Risk Based Arsenic Rational Sampling Design for Public and Environmental Health Management

Lihao Yin[a,b,*], Huiyan Sang[a,*], Douglas J. Schnoebelen[c,**], Brian Wels[c], Don Simmons[d], Alyssa Mattson[d], Michael Schueller[d], Michael Pentella[d] and Susie Y. Dai[e,***]

[a]Department of Statistics, Texas A&M University, College Station, Texas, 77843, USA
[b]Institute of Statistics and Big Data, Renmin University, Beijing, China
[c]University of Iowa, Iowa City, IA 52242
[d]State Hygienic Laboratory, University of Iowa, Coralville, Iowa, 52241, USA
[e]Department of Plant Pathology and Microbiology, Texas A&M University, College Station, TX, 77843, USA

ARTICLE INFO

Keywords:
private well
spatially clustered function model
resource management

ABSTRACT

Groundwater contaminated with arsenic has been recognized as a global threat, which negatively impacts human health. Populations that rely on private wells for their drinking water are vulnerable to the potential arsenic-related health risks such as cancer and birth defects. Arsenic exposure through drinking water is among one of the primary arsenic exposure routes that can be effectively managed by active testing and water treatment. From the public and environmental health management perspective, it is critical to allocate the limited resources to establish an effective arsenic sampling and testing plan for health risk mitigation. We present a spatially adaptive sampling design approach, based on a spatially varying estimation of the underlying contamination distribution. The method is different from traditional sampling design methods that often reply on a spatially constant or smoothly varying contamination distribution. In contrast, we propose a statistical regularization method to automatically detect spatial clusters of the underlying contamination risk from the currently available private well arsenic testing data in the USA, Iowa. This allows us to develop a sampling design method that is adaptive to the changes in the contamination risk across the identified regions. We provide the spatially adaptive sample size calculation and sampling locations determination for different acceptance precision and confidence levels for each cluster, to effectively mitigate the arsenic risk from the resource management perspectives. The model presents a framework that can be widely used for other environmental contaminant monitoring and sampling for public and environmental health.

1. Introduction

Arsenic (As) is ranked as the 20th most abundant element in the Earth's crust and has been studied internationally. Groundwater contaminated with arsenic has been recognized as a global treat, which negatively impacts human health [1, 2]. The primary human exposure to arsenic is through drinking water with additional contributors such as food and air [3–5]. Arsenic is a potent human carcinogen, which can cause blabber, lung, and skin cancers [6]. Furthermore, arsenic and its metabolites can cross the placental barrier and create risk for adverse maternal and fetal health leading to adverse birth outcomes [7]. The Safe Drinking Water Act (SDWA) by United States Environmental Protection Agency (USEPA) established 0.01 mg/L as the legal limit for arsenic in drinking water. In the USA, approximately 41.8 million (13% of the total US population) people obtain drinking water from private wells and the private wells are not regulated under the current EPA regulation [8]. The recent national Water-Quality Assessment Program from United States Geological Survey (USGS) reports that more than one out of five wells contain contaminants at concentrations exceeding the EPA maximum contaminated levels (MCLs) and/or USGS health-based screening levels. Among the various contaminants that exceed the EPA maximum contaminated levels, arsenic contamination is a common finding. Because private wells are not regulated in the US, in the Midwest region, a significant percentage of the population depends on private well for drinking water is at risk due to drinking water arsenic contamination. Arsenic testing in private well water represents a fundamental mean that helps mitigate the arsenic risk in the rural population for public

*Equally contributing authors.
**Current contact: U.S. Geological Survey, 5563 De Zavala Road, San Antonio, TX 78023
***Corresponding author. Department of Plant Pathology and Microbiology, College Station, TX, 77843, USA. E-mail address: sydai@tamu.edu (Susie Dai PhD).
ORCID(s):

and environmental health. In reality, many of the private wells are not tested, which presents a significant challenge for health risk mitigation. From the management perspective, a scientifically sound sampling plan to test a representative sample size is needed to evaluate the arsenic risk with limited resources.

Sampling theory can be used to guide a large number of chemical and biological analyses for environmental control and consumer safety [9]. As for arsenic testing, a systematic sampling plan is critical for risk assessment to draw science and data based conclusions and make the best usage of limited resources. The USEPA has published guidance for data quality objectives with regard to sampling design [10]. One of the key preparations for a sampling design is to determine the sample size and sampling error so that representative samples are collected.

Understanding sample statistical distribution parameters is critical when selecting a sampling method, sampling strategy and sample size. Application of probability distribution can help to develop a science-based sampling plan and estimate the chemical and biological hazards in the environment. Previously, binomial probability theory has been well studied for sample size determination for estimating a binomial proportion [11]. Application examples include the sampling plan in product inspection and surveillance [12], epidemiology [13], and medical diagnostics [14]. In many of these applications, a univariate binomial distribution is considered, that is, the underlying binomial proportion parameter is assumed to be a constant in the study. However, due to the spatially heterogeneity nature of arsenic distribution in the earth crust and groundwater, the traditional binomial sampling scheme based on a univariate binomial distribution may not be suitable for the survey of the target private well population. There is a great need to develop new sampling schemes that are capable of accounting for the spatially heterogeneity nature of arsenic distribution.

In terms of arsenic contamination, quite a few statistical and mathematical models have been used to estimate and predict arsenic concentration in groundwater and private wells. Logistic models for binomial distributions are widely adopted to estimate the spatial distribution of As contamination probability at both global and regional levels [15–21]. For instance, a logistic linear regression model has been used to predict high arsenic domestic well population in the US [22]. Furthermore, boosted regression trees models (weak -learner ensemble models) and traditional logistic linear models have been compared to estimate and predict arsenic distribution probabilities in drinking water wells in central valley, California [23]. Similarly to those statistical models, predictive variables are used to predict geogenic arsenic in drinking water wells in glacial aquifers, north-central USA [24]. Machine learning models have also been used to predict arsenic concentrations in ground water in Asia [25]. Nevertheless, all of the aforementioned models have been primarily focusing on the estimation and prediction of arsenic rather than the sampling design. Moreover, most of the methods often rely on a rich set of predictors and training data set to guarantee model accuracy. To the best of our knowledge, there is very limited work that combines the estimation of varying arsenic distribution with the binomial sampling design method.

To close this gap in the current literature for spatial binomial distribution sampling design, the current study proposes a spatially adaptive sampling design approach, by estimating a spatially clustered underlying contamination distribution. We apply this method to the determination of the data locations to understand arsenic contamination risk in private wells in Iowa. The method is different from traditional spatial sampling design methods [26, 27] that often assume continuous process spatial models for relatively smooth spatial fields. In contrast, we model the underlying contamination risk as a spatially clustered function for straightforward interpretation of the result . It also has the advantage of detecting discontinuous spatial heterogeneity in the arsenic distribution and then borrowing information within each identified spatially homogeneous cluster for sampling design. The method is built upon a graph fused lasso regularization method [28], which allows us to automatically detect clusters of spatial units and estimate the underlying spatially varying contamination distributions simultaneously. Thanks to the flexibility of graphs, the spatial clustering model enjoys several nice properties. First, it leads to very flexible cluster shapes naturally satisfying spatial contiguity constraints. Second, the method automatically learns the number of clusters from the data, relaxing the limitation in other clustering algorithms that require to specify the number of clusters a priori.

Another unique advantage of estimating a spatially clustered contamination distribution over other contamination distribution models lies in its easy integration with the traditional binomial sampling theory. Within each identified spatial cluster, the contamination distribution can be treated as having a common binomial proportion, for which we propose and compare two different sampling size determination methods at different levels of acceptance precision and confidence. Given the sample size calculations, a remaining sample design task is to determine the sampling locations. In our study, both the candidate wells and the available tested wells are distributed highly unevenly in the study region, which poses a challenge to design a sampling design with a balanced spatial coverage. To overcome this challenge, we propose a practical algorithm based on the spatial point process theory to distinguish areas that have been sufficiently-

sampled and insufficiently-sampled, and determine new sampling locations accordingly. This adaptive strategy will allow the practitioners to allocate sample collection efforts and resources more efficiently.

2. Materials and Methods

2.1. Sample Collection and Analysis

For the private well samples, the data used to build the model is collected as part of the Iowa Grants-to-Counties (GTC) program. The Iowa GTC program is established in 1987 after the Iowa legislature passed the Iowa Groundwater Protection Act to protect groundwater. Arsenic testing has been included as part of the GTC program based on Iowa Administrative Code [29] . A total of 14, 570 samples were collected and analyzed at University of Iowa State Hygienic Laboratory from July 1st, 2015 to June 16, 2020. As part of the GTC program, the local health department collects the private well samples by conducting a home visit, and sends to a laboratory for analysis. It should be noted that the selection of the laboratory is at the county's discretion. For all the samples analyzed at the State Hygienic Laboratory, the water sample is collected either at the tap faucet or outside of the house. Samples are collected in a 4 oz. HDPE plastic bottle containing 1 mL of 1 + 1 nitric acid as a preservative. Cooling is not required for sampling. Samples are screened for turbidity following Standard Methods 2130 B using a HACH model 2100N Turbidimeter. Samples exceeding 1 nephelometric turbidity units (NTU) are digested prior to analysis. The arsenic analysis is performed based on the Iowa State Hygienic Laboratory standard operating procedure (SOP), similar to the EPA 200.2 method. Briefly, a 50-mL aliquot is transferred from a well-mixed sample to a polypropylene digestion tube (Environmental Express #UC475-GN). One mL of 1+1 nitric acid and 0.5 mL of 1 + 1 hydrochloric acid (Fisher, Trace Metal Grade) are added to the tubes. Digestion is accomplished using a hot block (Environmental Express #SC154) at approximately 85°C. The sample volume is reduced to 10 mL, and then the sample is covered with a watch glass (Environmental Express #SC505), and refluxed for 30 minutes. The tubes are cooled and diluted to 25 mL with reagent water. The samples are further diluted to 50 mL using a mixture of 2% nitric acid and 1% hydrochloric acid. The samples are then analyzed for arsenic using an Agilent 7500 CE inductively coupled plasma mass spectrometer following EPA method 200.8. Approximately 5 mL of sample is transferred to a polypropylene autosampler tube for analysis. The instrument is calibrated using a multi-point calibration curve (0, 1, 5, 50, 100, 500 ug/L). Standards are matrix-matched to the sample. Thus, digested samples are not analyzed in the same run with direct analysis samples. Internal standards are introduced via a mixing tee at the instrument. Yttrium is used as the internal standard for arsenic. Results are not reported unless all quality controls pass their acceptance limits per the method.

2.2. Estimation of spatially clustered contamination risk

Let $y(\mathbf{s}_i)$ denote the binary variable at well location $\mathbf{s}_i$, for $i = 1, \dots, n$, coded as being 1 if the arsenic concentration is exceeding the EPA MCL (i.e., 0.01 mg/L), and 0 otherwise. Here, n is the total number of available tested wells. We propose a spatially varying binary logistic model for $y(\mathbf{s})$. Specifically, we assume

$$P\big(y(\mathbf{s}_i) = 1\big) \sim \text{Bernoulli}\big(p(\mathbf{s}_i)\big), \quad \text{for } i = 1, \dots, n \tag{1}$$

where $p(\mathbf{s}_i)$ is the probability of that the well located at $\mathbf{s}_i$ being contaminated. In the logistic regression model, we model the probability $p(\mathbf{s}_i)$ as

$$p(\mathbf{s}_i) = \frac{1}{1 + \exp\{-\beta(\mathbf{s}_i)\}}$$

or equivalently, $\log \frac{p(\mathbf{s}_i)}{1-p(\mathbf{s}_i)} = \beta(\mathbf{s}_i)$, where $\beta(\mathbf{s}_i)$ is interpreted as the log-odds of the arsenic contamination event that $y(\mathbf{s}_i) = 1$. Let $\boldsymbol{\beta} = \big(\beta(\mathbf{s}_1), \dots, \beta(\mathbf{s}_n)\big)$ be the stacked regression parameter for all the available tested well locations. It follows that the corresponding logistic regression log likelihood function takes the form:

$$\ell(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \log(1 + e^{\beta(\mathbf{s}_i)}) + \sum_{i=1}^{n} y(\mathbf{s}_i)\beta(\mathbf{s}_i) \tag{2}$$

We relax the assumption of having a constant contamination probability p, or equivalently, contamination log-odds, β, over the whole study region and instead assume it is varying over space. This assumption is reasonable for a large study region like Iowa due to the anticipated spatial heterogeneity in the arsenic concentration in groundwater

and private wells. Specifically, we assume $p(\mathbf{s})$ is a spatially clustered function, that is, there exists a number of geographical clusters such that $p(\mathbf{s})$ stay relatively homogeneous within each cluster but vary across clusters. This will facilitate the easy visualization and interpretation of the varying contamination risk in different identified sub-regions. We will show in Section 2.3 that this result leads to an efficient spatially adaptive sampling design strategy.

We consider a flexible regularization model for pursing the clustered pattern of $\beta(\mathbf{s})$ and $p(\mathbf{s})$. Regularization methods have gained a large popularity in modern high dimensional statistics and machine learning methods for various statistical learning tasks [30]. They have proved to be effective in imposing structural assumptions on model parameters such as sparsity, smoothness and clustering to avoid over-fitting problems. The regularization method for the Arsenic contamination model is performed in following steps:

1. Construct a spatial graph, denoted as $G = (V, E)$ where $V = v_1, v_2, ..., v_n$ is the vertex set with n vertices and E is the edge set. For a spatial problem, each vertex represents a spatial location. The choice of edge set is an important component of the method, which we will discuss later in this section.
2. Use this graph to construct a homogeneity pursuit regularization, also called the penalty function, for $\boldsymbol{\beta}$ as follows:

$$\lambda \sum_{(i,j)\in E} |\beta(\mathbf{s}_i) - \beta(\mathbf{s}_j))|. \tag{3}$$

3. Combine the penalty function in (3) with the logistic log-likelihood function in (2) to form a penalized objective function, which we minimize to obtain an estimator of $\boldsymbol{\beta}$ as follows:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = -\frac{1}{n}\ell(\boldsymbol{\beta}) + \sum_{(i,j)\in \mathbb{E}} |\beta(\mathbf{s}_i) - \beta(\mathbf{s}_j)|. \tag{4}$$

4. After obtaining $\hat{\boldsymbol{\beta}}$, calculate the estimates of the contamination probability from $\hat{p}(\mathbf{s}_i) = \dfrac{1}{1+\exp(-\hat{\beta}(\mathbf{s}_i))}$.

The regularization in step 2, referred to as the fused lasso penalty [28, 31], is to encourage homogeneity between the arsenic contamination probabilities at two locations if they are connected by an edge in E of the specified spatial graph. λ is a regularization tuning parameter determining the strength of fused lasso penalty and ultimately influencing the estimated number of clusters. The solution of l_1 penalty results in an exact fusion or separation between $\beta(\mathbf{s}_i) - \beta(\mathbf{s}_j)$, that is, the edges in the graph are classified into a set that corresponds to the non-zero elements of $\beta(\mathbf{s}_i) - \beta(\mathbf{s}_j)$, and another set that corresponds to the zero elements of $\beta(\mathbf{s}_i) - \beta(\mathbf{s}_j)$. As such, this regularization automatically leads to spatially clustered parameter estimates.

The choice of graph plays two important roles in the method; it not only reflects the prior information about the geological topology and clustering structure of the data, but also determines the computation complexity of the algorithm. Some natural graph choices for spatial data include the k nearest neighbor graphs, graphs connecting neighbors within a certain radius, and spatial Delaunay triangulation graphs (see, e.g., Li and Sang [32]). Alternatively, graphs can be constructed based on some preliminary estimates of parameters. For instance, the differences between the initial estimates of parameters can be used as the distance metric between vertices to replace the spatial Euclidean distance when constructing graphs. In this paper, we take a hybrid approach to construct the graph; the k nearest neighbor edge set connecting counties is determined based on the sample proportion within each county, and the k nearest neighbor edge set within each county is determined based on the Euclidean distance.

There are several advantages of using fused lasso penalty function for cluster detection. First, this penalization allows to detect clusters and estimate model parameters simultaneously. Second, this method guarantees to achieve a spatially contiguous clustering configuration such that only adjacent locations are clustered together. Another appealing property of this method is that the resulting clusters have very flexible shapes. We explain this point using the notion of connected components in graph theory; spatially contiguous clusters can be defined as the connected components of a graph G, and accordingly, a spatially contiguous partition of V can be defined as a collection of disjoint connect components such that the union of vertices is V. It is easy to show that any spatially contiguous partition with arbitrary cluster shapes can be recovered by removing a set of edges from a spatial graph [32]. In addition, the number of clusters does not need to be fixed a priori. Instead, we can determine it by a data-driven information criterion approach to be described later in this section. Finally, besides its capability to capture piece-wise constant coefficients, previous theoretical studies proved that this penalty has strong local adaptivity in that it is also capable of capturing piece-wise

Lipschitz continuous functions [33], which implies that the method can also approximate a spatially smoothly varying contamination probability reasonably well.

We now discuss how to solve the optimization in (4) to obtain the parameter estimation results. Note that $-\frac{1}{n}\ell(\boldsymbol{\beta})$ is convex and differentiable with respect to $\boldsymbol{\beta}$, and $\sum_{(i,j)\in\mathbb{E}} |\beta(\mathbf{s}_i) - \beta(\mathbf{s}_j)|$ is also convex. Therefore we propose an iterative algorithm combining the proximal gradient method [34] and the alternating direction method of multipliers (ADMM) [35] for this convex optimization problem. Specifically, given the current estimate $\boldsymbol{\beta}^{(t)}$, we let $\mathbf{g}^{(t)} = \boldsymbol{\beta}^{(t)} + (1/L)\frac{1}{n}\nabla\ell(\boldsymbol{\beta}^{(t)})$, where L is the Lipschitz constant of $-\frac{1}{n}\ell(\boldsymbol{\beta})$, and $\nabla\ell(\boldsymbol{\beta}^{(t)})$ is the first derivative of $\ell(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta}^{(t)}$). For the logistic regression model in (2), we can choose L to be $1/n$. Following the proximal gradient algorithm, we then update the value of $\boldsymbol{\beta}$ by solving:

$$\boldsymbol{\beta}^{t+1} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\left\|\boldsymbol{\beta} - \mathbf{g}^{(t)}\right\|_2^2 + \frac{\lambda}{L}\sum_{(i,j)\in\mathbb{E}} |\beta(\mathbf{s}_i) - \beta(\mathbf{s}_j)| \tag{5}$$

For the optimization in (5), we propose to solve by using the ADMM algorithm [36]. We will release the R code of our algorithm as a supplementary file upon acceptance of this manuscript for publication.

Finally, the parameter estimation algorithm involves the selection of the tuning parameter λ. In high dimensional statistics, data-dependent model selection criteria, such as generalized cross-validation [37], Bayesian information criterion (BIC) [38] and extended Bayesian information criterion [39] have been commonly used to determine the value of λ. Alternatively, one may choose a λ based on some prior knowledge on the number of clusters. For the numerical studies in this paper, we use BIC with the form, $\texttt{BIC} = -2\ell(\hat{\beta}) + k\log n$, to select the "optimal" λ that minimizes BIC from a candidate set.

2.3. Spatially adaptive sampling design

We now turn the attention to the sampling design problem for the determination of the sample size and locations of wells. Recall in Section 2.2 we have obtained a spatially clustered contamination probability $p(\mathbf{s})$, that is, within each identified spatial cluster, each sample is assumed to have the same probability of being contaminated from the population of wells in that cluster. This allows us to employ existing sampling design methods for the univariate binomial distribution with a constant p within each cluster, while adapting the value of p across clusters. The method leads to a simple but efficient preferential sampling strategy accounting for the spatial variations in $p(s)$.

Sample size determination and confidence interval construction methods for a constant-proportion binomial distribution have been well studied in the statistics literature. Popular methods include the Clopper-Pearson exact method, Wilson score method, Wald test, Bayesian Jeffreys method, and Agresti–Coull method, among others. For a review and comparison of different methods, see, for example, [40] and [11]. In this work, we consider two methods, the modified Jefferey and the Wilson score methods, following the recommendations by [11].

Consider a univariate binomial distribution where a random sample of size n is drawn from a large population, X is the number of 1's (e.g., the number of contaminated wells), and p is the probability of a randomly selected well is contaminated. We seek to find the sample size, n, such that, for a given p and acceptance precision level δ, the expected length of the confidence interval, $\mathrm{EL}_i(n,p) = E\left(\Delta_i(X)\right)$ is equal to 2δ, where $\Delta_i(X)$ is the length of confidence interval, and the expectation is taken over the binomial distribution of X. The modified Jefferey and the Wilson score methods are described below.

1. The Wilson score test confidence interval takes the form

$$\frac{2X + z_{1-\alpha/2}^2 \pm z_{1-\alpha/2}\sqrt{z_{1-\alpha/2}^2 + 4X(1-X/n)}}{2\left(n + z_{1-\alpha/2}^2\right)}$$

   This method is derived from Pearson's chi-square test, where the center of the interval is a weighted average of sample proportion and 1/2, such that it is more suitable than the commonly used Wald method for extreme probability or small sample sizes. The Wilson method also has the advantage of yielding an analytical formula for the sample size as follows

$$n_W = \frac{-z_{1-\alpha/2}^2[4\delta^2 - 2p(1-p)] + z_{1-\alpha/2}^2\sqrt{[4\delta^2 - 2p(1-p)]^2 - 4\delta^2\left(4\delta^2 - 1\right)}}{4\delta^2}$$

2. The modified Jeffreys method is derived from a Bayesian approach, which uses the non-informative Jeffrey's prior $\text{Beta}(1/2, 1/2)$ to derive the posterior credible interval for p, while modifying the formula at the boundary values. For $1 < X < n$, the credible interval is

$$\left[\text{Beta}_{\alpha/2}(X + 1/2, n - X + 1/2), \text{Beta}_{1-\alpha/2}(X + 1/2, n - X + 1/2)\right]$$

The expressions when X takes boundary values are provided in Table 1 of [11]. The modified Jeffreys method enjoys similar coverage properties as those of the Wilson score method. But it has an additional advantage of yielding a credible interval that is equal-tailed. For modified Jeffreys, sample size can be calculated by numerically solving $E\left(\Delta_i(X)\right) = 2\delta$. In practice, the sample size, is calculated by an approximated solution such that $|E\left(\Delta_i(X)\right) - 2\delta|$ is less than a certain tolerance.

Spatial sampling design not only involves the determination of sample size, but also the locations of sampling points. One simple and commonly used spatial sampling design is the uniform random sampling, where each location is chosen independently and uniformly within each cluster. However, two complications arise when applying this method for the Arsenic study. First, the number of all available candidate wells are not uniformly distributed in space. Second, a large number of wells have been tested where the sampling locations were arbitrarily chosen before the formal statistical sampling design, which results a highly unbalanced sampling in space with some areas over sampled and the other areas insufficiently sampled. The design for the new sample well locations need to exclude those previously tested wells. Our goal is to sample the candidate wells with the expectation that the combined new sample wells and the previously tested wells are of a spatially uniform distribution in each cluster except for the over-sampled areas. To achieve this goal, we utilize the connection between the uniform distribution in space and the spatial Poisson point process model, and adopt the thinning sampling idea from the latter. As a preliminary, we introduce the intensity function of the spatial point processes [41], which characterizes the probability that a point occurs in an infinitesimal ball around a given location. If there is a point process $\mathbf{X}$ on $D \subset \mathbb{R}^2$, then the intensity function $\lambda(u)$ at location $u \in D$ is defined as,

$$\lambda(u) = \lim_{|b(u)|\to 0} \frac{N(b(u))}{|b(u) \cap D|}$$

where $b(u)$ denote a small ball containing the u, measure $|\cdot|$ denotes the area and $N(B)$ indicates the expected number of points within a subset $B \subset \mathbb{R}^2$. If $\lambda(u) = \lambda$ is a constant for $u \in B$, then $\mathbf{X}$ is called a homogeneous point process on B, implying the point has the same probability to occur at each location in B. Besides, the intensity function determines the expected number of points on B by $E\{N(B)\} = \int_B \lambda(u)du$. It is known that, conditional on the number of points, the locations from a homogeneous Poisson point process are uniformly distributed on B. Therefore, we expect the sampled wells have the intensity function $\hat{\lambda}(u) = n_i/a_i$ for u located in cluster i, to render the sampled wells evenly-distributed. Here n_i and a_i denote the number of samples and the area in cluster i respectively.

The detailed sampling algorithm is described below. First, we use the nonparametric intensity estimation approach via `R` function `density.ppp` in package `spatstat` to estimate the candidate well intensity function, denoted as $\hat{\lambda}^{candi}(u)$, and the previously tested well intensity, denoted $\hat{\lambda}^{exist}$. To exclude the previously tested wells in Iowa from new samples, we calculate the target intensity from $\hat{\lambda}^{targ} = \max(\hat{\lambda} - \hat{\lambda}^{exist}, 0)$. Locations that have negative $\hat{\lambda} - \hat{\lambda}^{exist}$ values correspond to the over-sampled areas where the intensity of peviously tested wells exceeds the required sampling density. We will leave them out when drawing new samples. Finally, for other areas, each candidate well will be selected with the probability $\hat{\lambda}^{targ}(u)/\hat{\lambda}^{candi}(u)$, where u is the location of the candidate well. The last step is based on the assumption that $\hat{\lambda}^{candi}$ is large enough to bound $\hat{\lambda}^{targ}$, and indeed there are adequate wells available in Iowa to meet this assumption. As a result, the algorithm guarantees that the combined new samples and existing samples other than the over-sampled areas will be (nearly) uniformly distributed, and the expected sample size meets the requirement in Table 1.

3. Results

3.1. Descriptive Statistical Analysis Results

The raw data amount to 14,570 previously collected observations of Arsenic tests in total (Figure 1). Based on the risk categories, we characterize the wells that contain higher than 0.01 mg/L arsenic as high risk wells. We exclude the observations whose location information is absent. We also aggregate the repeated measurements at the same locations into one single observation, by setting Y to be 1 if there is at least one concentration measurement exceeding MCL.

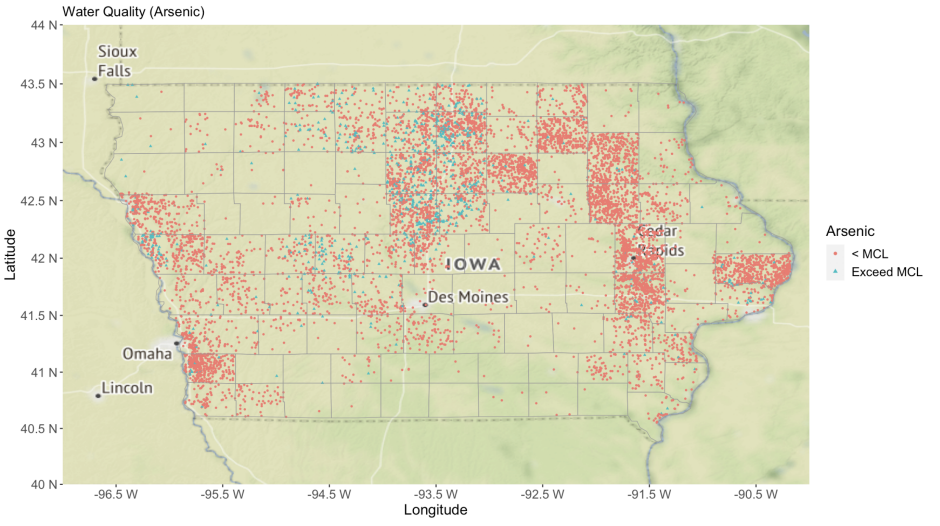

Figure 1: Spatial distribution of the Arsenic contamination presence/absence observations.

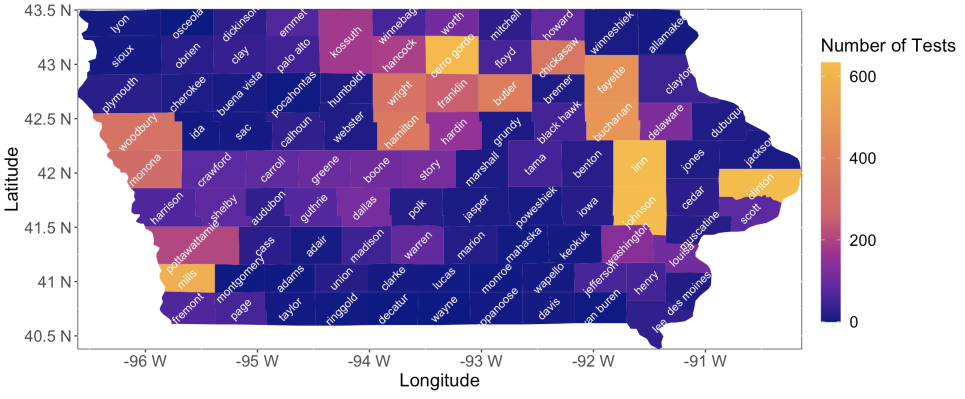

Figure 2: The number of tested wells in each county in Iowa;

After these pre-processing steps, there remain 9842 observations at different locations. Figure 1 shows the spatial distribution of the observations, and Figure 2 shows the spatial map of the number of observations in each county. From the existing tested data, the most tested regions include northern central Iowa, a few counties in the western central, southwestern and eastern central Iowa regions (Figure 2). Less than 20% of the counties have more than 100 tests per county (Figure 3). There are less tests per county in the southern, northeastern and northwestern regions. We show in Figure 4 the sample proportion $\hat{p}$ at each county by pooling the observations together at the county level, as a mean to visualize a rough estimate of the arsenic risk. Even though we see an uneven testing distribution, which means uneven sampling at the current testing scale, we observe that the arsenic risk characterization appears to be independent on the testing density (Figure 2 and 4).

3.2. Risk clusters and regional management

Ayotte et al. [22] uses a predictive logistic regression model to estimate arsenic presence in regions with limited arsenic data. In that study, a total of 20450 domestic well samples are used to develop the model to estimate for the whole conterminous US. Unique to our study, we do not aim to establish a predictive model to accurately predict the arsenic contamination in a given region, as the risk of As has been already recognized by the state and many local health risk management agencies. We aim to utilize the locally clustered arsenic risks to estimate a sampling size

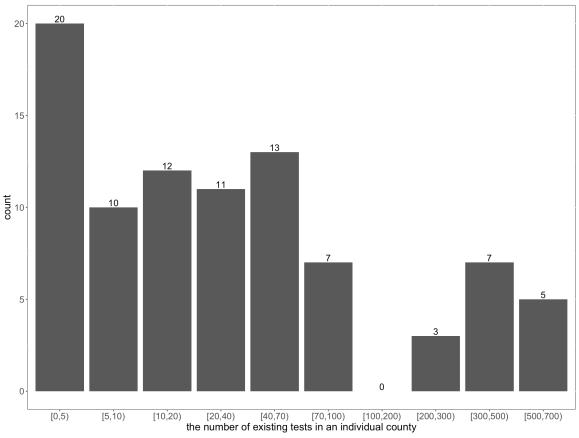

Figure 3: The frequency histogram of the number of observations in each county;

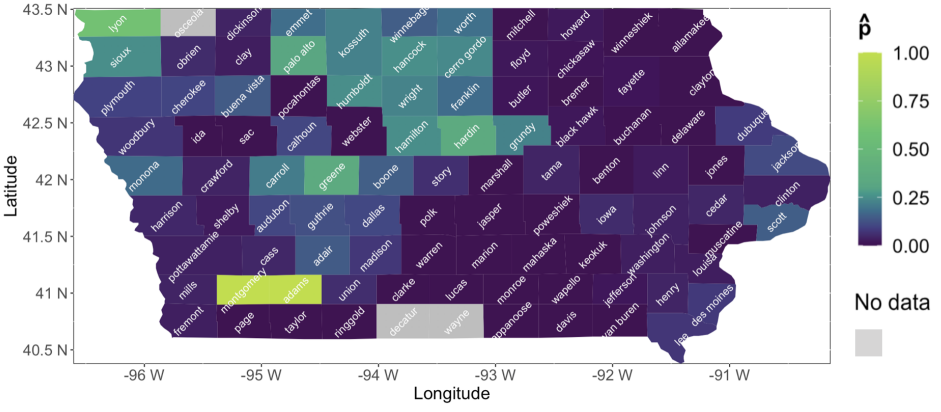

Figure 4: This pattern illustrates the $\hat{p}$ obtained directly by $\sum_{j=1}^{n_i} y_{i,j}/n_i$ on the county level; Here n_i denotes the number of observations in county i and $y_{i,j}$ is the jth individual observation in county i; Grey color means there is no observed data in the county.

with minimum bias, which can be managed with appropriately allocated resources. In order to do that, we define the binary existence of arsenic in a given private well is higher than 0.01 mg/L, which is the current EPA regulation level. In other words, we regard if the private well contains less than 0.01 mg/L arsenic, then the health risks are absent in a risk based sampling scheme. We first model and estimate the underlying contamination risk as a spatially clustered function following the method described in Section 2.2 for straightforward interpretation of the result and easy implementation of the sampling design. The optimization result partitions the whole state into three risk clusters based on the estimated arsenic presence probability (Figure 5). The three risk probabilities (p) are 0.03, 0.21, and 0.33 for cluster 1, 2, and 3, respectively. The risk cluster assignment is consistent with some previous observation and predictions. For example, cluster 1 is largely consistent with the estimations in [22]. Cluster 2 is also highlighted with potential high As contamination in the same study. Furthermore, a targeted As study performed in Cerro Gordo County (Northern Central Iowa) has sampled 68 wells over three years [42]. The study reveals one potential mechanism of As mobilization in shallow aquifer. The naturally occurring sulfide minerals (typically pyrite) in the bedrock aquifers could be the source of As. Under oxidizing condition, the As mobilization could happen from rocks to water. Significantly, the Cerro Gordo study has resulted in a policy change for arsenic testing and well completion locally. Interestingly, cluster 3 at the border of Iowa and Nebraska is identified as a new As "hotspot" in this current study.

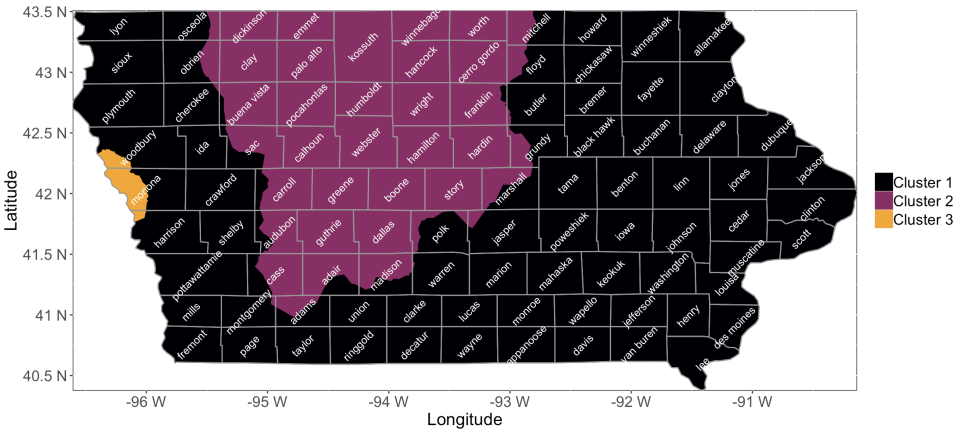

Figure 5: Partition of the map in terms of estimated p; In cluster 1, $\hat{p} = 0.02869485$; In cluster 2, $\hat{p} = 0.2088291$; In cluster 3, $\hat{p} = 0.3373494$; The number of observations in each cluster are 6482, 3194 and 166 respectively.

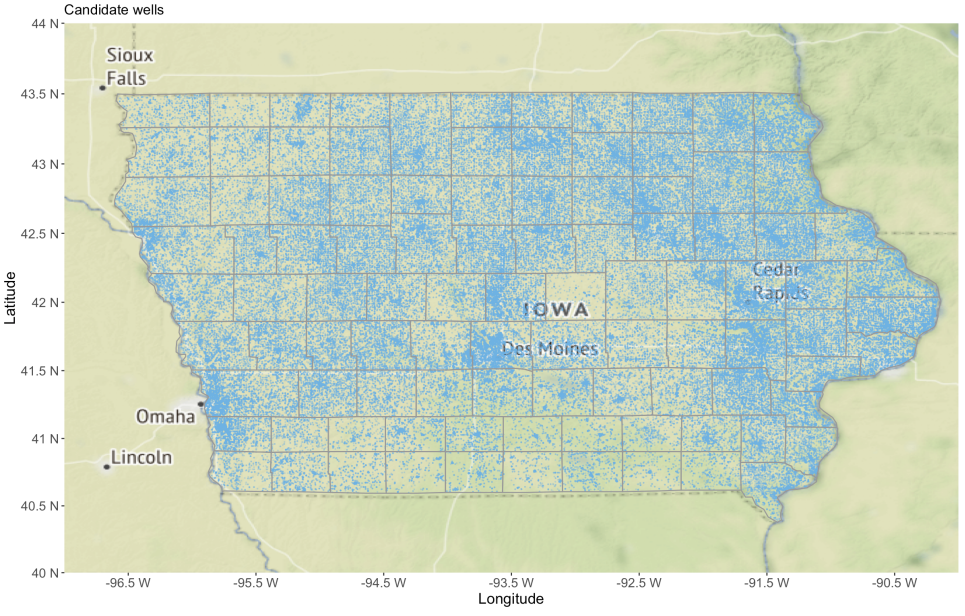

Figure 6: Locations of the 291, 882 candidate wells in Iowa, after discarding the wells in absence of their location information.

3.3. Sample design

Based on the estimated probability clusters, we further estimate the ideal sample size based on various acceptance precision and confidence level. Based on the public available database (Iowa Private Well Tracking System), it is estimated there are more than 300,000 private wells in Iowa. Among them, 291, 882 wells are geo-coded. The locations of the total geo-coded well population are shown in Figure 6, clearly indicating an uneven spatial distribution in Iowa. Based on the regional cluster risk probability, we thus define three different regions (region 1, 2, and 3) with different risk cluster ranks. For regions with reasonable testing coverage, we have three probabilities. For region 1, the estimated probability for As concentration higher than 0.01 mg/L probability is 0.03. For region 2 and 3, the probability is 0.21 and 0.34, respectively. If we define the precision acceptance to be 10% of the probability, the precision acceptance is 0.003 for region 1 (e.g. 10% of 0.03), 0.021 for region 2, and 0.034 for region 3 respectively. Table 1 provides the calculated required sample size for each cluster under three different confidence levels (90%, 95% and 99%) using both the Wilson and Jeffrey methods. For example, at 95% confidence interval, the estimated sample size based on the

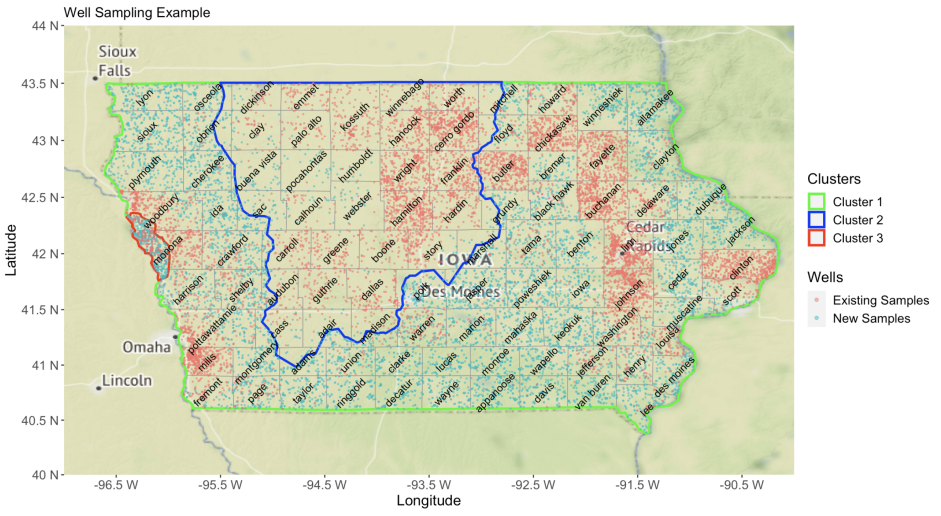

Figure 7: An example of sampling results; 8174, 313 and 586 additional wells are sampled in each cluster respectively in this example.

Jeffrey method is 12446 for region 1. Applying the same criteria to region 2 and 3, the estimated sample size would be 1442 for region 2 and 743 for region 3. The sample sizes calculated by the Wilson method only slightly differ from those of the Jeffrey method. Accordingly, at 99% confidence interval, we estimate that 21456, 2492 and 1282 samples are need for region 1, 2, and 3 respectively.

Table 1
Expected number of well sampling in each cluster;

Confidence Level	Method	Cluster 1	Cluster 2	Cluster 3
90%	Wilson	8766	1017	523
	Jeffrey	8746	1015	523
95%	Wilson	12446	1444	743
	Jeffrey	12420	1442	743
99%	Wilson	21497	2493	1282
	Jeffrey	21456	2492	1284

In the existing As data set, there are 6482, 3194, and 166 testing results already collected from cluster 1, 2 , and 3 respectively. It is noted that existing sample size within each cluster constitutes a large proportion or even exceeds the required sample size calculated in Table 1. However, we recognize the current As data collection is operated at the county level since the local environmental health jurisdiction resides in each county. This results in an uneven spatial distribution of sampling locations for the whole state as discussion in Section 3.2. Therefore, there exist areas that are over-sampled, but in the meantime, new samples still need to be collected at those places that are only sparsely sampled previously.

We follow the method presented in Section 2.3 to determine the locations of new sampling locations. We use the private wells in the current Iowa PWTS database as the target sampling population (Figure 6). The goal is to achieve a spatially balanced sampling design that meets the required sample size, while accounting for the facts that both the candidate wells and existing tested wells are distributed highly non-uniformly in space. To illustrate, we give an example of the sampling scenario using the sample size calculated from the Wilson method for the 95% CI in Figure 7. The previously over-sampled areas are revealed by the dense red point clouds in this Figure. Region 2 has the largest proportion of previously over-sampled areas. Only a relatively small number of additional wells (marked by green dots) need to be sampled, most of which are located in the middle west of this region. In contrast, most areas in Region 1 have not been sampled and tested previously, with exceptions in several counties (e.g., Buchanan, Butler,

and Clinton). In region 3, although the spatial coverage of the existing tested samples are nearly uniform, our method suggests that an additional number of wells need to be collected to achieve the desired confidence level and prevision accuracy. Overall, it is noted that the locations of samples in Figure 7 appear to be uniformly distributed except for the previously over-sampled areas. Looking more closely, we observe that the intensity/density of samples differs across the identified risk regions, as a consequence of adopting a spatially adaptive sampling design according to each region's own contamination risk.

4. Discussions

It is commonly recognized that there are many conditions such as geological, geochemical and hydrologic variables that impact arsenic presence in ground water. For example, It has been observed high arsenic concentrations are often found in more arid western US [22]. Furthermore, precipitation and recharge show significant correlations with arsenic concentrations in domestic wells in the conterminous US. Among various conditions, glaciated terrain, bedrock geology, soil hydrology, soil tile drainage, water table depth and climate factors can also impact arsenic concentrations in groundwater. Particularly, Iowa's groundwater resources are majorly surficial aquifers and bedrock aquifers. For a long history contacting with glaciers, many parts of Iowa soil/dirt contain glacier age materials that have moderate to low permeability. The water table beneath those materials occurs at rather shallow depths and varies from 3 to 30 feet below grounds [43].The micro-environment such as pH, soil and water bacterial activity, oxidation and reduction reactions (Redox), coexistence with other elements (e.g. iron) can also play a significant role in arsenic concentration in groundwater. Taking account of all those macro and micro-environmental conditions is a shared challenge for all current available predictive models to estimate arsenic concentrations at the county, state/province or region levels.

There are several benefits to adopt the proposed sampling design. First, the sample size estimate suggests future feasible random sampling targets, given the total Iowa private well population. As the sample sizes are dependent on the arsenic probabilities, we present options for the same probability with different sampling precision goals. We also recognize there are regions with too few or no data points (Figure 4, thus warrant further sampling for probability estimate). Second, the method developed in this study helps pinpoint future sampling locations with adequate statistical power. From the resource management perspective, future planning can prioritize the high risk well sampling, eliminate redundant testing, and collect representative samples for risk assessment purposes. In practice, sample collections and management are often conducted at certain administration levels. It is desired to develop a sampling design method that is easy and fast to implement at each administration unit. Third, this design presents future opportunities to investigate practical solutions on how to coordinate joint efforts across counties for the efficient implementation of the sampling design method.

Moving forward, this work could be further refined in several ways. First, the estimator we obtained by optimizing the regularized log-likelihood function does not come with an uncertainty measure. As such, the sample size calculation is only based on a point estimate of the contamination risk. A potential solution is to consider a Bayesian version of the method. In principle, the modified Jeffrey's method for sample size calculation can be adapted to account for the uncertainty in the estimate of p, where the expectation, $E\left(\Delta_i(X)\right)$, can be taken with respect to both p and X instead of X only. Second, the model can be further improved with more representative samples. As we noted, there are counties without testing data, which presents a gap for risk analysis. We expect collecting data in those regions helps build a more comprehensive evaluation of arsenic health risk at the state level. Overall, the current study presents a targeted approach to save cost and time for effective public health management strategy. The rational sampling design focuses on risk categories, which assures preventive measures and mitigation practices are implemented where most needed.

References

[1] J. Podgorski, M. Berg, Global threat of arsenic in groundwater, Science 368 (2020) 845–850.

[2] L. A. DeSimone, P. A. Hamilton, Quality of water from domestic wells in principal aquifers of the United States, 1991-2004, US Department of the Interior, US Geological Survey, 2009.

[3] K. S. Almberg, M. E. Turyk, R. M. Jones, K. Rankin, S. Freels, J. M. Graber, L. T. Stayner, Arsenic in drinking water and adverse birth outcomes in Ohio, Environmental research 157 (2017) 52–59.

[4] M. Vahter, Effects of arsenic on maternal and fetal health, Annual review of nutrition 29 (2009) 381–399.

[5] N. Sohel, L. Å. Persson, M. Rahman, P. K. Streatfield, M. Yunus, E.-C. Ekström, M. Vahter, Arsenic in drinking water and adult mortality: a population-based cohort study in rural Bangladesh, Epidemiology (2009) 824–830.

[6] M. Argos, H. Ahsan, J. H. Graziano, Arsenic and human health: epidemiologic progress and public health implications, Reviews on environmental health 27 (2012) 191–195.

[7] M. S. Bloom, S. Surdu, I. A. Neamtiu, E. S. Gurzau, Maternal arsenic exposure and birth outcomes: a comprehensive review of the epidemiologic literature focused on drinking water, International journal of hygiene and environmental health 217 (2014) 709–719.

[8] N. G. Association, Groundwater use in the United States of America, 2020. URL: https://www.ngwa.org/docs/default-source/default-document-library/groundwater/usa-groundwater-use-fact-sheet.pdf?sfvrsn=5c7a0db8_4.

[9] P. Minkkinen, Practical applications of sampling theory, Chemometrics and intelligent laboratory systems 74 (2004) 85–94.

[10] U. E. P. A. (USEPA), Guidance on choosing a sampling design for environmental data collection, 2002.

[11] L. Gonçalves, M. R. de Oliveira, C. Pascoal, A. Pires, Sample size for estimating a binomial proportion: comparison of different methods, Journal of Applied Statistics 39 (2012) 2453–2473.

[12] K.-M. Lee, T. J. Herrman, S. Y. Dai, Application and validation of a statistically derived risk-based sampling plan to improve efficiency of inspection and enforcement, Food Control 64 (2016) 135–141.

[13] N. Sepúlveda, C. Drakeley, Sample size determination for estimating antibody seroconversion rate under stable malaria transmission intensity, Malaria journal 14 (2015) 141.

[14] L. Joseph, C. Reinhold, Statistical inference for continuous variables, American Journal of Roentgenology 184 (2005) 1047–1056.

[15] M. Amini, K. C. Abbaspour, M. Berg, L. Winkel, S. J. Hug, E. Hoehn, H. Yang, C. A. Johnson, Statistical modeling of global geogenic arsenic contamination in groundwater, Environmental science & technology 42 (2008) 3669–3675.

[16] J. D. Ayotte, B. T. Nolan, J. R. Nuckols, K. P. Cantor, G. R. Robinson, D. Baris, L. Hayes, M. Karagas, W. Bress, D. T. Silverman, et al., Modeling the probability of arsenic in groundwater in New England as a tool for exposure assessment, Environmental science & technology 40 (2006) 3578–3585.

[17] L. Winkel, M. Berg, M. Amini, S. J. Hug, C. A. Johnson, Predicting groundwater arsenic contamination in Southeast Asia from surface parameters, Nature Geoscience 1 (2008) 536–542.

[18] J. E. Podgorski, S. A. M. A. S. Eqani, T. Khanam, R. Ullah, H. Shen, M. Berg, Extensive arsenic contamination in high-pH unconfined aquifers in the Indus Valley, Science advances 3 (2017) e1700935.

[19] L. H. Winkel, P. T. K. Trang, V. M. Lan, C. Stengel, M. Amini, N. T. Ha, P. H. Viet, M. Berg, Arsenic pollution of groundwater in vietnam exacerbated by deep aquifer exploitation for more than a century, Proceedings of the National Academy of Sciences 108 (2011) 1246–1251.

[20] L. Rodríguez-Lado, G. Sun, M. Berg, Q. Zhang, H. Xue, Q. Zheng, C. A. Johnson, Groundwater arsenic contamination throughout China, Science 341 (2013) 866–868.

[21] Q. Yang, H. B. Jung, R. G. Marvinney, C. W. Culbertson, Y. Zheng, Can arsenic occurrence rates in bedrock aquifers be predicted?, Environmental science & technology 46 (2012) 2080–2087.

[22] J. D. Ayotte, L. Medalie, S. L. Qi, L. C. Backer, B. T. Nolan, Estimating the high-arsenic domestic-well population in the conterminous United States, Environmental science & technology 51 (2017) 12443–12454.

[23] J. D. Ayotte, B. T. Nolan, J. A. Gronberg, Predicting arsenic in drinking water wells of the Central Valley, California, Environmental Science & Technology 50 (2016) 7555–7563.

[24] M. L. Erickson, S. M. Elliott, C. Christenson, A. L. Krall, Predicting geogenic Arsenic in Drinking Water Wells in Glacial Aquifers, North-Central USA: Accounting for Depth-Dependent Features, Water Resources Research 54 (2018) 10–172.

[25] Z. Tan, Q. Yang, Y. Zheng, Machine Learning Models of Groundwater Arsenic Spatial Distribution in Bangladesh: Influence of Holocene Sediment Depositional History, Environmental Science & Technology 54 (2020) 9454–9463.

[26] Z. Zhu, M. L. Stein, Spatial sampling design for prediction with estimated parameters, Journal of agricultural, biological, and environmental statistics 11 (2006) 24.

[27] P. J. Diggle, R. Menezes, T.-l. Su, Geostatistical inference under preferential sampling, Journal of the Royal Statistical Society: Series C (Applied Statistics) 59 (2010) 191–232.

[28] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2005) 91–108.

[29] I. D. of Public Health, Iowa Administrative Code 641, Chapter 24, Private Well Testing, Reconstruction, and Plugging- Grants to Counties , 2016.

[30] P. Bühlmann, S. Van De Geer, Statistics for high-dimensional data: methods, theory and applications, Springer Science & Business Media, 2011.

[31] R. J. Tibshirani, J. Taylor, et al., The solution path of the generalized lasso, The Annals of Statistics 39 (2011) 1335–1371.

[32] F. Li, H. Sang, Spatial homogeneity pursuit of regression coefficients for large datasets, Journal of the American Statistical Association 114 (2019) 1050–1062.

[33] O. H. Madrid Padilla, J. Sharpnack, Y. Chen, D. M. Witten, Adaptive nonparametric regression with the k-nearest neighbour fused lasso, Biometrika 107 (2020) 293–310.

[34] A. Beck, M. Teboulle, Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems, IEEE transactions on image processing 18 (2009) 2419–2434.

[35] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al., Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends® in Machine learning 3 (2011) 1–122.

[36] B. Wahlberg, S. Boyd, M. Annergren, Y. Wang, An ADMM algorithm for a class of total variation regularized estimation problems, IFAC Proceedings Volumes 45 (2012) 83–88.

[37] G. H. Golub, M. Heath, G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, Technometrics 21 (1979) 215–223.

[38] G. Schwarz, Estimating the dimension of a model, The annals of statistics 6 (1978) 461–464.

[39] J. Chen, Z. Chen, Extended Bayesian information criteria for model selection with large model spaces, Biometrika 95 (2008) 759–771.

[40] R. G. Newcombe, Two-sided confidence intervals for the single proportion: comparison of seven methods, Statistics in medicine 17 (1998) 857–872.

[41] P. Diggle, A kernel method for smoothing point process data, Journal of the Royal Statistical Society: Series C (Applied Statistics) 34 (1985) 138–147.

[42] D. J. Schnoebelen, S. Walsh, B. Hanft, O. E. Hernandez-Murcia, C. Fields, Elevated Arsenic in Private Wells of Cerro Gordo County, Iowa: Causes and Policy Changes., Journal of Environmental Health 79 (2017).

[43] J. C. Prior, J. L. Boekhoff, M. R. Howes, R. D. Libra, P. E. VanDorpe, Iowa's Groundwater Basics: A geological guide to the occurence, use, and vulnerability of Iowa's aquifers (2003).