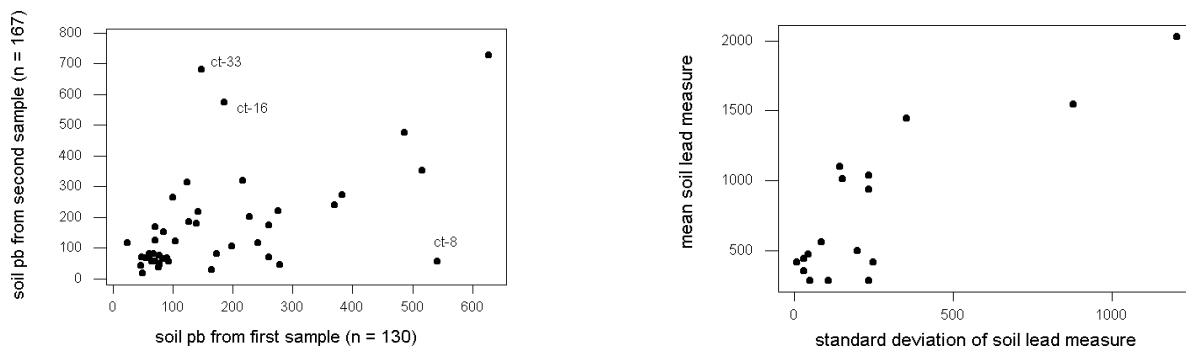
cation that would move it closer to the uniformly spaced hexagonal grid.

Precision of a sample statistic is important as well. This precision is achievable in a conventional random sampling context by increasing the sample size, with the minimum necessary sample size being in the interval (30, 100). Costs associated with satisfying this requirement more than likely will be prohibitive for geographic landscapes. For example, stratifying by the 61 census tracts of Syracuse would require at least 1830 soil samples. Furthermore, the effective sample size resulting from latent spatial autocorrelation in soil lead concentrations means that this number would need to be much larger. Regardless of spatial autocorrelation effects, issues of precision still need to be addressed. Consider the scatterplot comparison of the two samples for Syracuse presented in Figure 9. Not only does the displayed scatter of points deviate too much from a straight line, but three census tracts (i.e., CTs 8, 16, and 33) have estimate pairs that dramatically differ. Increasing sample sizes within these tracts, even at the expense of securing redundant information, would help circumvent this problem. Similarly, the scatterplot comparison of replicate samples for the smelter superfund site presented in Figure 9 suggests a relationship between arithmetic mean lead level and standard deviation; good precision would be accompanied by a standard deviation close to zero at all concentration levels. Of the 17 replicate locations, four are triplicated and 13 are duplicated. Again, a lack of precision is indicated by these replicates.

Therefore, even when considerable spatial autocorrelation characterizes a set of soil samples, the sampling distribution variability of mean surface maps for a landscape must be controlled with adequate sample size. Spatial autocorrelation can be used to help bolster precision; however, using it to decrease sample size too much may compromise precision.

## Maps With Holes: Missing Data Concerns

Another salient concern stems from the failure to have adequate geographic coverage after post-stratifying a sample. In the Syracuse, NY, example included in this article, 297 soil samples—



**Figure 9:** *Left:* Pairs of Soil Lead Geometric Means for 49 Census Tracts Appearing in Both of the Syracuse, New York, Samples. *Right:* Duplicate Soil Samples for 17 Locations in the Smelter Superfund Site.

some of which fall outside the city limits—are post-stratified into 61 census tracts, resulting in four tracts failing to contain any of the soil sample locations. Of course, one remedy is to secure a supplemental set of soil samples specifically from the four census tracts in question; of note is that a stratified random sampling design would avoid this problem. A much cheaper solution is to interpolate the missing data. Interpolation exploits latent spatial autocorrelation by borrowing duplicated information from nearby measures and emphasizes the normality assumption.

Use of the Box-Cox transformation,  $LN(Pb + \delta)$ , allows one to work with soil lead concentration measures that more closely mimic a normal curve. In turn, normal curve theory states that the expected value of Pb, say  $E(Pb)$ , is a function of the mean and the variance of  $LN(Pb + \delta)$ . More specifically,  $E(Pb)$  is given by  $-\hat{\delta} + EXP[\overline{LN(pb + \hat{\delta})}]H[\psi_n(\hat{s}_{LN(pb+\hat{\delta})}^2/2)]$ , where  $EXP$  denotes the anti-logarithm for  $LN$  and the back-transformation correction factor  $\psi_n$  is defined by Gilbert (1987). When spatial autocorrelation is present, the mean response also contains an autoregressive term. The SAR interpolation is used in rendering the pediatric blood-lead levels analysis back-transformed values: 116.7 ppm for CT 15, 102.9 ppm for CT 28, 125.7 ppm for CT 38, and 115.1 ppm for CT 42. Of note is that the back-transformation correction factors associated with these values range from 1.004 to 1.011, factors that are negligibly different from 1.

These interpolated values differ considerably from those obtained through kriging with the fitted Bessel semivariogram model. The effective range for this model is given by 4 times the range parameter ( $r$ ), which equals 0.584 units of standard distance in this case. Restricting attention to the local neighborhood of each of the four census tract centroids (which is operationalized as a circle of radius 0.070 to 0.085 units of standardized distance in order to avoid negative kriging weights) renders 164.7 ppm for CT 15, 59.0 ppm for CT 28, 60.7 ppm for CT 38, and 139.5 ppm for CT 42. Extending the radius to the range in all four cases, and hence accepting negative kriging weights, modifies these numbers to 152.6, 74.3, 57.0, and 151.1, respectively. Because the SAR interpolation is based on averages as well as the combined samples, whereas the kriging results are based on individual measures contained in the second sample, the SAR interpolations may be more reliable. Unfortunately, little work has been completed to date comparing these two interpolators.

## Conclusions

An answer can be put forth now to the question asking how the statistical and geographic distributions of soil lead concentration in the inhabited environment can be described. The statistical frequency distribution (or more formally, the statistical probability density function) appears to be a three-parameter log-normal distribution. Spatial statistical analysis reveals that the most appropriate semivariogram model appears to be the Bessel function, which links directly to the SAR spatial statistical model. The necessary geographic sampling design supporting estimation of these models needs to furnish good coverage of a landscape, with a sufficiently large sample size to ensure adequate precision of results. All of these findings should be transferable to places other than the four landscapes explored in this article. Furthermore, the methodology outlined here establishes expectations about geographic and non-geographic variability of soil lead concentrations in other landscapes, and furnishes a guide for quantifying and analyzing this variability in these other landscapes.

The importance of these findings is reflected in societal concerns such as pediatric lead poisoning. Being able to describe the statistical and geographic distributions of soil lead concentration in the inhabited environment furnishes a tool that can contribute to the solving of such problems. Scientific concerns contributed to by the knowledge of statistical and geographic distribution of soil lead concentration include methodological contributions for evaluating important quality features of a geographic sample and for plugging holes in maps.

---

## About The Author

**Daniel A. Griffith** is a professor of Geography at Syracuse University. His most recent work concerning pediatric lead poisoning in Syracuse was the topic of a feature article in the *Syracuse Herald-Journal*. He has published numerous spatial statistics articles in the geography, regional science, statistics, and mathematics literature. He has been a Fulbright Fellow, an American Statistical Association Research Fellow, and a Guggenheim Fellow. His biographical profile is listed in the Marquis *Who's Who in the World*.

Corresponding Address:

Daniel A. Griffith

Department of Geography, Syracuse University

Syracuse, NY 13244-1020

griffith@maxwell.syr.edu

---

## Acknowledgments

The Syracuse, NY soil lead concentration data have been made available to the author by Dr. David L. Johnson, Department of Chemistry, SUNY College of Environmental Sciences and Forestry, who with Ms. Jennifer K. Bretsch collected them. The flood plain soil lead concentration data have been made available to the author by Dr. Gerard Heuvelink; these data were originally collected by Dr. H. Leenaers, with financial assistance from the Department of Physical Geography, Utrecht University, the Dutch National Research Foundation NWO, and the Province of Limburg (NL). The superfund site soil lead concentration data have been made available to the author by Dr. Philip E. Goodrum, Research Scientist, and Mr. William C. Thayer, Syracuse Research Corporation. Dr. Susan Griffin, U.S. Environmental Protection Agency Region 8, furnished data on the original Murray superfund site.

---

## References

- Bretsch, J., 1998, *Soil Lead and Children's Blood-lead Levels in Syracuse, New York* (Syracuse, NY: unpublished master's thesis, Department of Chemistry, State University of New York College of Environmental Science and Forestry).
- Cressie, N., 1991, *Statistics for Spatial Data* (New York: John Wiley & Son).
- Gilbert, R., 1987, *Statistical Methods for Environmental Pollution Monitoring* (New York: van Nostrand Reinhold).
- Griffith, D. and L. Layne, 1999, *A Casebook for Spatial Statistical Data Analysis* (New York: Oxford University Press).
- Griffith, D. and M. Zhang, 1999, Computational Simplifications Needed for Efficient Implementation of Spatial Statistical Techniques in a GIS. *Geographic Information Sciences*, 5(2), 97-105.
- Griffith, D., P. Doyle, D. Wheeler, and D. Johnson, 1998a, A Tale of Two Swaths: Urban Childhood Blood-lead Levels across Syracuse, NY. *Annals Association of American Geographers*, 88(4), 640-665.
- Griffith, D., J. Paelinck, and R. van Gastel, 1998b, The Box-Cox Transformation: Computational and Interpretation Features of the Parameters. In Griffith, D. and C. Amrhein (Eds.), *Advances in Spatial Modelling and Methodology: Essays in Honor of Jean Paelinck* (Dordrecht: Kluwer), 45-56.
- Heuvelink, G., 1999, Aggregation and Error Propagation in GIS. In Lowell, K. and A. Jaton (Eds.), *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources* (Ann Arbor, MI: Ann Arbor Press), 219-225.
- Leenaers, H., 1991, Deposition and Storage of Solid-Bound Heavy Metals in the Floodplains of the River Geul (The Netherlands). *Environmental Monitoring Assessment*, 18, 79-103.
- McGarigle, B., 2000, GIS Draws a Bead on Lead Poisoning in Children. *Government Technology*, 12(2), 40, 42.
- Mielke, H., 1999, Lead in the Inner Cities. *American Scientist*, 87(1), 62-73.
- Millard, S. and N. Neerchal, 2001, *Environmental Statistics with S-Plus* (Boca Raton, FL: CRC Press).
- Rauci, J., 1999, Addressing the Problem of Elevated Pediatric Blood-lead Levels From a State Perspective, article presented to The Second Syracuse Lead Conference, Syracuse, NY, October 27.
- Reissman, D., F. Staley, G. Curtis, and R. Kaufmann, 2001, Use of Geographic Information System Technology to Aid Health Department Decision Making about Childhood Lead Poisoning Prevention Activities. *Environmental Health Perspectives*, 109(1), 89-94.
- Sen, A. and M. Srivastava, 1990, *Regression Analysis: Theory, Methods, and Applications* (Berlin: Springer-Verlag).
- Spake, A. and J. Couzin, 1999, In the Air that They Breathe: Lead Poisoning Remains a Major Health Hazard for America's Children. *U.S. News & World Report*, 127 (December 20), 54-57.
- Stehman, S. and W. Overton, 1996, Spatial Sampling. In S. Arlinghaus (Ed.), *Practical Handbook of Spatial Statistics* (Boca Raton, FL: CRC Press), 31-63.
- van Wijnen, J., P. Clausen, and B. Brunekreef, 1990, Estimating Soil Ingestion by Children. *Environmental Research*, 51, 147-162.