Chapter 8
The Role of Spatial Autocorrelation in Prioritizing Sites Within a Geographic Landscape

Superfund program legislation—primarily the U.S. Comprehensive Environmental Response, Compensation and Liability Act—and its public health motivations catapulted environmental contamination issues into the forefront of society's concerns. One outcome was a report by the U.S. National Research Council (NRC, 1994) examining principal methods considered or actually employed by federal and state government agencies to prioritize the remediation of hazardous waste sites. The emphasis was on between-site variation among locations, initially overlooking within-site variation for locations. The purpose of this paper is to extend more recent work on prioritizing the remediation of subregions within a given hazardous waste site, emphasizing within-site variation for locations. These extensions are illustrated with a case study of the Murray superfund site.

8.1 Introduction: The Problem

Because an enormous amount of money and people-years of effort are needed to complete the necessary environmental restoration targeted by superfund legislation, prioritizing schemes need to identify those sites in greatest need of remediation, followed by a determination of the extent to which a selected site needs to be remediated. The environmental evaluation involved consists of three stages (after NRC, 1994, p. 66): (1) identification of environmental landscapes and concomitant biomarkers indexing the risk of exposure; (2) estimation of the sources and magnitudes of contamination; and, (3) determination of appropriate remedial actions (e.g., soil removal, groundwater treatment). Heavy metal contaminants posing significant potential threats to human health, due to their known or suspected toxicity and their abundance at superfund sites, that have been identified by the USEPA's Office of Solid Waste and Emergency Response and the Agency for Toxic Substances and Disease Registry, with some being highlighted by the Centers for Disease Control (CDC, 2001), include: arsenic (As), barium (Ba), cadmium (Cd), chromium (Cr), copper (Cu), mercury (Hg), nickel (Ni), lead (Pb), and zinc (Zn). Both As and Pb are analyzed in the case study presented in this paper. As is naturally present in groundwater, and sometimes is a residue of industrial production; As is a poison that

D.A. Griffith, J.H.P. Paelinck, *Non-standard Spatial Statistics and Spatial Econometrics*, Advances in Geographic Information Science 1,
DOI 10.1007/978-3-642-16043-1_8, © Springer-Verlag Berlin Heidelberg 2011

is linked to, among other diseases, cancer and diabetes. Pb is a naturally occurring, ubiquitous element that human activities geographically concentrate in the environment far beyond its natural background level; Pb is a poison that is linked to neurological and developmental illnesses, especially in children.

The second prioritization stage involves the collection of soil, water and/or air samples—called extent of contamination samples—whose pollution contents are measured. If within-site subregions are to be identified, in order to help determine the extent to which remediation should be undertaken, then samples must be geocoded. Frequently the implemented geographic sampling design is poor, in that some subregions (e.g., hot spots—concentrations of excessively high levels of a pollutant) are oversampled while other subregions are undersampled. This outcome occurs mainly because the initial objective of sampling often is to find out which toxic materials are present, and to ascertain the site-wide extent of contamination. A subregion in which high levels of contamination are detected tends to be oversampled in order to verify the clustering of high levels. But budget constraints result in other subregions of a site being more sparsely sampled, sometimes causing their evaluations to be based upon too few samples, or even no samples when the wrong locations have been sampled.

Once measures of a contaminant have been made, the relative level of the contaminant can be established. EPA bases its exposure assessment guidelines on the upper 95% confidence limit (UCL) calculated using the mean and standard deviation of contaminant concentration computed with a site′s sample measures (Bowers et al., 1996). This criterion could suggest that a site should receive a low priority score for remediation, when in fact some subregions of the site should be assigned a high priority score. Or, this criterion could suggest that a site receive a high priority score, when not every subregion of the site is severely contaminated. Subregional assessment is further complicated by the presence of spatial autocorrelation (SA) in the sample data; nearby samples contain redundant contamination information, which in turn impacts upon the UCL that is calculated.

The research problem addressed here asks:

(1) What is the correct UCL calculation? and
(2) What method should be used to identify high priority subregions of a site?

Formulating answers to these two questions requires the use of both spatial statistics and geographic information systems (GISs). These answers are illustrated here in terms of the Murray superfund site.

8.2 The Murray Superfund Site: Part I

In all, 253 geocoded aggregated surface (0–2") soil samples—a number of nearly adjacent soil samples, whose assay results are pooled for a composite measure, and then tagged with a common georeferencing coordinate—were collected in a

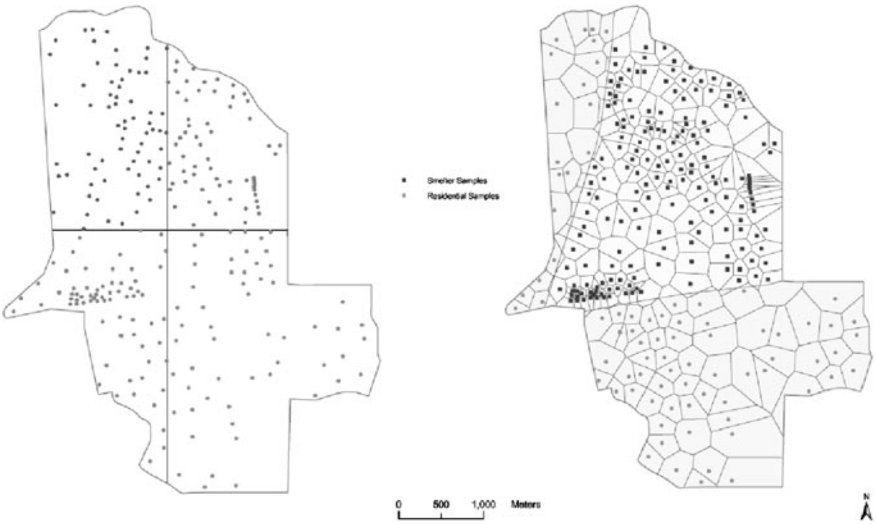

Fig. 8.1 Location of soil samples in the Murray superfund site. *Left* (**a**): division of the site into the four quadrants of the plane. *Right* (**b**): division of the site into the smelter parcel and residential neighborhoods, and the Thiessen polygon surface partitioning based upon soil sample locations

0.5 square mile area of Murray, Utah, and their concentrations of As and Pb measured. Of these, 173 were collected in an abandoned lead smelting facility superfund site, and 80 were collected in two of its adjacent residential neighborhoods located along the western and southern borders of the smelter site. Airborne emissions and placement of waste slag from the smelting process polluted this area. Sample Pb concentrations range from 37 parts per million (ppm) to 33,000 ppm. Sample As concentrations range from 5 ppm to 7,700 ppm. Besides differentiating geographic variability between the smelter site and its two adjacent neighborhoods, geographic variability also can be analyzed in terms of the four quadrants of the plane, which in counter-clockwise rotation respectively contain 63, 57, 68 and 65 soil sample locations. The geographic configuration of the sample points can be articulated with Thiessen polygons. These various features of this geographic landscape are portrayed in Fig. 8.1.

8.2.1 State-of-the-Art Practice

A considerable amount of effort has been devoted to handling the log-normal nature of most contamination measures—transforming a set of contamination measures by replacing them with their logarithm values results in a sample that more closely mimics a normal frequency distribution. The key analytical benefit here is reducing specification error attributable to wrongly assuming a normal distribution probability model for inferential purposes, one that does not characterize the raw data. The key communication complication here is the ability to discuss the UCL, which is based upon the normal probability model, in terms of the original measurements. Consequently, substantial effort has been expended on how to calculate accurate back-transformations (see Armstrong, 1992; Bowers et al., 1996). But whether the UCL is expressed in logarithm or raw-data measurement terms, it is severely limited when its calculation fails to accommodate SA that is latent in data.

In recognizing geographic pattern, several studies promote the use of spatial analysis for identifying high priority subregions of a contaminated site. Ginevan and Splitstone (1997) outline how kriging can be used to generalize a contamination surface from a set of sample points. Burmaster and Thompson (1997) outline the use of Thiessen polygons, with specific reference to incorporating spatial pattern of contamination into the UCL calculation; more specifically, they calculate a weighted average whose weights are the inverse areas of the Thiessen polygons.

The state-of-the-art practice illustrated by these researchers is to exploit SA in order to construct generalized contour maps, but otherwise to overlook SA, although not necessarily outcomes of the geographic configuration of sample data, in order to calculate the UCL. The methodology outlined in this paper corrects this second deficiency, incorporating SA into the UCL calculation through the use of a spatial simultaneous autoregressive (SAR) model specification, marries it to kriging based upon a semivariogram model that is consistent with the SAR model, and extends assessment to a bivariate situation. This extension satisfies Burmaster and Thompson's (1997) requirement of preserving the individual spatial patterns of, as well as the correlation between, two contaminant concentrations.

8.2.2 A Spatial Methodology: Stage 1, Spatial Sampling Data Collection and Preprocessing

The spatial methodology involves steps ranging from sample selection to identification of remediation regions. Sampling should be undertaken with two goals in mind. First, a site needs to be adequately covered. Second, pollution hot spots need to be verified. Stehman and Overton (1996) outline how to implement a hexagonal tessellation stratified random sample. This design ensures adequate coverage across a study site. It suggests that the first nearest neighbor statistic should be around 2, indicating a strong tendency for sample locations to be uniformly spaced; random selection within a hexagon avoids the sample being geographically systematic, and prevents this statistic from equaling its maximum (approximately 2.14). Often regions surrounding sample locations revealing high levels of a pollutant then are intensively sampled, in order to verify the existence of a hot spot. This second stage of the sampling process will further reduce the nearest neighbor statistic. Both of these stages would be well served by a model-informed sampling strategy that involves estimation of the nature and degree of latent SA in the geographic distribution of the pollutant. As sample intensity increases, SA tends to increase. As SA increases, total sample size should decrease, in order to minimize the collection of redundant information. An equilibrium between these two opposing trends is desirable.

The second step is to identify a variable transformation that converts the pollution measures into values that closely mimic a bell-shaped frequency distribution. Most all sample pollution measures exhibit a log-normal type of 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 to describe pollution measures well because they are bounded below at 0 and usually are strongly positively skewed. But a heavy metal such as Pb occurs naturally in all soils, implying that 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 Pb emissions dispersing from the smoke stack of a smelter. Relatively small amounts are deposited in most locations, while increasingly larger amounts are deposited in fewer and fewer locations (perhaps near the smoke stacks). If the process depositing pollution is repetitive, then with some stochastic fluctuation (e.g., wind pattern change), each layer of pollution has approximately 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 a log-normal distribution, and a transformed variable of the form

$$LN\,(\text{pollution concentration measure } + \delta)\,, \tag{8.1}$$

where *LN* denotes the natural logarithm, and δ is a translation parameter at least accounting for the naturally occurring background level of a pollutant.

Often real-world data, especially if they are georeferenced, contain considerable heterogeneity. This heterogeneity frequently is related to the magnitude of a measure. Equation (8.1) is equivalent to a Box-Cox power transformation with an exponent of zero. This zero exponent transforms positively skewed frequency distributions into ones that are more symmetric; it moves the left-hand frequency bump to the right, and squashes this bump downward, which forces the two tails to inflate. With regard to the raw measures, relatively speaking, this transformation shrinks very large values, magnifies very small values, and preserves intermediate values. Including the translation parameter, δ, primarily impacts upon one or both tails, modifying their inflation so that it better corresponds to that of a bell-shaped curve. At least some additional data heterogeneity can be accounted for by allowing δ to vary by the size of measures, or

$$LN\left[\text{pollution concentration measure} + \delta_0 + \delta_1 \left(\frac{\mathrm{r}_1}{\mathrm{n}+1}\right)^{\gamma_1} \left(\frac{\mathrm{r}_2}{\mathrm{n}+1}\right)^{\gamma_2}\right], \tag{8.2}$$

where r_1 and r_2 respectively are the ascending and descending rankings of the n pollution concentration measures, δ_0 is a translation parameter constant, δ_1 is a constant of proportionality, and γ_1 and γ_2 are exponents attached to the relative rankings. Equation (8.1) is the special case of $\delta_1 = 0$. The nonconstant translation parameter should have values contained within the range of the data, and should result in a closer alignment of the empirical and theoretical cumulative frequency distributions basically by stretching one or both of the tails of the empirical distribution. Additional heterogeneity can be accounted for by allowing the exponent to vary by the size of measures, or

$$\left[\text{pollution concentration measure} + \delta_0 + \delta_1 \left(\frac{r_1}{n+1}\right)^{\gamma_1} \left(\frac{r_2}{n+1}\right)^{\gamma_2}\right]^{\delta_2+\delta_3\left(\frac{r_1}{n+1}\right)^{\gamma_3}\left(\frac{r_2}{n+1}\right)^{\gamma_4}}, \tag{8.3}$$

where the terms of $\delta_2 + \delta_3 \left(\frac{r_1}{n+1}\right)^{\gamma_3} \left(\frac{r_2}{n+1}\right)^{\gamma_4}$ are defined in a similar fashion to those for Eq. (8.2). Equation (8.3) will tend to better align both the tails as well as the center of the empirical frequency distribution. Equations (8.1) and (8.2) are special case of $\delta_2 = \delta_3 = 0$. Equation (8.3) could have $\delta_1 = 0$, hence capturing heterogeneity solely with a nonconstant exponent. Equation (8.3) is suggested when the translation parameter values of Eqs. (8.1) and/or (8.2) fall outside the interval $(-y_{min}, y_{max})$, where y_{min} and y_{max} denote the extreme values of Y.

The third step is to krig values—spatial interpolation—and to produce the necessary quantities to calculate UCLs. Statistical analyses engaged in during this step should be nearly void of specification error, given the accommodation of assumptions of normality, constant variance, and observation independence. For remediation purposes, the important consideration is avoiding specification error. But for communication purposes, the important consideration is expressing decision criteria in understandable quantitative terms. Hence, this is the step in which a back-transformation could be calculated for communication purposes. The fourth, and final, step is to demarcate remediation subregions of a site. These third and fourth steps are spelled out in more detail in the ensuing sections of this paper.

8.3 The Murray Superfund Site: Part II

Griffith (2002b) reports a 1st nearest neighbor statistic of 0.06208 for the Murray site, indicating that the sample locations are highly clustered. Visual inspection of the maps in Fig. 8.1 suggests subregions that are under- or unsampled, subregions that are oversampled, and some apparent sampling transect.

Equation (8.2) was calibrated for both As and Pb. Normal distribution quantile plots appear in Fig. 8.2 and show the evolution of the transformed values. Both pollutants begin with the hooked quantile plot typifying untransformed log-normal data, and achieve their greatest conformity gains merely by being subjected to a simple logarithmic transformation. Inclusion of a constant translation parameter primarily better aligns the lower tails of the empirical cumulative frequency distributions with their theoretical normal cumulative frequency distribution counterpart. Capturing heterogeneity by letting the translation parameter vary with data value order ranking essentially aligns all but the largest two As values, and all of the

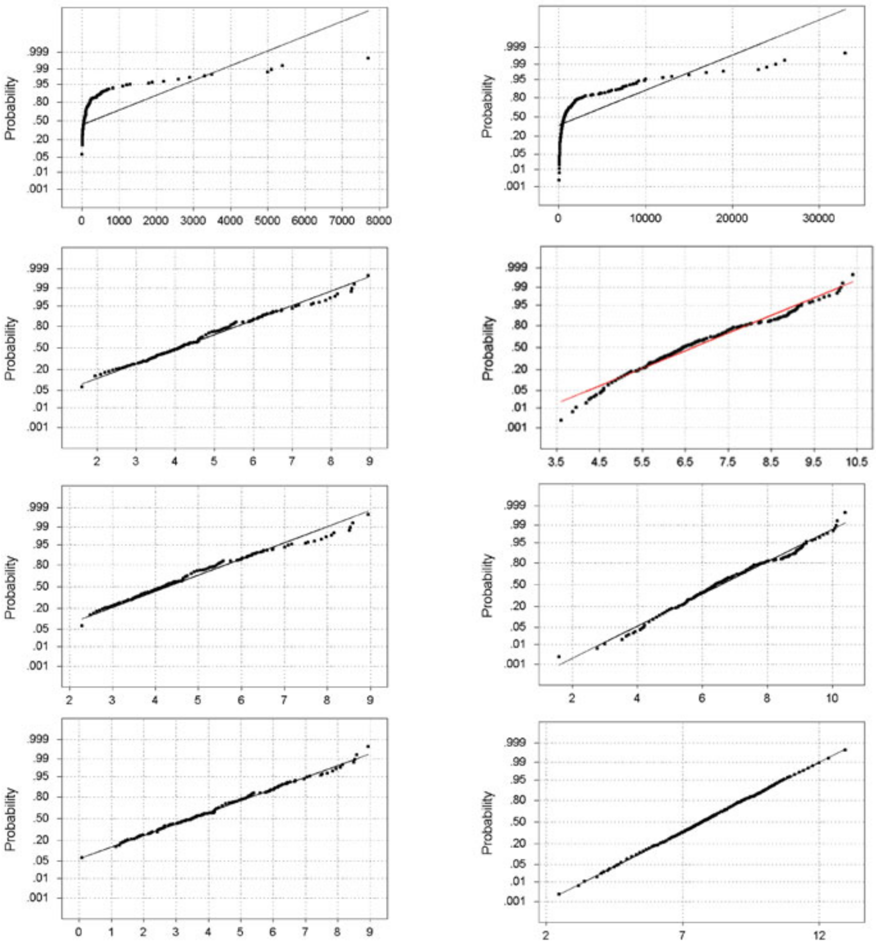

Fig. 8.2 Evolution of the pollutant data transformation quantile plots. *Left* (**a**): from top to bottom, raw As values, logarithmic As values, Eq. (8.1) As values, and Eq. (8.2) As values. *Right* (**b**): from top to bottom, raw Pb values, logarithmic Pb values, Eq. (8.1) Pb values, and Eq. (8.2) Pb values

Pb values. These results are corroborated by boxplots for these sequential transformations, which appear in Fig. 8.3. These latter graphics reveal that the frequency distribution bumps spread out from lower values toward high values, the highest values shrink toward the lower values, and improved symmetry emerges. Of note is that the As analysis is complicated by the presence of 32 measures occurring at the detection level of 5 ppm.

Geographic distributions of the relative transformation effects appear in Fig. 8.4. Both for As and Pb, conspicuous clusters of raw values are replaced by swaths of relatively high values that, for the most part, differentiate between the smelter site and the residential neighborhoods. Again, little difference is visually detectable

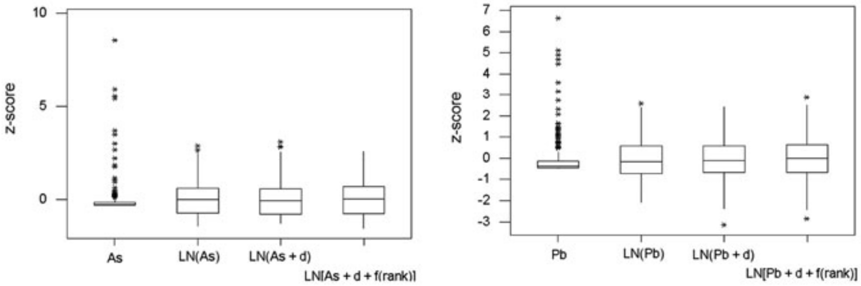

Fig. 8.3 evolution of the pollutant data transformation box plots: *top* (**a**): as results. *bottom* (**b**): pb results

between the application of a simple logarithm transformation and Eqs. (8.1) and (8.2). A Shapiro-Wilk (S-W) statistic indexing of conformity of these measures with a normal frequency distribution appears in Table 8.1; the null hypothesis value for S-W is 1. Each transformation increases S-W, with the largest increase attained by applying the simple logarithm transformation.

A quantification of geographic variability heterogeneity is summarized in Table 8.1. Homogeneity of variance for the various As and Pb measurement scales is evaluated with Bartlett's and Levene's (i.e., a non-normailty assuming diagnostic statistic used to assess the equality of variance in different samples) test statistics for equality of variance; each has a null hypothesis value of 0. These assessments are in

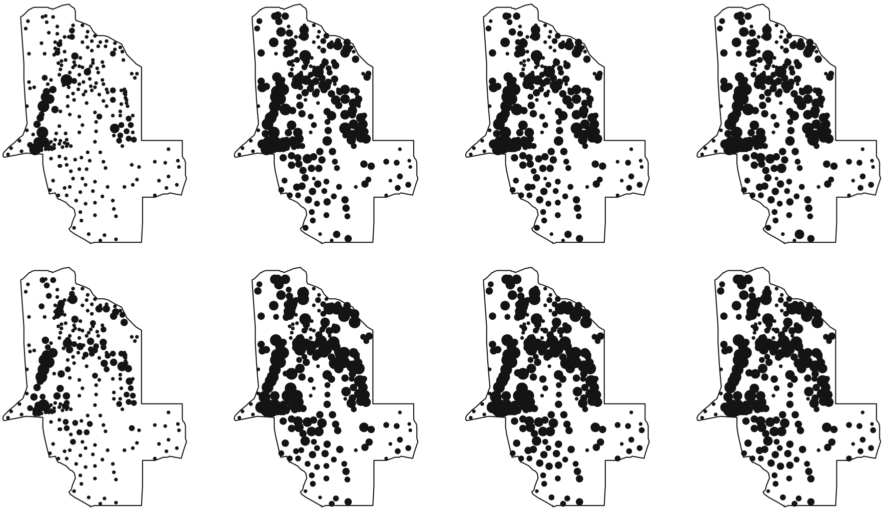

Fig. 8.4 Evolution of the pollutant data transformation relative values (i.e., proportional circles) maps. *Top* (**a**): from left to right, raw As values, logarithmic As values, Eq. (8.1) As values, and Eq. (8.2) As values. *Bottom* (**b**): from left to right, raw Pb values, logarithmic Pb values, Eq. (8.1) Pb values, and Eq. (8.2) Pb values

Table 8.1 Sequential construction of the variable transformations

	As ($5 \leq$ As $\leq 7{,}700$)				Pb ($37 \leq$ Pb $\leq 33{,}000$)		
Variable	Bartlett	Levene	S-W	R	Bartlett	Levene	S-W
Raw y	221.660***	9.017***	0.348***	0.589	34.161***	16.192***	0.523***
	221.161***	2.785***			168.109***	6.316***	
LN(y)	2.731***	23.600***	0.964***	0.740	3.005***	30.413***	0.976***
	9.011**	2.871**			28.067***	10.408***	
LN(y+δ)	3.201***	27.837***	0.970***	0.739	2.887***	26.912***	0.990*
	10.602**	2.607*			30.298***	10.297***	
LN[y+δ+f(r)]	2.465***	22.912***	0.972***	0.748	2.521***	19.819***	0.999
	7.860**	3.024**			28.361***	9.205***	
SAR residuals	1.842***	12.561***	0.995	0.706	2.236***	20.022***	0.995
	13.280***	6.098***			31.520***	13.304***	
Filter residuals	1.404	5.050**	0.996	0.688	1.442**	8.065***	0.991
	11.414**	3.550**			29.165***	9.702***	

NOTE 1: ***, **, * denote a significant difference from the null hypothesis value (0 for the Bartlett and Levene, and 1 for the S-W statistics) at, respectively, the 1, 5% and 10% level

NOTE 2: the first row Bartlett and Levene statistics test variance differences between the smelter site and neighboring residential neighborhoods

NOTE 3: the second row Bartlett and Levene statistics test variance differences between the four quadrants of the plane

NOTE 4: no evidence was found to support the presence of a heterogeneous transformation exponent

terms of both the smelter site/residential regions and the four quadrants of the plane (see Fig. 8.1). The Eq. (8.2) values display considerably less heterogeneity than do the raw values, and basically less than the simple logarithmically transformed values. But nonconstant geographic variance does persist, even though its magnitude is substantially less.

8.3.1 A Spatial Methodology: Stage 2, Spatial Statistics for Calculating UCLs

A spatial SAR model was fitted to the transformed data. A suitable surface tessellation for this purpose can be constructed using Thiessen polygons (see Fig. 8.1). The configuration of points depicted by this surface partitioning can be represented by a standard binary 0–1 geographic weights matrix, say **C**, where $c_{ij} = 1$ if two distinct points i and j share a Thiessen polygon boundary, and $c_{ij} = 0$ otherwise. The SAR model results allow the SA adjusted calculation of a mean, a standard error, and a t-statistic.

Calculation of a UCL requires an estimate of the mean, an estimate of the variance, and the number of degrees of freedom. The simplest, pure SAR model may be written as

$$\mathbf{Y} = \mu\,(1-\rho)\,\mathbf{1} + \rho\mathbf{W}\mathbf{Y} + \boldsymbol{\varepsilon}\,, \tag{8.4}$$

where $\mathbf{Y}$ is an n-by-1 vector of georeferenced values, $\mathbf{1}$ is an n-by-1 vector of ones, $\mathbf{W}$ is the row-standardized version of matrix $\mathbf{C}$, μ is the mean of Y, ρ is a SA parameter, $\mu(1–\rho)$ is the mean of $(\mathbf{Y} – \rho\mathbf{WY})$, and $\boldsymbol{\varepsilon}$ is an n-by-1 independent and normally distributed, constant variance random error vector. An estimate of the mean, corrected for the presence of SA, is given by $\hat{\mu}$ obtained with Eq. (8.4), which actually is the conditional mean of Y given $\mathbf{W}_i\mathbf{Y}$ (the average of surrounding values of Y for observation i). This interpretation is based upon two features of Eq. (8.4). First, if $\rho = 0$, then SA is absent and μ is calculated with independent observation values. Second, if $\mathbf{W}_i\mathbf{Y} = 0$, then the average of the surrounding values is 0. Although this second interpretation is weakened when 0 lies outside the range of the data, conceptually it is sensible; here the transformed As minimum is close to 0, equaling 0.1, while the transformed Pb minimum of 2.5 relates to the minimum value inflated by two-thirds via the translation parameter. While gathering additional sample data that include 0 would strengthen this latter interpretation of $\hat{\mu}$, such a data collection exercise often is impractical, if not impossible.

Meanwhile, the variance estimate corrected for the presence of SA is given by

$$\hat{\sigma}^2 = (\mathbf{Y} - \hat{\mu}\mathbf{1})^{\mathrm{T}}(\mathbf{I} - \hat{\rho}\mathbf{W})^{\mathrm{T}}(\mathbf{I} - \hat{\rho}\mathbf{W})(\mathbf{Y} - \hat{\mu}\mathbf{1})/(n - 2), \tag{8.5}$$

where $\mathbf{I}$ denotes the identity matrix, T denotes the matrix transpose operation, and division is by (n–2) because both μ and ρ are estimated. Because positive SA inflates the variance, the quantity yielded by Eq. (8.5) will tend to be less than its conventional counterpart of $s^2 = (\mathbf{Y} - \hat{\mu}\mathbf{1})^{\mathrm{T}}(\mathbf{Y} - \hat{\mu}\mathbf{1})/(n - 1)$; the variance inflation factor here is given by $TR\{[(\mathbf{I} - \hat{\rho}\mathbf{W})^{\mathrm{T}}(\mathbf{I} - \hat{\rho}\mathbf{W})]^{-1}\}/n$, where TR denotes the matrix trace operator. This variance inflation plays a critical role in determining the effective sample size—the number of independent observations to which a set of georeferenced observations are equivalent—say n*. In the presence of SA, as the distance between two sample locations decreases, their respective attribute values become increasingly similar, and their information content becomes increasingly redundant. Overlooking this redundant information introduces specification error into a data analysis. The purpose of calculating quantities like Eq. (8.5), using equations like (8.4), is to adjust for or remove impacts of the redundant information.

Next, consider the variance of the sampling distribution of the sample mean of variable Y, $\bar{y}$, when the variance of Y is unknown, which is given by

$$\{\mathbf{1}^{\mathrm{T}}[(\mathbf{I} - \hat{\rho}\mathbf{W})^{\mathrm{T}}(\mathbf{I} - \hat{\rho}\mathbf{W})]^{-1}\mathbf{1}/n\}\hat{\sigma}^2/n, \tag{8.6}$$

and which reduces to the conventional $\hat{\sigma}^2/n$ when $\rho = 0$. Rewriting Eq. (8.6) in terms of s^2 renders the following estimate of effective sample size;

$$n^* = n\, TR\{[(\mathbf{I}–\hat{\rho}\mathbf{W})^{\mathrm{T}}(I–\hat{\rho}\mathbf{W})]^{-1}\} /\mathbf{1}^{\mathrm{T}}[(I - \hat{\rho}\mathbf{W})^{\mathrm{T}}(I - \hat{\rho}\mathbf{W})]^{-1}\mathbf{1}, \tag{8.7}$$

which reduces to n when $\rho = 0$, and asymptotically converges on 1 as ρ approaches 1 (see Griffith and Zhang, 1999). Equation (8.7) allows determination of the appropriate t-statistic, which has n*–2 degrees of freedom.

Finally, normal curve theory states that the 95% UCL is given by

$$\bar{y} + t_{n-1,0.95}\frac{s}{\sqrt{n}}$$

which here translates into

$$\mathbf{1}^T\mathbf{Y}/n + \mathbf{t}_{n^*-2,0.95}\left(\{\mathbf{1}^T[(\mathbf{I}-\hat{\rho}\mathbf{W})^T(\mathbf{I}-\hat{\rho}\mathbf{W})]^{-1}1/n\}\times [(\mathbf{Y}-\hat{\mu}\mathbf{1})^T(\mathbf{I}-\hat{\rho}\mathbf{W})^T(\mathbf{I}-\hat{\rho}\mathbf{W})(\mathbf{Y}-\hat{\mu}\mathbf{1})/(n-2)]\right)^{1/2}/\sqrt{n},$$

or

$$\mathbf{1}^T\mathbf{Y}/n + \mathbf{t}_{n^*-2,0.95}\left(TR\{[[(\mathbf{I}-\hat{\rho}\mathbf{W})^T(\mathbf{I}-\hat{\rho}W)]^{-1}\mathbf{1}/n\}\times [(\mathbf{Y}-\hat{\mu}\mathbf{1})^T(\mathbf{I}-\hat{\rho}\mathbf{W})^T(\mathbf{I}-\hat{\rho}\mathbf{W})(\mathbf{Y}-\hat{\mu}\mathbf{1})/(n-2)]\right)^{1/2}/\sqrt{n^*}$$

As an aside, $\mathbf{1}^T[(\mathbf{I}-\hat{\rho}\mathbf{W})^T(\mathbf{I}-\hat{\rho}\mathbf{W})]^{-1}\mathbf{1}/n\} \approx e^{0.95\hat{\rho}/(1-0.91\hat{\rho})}$, $0 \le \hat{\rho} < 1$, which allows a quick, easier calculation of these expressions. Ignoring impacts of SA on the sampling distribution of $\bar{y}$ results in use of the incorrect expression

$$\mathbf{1}^T\mathbf{Y}/n + \mathbf{t}_{n-2,0.95}\left(TR\{[[(\mathbf{I}-\hat{\rho}\mathbf{W})^T(\mathbf{I}-\hat{\rho}\mathbf{W})]^{-1}\mathbf{1}/n\}\times ([[(\mathbf{Y}-\hat{\mu}\mathbf{1})^T(\mathbf{I}-\hat{\rho}\mathbf{W})^T(\mathbf{I}-\hat{\rho}\mathbf{W})(\mathbf{Y}-\hat{\mu}\mathbf{1})/(n-2)]\right)^{1/2}/\sqrt{n}.$$

This first expression renders UCL boundary values greater than or equal to (when $\rho = 0$) those calculated with this second expression. These are the equations used to calculate entries in Table 8.3.

8.4 The Murray Superfund Site: Part III

Results of fitting Eq. (8.4) to both the As and the Pb data appear in Table 8.2. In both cases a moderate level of positive SA is detected, with roughly a fifth of the variance in variable Y accounted for by variable $\mathbf{W}_i\mathbf{Y}$. Residual normality and variance homogeneity results appear in Table 8.1, and show close conformity with a normal distribution, but with the continued persistence of nonconstant geographic variance. The traditional predicted-versus-residual plots appear in Fig. 8.5, and suggest that, other than for the As = 5 complication, conspicuous deviations from conventional variance homogeneity are absent.

Table 8.2 Simultaneous autoregressive (SAR) modelling results

As				Pb			
		Residual				Residual	
$\hat{\rho}$	Adj-R^2	MC	GR	$\hat{\rho}$	Adj-R^2MC	GR	
0.532	0.245	–0.030	1.045	0.494	0.208	–0.034	1.085

Table 8.3 Quantities used to calculate, and the resulting, UCLs

	As		Pb	
Statistic	Uncorrected	Corrected	Uncorrected	Corrected
$\hat{\mu}$	3.46316	3.46316	7.69417	7.69417
standard error of $\hat{\mu}$	0.13344	0.25459	0.11539	0.20791
n*	253	68.9	253	77.6
Df	252	66.9	252	75.6
t-statistic for 0.95 level	1.6509	1.6680	1.6509	1.6653
UCL	3.68346	3.88782	7.88467	8.04040

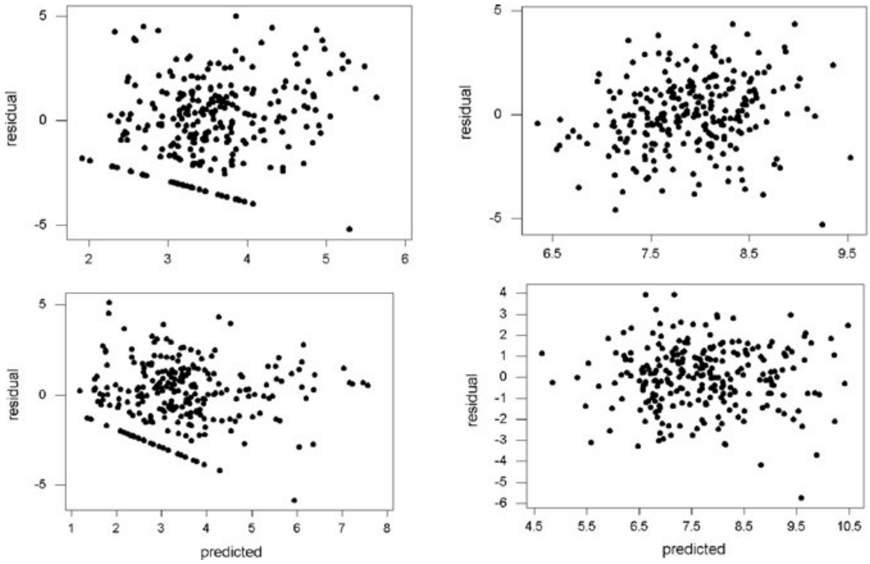

Fig. 8.5 Conventional homogeneity of variance scatter plots. *Left* (**a**): top, for spatial SAR model describing As; bottom, for spatial filter model describing As. *Right* (**a**): top, for spatial SAR model describing Pb; bottom, for spatial filter model describing Pb

The UCL results appear in Table 8.3. Variance inflation results in both the uncorrected means and their standard errors as larger than they should be, consequences that are compensatory to some degree since the mean is divided by the standard error. The presence of a moderate degree of positive SA results in effective sample sizes that are less than a third of n. This result has only a very modest impact upon the correct t-statistic, though, partially because a t-statistic converges upon a normal variate z-score as n goes to infinity; the only marked discrepancies are for values of n or n* very close to 1. The overall outcome is a UCL that expands by 2–6%. In other words, some subregions of the Murray superfund site would be misclassified as not being high priority remediation locations when in fact they are.

Geographic impacts of the changes in these UCLs include a shrinkage in area by about 8.3% of the As, and by about 14.2% of the Pb, subregions that qualify for remediation in the site. When SA is overlooked, roughly 38.5% of the Murray superfund site qualifies for remediation of As contamination, whereas roughly 35.9% of the site qualifies for remediation of Pb contamination. Respectively each of these percentages decreases to 35.3% and 30.8% once SA effects are taken into account. The marginal areas vulnerable to misclassification are located along the borders of the subregions identified with classical statistics.

8.4.1 A Spatial Methodology: Stage 3, Prioritizing Subregions for Remediation

The third step of the spatial methodology is to krig values produced by the most appropriate transformation equation [i.e., (8.1), (8.2), or (8.3)]. The semivariogram model selected for this spatial interpolation exercise needs to be consistent with the model selected for the spatial autoregressive analysis. Griffith and Layne (1999) argue that the SAR and Bessel function geostatistical semivariogram models conceptually and numerically are closely linked. This pair of models is used here to krig the As and Pb surfaces, and to compute the As and Pb UCLs.

The following Bessel function semivariogram model [Eq. (3.8), Sect. 3.2] was used to interpolate both the As and the Pb surfaces; the effective range is approximately 4r, where r denotes the range parameter. The graph of Eq. (3.8) displays a cusp in the neighborhood of $\overline{\mathrm{d}} = 0$, a characteristic of a second-order SA mechanism that also is captured by the spatial SAR model. Equation (3.8) is used to estimate the covariance (say, using matrix notation, $\mathbf{S}_{\mathrm{om}}$) between sample point pollutant measures and unsampled point pollutant measures, which are the ones to be interpolated. The m interpolated values are given by

$$\hat{\mathbf{Y}}_{\mathrm{m}} = \hat{\mu}\mathbf{1}_{\mathrm{m}} + \mathbf{S}_{\mathrm{mo}}^{\mathrm{T}}\mathbf{S}_{\mathrm{oo}}^{-1}(\mathbf{Y}_{\mathrm{o}} - \hat{\mu}\mathbf{1}_{\mathrm{o}}), \tag{8.8}$$

where the subscript m denotes values to be interpolated, the subscript o denotes observed sample values, and $\mathbf{S}_{\mathrm{oo}}$ denotes the variance-covariance matrix for observed sample values, the measures to which Eq. (3.8) is fitted (see Griffith and Layne, 1999). In effect, Eq. (8.8) spreads the information content in a sample over a map, much like spreading icing over the top of a cake. If SA does not exist in variable Y, then $\mathbf{S}_{\mathrm{oo}} = \mathbf{I}, \mathbf{S}_{\mathrm{om}} = \mathbf{0}$, and $\hat{\mathbf{Y}}_{\mathrm{m}} = \hat{\mu}\mathbf{1}_{\mathrm{m}}$; the conventional maximum likelihood estimate of a univariate missing value is the mean of the observed values.

8.5 The Murray Superfund Site: Part IV

Restricting attention to point pairs within a 1000-foot radius, Eq. (3.8) estimation results are as follows:

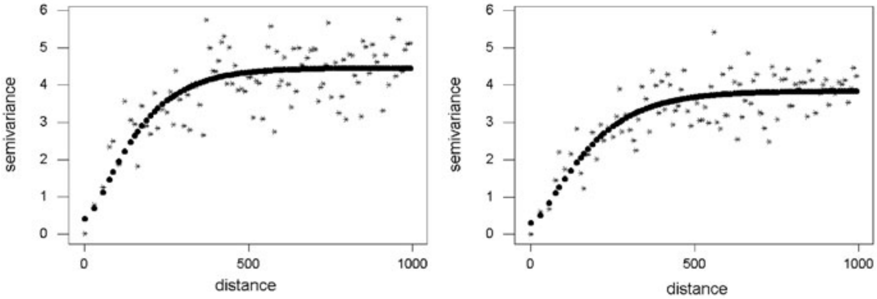

Fig. 8.6 Observed and Bessel function fitted semivariogram plots. *Top* (**a**): As results. *Bottom* (**b**): Pb results

$$\text{As: } \gamma(\bar{\mathrm{d}}) = 0.4119 + 4.0426\left[1 - \left(\frac{\bar{\mathrm{d}}}{109}\right)\mathrm{K}_1\left(\frac{\bar{\mathrm{d}}}{109}\right)\right], \text{RESS} = 0.448,$$

and

$$\text{Pb:}\gamma(\bar{\mathrm{d}}) = 0.2995 + 3.5355\left[1 - \left(\frac{\bar{\mathrm{d}}}{122}\right)\mathrm{K}_1\left(\frac{\bar{\mathrm{d}}}{122}\right)\right], \text{RESS} = 0.381,$$

where RESS denotes the relative error sum of squares (the error sum of squares divided by the total sum of squares adjusted for the mean). The semivariogram plots for these two situations appear in Fig. 8.6.

The fourth, and sometimes final, spatial methodology step is to demarcate remediation subregions of a site using both the kriged surface and the UCL of the adjusted mean. While a back-transformation can be used to compute the UCL in terms of the original pollutant measurement scale, the mapping exercise can and should retain precision by being done in the transformed variable space. The UCLs reported in Table 8.3 have been applied to the interpolation results based upon Eqs. (3.8) and (8.8). As is expected, the visually detectable swaths appearing in Fig. 8.4 reflect the high priority remediation subregions demarcated in Fig. 8.7. About half of the smelter site is ranked as high priority for remediation, as is much of the immediately adjacent residential neighborhoods, both for As and for Pb. Common to these subregions is a large portion of the western residential neighborhood, the southwest quadrant of the smelter site, and the northwester corner of the southern residential neighborhood. An additional feature of the remediation maps is the scattered set of isolated point UCLs. These locations signify subregions that are prime candidates for subsequent intensive sampling, but only when they are based upon the UCL adjusted for SA.

8.5.1 A Spatial Methodology: Stage 4, Covariation of Contaminants and Joint Pollutant Analyses

When contamination by more than a single pollutant is of concern, several additional aspects of the remediation prioritizing task arise. Foremost are covariations among

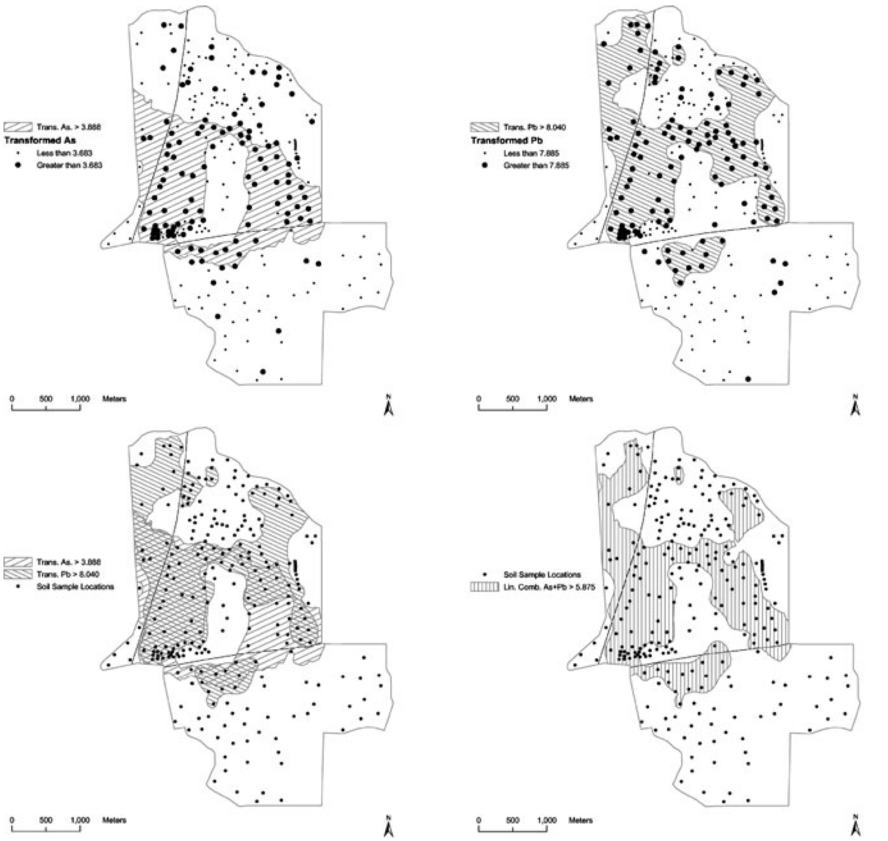

Fig. 8.7 Remediation subregions based on 95% UCLs. *Top, left* (**a**): As results. *Top, right* (**b**): Pb results. *Bottom, left* (**c**): map overlay of As and Pb results. *Bottom, right* (**d**): joint As and Pb results

pollutants. In a bivariate case, the focus is on correlation between the two pollutants as well as the SA contained in each pollutant.

Linear correlation measures are impacted upon by the log-normal nature of pollution data. Hence, correlations calculated with raw data values often do not accurately capture actual covariations. The more informative correlations are those calculated with Box-Cox transformed data values.

Meanwhile, SA also can disguise attribute covariations. Removing SA, either by dealing with the residuals of an SAR model, or the residuals from a spatial filter model, corrects for spatial dependency effects. Spatial filtering can be based upon the eigenfunctions of the numerator of the Moran Coefficient (MC) index of SA (see Griffith, 2000a), given by expression (5.8) [see Sect. 5.5]. Tiefelsdorf and Boots (1995) show that all of the eigenvalues of matrix expression (5.8) relate to specific MC values. The eigenvectors of expression (5.8) may be treated as synthetic variates, and interpreted in the context of map pattern as described in Sect. 6.2. Hence these n eigenvectors describe the full range of all possible mutually orthogonal map patterns, and may be interpreted as synthetic map variables. In the presence of positive SA, then, an analysis can employ those eigenvectors depicting map patterns exhibiting consequential levels of positive SA; operationally speaking, attention can be restricted to eigenvectors having MC ≥ 0.25, say.

One appealing property of expression (5.8) is that matrix **C** is constant for a given surface partitioning and adjacency definition, rendering the same set of eigenvectors for all attributes geographically distributed across a given surface partitioning. Another is that the eigenvectors can be used in a conventional, ordinary least squares regression analysis to account for SA. In other words,

$$\mathbf{Y} = \alpha_{\mathrm{Y}}\mathbf{1} + \mathbf{E}_{\mathrm{k}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\mathrm{Y}}, \tag{8.9a}$$

where k denotes the subset of eigenvectors that accounts for the SA contained in variable Y, α_{Y} is the conditional mean intercept term for variable Y, and $\boldsymbol{\varepsilon}_{\mathrm{Y}}$ is an independent random error term associated with variable Y. The correlation coefficient corrected for spatial dependency effects is calculated between $\boldsymbol{\varepsilon}_{\mathrm{X}}$ and $\boldsymbol{\varepsilon}_{\mathrm{Y}}$, where

$$\mathbf{X} = \alpha_{\mathrm{X}}\mathbf{1} + \mathbf{E}_{\mathrm{h}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\mathrm{X}}, \tag{8.9b}$$

and the terms in Eq. (8.9b) are defined like those in Eq. (8.9a), but with regard to X. The subset of eigenvectors contained in $\mathbf{E}_{\mathrm{h}}$ and $\mathbf{E}_{\mathrm{k}}$ most likely will not be the same. Any common eigenvectors will tend to inflate the linear correlation between X and Y; any non-common eigenvectors will tend to deflate this correlation. Of note is that these eigenvectors capture the separate X and Y map patterns that Burmaster and Thompson require to be preserved.

Finally, the joint treatment of X and Y require adjustments to the individual UCLs. Now two sources of redundant information exist: correlation between variables, and SA within each variable. Dutilleul (1993) updates the Richardson-Clifford discussion about how SA impacts upon the correlation coefficient. Extending his discussion reveals that covariation also has a variance inflation factor similar to that presented in Eq. (8.6), with this factor largely being compensated for by the individual variable variance inflation factors when a correlation coefficient is computed. Moreover, spatial dependency impacts upon a correlation coefficient increase as the correlation moves away from zero, and decrease again as the correlation approaches ±1. If the correlation between X and Y is zero, then Eqs. (8.9a) and (8.9b) would contain no common eigenvectors; if the correlation between X and Y is ±1, then Eqs. (8.9a) and (8.9b) would contain exactly the same set of eigenvectors. Meanwhile, constructing a weighted average of X and Y, say [wX + (1–w)Y] for $0 \leq \mathrm{w} \leq 1$, yields as the sampling distribution variance for $\mathrm{w}\bar{\mathrm{x}} + (1 - \mathrm{w})\bar{\mathrm{y}}$

$$\frac{\mathrm{w}^2\sigma_{\mathrm{X}}^2 + (1 - \mathrm{w})^2\sigma_{\mathrm{Y}}^2 + 2\mathrm{w}(1 - \mathrm{w})\rho_{\mathrm{XY}}\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}}{\mathrm{n}},$$

where σ^2_X and σ^2_Y respectively denote the variance of variables X and Y, ρ_{XY} denotes the product moment correlation between variables X and Y, and the term $2w(1-w)\rho_{XY}\sigma_X\sigma_Y$ adjusts for the presence of redundant attribute information in the bivariate georeferenced dataset.

In this bivariate case, effective sample size becomes a weighted average of the individual pollutant effective sample sizes that is adjusted for the correlation between X and Y. The numerator of Eq. (8.7) becomes

$$\frac{w^2\hat{\sigma}^2_X + (1-w)^2\hat{\sigma}^2_Y + 2w(1-w)\hat{\rho}_{XY}\hat{\sigma}_X\hat{\sigma}_Y}{w^2\hat{\sigma}^2_X + (1-w)^2\hat{\sigma}^2_Y}, \tag{8.10a}$$

times

$$w^2\hat{\sigma}^2_X TR\{[(\mathbf{I}-\hat{\rho}_X\mathbf{W})^T(\mathbf{I}-\hat{\rho}_X\mathbf{W})]^{-1}\} + (1-w)^2\hat{\sigma}^2_Y\{[(\mathbf{I}-\hat{\rho}_Y\mathbf{W})^T(\mathbf{I}-\hat{\rho}_Y\mathbf{W})]^{-1}\}, \tag{8.10b}$$

and the denominator of Eq. (8.7) becomes

$$w^2\hat{\sigma}^2_X\mathbf{1}^T[(\mathbf{I}-\hat{\rho}_X\mathbf{W})^T(\mathbf{I}-\hat{\rho}_X\mathbf{W})]^{-1}\mathbf{1} + (\mathbf{1}-w)^2\hat{\sigma}^2_Y\mathbf{1}^T[(\mathbf{I}-\hat{\rho}_Y\mathbf{W})^T(\mathbf{I}-\hat{\rho}_Y\mathbf{W})]^{-1}\mathbf{1}$$
$$+ 2w(1-w)\hat{\rho}_{XY}\hat{\sigma}_X\hat{\sigma}_Y\mathbf{1}^T[(I-\hat{\rho}_X\mathbf{W})^T(I-\hat{\rho}_Y\mathbf{W})]^{-1}\mathbf{1}. \tag{8.10c}$$

where $\hat{\rho}_X$ is the SA parameter estimate for variable X, $\hat{\rho}_Y$ is the SA parameter estimate for variable Y, and the sample statistics are $s^2_X = \hat{\sigma}^2_X$, $s^2_Y = \hat{\sigma}^2_Y$, and $r_{XY} = \hat{\rho}_{XY}$. Therefore, n^* equals n times expression (8.10a) times expression (8.10b) divided by expression (8.10c). If $\hat{\rho}_X = \hat{\rho}_Y = 0$, then this product of the three expressions reduces to n. If $w = 0$, $w = 1$ or $\hat{\rho}_X = \hat{\rho}_Y$, then this product of the three expressions reduces to Eq. (8.7). In other words, the bivariate effective sample size is a weighted average of the individual univariate effective sample sizes (i.e., it must be contained in the interval defined by them). And, as $\hat{\rho}_X$ and $\hat{\rho}_Y$approach 1, n^* approaches 1. The weighting is determined by both the relative variances of X and Y and the weights used in constructing a linear combination of X and Y, and is impacted little by the value of r_{XY}.

Therefore, Eqs. (8.9a) and (8.9b) can be used to compute $\hat{\rho}_X$, $\hat{\rho}_Y$, r_{XY}, s^2_Y and s^2_X, followed by expressions (8.10a)–(8.10c) being used to compute n^*. Variables X and Y can be averaged if contaminants X and Y are considered to be equally important for remediation prioritizing, rendering the statistic $\frac{1}{2}(\overline{X}+\overline{Y})$ and the need to construct a UCL for $\frac{1}{2}(\mu_X+\mu_Y)$, where μ_X and μ_Y respectively denote the means for variables X and Y. Results from this analysis identify the high priority subregions within a site in terms of X and Y jointly, a demarcation that may well differ from that identified by simply doing a map overlay of the UCL of X and the UCL of Y.

8.6 The Murray Superfund Site: Part V

Correlations between As and Pb are reported in Table 8.1. The logarithmic transformation markedly increases the linear correlation estimate from 0.589 to 0.740. Addition of the constant translation parameter, and then the heterogeneity translation parameters only slightly changes this result. Adjusting for the presence of SA in both As and Pb reduces the correlation to roughly 0.706. In addition, given $\hat{\rho}_X = 0.532$ and $\hat{\rho}_Y = 0.494$ (see Table 8.2), where As and Pb respectively were arbitrarily linked to X and Y, the effective sample size is $n^* = 72.5$, which is contained in the interval [68.9, 77.6].

Table 8.4 summarizes stepwise regression results for Eqs. (8.9a) and (8.b). Fourteen eigenvectors account for approximately 30 percent of the variance in As and in Pb; a graphical portrayal of these equations appears in Fig. 8.8. Eight of these eigenvectors are common to Eqs. (8.a) and (8.b); their map patterns appear in Fig. 8.9. Equations (8.9a) and (8.9b) furnish a modestly better data description than does the spatial SAR model specified in equation (8.4). The residuals produced by both Eqs. (8.4) and (8.9) appear to contain only trace levels of SA. Equations (8.9a) and (8.9b) suggest a slightly weaker correlation between X and Y than is obtained through the use of Eq. (8.4).

For a bivariate analysis, assuming equal importance of Pb and As for prioritizing (i.e., w = 0.5), normal curve theory states that the 95% UCL is needed for

$$[\mathbf{1}^T\mathbf{X}/n + \mathbf{1}^T\mathbf{Y}/n]/2,$$

Table 8.4 Stepwise spatial filter modeling results

	As					Pb				
				Residual					Residual	
Step	Eigen-vector	Coefficient probability	Adj-R^2	MC	GR	Eigen-vector	Coefficient probability	Adj-R^2	MC	GR
0	***	***	0	0.321	0.692	***	***	0	0.255	0.772
1	**3**	0.000	0.083	0.221	0.776	**3**	0.000	0.059	0.203	0.826
2	**1**	0.000	0.129	0.172	0.817	**12**	0.000	0.111	0.161	0.866
3	**28**	0.000	0.161	0.152	0.851	**28**	0.000	0.158	0.132	0.919
4	**12**	0.004	0.182	0.130	0.868	**20**	0.000	0.204	0.093	0.974
5	**31**	0.005	0.202	0.117	0.877	**17**	0.006	0.222	0.074	0.990
6	10	0.006	0.221	0.095	0.897	**6**	0.010	0.238	0.052	1.002
7	**20**	0.014	0.235	0.079	0.917	8	0.013	0.253	0.031	1.018
8	4	0.019	0.249	0.059	0.942	**31**	0.027	0.264	0.020	1.029
9	17	0.039	0.258	0.047	0.951	**2**	0.039	0.273	0.002	1.037
10	6	0.045	0.267	0.032	0.960	**1**	0.049	0.281	−0.014	1.050
11	35	0.045	0.276	0.023	0.970	**26**	0.060	0.288	−0.024	1.060
12	11	0.047	0.284	0.009	0.983	**34**	0.062	0.295	−0.033	1.064
13	33	0.082	0.290	0.002	0.992	**22**	0.072	0.302	−0.043	1.075
14	7	0.091	0.296	−0.011	1.004	**24**	0.091	0.307	−0.051	1.085

NOTE: common eigenvectors are highlighted with bold numbers

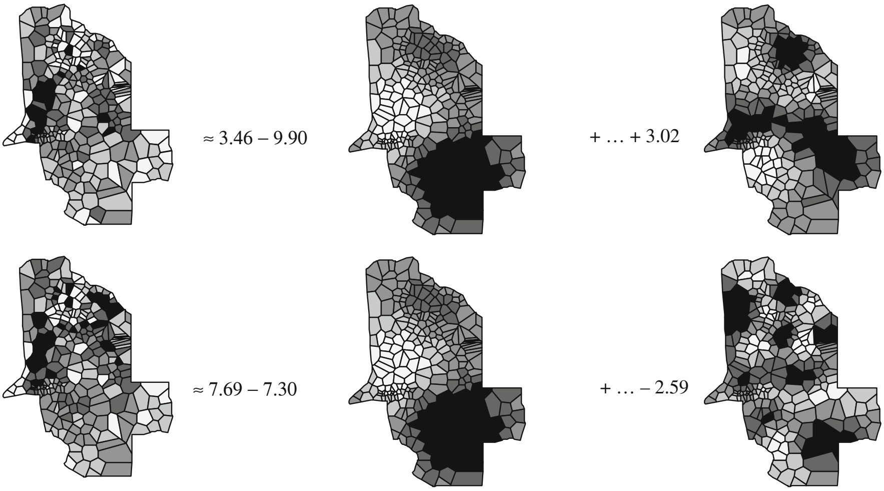

Fig. 8.8 Choropleth maps portraying the spatial filtering equation. *Top* (**a**): the As map together with eigenvector maps of E_3 and E_7. *Bottom* (**b**): the Pb map together with eigenvector maps of E_3 and E_{24}

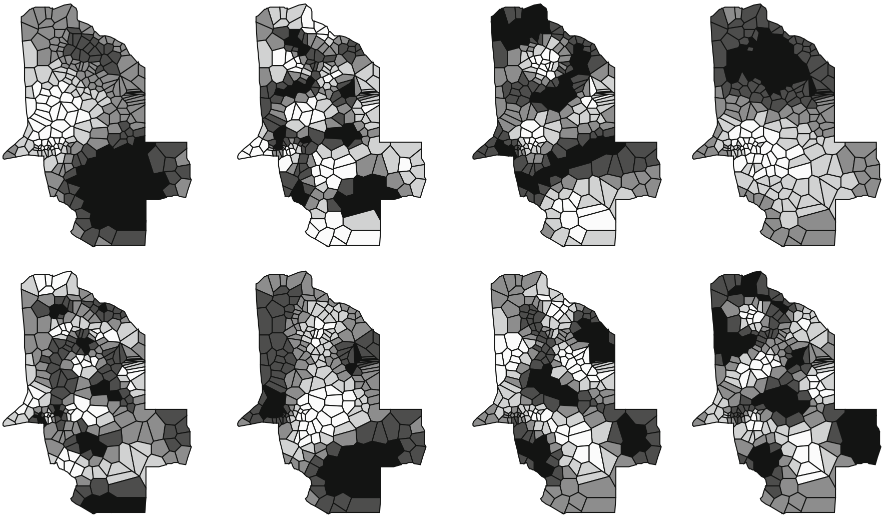

Fig. 8.9 Choropleth maps of common eigenvectors for the As and Pb spatial filter models that are highlighted in Table 8.4. In clockwise direction, beginning with the *top left*: E_1, E_3, E_6, E_{12}, E_{17}, E_{20}, E_{28}, and E_{31}

using the accompanying t-statistic of $t_{n*-4,0.95}$. The UCL value here increases from 5.77074 (the calculation result when ignoring latent SA) to 5.87521, an increase of nearly 2%.

The high priority subregions for remediation appear in Fig. 8.7. Figure 8.7d identifies those parts of the site whose joint As and Pb contamination meets the joint

95% UCL criterion. Figure 8.7c is the result of overlaying Figs. 8.7a, b. Of note is that Fig. 8.7d is not simply the union or intersection of Figs. 8.7a, b, as is Fig. 8.7c. Furthermore, the differences between Figs. 8.7c, d supports the need to do multivariate rather than univariate spatial analyses. The cost of substituting univariate overlay for multivariate spatial statistics would be undertaking remediation work on lower priority locations at the expense of consuming resources for undertaking remediation work on higher priority locations, perhaps even elsewhere.

8.7 Implications

Pollution remediation work within contaminated landscapes, such as superfund sites, may compromise remediation of only the highest priority polluted locations if SA latent in pollutants is overlooked. Researchers could believe they have more statistical information than actually is available to them, as well as more statistical precision than exists when calculating confidence intervals for demarcating subregions for remediation. The same is true if considerable heterogeneity is overlooked. In other words, the methodology presented in this paper furnishes an answer to the question asking what the correct UCL calculation is. In doing so, it highlights that spatial heterogeneity merits more attention when drawing a model-based geographic inference, the size of geographic samples generally is misunderstood, and ignoring SA reveals good but less efficient first-approximation priority subregions.

As is illustrated in the paper using the Murray superfund site, spatial autoregressive models or their spatial filtering counterparts can be used to establish the statistical thresholds for prioritizing remediation of locations. Geostatistical procedures can be used to interpolate pollution surfaces in order to identify subregions for remediation. And, when more than one heavy metal is of concern, the proper spatial statistical analysis is more than simply a map overlay exercise; neither the union nor the intersection of individual contaminant high priority subregions represents the joint contaminants high priority subregion. In other words, the methodology presented in this paper also furnishes an answer to the question asking what method should be used to identify high priority subregions of a polluted site. In doing so, it highlights that results for multiple contaminants should not be based simply on map overlays of individual contaminant results.

The most important finding of research summarized in this paper is that spatial statistics coupled with GIS offers an invaluable economic geography tool for helping allocate the enormous amount of money and people-years of effort needed to complete the necessary environmental restoration being undertaken by modern society. In other words, accounting for SA makes a difference!

Acknowledgment Part of this research was funded by the National Science Foundation, research grant #BCS-9905213. The superfund site data were made available by Dr. Philip E. Goodrum, Research Scientist, and Mr. William C. Thayer, Syracuse Research Corporation; Dr. Susan Griffin, U.S. E.P.A. Region 8, furnished the original Murray superfund site data. Help with preparation of the graphics by Mr. Michael Schramm, Department of Geography, Syracuse University, also is gratefully acknowledged.