PLOS ONE

RESEARCH ARTICLE

Spatial Prediction and Optimized Sampling Design for Sodium Concentration in Groundwater

Erum Zahid[1☯], Ijaz Hussain[1☯*], Gunter Spöck[2☯], Muhammad Faisal[3,4☯], Javid Shabbir[1‡], Nasser M. AbdEl-Salam[6‡], Tajammal Hussain[5‡]

1 Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan, **2** Department of Statistics, University of Klagenfurt, Klagenfurt, Austria, **3** Faculty of Health Studies, University of Bradford, BD7 1DP Bradford, United Kingdom, **4** Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, United Kingdom, **5** Department of Statistics, COMSATS Institute of Information Technology, Lahore, Pakistan, **6** Arriyadh Community College, King Saud University, Arriyadh 11437, Saudi Arabia

☯ These authors contributed equally to this work.
‡ These authors also contributed equally to this work.
* ijaz@qau.edu.pk

Citation: Zahid E, Hussain I, Spöck G, Faisal M, Shabbir J, M. AbdEl-Salam N, et al. (2016) Spatial Prediction and Optimized Sampling Design for Sodium Concentration in Groundwater. PLoS ONE 11(9): e0161810. doi:10.1371/journal.pone.0161810

Editor: David O. Carpenter, Institute for Health & the Environment, UNITED STATES

Received: August 28, 2015

Accepted: August 12, 2016

Published: September 28, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors would like to extend their sincere appreciation to the Deanship of Scientific Research at King Saud University for funding this work through Research Group no. RGP-210.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Sodium is an integral part of water, and its excessive amount in drinking water causes high blood pressure and hypertension. In the present paper, spatial distribution of sodium concentration in drinking water is modeled and optimized sampling designs for selecting sampling locations is calculated for three divisions in Punjab, Pakistan. Universal kriging and Bayesian universal kriging are used to predict the sodium concentrations. Spatial simulated annealing is used to generate optimized sampling designs. Different estimation methods (i.e., maximum likelihood, restricted maximum likelihood, ordinary least squares, and weighted least squares) are used to estimate the parameters of the variogram model (i.e, exponential, Gaussian, spherical and cubic). It is concluded that Bayesian universal kriging fits better than universal kriging. It is also observed that the universal kriging predictor provides minimum mean universal kriging variance for both adding and deleting locations during sampling design.

Introduction

Sodium is a mineral which is always present in drinking water through natural occurrences. The human body needs sodium to maintain the blood pressure and to control fluid levels. If it exceeds a threshold value (i.e., 200 mg/l), then it may change the taste of water. Moreover, it also creates severe medical problems for those who have high blood pressure. In most of the countries the level of sodium in water is less than 20 mg/l, however, in some countries it exceeds 250 mg/l.

About 1.1 billion people in the whole world are unable to access safe drinking water, and it causes a lot of deaths. Gadgil [1] provides some guidelines and highlights the presence of

required parameters in drinking water. Ferreccio et al. [2] evaluates the concentration of arsenic in drinking water which was about 860 mg/l during 1958 to 1970 in northern cities of Chile, however, later on it has been reduced to 40mg/l. They cross-examine two types of patients, smokers and non-smokers, and conclude that consumption of arsenic through water is the major cause of lung cancer. Gundogdu and Guney [3] use various kriging techniques to study the water levels by using spherical, tetraspherical, pentaspherical, exponential, Gaussian, rational quadratic, hole effect, K-bessel, J-bessel and stable variogram models. They conclude that universal kriging is better than ordinary kriging for interpolating water levels. Neuman et al. [4] described and implemented Bayesian model averaging, and maximum likelihood version of Bayesian model averaging which does not require any prior knowledge about parameters. It updates posterior probabilities as well as model parameters on the basis of new data, and is consistent with modern statistical methods of hydrologic model calibration. Mehrjardi et al. [5] show that co-kriging and kriging methods are better than inverse distance weighting techniques for predicting the spatial distribution of some characteristics of groundwater quality. Nas [6] concludes that ordinary kriging provides accurate patterns of groundwater quality parameters in Konya, Turkey. Sarukkalige [7] also uses kriging techniques for the analysis of the quality of groundwater in Western Australia. Andrade and Stigter [8] model the spatio-temporal variation of arsenic concentration in groundwater by using geostatistical and multivariate methods. They show that the concentration of arsenic has strong correlation with rice culture and that indicator kriging provides appropriate maps of arsenic concentrations.

Dhar and Datta [9] developed methodology based on inverse distance weighting method to reduce redundancy in monitoring network. Siri et al. [10] establish a sampling scheme for generating samples simultaneously. They prove that GPS units and a pseudo-sampling frame are more effective than old sampling methods. Brus and Heuvelink [11] validate that universal kriging performs better than ordinary kriging in terms of smaller mean universal kriging variance (MUKV) for spatial sampling design. Optimal spatial sampling design is a core issue in environmental studies when exploring low cost and greater efficiency samples. The sampling scheme can be optimized through prior knowledge and sampling constraints. Van Groenigen and Stein [12] conclude that Spatial Simulated Annealing (SSA) is better than classical methods for optimizing sampling designs. SSA is useful for those studies that have several sampling constraints. Zhu and Stein [13] study spatial sampling design for the estimation of covariance parameters by the maximum likelihood method.

In present paper, spatial behavior of sodium in drinking water is determined and optimized sampling patterns by both, the optimal deletion and subsequent addition of locations is generated for three divisions of Punjab, Pakistan. Several variogram models are used for modeling the spatial dependence existing in the data. Maximum likelihood (ML), restricted maximum likelihood (REML), ordinary least square (OLS), and weighted least square (WLS) are used to estimate the parameters of the variogram. Universal kriging and Bayesian kriging with varying trend are used for the prediction of sodium concentration at unobserved locations. Moreover, spatial simulated annealing optimized sampling patterns are generated.

Materials and Methods

Study Area

Three divisions of Punjab (Bahawalpur, Dera Ghazi Khan and Multan) are investigated for study purpose. The Bahawalpur division is a second-order administrative division that includes the districts Bahawalnagar, Bahawalpur and Rahimyar Khan. It is located at an elevation of 92 meters above sea level. Its latitude is 28°30′0" North and longitude is 71°30′0" East in Degrees, Minutes and Seconds (DMS) or 28.5 and 71.5 in decimal degrees. The Dera Ghazi Khan

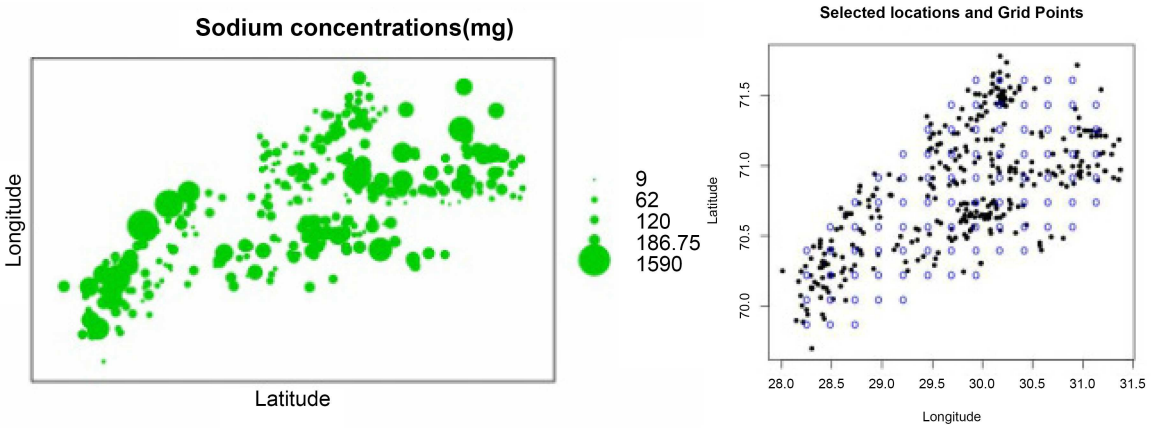

Fig 1. Level of sodium concentration (mg/L) at observed locations in threedivisions of Punjab, Pakistan (left), and observed locations and gridded locations (right).

doi:10.1371/journal.pone.0161810.g001

division has four districts (Dera Ghazi Khan, Layyah, Muzaffargarh and Rajanpur). Geographical coordinates of Dera Ghazi Khan are 30°3′22" North, 70°38′4" East. Multan division has also four districts (Khanewal, Lodhran, Multan and Vehari). Multan district is spread over an area of 3721 square km. Its latitude is 30°12′54" North and longitude is 71°35′27" East.

The Pakistan Council of Research in Water Resources (PCRWR) collected and analyzed two samples from each union council (sub-sub-sub district). The water samples were collected from easily and frequently available sources such as, hand pump, tab-water, tube-well, and water supply because it depends on inhabitants of the region. According to the PCRWR survey (2008) the samples of 370 selected sites are shown in left panel of Fig 1 and the gridded unsampled locations are presented in right panel of Fig 1. The samples of water were collected in 500 ml polystyrene bottles from the two selected sites of each union council according to a standardized method. The details about selection of samples laboratory analysis is provided in [14]. Various properties of the selected samples were analyzed according to the 2540C APHA (1992) standard.

Spatial Interpolation Methods

Kriging is extensively used in spatial analysis and its main objective is to predict the value at an unobserved location by calculating the weighted average of samples at observed locations, see Matheron [15].

Universal kriging. Ordinary kriging assumes that the mean is constant in the study region, however, in universal kriging the mean is a function of the coordinates (trend). The trend can be modeled through linear or polynomial functions. Let $Y = (Y_1, \ldots, Y_n)^T$ be a vector of response variables measured at observed locations $x_1, x_2, \ldots, x_n$, and let its distribution be as follows:

$$Y \sim N(\mu, \Sigma_Y), \tag{1}$$

where $\mu = (\mu(x_1), \mu(x_2), \ldots, \mu(x_n))^T$, and

$$\mu(x_i) = \sum_{k=1}^{p} \beta_k f_k(x_i), \quad \text{for } i = 1, 2, \ldots, n. \tag{2}$$

Here β_k, $k = 1, 2, \ldots, p$, are unknown regression coefficients and the $f_k(x)$ are known functions of the spatial coordinates x to $\Re^1$; p is the number of functions that are used to model the

trend component. The covariance matrix Σ_Y is defined as

$$\Sigma_Y = \sigma^2((1-\tau^2)R_\alpha + \tau^2 I), \tag{3}$$

where R_α is a correlation matrix generated from one of the well known positive definite correlation function models, i.e. Gaussian, exponential or spheric. Σ_Y depends on a vector-valued parameter $\theta = (\sigma^2 = \text{total sill}, \alpha = \text{range}, \tau^2 = \text{relative nugget})$. Equivalently one may define also the matrix of semivariogram values

$$\Gamma_Y = \sigma^2 \mathbf{1} - \Sigma_Y,$$

where $\mathbf{1}$ is the nxn-matrix of ones. The covariance or semivariogram parameters θ are estimated using estimation methods like ML, restricted-ML or empirical semivariogram estimation and weighted-least-squares-fitting. The universal kriging predictor at an unsampled location x_0 is a linear predictor $\hat{Y}(x_0) = \sum_{i=1}^{n} \lambda_i Y_i$. Minimizing the mean squared error of prediction subject to the condition of unbiasedness results in the following system of simultaneous equations for the weighting vector $\lambda = (\lambda_1, \lambda_2, \ldots \lambda_n)^T$:

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & f_1^1 & \cdots & f_p^1 \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} & f_1^2 & \cdots & f_p^2 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} & f_1^n & \cdots & f_p^n \\ f_1^1 & f_1^2 & \cdots & f_1^n & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ f_p^1 & f_p^2 & \cdots & f_p^n & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ l_1 \\ \vdots \\ l_p \end{bmatrix} = \begin{bmatrix} \gamma_{01} \\ \gamma_{02} \\ \vdots \\ \gamma_{0n} \\ f_1^0 \\ \vdots \\ f_p^0 \end{bmatrix} \tag{4}$$

where l_k, $k = 1, 2, \ldots, p$ are the Lagrange multipliers for the conditions of unbiasedness, γ_{ij} are the semivariogram values between locations x_i and x_j, $i, j = 0, 1, 2, \ldots, n$ and $f_k^i = f_k(x_i)$, $i = 0, 1, 2, \ldots, n$, $k = 1, 2, \ldots, p$. The weights vector λ and Lagrange multiplier l are calculated by solving the above system of equations and are utilized also for calculating the universal kriging variance:

$$\sigma_{uk}^2 = \left\{ \sum_{i=1}^{n} \lambda_i \gamma_{i0} \right\} + \left\{ \sum_{k=1}^{p} l_k f_k(x_0) \right\}. \tag{5}$$

Bayesian Kriging. Bayesian kriging makes use of the Bayes theorem which involves the likelihood function $l(\theta;Y)$ and the prior distribution $\pi(\theta)$ of the respective parameters $\theta =$ (total sill = σ^2,range = α,relative nugget = τ^2,trend = β), see Omre [16]. The posterior distribution of the parameter vector θ can be expressed as:

$$\pi(\theta|Y) = \frac{l(\theta; Y).\pi(\theta)}{\int l(\theta; Y).\pi(\theta) d\theta} \tag{6}$$

All model parameters are considered uncertain, and the prior distribution of parameters is given as follows:

$$\pi(\beta, \sigma^2, \tau^2, \alpha) = \pi(\beta|, \sigma^2, \tau^2, \alpha)\pi(\sigma^2|\tau^2, \alpha)\pi(\tau^2, \alpha), \tag{7}$$

whereas noninformative prior for β is used i.e $\pi(\beta|, \sigma^2, \tau^2, \alpha) \propto 1$. Now, using the above prior

distribution of parameters, the posterior distribution is obtained as

$$[\beta, \sigma^2, \tau^2, \alpha|Y] = [\beta|\sigma^2, \tau^2, \alpha, Y][\sigma^2|\tau^2, \alpha, Y][\tau^2, \alpha|Y], \tag{8}$$

where $[\beta|\sigma^2, \tau^2, \alpha, Y]$ is Gaussian with mean the generalized least squares estimate of β and covariance matrix the corresponding generalized least squares covariance matrix. The posterior density for τ^2 and α is given as

$$\pi(\tau^2, \alpha|Y) \propto \pi(\tau^2, \alpha)|\mathbf{F}^T\mathbf{R}_{\tau^2,\alpha}^{-1}\mathbf{F}|^{\frac{1}{2}}|\mathbf{R}_{\tau^2,\alpha}|^{-\frac{1}{2}}S^{2\frac{n-p}{2}}, \tag{9}$$

where $\mathbf{F}$ is the design matrix of regression functions, $\mathbf{R}_{\tau^2,\alpha}$ is the correlation matrix between the data and

$$S^2 = \frac{1}{n-p}(\mathbf{Y} - \mathbf{F}\hat{\beta})^T\mathbf{R}_{\tau^2,\alpha}^{-1}(\mathbf{Y} - \mathbf{F}\hat{\beta}), \tag{10}$$

with $\hat{\beta}$ the generalized least squares estimate of β. For the case that a noninformative prior $\pi(\sigma^2|\tau^2, \alpha) \propto 1/\sigma^2$ is used, the posterior of σ^2 is given as scaled inverse chi-square distribution:

$$\sigma^2|\tau^2, \alpha, Y \sim \chi^{-2}(n-p, S^2), \tag{11}$$

which is equivalent to

$$\frac{(n-p)S^2}{\sigma^2}|\tau^2, \alpha, Y \sim \chi^2(n-p), \tag{12}$$

To predict values at unsampled locations the predictive distribution is used. Proposed by Diggle and Lophaven [17] and described in Diggle and Ribeiro [18] the predictive distribution for the unobserved signal process $Y_0 = Y(x_0)$ at location x_0 is specified as:

$$[Y_0|Y] = \int\int\int [Y_0|\sigma^2, \alpha, \tau^2, Y][\sigma^2, \alpha, \tau^2|Y]d\sigma^2 d\alpha d\tau^2 \tag{13}$$

Diggle and Ribeiro [18] use a discrete prior for the covariance parameters. This way the posterior becomes also discrete and parameters can be simulated easily from the posterior by means of multinomial sampling. Sampling from the predictive distribution is performed by first sampling a parameter vector from the discrete posterior and then with the parameter vector fixed sampling from $[Y_0|\sigma^2, \alpha, \tau^2, Y]$, which is a Gaussian distribution with mean, the universal kriging predictor, and variance, the universal kriging variance. If no discrete prior $\pi(\tau^2, \alpha)$ is used but a continuous one then one has to use MCMC methods to sample from the posterior $\pi(\tau^2, \alpha|Y)$.

Spatial Sampling Design

Spatial sampling is a process in which a number of samples are used to evaluate the content of a larger geographical region. Every point in a sample exhibits information about the variable of interest at an unsampled spatial location. The sampling process has many advantages over the complete enumeration; for example, low cost, greater speed and higher scope, see Cochran [19]. The major objective of spatial sampling is to get the desired results with high precision at low cost. This can be achieved by allocating samples to locations based on minimizing an objective function, i.e., the mean universal kriging variance, see Wang et al. [20].

Spatial Simulated Annealing. Here, SSA is used for the optimization of the sampling design, see Van Groenigen and Stein [12]. In SSA locations are candidate measurements that are either removed or added iteratively and are optimized by minimizing the Mean Universal Kriging Variance (MUKV). The MUKV is an average of all kriging variances over a fine square

grid of locations in the study region i.e.

$$MUKV = \frac{\sum \sigma_i^2}{n}$$

where σ_i^2 is the variance of universal kriging. If the trend is constant, the algorithm uses the ordinary kriging variance. As the trend varies, it uses the universal kriging variance. The SSA-algorithm used here is just standard simulated annealing algorithm as described i.e. in Kirkpatrick [21] but adapted to the spatial context by following Brus and Heuvelink [11]. In SSA two different design tasks can be accomplished by: i.) Adding n new sample locations to an existing monitoring-network of m locations. ii.) Selecting $n < m$ samples from an existing network of m locations.

SSA can be described as; i): The algorithm is initialized by randomly selecting n design-locations from the spatial region and by calculating the MUKV based on these random locations and maybe other available m given locations. Let's denote this MUKV as $MUKV_0$. As the next step these n previously selected locations are randomly perturbed. Again the MUKV is calculated for n new locations i.e $MUKV_1$. If $MUKV_1$ is smaller than $MUKV_0$ then it can be assumed that there is an improvement with last design and is stored for memorizing. If $\Delta_{MUKV} = MUKV_1 - MUKV_0 > 0$ then the previous design is accepted only with a certain probability i.e.

$$p = exp\left(-\frac{\Delta_{MUKV}}{T}\right) \tag{14}$$

T is called temperature and at the beginning of the algorithm can be large. The larger T the higher is the probability that a worsening design will be accepted. The algorithm now starts to iterate: Every current design is randomly perturbed as described above; its MUKV is calculated and compared to the $MUKV$ of the so far best design. Improvements are always accepted and stored and worsenings are accepted with the above probability, where $MUKV_0$ takes over the role of the $MUKV$ of the so far best design. Actually, the probable acceptance of worsenings prevents the algorithm from being trapped in local minima of the MUKV. Further accuracy can be achieved by accounting uncertainty of variogram parameters and kriging predictors.

Cross Validation

Cross validation statistics are used to compare the performance of different fitted models. In the present study leave-one-out cross-validation is used for selecting appropriate variogram models, parameter estimation methods and kriging predictors. The Root Mean Squared Prediction Error(RMSPE) is used as performance measure:

$$\mathrm{RMSPE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}(x_i) - Y(x_i))^2}{n}}$$

In the above statistic $\hat{Y}(x_i)$ are the cross-validated, predicted values from a specific model and $Y(x_i)$ are the observed values of the data. Finally, the method with minimum RMSPE is used for predicting the response variable at unobserved locations.

Results and Discussion

Exploratory Spatial Data Analysis

Exploratory spatial data analysis is performed on the sodium concentration data by using the geoR package of Ribeiro and Diggle [22] and R software Team [23]. This analysis is carried out to explore the spatial auto-correlation and the assumption of normality which is necessary for

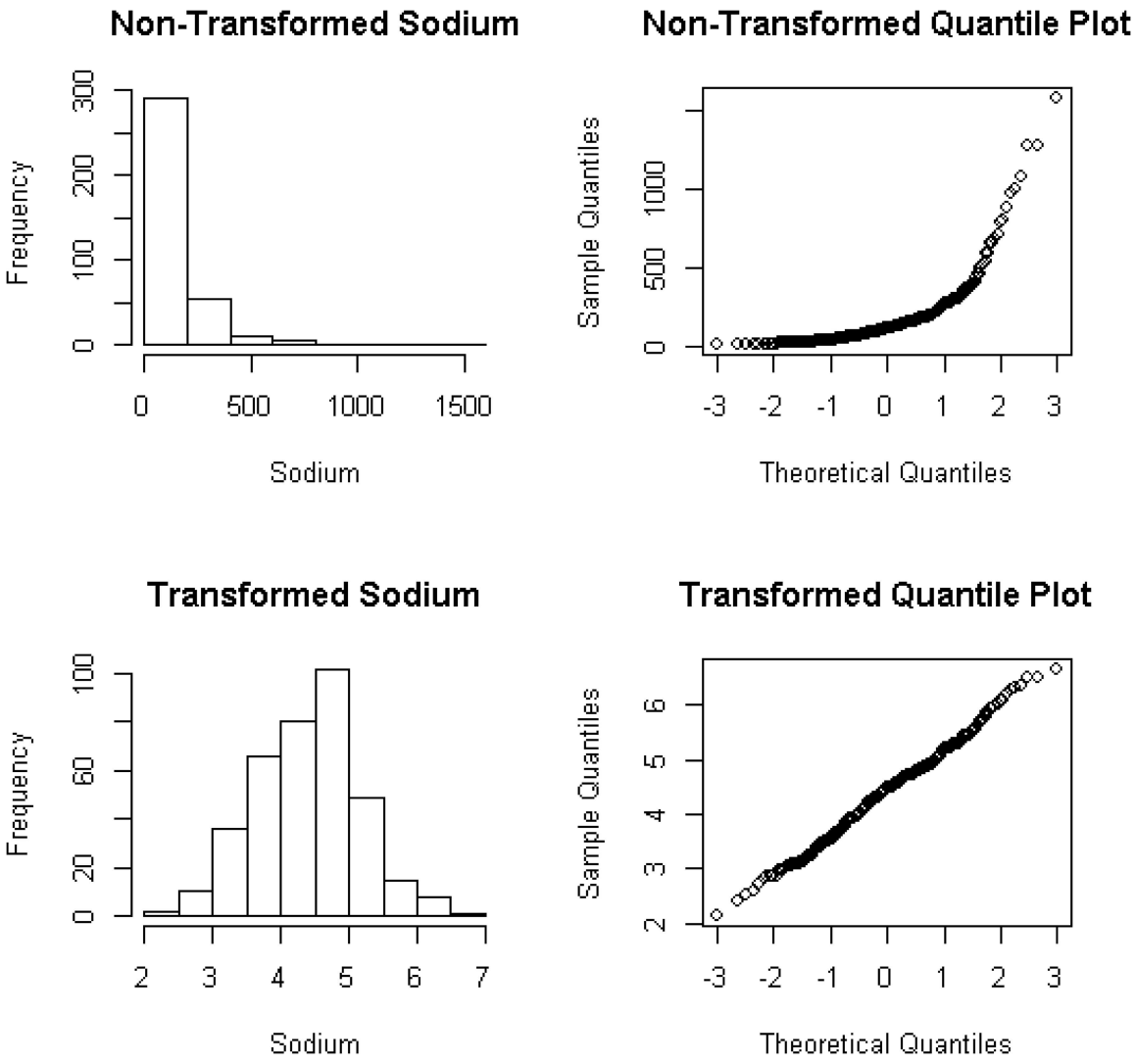

Fig 2. Distributions of non-transformed and transformed response variable withtheir respective quantile plots.

doi:10.1371/journal.pone.0161810.g002

most of the kriging methods. It is observed that the sodium concentration data violate the assumption of normality. The Box-Cox transformation with parameter $\lambda = -0.028$ is used to normalize the data, see Fig 2. The transformed data are used for modeling the spatial distribution of sodium concentrations and for selecting optimal sampling designs.

Results for Universal Kriging

From exploratory analysis it is observed that there exists spatial dependence between sodium concentrations and coordinates, see left panel of Fig 1. Universal kriging with linear trend can take into account this spatial dependence. Since modeling the variogram is essential for prediction by kriging, various variogram models are fitted through different estimation methods (results of RMSPE given in Table 1). Table 1 shows that REML provides minimum RMSPE for covariance (see column 2-5 of Table 1). REML and universal kriging provide minimum RMSPE as REML has the advantage of removing the trend during variogram estimation. From the RMSPE presented in Table 1 it can be concluded that universal kriging with a spherical

Table 1. RMSPE of spatial covariance models subject to different methods of estimation (ML, REML, OLS, and WLS) and universal kriging.

Models	ML	REML	OLS	WLS
		Universal kriging		
Expon.	178.227	177.208	179.242	187.428
Gaussian	179.244	178.612	180.022	182.425
Spherical	177.998	176.603	181.518	181.578
Cubic	178.731	178.142	179.551	179.993

doi:10.1371/journal.pone.0161810.t001

variogram model fits well under REML. The estimated parameters of the best fitting model are given in Table 2. Thus, the spherical model is used for the prediction of sodium concentrations in drinking water. The contour plots of predicted values and prediction variances are given in left panels of Figs 3 and 4, respectively. Left panel of Fig 3 shows that the concentration of sodium is highly exceeding the limit, which is $200mg/l$ in the region between (28.3° – 28.5° latitude and 70° – 70.5° longitude), (30.2° – 30.4° latitude and 70.2° – 70.4° longitude) and (30.1° – 31.1° latitude and 70.9° – 71.5° longitude). Left panel of Fig 4 shows that in most of the area the variance of prediction is small except in the region between (30.2° – 30.7° latitude and 70.9° – 71.2° longitude).

Results for Bayesian Kriging

Bayesian kriging with linear trend is used with spherical and exponential variogram models. The prior distributions are specified as follows: for the mean parameter β a uniform prior is used, i.e $p(\beta) \propto 1$; for the range parameter α a uniform prior is used, too, i.e $p(\alpha) \propto 1$; the prior for σ^2 (total sill) is scaled inverse chi-square with 367 degrees of freedom; and for the relative nugget parameter τ^2 a uniform prior is used, too. The exponential and spherical covariance models are investigated. Leave-one-out cross-validation is used to estimate the RMSPE for each model. The RMSPE presented in Table 3 shows that the spherical model fits well (minimum RMSPE) for Bayesian kriging with linear trend. For predicting the sodium concentration at unsampled locations the posterior predictive distribution is used to draw contour maps. Right panel of Fig 3, showing the posterior predictive means, indicates that sodium concentrations are high in the regions (28.3° – 28.6° latitude and 69.9° – 70.5° longitude), (29° – 29.7° latitude and 70.1° – 70.7° longitude), and concentration of sodium is a serious issue in the regions (30.1° – 31.1° latitude and 70.9° – 71.5° longitude) and (30.3° – 30.5° latitude and 70.2° –70.4° longitude). From right panel of Fig 4 it can be observed that in most of the area the posterior predictive variance is similar except in the region (30.5° – 31.0° latitude and 71.4° – 71.6° longitude), where it is extremely high.

Comparison of Spatial Interpolation Methods

Root mean squared errors of prediction of the spatial interpolation methods are given in Table 4. The RMSPE of Bayesian universal kriging is less than the RMSPE of universal kriging. It can be conclude that Bayesian universal kriging performs better than universal kriging. This

Table 2. Parameters of the variogram model subject to REML estimation and universal kriging.

Spherical Model			
Methods	Sill	Range	Nugget
Universal kriging	0.7078	1.1709	0.43

doi:10.1371/journal.pone.0161810.t002

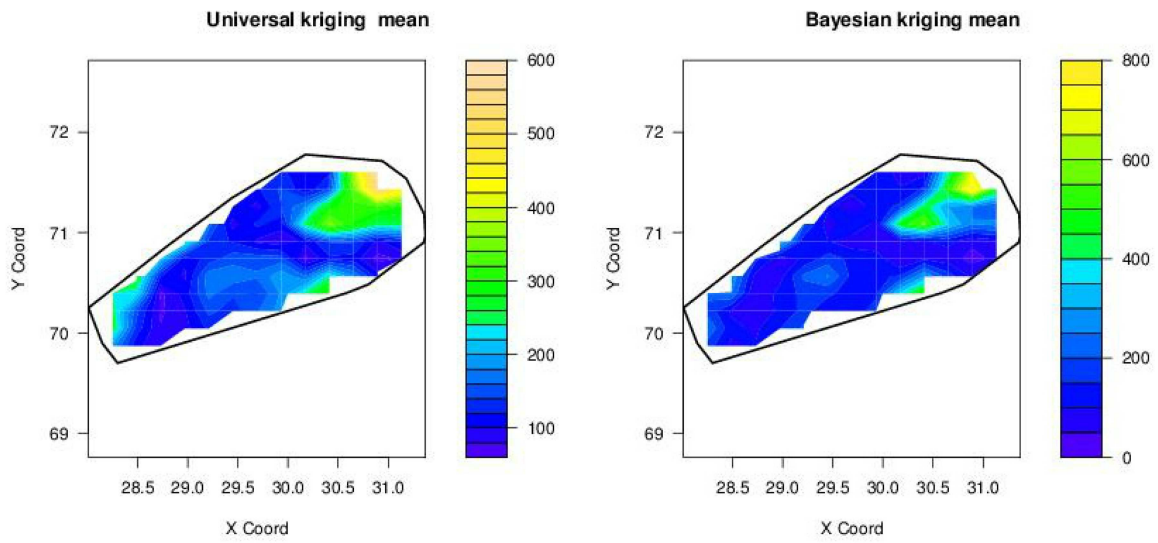

Fig 3. Maps of predicted sodium concentrations (mg/L) by using universal kriging(left) and Bayesian universal kriging (right), here X-coord = Latitude and Y-coord = Longitude.

doi:10.1371/journal.pone.0161810.g003

is due to the use of additional information, data information as well as prior information, and more statistical modeling flexibility of Bayesian universal kriging.

Optimized Sampling Design

One of the objectives of the present paper is to obtain an optimum allocation of sampling locations to minimize cost and prediction error. Since the kriging prediction variance depends on both the sample size, the variance and covariance of the sample. Optimum allocation can be obtained only by considering all of these factors. Different variogram models are fitted as it is a pre-requisite for prediction by kriging. The variogram model providing minimum RMSPE is

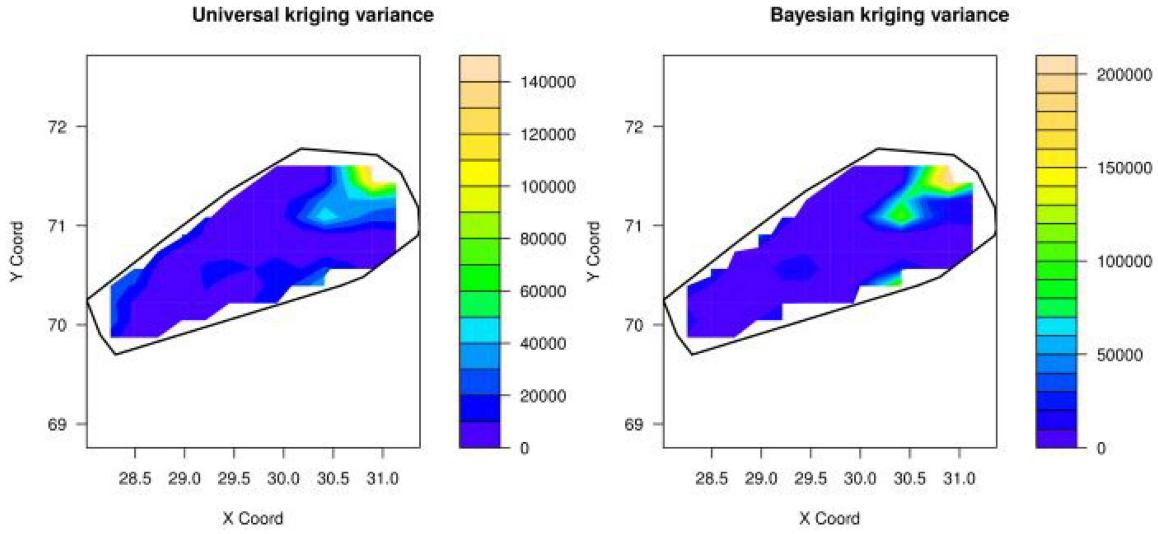

Fig 4. Maps of prediction variance by using universal kriging (left) and Bayesianuniversal kriging (right),here X-coord = Latitude and Y-coord = Longitude.

doi:10.1371/journal.pone.0161810.g004

Table 3. RMSPE of spatial covariance models subject to Bayesian universal kriging.

RMSPE		
Methods	**Exponential**	**Spherical**
Bayesian univ. kriging	177.195	175.601

doi:10.1371/journal.pone.0161810.t003

Table 4. Comparison of the interpolation methods.

Methods	RMSPE
Universal kriging	176.603
Bayesian universal kriging	175.601

doi:10.1371/journal.pone.0161810.t004

Table 5. MUKV vs. sample size using spherical variogram model.

Method	Sample Size	MUKV
Ordinary kriging	370	24018
Universal kriging	370	23930

doi:10.1371/journal.pone.0161810.t005

considered as appropriate for prediction and sampling design. For selecting the optimal sampling design optimal deletion and subsequent addition of locations are performed. The diameter was fixed by k-means i.e $k = 6$ and temperature was fixed by following Brus and Heuvelink [11].

Table 5 shows the MUKV using the observed 370 data locations for both, ordinary and universal kriging, and a spherical variogram model which comes out to be the appropriate model.

Deleting Locations using Spatial Simulated Annealing

To minimize sampling cost one possible option is to reduce the number of sampling locations. Spatial simulated annealing (SSA) is used to optimize the effect of deleting sampling locations upon MUKV. The MUKV is calculated using ordinary and universal kriging. The MUKV increases in both kriging techniques as the number of data locations goes down, see Table 6 and Fig 5. Subsequently deleting 20 and 40 locations from the 370 locations. When deleting 50 locations from the 370 locations the MUKV using ordinary kriging is 24560 and the MUKV for universal kriging is 24480, see Tables 5 and 6. According to the MUKV criterion it can be inferred that universal kriging theoretically seems to perform better than ordinary kriging. The sampling patterns for deleting locations using ordinary kriging and universal kriging are

Table 6. MUKV vs. sample size using spherical variogram model for deleting locations.

Method	Sample Size	MUKV
Ordinary Kriging	370-10	24538
	370-20	24539
	370-30	24542
	370-40	24549
	370-50	24560
Universal Kriging	370-10	24456
	370-20	24457
	370-30	24461
	370-40	24467
	370-50	24480

doi:10.1371/journal.pone.0161810.t006

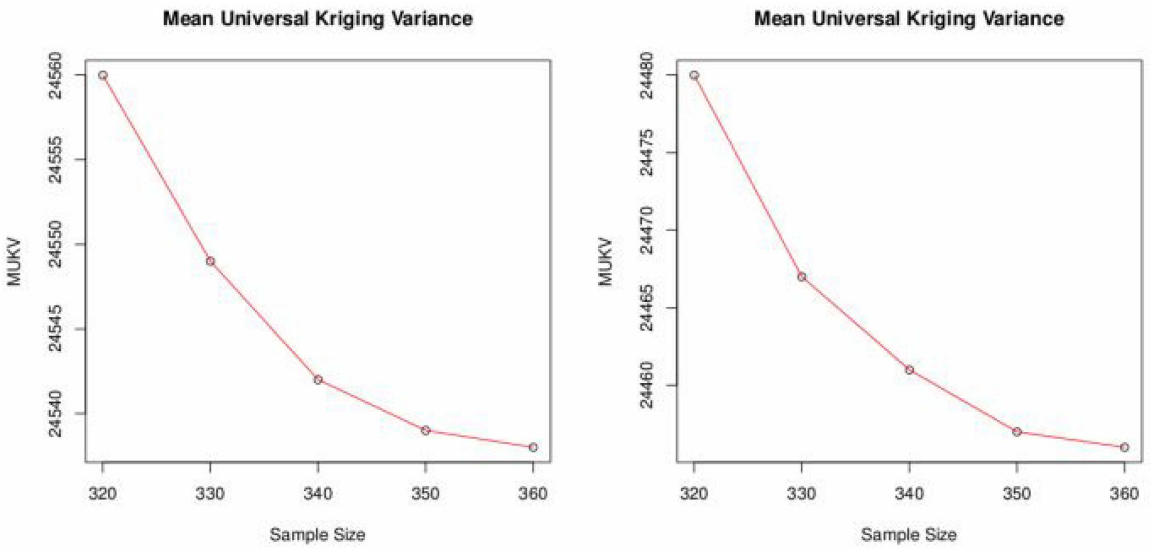

Fig 5. MUKV vs. sample size using ordinary and universal kriging.

doi:10.1371/journal.pone.0161810.g005

shown in Figs 6 and 7. Algorithm stops on getting the optimum criterion i.e., MUKV. It can be observed that number of iterations performed by the algorithm are different in every sub figure. Obviously, only redundant locations in densely sampled areas are deleted, resulting in only a slight increase of the MUKV.

Adding Locations using Spatial Simulated Annealing

Like before the objective here is to find the sampling pattern by adding locations that has minimum MUKV. If the number of locations to be added are increased in the SSA-algorithm then the MUKV decreases for both, ordinary and universal kriging, see Fig 8 and Table 7. However, the mean universal kriging variance is smaller than the ordinary kriging one. The sampling

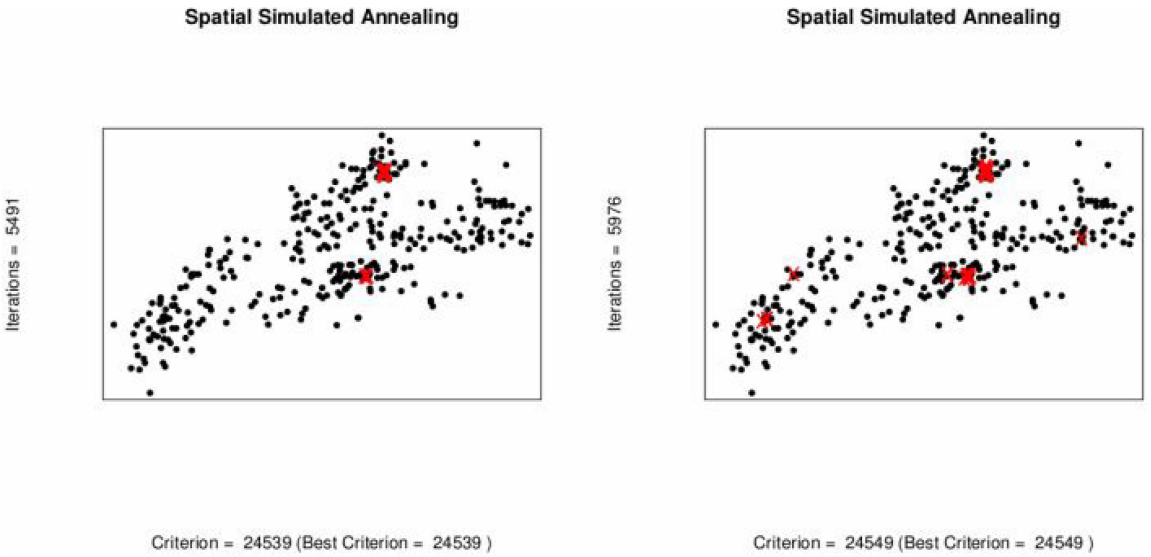

Fig 6. Optimal sampling design for ordinary kriging when locations are deleted,(Red).

doi:10.1371/journal.pone.0161810.g006

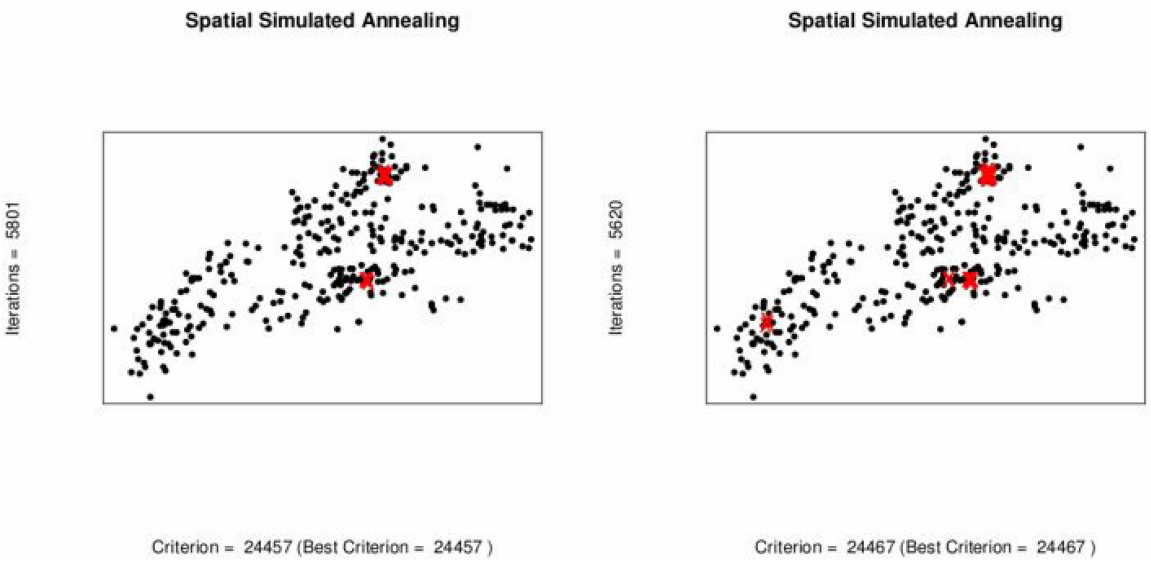

Fig 7. Optimal sampling design for universal kriging when locations are deleted (Red).

doi:10.1371/journal.pone.0161810.g007

patterns after adding locations using ordinary kriging and universal kriging are shown in Figs 9 and 10. Both sampling designs look quite similar and space-filling.

Conclusion

The present paper is focused on two objectives: (i) to model the spatial distribution of sodium concentration, and (ii) to generate optimized spatial sampling designs for three divisions of Punjab, Pakistan. Universal kriging and Bayesian kriging with varying linear trend are used for modelling the spatial distribution of sodium concentrations in drinking water. To take account of the spatial dependence in the response variable Gaussian, exponential, spherical and cubic

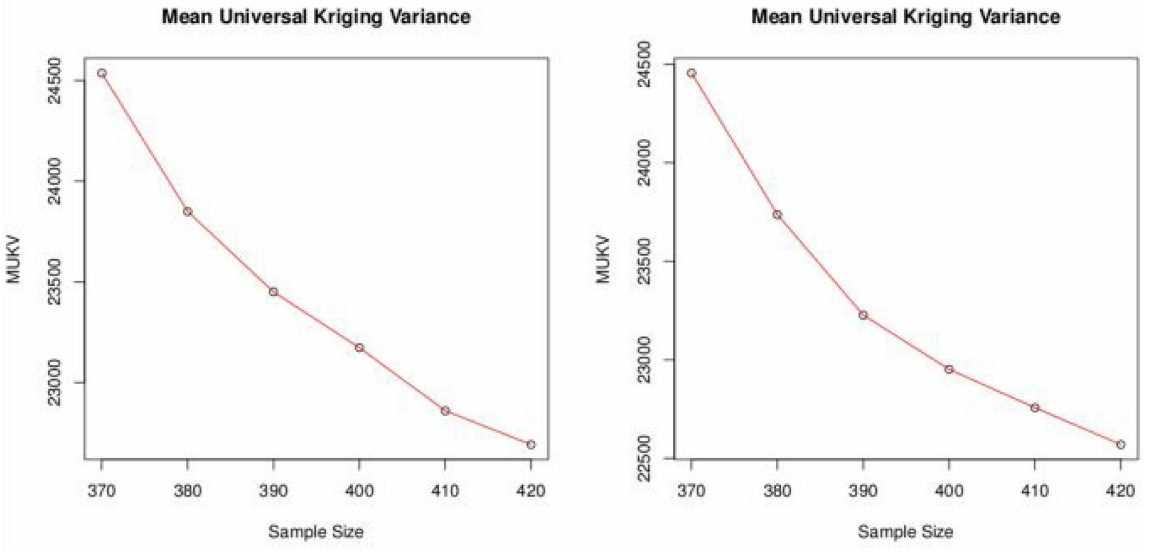

Fig 8. MUKV vs. sample size using ordinary and universal Kriging.

doi:10.1371/journal.pone.0161810.g008

Table 7. MUKV by varying the sample size using a spherical model for adding locations.

Method	Sample Size	MUKV
Ordinary Kriging	370+10	23849
	370+20	23450
	370+30	23173
	370+40	22860
	370+50	22692
Universal Kriging	370+10	23737
	370+20	23227
	370+30	22952
	370+40	22757
	370+50	22570

doi:10.1371/journal.pone.0161810.t007

variogram models are used. It is concluded that the spherical variogram model provides smaller mean squared error of prediction than other variogram models. Since Bayesian universal kriging has the advantage of utilizing prior information about model parameters and is statistically more flexible, it performs better than universal kriging.

A comparison of the used kriging methods can also be based on credible intervals. The 95% credible intervals are estimated from the simulated values of sodium concentration. Plots of credible intervals as presented in Fig 11 show that Bayesian universal kriging has shorter intervals than universal kriging. In Bayesian universal kriging most of the actual data values lie within the 95% predictive intervals. Thus, Bayesian universal kriging gives more reliable predictions than universal kriging. This is also due to the fact that Bayesian universal kriging does take into account the uncertainty of the covariance model. Universal kriging does not; once the covariance model is estimated it is fixed and its uncertainty is discarded during prediction. Prediction maps of sodium concentrations are generated based on Bayesian universal and universal kriging; locations having sodium concentration above the threshold value of 200mg are

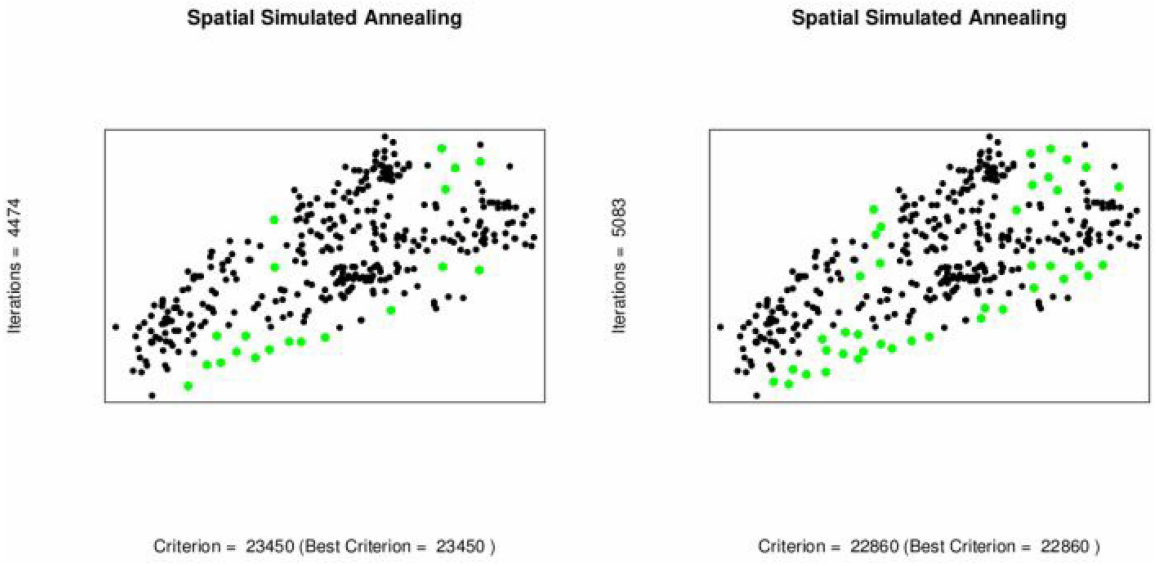

Fig 9. Sub-optimal 370+40 point design, ordinary kriging. Green: added locations.

doi:10.1371/journal.pone.0161810.g009

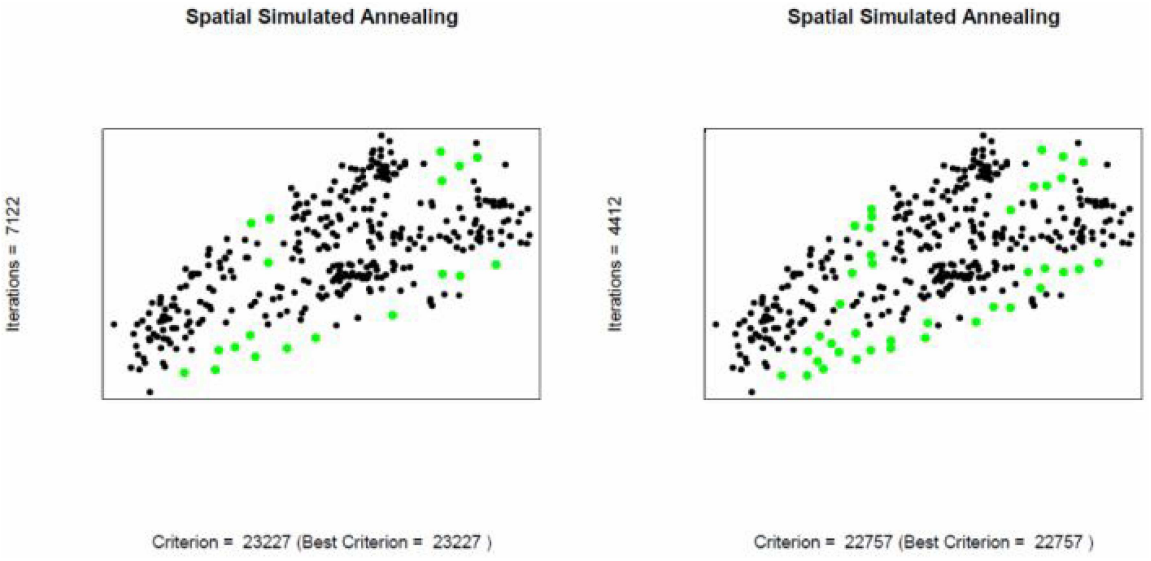

Fig 10. Suboptimal 370+40 point design, universal kriging. Green: added (20 &40) locations.

doi:10.1371/journal.pone.0161810.g010

identified. Prediction maps show that sodium concentrations in drinking water are increasing to the southern side of the study area.

Spatial simulated annealing is used to generate optimized spatial designs for adding and deleting locations. For optimization the MUKV is used as objective function subject to ordinary kriging or universal kriging. When deleting locations the MUKV increases; however, the increase is not much because redundant locations are deleted. If locations are added the MUKV decreases; the designs themselves have a space-filling character, when adding locations by means of SSA. Spatial simulated annealing optimize the patterns, which are generated by deleting and adding locations. Time and cost may be saved by deleting the redundant locations, and the variation can be minimized by adding locations that was unfilled. Recently, Junez-Ferreira and Herrera [24] used Kalman filter to sequentially optimize space-time monitoring points. Their suggested method can minimize the prediction error in better way as compared

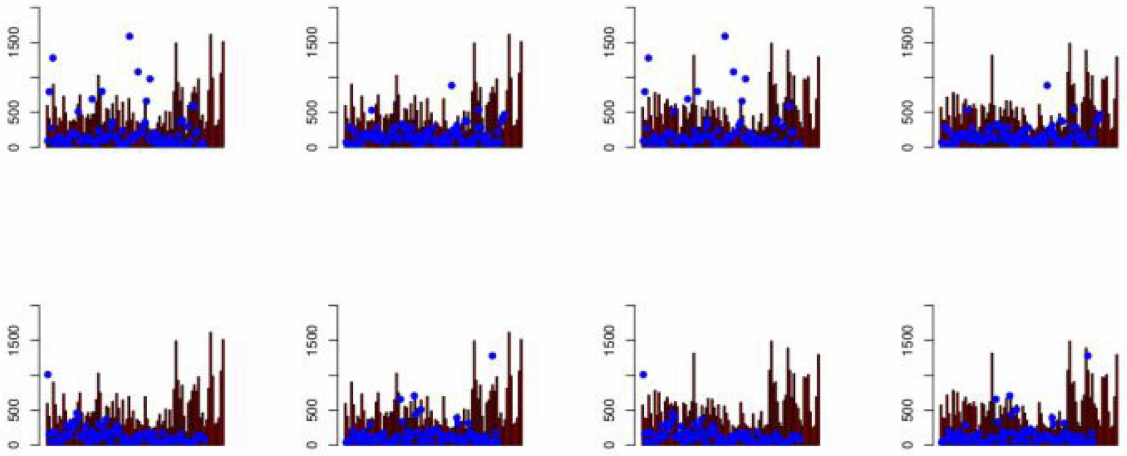

Fig 11. The 95% predictive interval plots and actual values of data for universalkriging (left) and Bayesian Universal Kriging (right), the dots represent to actual interval and bars represent to 95% credible intervals.

doi:10.1371/journal.pone.0161810.g011

with simulated annealing algorithm. For further improvement in optimizing sampling design in present paper the method suggested in [24] can be used.

Supporting Information

S1 File. R-Language program used for statistical analysis, "S1 File.r".
(R)

S2 File. Data used in the manuscript "S2 File.csv".
(CSV)

Acknowledgments

The authors are grateful to the two anonymous referee for their valuable comments and feedback. The authors would like to extend their sincere appreciation to the Deanship of Scientific Research at King Saud University for funding this work through Research Group no. RGP-210.

Author Contributions

Conceived and designed the experiments: IH.

Performed the experiments: EZ.

Analyzed the data: EZ GS.

Contributed reagents/materials/analysis tools: JS TH.

Wrote the paper: GS MF.

Corrected by: JS.

Visualized the results, made final corrections and funded for preparation of manuscript: NMAS.

References

1. Gadgil A. Drinking water in developing countries. Annual Review of Energy and the Environment. 1998; 23(1):253–286. doi: 10.1146/annurev.energy.23.1.253
2. Ferreccio C, González C, Milosavjlevic V, Marshall G, Sancha AM, Smith AH. Lung cancer and arsenic concentrations in drinking water in Chile. Epidemiology. 2000; 11(6):673–679. PMID: 11055628
3. Gundogdu KS, Guney I. Spatial analyses of groundwater levels using universal kriging. Journal of Earth System Science. 2007; 116(1):49–55. doi: 10.1007/s12040-007-0006-6
4. Neuman SP, Xue L, Ye M, Lu D. Bayesian analysis of data-worth considering model and parameter uncertainties. Advances in Water Resources. 2012; 36:75–85. doi: 10.1016/j.advwatres.2011.02.007
5. Mehrjardi R, Jahromi M, Mahmodi S, Heidari A. Spatial distribution of groundwater quality with geostatistics (Case Study: Yazd-Ardakan plain). World Applied Sciences Journal. 2008; 4(1):9–17.
6. Nas B. Geostatistical approach to assessment of spatial distribution of groundwater quality. Polish Journal of Environmental Studies. 2009; 18:1073–1082.
7. Sarukkalige R. Geostatistical Analysis of Groundwater Quality in Western Australia. IRACST. 2012; 2 (4):2250–3498.
8. Andrade A, Stigter T. The distribution of arsenic in shallow alluvial groundwater under agricultural land in central Portugal: Insights from multivariate geostatistical modeling. Science of the Total Environment. 2013; 449:37–51. doi: 10.1016/j.scitotenv.2013.01.033 PMID: 23410893
9. Dhar A, Datta B. Logic-based design of groundwater monitoring network for redundancy reduction. Journal of water resources planning and management. 2009; 136(1):88–94. doi: 10.1061/(ASCE)0733-9496(2010)136:1(88)

10. Siri JG, Lindblade KA, Rosen DH, Onyango B, Vulule JM, Slutsker L, et al. A census-weighted, spatially-stratified household sampling strategy for urban malaria epidemiology. Malaria Journal. 2008; 7 (1):39. doi: 10.1186/1475-2875-7-39 PMID: 18312632

11. Brus DJ, Heuvelink G. Optimization of sample patterns for universal kriging of environmental variables. Geoderma. 2007; 138(1):86–95. doi: 10.1016/j.geoderma.2006.10.016

12. Van Groenigen J, Stein A. Constrained optimization of spatial sampling using continuous simulated annealing. Journal of Environmental Quality. 1998; 27(5):1078–1086. doi: 10.2134/jeq1998.00472425002700050013x

13. Zhu Z, Stein ML. Spatial sampling design for parameter estimation of the covariance function. Journal of Statistical Planning and Inference. 2005; 134(2):583–603. doi: 10.1016/j.jspi.2004.04.017

14. Hussain I, Shakeel M, Faisal M, Soomro ZA, Hussain M, Hussain T. Distribution of total dissolved solids in drinking water by means of bayesian kriging and gaussian spatial predictive process. Water Quality, Exposure and Health. 2014; 6(4):177–185. doi: 10.1007/s12403-014-0123-9

15. Matheron G. Principles of geostatistics. Economic Geology. 1963; 58(8):1246–1266. doi: 10.2113/gsecongeo.58.8.1246

16. Omre H. Bayesian kriging—merging observations and qualified guesses in kriging. Mathematical Geology. 1987; 19(1):25–39. doi: 10.1007/BF01275432

17. Diggle P, Lophaven S. Bayesian geostatistical design. Scandinavian Journal of Statistics. 2006; 33 (1):53–64. doi: 10.1111/j.1467-9469.2005.00469.x

18. Diggle PJ, Ribeiro PJ. Model-based Geostatistics. Springer; 2007.

19. Cochran W. Sampling Techniques. New York, Wiley and Sons. 1977; 98:259–261.

20. Wang JF, Stein A, Gao BB, Ge Y. A review of spatial sampling. Spatial Statistics. 2012; 2:1–14. doi: 10.1016/j.spasta.2012.08.001

21. Kirkpatrick V Gelatt. Optimization by simulated annealing. Science. 1983; 220(4598):671–680. doi: 10.1126/science.220.4598.671 PMID: 17813860

22. Ribeiro PJ, Diggle PJ. geoR: A package for geostatistical analysis. R News. 2001; 1(2):14–18.

23. Team RC. R: A language and environment for statistical computing. 2005;.

24. Júnez-Ferreira H, Herrera G. A geostatistical methodology for the optimal design of space—time hydraulic head monitoring networks and its application to the Valle de Querétaro aquifer. Environmental monitoring and assessment. 2013; 185(4):3527–3549. doi: 10.1007/s10661-012-2808-5 PMID: 22936025