International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 6 (Special Issue), 2007, pp 89-100

Location Mapping Implications for Spatial Prediction of Low-Concentration Pollutants and the Arsenic Problem

Munni Begum
Dept. of Mathematical Sciences
Ball State University
Muncie, IN 47306-0490
U.S.A

Pranab K. Sen
Dept. of Biostatistics, Statistics and Operational Research
University of North Carolina at Chapel Hill, NC 27599-7420
U.S.A

[Received November 4, 2005; Revised June 20, 2007; Accepted July 8, 2007]

Abstract

Exposure assessment is an important task in environmental risk assessment and should be done prior to formulating any dose response relationship. Understanding the presence or absence of an agent, and its concentration and distribution is the primary focus of an exposure assessment. Predictive models and environmental monitoring can be used to determine the levels of exposure at different points. Spatial prediction and interpolation of pollutant concentrations is considered here as a mode of exposure assessment. Models based on spatial covariance matrices with unknown elements are considered to construct a predictor of pollutant concentrations at points of interests. Bayesian approach with Markov Chain Monte Carlo (MCMC) techniques is implemented to estimate model parameters. An extensive simulation analysis is conducted to demonstrate the whole process of *kriging* for predicting the arsenic concentration in ground water used for personal usage.

Keywords and Phrases: Exposure Assessment, Groundwater Arsenic Contamination, Spatial Prediction, Gaussian Process, Markov Chain Monte Carlo.

AMS Classification: 62P12.

1 Introduction

One of the major tasks involved in an environmental risk assessment of a toxic chemical is commonly known as exposure assessment. Understanding the presence or absence of an environmental agent and its concentration and distribution is the primary focus of exposure assessment. Measurements of the exposure levels is not always straightforward as it involves measurements to which individuals are exposed, at different locations and/or at different time points. Although measurement error could be regarded as one of the potential sources of variation in the pollutant concentration, it is important to assess spatial and temporal variabilities as well. Investigations on such variabilities in environmental contaminants scramble into limited environmental tools and that in turn produce extra uncertainties for tracing out the environmental stressors and their impact in the midst of significant noise background. As the most noteworthy case, we may refer to groundwater contamination by inorganic as well as organic arsenite compounds which prevails in many parts of the world. There is also a substantial lack in applying proper statistical methods in investigating the sources of variation in exposure levels. Proper assessment of spatial fluctuations, temporal variabilities along with measurement error might provide useful implications for adopting remedial measures. Selection of study sites for sampling purpose and the time period for actual data collection are usually attributed to spatial and temporal variabilities.

The concentration pattern and level of many ambient pollutants may be hard to assess properly; uneven spatial variation may be one of the most significant problem in measuring concentration levels fairly accurately over a region, and generally, based on suitably chosen sites or locations, gridded measures are used for spatial prediction. We illustrate this with the groundwater arsenic contamination episode. Globally, more so in rural areas, a large proportion of people still use untreated groundwater as their principal source of drinking water or for other household usage. In the United States, 16 percent of the population rely on their own private well for drinking water [8]. In other parts of the world - much less wealthy countries like Bangladesh and other parts of South Asia, majority of the population use untreated groundwater for drinking and other usages. As such, we think that it is timely to appraise the limitations of selection of too many sites or locations as well as their spatial distributions for examining the implications of location mapping for low concentration arsenite pollutants exposure and after effect studies. Arsenic in drinking water poses serious public health problem in Bangladesh, West Bengal, Taiwan, Northern China, Vietnam, Argentina, Mexico, and parts of the United States. In most cases the occurrence of arsenic in the water sediment is geologic in nature. Recent studies [4, 2, 10] provide evidences of the presence of high degree of spatial variability of groundwater arsenic concentration in Bangladesh and other parts of South Asia.

1.1 ***Arsenic Risk From Ground Water Contamination***: Although arsenic contamination of ground water occur mostly due to complicated geological process, anthropogenic assault may also aggravate the whole contamination process. In arsenic endemic regions, mostly in low-lands or coastal planes with subsoil groundwater level very close to the surface, as is particularly the case with Bangladesh, southern parts of West Bengal and Orissa and some other South Asian countries, not only inorganic arsenites prevail at the subsoil level but also often organic arsenic compounds contaminate groundwater layers and sources (e.g. wells, ponds or lakes, streams or canals, and even some rivers). The sources of human exposures to such arsenites can be primarily listed as follows; this listing is by no means exhaustive.

i) Well water [Bangladesh, Taiwan, Inner Mongolia, Southwest USA, West Bengal, India],

ii) Pesticide production plants near residential area [Calcutta, India],

iii) Natural origin and geochemical environments [USA],

iv) Animal corps, dead fishes, and even human bodies buried under ground with insufficient protections, contaminate organic arsenites (although in low levels compared to heavy metal arsenites), and eventually may have significant impact.

Exposure assessment and prediction procedure are likely to vary depending on the study site. Environmental exposure assessment studies mostly collect routine monitoring data that has limited use in assessing the magnitude of spatio - temporal variability [9]. Data on individual level exposure over a region or at different locations in a region for a certain period of time can be used in assessing spatio - temporal distribution. For example, let us consider source (i) in the above list. People are exposed to arsenic primarily through drinking water extracted from the aquifer using shallow or deep wells. The variation in the arsenic concentration to which an individual is exposed, may depend on concentration in neighboring wells along with specific well and its depth, study site accounting for regional geology, seasonal change and even the duration of using the well as a source of drinking water. Concentration levels are observed as positive quantities and usually exhibit skewed distribution with a long right tail. Log and/or power transformation could be applied to the observations in order to attain some degree of symmetry to the distribution.

Spatial variability in the distribution of arsenic concentration over a target region may often be present to a significant degree. The nature of the spatial variability could be complex with both large differences between neighboring wells and regional trends. Moreover occasional arsenic ‘hot spots’ in the generally low arsenic regions complicates the situation further. Due to large degree of well-to-well variation within a village, it would be difficult to predict arsenic concentration from the available concentrations in neighboring wells. Regional geology also determines the trend, as such, young holocene aquifer along with depth of a well is usually associated with high arsenic concentration [6]. On the other hand pleistocene sediments of the uplifted regions provide low-arsenic water. However, presence of iron oxyhydroxides coating in an aquifer under oxic conditions keeps ground water arsenic level low. But in river floodplains, coastal areas, and delta sediments these oxyhydroxides dissolve due to microbial action and cause increases in the concentration of arsenic in ground water in those areas. Human activities, such as dumping arsenic containing chemicals directly into the ground may also cause increased amount of leaching of arsenic into the deep aquifer. In any case, understanding spatial pattern of arsenic concentration at a particular region precisely, has several implications for mitigation. Most important of these implications is mapping locations with low arsenic concentration that would reduce the extent of exposure, at least in the short term.

2 Spatial Prediction of Pollutant Concentration at a Single Point

As has been discussed earlier, pollutant-concentrations are measured at selected sites, scattered over a region, and are used to estimate the spatial pattern for the entire area. For air-borne pollutants, some measure of the particle density in a small grid is to be recorded at various times and for different sites as well. Typically, such measures are nonnegative and rarely conform to a normal distribution. For the arsenic in the groundwater problem a small volume of water or soil is to be appraised for the pollution-concentration, and above a threshold level, it is regarded as unsafe for human reaction. Apart from the difficulties in getting the proper sample from a site, it is to be noted that considerable variation is usually perceived not only due to measurement problems but also non-uninform concentration of the pollutants across different layers of depth as well as time of the day and season. Dealing with this nonstandard problem, the first task is to have these measurement in a standardized way so as to control measurement errors as far as possible. Further, suitable transformations on the measurements to induce more symmetry of their distributions, if not more closer approximation of normality, is generally essential. As such, we shall assume that based on past experience, we have a transformed variable recorded at different sites and different time-points and these have closely Gaussian distributions. It is possible that concentration level below a minimal threshold value may either be imperceptible or their measurement could be very imprecise. This may invite left censoring or truncation in our model. A right censoring or truncation may similarly arise if an upper threshold value create similar measurement problems.

In order to incorporate spatial dependence as described above we consider spatial prediction process using parametric covariance models. The process is technically known as *kriging* named after a South African mining engineer D. G. Krige who developed the technique to predict ore reserves accurately. The process usually involves construction of a spatial predictor in terms of unknown model parameters. Let us consider a set of observations of a vector random field, $Z(s_1), \ldots, Z(s_n)$, at a set of points $s_1, \ldots, s_n$. Here $\{Z(s), s \in D\}$ is a stochastic process with $D \subset \mathcal{R}^d$, where $\mathcal{R}^d$ is a d dimensional euclidian space. For example, $Z(s)$ may denote average daily ozone concentration during summer at a location s. The problem is to predict $Z(s_0)$ for some $s_0 \notin \{s_1, \ldots, s_n\}$.

Let

$$\mathbf{Z} \;=\; [Z(s_1), \ldots, Z(s_n)]^T; \quad z_0 = Z(s_0),$$

with $\mathrm{Cov}(\mathbf{Z}) = \boldsymbol{\Sigma}$, $\mathrm{Cov}(\mathbf{Z}, z_0) = \boldsymbol{\tau}$ and $\mathrm{Var}(z_0) = \sigma_0^2$. Then a basic model to start with is,

$$\mathbf{Z} \;=\; \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\; z_0 = \mathbf{x}_0^T\boldsymbol{\beta} + \epsilon_0;$$

where $\mathbf{X}$ is some matrix of covariates at the points $s_1, \ldots, s_n$, and $\mathbf{x}_0$ is a given vector of covariates at s_0. Under the assumption that the process is Gaussian, $Z(s_1), \ldots, Z(s_n)$ has a multivariate normal distribution with mean vector $\mathbf{X}\boldsymbol{\beta}$ and variance covariance matrix $\boldsymbol{\Sigma}$. Following the classical result of multivariate analysis [7], the conditional distribution of z_0 given $\mathbf{Z}$ is normal with mean $\mathbf{x}_0^T\boldsymbol{\beta} + \boldsymbol{\tau}^T\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})$ and variance $\sigma_0^2 - \boldsymbol{\tau}^T\boldsymbol{\Sigma}\boldsymbol{\tau}$. We assumed $\boldsymbol{\Sigma}$ to be an unknown positive-definite matrix of variance and covariances. Similarly, $\boldsymbol{\tau}$ and σ_0^2 are also assumed to be unknown. We consider a structured form for $\boldsymbol{\Sigma}$, based on the correlations between two observations at two different locations. Observations taken at two locations closer to each other are expected to vary in similar pattern, which may not hold for those far apart. A natural model to handle such a covariance structure is an exponential correlation model for $\boldsymbol{\Sigma}$ as follows,

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \exp(-\phi d_{12}) & \exp(-\phi d_{13}) & \cdots & \exp(-\phi d_{1n}) \\ & 1 & \exp(-\phi d_{23}) & \cdots & \exp(-\phi d_{2n}) \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ & & & & \exp(-\phi d_{n-1n}) \\ \ldots & \ldots & \ldots & \cdots & 1 \end{bmatrix} = \sigma^2\mathbf{B},$$

where $d_{jk} = |s_j - s_k|,\; j \neq k$, euclidian distance between locations s_j and s_k. Similarly we consider,

$$\begin{aligned} \boldsymbol{\tau} &= \mathrm{Cov}(\mathbf{Z}, z_0) = \mathrm{Cov}\left((z(s_1), \ldots, z(s_n)), z_0\right) \\ &= \left[\mathrm{Cov}\left(z(s_1), z_0\right), \mathrm{Cov}\left(z(s_2), z_0\right), \ldots, \mathrm{Cov}\left(z(s_n), z_0\right)\right] \\ &= \sigma^2\left[\exp(-\phi d_{01}), \exp(-\phi d_{02}), \ldots, \exp(-\phi d_{0n})\right] \\ &= \sigma^2\mathbf{A}, \end{aligned}$$

where $d_{0j} = |s_0 - s_j|\; j = 1, \ldots, n$ and $\mathrm{Var}(z_0) = \sigma_0^2$.

The correlation model assumes that the correlation between two pairs of observations at two points in an area, decays toward zero as the distance between the points increases, also that the correlation tends to 1 as the distance tends to zero. This assumption is reasonable as pollutant concentrations measured at points with greater proximity tend to show similar pattern. It is also to be noted that the number of covariance parameters is reduced from $\frac{n(n+1)}{2}+2$ to only three σ^2, ϕ, and σ_0^2. However, iterative algorithms are still needed to estimate these parameters and the regression coefficients $\boldsymbol{\beta}$. The prediction of $z(s_0)$ is then straight forward to carry out by plugging in the estimates of $\boldsymbol{\beta}$ and σ^2, σ_0^2 and ϕ in the following expression.

$$\begin{aligned}\hat{z}(s_0) &= E_{z_0}[z_0|\mathbf{Z}] \\ &= (\mathbf{x}_0 - \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau})^T\hat{\boldsymbol{\beta}} + \boldsymbol{\tau}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}. \qquad (1)\end{aligned}$$

The regression parameters and covariance parameters are estimated simultaneously using Bayesian techniques with the implementation of Markov Chain Monte Carlo (MCMC) tools. These tools are presented in the following subsection.

2.1 MCMC Techniques for Predicting Pollutant Concentration

The basic model we have is as under,

$$\mathbf{z}^* = \begin{pmatrix} \mathbf{Z}_n \\ z_0 \end{pmatrix} \sim \mathcal{N}_{n+1}\left[\begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{x}_0^T\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\tau} \\ \boldsymbol{\tau}^T & \sigma_0^2 \end{pmatrix} \right].$$

Therefore,

$$z_0|\mathbf{Z} \sim \mathcal{N}[\mathbf{x}_0^T\boldsymbol{\beta} + \boldsymbol{\tau}^T\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}),\ \sigma_0^2 - \boldsymbol{\tau}^T\boldsymbol{\Sigma}\boldsymbol{\tau}].$$

To carry out the MCMC method under a Bayesian setup, we need to specify the prior distributions for the unknown parameters, $\boldsymbol{\beta}$, σ^2, ϕ, and σ_0^2. One may specify either informative or non-informative priors for the parameters. Non-informative prior distributions are preferable in the sense that the inference mimics likelihood based inference carried out by frequentist approach. We specify the priors for regression parameters $\boldsymbol{\beta}$ and variance parameters as under,

$$\boldsymbol{\beta} \sim \mathcal{N}_q(\boldsymbol{\beta}_0,\ \boldsymbol{\Sigma}_0).$$

The hyper-covariance matrix $\boldsymbol{\Sigma}_0$ is taken as diagonal with large diagonal values to make the prior specification for $\boldsymbol{\beta}$ as non-informative. Instead of the variance parameters σ^2 and σ_0^2, we specify priors for the precision parameters $\tau_1 = \frac{1}{\sigma^2}$ and $\tau_0 = \frac{1}{\sigma_0^2}$. The prior specifications for τ_0, τ_1 and the correlation parameter ϕ is considered as follows,

$$\begin{aligned}\tau_0 &\sim \text{Gamma}\left(\frac{a}{2}, \frac{b}{2}\right), \\ \tau_1 &\sim \text{Gamma}\left(\frac{c}{2}, \frac{d}{2}\right), \\ \phi &\sim \text{Uniform}(0, 1).\end{aligned}$$

Gamma priors are reasonable choices for the non-negative precision parameters while a Uniform prior with range (0,1) would be appropriate for the correlation parameter ϕ. With these prior specifications, the posterior distribution of the unknown quantities can be written up to the multiplicative constant as under,

$$\begin{aligned} & p(\boldsymbol{\beta},\ \tau_0,\ \tau_1,\ \phi|\mathbf{z}^*) \propto \\ & \quad (\sigma_0^2 - \boldsymbol{\tau}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(z_0 - \mathbf{x}_0^T\boldsymbol{\beta} - \boldsymbol{\tau}^T\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}))^2\right] \\ & \times \quad |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})\right] \times |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} \\ & \times \quad \exp\left[-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right] \times \tau_0^{\frac{a}{2}-1} \exp\left[-\frac{b}{2}\tau_0\right] \times \tau_1^{\frac{c}{2}-1} \exp\left[-\frac{d}{2}\tau_1\right] \times 1. \end{aligned}$$

Clearly the above expression does not follow a closed form. Therefore, we derive full conditional distributions for each parameters, to carry out MCMC techniques by drawing samples from these marginal posterior distributions. The full conditionals of $\boldsymbol{\beta}$, τ_0, τ_1, ϕ can be written up to a multiplicative constant as under,

$$p(\boldsymbol{\beta}|\mathbf{z}^*,\ \tau_0,\ \tau_1,\ \phi) \ \propto \ \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Lambda}_2\right)^T \boldsymbol{\Lambda}_1 \left(\boldsymbol{\beta} - \boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Lambda}_2\right)\right], \tag{2}$$

$$p(\tau_0|\mathbf{z}^*,\ \tau_1,\ \boldsymbol{\beta},\ \phi) \ \propto \ \left(1 - \tau_0\sigma^2\mathbf{A}^T\mathbf{B}^{-1}\mathbf{A}\right)^{-\frac{1}{2}} \tau_0^{\frac{a+1}{2}-1} \exp\left(-\frac{b}{2}\tau_0\right), \tag{3}$$

$$\begin{aligned} & p(\tau_1|\mathbf{z}^*,\ \tau_0,\ \boldsymbol{\beta},\ \phi) \propto \\ & \quad \left(\tau_1\tau_0^{-1} - \mathbf{A}^T\mathbf{B}^{-1}\mathbf{A}\right)^{-\frac{1}{2}} \tau_1^{\frac{1}{2}c-1} \times \exp\left[-\frac{\tau_1}{2}\left\{d + (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{B}^{-1} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})\right\}\right], \end{aligned} \tag{4}$$

$$\begin{aligned} & p(\phi|\mathbf{z}^*,\ \boldsymbol{\beta},\ \tau_0,\ \tau_1) \propto \\ & \quad \left(\tau_0^{-1} - \tau_1^{-1}\mathbf{A}^T\mathbf{B}^{-1}\mathbf{A}\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(z_0 - \mathbf{x}_0^T\boldsymbol{\beta} - \mathbf{A}^T\mathbf{B}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})\right)^2\right] \\ & \quad \times |\mathbf{B}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\,\tau_1 \left(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}\right)^T \mathbf{B}^{-1} \left(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}\right)\right]. \end{aligned} \tag{5}$$

The right hand side in expression (2) is a kernel of a q variate normal distribution with mean vector $\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Lambda}_2$ and variance covariance matrix $\boldsymbol{\Lambda}_1^{-1}$. The expressions for $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ are given as below,

$$\begin{aligned} \boldsymbol{\Lambda}_1 &= (\mathbf{x}_0 - \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau})(\mathbf{x}_0 - \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau})^T + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} + \boldsymbol{\Sigma}_0^{-1} \\ \boldsymbol{\Lambda}_2 &= (\mathbf{x}_0 - \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}) + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0. \end{aligned}$$

The full conditionals of τ_0, τ_1 and ϕ in (3), (4) and (5) do not have closed forms. To draw samples from these full conditional distributions, the Metropolis- Hastings algorithm is invoked. Logged versions of (3), (4) and (5) are used in the Metropolis algorithm.

2.2 Implementation of MCMC technique through Gibbs Algorithm

The central idea of Monte Carlo methods lies in the use of dependent samples generated by Markov Chains whose limiting distribution is assumed to be the true posterior distribution. Two widely applied methods for generating Markov Chains are Gibbs algorithm of Geman and Geman (1984) [3] extended by Gelfand and Smith (1990) [1] into the statistical context, and Metropolis algorithm [5]. Implementation of the Gibbs algorithm is straightforward for the regression parameters ($\boldsymbol{\beta}$) as their full conditional distribution is of closed form. However, the full conditionals for precision and correlation parameters do not follow closed forms. Also these parameters are constrained to take values within certain range. For instance, τ_0 and τ_1 take strictly non negative values whereas ϕ takes values in the interval $[0, 1]$. Metropolis-Hastings algorithm is used to sample from non-closed form full conditionals. The constraints are addressed in the Metropolis-Hastings algorithm with proposal densities satisfying these. A detailed sampling scheme for generating samples for all the parameters jointly is described below.

Step 0) Initialize regression, precision and correlation parameters. Let $\boldsymbol{\theta}^0 = (\boldsymbol{\beta}^0, \tau_0^0, \tau_1^0, \phi^0)$ be the initial values that correspond to these parameters.

Step 1a) Draw $\boldsymbol{\beta}^1 \sim [\boldsymbol{\beta}|\tau_0^0, \tau_1^0, \phi^0]$. This step is straight forward to carry out since the full conditional of $\boldsymbol{\beta}$ has a closed-form multivariate normal distribution.

Step 1b) $\tau_0^1 \sim [\tau_0|\boldsymbol{\beta}^1, \tau_1^0, \phi^0]$. A Metropolis algorithm with a two parameter Gamma proposal density is used to generate samples from the full conditional distribution of τ_0 as given below,

- $\tau_0 \sim \tau_0^{\tau_0^0 - 1} \exp\left(-\frac{\tau_0}{\alpha}\right)$,
- $r = \frac{p(\tau_0)}{p(\tau_0^0)} = \frac{L(\tau_0)\pi(\tau_0)}{L(\tau_0^0)\pi(\tau_0^0)}$. The logged version of the expression in (3) is evaluated at τ_0 and τ_0^0 to compute the numerator and the denominator respectively.
- if $r \geq 1$ set $\tau_0^1 = \tau_0$
- if $r < 1$ set

$$\tau_0^1 = \begin{cases} \tau_0 & \text{with probability } r, \\ \tau_0^0 & \text{with probability } 1 - r. \end{cases}$$

Return τ_0^1.

Step 1c) $\tau_1^1 \sim [\tau_1|\boldsymbol{\beta}^1, \tau_0^1, \phi^0]$. This is also a Metropolis step evaluating the logged version of the expression in (4) as the odds ratio.

Step 1d) $\phi^1 \sim [\phi|\boldsymbol{\beta}^1, \tau_0^1, \tau_1^1]$. To insure that ϕ takes values in the interval $[0, 1]$ we consider a Uniform proposal density in the Metropolis step to generate samples from the full conditional of ϕ. That is, $\phi \sim U\left(0, \frac{1}{2}\right)$. The odds ratio for this M-step is calculated by evaluating the logged version of the expression in (5) at ϕ and ϕ^0. Up to this point, one full Gibbs cycle with the Metropolis scheme embedded within Gibbs algorithm is completed. The whole scheme is iterated long enough time to insure convergence of the posterior distribution to its limiting distribution.

The whole setup is dependent on the normality assumptions on the distributions of the suitably transformed pollutant concentrations. However, it may be the case that even after transformation, normality assumptions are not quite satisfied or we would like to work with original dataset. In such situations, we are required to relax the normality assumptions and the methodology presented here won't work quite well. We may need to introduce some semi-parametric or even nonparametric methods instead of a parametric modeling approach.

2.3 Numerical Example: Prediction of Arsenic Concentration

Arsenic concentrations at one or more neighboring locations are assumed to be available. Prediction of concentration levels at unknown location(s) are of interest. Spatial prediction has implications in mapping locations with low (high) arsenic concentrations. Mapping locations with low concentrations may serve as an inexpensive prevention for short term. Data needed to implement spatial prediction of arsenic concentrations include,

i) known concentrations at neighboring locations of one or more unknown location(s),

ii) distances among all locations, in case exponential correlation model is assumed for the covariance matrix,

iii) covariates associated with available concentrations.

Spatial and/or temporal variability in arsenic concentrations, if present and significant, should be considered in the exposure assessment in order to increase accuracy in the estimates. We have addressed spatial variability here and the methods presented can be extended to address temporal variability as well. In the absence of real data from a well designed study, simulated data are used to implement the spatial prediction method described in the previous section. The arsenic concentrations were generated using the model

$$\begin{aligned} Z(s_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i \\ & \quad e_i \sim N(0, \sigma_e^2), \end{aligned}$$

where, x_{1i} = Seasonal index,
x_{2i} = Duration served as water consumption,
x_{3i} = Indicator for nearby chemical dumping site.

The data generation was repeated for three sets of values of the parameters in the above model, $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma_e^2)$, as well as for two different distance matrices. The estimated values of the regression parameters and three variance - covariance parameters in the prediction function along with their standard errors are presented in Table 1.

Table 1 Estimated parameters with standard errors

	10×10 **distance matrix**		
	(1, 1, 1, 0.5, 0.1)	(2, 2, 3, 1.5, 0.5)	(2, 2, 3, 1.5, 1)
β_0	0.8868 (0.0256)	1.9787 (0.1234)	1.8654 (0.3409)
β_1	1.198 (0.0785)	2.0345 (0.5411)	2.1132 (0.7599)
β_2	0.9678 (0.0879)	2.8472 (0.6712)	2.8612 (0.8454)
β_3	0.4473 (0.1289)	1.4364 (0.1113)	1.6353 (.2113)
τ_0	1.8006 (0.4987)	1.7754 (0.4543)	1.8232 (0.3934)
τ_1	2.3362 (0.00813)	2.1225 (0.01781)	2.0034 (0.01123)
ϕ	0.1989 (2.42×10^{-06})	0.2023 (1.44×10^{-04})	0.1897 (3.54×10^{-05})
	20×20 **distance matrix**		
	(1, 1, 1, 0.5, 0.1)	(2, 2, 3, 1.5, 0.5)	(2, 2, 3, 1.5, 1)
β_0	0.9378 (0.0346)	1.7887 (0.2234)	1.7454 (0.1405)
β_1	1.113 (0.0965)	1.9945 (0.3465)	2.0232 (0.6529)
β_2	1.0218 (0.0489)	2.9232 (0.1212)	3.3312 (0.4213)
β_3	0.5573 (0.0893)	1.5364 (0.2114)	1.5863 (.1233)
τ_0	1.0766 (0.3977)	1.8854 (0.5542)	1.7242 (0.2934)
τ_1	2.6232 (0.0613)	1.4225 (0.1781)	2.3421 (0.1823)
ϕ	0.2089 (3.75×10^{-04})	0.3083 (2.65×10^{-03})	0.2197 (4.98×10^{-04})

The simulation results for Bayesian Monte Carlo approach for spatial prediction, show that the regression parameters were estimated with relatively less bias for each set of parameter values. The distance matrices were generated to address moderate amount of spatial correlation which is reflected in the estimates of ϕ. The simulation results show that, the estimates of the regression parameters are almost similar even if the dimension of the distance matrix is doubled. That is, the estimates do not change a lot if we gather concentrations from larger number study sites instead of relatively smaller number of sites. However, in order to draw any conclusion we need to validate this finding through real data on concentration measurements, from a well organized study. The prediction of $z(s_0)$, the arsenic concentration at an unknown point can be carried out straightforwardly by plugging in the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}, \boldsymbol{\tau}, \sigma_0^2$ in the prediction formula given as equation (1) in section 2. Such prediction based on neighboring concentration levels has useful implications in digging new wells for drinking water in arsenic inflicted regions.

3 Discussion

Exposure assessment is an important task of environmental risk assessment that should be done prior to formulating any dose response relationship. Understanding the presence or absence of an agent and its concentration and distribution is the primary focus of exposure assessment. Predictive models and environmental monitoring can be used to determine the levels of exposure at different points. Interpolation of pollutant concentration and spatial prediction method, technically known as kriging, is discussed. Kriging is used to construct a predictor of unobserved concentrations based on the available observed concentrations. Spatial pattern, if any, is captured in the elements of variance covariance matrix as functions of the distances between points at which concentrations are measured. This leads to a structured covariance model among the observations and facilitates estimating only a fixed number of unknown parameters for the predictive model. Numerical implementation of the spatial prediction method through simulated data is presented. However, no real data were available for demonstrating the application of the prediction method. The normality assumption for a suitably transformed pollutant concentration may not hold always. In addition, it is possible that we may have left- or right-censored concentrations in the case of undetectable or imprecise lowest or highest measurements. We will address these more practical issues in the subsequent research.

A generalization to the prediction of pollutant concentrations at several points is straightforward to carry out. However, additional parameters in the variance - covariance matrices are to be estimated jointly with the regression parameters. These computations can be carried out using MCMC techniques as before in the case of prediction at a single point. Extension toward including temporal variability is also possible. An important implication of exposure assessment incorporating spatio - temporal variation is to estimate bioconcentration factor as an unobserved latent factor and integrate this into dose-response modeling of the pollutant. Such an integration is significant in the sense that the bio-concentration factor determines actual response from the exposure to environmental pollutants. Integration of this important determinant in the dose-response model will eliminate uncertainties while extrapolating the results from low dose to high dose or even from species to another. We will address this issue in our future research.

References

1. Gelfand A.E. and A.F.M. Smith; *Sample based approaches to calculating marginal densities*; Journal of the American Statistical Association; **85** (1990); No. 410, 398-409.

2. Gelman A., Trevisani M., Lu H., and Green A. V.; *Direct Data Manipulation for Local Decision Analysis as Applied to the Problem of Arsenic in Drinking Water from Tube Wells in Bangladesh*; Risk Analysis; **24**(2004); No. 6, 1597 - 1612.

3. Geman,S. and Geman,D.; *Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images*; IEEE Trans. Pattern Anal. Mach. Intell; **6** (1984); 721-740.

4. Green A. V., Zheng Y., Versteeg R., Stute M., Horneman A., Dhar R., Steckler M., Gelman A., Small C., Ahsan H., Graziano J.H., Hussain I., and Ahmed K. M.; *Spatial Variability of Arsenic in 6000 Tube Wells in 25 km^2 Area of Bangladesh*; Water Resources Research; **39** (2003); No. 5, HWC 3-1 HWC 3-16

5. Hastings W.K.; *Monte Carlo sampling methods using Markov chain and their applications*; Biometrika; **87** (1970); 97-109.

6. Kinniburg D.G. and Smedley P.L. (editors); *Arsenic contamination of groundwater in Bangladesh*; BGS Technical Report, Volume 1. February 2001.

7. Mardia K.V., Kent J.T. and Bibby J.M.; *Multivariate Analysis*; New York: Academic Press (1979).

8. Solly, W.B., R.R. Pierce, and H.A. Perlman; *Estimated Use of Water in the United States in 1995*; US Geological Survey, 1200; 1998.

9. Symanski E., Savitz D. A., Singer P. C.; *Assessing spatial fluctuations, temporal variability, and measurement error in estimated levels of disinfection by-products in tap water: implications for exposure assessment*; Occupational and Environmental Medicine; **61** (2004); 65 - 72.

10. Winston H. Yu, Charles M. Harvey, Charles F. Harvey; *Arsenic in Groundwater in Bangladesh: A geostatistical and epidemiological framework for evaluating health effects and potential remedies*; Water Resources Research; **39** (2003); No. 6, WES 1-1 WES 1-17.