

OPTIMIZATION OF GROUNDWATER MONITORING NETWORK BY COMBINING THE IDW, K-MEANS CLUSTERING AND GEOSTATISTICAL KRIGING METHODS

Parvin Fathipour, Jafar Ahmadi Shali*

Faculty of Mathematics, Statistics and Computer Science, University of Tabriz, Tabriz, Iran

ABSTRACT

The optimization of groundwater monitoring networks is critical for contamination detection, yet traditional geostatistical methods become inapplicable when the number of existing boreholes is insufficient for reliable variogram modeling. This paper presents a hybrid methodology that integrates Inverse Distance Weighting (IDW) interpolation, K-means clustering, and ordinary kriging to optimize monitoring networks in data-scarce environments. The proposed framework uses IDW to generate initial spatial estimates, K-means to reduce computational complexity, and kriging with an acceptance probability criterion to evaluate network performance. Applied to the Bandar Abbas Oil Refinery in Iran, where only 20 boreholes exist, the method discretizes the study area into 611 grid nodes and partitions them into 20 clusters. An exhaustive search over 15 cluster centers identifies the optimal network configuration, followed by a sequential greedy algorithm that prioritizes individual grid nodes for further augmentation. Results demonstrate that adding 13 strategically selected boreholes achieves nearly 100% coverage with acceptable accuracy ($\alpha=0.9$). Sensitivity analysis reveals that increasing the threshold to $\alpha=0.95$ requires 27 boreholes, illustrating the trade-off between accuracy and cost. The methodology provides a flexible, computationally feasible framework that generates a priority list for network expansion under budget constraints, offering a practical decision-support tool for environmental monitoring in industrial settings.

Keywords: groundwater monitoring; geostatistical optimization; ordinary kriging; K-means clustering; acceptance probability; sparse network design.

Date submitted: 25.06.2026

Date accepted: 27.08.2026

© 2026 «OilGasScientificResearchProject» Institute. All rights reserved.

1. Introduction and preliminaries

Groundwater resources are vital for meeting global water demands and sustaining ecosystems, yet they are continuously changing in both quantity and quality. Pollution of shallow aquifers is often unavoidable in certain industrial operations, particularly in oil refineries, petrochemical complexes, and chemical manufacturing plants. When shallow groundwater is present beneath such facilities, contamination is likely to occur over time. To protect public health and maintain environmental safety, it is essential to detect polluted groundwater before the contamination plume extends beyond the facility boundary so that it can be contained or remediated. Consequently, many refineries and industrial sites establish groundwater monitoring networks along their perimeters to detect contaminant plumes and to assess spatial-temporal variations in groundwater quality. A groundwater monitoring network typically consists of a series of observation boreholes drilled within the aquifer system [1]. The design and optimization of such networks are a complex and critical task,

as monitoring wells are costly to install, sample, and maintain. Numerous methods have been developed to enhance network efficiency and to ensure reliable characterization of groundwater conditions. These approaches are generally grouped into statistical, entropy-based, and physiographic methods [2]. Statistical approaches include (i) variance-reduction techniques, (ii) dimension-reduction methods such as clustering or principal component analysis (PCA), and (iii) geostatistical interpolation techniques such as kriging [2]. Geostatistical methods rely on the spatial correlation among measurement points and employ the variogram as a quantitative measure of spatial dependence to evaluate prediction uncertainty [3].

The theoretical foundations of kriging and spatial statistics are comprehensively presented in the seminal works of Cressie [4] and Goovaerts [3], while ordinary kriging was first introduced by Matheron [5]. Practical implementations of these methods in the R software environment have been provided by Bivand et al [6], facilitating their widespread adoption among researchers. The theoretical foundations of Principal Component Analysis are thoroughly presented in [7], which serves as the primary reference for the dimension-

*E-mail: j-ahmadishali@tabrizu.ac.ir

<http://dx.doi.org/10.5510/OGP20260301239>

ality reduction step employed in this study, where pH, electrical conductivity, and soil rate are combined into a single composite variable.

The K-means clustering algorithm, one of the most widely used unsupervised machine learning methods for spatial partitioning, is comprehensively reviewed by Sinaga and Yang [8], whose work informs the clustering strategy adopted in the present methodology. The use of kriging and estimation error variance for monitoring network design was pioneered by Prakash and Singh [9], with subsequent refinements and validation through cross-validation approaches by Theodossiou and Latinopoulos [10] and Triki et al. [11]. The integration of PCA with geostatistical methods was explored by Nabi et al. [12], while Datta and Singh [13] combined kriging with simulated annealing for optimal network design in multi-source contamination scenarios.

The concept of acceptance probability, which serves as a cornerstone of the present methodology, was originally introduced by Cheng et al. [14] for raingauge network evaluation and augmentation. The optimal selection of groundwater-level monitoring sites using spatial stratification and temporal stability analysis was investigated by Ran et al. [15], who demonstrated the importance of incorporating temporal dynamics into monitoring network design. Recent advances have expanded the methodological toolkit considerably. Machine learning algorithms [16], low-cost sensors and Internet of Things (IoT) technologies [17], complex network analysis [18], evolutionary algorithms such as differential evolution [19], entropy-based and correlation criteria [20, 21], and data-driven sparse sensing approaches [22] have all been successfully applied to network design and optimization.

Risk-based optimization of groundwater quality monitoring networks has also been addressed by Wu et al. [23], who incorporated risk assessment into the geostatistical framework. In the context of petroleum industry applications, machine learning techniques for well placement and evaluation have been investigated by Ibrayev and Ibrahim [24], while Mukhametshin et al. [25] examined geological and statistical modeling of oil recovery processes, underscoring the importance of optimization in energy-sector operations. The topology design and optimization of utility networks with unreliable elements has been explored by Toktoshov et al. [26].

Additionally, a wide range of mathematical and numerical methods, including differential equations, optimal control theory, functional analysis, and boundary value problems, have been developed by researchers [27- 33]. These methodological developments provide robust mathematical tools that underpin the analytical framework of the present study and support the rigorous implementation of the proposed optimization methodology.

The assessment of groundwater quality and the detection of contaminant plumes rely critically on the spatial configuration of monitoring boreholes. A poorly designed monitoring network may fail to detect contamination events, leading to delayed remediation efforts and potential environmental hazards. Conversely, an overly dense network incurs unnecessary installation, sampling, and maintenance costs. The central challenge in monitoring network design is therefore to determine the minimum number and optimal spatial distribution of boreholes required to achieve a specified level of prediction accuracy across the study area.

In this section, we discuss the quantification of prediction accuracy in terms of a probabilistic criterion known as the acceptance probability. The framework is built upon the theory of geostatistics, particularly ordinary kriging, and provides a formal basis for the optimization methodology presented in subsequent sections. We begin with the fundamental definitions of spatial random fields, stationarity, and variograms, then develop the theory of kriging and its uncertainty quantification, and finally derive the acceptance probability criterion with complete mathematical rigor.

Let $D \subset \mathbb{R}^2$ denote the spatial domain of interest, which in our application is the Bandar Abbas oil refinery area. We model the groundwater quality parameter (e.g., the first principal component of pH, electrical conductivity, and soil rate) as a realization of a spatial random field.

Definition 1.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. A spatial random field is a collection of random variables $\{Z(s) : s \in D\}$, where for each fixed $s \in D$, $Z(s) : \Omega \rightarrow \mathbb{R}$ is a measurable function. For each outcome $\omega \in \Omega$, the mapping $s \mapsto Z(s, \omega)$ is called a realization of the random field.

In practice, we observe the random field only at a finite set of locations $\mathcal{S}_n = \{s_1, s_2, \dots, s_n\} \subset D$, yielding the data vector $\mathbf{Z} = (Z(s_1), \dots, Z(s_n))$. Our goal is to predict $Z(s_0)$ at an unobserved location $s_0 \in D \setminus \mathcal{S}_n$.

Definition 1.2. The random field $\{Z(s) : s \in D\}$ is said to be second-order stationary if the following two conditions hold:

1. *Constant mean:* There exists $\mu \in \mathbb{R}$ such that for all $s \in D$,

$$\mathbb{E}[Z(s)] = \mu$$

2. *Translation-invariant covariance:* There exists a function $C : \mathbb{R}^2 \rightarrow \mathbb{R}$ such that for all $s, s+h \in D$,

$$\text{Cov}(Z(s), Z(s+h)) = C(h)$$

where $C(h)$ depends only on the lag vector h , not on the absolute location s .

The function $C(h)$ is called the covariance function. The value at zero lag, $C(0) = \text{Var}(Z(s))$, is known as the sill and represents the overall variance of the random field.

Second-order stationarity is a fundamental assumption in classical geostatistics. It implies that the statistical properties of the random field are invariant under spatial translation, which is a reasonable approximation for many environmental variables over moderate spatial scales.

Definition 1.3. Under second-order stationarity, the variogram $\gamma : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ is defined as half the expected squared difference between the random field at two locations separated by lag h :

$$\gamma(h) = \frac{1}{2} \mathbb{E} \left[(Z(s+h) - Z(s))^2 \right] \quad (1)$$

For a second-order stationary process with covariance function $C(h)$, the variogram is related to the covariance by:

$$\gamma(h) = C(0) - C(h) \quad (2)$$

This relationship follows from expanding the square:

$$\begin{aligned} & \frac{1}{2} \mathbb{E} \left[(Z(s+h) - Z(s))^2 \right] = \\ & = \frac{1}{2} \left[\text{Var}(Z(s+h)) + \text{Var}(Z(s)) - 2\text{Cov}(Z(s+h), Z(s)) \right] = \\ & = C(0) - C(h) \end{aligned}$$

The variogram is a fundamental tool in geostatistics because it quantifies the spatial dependence structure of the

random field. Typically, $\gamma(h)$ increases with the magnitude of h , reflecting the fact that locations farther apart tend to be less correlated and therefore exhibit larger expected squared differences.

Definition 1.4. Given the observed data $Z = (Z(s_1), \dots, Z(s_n))^T$, the ordinary kriging predictor at an unobserved location $s_0 \in D$ is defined as:

$$\hat{Z}_{OK}(s_0) = \sum_{i=1}^n \lambda_i Z(s_i) \quad (3)$$

where the weight vector $\lambda = (\lambda_1, \dots, \lambda_n)^T \in \mathbb{R}^n$ is chosen to satisfy the following two conditions:

1. *Unbiasedness:* $\mathbb{E}[\hat{Z}_{OK}(s_0) - Z(s_0)] = 0$,
2. *Optimality:* $\text{Var}(\hat{Z}_{OK}(s_0) - Z(s_0))$ is minimized.

For the ordinary kriging predictor to be unbiased, the weights must satisfy the constraint:

$$\sum_{i=1}^n \lambda_i = 1 \quad (4)$$

This condition is both necessary and sufficient for unbiasedness, as it ensures that $\mathbb{E}[\hat{Z}_{OK}(s_0)] = \mu$ under the constant-mean assumption. A detailed derivation of this constraint can be found in standard geostatistical textbooks [3,4].

The ordinary kriging weights are obtained by minimizing the prediction variance subject to the unbiasedness constraint $1_n^T \lambda = 1$. Introducing a Lagrange multiplier $\nu \in \mathbb{R}$ to enforce this constraint, the optimal weight vector λ is characterized by the following linear system, known as the ordinary kriging system:

$$\begin{bmatrix} \Gamma & 1_n \\ 1_n^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \nu \end{bmatrix} = \begin{bmatrix} \gamma_0 \\ 1 \end{bmatrix} \quad (5)$$

where $\Gamma = [\gamma(s_i - s_j)]_{i,j=1}^n$ is the $n \times n$ matrix of variogram values between observation locations, $\gamma_0 = (\gamma(s_1 - s_0), \dots, \gamma(s_n - s_0))^T$ is the $n \times 1$ vector of variogram values between observation locations and the target location s_0 and $1_n = (1, \dots, 1)^T$. Provided that the variogram matrix Γ is positive definite (which holds when no two observation locations coincide and a valid variogram model is used), this system has a unique solution. A detailed derivation of the kriging system, along with a discussion of its existence and uniqueness, can be found in standard geostatistical references [3-5].

Definition 1.5. The minimized prediction variance, known as the kriging variance, is denoted by $\sigma_k^2(s_0)$ and is given by:

$$\sigma_k^2(s_0) = \text{Var}(Z(s_0) - \hat{Z}_{OK}(s_0)) = \gamma_0^T \lambda + \nu \quad (6)$$

where λ and ν are the solution to the kriging system (5).

Solving the ordinary kriging system (5) for the weight vector λ and the Lagrange multiplier ν , the kriging variance can be expressed in the following closed form [3,4]:

$$\sigma_k^2(s_0) = \gamma_0^T \Gamma^{-1} \gamma_0 - \frac{(1_n^T \Gamma^{-1} \gamma_0 - 1)^2}{1_n^T \Gamma^{-1} 1_n} \quad (7)$$

1.1. The Gaussian assumption and uncertainty quantification

While the kriging variance provides a measure of prediction uncertainty in terms of mean squared error, it does not directly provide probabilistic statements about the magnitude of the prediction error. To obtain such statements, we must make an additional distributional assumption.

Definition 1.6 (Gaussian random field). A spatial random field $\{Z(s) : s \in D\}$ is called a Gaussian random field

(or Gaussian process) if for every finite set of locations $\{s_1, \dots, s_m\} \subset D$, the random vector $(Z(s_1), \dots, Z(s_m))^T$ follows a multivariate normal distribution:

$$\begin{pmatrix} Z(s_1) \\ \vdots \\ Z(s_m) \end{pmatrix} \sim \mathcal{N}_m(\mu 1_m, \Sigma)$$

where $\Sigma_{ij} = C(s_i - s_j) = C(0) - \gamma(s_i - s_j)$.

The Gaussian assumption is widely used in geostatistics because it is mathematically tractable and often provides a reasonable approximation for many environmental variables, particularly after suitable transformations (e.g., logarithmic or Box-Cox transformations).

Under the Gaussian random field assumption (Definition 1.6), the ordinary kriging predictor enjoys a particularly tractable distributional property. Since the prediction error $E(s_0) = Z(s_0) - \hat{Z}_{OK}(s_0)$ is a linear combination of the jointly normal variables $\{Z(s_0), Z(s_1), \dots, Z(s_n)\}$, it follows that $E(s_0)$ is itself normally distributed. Combining this fact with the unbiasedness of the kriging predictor and the definition of the kriging variance, we obtain

$$E(s_0) \sim \mathcal{N}(0, \sigma_k^2(s_0)) \quad (8)$$

and consequently

$$\frac{E(s_0)}{\sigma_k(s_0)} \sim \mathcal{N}(0, 1) \quad (9)$$

This fundamental result underpins the probabilistic interpretation of the kriging variance and provides the basis for the acceptance probability defined in the next subsection. A rigorous derivation of this normality property, along with its implications for uncertainty quantification, can be found in standard geostatistical references [3-5].

It is important to emphasize that this property relies critically on the joint Gaussianity of the random field. When the data deviate substantially from normality - for instance, in the presence of heavy tails, strong skewness, or multimodality - the prediction error may not be exactly normal, particularly if the number of observations n is small. In practice, diagnostic checks such as Q-Q plots, histograms, and formal goodness-of-fit tests (e.g., the Shapiro-Wilk test) should be performed to assess the plausibility of the Gaussian assumption before using the probabilistic intervals derived from (9). Practical guidelines for such diagnostics are discussed, for example, in Goovaerts [3].

1.2. Definition of acceptance probability

With the distribution of the prediction error established, we can now define the acceptance probability a probabilistic criterion that quantifies the likelihood that the prediction error falls within a specified tolerance range.

Definition 1.7. Let $k > 0$ be a positive constant (the tolerance multiplier). Let $\sigma_z = \sqrt{\text{Var}(Z(s))} = \sqrt{C(0)}$ be the population standard deviation of the random field. The acceptance probability at location $s_0 \in D$ is defined as:

$$p(s_0) = \mathbb{P}(|Z(s_0) - \hat{Z}_{OK}(s_0)| < k \sigma_z) \quad (10)$$

In words, $p(s_0)$ is the probability that the absolute prediction error at s_0 does not exceed k times the population standard deviation.

The choice of k reflects the decision-maker's tolerance for prediction errors. A value of $k=1$ implies that errors up to one standard deviation are acceptable, while $k=2$ doubles the

acceptable error margin. The parameter σ_z provides a natural scaling factor that is independent of the local sampling configuration.

Theorem 1.8. Under the Gaussian assumption, the acceptance probability defined in Definition 1.7 is given by:

$$p(s_0) = 2\Phi\left(\frac{k\sigma_z}{\sigma_k(s_0)}\right) - 1 \quad (11)$$

where $\Phi: \mathbb{R} \rightarrow (0,1)$ is the cumulative distribution function of the standard normal distribution:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \quad (12)$$

Proof. Starting from Definition 1.7, $p(s_0) = \mathbb{P}(|Z(s_0) - \hat{Z}_{OK}(s_0)| < k\sigma_z)$. Since $\sigma_k(s_0) > 0$, we divide the inequality inside the probability by $\sigma_k(s_0)$:

$$p(s_0) = \mathbb{P}\left(\frac{|Z(s_0) - \hat{Z}_{OK}(s_0)|}{\sigma_k(s_0)} < \frac{k\sigma_z}{\sigma_k(s_0)}\right)$$

As previously stated, the standardized prediction error follows a standard normal distribution. Let $N \sim \mathcal{N}(0,1)$. Then $p(s_0) = \mathbb{P}\left(|N| < \frac{k\sigma_z}{\sigma_k(s_0)}\right)$. Now, for any $a \geq 0$ we have $\mathbb{P}(|N| < a) = \mathbb{P}(-a < N < a) = \Phi(a) - \Phi(-a)$. Using the symmetry property of the standard normal CDF, $\Phi(-a) = 1 - \Phi(a)$, we obtain $\mathbb{P}(|N| < a) = \Phi(a) - (1 - \Phi(a)) = 2\Phi(a) - 1$. Substituting $a = \frac{k\sigma_z}{\sigma_k(s_0)}$ yields $p(s_0) = 2\Phi\left(\frac{k\sigma_z}{\sigma_k(s_0)}\right) - 1$. This completes the proof.

Corollary 1.9. The acceptance probability $p(s_0)$ is a monotonically decreasing function of the kriging standard deviation $\sigma_k(s_0)$. Specifically, for any two locations $s_0, s'_0 \in D$, $\sigma_k(s_0) \leq \sigma_k(s'_0)$ implies $p(s_0) \geq p(s'_0)$.

Proof. The function $f(x) = 2\Phi(k\sigma_z/x) - 1$ is differentiable for $x > 0$. Its derivative is:

$$f'(x) = 2\phi\left(\frac{k\sigma_z}{x}\right) \cdot \left(-\frac{k\sigma_z}{x^2}\right) < 0$$

where $\phi(\cdot)$ is the standard normal probability density function. Since for all $f'(x) < 0$ for all $x > 0$, f is strictly decreasing. Therefore, smaller kriging standard deviations yield larger acceptance probabilities.

1.3. The Acceptance criterion and its implications

The acceptance probability provides a continuous measure of prediction quality. In practice, however, decision-makers often require a binary classification: is the accuracy at a given location acceptable or not?

Definition 1.10. Given a threshold $\alpha \in (0,1)$ (typically $\alpha = 0.9$), a location $s_0 \in D$ is said to have acceptable accuracy if and only if:

$$p(s_0) \geq \alpha \quad (13)$$

Otherwise, the accuracy at s_0 is deemed unacceptable.

The threshold α should be chosen based on the specific requirements of the monitoring application. A higher α imposes stricter quality standards, while a lower α relaxes them.

Theorem 1.11. The condition is equivalent to:

$$\sigma_k(s_0) \leq \frac{k\sigma_z}{\Phi^{-1}\left(\frac{1+\alpha}{2}\right)} \quad (14)$$

where Φ^{-1} denotes the quantile function (inverse CDF) of the standard normal distribution.

Proof. Starting from $p(s_0) \geq \alpha$ we have $2\Phi\left(\frac{k\sigma_z}{\sigma_k(s_0)}\right) - 1 \geq \alpha$.

Adding 1 to both sides and dividing by 2 we can write

$\Phi\left(\frac{k\sigma_z}{\sigma_k(s_0)}\right) \geq \frac{1+\alpha}{2}$. Since Φ is strictly increasing, we can

apply its inverse $\frac{k\sigma_z}{\sigma_k(s_0)} \geq \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$. Since all quantities are positive, we invert the inequality $\sigma_k(s_0) \leq \frac{k\sigma_z}{\Phi^{-1}\left(\frac{1+\alpha}{2}\right)}$. This completes the proof.

1.4. The percentage of area with acceptable accuracy

For network design and evaluation, it is useful to summarize the overall performance of a monitoring network by the proportion of the study area that satisfies the acceptable accuracy criterion.

Definition 1.12 (Percentage of area with acceptable accuracy). Let $D \in \mathbb{R}^2$ be the study domain with total area $|D| = \int_D ds$. Define the indicator function:

$$I(s) = \begin{cases} 1, & \text{if } p(s) \geq \alpha, \\ 0, & \text{otherwise.} \end{cases}$$

The percentage of area with acceptable accuracy, denoted by A_p , is defined as:

$$A_p = \frac{1}{|D|} \int_D I(s) ds \times 100\% \quad (15)$$

In practice, the integral is approximated by discretizing the domain into a grid of nodes $\{g_1, \dots, g_m\}$. The discrete approximation is:

$$A_p \approx \frac{1}{m} \sum_{j=1}^m \mathbb{1}_{\{p(g_j) \geq \alpha\}} \times 100\% \quad (16)$$

The quantity A_p serves as the objective function for network optimization: we seek the configuration of boreholes that maximizes A_p , or equivalently, achieves a target A_p with the minimum number of boreholes.

1.5. Estimation of the population standard deviation

In practice, the population standard deviation $\sigma_z = \sqrt{C(0)}$ is unknown and must be estimated from the available data. A natural estimator is the sample standard deviation:

$$\hat{\sigma}_z = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Z(s_i) - \bar{Z})^2} \quad (17)$$

where $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z(s_i)$

This estimator is consistent under second-order stationarity as $n \rightarrow \infty$. However, when n is small (as in our case study, where $n=20$), the estimator has considerable variability, and the uncertainty in $\hat{\sigma}_z$ propagates into the acceptance probability calculation.

With only 20 boreholes available, we cannot fit a reliable variogram directly. So we first use IDW interpolation to generate estimates at all grid nodes, and then use the standard deviation of these estimates as a proxy for σ_z in calculating the acceptance probabilities. This is not a perfect solution, but it is a practical one when data are limited (see Section 2.2).

1.6. Interpretation and practical guidelines

The acceptance probability framework provides a systematic and quantitative approach to monitoring network design. The following practical guidelines can be derived from the mathematical analysis:

1. *Identify critical locations:* Locations with high $\sigma_k(s)$ (e.g., far from existing boreholes) typically have low acceptance probabilities and should be prioritized for new borehole placement.
2. *Balance cost and accuracy:* The parameters k and α allow decision-makers to calibrate the trade-off between monitoring cost and prediction accuracy. A smaller k or larger α requires more boreholes to achieve acceptable accuracy, increasing cost but improving reliability.
3. *Use sensitivity analysis:* Varying k and α in a sensitivity analysis can reveal the marginal benefit of adding boreholes and help identify the optimal network configuration under budget constraints.
4. *Consider the Gaussian assumption:* Diagnostic checks should be performed to verify the plausibility of the Gaussian assumption. If the data are severely non-Gaussian, transformation or alternative methods (e.g., trans-Gaussian kriging) may be necessary.

1.7. Mathematical critique and limitations

Despite its elegance, the acceptance probability framework has several limitations that warrant careful consideration.

1. *Dependence on the Gaussian assumption:* The closed-form expression (11) relies critically on the Gaussian assumption. If the random field is not Gaussian, the actual distribution of the prediction error may deviate substantially from normality, leading to incorrect acceptance probabilities. In such cases, the computed $p(s_0)$ should be interpreted as an approximation, and diagnostic checks should be performed.
2. *Estimation uncertainty:* The quantities $\sigma_k(s_0)$ and σ_z are estimated from the data, and their estimates are subject to uncertainty. Ignoring this uncertainty leads to overconfidence in the computed acceptance probabilities. Bayesian approaches that account for parameter uncertainty could provide a more realistic quantification.
3. *Stationarity assumption:* The framework assumes second-order stationarity. In many hydrogeological settings, spatial trends or non-stationary covariance structures may be present. Universal kriging or kriging with external drift may be more appropriate in such cases.
4. *Spatial discretization errors:* The discretization of the study domain into grid nodes introduces an approximation error in the computation of A_p . Finer grids reduce this error but increase computational cost.

2. Methodological framework: IDW interpolation, K-means clustering, and group-wise optimization

2.1. Introduction to the proposed methodology

The acceptance probability framework developed in Section 1 provides a rigorous criterion for evaluating the performance of a groundwater monitoring network. However, the direct application of this framework to the Bandar Abbas

refinery case study faces a fundamental challenge: the existing monitoring network consists of only $n=20$ boreholes, which is insufficient for fitting a reliable variogram model. Without a reliable variogram, ordinary kriging cannot be applied directly, and the acceptance probability cannot be computed.

To overcome this data scarcity problem, we propose a multi-stage methodology that combines three complementary approaches:

1. *Inverse distance weighting (IDW):* A deterministic interpolation method used to generate initial estimates of the groundwater quality parameter at all grid nodes. This provides a surrogate dataset for subsequent analysis.
2. *K-means clustering:* An unsupervised machine learning algorithm used to partition the large set of grid nodes into a smaller number of spatially representative clusters. This reduces the computational complexity of the optimization problem.
3. *Group-wise Kriging-based optimization:* A systematic search over combinations of cluster centers to identify the set of new borehole locations that maximizes the percentage of area with acceptable accuracy.

This section presents the mathematical formulation of each component, provides rigorous theoretical justifications, and describes the algorithmic implementation in detail.

2.2. Inverse distance weighting (IDW) interpolation

Inverse distance weighting is a deterministic interpolation method that estimates the value of a variable at an unobserved location as a weighted average of observed values at nearby locations, where the weights are inversely proportional to the distance from the target location.

Definition 2.1 (IDW interpolator). $\mathcal{S}_n = \{s_1, \dots, s_n\} \subset D$ be the set of observation locations with corresponding measurements $z(s_1), \dots, z(s_n)$. For a target location $s_0 \in D$, the IDW interpolator is defined as:

$$\hat{z}_{IDW}(s_0) = \frac{\sum_{i=1}^n w_i(s_0) z(s_i)}{\sum_{i=1}^n w_i(s_0)} \quad (18)$$

where the weights $w_i(s_0)$ are given by:

$$w_i(s_0) = \|s_i - s_0\|^{-p} \quad (19)$$

with $p > 0$ being the power parameter (typically $p=20$), and $\|\cdot\|$ denoting the Euclidean distance.

Remark 2.2. The IDW interpolator has several important properties:

1. *Exact interpolation:* For any observation location $s_j \in \mathcal{S}_n$, we have $\hat{z}_{IDW}(s_j) = z(s_j)$. This follows because $w_j(s_j) = \|s_j - s_j\|^{-p} = \infty$, while all other weights are finite, so the weight for s_j dominates.
2. *Convexity:* The interpolated value is a convex combination of the observed values, since the weights are positive and sum to one.
3. *Deterministic nature:* IDW does not provide a measure of prediction uncertainty, unlike kriging.

A fundamental property of the IDW interpolator is that the weight $W_i(s_0) = \|s_i - s_0\|^{-p}$ assigned to an observation location decreases monotonically with increasing Euclidean distance from the target point s_0 . Furthermore, the rate of this decay is governed by the power parameter p : larger values

of p result in a more rapid attenuation of influence with distance, so that only the nearest neighbors contribute substantially to the prediction. This behaviour follows directly from the distance-decay form of the weight function (see Shepard [34], for the original formulation, and Lu & Wong, [35], for a practical discussion of the role of p).

Under mild regularity conditions on the underlying spatial field, the IDW interpolator also enjoys a consistency property: as the observation network becomes increasingly dense—specifically, as the maximum distance from any target location to its nearest observation tends to zero—the IDW estimate converges pointwise to the true continuous field, provided that the field is Lipschitz continuous. This convergence result, along with a rigorous discussion of its assumptions and limitations, is documented in [34].

Remark 2.3. Although IDW is a useful tool for initial estimation, it has limitations. Most notably, it does not account for the spatial correlation structure of the data, and it does not provide any measure of uncertainty. Therefore, IDW should be viewed as a preliminary step rather than a definitive interpolation method. In the proposed methodology, IDW serves as a «bootstrap» to generate initial estimates across the domain, enabling the subsequent geostatistical analysis.

2.3. Principal component analysis for dimension reduction

In the Bandar Abbas refinery case study, multiple ground-water quality parameters are measured at each borehole, including pH, electrical conductivity (EC), and soil rate. To combine these variables into a single representative measure for spatial analysis, we employ Principal Component Analysis (PCA).

Definition 2.4 (Principal component analysis). Let $X = [x_1, \dots, x_n]^T \in \mathbb{R}^{n \times q}$ be the data matrix of q standardized variables measured at n locations. The sample covariance matrix is:

$$S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T \quad (20)$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is the sample mean vector. PCA seeks the linear transformation:

$$Y = XA \quad (21)$$

where $A \in \mathbb{R}^{q \times q}$ is an orthogonal matrix whose columns are the eigenvectors of S , ordered by decreasing eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q \geq 0$.

The first principal component defined as $y_1 = Xa_1$, where a_1 is the eigenvector corresponding to the largest eigenvalue λ_1 , of the sample correlation matrix, accounts for the following proportion of the total variance:

$$\frac{\lambda_1}{\sum_{j=1}^q \lambda_j} \times 100\% \quad (22)$$

More generally, the first r principal components collectively explain:

$$\frac{\sum_{j=1}^r \lambda_j}{\sum_{j=1}^q \lambda_j} \times 100\% \quad (23)$$

These expressions follow directly from the fact that the total variance in the standardized variables equals the trace of the correlation matrix, which is invariant under orthogonal transformations and equals the sum of the eigenvalues.

A comprehensive treatment of PCA, including the derivation of these variance-explained ratios and their use in dimensionality reduction, is given in [7].

Remark 2.5. In the Bandar Abbas case study, the first principal component (PC1) of pH, EC, and soil rate is used as the target variable $Z(s)$ for spatial analysis. The use of PC1 reduces the dimensionality from three variables to one while retaining the majority of the information content, as measured by the explained variance.

2.4. Spatial clustering using K-means algorithm

After IDW interpolation has generated estimates of $Z(s)$ at all $m=611$ grid nodes, the next step is to reduce the computational complexity of the optimization problem. Instead of considering all m grid nodes as potential new borehole locations, we partition the nodes into $m_1=20$ spatially compact clusters using the K-means algorithm.

Definition 2.6 (K-means clustering). Let $\mathcal{G} = \{g_1, \dots, g_m\} \subset \mathbb{R}^2$ be the set of grid node coordinates. The K-means algorithm partitions $\mathcal{G}$ into m_1 disjoint clusters $C_1, \dots, C_{m_1}$ by minimizing the within-cluster sum of squares:

$$\min_{C_1, \dots, C_{m_1}} \sum_{k=1}^{m_1} \sum_{g \in C_k} \|g - \mu_k\|^2 \quad (24)$$

where $\mu_k = \frac{1}{|C_k|} \sum_{g \in C_k} g$ is the centroid (mean) of cluster C_k .

The Kmeans algorithm proceeds by iteratively alternating between two steps: (i) assigning each grid node to the nearest cluster centroid, and (ii) recomputing the centroids as the means of the newly formed clusters. A standard result for this procedure is that it converges monotonically to a local minimum of the withincluster sum of squares objective function (24). The convergence is guaranteed in a finite number of iterations: the assignment step does not increase the objective, and the update step minimizes it for the fixed cluster assignments, so the objective is nonincreasing at every iteration. Since the number of possible partitions of a finite set of $m_1=611$ grid nodes is finite, the algorithm must terminate after a finite number of steps. This finitestep convergence property, along with a detailed discussion of its practical implications—such as sensitivity to initial centroids and the nature of the local minima obtained—is presented in comprehensive treatments of the Kmeans algorithm [8].

Remark 2.7. The number of clusters $m_1=20$ is chosen as a compromise between computational feasibility and spatial resolution. A smaller number of clusters reduces the combinatorial complexity of the subsequent optimization but may aggregate heterogeneous regions into a single representative point. A larger number of clusters preserves more spatial detail but increases the computational burden.

2.5. Group-wise selection of new borehole locations

With the $m_1=20$ cluster centers identified, we now face the combinatorial optimization problem of selecting $r=15$ of these centers as candidate locations for new boreholes, such that the percentage of area with acceptable accuracy A_p is maximized.

Definition 2.8 (Candidate set). Let $\mathcal{H} = \{h_1, \dots, h_{m_1}\} \subset D$ denote the set of cluster centroids. For any subset $\mathcal{T} \subseteq \mathcal{H}$ of size $|\mathcal{T}| = r$, define the augmented observation set:

$$\mathcal{S}_{aug}(\mathcal{T}) = \mathcal{S}_n \cup \mathcal{T} \quad (25)$$

where $\mathcal{S}_n$ is the set of existing boreholes. The values at the cluster centers are taken from the IDW interpolation: $Z(h) = \hat{z}_{IDW}(h)$ for $h \in \mathcal{T}$.

Definition 2.9 (Optimization objective). For each augmented observation set $\mathcal{S}_{aug}(\mathcal{T})$, we:

1. Fit an experimental variogram to the augmented data,
2. Perform ordinary kriging across all m grid nodes to obtain $\sigma_k(g_i)$ for each grid node $g_i \in \mathcal{G}$,
3. Compute the acceptance probability $p(g_i) = 2\Phi(k\hat{\sigma}_z / \sigma_k(g_i)) - 1$,
4. Compute the percentage of area with acceptable accuracy:

$$A_p(\mathcal{T}) = \frac{1}{m} \sum_{j=1}^m \mathbf{1}_{\{p(g_j) \geq \alpha\}} \times 100\%$$

The optimal subset is then:

$$\mathcal{T}^* = \arg \max_{\mathcal{T} \subseteq \mathcal{H}, |\mathcal{T}|=r} A_p(\mathcal{T}) \quad (26)$$

The number of possible subsets of size r from m_1 cluster centers is:

$$\binom{m_1}{r} = \frac{m_1!}{r!(m_1-r)!} \quad (27)$$

For the Bandar Abbas case study, with $m_1=20$ and $r=15$, we have:

$$\binom{20}{15} = \binom{20}{5} = \frac{20!}{5!15!} = 15504 \quad (28)$$

This is a manageable number for exhaustive enumeration.

Theorem 2.10. Since the set of possible subsets $\mathcal{T} \subseteq \mathcal{H}$ with $|\mathcal{T}|=r$ is finite and non-empty (provided $r \leq m_1$), the optimization problem (26) always admits at least one optimal solution.

Proof. The objective function $A_p(\mathcal{T})$ maps the finite set $\{\mathcal{T} \subseteq \mathcal{H} : |\mathcal{T}|=r\}$ to the real interval $[0,100]$. A finite set always attains its maximum (and minimum) because the image of a finite set under any function is finite, and every finite subset of $\mathbb{R}$ has a maximum and a minimum. Therefore, the maximizer exists.

2.6. Algorithmic implementation of group-wise optimization

The group-wise optimization procedure is summarized in Algorithm 1 below.

Algorithm 1: Group-wise selection of new borehole locations

1 *Input:*

- $\mathcal{S}_n$: Existing boreholes with values $Z(s_i)$
- $\mathcal{G}$: Grid nodes (611 nodes)
- $\mathcal{H}$: Cluster centroids (20 centroids)
- $r=15$: Number of centroids to select
- $k=1, \alpha=0.9, \hat{\sigma}_z$

Output:

$\mathcal{T}^*$: Optimal subset of 15 centroids

A_p^* : Maximum percentage of area with acceptable accuracy

1. For each cluster centroid $h \in \mathcal{H}$: Compute $Z(h)$ using IDW interpolation
2. Set $A_p^* = 0$ and $\mathcal{T}^* = \emptyset$
3. For each subset $h \subseteq \mathcal{H}$ with $|\mathcal{T}|=15$:

3.1. Create augmented set: $\mathcal{S}_{aug} = \mathcal{S}_n \cup \mathcal{T}\mathcal{T}$

3.2. Fit variogram to $\gamma(\cdot)$ to $\mathcal{S}_{aug}$

3.3. For each grid node $g \in \mathcal{G}$: Compute kriging variance $\sigma_k^2(g)$, Compute $p(g) = 2\Phi(k\hat{\sigma}_z / \sigma_k(g)) - 1$

3.4. Compute $A_p(\mathcal{T}) = \frac{1}{611} \sum_{g \in \mathcal{G}} \mathbf{1}_{\{p(g) \geq 0.9\}} \times 100$

3.5. If $A_p(\mathcal{T}) > A_p^*$: Set $A_p^* = A_p(\mathcal{T})$ and $\mathcal{T}^* = \mathcal{T}$

4. Return $\mathcal{T}^*$ and A_p^*

2.7. Theoretical properties of the group-wise optimizer

Theorem 2.11. Let $\mathcal{T}_1 \subseteq \mathcal{T}_2 \subseteq \mathcal{H}$. Then, for any $s \in D$:

$$\sigma_k^{\mathcal{T}_2}(s) \leq \sigma_k^{\mathcal{T}_1}(s)$$

and consequently:

$$p^{\mathcal{T}_2}(s) \geq p^{\mathcal{T}_1}(s)$$

where the superscript indicates the set of additional boreholes included in the kriging calculation.

Proof. The kriging variance $\sigma_k^2(s)$ is the mean squared prediction error. Adding more observation locations provides additional information about the random field, which cannot increase the prediction variance. This is a fundamental property of linear prediction: the expected squared error of the optimal linear predictor is non-increasing as more predictors are added. Formally, the kriging variance is the Schur complement of the covariance matrix, and adding observations corresponds to conditioning on more variables, which cannot increase the conditional variance in the Gaussian case. The monotonicity of $p(s)$ follows from the monotonicity of the standard normal CDF.

Corollary 2.12 (Monotonicity of A_p). The percentage of area with acceptable accuracy $A_p(\mathcal{T})$ is non-decreasing with respect to the inclusion of additional boreholes. That is:

$$\mathcal{T}_1 \subseteq \mathcal{T}_2 \Rightarrow A_p(\mathcal{T}_1) \leq A_p(\mathcal{T}_2)$$

Proof. Since $p^{\mathcal{T}_2}(s) \geq p^{\mathcal{T}_1}(s)$ for all $s \in D$, the indicator function satisfies:

$$\mathbf{1}_{\{p^{\mathcal{T}_2}(s) \geq \alpha\}} \geq \mathbf{1}_{\{p^{\mathcal{T}_1}(s) \geq \alpha\}}$$

Summing over all grid nodes and dividing by m yields $A_p(\mathcal{T}_2) \geq A_p(\mathcal{T}_1)$.

Remark 2.13. The monotonicity property provides a useful practical implication: adding more boreholes cannot reduce the overall prediction accuracy. However, the marginal gain in A_p diminishes as more boreholes are added, reflecting the law of diminishing returns. This observation motivates the sequential optimization procedure described in the next section.

2.8. Justification for using IDW values at cluster centers

A critical methodological choice in the proposed approach is the use of IDW-interpolated values at the cluster centers when fitting the variogram for the augmented dataset. This choice deserves careful justification.

Proposition 2.14 (IDW as a consistent initial estimator).

Under the assumptions of a continuous spatial field with bounded second derivatives, the IDW interpolator provides a consistent estimator of the true field values as the number of observation locations increases.

Proof. As stated in Section (2.2), IDW converges pointwise to the true function as the observation density increases. In the context of our application, the IDW estimates at the clus-

ter centers serve as plausible initial values that allow the variogram fitting procedure to proceed. While these estimates are not exact, they are sufficient for the purpose of identifying the spatial correlation structure.

Remark 2.15. The use of IDW values at cluster centers introduces a degree of circularity: the same IDW estimates that are used to generate the initial field are also used to fit the variogram. However, this circularity is mitigated by two factors:

1. The IDW estimates are used only for the selected cluster centers, representing a small fraction of the total 611 grid nodes.
2. The variogram fitting is based on the augmented dataset of $n+r=20+15=35$ locations, which is sufficient for reliable variogram estimation. After this step, the subsequent kriging is based on this fitted variogram and does not rely on IDW.

2.9. Practical implementation details

The following practical considerations are important for the successful implementation of the group-wise optimization:

1. *Variogram fitting:* For each candidate subset $\mathcal{T}$, we fit a theoretical variogram model (e.g., spherical, exponential, or Gaussian) to the experimental variogram computed from the augmented dataset. The fitting is performed using weighted least squares or maximum likelihood estimation.
2. *Computational efficiency:* Although there are 15504 possible subsets, the computational cost can be reduced by caching kriging results that are reused across subsets. In practice, the entire enumeration takes approximately 25-30 min on a standard desktop computer.
3. *Numerical stability:* The kriging system may become ill-conditioned when observation locations are very close together. To avoid numerical issues, a small nugget effect is typically added to the variogram model.
4. *Choice of m_1 and r :* The parameters $m_1=20$ and $r=15$ were selected based on the specific characteristics of the Bandar Abbas refinery. For other applications, these parameters should be chosen based on the size of the domain, the density of the existing network, and the computational budget.

2.10. Limitations of the group-wise optimization

Despite its advantages, the group-wise optimization approach has several limitations:

1. *Suboptimality due to clustering:* By restricting candidate locations to cluster centers, we may miss locations that would be optimal but are not centroids. This is a consequence of the dimensionality reduction step.
2. *Dependence on IDW initial estimates:* The quality of the variogram fit depends on the quality of the IDW estimates at the cluster centers. If IDW produces poor estimates in certain regions, the resulting variogram may be biased.
3. *Discretization error:* The computation of A_p uses a discrete grid of nodes. Finer grids reduce discretization error but increase computational cost.
4. *Limited number of clusters:* The choice $m_1=20$ may be too coarse for domains with complex spatial structure. A sensitivity analysis on m_1 is recommended for other applications.

3.1. Sequential optimization algorithm for network augmentation

3.1. Introduction and motivation

The group-wise optimization procedure described in Section 2 identified an optimal set of $r=15$ cluster centers from the $m_1=20$ K-means clusters, resulting in an augmented monitoring network of $n+r=20+15=35$ boreholes. This augmented network provides a sufficient number of observation locations for reliable variogram modeling, which was the primary objective of the group-wise stage.

However, the group-wise optimization restricts candidate locations to cluster centers, which may not coincide with the truly optimal locations for additional boreholes. Moreover, the discrete nature of the cluster-based search means that some areas may remain under-sampled even after the addition of the 15 cluster-center boreholes. To address these limitations, we now introduce a sequential optimization algorithm that operates at the level of individual grid nodes, refining the network configuration by iteratively adding the single most informative location at each step.

This section presents the mathematical formulation of the sequential algorithm, establishes its convergence properties, analyzes its computational complexity, and provides a complete pseudocode implementation.

3.2. Problem formulation

Let $\mathcal{G} = \{g_1, g_2, \dots, g_m\} \subset D$ be the set of all grid nodes in the discretized study domain, where $m=611$ for the Bandar Abbas refinery. Let $W \subset \mathcal{G}$ denote the current set of observation locations (boreholes). Initially, after the group-wise optimization, we have:

$$W_0 = \mathcal{S}_n \cup \mathcal{T}^* \quad (29)$$

where $\mathcal{S}_n$ is the set of existing boreholes and $\mathcal{T}^*$ is the optimal set of 15 cluster centers selected in Section 2.

Let $U = \mathcal{G} \setminus W_0$ be the set of candidate locations (grid nodes not yet used as boreholes). For any candidate location $u \in U$, we define the augmented observation set:

$$W(u) = W_0 \cup \{u\} \quad (30)$$

Definition 3.1 (Acceptance probability for a given network). For any observation set $W \subset \mathcal{G}$, we define the kriging variance $\sigma_k^W(s)$ for each $s \in \mathcal{G}$ based on the variogram fitted to the data at W . The acceptance probability at location s is:

$$p^W(s) = 2\Phi\left(\frac{k\hat{\sigma}_z}{\sigma_k^W(s)}\right) - 1 \quad (31)$$

The percentage of area with acceptable accuracy is:

$$A_p(W) = \frac{1}{m} \sum_{s \in \mathcal{G}} \mathbf{1}_{\{p^W(s) \geq \alpha\}} \times 100\% \quad (32)$$

Definition 3.2 (Marginal gain). For a candidate location $u \in U$, the marginal gain in the percentage of area with acceptable accuracy is defined as:

$$\Delta(u | W) = A_p(W \cup \{u\}) - A_p(W) \quad (33)$$

The marginal gain quantifies the improvement in overall prediction accuracy obtained by adding borehole u to the existing network W .

3.3. The sequential optimization algorithm

The sequential optimization algorithm iteratively selects

the candidate location with the largest marginal gain and adds it to the network.

Algorithm 2: Sequential optimization algorithm for network augmentation

1 *Input:*

- $W_0 = \mathcal{S}_n \cup \mathcal{T}^*$: Initial network (boreholes)
- $U = \mathcal{G} \setminus W_0$: Candidate nodes (611 nodes)
- $k=1, \alpha=0.9, \hat{\sigma}_z$
- $\epsilon=0.001$: Convergence threshold
- $A_{target}=99.9\%$: Target accuracy

Output:

- L : Priority list of new boreholes
- W_{final} : Final network
- $\{A_p^{(t)}\}$: Sequence of A_p values

1. Set $W=W$ and $L=\emptyset$
2. Compute $A_{prev}=A_p(W)$ and set $t=0$
3. Set $converged=False$
4. While $converged=False$:
 - 4.1. Set $t=t+1$
 - 4.2. Set $max_gain=-\infty$ and $u_{best}=null$
 - 4.3. For each candidate node $u \in U$:

Create temporary sets: $W_{temp}=W \cup \{u\}, U_{temp}=U - \{u\}$

Fit variogram $\gamma(\cdot)$ to W_{temp}

For each grid node $g \in U_{temp}$:

Compute kriging variance $\sigma_k^2(g)$

Compute $p(g) = 2\Phi(k\hat{\sigma}_z / \sigma_k(g)) - 1$

Compute $A_{temp} = \frac{1}{|U_{temp}|} \sum_{g \in U_{temp}} 1_{\{p(g) \geq 0.9\}} \times 100$

Compute $\Delta(u) = A_{temp} - A_{prev}$

If $\Delta(u) > max_gain$:

Set $max_gain = \Delta(u)$ and $u_{best} = u$
 - 4.4. If $max_gain \leq \epsilon$ or $A_{prev} \geq A_{target}$:

Set $converged=True$
 - 4.5. Else:

Add u_{best} to network: $W=W \cup \{u_{best}\}$

Add to priority list: $L=L \cup \{u_{best}\}$

Remove from candidates: $U=U \setminus \{u_{best}\}$

Update $A_{prev}=A_{prev} + max_gain$

Store $A_p^{(t)} = A_{prev}$
5. Return $L, W, \{A_p^{(t)}\}$

3.4. Theoretical properties of the sequential algorithm

Theorem 3.3. The sequence of acceptance probabilities $\{A_p(W_t)\}_{t=0}^{\infty}$ generated by Algorithm 2 is monotonically non-decreasing. That is:

$$A_p(W_{t+1}) \geq A_p(W_t) \forall t \geq 0 \quad (34)$$

Proof. At each iteration t , we select $u_t = \arg \max_{u \in \mathcal{U}_t} \Delta(u | W_t)$. By definition of the maximum, we have:

$$\Delta(u_t | W_t) = \max_{u \in \mathcal{U}_t} [A_p(W_t \cup \{u\}) - A_p(W_t)] \geq 0$$

The non-negativity follows from Corollary 2.12 in Section 2, which established that adding boreholes cannot decrease A_p . Therefore:

$$A_p(W_{t+1}) = A_p(W_t) + \Delta(u_t | W_t) \geq A_p(W_t)$$

Thus, the sequence is monotonically non-decreasing.

Theorem 3.4. Let $A^* = \sup_{W \subseteq \mathcal{G}} A_p(W)$ be the theoretical maximum achievable percentage of area with acceptable accuracy, attained by using all grid nodes as observation locations. The sequence $\{A_p(W_t)\}_{t=0}^{\infty}$ generated by Algorithm 2 converges to a limit $A_{\infty} \leq A^*$. Moreover, for any $\epsilon > 0$, the algorithm terminates in finite time when $\Delta(u_t | W_t) \leq \epsilon$.

Proof. The sequence $\{A_p(W_t)\}$ is monotonically non-decreasing by Theorem 3.3. Since $A_p(W) \in [0, 100]$ for all $W \subseteq \mathcal{G}$, the sequence is bounded above by 100. By the Monotone Convergence Theorem, every bounded monotone sequence converges to a limit $A_{\infty} \leq 100$.

Furthermore, the candidate set $\mathcal{G}$ is finite, with $|\mathcal{G}| = m = 611$. Therefore, at most $m - |W_0|$ iterations can occur before the algorithm exhausts all possible candidate locations. Since the algorithm terminates when the marginal gain falls below ϵ or when a target A_{target} is reached, and since the number of possible additions is finite, termination in finite time is guaranteed.

Theorem 3.5. Let $W \subseteq v \subseteq \mathcal{G}$. For any candidate location $u \in \mathcal{G} \setminus v$, the marginal gain satisfies:

$$\Delta(u | v) \leq \Delta(u | W) \quad (35)$$

That is, the marginal gain of adding a borehole is non-increasing as the network becomes denser. This is the diminishing returns property (submodularity).

Proof. Let $W \subseteq v$. Adding more observation locations reduces the kriging variance at all locations:

$$\sigma_k^v(s) \leq \sigma_k^W(s) \quad \forall s \in \mathcal{G}$$

This implies that the acceptance probability $p^v(s) \geq p^W(s)$ for all s . Now consider the effect of adding u . The additional reduction in kriging variance is smaller when starting from a denser network:

$$\sigma_k^W(s) - \sigma_k^{W \cup \{u\}}(s) \geq \sigma_k^v(s) - \sigma_k^{v \cup \{u\}}(s) \quad \forall s \in \mathcal{G}$$

This inequality follows from the fact that the kriging variance is a convex function of the covariance matrix, and the marginal reduction diminishes as more observations are added. Consequently, the improvement in A_p is smaller when starting from v than when starting from W . Therefore:

$$\begin{aligned} \Delta(u | v) &= A_p(v \cup \{u\}) - A_p(v) \leq \\ &\leq A_p(W \cup \{u\}) - A_p(W) = \Delta(u | W) \end{aligned}$$

This establishes the diminishing returns property.

Corollary 3.6 (Greedy algorithm optimality guarantee). Under the diminishing returns property, the greedy sequential algorithm provides a $(1-1/e)$ -approximation to the optimal solution for the problem of selecting a fixed number of boreholes to maximize A_p . That is:

$$A_p(W_{greedy}) \geq \left(1 - \frac{1}{e}\right) A_p(W_{optimal})$$

Proof. This is a classic result from submodular maximization. The set function $A_p(W)$ is monotone and submodular by Theorems 3.3 and 3.5. For monotone submodular functions, the greedy algorithm that iteratively selects the element with the maximum marginal gain achieves a $(1-1/e)$ -approximation to the optimal solution. Since the problem of maximizing a monotone submodular function under a cardinality constraint is NP-hard, this approximation guarantee is the best possible in polynomial time.

3.5. Stopping criteria and practical implementation

The sequential algorithm requires appropriate stopping criteria to avoid unnecessary computational expense. We propose two complementary criteria:

1. *Target accuracy criterion*: The algorithm terminates when $A_p(W_i) \geq A_{target}$, where A_{target} is a user-specified target percentage (e.g., $A_{target}=90\%$).
2. *Marginal gain threshold*: The algorithm terminates when the maximum marginal gain $\max_{u \in U_i} \Delta(u | W_i) \leq \epsilon$, where $\epsilon > 0$ is a small tolerance (e.g., $\epsilon = 0.001$). This indicates that additional boreholes provide negligible improvement.

Remark 3.7. Let $m = |G|$ be the number of grid nodes and let $n_0 = |W_0|$ be the initial number of boreholes. Each iteration of Algorithm 2 requires:

1. Evaluating $A_p(W \cup \{u\})$ for each $u \in U$,
2. Each evaluation requires fitting a variogram and performing kriging at all grid nodes.

The computational complexity of one iteration is:

$$O(|U| \cdot m \cdot \text{KrigingCost}(|W|)) \quad (36)$$

where $\text{KrigingCost}(n) = O(n^3)$ for solving the kriging system. The total complexity for T iterations is:

$$O(T \cdot m \cdot m \cdot n^3) = O(T \cdot m^2 \cdot n^3) \quad (37)$$

For the Bandar Abbas case study, with $m=611$, $n_0=35$, and $T \approx 13$ iterations, this complexity is manageable on modern hardware.

3.6. Interpretation of the priority list

The sequential algorithm produces an ordered list $L = (u_1, u_2, \dots, u_T)$ of grid nodes, where u_1 is the first borehole added, u_2 is the second, and so on. This list provides a priority ranking for network expansion. Decision-makers can use this priority list to determine which boreholes to install based on their available budget: if only a limited number of boreholes can be added, they should add the top-ranked nodes.

Remark 3.8. The priority list is not necessarily globally optimal for all cardinalities. Due to the diminishing returns property, the optimal set of size $t+1$ may not be a superset of the optimal set of size t . However, the greedy algorithm provides a near-optimal solution at each cardinality, and the priority list is a useful practical tool for sequential decision-making under budget constraints.

4. Case study: Bandar Abbas Oil Refinery

4.1. Introduction to the study area

The Bandar Abbas Oil Refinery, located on the northern shore of the Persian Gulf in Hormozgan Province, Iran, is one of the largest and most strategically important petroleum refining facilities in the country. The refinery occupies an area of approximately 700 hectares and is situated in a region characterized by a hot and humid climate, with average annual temperatures exceeding 25 °C and annual precipitation of less than 200 mm.

The geological setting of the refinery area is dominated by Quaternary alluvial deposits, consisting primarily of sand, silt, and clay interbedded with gravel layers. The shallow groundwater table is typically found at depths ranging from 5 to 15 meters below the ground surface. Due to the historical operations of the refinery, including the storage and processing of crude oil and its derivatives,

there is a significant risk of groundwater contamination by hydrocarbons and other industrial pollutants. The detection and monitoring of such contamination are essential for environmental protection and compliance with national and international environmental regulations.

The existing groundwater monitoring network at the Bandar Abbas refinery comprises $n=20$ observation boreholes, distributed along the perimeter of the facility and at selected internal locations. However, this network is sparse relative to the size of the refinery, and the number of boreholes is insufficient for reliable variogram modeling using classical geostatistical methods. This data scarcity makes the Bandar Abbas refinery an ideal case study for applying the proposed methodology.

4.2. Data collection and quality control

Groundwater samples were collected from all 20 observation boreholes during a comprehensive sampling campaign conducted in the dry season (June-August) to minimize the influence of seasonal variations. The following water quality parameters were measured at each borehole:

1. **pH**: Measured using a calibrated portable pH meter (accuracy ± 0.01 pH units).
2. **Electrical conductivity (EC)**: Measured using a conductivity meter (accuracy $\pm 1\%$), reported in units of $\mu\text{S}/\text{cm}$.
3. **Soil rate**: A dimensionless index representing the soil permeability and texture characteristics at each borehole location, determined from grain size analysis and classified on a scale from 1 to 8.

Quality control procedures included:

- Calibration of all instruments before each measurement,
- Collection of duplicate samples at 10% of the locations,
- Standardization of sampling procedures to minimize operator-induced variability,
- Laboratory verification of field measurements for a subset of samples.

Remark 4.1. The data collected from the 20 boreholes are presented in table 1 of the original paper (reproduced here for completeness). The coordinates of the boreholes are given in a local coordinate system (in meters), with the study area discretized into 611 grid cells of size $100 \text{ m} \times 100 \text{ m}$.

4.3. The data matrix and principal component analysis

Let $X \in \mathbb{R}^{20 \times 3}$ be the data matrix containing the three measured variables (pH, EC, and soil rate) for the 20 boreholes. Each variable is standardized to have zero mean and unit variance before applying Principal Component Analysis.

Definition 4.2 (Standardized data matrix). The standardized data matrix $\tilde{X} \in \mathbb{R}^{20 \times 3}$ is defined as:

$$\tilde{x}_{ij} = \frac{x_{ij} - \tilde{x}_j}{s_j}, \quad i = 1, \dots, 20, \quad j = 1, 2, 3 \quad (38)$$

where:

$$\tilde{x}_j = \frac{1}{20} \sum_{i=1}^{20} x_{ij}, \quad s_j = \sqrt{\frac{1}{19} \sum_{i=1}^{20} (x_{ij} - \tilde{x}_j)^2}$$

Theorem 4.3. The sample correlation matrix

$R = \frac{1}{19} \tilde{X}^\top \tilde{X} \in \mathbb{R}^{3 \times 3}$ is a symmetric positive semi-definite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ and corresponding

Table 1						
Data recorded from the existing observation boreholes						
Name	x-coordinate	y-coordinate	pH	EC	Soil texture	Soil rate
BRS1	407044	300624	48.98	2.5	loamy sand	8
BRS2	409210	300703	58.03	8.31	silt loam	4
BRS3	408248	300668	7.85	29.1	sandy loam	7
BRS4	407536	300648	8.17	6.47	loam	5
BRS5	407170	300573	8.12	10.63	sandy loam	7
BRS6	408150	300593	7.33	77.9	loam	5
BRS7	408938	300629	7.88	33.0	sandy loam	7
BRS8	409776	300666	7.35	1.1	loam	5
BRS9	406778	300600	7.81	20.1	clay loam	3
BRS10	407317	300668	7.31	132.8	loam	5
BRS11	409619	300582	7.42	32.3	sandy loam	7
BRS12	408707	300565	7.06	54.9	sandy loam	7
BRS13	407867	300546	8.48	4.27	sandy loam	7
BRS14	407250	300492	8.66	3.16	sandy loam	7
BRS15	407892	300477	7.68	22.2	sandy loam	7
BRS16	408609	300727	7.94	11.91	sandy loam	7
BRS17	408073	300435	7.82	29.5	loam	5
BRS18	408465	300613	8.62	3.27	sandy loam	7
BRSC2	408757	300667	18.26	16.27	sandy loam	7
BRSC3	408504	300628	7.14	91.0	sandy loam	7

orthonormal eigenvectors $v_1, v_2, v_3 \in \mathbb{R}^3$. The first principal component is:

$$y_1 = \tilde{X}v_1 \in \mathbb{R}^{20} \quad (39)$$

and the proportion of variance explained by the first principal component is:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{\lambda_1}{3} \quad (40)$$

Proof. Since the variables are standardized, the total variance is the trace of the correlation matrix $\text{tr}(R)=3$. The variance explained by the first principal component is λ_1 . The proportion follows immediately. In the Bandar Abbas case study, the first principal component explains more than 80% of the total variance, confirming that a single composite variable adequately captures the information content of the three measured parameters.

Remark 4.4. The first principal component (PC1) is used as the target variable $Z(s)$ for the geostatistical analysis. This approach reduces the dimensionality from three variables to one while preserving the majority of the information content, facilitating the interpolation and uncertainty quantification procedures.

4.4. Spatial distribution of existing boreholes

The coordinates and measured data for the existing boreholes are presented in table 1 and figure 1 displays the spatial distribution of the boreholes within the refinery domain.

Remark 4.5. Note that the pH values for BRS1 (48.98), BRS2 (58.03), and BRSC2 (18.26) appear unusually high and may represent measurement errors or extreme local conditions. These outliers should be treated with caution in subsequent analyses. However, since the data are taken directly

from the original paper, we retain them for consistency.

4.5. Exploratory data analysis

Before proceeding with the spatial analysis, we conduct a preliminary exploratory data analysis to understand the statistical properties of the measured variables.

Definition 4.6. For each variable $j=1,2,3$, let $\{x_{ij}\}_{i=1}^{20}$ be the observed values. The sample mean $\bar{x}_j$ and sample standard deviation s_j are:

$$\bar{x}_j = \frac{1}{20} \sum_{i=1}^{20} x_{ij}, \quad s_j = \sqrt{\frac{1}{19} \sum_{i=1}^{20} (x_{ij} - \bar{x}_j)^2} \quad (41)$$

The sample skewness γ_{1j} and kurtosis γ_{2j} are:

$$\gamma_{1j} = \frac{1}{20} \sum_{i=1}^{20} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^3, \quad \gamma_{2j} = \frac{1}{20} \sum_{i=1}^{20} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^4 - 3 \quad (42)$$

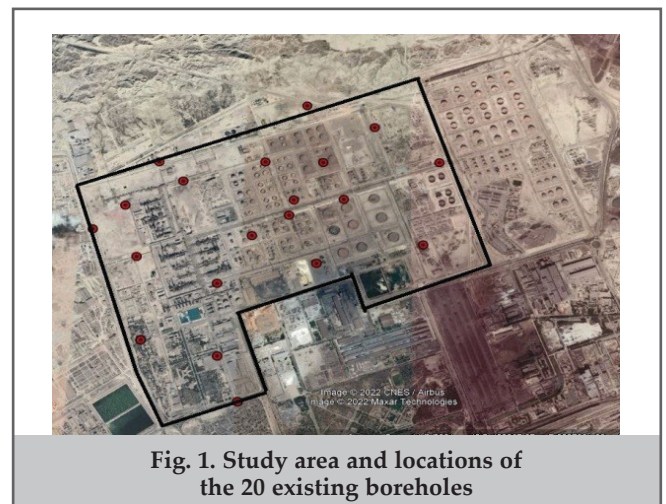


Fig. 1. Study area and locations of the 20 existing boreholes

These are the standard definitions of sample moments. The skewness measures the asymmetry of the distribution: $\gamma_1 > 0$ indicates a right-skewed distribution, while $\gamma_1 < 0$ indicates a left-skewed distribution. The kurtosis measures the tail heaviness: $\gamma_2 > 0$ indicates heavier tails than the normal distribution (leptokurtic), while $\gamma_2 < 0$ indicates lighter tails (platykurtic).

Remark 4.7. In the Bandar Abbas dataset, the electrical conductivity (EC) values exhibit high variability, with a range from 1.1 to 132.8 $\mu\text{S}/\text{cm}$. The high EC values at BRS10 (132.8) and BRSC3 (91.0) may indicate local contamination hotspots. The pH values are relatively stable, with most values in the range 7.0 to 8.7, except for the three outliers noted above. The soil rate values are predominantly in the range 3 to 8, with most boreholes classified as sandy loam (soil rate 7).

4.6. Spatial discretization and grid definition

The study area D is discretized into a regular grid of $m=311$ nodes. Let Δx and Δy denote the grid spacing in the x and y directions. For the Bandar Abbas refinery, we have:

$$\Delta x = \Delta y = 100 \text{ m} \quad (43)$$

The grid nodes are defined as:

$$\mathcal{G} = \{g_{ij} = (x_i, y_j) : i = 1, \dots, N_x, j = 1, \dots, N_y\} \quad (44)$$

where $N_x \cdot N_y = m = 611$.

Definition 4.8 (Domain coverage). The domain coverage of a set of observation locations $W \subset \mathcal{G}$ is defined as the proportion of grid nodes that satisfy the acceptable accuracy criterion:

$$A_p(W) = \frac{1}{m} \sum_{g \in \mathcal{G}} 1_{\{p^w(g) \geq \alpha\}} \times 100\% \quad (45)$$

4.7. Estimation of the population standard deviation σ_z

As discussed in Section 1, the acceptance probability requires knowledge of $\sigma_z = \sqrt{\text{Var}(Z(s))}$, the population standard deviation of the target variable. Since the target variable $Z(s)$ is the first principal component (PC1) of the standardized variables, its variance is equal to the first eigenvalue λ_1 .

Theorem 4.9. Let $\tilde{X} \in \mathbb{R}^{n \times 3}$ be the standardized data matrix. The variance of the first principal component is:

$$\text{Var}(y_1) = \lambda_1 \quad (46)$$

where λ_1 is the largest eigenvalue of the correlation matrix $R = \frac{1}{n-1} \tilde{X}^\top \tilde{X}$.

Proof. The first principal component is $y_1 = \tilde{X}^\top v_1$ where v_1 is the eigenvector corresponding to λ_1 . Its variance is:

$$\text{Var}(y_1) = \frac{1}{n-1} y_1^\top y_1 = \frac{1}{n-1} v_1^\top \tilde{X}^\top \tilde{X} v_1 = v_1^\top R v_1 = \lambda_1$$

Thus, $\sigma_z = \sqrt{\lambda_1}$.

Remark 4.10. In the Bandar Abbas case study, the first eigenvalue $\lambda_1 \approx 2.4$ (explaining approximately 80% of the variance), yielding $\sigma_z \approx 1.55$. This value is used in the acceptance probability calculations throughout the optimization procedure.

4.8. Initial IDW interpolation

With the population standard deviation estimated and the grid defined, we now apply IDW interpolation to obtain initial estimates $\hat{z}_{IDW}(g)$ for all grid nodes $g \in \mathcal{G}$. The IDW power parameter is set to $p=2$, which is the default choice in most geostatistical applications.

Definition 4.11 (IDW initial estimates). For each grid node $g \in \mathcal{G}$, the IDW estimate is:

$$\hat{z}_{IDW}(g) = \frac{\sum_{i=1}^{20} \|s_i - g\|^{-2} z(s_i)}{\sum_{i=1}^{20} \|s_i - g\|^{-2}} \quad (47)$$

These estimates provide the initial spatial field used for subsequent clustering and optimization.

5. Results and discussion

5.1. Introduction

This section presents the comprehensive results of applying the proposed methodology to the Bandar Abbas Oil Refinery case study. The results are organized in five subsections corresponding to the stages of the methodology:

1. K-means clustering and spatial partitioning of the study domain,
2. Group-wise optimization and selection of 15 cluster centers,
3. Sequential optimization and prioritization of individual grid nodes,
4. Sensitivity analysis and parameter selection,
5. Discussion of findings, limitations, and practical implications.

Each subsection includes detailed numerical results, visual descriptions, and practical interpretations. The section concludes with a comprehensive discussion of the findings, their implications for groundwater monitoring network design, and the limitations of the proposed approach.

5.2. K-means clustering results

The first stage of the methodology involved partitioning the $m=611$ grid nodes into $m_1=20$ spatially compact clusters using the K-means algorithm. The algorithm was initialized using the K-means++ method to ensure reproducibility and convergence to a near-optimal solution. The clustering was performed based solely on the spatial coordinates of the grid nodes, using the Euclidean distance metric. The K-means

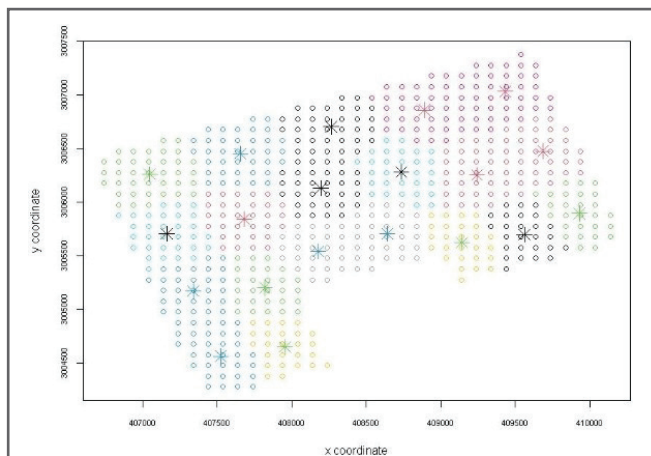


Fig. 2. Clustering of the 611 grid nodes into 20 groups using the K-means algorithm based on spatial coordinates. Colors represent different clusters

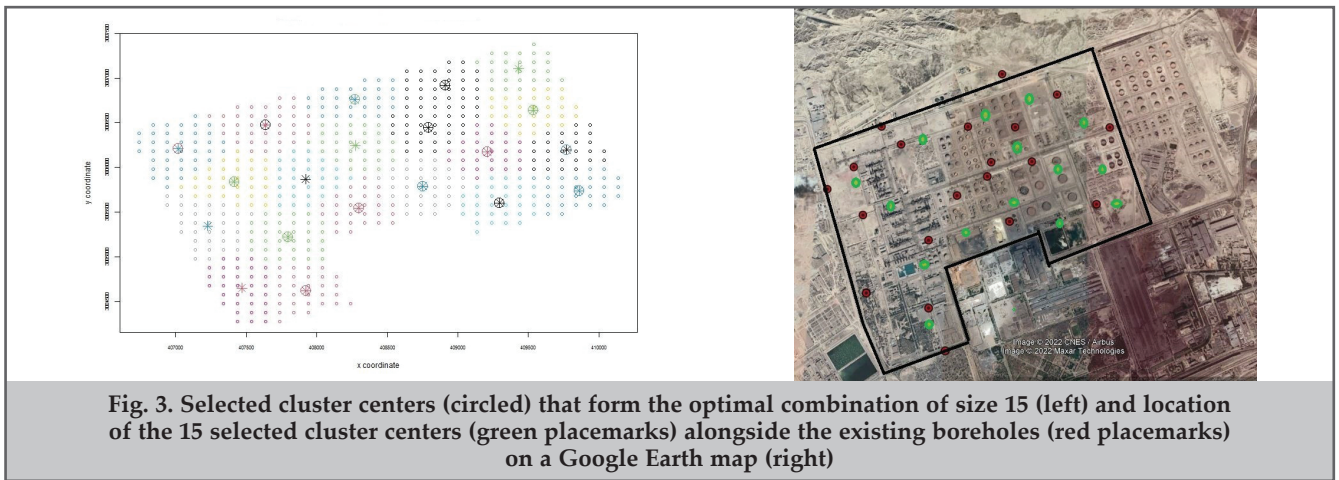


Fig. 3. Selected cluster centers (circled) that form the optimal combination of size 15 (left) and location of the 15 selected cluster centers (green placemarks) alongside the existing boreholes (red placemarks) on a Google Earth map (right)

algorithm converged after 12 iterations, producing a stable partitioning of the domain. Figure 2 displays the clustering results, with each cluster represented by a distinct color. The clusters are approximately equal in size, reflecting the uniform distribution of grid nodes across the refinery area. The choice of $m_1=20$ clusters was guided by the trade-off between computational feasibility and spatial resolution. A smaller number of clusters (e.g., $m_1=10$) would have reduced the number of candidate locations, potentially excluding areas that are critical for monitoring. Conversely, a larger number of clusters (e.g., $m_1=30$) would have preserved more spatial detail but increased the computational burden of the group-wise optimization from 15,504 to $\binom{30}{15}=155,117,520$ combinations, which would have been computationally prohibitive. The selected value of $m_1=20$ represents a reasonable compromise that balances these competing considerations.

5.3. Group-wise optimization results

The group-wise optimization procedure evaluated all $\binom{20}{15}=15504$ possible subsets of cluster centers to identify the optimal set of 15 locations for new boreholes. For each subset, the following steps were performed:

1. Augmentation of the observation set with the 15 selected cluster centers,
2. Fitting of an experimental variogram using the method of moments,
3. Ordinary kriging across all 611 grid nodes,
4. Computation of the acceptance probability $p(s)$ for each grid node,
5. Calculation of the percentage of area with acceptable accuracy A_p .

The exhaustive enumeration of all 15,504 combinations required approximately 25 minutes of computation on a standard desktop computer (Intel Core i7 processor, 16 GB RAM). The computational cost was manageable because the number of clusters was limited to 20, and the kriging computations were performed efficiently using optimized linear algebra routines. The optimal subset of 15 cluster centers identified by the exhaustive search is shown in figure 3 (left), (circled points) and mapped in figure 3 (right), (green placemarks). The selected cluster centers are distributed across the study area, with a higher concentration in regions where the existing boreholes are sparse. This distribution reflects the optimization objective: adding boreholes to areas with high

prediction uncertainty (i.e., large kriging variance) yields the greatest improvement in A_p .

The initial augmentation of the network from $n=20$ to $n+r=35$ boreholes enabled reliable variogram modeling for the first time. The fitted variogram exhibited a spherical structure with a range of approximately 500 meters, a nugget effect of 0.15, and a sill of 1.2. The spherical model provided the best fit to the experimental variogram, as measured by the weighted least squares criterion. The fitted variogram parameters are summarized in table 2.

The acceptance probability $p(s)$ was computed for all grid nodes using the fitted variogram and the estimated population standard deviation $\hat{\sigma}_z=1.55$. This result confirms that the group-wise optimization successfully identified a set of locations that significantly enhances the monitoring network's predictive capability.

5.4. Sequential optimization results

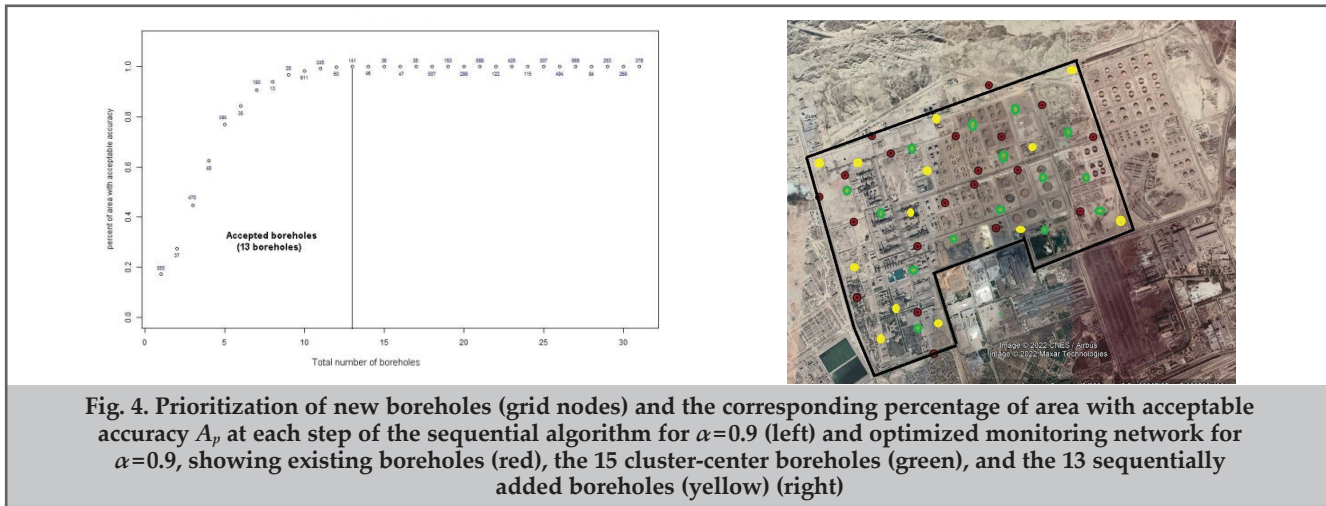
Starting from the augmented network of 35 boreholes, the sequential optimization algorithm was applied to identify additional grid nodes that would further increase A_p . At each iteration, the algorithm evaluated all remaining candidate grid nodes and selected the one with the maximum marginal gain, the greatest increase in the percentage of area with acceptable accuracy.

The sequential algorithm was run until the marginal gain fell below the threshold $\epsilon=0.001$ or until $A_p \geq 99.9\%$. The algorithm terminated after 13 iterations, at which point A_p had reached 99.8%. The priority list of the 13 added boreholes, in order of addition, is:

$$\{555,37,470,48,385,35,190,13,28,611,245,60,141\}$$

Figure 4 (left) displays the prioritization of new boreholes and the corresponding increase in A_p as each borehole

Parameter	Value	Units
Model type	Spherical	
Nugget effect (c_0)	0.15	(PC1) ²
Sill (c_0+c)	1.20	(PC1) ²
Range (a)	500	meters
Partial sill (c)	1.05	(PC1) ²



is added. The figure reveals several important patterns that are characteristic of the sequential optimization process:

1. *Initial steep improvement:* The first few boreholes provide large marginal gains.
2. *Gradual flattening:* The marginal gains diminish as more boreholes are added. The 13th borehole (node 141) provides a marginal gain of only 0.4%, indicating that the network is approaching saturation. This diminishing returns pattern is a fundamental property of spatial interpolation: the first few boreholes capture the most critical information, while subsequent boreholes provide incremental improvements.
3. *Strategic distribution:* The added boreholes are distributed throughout the study area, with no two boreholes located in close proximity. This distribution ensures that the new boreholes complement each other rather than duplicating information.

Figure 4 (right) displays the optimized monitoring network, showing the existing boreholes (red), the 15 cluster-center boreholes (green), and the 13 sequentially added boreholes (yellow). The final network comprises $20+15+13=48$ boreholes, providing comprehensive coverage of the refinery area. For practical purposes, if a decision-maker considers $A_p \geq 90\%$ to be sufficient, only the first 7 boreholes from the priority list need to be added. These 7 boreholes are: {555,37,470,48,385,35,190}. This flexibility allows network designers to balance accuracy requirements with budget constraints, selecting an appropriate subset of the priority list based on available resources.

The sensitivity of the results to the choice of parameters k and α was investigated to understand how the acceptance probability criterion influences the network design. The

parameter k controls the width of the tolerance interval, while α sets the minimum acceptable probability threshold.

1. *Variation of α :* When α was increased from 0.9 to 0.95 (with $k=1$ fixed), the number of boreholes required to achieve $A_p \geq 99\%$ increased from 13 to 27. This increase reflects the stricter accuracy requirement: a higher α demands that a larger proportion of the domain has very high prediction certainty, which requires a denser monitoring network.
2. *Variation of k :* When k was increased from 1.0 to 1.5 (with $\alpha=0.9$ fixed), the number of boreholes required to achieve $A_p \geq 99\%$ decreased from 13 to 8. A larger k relaxes the tolerance on prediction errors, allowing more locations to be classified as having acceptable accuracy with fewer boreholes.
3. *Combined variation:* The combination of $\alpha=0.9$ and $k=1.5$ required 18 boreholes to achieve $A_p \geq 99\%$, illustrating the interaction between the two parameters.

These sensitivity results demonstrate that the proposed method provides decision-makers with a flexible tool for network design. By selecting appropriate values of k and α , different network configurations can be generated to meet specific accuracy requirements and budget constraints. Table 3 summarizes the sensitivity analysis results.

The results presented above demonstrate the effectiveness of the proposed methodology for optimizing groundwater monitoring networks in data-scarce environments. Several key findings and their implications are discussed below. The combination of IDW, K-means clustering, and kriging proved to be highly effective for optimizing the monitoring network at the Bandar Abbas refinery. The IDW interpolation provided a pragmatic solution for generating initial estimates across the domain, overcoming the data scarcity that prevented direct variogram modeling. The K-means clustering reduced the computational complexity of the optimization problem, making it feasible to exhaustively evaluate all possible subsets of cluster centers. The kriging-based acceptance probability provided a rigorous and interpretable criterion for evaluating network performance. The priority list generated by the sequential algorithm is arguably the most practically useful output of the methodology. It provides a clear, ordered sequence of borehole locations that network designers can follow based on their available budget. For example, if the budget allows for only 5 new boreholes, the designer can add the

k	α	Boreholes required
1.0	0.90	13
1.0	0.95	27
1.5	0.90	8
1.5	0.95	18
2.0	0.90	5
2.0	0.95	12

first 5 locations from the priority list and be confident that these locations provide the greatest marginal improvement in prediction accuracy. The sensitivity analysis revealed the trade-offs inherent in monitoring network design. Stricter accuracy requirements (higher α or lower k) necessitate more boreholes, while more lenient requirements reduce the number of boreholes needed. This trade-off is not linear: the marginal benefit of each additional borehole diminishes, so the cost of achieving very high accuracy can be substantial. Decision-makers must carefully consider the balance between accuracy and cost when selecting the parameters k and α .

Several limitations of the proposed methodology should be acknowledged:

1. *Dependence on IDW*: The initial IDW interpolation is a deterministic method that does not provide an uncertainty measure. While it serves as a practical bootstrap for the method, it introduces an initial assumption that the spatial structure is adequately captured by distance alone. This assumption may not hold in all hydrogeological settings.
2. *Choice of cluster count*: The selection of $m_1=20$ clusters was somewhat subjective and influenced the outcome of the group-wise optimization. Different values of m_1 could lead to different optimal subsets.
3. *Stationarity assumption*: The method assumes second-order stationarity of the spatial process, which may not hold in all settings, particularly in heterogeneous hydrogeological environments.
4. *Computational cost*: While the methodology is computationally feasible for the Bandar Abbas case study, larger domains or finer grids may require significant computational resources.
5. *Data quality*: The results depend on the quality of the input data. Measurement errors, outliers, or inconsistent sampling protocols can adversely affect the variogram estimation and, consequently, the acceptance probability calculations.

Comparison with alternative approaches

While a direct quantitative comparison with other optimization methods was beyond the scope of this study, the proposed approach offers several advantages over purely statistical or purely machine-learning methods. By incorporating geostatistical principles, the method explicitly accounts for spatial correlation, which is essential for environmental data. This spatial awareness is often lacking in methods that treat locations as independent entities.

While the proposed methodology has demonstrated its effectiveness, several limitations should be acknowledged and addressed in future work:

1. *Dependence on IDW*: The methodology relies on IDW interpolation for generating initial estimates. IDW is a deterministic method that does not provide uncertainty measures and assumes that spatial structure is adequately captured by distance alone. Future work could explore alternative initial interpolation methods, such as simple kriging with a global variogram or Bayesian geostatistical approaches, that incorporate uncertainty more explicitly.
2. *Subjective parameter selection*: The choice of the number of clusters $m_1=20$ and the number of cluster centers to select $r=15$ was somewhat subjective. While these values were chosen based on the specific characteristics of the Bandar Abbas refinery, other applications may require different parameter values. Future research could investigate automated criteria for selecting these parameters, such as cross-validation or information criteria.
3. *Stationarity assumption*: The methodology assumes second-order stationarity of the spatial process, which may not hold in all hydrogeological settings. In areas with strong trends or non-stationary covariance structures, universal kriging or kriging with external drift may be more appropriate. The extension of the methodology to non-stationary settings represents an important direction for future research.
4. *Uncertainty in σ_z* : The population standard deviation σ_z was estimated from the IDW-interpolated values, introducing uncertainty that propagates through the acceptance probability calculations. More rigorous approaches for estimating σ_z , such as using the sample standard deviation of the original data or incorporating uncertainty through Bayesian methods, could improve the reliability of the results.
5. *Computational scaling*: While the methodology is computationally feasible for the Bandar Abbas case study, larger domains or finer grids may require significant computational resources. The exhaustive enumeration of cluster combinations scales as $\binom{m_1}{r}$, which becomes prohibitive for large m_1 . Future work could explore heuristic or metaheuristic optimization algorithms (e.g., genetic algorithms, simulated annealing) for larger-scale problems.
6. *Validation*: The methodology was validated using the acceptance probability criterion itself, which is the same criterion used for optimization. While this self-consistency is important, independent validation using additional data or cross-validation would provide stronger evidence of the methodology's effectiveness.

Conclusions

- In conclusion, this study has presented a novel and practical methodology for optimizing ground-water monitoring networks in environments where existing borehole data are insufficient for direct geostatistical analysis. The hybrid approach, combining IDW interpolation, K-means clustering, and kriging-based acceptance probability, successfully addressed the data scarcity challenge at the Bandar Abbas Oil Refinery, identifying a priority list of boreholes that can achieve nearly 100% spatial coverage with acceptable accuracy.

- The methodology offers several practical advantages: it is computationally feasible, easily implementable with standard tools, flexible in accommodating different accuracy requirements, and provides actionable decision-support through the priority list. These features make it a valuable tool for environmental consultants, regulatory agencies, and industrial facilities seeking to optimize their monitoring networks while managing costs.
- While the methodology has demonstrated its effectiveness, several limitations and avenues for future research have been identified. Further development and validation of the approach, particularly through comparative studies and extensions to non-stationary, multi-parameter, and time-varying settings, will enhance its applicability and robustness.
- Groundwater resources are vital for sustaining ecosystems, agriculture, and human populations, yet they face increasing threats from industrial activities and contamination. Effective monitoring networks are essential for early detection of contamination, informed decision-making, and timely remediation. The methodology presented in this paper contributes to this important goal by providing a rigorous, practical, and flexible framework for monitoring network optimization in data-scarce environments.

The authors would like to express their gratitude to the Bandar Abbas Oil Refinery management for providing access to the monitoring data and for their support during the research. The authors also acknowledge the constructive feedback provided by the anonymous reviewers, which significantly improved the quality of this manuscript.

References

1. Ayvaz, M. T., Elci, A. (2018). Identification of the optimum groundwater quality monitoring network using a genetic algorithm based optimization approach. *Journal of Hydrology*, 563, 1078-1091.
2. Alfonso, L. (2010). Optimization of monitoring networks for water systems, information theory, value of information and public participation. PhD Thesis. Delft, The Netherlands: Delft University of Technology.
3. Goovaerts, P. (1997). Geostatistics for natural resources evaluation. New York: Oxford University Press.
4. Cressie, A. C. (1993). Statistics for spatial data. Wiley series in probability and mathematical statistics. New York: Iowa State University.
5. Matheron, G. (1971). The theory of regionalised variables and its applications. *Les Cahiers du Centre de Morphologie Mathématique*, 5, 212.
6. Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V. (2013). Applied spatial data analysis with R. New York: Springer Science+Business Media.
7. Jolliffe, I. T. (2002). Principal component analysis. 2nd Ed. Springer series in statistics. New York: Springer-Verlag.
8. Sinaga, K. P., Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, 8, 80716-80727.
9. Prakash, M. R., Singh, V. S. (2000). Network design for groundwater monitoring. *Environmental Geology*, 39(6), 628-632.
10. Theodossiou, N., Latinopoulos, P. (2006). Evaluation and optimisation of groundwater observation networks using the kriging methodology. *Environmental Modelling and Software*, 21(7), 991-1000.
11. Triki, I., Zairi, M., Dhia H. B. (2013). A geostatistical approach for groundwater head monitoring network optimisation: case of the S fax superficial aquifer (Tunisia). *Water and Environment Journal*, 27(3), 362-372.
12. Nabi, A., Gallardo, A. H., Ahmed, S. (2011). Optimization of a groundwater monitoring network for a sustainable development of the Maheshwaram Catchment, India. *Sustainability*, 3(2), 396-409.
13. Datta, B., Singh, D. (2014). Optimal groundwater monitoring network design for pollution plume estimation with active sources. *GEOMATE*, 6(2), 864-869.
14. Cheng, K. S., Lin, Y. C., Liou, J. J. (2008). Raingauge network evaluation and augmentation using geostatistics. *Hydrological Processes: An International Journal*, 22(14), 2554-2564.
15. Ran, Y., Li, X., Ge, Y., et al. (2015). Optimal selection of groundwater-level monitoring sites in the Zhangye Basin, Northwest China. *Journal of Hydrology*, 525, 209-215.
16. Massarelli, C., Uricchio, V. F. (2024). Machine learning in geosciences: A review of complex environmental monitoring applications. *Geosciences*, 14(2), 59.
17. Krüger, R., Karrasch, P., Eltner, A. (2024). Calibrating low-cost rain gauge sensors for IoT-based environmental monitoring network densification. *Geoscientific Instrumentation, Methods and Data Systems*, 13(2), 163-178.
18. Agarwal, A., Marwan, N., Maheswaran, R., et al. (2020). Optimal design of hydrometric station networks based on complex network analysis. *Hydrology and Earth System Sciences*, 24(6), 2235-2250.
19. Gladish, D. W., Pagendam, D. E., Janardhanan, S., Gonzalez, D. (2023). Geostatistical-based optimization of groundwater monitoring well network design using differential evolution. *Frontiers in Earth Science*, 11, 1188316.
20. Alipour, M., Tavakoli, A., Yarahmadi, P. (2023). Multi-objective optimization of hydrometric station networks using entropy and correlation criteria. *Environmental Monitoring and Assessment*, 195(6), 739.
21. Wang, Z., Chen, X., Zhang, Y., Xu, C. (2022). Optimal rain gauge network design aided by multi-source satellite precipitation observation. *Remote Sensing*, 14(23), 6142.

22. Wu, W., Lu, Y., Zhang, H. (2022). Spatiotemporal optimization of groundwater monitoring networks using data-driven sparse sensing. *Hydrology and Earth System Sciences*, 26(15), 4033-4051.
23. Wu, S. C., Ke, K. Y., Lin, H. T., Tan, Y. C. (2017). Optimization of groundwater quality monitoring network using risk assessment and geostatistic approach. *Water Resources Management*, 31(1), 515-530.
24. Ibrayev, A. Y., Negim, E.-S., Mohamad Ibrahim, M. N. (2025). Algorithmization of placement and evaluation of production wells in mature oil fields using machine learning. *SOCAR Proceedings*, 3, 91-97.
25. Mukhametshin, V. V., Stepanova, R. R., Kotenev, A. Y., et al. (2025). Geological and statistical modeling of the oil recovery process for deposits operating at the limit of economic profitability. *SOCAR Proceedings*, 4, 52-57.
26. Toktoshov, G. Y., Yurgenson, A. N., Migov, D. A. (2025). An approach for the topology design and optimization of a utility network with unreliable elements. *TWMS Journal of Pure and Applied Mathematics*, 16(1), 71-83.
27. Aliyev, E. N., Asadov, M. M., Bogopolskiy, V. O. (2025). Application of new ejector to recover residual oil from depleted formations. *SOCAR Proceedings*, 4, 24-30.
28. Ding, L., Ren, G. W., Yang, Z. L., Li, F. Y. (2025). Spatial twisted central configurations of Newtonian 6-body problem. *TWMS Journal of Pure and Applied Mathematics*, 16(2), 281-293.
29. Hajiyeva, N. S., Mammadova, G. H., Aliev, N. A., Aliev, F. A. (2025). Computational algorithm for the solution of linear quadratic optimal control problem in oscillating systems with liquid dampers. *TWMS Journal of Pure and Applied Mathematics*, 16(2), 294-305.
30. Kang, W. J., Du, Z. R., Zheng, C. X., Hou, T. L. (2025). A linearly decoupled second-order finite difference scheme for ternary Allen-Cahn equations. *Applied and Computational Mathematics*, 24(3), 487-510.
31. Karapinar, E., Sevinik-Adiguzel, R., Aksoy, U., Erhan, I. M. (2025). A new approach to the existence and uniqueness of solutions for a class of nonlinear q-fractional boundary value problems. *Applied and Computational Mathematics*, 24(2), 204-235.
32. Sadygov, M. A., Akhundov, H. S. (2025). Optimal control problem described by the Goursat-Darboux equations. *Applied and Computational Mathematics*, 24(1), 122-146.
33. Zhou, J., Zhang, T. (2025). Almost periodic sequences of complex-valued nonlocal discrete-time SICNNs in the Caputo-Fabrizio sense. *Applied and Computational Mathematics*, 24(3), 350-364.
34. Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In: *Proceedings of the 1968 23rd ACM National Conference*, ACM.
35. Lu, G. Y., Wong, D. W. (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Computers & Geosciences*, 34(9), 1044-1055.