Hindawi
Journal of Probability and Statistics
Volume 2023, Article ID 1328265, 10 pages
https://doi.org/10.1155/2023/1328265

Research Article

Applications of Robust Methods in Spatial Analysis

Selvakkadunko Selvaratnam

Department of Statistical Sciences, University of Toronto, Toronto, Canada

Correspondence should be addressed to Selvakkadunko Selvaratnam; selva.selvaratnam@utoronto.ca

Received 19 May 2022; Revised 22 September 2022; Accepted 20 April 2023; Published 10 May 2023

Academic Editor: Muhammad Ahsan

Spatial data analysis provides valuable information to the government as well as companies. The rapid improvement of modern technology with a geographic information system (GIS) can lead to the collection and storage of more spatial data. We developed algorithms to choose optimal locations from those permanently in a space for an efficient spatial data analysis. Distances between neighboring permanent locations are not necessary to be equispaced distances. Robust and sequential methods were used to develop algorithms for design construction. The constructed designs are robust against misspecified regression responses and variance/covariance structures of responses. The proposed method can be extended for future works of image analysis which includes 3 dimensional image analysis.

1. Introduction

Companies can learn consumer behavior to increase their profits through spatial data analysis. A common consumer behavior pattern can be identified in neighborhoods. When these patterns are identified, companies can reduce expenditures and wastage. Recently, massive amount of spatial data have been collected through remote sensing techniques, magnetic resonance imaging (MRI) scanners, X-ray machines, cameras, governments, and companies. These types of data are mostly nonexperimental observational data [1]. Jaworski et al. [2] noted that data analysis with a large sample is a time-consuming and expensive procedure. The subsampling method can overcome this obstacle and was developed by many authors including Rocke and Dai [3] and Salloum et al. [4]. Moreover, Wang et al. [5] and Yao and Wang [6], among others, discussed the optimal subsampling method for nonexperimental data.

Groundwater contamination started with the industrial revolution. Gas sectors, mining industries, and industrial waste are the main sources of groundwater contamination. Water pollution brings risks to human health. Therefore, groundwater monitoring is important to identify potential water contamination. The selection of optimal wells from a large number of wells lead an efficient understanding of groundwater pollution and a cost reduction in groundwater monitoring [7]. Naturally, the levels of contamination in water are highly correlated if two wells are close to each other. In this paper, we accommodate these kinds of correlations among responses in design construction.

The robustness including correlation structure among responses is discussed in many studies; for instance, see Shi et al. [8] Wiens [9] and Wiens [10] on the construction of designs. The misspecified variance/covariance structure was considered by Wiens [10] in the development of a robust method to construct designs for spatial analysis. However, they developed algorithms to choose optimal locations from equispaced locations. Wiens [11] included the misspecified variance/covariance structure in the model by incorporating robust methods. The universal kriging estimate was used in his development of the loss function for design construction. In this paper, theoretical works of Wiens [11] were applied to establish an algorithm to select optimal locations from permanent locations that are not necessarily equispaced locations.

The rest of this paper is organized as follows. We describe the model formulation and methods in §2. In §3, an algorithm is described using the sequential method, and the proposed algorithm for the design construction is validated by some test cases. In §4, we outline an algorithm to choose optimal locations from fixed permanent locations and give an example using the algorithm. Also, the discussed robust method was applied to 'coal-ash' data in the same section. We summarize our findings in §5.

2. Materials and Methods

The material in this section is based on the theory in Wiens [11]. We discuss how to find n optimal locations from a design space $\Omega = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ where $\mathbf{x}_i \in \mathfrak{R}^q$ with $\mathbf{x}_i = (x_{i0}, x_{i1}, x_{i2}, \ldots, x_{ip})'$ and $q = (p+1)$. $\mathbf{x}_i$ contains information regarding ith spatial location. We assume that the relationship between responses and locations can be expressed by a linear model. We include robustness by considering the model misspecification and correlations among responses in the construction of designs. We consider the following approximately linear model:

$$Y(\mathbf{x}_i) = \mathbf{f}'(\mathbf{x}_i)\boldsymbol{\theta} + \psi(\mathbf{x}_i) + \varepsilon(\mathbf{x}_i), \quad i = 1, 2, \ldots, N, \qquad (1)$$

for some small model error $\psi(\mathbf{x}_i)$, and $\varepsilon(\mathbf{x}_i)$ is a homoscedastic measurement error with $\mathrm{Var}[\varepsilon(\mathbf{x}_\mathrm{i})] = \sigma_\varepsilon^2$, q-dimensional vector regressors $\mathbf{f}(\mathbf{x})$, and model parameters θ. However, the experimenter assumes the incorrect model $E[Y(\mathbf{x}_i)] = \mathbf{f}'(\mathbf{x}_i)\theta$. Based on this assumption, the true unknown parameters can be obtained by

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left(E[Y(\mathbf{x}_i)] - \mathbf{f}'(\mathbf{x}_i)\boldsymbol{\theta}\right)^2. \qquad (2)$$

Define the $N \times q$ matrix $\mathbf{F}$ having rows $\{\mathbf{f}'(\mathbf{x}_i)\}_{i=1}^{N}$and $N \times 1$ vector ψ_N with elements $\{\psi(\mathbf{x}_i)\}_{i=1}^{N}$. We assume that $\mathbf{F}$ has full column rank. Condition (2) leads to the following orthogonality requirement:

$$\mathbf{F}'\boldsymbol{\psi}_N = \mathbf{0}_{q\times 1}. \qquad (3)$$

Responses $Y(\mathbf{x}_i)\, i = 1, 2, \ldots, N$ are correlated having the following covariance matrix:

$$\mathbf{C}_N = \mathrm{COV}[\mathbf{Y}] \overset{def}{=} (\sigma_{\mathrm{ij}})_{i,j=1,2,\ldots,N}, \qquad (4)$$

where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$.

In general, the experimenter has an objective to measure $n_i \geq 0$ responses $\{Y_{i_k}(\mathbf{x}_i)\}_{i_k=1}^{n_i}$ at the location $\mathbf{x}_i$. We assume that covariances among responses have the following structure;

$$\mathrm{COV}\left[Y_{\mathrm{i_k}}, Y_{\mathrm{j_l}}\right] = \sigma_{\mathrm{ij}} + \begin{cases} \sigma_\varepsilon^2, & (i,k) = (j,l), \\ 0, & \text{otherwise}. \end{cases} \qquad (5)$$

We impose the following conditions:

$$\begin{aligned} &\text{(i)} \left\|\boldsymbol{\psi}_N\right\|^2 \leq \frac{\alpha^2}{n}, \\ &\text{(ii)} \left\|\mathbf{C}_N\right\|_\mathrm{M} \leq \frac{\beta^2}{n}, \end{aligned} \qquad (6)$$

where α and β are constants, $\|\cdot\|_M$ is an induced matrix norm. The experimenter has a plan to collect data $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$.

Let Ψ be a class of functions $\psi(\cdot)$ satisfying conditions (3) and (6) and $\mathscr{C}$ be the class of positive semi-definite matrices $\mathbf{C}$ satisfying condition (6). The model misspecification is accounted by a function in Ψ and covariance matrix in $\mathscr{C}$. We define the covariance matrix of $\mathbf{y}$ by

$$\mathbf{C}_n = \mathrm{COV}[\mathbf{y}]\colon n \times n. \qquad (7)$$

Also, we define the incidence matrix $\mathbf{E}$ to express $\mathbf{C}_n$ in terms of $\mathbf{C}_N$ and it is described as follows:

$$\mathbf{E} = \begin{pmatrix} \mathbf{e}_1' \\ \vdots \\ \mathbf{e}_N' \end{pmatrix},$$

$$\text{where } \mathbf{e}_i' = \begin{cases} \mathbf{0}_{1\times n}', & n_i = 0, \\ \left(\mathbf{0}_{\sum_{j<i} n_j}' \vdots\, \mathbf{1}_{n_i}' \vdots\, \mathbf{0}_{\sum_{j>i} n_j}'\right), & n_i > 0. \end{cases} \qquad (8)$$

Thus, the covariance matrix $\mathbf{C}_n$ can be expressed by

$$\mathbf{C}_n = \mathbf{E}'\mathbf{C}_N\mathbf{E} + \sigma_\varepsilon^2 \mathbf{I}_n. \qquad (9)$$

The optimal linear predictors $\widehat{\mathbf{Y}}$ of the random quantities $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$ can be obtained by the universal kriging [12]. This task can be achieved by minimizing the prediction mean squared error (PMSE) that is defined by

$$\mathrm{PMSE} = \sum E\left(Y_i - \widehat{Y}_i\right)^2 = E\left(\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2\right). \qquad (10)$$

By using Theorem 1 of Wiens [11]; the PMSE can be written as follows:

$$\mathrm{PMSE} = \left\|\mathbf{A}_0\boldsymbol{\psi}_N\right\|^2 + \mathrm{tr}\{\mathbf{A}_0\mathbf{C}_N\mathbf{A}_0'\} + \sigma_\varepsilon^2 \mathrm{tr}\{\mathbf{L}_0\mathbf{L}_0'\}, \qquad (11)$$

for any function ψ_N in Ψ and covariance matrix $\mathbf{C}_N$ in $\mathscr{C}$, where

$$\begin{aligned} \mathbf{L}_0 &= \left[(\mathbf{I}_N - \mathbf{C}_N\mathbf{V}_0)\mathbf{F}\left(\mathbf{F}'\mathbf{V}_0\mathbf{F}\right)^{-1}\mathbf{F}' + \mathbf{C}_N\right]\mathbf{E}\mathbf{C}_n^{-1}, \\ \mathbf{A}_0 &= (\mathbf{I}_N - \mathbf{C}_N\mathbf{V}_0)\left(\mathbf{I}_N - \mathbf{F}\left(\mathbf{F}'\mathbf{V}_0\mathbf{F}\right)^{-1}\mathbf{F}'\mathbf{V}_0\right), \\ \mathbf{V}_0 &= \mathbf{E}\mathbf{C}_n^{-1}\mathbf{E}'. \end{aligned} \qquad (12)$$

Theorem 1. *Let $\mathscr{C}_\gamma = \left\{(1-\gamma)\mathbf{C}_{0;N} + \gamma\mathbf{C}_N \| \|\mathbf{C}_N\|_M \leq \beta^2/n\right\}$ for $\gamma \in [0,1]$ be a family of covariance structures, where $\mathbf{C}_{0;N}$ is the true covariance matrix. Then under the assumptions of Ψ and $\mathscr{C}$ satisfying condition (6), the maximum value of PMSE over Ψ and $\mathscr{C}$ is $(\alpha^2 + \beta^2\gamma + \sigma_\varepsilon^2 + 1 - \gamma)/n$ times*

$$\begin{aligned}\mathscr{L}(\xi) = {} & (1-a-b-c)\operatorname{ch}_{\max}\mathbf{A}_0\mathbf{A}_0' + a \cdot \operatorname{tr}\{\mathbf{A}_0\mathbf{A}_0'\} + b \cdot \operatorname{tr}\left\{n\mathbf{A}_0\mathbf{C}_{0;N}\mathbf{A}_0'\right\} \\ & + c \cdot \operatorname{tr}\{n\mathbf{L}_0\mathbf{L}_0'\},\end{aligned} \tag{13}$$

where $a = \beta^2\gamma/(\alpha^2 + \beta^2\gamma + \sigma_\varepsilon^2 + 1 - \gamma), b = (1-\gamma)/(\alpha^2 + \beta^2\gamma + \sigma_\varepsilon^2 + 1 - \gamma)$, $c = \sigma_\varepsilon^2/(\alpha^2 + \beta^2\gamma + \sigma_\varepsilon^2 + 1 - \gamma)$, $\sigma_\varepsilon = \delta\sqrt{c/b}$ *with* $0 < \delta < 1$.

The proof of Theorem 1 follows directly from Theorem 2 and Remark 1 of Wiens [11]. In this study, the loss function $\mathscr{L}(\cdot)$ in Theorem 1 is used for design constructions. We will discuss two types of correlation functions in the next section.

2.1. Correlation Matrix. We assume two correlation functions: (i) the isotropic Gaussian correlation function $\rho_{\mathrm{ij}} = \operatorname{corr}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\lambda\|\mathbf{x}_i - \mathbf{x}_j\|^2\right\}$ and (ii) the anisotropic Gaussian correlation function $\rho_{\mathrm{ij}} = \operatorname{corr}(\mathbf{x}_i - \mathbf{x}_j) = \exp\left\{-\lambda\left((\mathbf{x}_i - \mathbf{x}_j)^T \operatorname{diag}(1,5)(\mathbf{x}_i - \mathbf{x}_{\mathrm{j}})\right)^{(1/2)}\right\}$ for $i, j = 1, 2, \ldots, N$, where $\|\cdot\|$ is Euclidean norm [9]. Also, the true correlation matrix $\mathbf{P}_{0;N}$ has the following form:

$$\mathbf{P}_{0;N} = \begin{array}{c} \\ x_1 \\ x_2 \\ \cdots \\ x_{N-1} \\ x_N \end{array} \left[\begin{array}{ccccc} x_1 & x_2 & \cdots & x_{N-1} & x_N \\ \rho_{11} & \rho_{12} & \cdots & \rho_{1(N-1)} & \rho_{1N} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2(N-1)} & \rho_{2N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{(N-1)1} & \rho_{(N-1)2} & \cdots & \rho_{(N-1)(N-1)} & \rho_{(N-1)N} \\ \rho_{N1} & \rho_{N2} & \cdots & \rho_{N(N-1)} & \rho_{\mathrm{NN}} \end{array}\right]. \tag{14}$$

Wiens [9] suggests the value of λ is 0.9 that is the nearest neighbor correlation. We will use the same value of λ in the construction of optimal locations in §3 and §4.1. The true covariance matrix $\mathbf{C}_{0;N}$ can be evaluated by $\mathbf{C}_{0;N} = \sigma_0^2\mathbf{P}_{0;N}$ for a specified constant σ_0^2.

3. Design Construction

In this section, we will discuss how to choose n optimal locations from N locations in a two dimensional space. We consider the approximately linear model

$$E[Y(\mathbf{x}_i)] = \boldsymbol{\theta}'\mathbf{x}_i + \psi(\mathbf{x}_i), \quad i = 1, 2, \ldots, N, \tag{15}$$

where $\mathbf{x}_i' = (1, x_{i1}, x_{i2})$, $\theta' = (\theta_0, \theta_1, \theta_2)$, $Y(\mathbf{x}_i)$ is a response observed at $\mathbf{x}_i$, and $\psi(\mathbf{x}_i)$ is a small departure from the assumed model by the experimenter. We suppose that a region, $R1$, is a two dimensional square with 1 unit length. The region $R1$ consists of vertices (0, 0), (0, 1), (1, 0), and (1, 1). Let x_1 and x_2 be the horizontal and vertical distance (in units) from origin (0, 0). In the next subsection, brute-force procedures were applied to pick optimal locations from all possible subsamples for a given set of parameters that are required to compute the loss function.

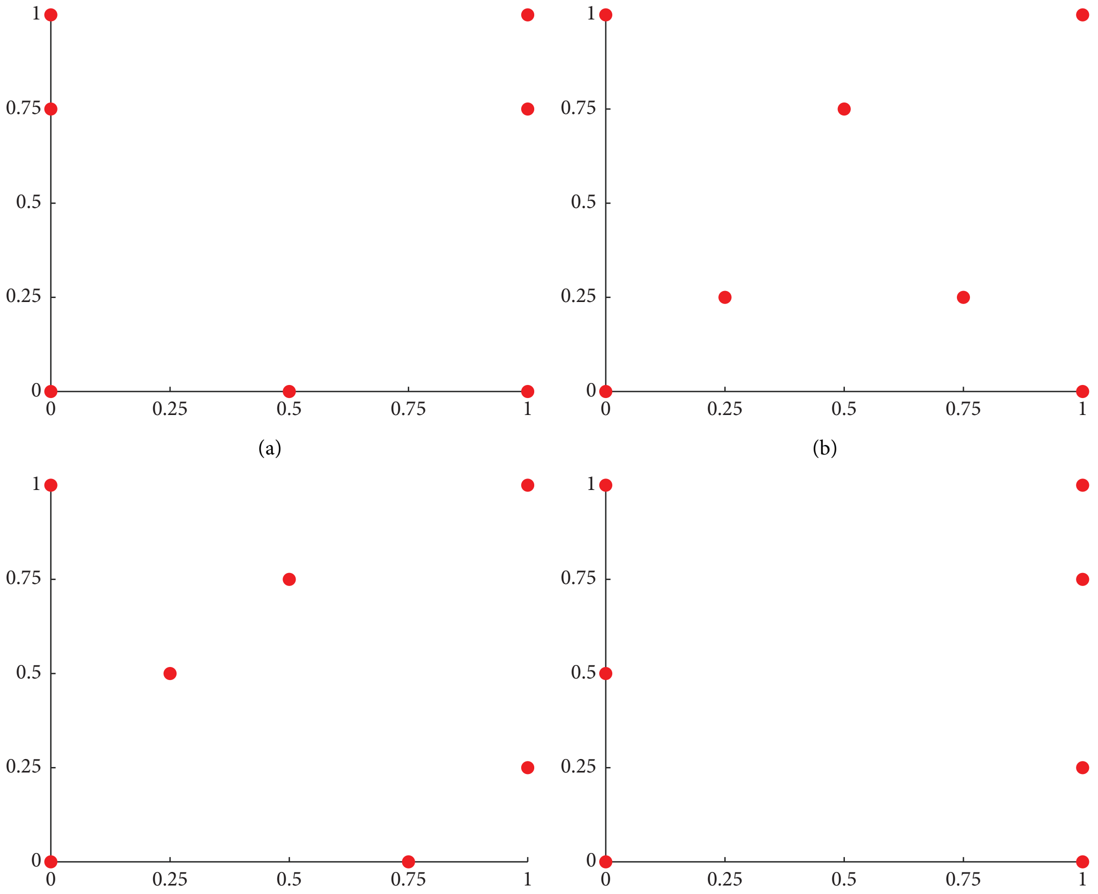
3.1. Brute-Force Search. The loss function (13) depends on parameters σ_0, δ, a, b, and c. There are restrictions among these parameters which are $0 < a + b + c < 1$, $0 < \delta < 1$, and $\sigma_0 > 0$. We chose these parameters to include a wide range of possible scenarios. Also, small or moderate values were selected for σ_0 from the interval [0.3, 2]. The selected set of parameters were reported in Table 1.

Eight different values of the parameters were taken to evaluate the proposed algorithm in §3.2 through brute-force sequential search. These values are shown in Table 1. In this section, we construct some test cases to evaluate the performance of Algorithm 1 that is discussed in §3.2. Let n_0 be the required number of locations to an investigator. These test cases can be constructed by the brute-force search of $\binom{N}{n_0}$ all possible subsample locations. We display four test cases with the assumption of isotropic Gaussian correlations structure, $N = 25$ and $n_0 = 7$ in Figure 1. In this case, 480,700 possible subsample locations were checked to obtain optimal locations. In Figure 2, we show four test cases with the assumption of anisotropic Gaussian correlations structure, $N = 36$ and $n_0 = 8$. In this scenario, 30,260,340 possible subsample locations were verified to select optimal locations. We used the MATLAB command "nchoosek" to take all possible subsample locations from N locations for the brute-force search. If the number of subsample locations is greater than 10^8, we cannot apply the command "nchoosek" for the brute-force search to choose optimal locations. Thus, further research is needed to apply the brute-force search if the number of subsample locations is greater than 10^8. However, Algorithm 1 in §3.2 and Algorithm 2 in §4.1 work for any N and $n_0\,(< N)$.

3.2. Sequential Method. The sequential method is widely applied in the area of the construction of optimal designs; for instance, Wiens [10] developed algorithms using the sequential approach to choose optimal designs. In the sequential method, one design point at a time is added to the current design. We collect spatial locations. Therefore, locations are chosen without replacement. Next, we discuss Algorithm 1, which will be based on the sequential approach.

Table 1: The selected values of parameters and the minimum loss (minloss) that was computed using optimal locations for the brute-force search and Algorithm 1.

Figure	σ_0	δ	a	b	c	Minloss
1(a)	1.0	0.6	0.4	0.3	0.2	26.560
1(b)	2.0	0.6	0.4	0.3	0.2	33.664
1(c)	1.0	0.9	0.1	0.7	0.1	14.111
1(d)	1.0	0.4	0.4	0.3	0.2	28.478
2(a)	1.0	0.1	0.4	0.3	0.2	64.328
2(b)	0.5	0.1	0.4	0.3	0.2	51.646
2(c)	0.8	0.6	0.7	0.1	0.1	41.002
2(d)	0.3	0.4	0.1	0.2	0.5	44.903



(a)

(b)

(c)

(d)

Figure 1: The x-axis and y-axis represent the variable x_1 and x_2, respectively. The values of parameters and minimum loss (minloss) for optimal locations are reported in Table 1. The isotropic Gaussian correlation structure was assumed in the construction of optimal locations.

4. Applications

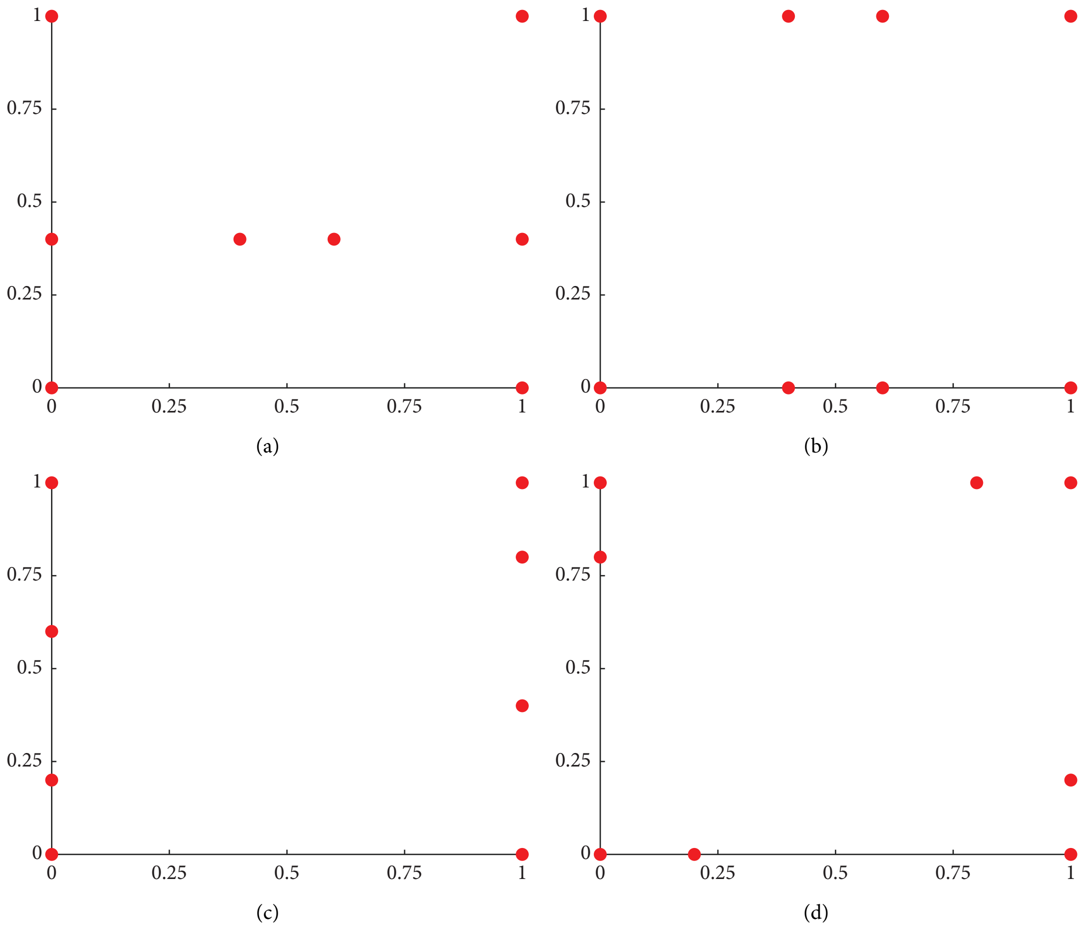
In this section, we discuss how to choose optimal locations from permanent locations. The procedure is described in Algorithm 2 and this algorithm is explained in §4.1. In §4.2, we apply Algorithm 1 to the "coal-ash" data.

4.1. Application 1

$$j_n = \underset{j\in\mathcal{J}^*-\mathcal{J}^*_{n-1}}{\operatorname{argmin}} \left\|\mathbf{x}_{i_n} - \mathbf{x}^*_j\right\|, \qquad (16)$$

where the empty set $\mathcal{J}^*_0 = \varnothing$ and $\mathcal{J}^*_n = \{j_1, j_2, \ldots, j_n\}$. Thus, the set $\mathcal{S}^*_{n_0} = \left\{\mathbf{x}^*_{j_1}, \mathbf{x}^*_{j_2}, \ldots, \mathbf{x}^*_{j_{n_0}}\right\}$ contains the chosen optimal permanent locations.

We simulated $r_0 = 90$ permanent locations in a square that has vertices (0, 0), (0, 1), (1, 0), and (1, 1). These permanent locations are displayed in Figure 3(a). Algorithm 2 was applied to choose $n_0 = 11$ optimal locations from these 90 permanent locations. Equispaced locations were generated with size $N = 121$. These locations are shown in Figure 3(b). The isotropic Gaussian correlation structure was



(a)

(b)

(c)

(d)

FIGURE 2: The x-axis and y-axis represent the variable x_1 and x_2, respectively. The values of parameters and the minimum loss (minloss) of an optimal design are reported in Table 1. The anisotropic Gaussian correlation structure was assumed in the construction of optimal locations.

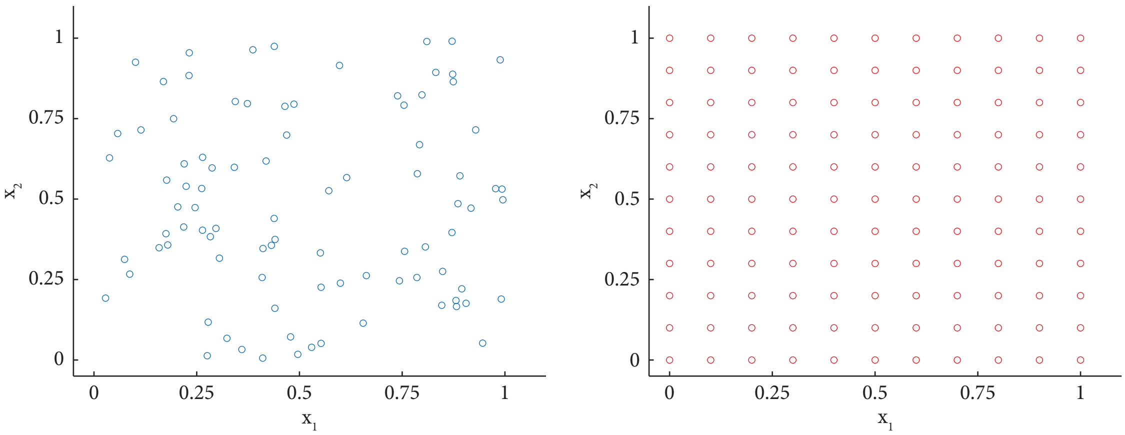
```
Step 1: Collect $n = n_1$ locations randomly without replacement from the design space $\Omega$ and let $\mathcal{S}_n = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_n}\}$ be the collected locations and $\mathcal{I}_n = \{i_1, i_2, \ldots, i_n\}$ be the corresponding index set, where $i_j \in \mathcal{I}$.
Step 2: Sequentially select $n = n_1 + 1, n_1 + 2, \ldots, n_0$ location such that
$i_n = \underset{j \in \mathcal{I} - \mathcal{I}_{n-1}}{\operatorname{argmin}} \mathcal{L}(\xi_{n,j})$
where $\xi_{n,j} \overset{def}{=} ((n-1)\xi_{n-1} + (0, \ldots, 0, \overset{\downarrow j}{1} \overset{j}{\downarrow}, 0, \ldots, 0)'/n)$. Thus, we have $\mathcal{S}_{n_0} = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_{n_0}}\}$ the chosen locations and corresponding index set $\mathcal{I}_{n_0}$.
Step 3: Remove the initial locations $\mathcal{S}_{n_1}$ from the set $\mathcal{S}_{n_0}$. So, the collected locations are $\mathcal{S}_{n_0} - \mathcal{S}_{n_1} = \{\mathbf{x}_{i_{n_1+1}}, \mathbf{x}_{\mathrm{i}_{n_1+2}}, \ldots, \mathbf{x}_{i_{n_0}}\}$ and the corresponding index set is $\mathcal{I}_n = \{i_{n_1+1}, i_{n_1+2}, \ldots, i_{n_0}\}$ with $n = n_0 - n_1$. Also, we have $\xi_n = \xi_{n_0} - \xi_{n_1}$.
Step 4: Again sequentially choose $n^* = 1, 2, \ldots, n_1$ location such that
$i_{n^*} = \underset{j \in \mathcal{I} - \mathcal{I}_{n-1}}{\operatorname{argmin}} \mathcal{L}(\xi_{n,j})$, where $n = n_0 - n_1 + n^*$,
$\mathcal{I}_n = \{i_{n_1+1}, i_{n_1+2}, \ldots, i_{n_0}, i_1, \ldots, i_{n^*}\}$, $\xi_{n,j} \overset{def}{=} ((n-1)\xi_{n-1} + (0, \ldots, 0, \overset{\downarrow j}{1} \overset{j}{\downarrow}, 0, \ldots, 0)'/n)$.
Finally, the set $\mathcal{S}_{n_0} = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_{n_0}}\}$ contains the selected optimal locations.
```

ALGORITHM 1: Let n_0 be the required number of optimal locations and $\mathcal{I} = \{1, 2, \ldots, N\}$ be the index set. The set $A - B$ denotes the set difference of A and B. We follow steps 1–4 to get n_0 optimal locations.

assumed in the construction of 11 grid-based optimal locations. An initial design is required to run Algorithm 2. The number of initial locations $n_1 = 6$ was used to run Algorithm 2. Although initial locations were removed and new locations were chosen instead of initial locations at the end of Algorithm 2, the choice of the final locations slightly depends on the initial locations. Thus, we considered 100 runs using Algorithm 2 to obtain the grid-based optimal locations.

```
Step 1: Let $N = (\lfloor \sqrt{r_0} \rceil + 1)^2$, where $\lfloor \bullet \rceil$ is the nearest integer function.
Step 2: Identify a rectangle that includes all permanent locations.
Step 3: Generate equispaced locations with size $N$ in the rectangle. We assume that the generated design space is $\Omega = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$,
        where $\mathbf{x}_i$ contains information that is related to $i$ th generated location for $i = 1, 2, \dots, N$.
Step 4: Choose optimal locations with size $n_0$ using Algorithm 1 from $\Omega$. Let $\mathcal{S}_{n_0} = \left\{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_{n_0}}\right\}$ be the selected optimal locations.
Step 5: Sequentially pick $n = 1, 2, \dots, n_0$ location such that
```

Algorithm 2: We suppose that r_0 permanent locations are in a two dimensional space, for instance, water wells in a region. The experimenter is interested in choosing n_0 optimal locations from these permanent locations. Let $\Omega^* = \left\{\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_{r_0}^*\right\}$ be a design space, where $\mathbf{x}_j^*$ contains information about j th permanent location for $j = 1, 2, \dots, r_0$. Define the index set $\mathcal{I}^* = \{1, 2, \dots, r_0\}$.



(a)

(b)

Figure 3: (a) Simulated permanent locations and (b) generated equispaced locations are displayed. The variables x_1 and x_2 are defined in §3.

Figure 4(b) shows the losses of 100 runs, and the minimum loss for the grid-based optimal locations was 133.316 and it occurred in the 32th run. So, we finally chose 11 locations that were generated in the 32th run. These selected grid-based locations are shown in Figure 4(a). The cluster of permanent locations that was nearest to the cluster of grid-based optimal locations was picked as the permanent optimal locations. These permanent optimal locations are displayed in Figure 4(a). In fact, Algorithm 2 can be used to identify optimal locations for an image, for instance, X-rays, a large number of water wells in a region, or a soil test for a given area.

4.2. Application 2. In this section, we study 'coal-ash' data to investigate the performance of the discussed method. The coal-ash core measurements were collected from 208 locations in the Pittsburgh coal seam. These locations are with an approximately 2500 feet equispaced distance [12]. Wiens [10] applied his developed method to choose optimal locations for the 'coal-ash' study. The values of the parameters $\sigma_\varepsilon^2, \sigma_0^2$, and λ are essential to constructing optimal locations for this study. The previous study results are a solution to overcome this problem [13]. We used the information on the final optimal locations with size 30 of Wiens [10] to obtain the values for these parameters. The coal-ash core measurement 17.61 was an outlier at location (5, 6) in this information. Generalized least squares estimate performs poorly if there is an outlier in a data set [10]. However, although a data set contains outliers, M-estimators are robust and efficient [14]. Thus, we preferred M-estimate in this application. These M-estimate are $(\sigma_\varepsilon^2, \sigma_0^2, \lambda) = (0.94, 0.77, 0.027)$.

The performance of the constructed optimal locations was evaluated by the root mean squared error (RMSE) and it is defined by

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)^2}{N}}, \qquad (17)$$

$$\text{where } \widehat{Y}_i = \widehat{\boldsymbol{\theta}}' \mathbf{x}_i,$$

where $\widehat{\theta} = (\widehat{\theta}_0, \widehat{\theta}_1, \widehat{\theta}_2)$ are the M-estimates of the unknown true parameters $\theta_0 = (\theta_{00}, \theta_{10}, \theta_{20})$.

Data collection from a small number of locations yields saving expenditure, reduction of time for an experiment, and fast statistical analysis. Thus, the small number of locations, n_0, were considered to verify the performance of our proposed method. The various sizes of $n_0 = 10, 20, 30, 50$ were taken to compare information obtained from optimal locations with full locations having size $n_0 = 208$. These results were reported in Table 2. The value of RMSE for the full locations is 1.1220. The maximum difference between RMSE for optimal locations and full locations is 0.0847. Meanwhile, the minimum difference between RMSE for optimal locations and full locations is 0.0082. Therefore, RMSE for full locations is

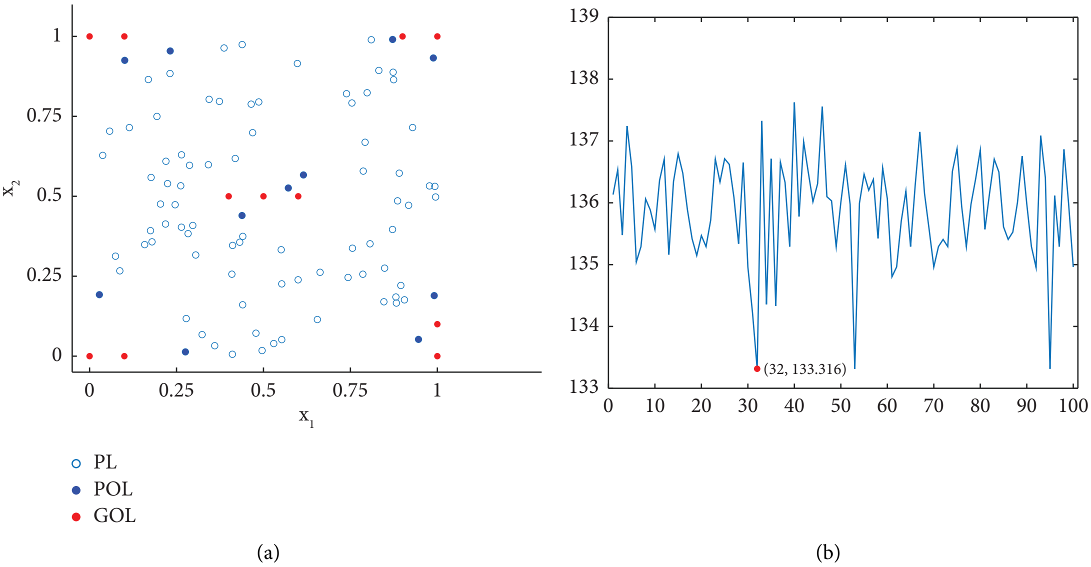


FIGURE 4: (a) Generated permanent locations (PL), selected permanent optimal locations (POL), grid-based optimal locations (GOL), and (b) losses of 100 sequential runs for 11grid-based locations are displayed for the values of parameters $a = 0.4, b = 0.3, c = 0.2, \delta = 0.6$. In subplot (b), the x-axis and y-axis represent minimum loss and run, respectively. The variables x_1 and x_2 are defined in §3.

TABLE 2: M-estimates of the model parameters, standard errors in parentheses, and RMSE for the coal-ash study.

n_0	$\widehat{\theta}_0$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	RMSE
10	10.1990 (0.8765)	−0.1823 (0.0838)	0.0773 (0.0511)	1.2067
20	11.0270 (0.6505)	−0.1776 (0.0660)	0.0277 (0.0406)	1.1302
30	10.6620 (0.5085)	−0.1593 (0.0520)	0.0272 (0.0316)	1.1368
50	10.6110 (0.3479)	−0.1428 (0.0359)	0.0105 (0.0219)	1.1431
208	11.0390 (0.1986)	−0.1781 (0.0222)	−0.0011 (0.0124)	1.1220

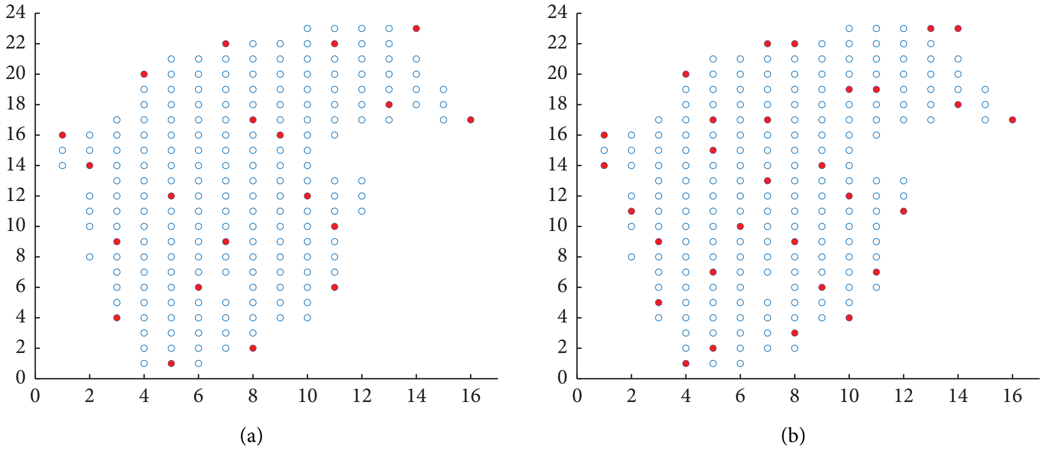


FIGURE 5: Continued.

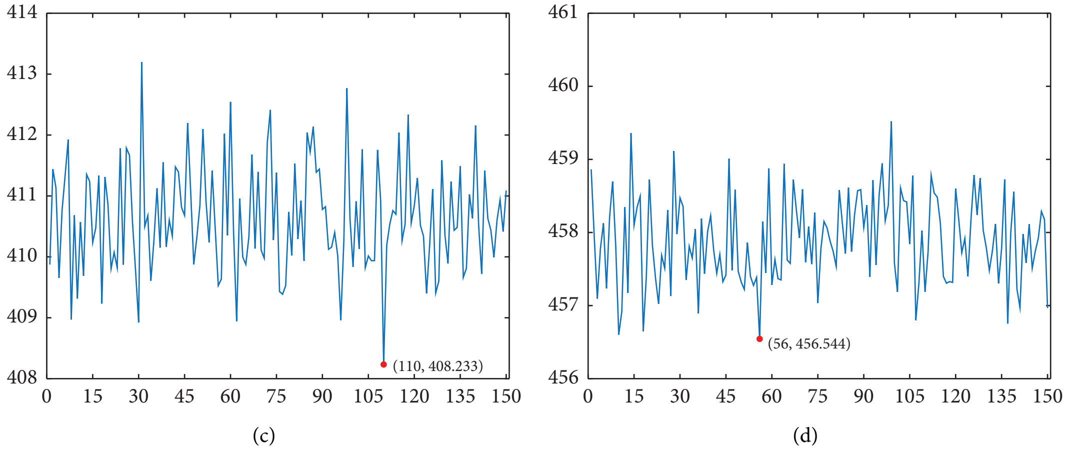


Figure 5: In subplots (a) and (b), the x-axis and y-axis indicate the position of a location in the east-west and north-south directions respectively. (a) 20 optimal locations; (b) 30 optimal locations; (c) losses of 150 runs for 20 optimal locations; and (d) losses of 150 runs for 30 optimal locations for the coal-ash study. In subplot (c) and (d), the x-axis and y-axis represent minimum loss and run, respectively.

Table 3: M-estimates of the model parameters; standard errors are in parentheses and RMSE for $n_0 = 30$, $\delta = 0.79$ and various combinations of a, b, and c for the coal-ash study.

Scenarios	σ_ε^2	a	b	c	$\widehat{\theta}_0$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	RMSE
S1	0.94	0.0	0.2	0.3	10.84 (0.4725)	−0.2347 (0.0492)	0.0593 (0.0289)	1.1484
S2	112.34	0.0	0.005	0.9	9.869 (0.3556)	−0.1438 (0.0363)	0.0652 (0.0215)	1.2743
S3	0.94	0.3	0.2	0.3	10.662 (0.5085)	−0.1593 (0.0520)	0.0272 (0.0316)	1.1368
S4	0.0007	0	0.92	0.001	10.482 (0.4923)	−0.1321 (0.0524)	0.0173 (0.0318)	1.1528
S5	12.482	0	0.01	0.2	10.15 (0.4014)	−0.1374 (0.0390)	0.0534 (0.0243)	1.1950

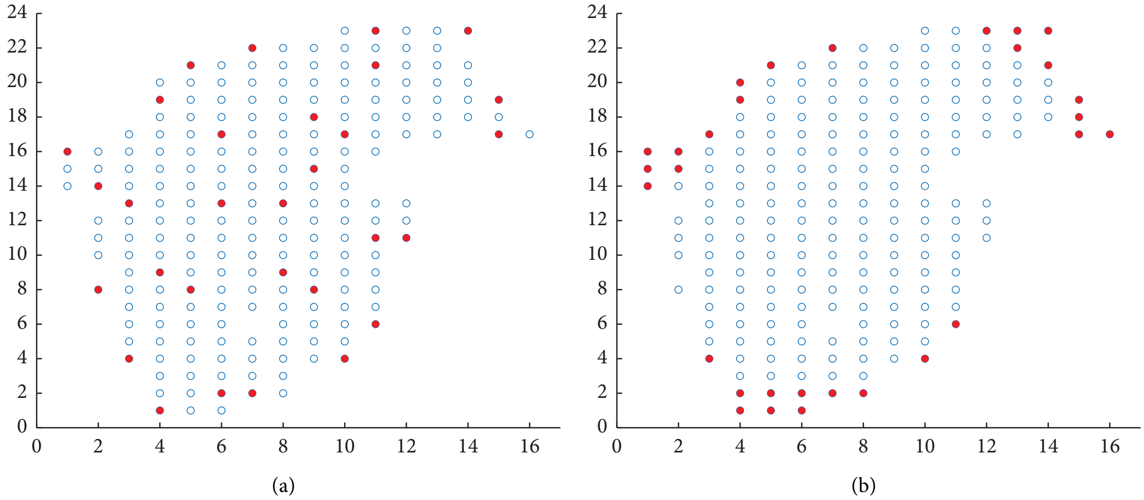


Figure 6: Continued.

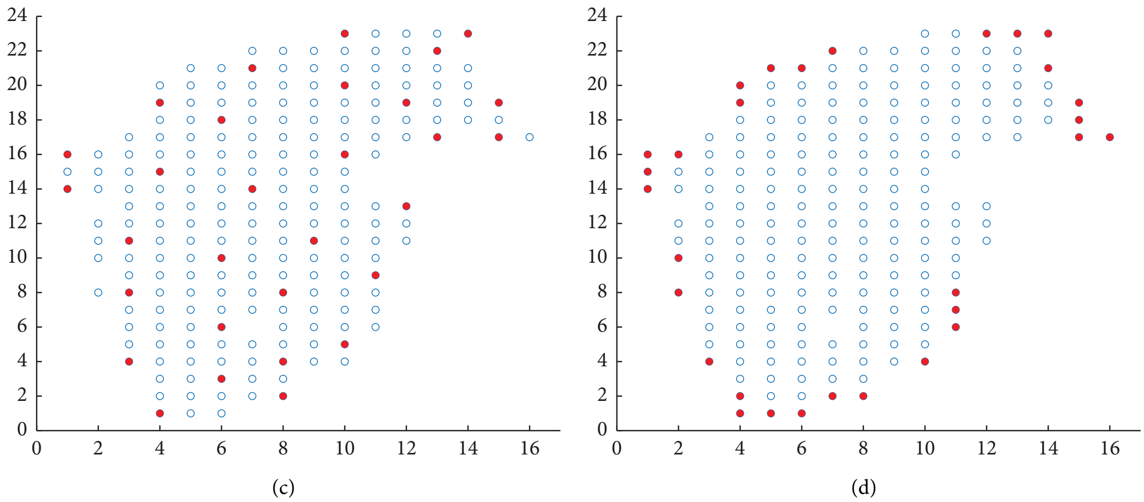


Figure 6: The x-axis and y-axis indicate the position of a location in the east-west and north-south directions respectively. 30 optimal locations are displayed in (a)–(d) for scenario S1, S2, S3, and S5, respectively. These scenarios are described in Table 3.

approximately equal to RMSE for optimal locations. That is, the information obtained from optimal locations is approximately the same as information obtained from full locations. Therefore, when we conduct an experiment in the optimal locations, expenditure can be reduced without losing information. Optimal locations having size $n_0 = 20$ and $n_0 = 30$ are displayed in Figure 5.

We selected 5 sets of parameters a, b, and c to observe patterns and test the effectiveness of optimal locations. These sets of parameters are reported in Table 3. σ_ε^2 is the variance of a homoscedastic measurement error $\varepsilon(\mathbf{x}_i)$ and it depends on the parameters b, c and δ. Therefore, we computed the values of σ_ε^2 and these values are in Table 3. Also, the value of σ_ε^2 $(= 0.94)$ was taken from the paper of Wiens [10] for scenario 1 (S1) and that value was computed using the final 30 optimal locations. We used $b = 0.2$ and $c = 0.3$ for S1. The value of δ can be calculated by the formula $\sigma_\varepsilon \sqrt{b/c}$ and the calculated value of δ was 0.79. We used this value for all scenarios in Table 3.

The optimal locations are condensed in the border of the target region (see Figures 6(b) and 6(d)) to the large values of σ_ε^2. Meanwhile, the optimal locations are scattered in the target region (see Figures 6(a) and 6(c)) to the small values of σ_ε^2. Also, the values of RMSE for the optimal locations are faraway from the value of RMSE for full locations when we assume a large value of σ_ε^2. In contrast, the values of RMSE for the optimal locations are approximately the same as the value of RMSE for full locations when we use a small value of σ_ε^2.

5. Summary and Conclusion

We have discussed the robust method to construct optimal locations for spatial data analysis. The design constructions are robust against model misspecifications regarding regression responses and variance/covariance structures of responses. The prediction mean squared error was considered to form the loss function. The loss function was obtained by maximizing the misspecified regression function and variance/covariance matrix of responses. Algorithm 1 was developed using the sequential method to choose optimal locations from equispaced locations. However, Algorithm 2 works for the nonequispaced locations. Therefore, Algorithm 2 can be used to choose optimal permanent locations from a two-dimensional space. The proposed approach can be used to answer a scientific question through an effective spatial analysis that includes minimum cost and time. Thus, the proposed sequential method can be applied to choose optimal locations from the Earth for water and soil monitoring, X-rays for diagnosing a disease, and a region for business analytics. The brute-force search only works If the number of subsample locations is less than or equal to 10^8. So, further research regarding the brute-force search should be done for any number of subsample locations. However, the proposed sequential method can be applied to select the optimal location from a large number of locations. We can reduce measurement error in data collection when we focus on a small number of optimal locations. Also, an efficient spatial data analysis can be done with optimal locations without losing any information. Optimal locations can be collected regardless of the shape of a region using the proposed method. Also, the proposed method is a way to conduct big data analytics as fast and efficiently as possible. However, it should be verified through future research for image analysis.

Data Availability

The data that was used in Application 2 can be found in Cressie (2015).

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work has been benefited by the comments of Douglas P. Wiens, University of Alberta. The processing fee for the manuscript was granted by the Governing Council of the University of Toronto.

References

[1] D. Sharma, C. Willy, and J. Bischoff, "Optimal subset selection for causal inference using machine learning ensembles and particle swarm optimization," *Complex & Intelligent Systems*, vol. 7, no. 1, pp. 41–59, 2021.

[2] L. Jaworski, S. Jansen, W. P. Pfitzner et al., "Comparative analysis of subsampling methods for large mosquito samples," *Parasites & Vectors*, vol. 12, no. 1, p. 354, 2019.

[3] D. M. Rocke and J. Dai, "Sampling and subsampling for cluster analysis in data mining: with applications to sky survey data," *Data Mining and Knowledge Discovery*, vol. 7, no. 2, pp. 215–232, 2003.

[4] S. Salloum, J. Z. Huang, and Y. He, "Exploring and cleaning big data with random sample data blocks," *Journal of Big Data*, vol. 6, no. 1, p. 45, 2019.

[5] H. Wang, R. Zhu, and P. Ma, "Optimal subsampling for large sample logistic regression," *Journal of the American Statistical Association*, vol. 113, no. 522, pp. 829–844, 2018.

[6] Y. Yao and H. Wang, "A Review on optimal subsampling methods for massive datasets," *Journal of Data Science*, vol. 19, pp. 151–172, 2021.

[7] S. Janardhanan, D. Gladish, D. Gonzalez, D. Pagendam, T. Pickett, and T. Cui, "Optimal design and prediction-independent verification of groundwater monitoring network," *Water*, vol. 12, no. 1, p. 123, 2019.

[8] P. Shi, J. Ye, and J. Zhou, "Discrete minimax designs for regression models with autocorrelated MA errors," *Journal of Statistical Planning and Inference*, vol. 137, no. 8, pp. 2721–2731, 2007.

[9] D. P. Wiens, "Robustness in spatial studies I: minimax prediction," *Environmetrics*, vol. 16, no. 2, pp. 191–203, 2005.

[10] D. P. Wiens, "Robustness in spatial studies II: minimax design," *Environmetrics*, vol. 16, no. 2, pp. 205–217, 2005.

[11] D. P. Wiens, "Minimax prediction designs, robust against misspecified response and error structures," 2019, https://sites.ualberta.ca/%7Edwiens/home%20page/publist.htm.

[12] N. Cressie, *Statistics for Spatial Data*, Wiley Series in Probability and Statistics, Washington, DC, USA, 2015.

[13] M. R. Lange and H. Schmidli, "Optimal design of clinical trials with biologics using dose-time-response models," *Statistics in Medicine*, vol. 33, no. 30, pp. 5249–5264, 2014.

[14] D. Q. F. de Menezes, D. M. Prata, A. R. Secchi, and J. C. Pinto, "A review on robust *M*-estimators for regression analysis," *Computers & Chemical Engineering*, vol. 147, Article ID 107254, 2021.