Simulating Spatial Distributions, Variability and Uncertainty of Soil Arsenic by Geostatistical Simulations in Geographic Information Systems

Y.P. Lin*  

Department of Bioenvironmental Systems Engineering, National Taiwan University, Taiwan

Abstract: This study quantifies and delineates the spatial distributions, variability and uncertainties of soil arsenic (As) in the northern part of Changhua County in central Taiwan by using kriging, sequential Gaussian simulation (SGS) and simulated annealing simulation (SAS) in geographic information systems. Thousand realizations of soil As are simulated by using SGS and SAS. The impacts of the number of generated realizations on the standard deviation of the soil arsenic distributions simulated by SGS and SAS were performed for assessing and mapping spatial variability and distributions of soil arsenic. The semivariogram results show that As data exhibited small scale variation in the study area. Kriging captures spatial distribution of soil As, but underestimate high As concentration area. However, both SGS and SAS well capture spatial distributions and variability of soil arsenic in the study area, but SGS results in higher average standard deviation than SAS with the same number of realizations. 40 realizations of SAS are reliable to simulate and map the distributions, variability and uncertainty of soil arsenic, but 100 realizations are needed by using SGS. Both estimates and simulates demonstrated that the high As concentration area distributed around the irrigation ditch systems and industrial plants in the study site. Finally, the cumulative distribution of soil As of 100 SGS realizations is obtained and can be used for further risk assessment.

Keywords: Soil arsenic, Geostatistical simulation, Kriging, spatial distribution, Spatial variability, Spatial uncertainty.

INTRODUCTION

Effective environmental management and risk assessment in soil-contaminated sites rely on the precise delineation and mapping of spatial distributions, polluted areas and the variability of soil pollutants. However, soil pollution data such as on heavy metals in soils occasionally exhibit small-scale discontinuities or variations which can cause difficulty in delineating the characteristics of soil pollutants. These are a result of natural phenomena and human activity. One heavy metal of great concern, arsenic, is a ubiquitous naturally occurring element which is present in almost all soils [2]. The arsenic content of soils is related to the geological substratum, and a rather wide range of arsenic levels have been found in soils around the world with an average of 5-10 mg/kg in uncontaminated soils [2, 3]. Arsenic may accumulate in soil because of human activity such as the discharge of waste from metal processing plants, the burning of fossil fuels, the mining of arsenic-containing ores and the use of arsenical pesticides [2, 3]. It has toxic effects on plants and may accumulate in plants, possibly entering the animal and human food chain [4]. Therefore, accurately identifying the spatial characteristics of soil arsenic promotes accurate and effective environmental management, risk assessment and remediation in agricultural fields.

Geostatistics can be used to characterize and quantify spatial variability, perform rational interpolation, and estimate the variance of the interpolated values [1]. Kriging, a geostatistical method, is a linear interpolation procedure that provides a best linear unbiased estimator (BLUE) for quantities that vary in space. The procedure is used to obtain estimates for unsampled sites. Unlike kriging, stochastic simulation does not aim to minimize local error variance but focuses on the reproduction of statistics such as a sample histogram or the semivariogram model, while honoring data values [5-8]. A simulated map, which is a realization of the adopted RF (Random Function) model, looks more "realistic" than the map of the statistically "best" estimates, because it reproduces the spatial variability from the sample information [7]. Stochastic simulation is thus increasingly preferred to kriging for applications that require that the spatial variation of the measured field is preserved [7, 9].

Stochastic conditional simulation - geostatistics, such as sequential Gaussian (SGS) and simulated annealing (SAS) - can be adopted to generate multiple realizations, including an error component, absent from classical interpolation approaches. The described simulation techniques generate a set of values with a specified mean and covariance, and also reproduce measured data at various locations. However, these realizations match similar sample statistics and the conditioning data provide a visual and quantitative measure of spatial uncertainty [10]. For example measurements made can be used, along with simulated values, to analyze the spatial distribution of the variable of interest. In the SGS process, the Gaussian transformation of the available measurements is simulated, such that each simulated value is conditional on the original data and all previous simulated values [1, 11]. Basically, in geostatistics, simulated annealing simulation is a brute-force Monte Carlo simulation method that perturbs or updates, pixelwise, over many iterations, an ini-
tial pixel grid, until the pixel values honor the proposed histogram and semivariogram model [12]. Geostatistical conditional simulations have been recently applied to simulate the spatial variability and distribution of heavy metal pollution in soil to identify polluted areas, support risk assessments and support uncertainty analysis [13-20]. Theories concerning such geostatistical simulation techniques can be found in various books and articles [11, 8, 21].

MATERIALS AND METHODS

In this study, semivariogram models of measured soil As and normalized As data were fitted within a geostatistical software GS+ [22]. Ordinary kriging, SGS and SAS were performed for As data using OKB2DM (ordinary kriging 2-dimension), SGSIM (sequential Gaussian simulation) and SASIM (simulated annealing simulation) in GSLIB (Geostatistical Software Library) [11]. A thousand realizations of As were simulated and estimates were mad using SGS, SAS and kriging in a grid with 38 columns x 28 rows, and therefore 1064 cells. Cells were 50m by 50m squares. The impacts of the number of generated realizations on the standard deviation of the soil arsenic distributions simulated by SGS and SAS were performed for evaluating the simulation consistencies of SGS and SAS, and then the variability and mapping distributions of As. The estimated and simulated results were transferred into Arcview 3.0 to identify the spatial distributions, area of potentially high As concentration and high variability of soil As at this site [23]. Finally, the area of high As concentration was delineated from maps of As distributions and variability in the Arcview system. The As concentration in the area was simulated by SGS. The cumulative distribution of 100 SGS realizations of soil As in high As concentration blocks was obtained to support further risk assessment.

Data

In 1983, the Environmental Protection Administration (EPA) of the Republic of China began a collaborative research program to establish the presence of As, Cd, Cu, Cr, Hg, Ni, Pb and Zn trace elements in soil in Taiwan [13]. The program also sought to detect soil properties, such as cation-exchange capacity and pH. These studies sampled 878 sites across Taiwan.

In this study, 194 topsoil samples of As, collected between 1981 and 1986, were taken from the program described above. They were taken from geographically distributed sites at target intervals of 100 m and 50 m in a network formation. Topsoil was sampled at depths of 0-15 cm. Table 1 presents descriptive statistics for As in these 194 topsoil samples. The county from which the samples were selected, Chunghua County, is one of the most important agricultural counties in Taiwan. Fig. (1) presents the 194 sampling sites, located in the northern part of Chunghua County. Fig. (2) displays the industrial plants and irrigation system in the study area. Most of these industrial plants serve the metalwork, electroplating and metal surface treatment industries, as indicated in Fig. (2).

Variography

In geostatistics, semivariograms are used to quantify the observed relationships between sample values [16]. An experimental semivariogram (semivariance) for an interval lag distance class \( h \), \( \gamma_z(h) \), is given by

\[
\gamma_z(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} [Z(x_i + h) - Z(x_i)]^2
\]  

where \( h \) is the lag distance that separates pairs of points; \( Z(x_i) \) is the hydraulic conductivity at location \( x \); \( Z(x + h) \) is the hydraulic conductivity at location \( x + h \), and \( n(h) \) is the number of pairs that are separated by lag distance \( h \).

The experimental cross-semivariogram \( \gamma_{xy}(h) \) of the second-order stationary regionalized variables, \( Z(x) \) and \( Y(x) \) (second variable), is given by

\[
\gamma_{xy}(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} [Z(x_i + h) - Z(x_i)][Y(x_i + h) - Y(x_i)]
\]  

Kriging

Kriging is performed using the weighted sums of the adjacent sampled concentrations. The weights depend on the structure of the correlation exhibited. The criterion used to determine the weights is whether they minimize the
estimation variance. In this context, kriging estimates (BLUE) are regarded as the most accurate of all linear estimates. Hence, kriging was adopted herein to estimate the value of the random variable at an unsampled location $x_0$ based on the measured values in linear form:

$$Z'(x_0) = \sum_{i=1}^{N} \lambda_i Z(x_i),$$

where $Z'(x_0)$ is the estimated value at location $x_0$; $\lambda_i$ is the weight in the kriging estimate of $Z(x_i)$; $x_i$ is the location of the sampling site for variable $Z$, and $N$ is the number of variable $Z$ involved in the estimation.

Based on non-bias constraints and minimized estimation variance, the variance in kriging estimate is given by

$$\sigma^2_{kriging} = \sum_{i=1}^{N} \lambda_i \gamma_i(x_i) + \mu$$

where $\mu$ is the Lagrange Multiplier.

**Sequential Gaussian Simulation**

SGS assumes a Gaussian random field, such that the mean value and covariance completely characterize the conditional cumulative density function (cdf) [24]. In the SGS process, the Gaussian transformation of the available measurements is simulated, such that each simulated value is conditional on the original data and all previously simulated values [1, 11]. A value simulated at a one location is randomly selected from the normal distribution function that is defined by the kriging mean and variance based on neighborhood values. Finally, the simulated normal values are back-transformed into simulated values to yield the original variable. The simulated value at the new randomly visited point value depends on both the original data and the previously simulated values. This process is repeated until all points have been simulated.

**Simulated Annealing**

The annealing algorithm perturbs the image by simulating thermal perturbation [1, 25, 26]. In SAS, swapping the values in pairs of grid nodes that do not contain a conditioning datum, modifies the initial field, reproducing a histogram [1, 11]. The target constraint is the reproduction of semivariograms, and the corresponding objective function must be lowered using the annealing schedule [8]. The objective function is defined as a mean squared difference between the experimental and given semivariograms. The objective function ($O$) is defined as,

$$O = \sum_h \frac{[\gamma'(h) - \gamma(h)]^2}{\gamma(h)^2},$$

where $\gamma(h)$ is the pre-specified semivariogram, and $\gamma'(h)$ represents the semivariogram of the simulated realization.

A temperature function (the Boltzman distribution) in a simulated annealing procedure controls the rate of reduction of the optimization function by allowing some switches to increase its value [11, 27]. The parameter of the temperature function is called the temperature in the annealing procedure [1]. Increasing the temperature increases the probability that an unfavorable swap is accepted [11].

**RESULTS AND DISCUSSION**

**Semivariogram Model of As**

The sill is the upper limit toward which a semivariogram tends at a large distance (the range of a semivariogram model). The sill of the semivariogram model measures the variability of the investigated variable; a higher sill corresponds to greater variability. The range of a semivariogram model denotes the distance above which measured variables become independent each other and are more likely to be random. Therefore, a greater range of the semivariogram model corresponds to greater continuity of the soil heavy metals. Given the large degree of spatial variability, the semivariograms of soil heavy metals commonly exhibit a marked nugget effect that is caused by field test errors and small-scale variability. In this study, experimental semivariograms were obtained for As and normalized As at the same active lag and lag interval. A relatively consistent set of best-fit models with minimal RSS (Model Reduced Sum of Squares) and maximal $r^2$ values were obtained by the least-squares fitting of these semivariograms. Figs. (2,3) display the best-fit models for As and normalized As.

![Fig. (2). Semivariogram of (a) measured As; (b) normalized As.](image-url)

A spherical model with nugget effect = 5.29 (mg/kg)$^2$, sill = 17.32 (mg/kg)$^2$ and range = 704.00m was the best-fit model for the As data ($r^2$=0.93 And RSS=6.84). The semivariogram of normalized As was an exponential model with nugget effect = 0.14, sill = 0.94 and range = 186.00m (effective range = 558.00m) ($r^2$=0.9 And RSS=0.03). The
small-scale variations (nugget effects) of As and normalized As were, respectively, 30.50% and 14.89% of the total variation of the As data and the normalized As with ranges 704.00 m and 558.00 m, respectively. The variographic results revealed that the semivariogram models of As with a high nugget effect exhibited a large small-scale variation or measurement error in the As data in the study area. Additionally, the ratio of the nugget effect to the total sill of As is greater than that of normalized As, revealing that the measured As exhibited a greater small-scale variation than did the normalized As. The ranges of the semivariogram models of As and normalized As indicated that the continuity of measured As data slightly exceeded that of the normalized As.

Statistics and Spatial Dependence of Estimation and Simulations of As Concentration

The ordinary kriging estimates, and SGS and SAS simulations were based on the above semivariogram models and the 194 soil As and normalized As data. Table 1 summarizes the descriptive statistics for ordinary kriging, three realizations of SGS and three realizations of SAS results. This comparison reveals that conditional simulations, such as using SGS and SAS, produce statistics for the empirical data concerning As that are better than those obtained by kriging.

Table 1. Descriptive Statistics of Kriging Estimates and Simulations of As

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Med.</th>
<th>Std. D</th>
<th>Min</th>
<th>Max</th>
<th>Q25th</th>
<th>Q75th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measured</td>
<td>10.93</td>
<td>10.00</td>
<td>3.86</td>
<td>5.23</td>
<td>30.00</td>
<td>8.75</td>
<td>12.18</td>
</tr>
<tr>
<td>Ok</td>
<td>11.12</td>
<td>10.69</td>
<td>2.81</td>
<td>6.48</td>
<td>25.94</td>
<td>9.25</td>
<td>12.18</td>
</tr>
<tr>
<td>SGS1</td>
<td>11.07</td>
<td>10.00</td>
<td>4.15</td>
<td>5.23</td>
<td>30.00</td>
<td>8.83</td>
<td>12.00</td>
</tr>
<tr>
<td>SGS2</td>
<td>11.16</td>
<td>10.00</td>
<td>3.68</td>
<td>5.23</td>
<td>30.00</td>
<td>8.83</td>
<td>12.92</td>
</tr>
<tr>
<td>SGS3</td>
<td>11.11</td>
<td>10.00</td>
<td>3.93</td>
<td>5.23</td>
<td>30.00</td>
<td>8.84</td>
<td>12.64</td>
</tr>
<tr>
<td>SAS1</td>
<td>10.83</td>
<td>10.00</td>
<td>3.67</td>
<td>5.23</td>
<td>30.00</td>
<td>8.70</td>
<td>12.00</td>
</tr>
<tr>
<td>SAS2</td>
<td>11.00</td>
<td>10.00</td>
<td>4.16</td>
<td>5.23</td>
<td>30.00</td>
<td>8.76</td>
<td>12.00</td>
</tr>
<tr>
<td>SAS3</td>
<td>10.85</td>
<td>10.00</td>
<td>3.87</td>
<td>5.23</td>
<td>30.00</td>
<td>8.62</td>
<td>12.00</td>
</tr>
</tbody>
</table>

SGS #: SGS simulation #; SAS #: SAS simulation #; Q25th: 25th percentile; Q75th: 75th percentile; Unit: (mg/kg).

The experimental semivariograms of three realizations of SGS and SAS were also calculated using GS+, as presented in Fig. (3). All experimental semivariograms based on the simulations are consistent with the experimental semivariogram for As data. These experimental semivariograms show that SGS and SAS effectively reproduce the spatial structure, spatial dependence and spatial continuity (experimental semivariogram) of the investigated As values, while ordinary kriging does not perform well.

Local and Global Spatial Variability and Uncertainty

The extent of the space of uncertainty increases with the number of realizations generated [8]. In this study, Fig. (4) plots the relationships between the (average standard deviation of realizations of soil As simulated by SGS and SAS) and the number of realizations to determine the spatial variability and distributions of As. When SGS is used to simulate As concentration, the average standard deviation reaches a plateau when the number of realizations equals 100. However, when SAS is used, the average standard deviation approaches an asymptotic value when the number of realizations is 40. Moreover, SGS yielded a larger average standard deviation than SAS with the same number of realizations (Fig. 4). The relational analysis revealed that SAS consistently produces narrower output distributions of soil As than does SGS.

Differences between realizations of soil As provide a model for uncertainty in the spatial distribution of measured values [8]. Accordingly, the spatial distributions of standard deviations of 100 realizations simulated by SGS and 40 by SAS at 1064 grid cells in the study area are mapped in GIS, as shown in Figs. (5,6). The figures show that the area of high variability area is along the irrigation systems, according to the SGS simulations (Fig. 5). The areas of high variability according to the SAS simulations are in the north part and east part of the study area. However, the north part of the study area with high variability, identified by both SAS and SGS, must be considered in further sampling, and assess-

Fig. (3). Experimental semivariogram of measured and simulated As.

Fig. (4). Number of realization associated with the average standard deviation of simulated As concentration using SGS and SAS.
ments of soil arsenic pollution. Figs. (5,6) also verify that SAS consistently simulates not only the global but also the local distributions of soil As in the study area.

**Spatial Distribution of Estimation and Simulations**

Estimates and simulations of soil As were imported and mapped in Arcview 3.0a [28]. Fig. (7) presents the kriging map of As, in which the high concentration areas were in the north and northeast of the investigation area. The ordinary kriging estimate map confirmed that kriging tended to smooth extreme values of the empirical As data set. Although they show the spatial patterns of these As concentrations, kriging results may overestimate the size of areas of high and low concentrations of heavy metals in the soil, and underestimate the sizes of areas with extremely high and low concentrations of soil As. Stochastic simulation does not seek to minimize a local error variance but focuses on the reproduction of the sample histogram or the semivariogram model in addition to honoring the data values [8]. Simulated maps of soil arsenic look more realistic than the map of the statistically best estimates, because they reproduce the spatial variability from sample information [5-7, 8]. Simulated maps of soil arsenic are shown in Figs. (8-10). These maps also plot distributions of As concentration over the study area. The distributions of estimated and simulated As based on kriging, SGS and SAS can be verified by the measured As at 194 samples in Fig. (1b). Figs. (1b,7-10) also demonstrated that kriging, SGS and SAS can estimate and simulate spatial distributions of As over the study area. Most areas with high As concentration (> 10 mg/kg) are strongly associated with the location of industrial plants and the irrigation systems of the study area, as displayed in Figs. (8-10).
spatial patterns of soil As are related not only to the industrial plants but also to the irrigation systems in the study area, as presented in Figs. (7-10). These simulated maps also reveal that the spatial distributions and spatial variability of soil As can be captured by 100 SGS realizations and 40 SAS realizations (Figs. 1b, 8-10). The simulation results also indicate why stochastic simulation is increasingly preferred to kriging in applications in which the spatial variation of the measured field must be preserved, such as in the delineation of the contaminated area [8]. Additionally, a set of simulations can be conducted with remediation or the land-use policy to evaluate the uncertainty of the responses, such as remediation efficiency or soil productivity [8].

The simulation-based approach has several advantages over kriging: (1) it provides a model of spatial uncertainty, e.g. the probability that a given threshold is exceeded jointly at several locations can be readily computed, (2) conditional cumulative distribution function for support larger than the measurement support (e.g. remediation units) can be numerically approximated by the cumulative distribution of block simulated values that are obtained by averaging values simulated within block, and (3) the set of realizations allows one to study the propagation of uncertainty through global GIS operations [21]. In this study, a block of high concentration was identified in Arcview based on the average values of 40 SAS realizations and 100 SGS realizations, as presented in Fig. (11). The block is located in the north part of the study area. Fig. (12) plots the cumulative distribution of the 100 SGS realizations of As values. The cumulative distribution reveals that the probability of exceeding an As concentration of 10 (mg/kg) in the block is about 0.7. The cumulative distribution (Fig. 12) can be used for further assessing the risk associated with soil arsenic.

Fig. (9). Average As concentration of (a) 100 SGS realizations; (b) 1000 SGS realizations.

CONCLUSIONS

This study performed kriging estimation, as well as SGS and SAS simulations to quantify the spatial distributions, variability and uncertainties of soil arsenic in the northern part of Changhua County in central Taiwan. Kriging can estimate the spatial distribution of soil As, but fails to capture and identify high soil As concentrations and the areas in which the concentrations are high. However, 1000 SGS realizations and 40 SAS realizations of soil arsenic can capture both spatial distributions and high As concentration area. The kriging results and simulations revealed that the high As area was around the irrigation ditch systems and industrial plants in the study site. The cumulative distribution of simulated realizations and the estimated and simulated maps of soil As can be used in further risk assessment and management.
ACKNOWLEDGEMENT
The author would like to thank Dr. Tsun-Kuo Chang for supporting data and the National Science Council of the Republic of China for financially supporting this research, under Contract Nos. NSC 87-2621-P-002-012 and NSC 89-2621-B-002-004.

REFERENCES


© Y.P. Lin; Licensee Bentham Open.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.5/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.