Published in final edited form as:
Environmetrics. 2013 September 1; 24(6): 377–386. doi:10.1002/env.2218.

Bayesian Spatial Design of Optimal Deep Tubewell Locations in Matlab, Bangladesh

Joshua L. Warren[a,*], **Carolina Perez-Heydrich**[a,b], and **Mohammad Yunus**[c]

[a]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420

[b]Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420

[c]International Centre for Diarrhoeal Disease Research, GPO Box 128, Dhaka 1000, Bangladesh

Summary

We introduce a method for statistically identifying the optimal locations of deep tubewells (dtws) to be installed in Matlab, Bangladesh. Dtw installations serve to mitigate exposure to naturally occurring arsenic found at groundwater depths less than 200 meters, a serious environmental health threat for the population of Bangladesh. We introduce an objective function, which incorporates both arsenic level and nearest town population size, to identify optimal locations for dtw placement. Assuming complete knowledge of the arsenic surface, we then demonstrate how minimizing the objective function over a domain favors dtws placed in areas with high arsenic values and close to largely populated regions. Given only a partial realization of the arsenic surface over a domain, we use a Bayesian spatial statistical model to predict the full arsenic surface and estimate the optimal dtw locations. The uncertainty associated with these estimated locations is correctly characterized as well. The new method is applied to a dataset from a village in Matlab and the estimated optimal locations are analyzed along with their respective 95% credible regions.

Keywords

Approximate likelihood; Environmental health; Markov chain Monte Carlo

1. Introduction

Spatial modeling of the access to public health resources has gained considerable attention in recent years as geographic databases that incorporate spatial, health, and demographic data have become more widespread (e.g., MEASURE DHS, www.measuredhs.com). The optimal allocation of resources associated with disease/exposure mitigation, however, has

[*]Correspondence to: Joshua L. Warren, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420. joshuawa@email.unc.edu.

This paper has been submitted for consideration for publication in *Environmetrics*

Supplementary Materials: Additional information and supplementary material for this article is available online at the journal's website.

been less formally addressed. New statistical methods are required to efficiently allocate resources across an area where demand, defined jointly by exposure risk and population size, varies spatially. In this paper, we develop these statistical methods within the context of arsenicosis mitigation in Matlab, Bangladesh, and show how the proposed approach can be implemented within a resource planning framework.

1.1. Case Study

The population of Bangladesh was inadvertently exposed to high levels of arsenic in the 1970s through the consumption of contaminated groundwater, which went unidentified until the early 1990s. Increased exposure to arsenic concentrations is well known to adversely affect multiple systems of the body including but not limited to the gastrointestinal, renal, cardiovascular, neurological, respiratory, and reproductive systems. Arsenic exposure has also been linked to certain cancers, with the strongest associations found for the skin, lung, and bladder (ATSDR, http://www.atsdr.cdc.gov/). The World Health Organization (WHO) and Environmental Protection Agency established that arsenic concentrations in drinking water exceeding 10 micrograms per liter (μg/l) should be considered unfit for human consumption.

Early in the 1970s, tubewells were installed by the United Nations Children's Fund and the Department of Public Health Engineering in order to provide drinking water to the Bangladesh population that was free of enteric pathogens (Ebi et al. 2005). At the time of installation, the tubewells were not tested for arsenic contamination (Smith et al. 2000); however, at depths of less than 200 meters (m), these tubewells were extracting water that contained high levels of arsenic. Although diarrhea-mediated deaths declined as a result of this shift to groundwater, health problems associated with exposure to naturally occurring arsenic at these aquifer depths raised a new public health concern throughout the country. Smith et al. (2000) described the poisoning of the people of Bangladesh as the largest environmental disaster to ever strike a population, more drastic than the accidents at Chernobyl, Ukraine and Bhopal, India. After testing nearly five million tubewells, it was determined that approximately 30% of the population was exposed to arsenic levels exceeding the WHO standard, while half of those were exposed to levels exceeding the Bangladesh standard of 50 μg/l (BG and BDPHE 2001).

The most effective arsenic mitigation strategy involves the installation of deep tubewells (dtws), which tap into the aquifer at a depth of 203 m or more. Over 165, 000 dtws have been installed throughout the country since 2000 (DPHE and JICA 2009). Dtws can be public or privately-owned, and are relatively expensive to install. Public wells, installed by the government of Bangladesh and non-governmental organizations, are often located near roads to facilitate access for neighboring residents. Installation of these public wells are motivated primarily by arsenic measurements obtained from shallow tubewells (BG and BDPHE 2001).

1.2. Approach

One purpose of this study is to mathematically define theoretically optimal spatial locations for these dtws throughout Matlab, Bangladesh, based on environmental arsenic levels and population sizes/locations. The other is to estimate these optimal dtw locations, based on this definition, using Bayesian spatial statistical modeling. We work in the Bayesian setting because it allows us to correctly characterize the uncertainty associated with the estimates in a flexible manner through the use of hierarchical modeling.

Due to the high expense of installing dtws, it is important to ensure that they are placed in the areas with the most need. This includes areas that have elevated arsenic levels and are located near highly populated baris (patrilineally-related clusters of households). Given only a sample of arsenic concentrations across the region, however, it is difficult to determine where areas of increased arsenic concentrations exist and thus, where dtws should be optimally located. We present an objective function, that when minimized over a region of interest, creates intuitively placed optimal dtw locations. Once the optimal locations are defined theoretically through use of this objective function, we work in the Bayesian setting to statistically model the arsenic concentrations and ultimately obtain samples from the joint posterior distribution of the optimal dtw locations given the observed arsenic data. We then summarize these samples in order to obtain posterior point estimates of the optimal dtw locations as well as 95% credible regions for each estimated location.

Spatial network design problems generally focus on determining the position of future environmental monitors (e.g., air pollution, meteorological) in order to improve the utility of the entire network. The traditional approach focuses on monitor placement which minimizes the average kriging variance (mean squared prediction error) of future predictions over the domain (Zhu and Stein 2005; Zimmerman 2006). More generally, the process attempts to optimize some aspect of the resulting parameter inference. This also includes optimal estimation of the spatial covariance parameters and regression coefficients (Russo 1984; Warrick and Myers 1987; Zimmerman and Homer 1991; Muller and Zimmerman 1999; Zhu and Stein 2005; Zimmerman 2006). Diggle and Lophaven (2006) used a Bayesian analysis to optimize spatial predictions while accounting for the fact that the model parameters were unknown. Entropy based designs have also been utilized in multiple settings (Fuentes et al. 2007). Computationally, Monte carlo techniques have been used in a number of studies to optimize the various objective functions (van Groenigen and Stein 1998; van Groenigen et al. 2000; Lark 2002; Muller et al. 2004).

In this paper, our interest is not in monitoring the arsenic process over the domain. Instead, we seek to minimize the risk of exposure to elevated arsenic levels near highly populated baris by placing dtws in the areas of most need. In the traditional spatial design approach, the kriging variance does not depend on the data. Assuming the spatial covariance parameters are known, the design of the network is completely determined without uncertainty. In our setting, the optimal locations depend on the entire arsenic surface over the domain, which is unknown other than the sampled locations. If the arsenic surface were completely known over the region of interest, there would be no uncertainty associated with the final optimization results and optimization would proceed similarly to the traditional approach. Our model simultaneously incorporates the uncertainty in the unobserved arsenic surface as well as the unknown model parameters, including the spatial covariance parameters, in order to estimate the optimal dtw locations. Working in the Bayesian setting allows us to correctly characterize this introduced uncertainty and ultimately obtain samples from the posterior distribution of the optimal dtw locations.

We begin by exploring the introduced objective function and demonstrating its usefulness over simpler functions that consider only arsenic or population alone. Next, we introduce the statistical model for the arsenic concentrations across Matlab, Bangladesh. Using the modeling results, we design a relevant simulation study to display the benefits of the newly introduced objective function and why incorporating the spatial correlation observed in the arsenic dataset is necessary in order to properly estimate the optimal locations as well as characterize the uncertainty associated with the estimates. We finish with the application of our newly introduced method to the Matlab dataset, identifying the optimal locations of the dtws in a Matlab village assuming that no dtws have been previously installed as well as determining where the next dtw should optimally be located given the actual dtw locations.

The statistical model is introduced in Section 2 and information regarding the developed objective function is discussed in Section 3. Sections 4 and 5 present the included simulation study and application to the Matlab dataset, respectively. We close in Section 6 with the conclusions and discussion.

2. Statistical Model

We model the observed arsenic concentrations on the log scale, with a correction factor at zero, such that

$$Y(\boldsymbol{s}_i)=\ln\{A(\boldsymbol{s}_i)+0.5\}=\mathbf{x}(\boldsymbol{s}_i)^T\boldsymbol{\beta}+w(\boldsymbol{s}_i)+\varepsilon(\boldsymbol{s}_i), i=1,\ldots,n, \quad (1)$$

where $A(\boldsymbol{s}_i)$ is the arsenic concentration at location $\boldsymbol{s}_i$, $\mathbf{x}(\boldsymbol{s}_i)$ is a vector of covariates specific to location $\boldsymbol{s}_i$, $\boldsymbol{\beta}$ represents the vector of regression parameters which relate the covariates to the response, $w(\boldsymbol{s}_i)$ is the purely spatial random error component, and $\boldsymbol{\varepsilon}(\boldsymbol{s}_i)$ is the residual white noise error component. Tables 1 and 2 of the Supplementary Materials Section present results of a sensitivity analysis addressing how the choice of a correction factor, e.g. 0.5 in (1), influences model results and predictions. We find that the results and predictions are not overly sensitive to the choice of this correction factor.

The vector of spatially correlated errors, $w = \{w(\boldsymbol{s}_1), \ldots, w(\boldsymbol{s}_n)\}^T$, is given a prior distribution such that $w \sim$ MVN $(0, \Sigma)$ with $\sum_{i,j}=\text{Cov}\{w(\boldsymbol{s}_i), w(\boldsymbol{s}_j)\}=\sigma_w^2\rho(\|\boldsymbol{s}_i-\boldsymbol{s}_j\||\phi)$. The $\boldsymbol{\rho}(.|\boldsymbol{\phi})$ function depends on the unknown spatial decay parameter, $\boldsymbol{\phi}$, and represents the specified isotropic spatial correlation function, where the correlation between errors depends only on the Euclidean distance between the locations. The choice of the specific correlation function is further explained in Section 5. The remaining error terms are assumed to arise independently from a normal distribution such that $\varepsilon(\boldsymbol{s}_i)\overset{iid}{\sim}\text{N}(0,\sigma_\varepsilon^2)$, where σ_ε^2 represents the nugget effect of the process. The $\mathbf{x}(\boldsymbol{s}_i)$ vector contains an intercept term, the depth of the measurement taken at location $\boldsymbol{s}_i$, and the square of the depth. The form of $\mathbf{x}(\boldsymbol{s}_i)$ is also explained in Section 5. The basic form of the initial arsenic statistical model in (1) was shown to be effective in modeling North Carolina (NC) arsenic levels by Kim et al. (2011).

We define the jointly optimal locations of the dtws within region $\boldsymbol{B}$ as the set of locations, $\boldsymbol{S}$*, which minimizes the objective function

$$Z(\boldsymbol{S}^{*})=\frac{1}{|\boldsymbol{B}|}\int_{\boldsymbol{B}}\gamma(\boldsymbol{s})\min\{\|\boldsymbol{s}_{i}^{*}-\boldsymbol{s}\|:i=1,\ldots,m\}d\boldsymbol{s}, \quad (2)$$

where $\boldsymbol{S}^{*}=(\boldsymbol{s}_{1}^{*},\ldots,\boldsymbol{s}_{m}^{*})^{T},\boldsymbol{s}_{i}^{*}\in\boldsymbol{B}$, m represents the fixed number of dtws to be installed in region $\boldsymbol{B}$, and

$$\gamma(\boldsymbol{s})=\frac{A(\boldsymbol{s})\text{Popn}\{\boldsymbol{b}_{(\min)}\}}{\min\{\|\boldsymbol{b}_{j}-\boldsymbol{s}\|:j=1,\ldots,n_{b}\}+\delta}. \quad (3)$$

The $\boldsymbol{b}_j$ term represents the location of bari j in region $\boldsymbol{B}$, $\boldsymbol{b}_{(\min)}$ is the bari location which minimizes $\|\boldsymbol{b}_{\text{j}} - \boldsymbol{s}\|$ for a given location $\boldsymbol{s}$, Popn {.} is a function that returns the population of the input bari location, and n_b is the number of baris in region $\boldsymbol{B}$. The $\delta > 0$ term allows γ ($\boldsymbol{s}$) to be defined for all $\boldsymbol{s}$, even in the case where $\boldsymbol{s} = \boldsymbol{b}_j$. The set of optimal dtw locations is denoted by $\boldsymbol{S}_{(\text{opt})}^{*}=\min^{-1}\{Z(\boldsymbol{S}^{*}):\boldsymbol{s}_{i}^{*}\in\boldsymbol{B},i=1,\ldots,m\}$. More information regarding the form of γ ($\boldsymbol{s}$) as well as the intuitive benefits and properties of the optimal locations obtained by minimizing the objective function in (2) with respect to a fully specified arsenic surface are discussed in Section 3.

In real data settings, we do not observe a sufficient number of arsenic observations within region to obtain an accurate approximation of $\boldsymbol{S}_{(\text{opt})}^{*}$, which requires full knowledge of the arsenic surface. We overcome this by statistically modeling and predicting the arsenic surface across the region. In the Bayesian setting, our interest lies in $f\{\boldsymbol{S}_{(\text{opt})}^{*}|\boldsymbol{Y}\}$, the posterior distribution of the set of jointly optimal dtw locations. Using Markov chain Monte Carlo (MCMC) techniques, we are able to obtain samples from this posterior distribution and then summarize them to conduct inference on the unknown optimal locations of interest.

2.1. Prior Specification

We complete the model specification by assigning prior distributions to the unknown model parameters. The $\boldsymbol{\beta}$ parameters are given independent, normal prior distributions with large, fixed prior variances. This essentially results in a flat prior for these parameters. The nugget effect parameter, σ_{ε}^{2}, and the partial sill parameter, σ_{w}^{2}, are each given independent, vague Uniform(0.1, 5.0) prior distributions to reflect our lack of prior information about their values. For reference, the sample variance of the observed transformed arsenic data is 5.59 (mean: 3.97, median: 5.17) and our priors for σ_{ε}^{2} and σ_{w}^{2} allow for the prior variance to vary from 0.2 to 10. We perform a sensitivity analysis, refitting the model using independent gamma prior distributions with a mean and variance of one for σ_{ε}^{2} and σ_{w}^{2}. No substantial differences are seen in the posterior distributions of the model parameters when theses priors are used (results shown in Table 3 of the Supplementary Materials Section). The inverse of the spatial autocorrelation parameter, $\boldsymbol{\phi}$, is given an independent Uniform(0.006, 6.214) prior distribution which also reflects our lack of initial information regarding the parameter. This allows the prior correlation at the average observed distance between shallow tubewells (7.591 kilometers (km)) to vary from 0.00 to 0.95, where distances range from 0.002 km to 21.789 km. This also allows the prior spatial range $(\approx \frac{3}{\phi})$ to vary from 0.5 km to 482.8 km.

2.2. Inference for Optimal Deep Tubewell Locations

Once we obtain samples from the posterior distribution of the optimal dtw locations, $f\{\boldsymbol{S}^*_{(\mathrm{opt})}|\boldsymbol{Y}\}$, we must summarize them in order to create posterior point estimates and credible regions for the optimal locations. We use the k-means procedure (Hartigan and Wong 1979) to identify cluster centers, which serve as posterior point estimates for the optimal locations. The k-means method is a clustering algorithm which partitions the data into a fixed number of groups where each datapoint is associated with the closest cluster centroid. Typically with k-means, a common concern is choosing the appropriate number of clusters in the data. In our situation, we avoid this complication by knowing beforehand the specified number of dtw locations being considered. This value is *m* in (2), the number of dtws to be installed in the region.

Once point estimates are obtained, we also require measures of uncertainty related to these optimal locations. To achieve this, we calculate 95% elliptical credible regions for each cluster of the form $\{\boldsymbol{S}^*_{(\mathrm{opt})}(i) - \hat{\boldsymbol{\mu}}_i\}^T \hat{\sum}_i^{-1} \{\boldsymbol{S}^*_{(\mathrm{opt})}(i) - \hat{\boldsymbol{\mu}}_i\} \leq c$ where *c* is chosen such that 95% of the posterior samples associated with a cluster are contained in the region, $\hat{\boldsymbol{\mu}}_i$ is the estimated mean vector of cluster *i*, $\hat{\Sigma}_i$ is the estimated covariance matrix of cluster *i*, and $\boldsymbol{S}^*_{(\mathrm{opt})}(i)$ is the optimal location of cluster *i*. The estimation of the mean vector and covariance matrix improves as the number of posterior samples obtained increases and therefore, the resulting uncertainty in these estimates can be made arbitrarily small. We repeat this process and obtain a credible region for each optimal dtw location. We study the coverage probabilities of the proposed regions in the simulation study of Section 4. Figure 1 in the Supplementary Materials Section displays the posterior samples from our model application along with the posterior point estimates identified by the k-means algorithm and the calculated 95% elliptical credible regions.

2.3. Fitting Algorithm

We use an approximate likelihood method introduced by Vecchia (1988) and further developed by Pardo-Igúzquiza and Dowd (1997) due to the computational demand of working with 10,376 spatially referenced arsenic observations in our analysis. This method allows for the use of the entire dataset as opposed to more common subsampling techniques which ease the computational burden by simply removing a portion of the data. In the usual Bayesian spatial model, the inverse and determinant of an *n* by *n* matrix is required during each MCMC iteration, where *n* represents the sample size of the dataset. These calculations can be problematic for a large *n*. The approximate likelihood method avoids this complication by replacing the original likelihood, $f(\boldsymbol{Y}|\boldsymbol{\theta})$, with $\prod_{i=1}^{n} f\{Y(s_i)|\boldsymbol{\theta}, z_i\}$, where $\boldsymbol{z}_i$ is the vector of observations from the *l* closest (spatially) observed locations to location $\boldsymbol{s}_i$ within the dataset and $\boldsymbol{\theta}=(\beta^T, \sigma_w^2, \sigma_\varepsilon^2, \phi)^T$. This approximation results in the need for *n* inverse and determinant calculations of l by l matrices for each MCMC iteration, which represents a computational improvement over the original situation. The idea behind this approximation is that after a certain distance, the remaining information regarding the spatial structure of the data is often superfluous or redundant. The key for this approximation to be successful is in choosing l large enough to account for the relevant neighboring observations but small enough to allow for computational efficiency. Results from Pardo-Igúzquiza and Dowd (1997) suggest that $l = 10$ to 15 is adequate. Based on this recommendation and exploring subsets of our data, we choose $l = 15$ in the analysis.

We begin by modeling the observed arsenic concentrations given the model in (1) while implementing the approximate likelihood method. From this modeling, we obtain samples from $f(\boldsymbol{\theta}|\boldsymbol{Y})$. Using these samples, we obtain samples from the posterior predictive distribution (ppd) of $\boldsymbol{Y}_0|\boldsymbol{Y}$, a vector of transformed arsenic concentrations at unobserved locations across the region of interest. This ppd is given as $f(\boldsymbol{Y}_0|\boldsymbol{Y}) = \int f(\boldsymbol{Y}_0|\boldsymbol{Y}, \boldsymbol{\theta})\, f(\boldsymbol{\theta}|\boldsymbol{Y})\, d\boldsymbol{\theta}$ where $f(\boldsymbol{Y}_0|\boldsymbol{Y}, \boldsymbol{\theta})$ is the probability distribution function of a multivariate normal distribution with known mean and covariance. We once again implement the approximate likelihood method since prediction requires the inverse of an n by n matrix during each iteration. We replace $f(\boldsymbol{Y}_0|\boldsymbol{Y}, \boldsymbol{\theta})$ with $f(\boldsymbol{Y}_0|\boldsymbol{\theta}, \boldsymbol{z}_0)$, where $\boldsymbol{z}_0$ is the vector of the l observations from the closest observed spatial locations to each entry of the $\boldsymbol{Y}_0$ vector, with duplicates removed. Therefore, $\boldsymbol{z}_0$ can contain anywhere from l to $l * n_0$ observations, where n_0 is length of the $\boldsymbol{Y}_0$ vector. Using composition sampling (Banerjee et al. (2004)), we are able to obtain a sample from this ppd using a sample from the posterior distribution of $\boldsymbol{\theta}$. By applying the inverse of the transformation function used in (1), we obtain ppd samples from $f(\boldsymbol{A}_0|\boldsymbol{Y})$, arsenic concentrations at unobserved locations in the region.

These ppd samples are obtained from a large number of locations within a specified region resulting in the usual interpolated surface of arsenic concentrations found using the method of Bayesian kriging (Handcock and Stein 1993). Spatially filling in the domain allows us to then accurately approximate $\boldsymbol{S}^*_{(\text{opt})}$. We use the OPTIM function within R's *utils* package (R Development Core Team 2012) to minimize $Z(\boldsymbol{S}^*)$ with respect to $\boldsymbol{S}^*$ for each set of joint ppd samples. Therefore, we once again use composition sampling to obtain a posterior sample from $f\{\boldsymbol{S}^*_{(\text{opt})}|\boldsymbol{Y}\}$ for each ppd sample of jointly predicted arsenic surface.

3. Optimal Deep Tubewell Objective Function

Minimizing the introduced objective function in (2), with respect to $\boldsymbol{S}^*$, creates intuitively located dtws across a selected region. When creating the objective function, we attempt to define the most at risk locations across a region and then construct the objective function around this definition. Placement of dtws in these high risk areas ensures that nearby populations have access to an arsenic free water supply since they tap into the aquifer at depths of at least 203 m. We decide that the locations with the highest risk are those with high arsenic concentrations, located only a short distance from a highly populated bari. The introduced objective function attempts to take this definition into account through the use of $\gamma(\boldsymbol{s})$ in (3). This term is large when the arsenic level at location s is high and/or the location is very close to a bari with a large population.

In the denominator of $\gamma(\boldsymbol{s})$, min $\{\|\boldsymbol{b}_\mathrm{j} - \boldsymbol{s}\| : j = 1, \ldots, n_\mathrm{b}\}$ decreases as $\boldsymbol{s}$ gets closer to a bari location, causing $\gamma(\boldsymbol{s})$ to increase. This leads to dtw placements generally closer to a bari location. At the same time, if the population of this closest bari to location $\boldsymbol{s}$ (Popn $\{\boldsymbol{b}_{(\mathrm{min})}\}$) is large, $\gamma(\mathbf{s})$ will increase further which favors dtw locations near highly populated baris. Finally, larger arsenic values ($A(\boldsymbol{s})$) will also increase the numerator of $\gamma(\boldsymbol{s})$, leading to dtw placements closer to these regions with increased arsenic levels. Given two locations the same distance from their closest baris which have the same population, $\gamma(\boldsymbol{s})$ will give an elevated value to the location in the higher arsenic area. This can be seen since only the $A(\boldsymbol{s})$ term will be different in their respective calculations. Similar intuitive properties emerge as the other terms in $\gamma(\boldsymbol{s})$ are fixed and a single factor is allowed to vary.

For large values of $\gamma(\boldsymbol{s})$, we prefer a dtw to be placed very close to location $\boldsymbol{s}$, thereby causing $\min\{\|\boldsymbol{s}_i^* - \boldsymbol{s}\|:i=1,\ldots,m\}$ in (2) to be small. For smaller values of $\gamma(\boldsymbol{s})$, we can afford to place dtws further from location $\boldsymbol{s}$ since $\gamma(\boldsymbol{s})$ will weaken the impact of $\gamma(\boldsymbol{s})\min\{\|\boldsymbol{s}_i^* - \boldsymbol{s}\|:i=1,\ldots,m\}$ in (2).

Examples of simulated, full arsenic surfaces (log scale) and the corresponding optimal placement of dtws using this objective function are shown in Figures 1 and 2. We show how five dtws would be positioned using our objective function under different simulated arsenic surfaces and specified locations/population sizes of 10 baris. These examples assume the entire arsenic surface is known, therefore, no statistical modeling is required.

We also display the optimal locations identified using an objective function which only considers bari locations/populations by allowing $\gamma_1(\boldsymbol{s}) = \text{Popn}\ \{\boldsymbol{b}_{(\mathrm{min})}\} / \min\{\|\boldsymbol{b}_j - \mathbf{s}\| : j = 1,\ldots, n_b\}$, and an objective function which only considers arsenic, by allowing $\gamma_2(\boldsymbol{s}) = A(\boldsymbol{s})$. The intuitive benefits of the newly constructed $\gamma(\mathbf{s})$ function in (3) are evident when compared to these similar functions which only consider a single factor at a time.

Figure 1 shows the optimal locations identified using the $\gamma_1(\boldsymbol{s})$, $\gamma_2(\boldsymbol{s})$, and $\gamma(\boldsymbol{s})$ functions under the assumption of a constant arsenic surface over the region. In this situation, minimizing $Z(\boldsymbol{S}^*)$ results in a geographically balanced set of locations across the region for the objective function using $\gamma_2(\boldsymbol{s})$. This is true since $\gamma_2(\boldsymbol{s})$ ignores the bari locations/populations and considers arsenic concentrations alone. The optimal locations for $\gamma(\boldsymbol{s})$ and $\gamma_1(\boldsymbol{s})$ are identical since under the constant arsenic surface assumption $\gamma(\boldsymbol{s}) = c * \text{Popn}\ \{\boldsymbol{b}_{(\mathrm{min})}\} / \min\{\|\boldsymbol{b}_j - \mathbf{s}\| : j = 1, \ldots, n_b\} = c * \gamma_1(\boldsymbol{s})$ where $c = A(\boldsymbol{s})\ \forall \boldsymbol{s}$. These optimal dtw locations are therefore based on the population and locations of the baris in the region alone. The results seen using $\gamma(\boldsymbol{s})$ are what we would expect under the constant arsenic surface since they depend on the baris alone.

Figure 2 shows the dtw placement under a more realistic simulated arsenic surface. The optimal locations are now shifted towards the higher arsenic areas for $\gamma_2(\boldsymbol{s})$, ignoring the bari populations/locations. For $\gamma_1(\boldsymbol{s})$, the optimal locations are identical to the Figure 1 results since arsenic is not a factor, only population. For $\gamma(\boldsymbol{s})$, it is clear that the locations are shifted towards the regions with higher arsenic concentrations while also favoring close proximity to the baris. Using $\gamma(\boldsymbol{s})$ places dtws in these highly populated/high arsenic areas and represents an intuitive improvement over $\gamma_1(\boldsymbol{s})$ and $\gamma_2(\boldsymbol{s})$.

Figures 1 and 2 graphically display the interpretability of the introduced metric and its ability to place dtws in locations which are in the most need. In these examples, we simulate the data over the entire region of interest, generating enough observations to heavily fill in the region and accurately approximate $\boldsymbol{S}^*_{(\mathrm{opt})}$. Increasing the number of simulated observations improves the approximation since $\hat{Z}(\boldsymbol{S}^*)=\sum_{j=1}^{n}\gamma(\boldsymbol{s}_j)\min\{\|\boldsymbol{s}_i^*-\boldsymbol{s}_j\|:i=1,\ldots,m;\boldsymbol{s}_j\in\boldsymbol{B}\}/n$ is a Monte Carlo integration for $Z(\boldsymbol{S}^*)$.

4. Simulation Study

We attempt to create scenarios which are similar to results seen when working with the arsenic dataset as we choose our simulation settings. In the Matlab dataset, the median number of dtws which are currently installed within a single village is two. We therefore consider two dtw locations in the simulation study ($m = 2$). We allow the number of observed arsenic measurements within the village to vary in the study by introducing three sample size settings (SS) such that

- SS 1: 3 arsenic measurements observed within the village (25th percentile of the number of arsenic measurements within the villages),
- SS 2: 63 arsenic measurements observed within the village (50th percentile of the number of arsenic measurements within the villages),
- SS 3: 263 arsenic measurements observed within the village (75th percentile of the number of arsenic measurements within the villages).

We use the true bari locations and population sizes for an actual village in Matlab.

We begin by simulating the true arsenic surface across the village of interest. First, we create the spatially correlated and white noise error processes from (1) using results from the modeling of the observed arsenic data. We choose the exponential spatial correlation structure and the mean of the arsenic surface using the analysis of the observed data as a reference. We allow $E\{Y(\boldsymbol{s})\} = \beta_0 + \beta_1 d(\boldsymbol{s}) + \beta_2 d(\boldsymbol{s})^2$, where $d(\boldsymbol{s})$ is the depth (m) of the arsenic measurement at location $\boldsymbol{s}$. The specific settings of $\boldsymbol{\beta}$, σ_w^2, σ_ε^2, and $\boldsymbol{\phi}$ can be seen in the posterior mean column of Table 3. The choice of covariance function and mean form are further discussed in Section 5.

Two versions of the mean surface are created. First we choose $d(\boldsymbol{s})$ based on the distribution of depths seen in the observed data using a N (120, 20) distribution as an approximation. Next, we set $d(\boldsymbol{s}) \equiv 98$ for all $\boldsymbol{s}$. We choose 98 m because this is the median depth of all shallow tubewells in the observed data. Therefore, this is a common depth from which villagers across Matlab obtain drinking water. In the application of our method, we observe arsenic concentrations at varying depths across a region. Thus, the effect of measurement depth must be controlled for when predicting the arsenic concentration across a region. We therefore work at the 98 m level to determine where high arsenic clusters are located, after controlling for the depth of the measurement.

In total, we simulate 20,000 arsenic concentrations, enough to thoroughly cover the specified village, and also simulate 1,000 values within surrounding villages. For the purposes of our simulation study, we select one of the 142 villages included in the dataset (shown in Figure 3). Choosing an actual village ensures that we are working with distances that are similar to those observed in the data. Using the complete set of simulated within-village arsenic values, we are able to accurately approximate $\boldsymbol{S}^*_{(\mathrm{opt})}$ and treat these locations as the truth for our simulation study. This approximation of the truth is based on the arsenic surface with depth set at 98 m. Next, we randomly sample arsenic values within the region using a particular sample size setting (SS 1, SS 2, SS 3). These sampled arsenic concentrations are located at various depths across the village, not only at d ($\boldsymbol{s}$) = 98. Therefore, this simulated dataset closely resembles the scenario seen in the actual observed arsenic dataset. Given this sample of values, we then apply each method to the dataset. We first predict the arsenic surface at d ($\boldsymbol{s}$) = 98 and then estimate $\boldsymbol{S}^*_{(\mathrm{opt})}$ as described in Section 2.

We compare the new method for determining optimal dtw locations with two alternative techniques. All methods are compared across different sample size settings within a region of interest. The three methods include

- Method 1: Using only the observed arsenic values within the region to estimate the optimal dtw locations,
- Method 2: Newly developed method based on Bayesian modeling and spatial interpolation of the arsenic surface within a region and obtaining the ppd of the optimal dtw locations,
- Method 3: Spatially independent method based on Bayesian modeling and interpolation of the arsenic surface (ignoring spatial correlation) within a region and obtaining the ppd of the optimal dtw locations.

Each of these methods utilize some form of the objective function in (2) to determine the optimal locations. Method 1 does not directly involve statistically modeling and/or predicting the arsenic values across the region. It represents a deterministic technique for determining the optimal dtw locations within a region with Z ($\boldsymbol{S}$*) being replaced by $\hat{Z}$ ($\boldsymbol{S}$*) such that $\hat{Z}(\boldsymbol{S}^*)=\sum_{j=1}^{n}\gamma(\boldsymbol{s}_j)\min\{\|\boldsymbol{s}_i^*-\boldsymbol{s}_j\|:i=1,\ldots,m;\boldsymbol{s}_j\in\boldsymbol{B}\}/n$, where n is the number of observed arsenic values within region $\boldsymbol{B}$. As the sample size within the region increases, the approximation to the true optimal dtw locations of Method 1 will improve as described in Section 3. With smaller, more realistic sample sizes however, this method will struggle tremendously. Method 1 represents a naive attempt at determining the optimal locations without considering the statistical modeling of the arsenic surface and serves as a baseline method for comparison purposes.

Method 2 represents the new technique described thoroughly in Section 2. Method 3 is similar to Method 2 in that it attempts to statistically model and predict the unknown arsenic surface using the observed arsenic values. The difference is that Method 3 ignores the spatial correlation present in the arsenic data and therefore represents a basic multiple regression fitting and prediction of the data. This method demonstrates why care should be taken to correctly characterize the spatial association present in the arsenic data.

Once each method is applied to a simulated dataset, we collect three different pieces of information from the output. For each method, we determine the average distance (km) that the posterior point estimates of the optimal dtw locations are located from the true optimal locations. These data give us insight into how well each method is able to estimate the optimal dtw locations. For methods 2 and 3, we also collect the average area of the elliptical 95% credible regions and the number of the true dtw locations which fall within these regions. This allows us to determine if the resulting inference regarding the dtw locations is accurate for these methods. We prefer credible regions with smaller areas while maintaining the correct coverage probability.

In order to compare the performance of the three methods, we investigate the association between performance metrics (i.e., distance, area, and coverage) and method type. Data associated with each performance metric are analyzed separately using mixed effects models with a random effect corresponding to the simulated dataset. This allows observations from different methods, but the same sample size setting and dataset, to be positively correlated. Including a random effect term, therefore, controls for the fact that each method was applied to the same generated dataset under a particular sample size setting.

In Table 1, we display the results from the average distance analysis for each method and sample size setting. There is a significant interaction effect between method and sample size present in the data; however, Method 2 most consistently estimates optimal locations that are closest to the true optimal locations. Each of the estimates from Method 2 in Table 1 are significantly smaller than the respective estimates obtained from Method 1. As expected, Method 1 struggles greatly for small sample sizes but improves steadily as sample size increases, though it still is outperformed by Method 2. Method 3 stays fairly consistent for each sample size setting, and is outperformed by Method 2 under SS 2 and SS 3. For SS 1, the estimates from methods 2 and 3 are not statistically different. Method 2 is never statistically outperformed by either of the alternative methods.

In Table 2, the estimates for the average area (m^2) of the credible regions are displayed. In the ideal situation, we prefer a region with small area and excellent coverage probabilities. It is clear that Method 2 produces larger credible regions than Method 3 but as a result, the coverage probabilities are significantly higher for Method 2 even at the largest sample size setting (Method 2, SS3: 0.93 (0.02); Method 3, SS3: 0.51 (0.04)). As the sample size increases however, estimates from both methods become more comparable. For SS 3, the estimates are not significantly different while for SS 1 and SS 2 Method 3 produces significantly smaller credible regions. Overall, combining the area and coverage analysis results for the credible regions, it is clear that Method 2 is preferred due to its ability to consistently cover the truth. Method 2 also produces posterior estimates which are much closer to the truth on average.

5. Application to Matlab, Bangladesh Dataset

We analyze a dataset of arsenic concentrations from tubewells across the 142 villages in Matlab, a rural region in Bangladesh with a population of over 220,000. Within villages, households are situated in patrilineal clusters called baris, which are all included in a geographic database of the region. Records of all Matlab residents upon birth or migration into the study area have been maintained by a health and demographic surveillance system (HDSS) since 1966. We obtain population estimates from this HDSS. Between 2002 and 2004, a comprehensive survey of Matlab's 12,018 shallow tubewells was conducted. Surveyors collected tubewell locations using global positioning system receivers and information on tubewell depth from well owners. Samples from these tubewells were tested for arsenic. In 2009, community health workers collected information on the locations of dtws by asking householders to identify any deep community tubewell in or adjacent to their bari. Dtws were assigned the same geographic coordinates as the closest bari, and were added to the existing tubewell database.

Tubewells with missing depth data are omitted from the analysis. The final dataset of observed arsenic concentrations consists of 10,376 observations. The mean arsenic concentration across shallow tubewells for all of Matlab is 209.59 μg/l (sd: 216.70; median: 175.75). Arsenic concentrations range from 0 to 3,644 μg/l. The mean depth of the shallow tubewells from which the arsenic concentrations were obtained is 132.01 m (sd: 94.74; median: 98), and the measurement depth ranges from 0 to 980 m.

We apply the newly introduced method to the Matlab dataset. All of the displayed results are based on 5,000 samples from the posterior distribution of the optimal dtw locations after a burnin period of 10,000 samples. The analysis is carried out using R statistical software (R Development Core Team 2012).

We begin by fitting the proposed arsenic statistical model in (1), using the approximate likelihood method detailed in Section 2.3. An important indicator of the arsenic concentration at a shallow tubewell is the depth of the tubewell. We include an intercept term, the depth of the shallow tubewell, and the quadratic depth term in the mean function of the arsenic model. We select the isotropxic exponential spatial covariance function based on deviance information criterion (Spiegelhalter et al. 2002) comparisons with other common covariance structures (Table 4 of the ays the posterior Supplementary Materials Section) as well as its use by Kim et al. (2011) in modeling NC arsenic data. This leads to $\text{Cov}\{w(\boldsymbol{s}_i), w(\boldsymbol{s}_j)\}=\sigma_w^2\text{exp}\{-\phi\|\boldsymbol{s}_i-\boldsymbol{s}_j\|\}$. Table 3 displays posterior summaries for the model parameters while Figure 2 in the Supplementary Materials Section displays the posterior means and standard deviations of the predicted transformed arsenic surface over the village.

Once we model the arsenic surface, we then determine the optimal locations for dtws within a village in Matlab. We use the current placement of the dtws to determine how many wells to install in the village (choosing *m).* We therefore allow $\boldsymbol{B}$, from (2), to represent the specific village of interest. The results from the new method are shown in Figure 3 with the original map of current dtw locations also displayed. Each figure includes the predicted arsenic surface (log scale) along with the bari locations within the village (represented by clear circles). The larger circles indicate higher population within the baris. The cluster centers, identified by k-means, are displayed as white circles and represent the optimal posterior point estimate locations in Figure 3. Overall, it is clear that the new method favors a balance of tubewells over the specified village with respect to arsenic concentrations and population. The current dtws appear to be placed with respect to population size alone, failing to account for the elevated arsenic exposures observed in the northeastern part of the village.

In Figure 4, we display the posterior point estimate, actual dtw location, and 95% elliptical credible region for each optimal location. Figure 4 gives us better insight into the uncertainty associated with the posterior distribution of the optimal dtw locations. Five of the current dtw locations fall within the 95% credible regions, but based on the newly introduced objective function, two of the dtws are likely installed in suboptimal locations. This method allows us to correctly characterize the uncertainty associated with our estimates rather than relying on a deterministic method where only point estimates are produced, such as Method 1.

Next, we determine the optimal location of a proposed dtw installation given the locations of other dtws in the area. This allows our method to be adapted based on the current locations of dtw installations. We modify our method by setting the current dtw locations as fixed in (2) and allowing only one location to vary across the village. These results can be seen in Figure 5 which displays the current dtw locations along with the posterior mean estimate of the optimal location for the next monitor to be installed. As expected, the next monitor is estimated to be placed among the baris in the higher arsenic areas. This process could easily be extended to estimate more than one future optimal dtw given the current dtw locations as well.

6. Discussion

We presented a statistical model in the Bayesian setting which allows for the joint identification of optimal dtw locations across a region of interest. Through the simulation study results, we showed that the newly presented approach outperforms competing methods in terms of location estimation and coverage probability of the credible regions. The application to the Matlab dataset also demonstrated its utility in finding optimal locations for dtw installations in areas that either have no installations in place or are adding to their existing assemblage of dtws.

More generally, this modeling framework could be used in devising planning strategies for the efficient distribution of resources across a variety of disease mitigation settings. Such examples include the distribution of sanitation resources with respect to waterborne disease risk, or bed nets with respect to malaria risk. Essentially, this method could be applied to any system for which the optimal distribution of resources is defined jointly according to a risk profile interpolated from spatially explicit exposure data, and covariates expected to be important to the decision-making process. In our study, we defined arsenic as the exposure of interest and population size as a covariate driving the placement of resources (i.e., dtws), but this can also be extended in future work to include other factors of interest such as distance from a main road and/or diarrheal disease incidence.

One such system for which considerable work has been conducted on the issue of optimal design is air pollution monitoring networks. For example, alternative methods for determining the optimal placement of air pollution monitors have been previously suggested (Kanaroglou et al. 2005) which can also incorporate factors such as land use, transportation infrastructure, and population density in the determination of optimal sampling locations. However, whereas the focus of the present study was on optimizing resource allocations, the focus of air pollution monitoring studies has been on optimizing sampling design. Generally, previous studies have focused on defining monitor placement in terms of reducing measurement error (Roberts 1984), and have incorporated the spatial covariance structure of air quality patterns to determine the location of monitors while minimizing redundancies in air pollution measurements (Langstaff et al. 1967; Shindo et al. 1990; Arbeloa et al. 1993; Mofarrah and Husain 2010).

A possible limitation of the newly introduced method is the computational expense associated with the particular choice of m from (2). As m increases, the amount of time needed to optimize the objective function over the domain of interest also increases. This can be problematic since this optimization must take place for each MCMC sample obtained in the analysis. We also assume that m is known *a priori*. Future work could attempt to leave m as a random quantity to be estimated from the data, though this is not trivial since a threshold definition would have to be created based on the optimized value of the objective function. In practice, however, the value of m is likely to be fixed and determined by budgetary constraints.

Overall, the introduced method represents a statistically sound way to estimate optimal dtw locations across a region while also correctly characterizing the uncertainty associated with these locations. Providing credible regions allows flexibility in determining the exact location of the installation while still ensuring probable optimality of the design. These optimal dtws are placed in intuitive locations across a region with respect to arsenic and population size. The current process of installing dtws appears to favor population and proximity to baris alone which could be problematic for people living in less populated, high arsenic areas. Targeting exposure mitigation efforts where they will be most effective will likely involve consideration of exposure risk along with at-risk population characteristics. The framework we provide here incorporates these factors along with the spatial uncertainty associated with the risk profile, and can be widely applied across resource planning settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants from the National Institute of Environmental Health Sciences (T32ES007018) and National Institute of Child Health and Human Development (R24 HD050924), and by core donors which provide unrestricted support to icddr,b for its operations and research. Current donors providing unrestricted support include: Australian Agency for International Development (AusAID), Government of the Peoples Republic of Bangladesh; Canadian International Development Agency (CIDA), Swedish International Development Cooperation Agency (Sida), and the Department for International Development, UK (DFID). We gratefully acknowledge these donors for their support and commitment to icddr,b's research efforts.

References

Arbeloa FJS, Caseiras CP, Andres PML. Air-quality monitoring – optimization of a network around a hypothetical potash plant in open countryside. Atmospheric Environment Part A – General Topics. 1993; 27(5):729–738.

Banerjee, S.; Carlin, BP.; Gelfand, AE. Hierarchical Modeling and Analysis for Spatial Data. Boca Raton: Chapman & Hall/CRC; 2004.

British Geological Survey and Bangladesh Department of Public Health Engineering. Final Technical Report WC/00/19. British Geological Survey; Keyworth, UK: 2001. Arsenic contamination of groundwater in Bangladesh.

Department of Public Health Engineering and Japan International Cooperation Agency. Report on Situation Analysis of Arsenic Mitigation. Dhaka, Bangladesh: Tech. rep., Department of Public Health Engineering; 2009.

Diggle P, Lophaven S. Bayesian geostatistical design. Scandinavian Journal of Statistics. 2006; 33(1): 53–64.

Ebi, L.; Mills, D.; Smith, J. chap A Case Study of Unintended Consequences: Arsenic in Drinking Water in Bangladesh. London: Taylor & Francis.; 2005. Integration of Public Health with Adaptation to Climate Change: Lessons Learned and New Directions; p. 72-90.

Fuentes M, Chaudhuri A, Holland DM. Bayesian entropy for spatial sampling design of environmental data. Environmental and Ecological Statistics. 2007; 14(3):323–340.

Handcock MS, Stein ML. A Bayesian analysis of kriging. Technometrics. 1993; 35(4):403–410.

Hartigan J, Wong M. Algorithm AS 136: a k-means clustering algorithm. Journal of the Royal Statistical Society Series C (Applied Statistics). 1979; 28:100–108.

Kanaroglou PS, Jerrett M, Morrison J, Beckerman B, Arain MA, Gilbert NL, Brook JR. Establishing an air pollution monitoring network for intra-urban population exposure assessment: a location-allocation approach. Atmospheric Environment. 2005; 39(13):2399–2409.

Kim D, Miranda M, Tootoo J, Bradley P, Gelfand A. Spatial modeling for groundwater arsenic levels in North Carolina. Environmental Science and Technology. 2011; 45:4824–4831. [PubMed: 21528844]

Langstaff J, Seigneur C, Mei-Kao L, Behar J, McElroy JL. Design of an optimum air monitoring network for exposure assessments. Atmospheric Environment (1967). 1967; 21(6):1393–1410.

Lark R. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. Geoderma. 2002; 105(1-2):49–80.

Mofarrah A, Husain T. A holistic approach for optimal design of air quality monitoring network expansion in an urban area. Atmospheric Environment. 2010; 44(3):432–440.

Muller P, Sanso B, De Iorio M. Optimal Bayesian design by inhomogeneous Markov chain simulation. Journal of the American Statistical Association. 2004; 99(467):788–798.

Muller W, Zimmerman D. Optimal designs for variogram estimation. Environmetrics. 1999; 10(1):23–37.

Pardo-Igúzquiza E, Dowd PA. AMLE3D: A computer program for the inference of spatial covariance parameters by approximate maximum likelihood estimation. Computers and Geosciences. 1997; 23(7):793–805.

R Development Core Team. R: A Language and Environment for Statistical Computing. Vol. ISBN 3-900051-07-0. Vienna, Austria: R Foundation for Statistical Computing; 2012. URL http://www.R-project.org

Roberts EM. Design methodology for optimum dosage air monitoring site selection. Atmospheric Environment (1967). 1984; 18(6):1243–1244.

Russo D. Design of an optimal sampling network for estimating the variogram. Soil Science Society of America Journal. 1984; 48(4):708–716.

Shindo J, Oi K, Matsumoto Y. Considerations on air pollution monitoring network design in the light of spatio-temporal variations of data. Atmospheric Environment Part B Urban Atmosphere. 1990; 24(2):335–342.

Smith A, Lingas E, Rahman M. Contamination of drinking-water by arsenic in Bangladesh: a public health emergency. Bulletin of the World Health Organization. 2000; 78(9):1093–1103. [PubMed: 11019458]

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002; 64(4):583–639.

van Groenigen J, Pieters G, Stein A. Optimizing spatial sampling for multivariate contamination in urban areas. Environmetrics. 2000; 11(2):227–244.

van Groenigen J, Stein A. Constrained optimization of spatial sampling using continuous simulated annealing. Journal of Environmental Quality. 1998; 27(5):1078–1086.

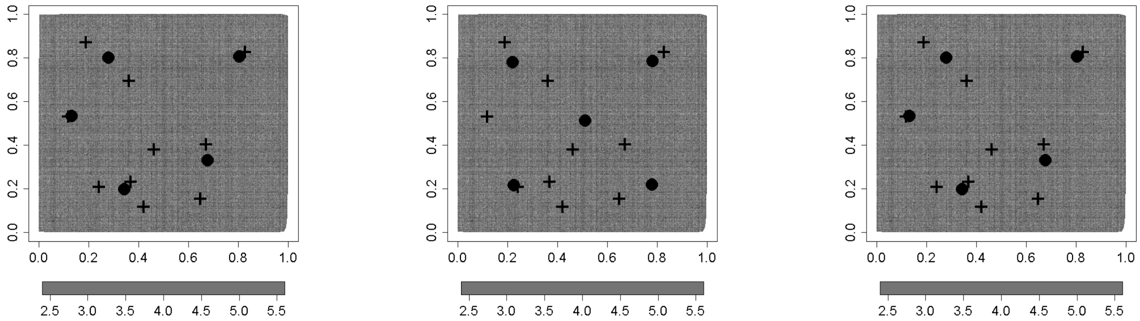
Vecchia AV. Estimation and model identification for continuous spatial processes. Journal of the Royal Statistical Society Series B (Methodological). 1988; 50(2):297–312.

Warrick A, Myers D. Optimization of sampling locations for variogram calculations. Wather Resources Research. 1987; 23(3):496–500.

Zhu Z, Stein M. Spatial sampling design for parameter estimation of the covariance function. Journal of Statistical Planning and Inference. 2005; 134(2):583–603.

Zimmerman D, Homer K. A network design criterion for estimating selected attributes of the semivariogram. Environmetric. 1991; 2(4):425–441.

Zimmerman DL. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. Environmetrics. 2006; 17(6):635–652.



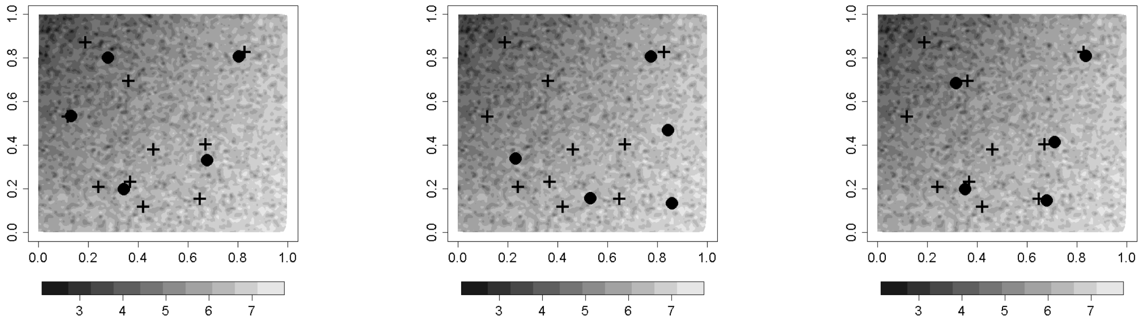
(a) Population Only.

(b) Arsenic Only.

(c) Population and Arsenic.

Figure 1.

True optimal dtw locations (circles) and bari locations (crosses), under the assumption of a constant arsenic surface (log scale) over the region, for the introduced objective function using $\gamma_1(\boldsymbol{s})$ (A), $\gamma_2(\boldsymbol{s})$ (B), and $\gamma(\boldsymbol{s})$ (C).



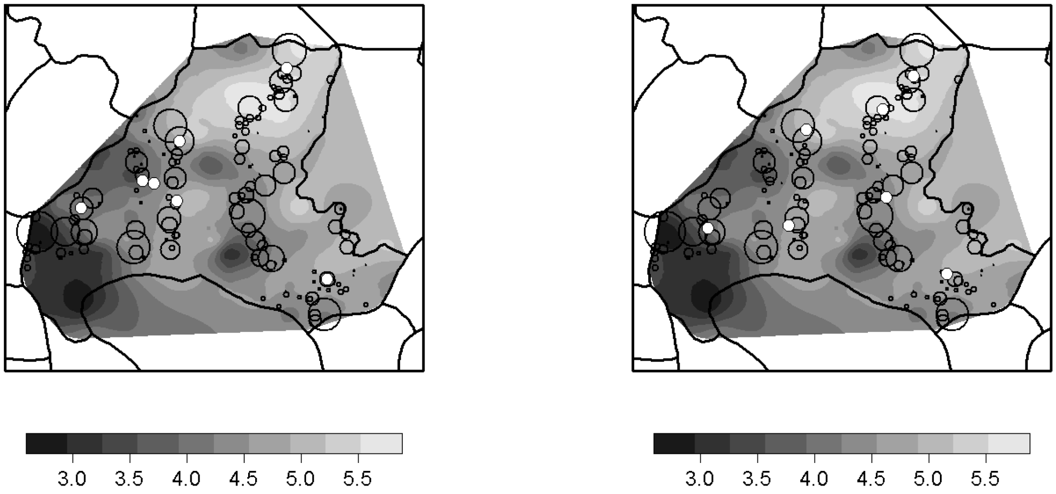
(a) Population Only.

(b) Arsenic Only.

(c) Population and Arsenic.

Figure 2.

True optimal dtw locations (circles) and bari locations (crosses), under a realistically simulated arsenic surface (log scale) over the region, for the introduced objective function using $\gamma_1(\boldsymbol{s})$ (A), $\gamma_2(\boldsymbol{s})$ (B), and $\gamma(\boldsymbol{s})$ (C).



(a) Actual DTW Placement.

(b) Estimated DTW Placement.

Figure 3.

Actual (a) and estimated (b) locations (white circles) and bari locations (clear circles) across the Matlab village used in the analysis. The larger bari circles indicate higher population of the bari. Predicted log scale arsenic surface is also displayed. Optimal dtw locations, estimated by k-means analysis of the posterior samples, are displayed in (b).

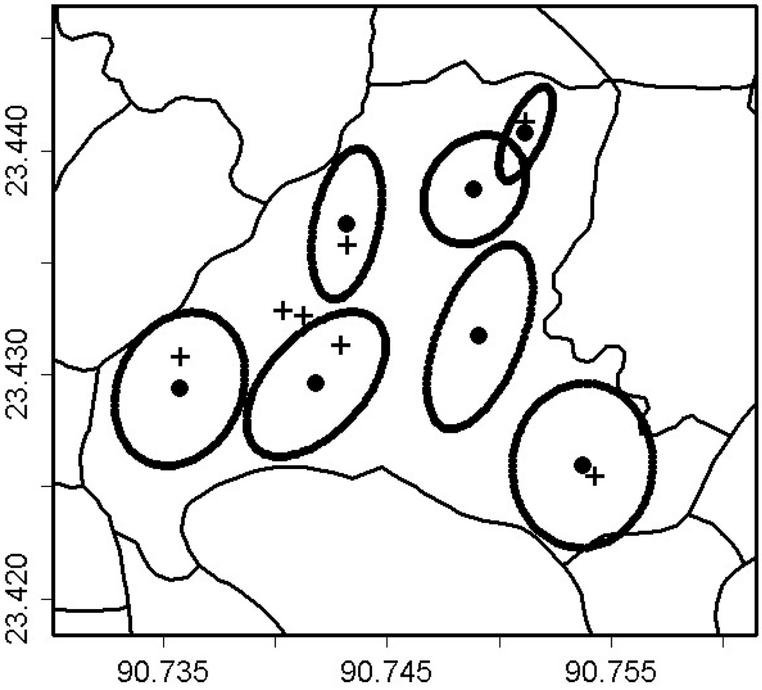


Figure 4.

Current (crosses) and estimated (circles) dtw locations. 95% elliptical credible regions are displayed for each estimated optimal dtw location.

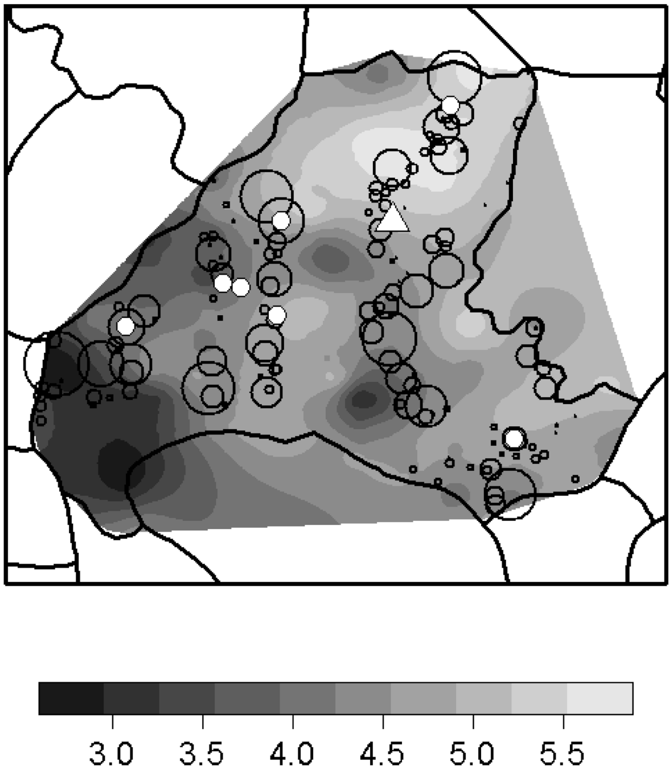


Figure 5.

Current dtw locations (white circles) and optimal location of the next to be installed dtw (triangle). Bari locations (clear circles) are displayed with larger circles indicating higher population, along with the predicted log scale arsenic surface.

Table 1

Average distance (km) from true optimal locations simulation study results. The standard error for each displayed estimate is 0.02. (M: Method, SS: Sample Size).

	M1	M2	M3
SS 1:	0.515	0.232	0.251
SS 2:	0.275	0.178	0.260
SS 3:	0.183	0.119	0.210

Table 2

Average area (m^2) of credible regions simulation study results. The standard error for each displayed estimate is 1.89. (M: Method, SS: Sample Size).

	M2	**M3**
SS 1:	80.38	24.82
SS 2:	51.56	23.38
SS 3:	29.29	24.44

Table 3

Included covariate results for the arsenic model in (1) while implementing the approximate likelihood method.

			Percentiles		
Parameter	**Mean**	**SD**	**0.025**	**0.50**	**0.975**
β_0 : Intercept	7.33	0.14	7.06	7.33	7.61
β_1 : Depth	–0.0284	0.0004	–0.0291	–0.0284	–0.0276
β_2 : (Depth)^2	0.0000255	0.0000005	0.0000245	0.0000255	0.0000265
σ_{ε}^{2} : Nugget	1.32	0.02	1.27	1.32	1.36
σ_{w}^{2} : Partial Sill	1.13	0.09	0.98	1.12	1.34
ϕ : Spatial Decay	2.81	0.33	2.23	2.80	3.46