**Technical University of Crete**
**School of Electrical and Computer Engineering**

# DIPLOMA THESIS

# Applications of Machine Learning Methods in Spatial Analysis of Zinc Data

Maria-Konstantina Germanou

## Committee:

Prof. Dionysios Hristopulos (Supervisor)
Prof. Michail G. Lagoudakis
Assistant Prof. Emmanouil Varouchakis (School of MRE)

Chania, December 2024

Πολυτεχνείο Κρήτης
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

# Εφαρμογές Μεθόδων Μηχανικής Μάθησης στη Χωρική Ανάλυση Δεδομένων Ψευδαργύρου

Γερμανού Μαρία-Κωνσταντίνα

Επιτροπή:

Καθ. Διονύσιος Χριστόπουλος (Επιβλέπων)
Καθ. Μιχαήλ Λαγουδάκης
Αν. Καθ. Εμμανουήλ Βαρουχάκης (Σχολή ΜΗΧΟΠ)

Χανιά, Δεκέμβριος 2024

**Dedication**

I dedicate this thesis to my family, for their unwavering support, love, and guidance in every step of my journey. Thank you for always being there to inspire me and encourage me to pursue my dreams.

**Acknowledgments**

I would like to express my sincere gratitude to my supervisor Prof. Dionysios Hristopulos and Assistant Prof. Emmanouil Varouchakis for their invaluable guidance and support throughout this research project. Their insights and encouragement have been instrumental in the completion of this thesis.

I am deeply indebted to my thesis committee members, for their valuable feedback and suggestions which significantly improved the quality of this work.

My heartfelt thanks go to Andrew Pavlides for his invaluable help with my thesis and for aiding my understanding of the field.

Additionally, I am grateful to the technical and administrative staff of the University for their assistance and support throughout the duration of my studies.

Lastly, my heartfelt thanks to all my family and friends who have been a constant source of encouragement and motivation, especially during challenging times.

**Abstract**

This thesis explores the integration of geostatistical methods with advanced machine learning techniques for improving spatial data modeling with emphasis on the prediction of zinc concentration. It also aims to establish the effectiveness of these methods in refining uncertainty quantification and providing insights into complex spatial dependencies. The core methodologies that are employed include Ordinary Kriging (OK), Gaussian Process Regression (GPR), and Self-Organizing Maps (SOMs), selected for their complementary strengths in spatial modeling. OK is a powerful geostatistical baseline that uses variogram modeling to capture spatial correlations, thus providing consistent and interpretable predictions. However, because of the assumptions of linearity it offers limited adaptability for non-linear or heterogeneous spatial structures. GPR introduces a very flexible, non-parametric approach for modeling complex spatial relationships. GPR with an Automatic Relevance Determination Exponential Kernel captures anisotropic spatial dependencies by modeling correlations between locations as a function of their separation distance. The optimized parameters of the kernel through Maximum Likelihood Estimation, ensure accurate predictions of zinc concentrations at unsampled locations and estimates of prediction uncertainty. Moreover, SOMs adds value by providing localized neuron-based data clustering leading to neighborhood-specific kriging predictions and enhancing localized anomaly detection capability (although its overall accuracy is similar to that of OK). These results highlight that OK, GPR, and SOM provide an efficient and flexible framework for analyzing spatial data. The combination of these approaches, harmoniously balances predictive accuracy, uncertainty quantification, and detection of local patterns, thus offering a comprehensive toolkit to address challenging tasks of spatial modeling. In conclusion, this thesis demonstrates how geostatistical techniques in combination with machine learning techniques can push the frontiers of spatial analysis. In addition, refining methodological approaches and optimization of parameters will make such techniques amenable to better decision-making for environmental monitoring and resource management. Future work could also focus on further enhancements of these methods or explore novel hybrid models for answering challenges in spatial data analysis.

## ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία διερευνά την ενοποίηση γεωστατιστικών μεθόδων με σύγχρονες τεχνικές μηχανικής μάθησης για τη βελτίωση της χωρικής μοντελοποίησης δεδομένων, με έμφαση στην πρόβλεψη συγκεντρώσεων ψευδαργύρου. Στόχος είναι η αξιολόγηση της αποτελεσματικότητας αυτών των μεθόδων στη βελτίωση των χωρικών προβλέψεων, την ακριβέστερη ποσοτικοποίηση της αβεβαιότητας και την εξαγωγή πολύτιμων πληροφοριών για σύνθετες χωρικές εξαρτήσεις. Οι βασικές μέθοδοι που χρησιμοποιήθηκαν περιλαμβάνουν το Ordinary Kriging (OK), την παλινδρόμηση με Γκαουσσιανές στοχαστικές διαδικασίες, Gaussian Process Regression (GPR), και τους αυτοοργανούμενους χάρτες, Self-Organizing Maps (SOMs). Οι μέθοδοι αυτές επιλέχθηκαν για τις συμπληρωματικές τους δυνατότητες στη χωρική μοντελοποίηση. Το OK αποτελεί μια ισχυρή βασική μεθοδολογία, η οποία αξιοποιεί τη μοντελοποίηση βαριογραμμάτων για την αποτύπωση χωρικών συσχετισμών και την παροχή συνεπών και ερμηνεύσιμων προβλέψεων. Ωστόσο, οι υποθέσεις γραμμικότητας περιορίζουν την προσαρμοστικότητά της σε δεδομένα με μη γραμμικές ή ετερογενείς χωρικές δομές. Το GPR αποτελεί μια ευέλικτη, μη παραμετρική προσέγγιση, ικανή να μοντελοποιήσει σύνθετες χωρικές σχέσεις. Χρησιμοποιώντας τον εκθετικό ARD πυρήνα, το GPR απέδωσε ανισοτροπικές χωρικές εξαρτήσεις, αποτυπώνοντας τις συσχετίσεις μεταξύ τοποθεσιών ως συνάρτηση της απόστασής τους. Η βελτιστοποίηση των παραμέτρων του πυρήνα μέσω της μεθόδου Μέγιστης Πιθανοφάνειας (MLE) εξασφάλισε αξιόπιστες προβλέψεις σε μη δειγματοληπτημένες τοποθεσίες, με ταυτόχρονη εκτίμηση αβεβαιότητας. Επιπλέον, το SOM προσέθεσε αξία με την τοπική ομαδοποίηση δεδομένων σε νευρωνικούς κόμβους, επιτρέποντας προβλέψεις kriging βασισμένες σε πληροφορίες γειτνίασης. Αυτή η προσέγγιση βελτίωσε την ανίχνευση τοπικών ανωμαλιών και λεπτομερών χωρικών μοτίβων, παρόλο που η συνολική ακρίβεια παρέμεινε συγκρίσιμη με εκείνη του OK. Τα αποτελέσματα υπογραμμίζουν ότι οι μέθοδοι OK, GPR και SOM αποτελούν ένα αποτελεσματικό και ευέλικτο πλαίσιο για την ανάλυση χωρικών δεδομένων. Το OK προσφέρει μια αξιόπιστη προσέγγιση για γενική χωρική παρεμβολή, το GPR ενισχύει την ευελιξία για τη μοντελοποίηση μη γραμμικών τάσεων, και το SOM παρέχει τοπικές βελτιώσεις μέσω ομαδοποίησης. Συνδυαστικά, οι μέθοδοι αυτές εξισορροπούν την ακρίβεια προβλέψεων, την ποσοτικοποίηση αβεβαιότητας και την ανάλυση τοπικών μοτίβων, αποτελώντας ένα ολοκληρωμένο εργαλείο για την αντιμετώπιση απαιτητικών προκλήσεων χωρικής μοντελοποίησης. Συμπερασματικά, η διπλωματική εργασία αποδεικνύει πως η συνδυαστική χρήση γεωστατιστικών μεθόδων και τεχνικών μηχανικής μάθησης μπορεί να επεκτείνει τα όρια της χωρικής ανάλυσης. Η συνεχής βελτίωση μεθοδολογικών προσεγγίσεων και η βελτιστοποίηση παραμέτρων καθιστούν αυτές τις τεχνικές πλέον κατάλληλες για τη λήψη αποφάσεων σε περιβαλλοντική παρακολούθηση και διαχείριση πόρων. Μελλοντική έρευνα μπορεί να επικεντρωθεί στην περαιτέρω εξέλιξη αυτών των μεθόδων ή στην ανάπτυξη νέων υβριδικών μοντέλων που να ανταποκρίνονται στις αυξανόμενες απαιτήσεις χωρικής ανάλυσης δεδομένων.

# Contents

# ListofFigures

# 1 Introduction

The integration of geostatistical methods with machine learning techniques has become a powerful and versatile approach for analyzing and predicting complex geospatial data, especially in mining, environmental monitoring, and resource management. These datasets normally present intricate spatial patterns and dependencies that are crucial for accurate analysis yet challenging to model effectively. Traditional methods such as Ordinary Kriging are well known for their value in spatial prediction based on spatial correlations in the observed data, using variogram modeling to make unbiased predictions. However, despite their effectiveness, OK and similar methods face significant limitations when handling nonlinear relationships and nonstationary processes that are frequently encountered in complex datasets. These challenges have been extensively discussed by Cressie [1] and Goovaerts [2] in their seminal works.

To overcome these drawbacks, Rasmussen and Williams [3] emphasized Gaussian Process Regression (GPR) as a flexible probabilistic framework that is able to capture non-linear spatial patterns while also providing robust uncertainty quantification. Further to this, Hengl, Heuvelink, and Stein [8] discussed the application of GPR in the prediction of soil properties like organic carbon, texture classes, and nutrient distribution. Similarly, Liu and Weisberg [9] focused on the flexibility of GPR in modeling groundwater quality and ocean current variability, demonstrating its applicability for a wide range of geospatial contexts.

Moreover, Kohonen's book [4] introduced the Self-Organizing Maps methodology, which extends geospatial analysis by clustering and visualizing high-dimensional data into interpretable structures. In addition, Bação, Lobo, and Painho [7] applied SOMs to real-world problems in soil variability studies as an alternative to traditional K-Means clustering. Vesanto and Alhoniemi [10] illustrated the capability of SOMs in clustering high-dimensional environmental data, which is useful in identifying sources of groundwater contamination. Liu and Weisberg
citeLiu2011 reviewed various applications of SOMs in meteorology and oceanography, pointing out their usefulness in environmental sciences.

Bridging these methodologies, Hengl et al. [8] proposed regression-kriging frameworks that combine GPR with geostatistical methods, thereby increasing the prediction of soil variables. Such a combination is a good example of how the integration of geostatistics with machine learning will lead to more accurate and robust models. Similarly, in hydrology and water resource management, SOM-based clustering has been instrumental in pollution pattern assessment, as demonstrated by Liu and Weisberg [9] and Vesanto and Alhoniemi [10]. These studies, along

with others, point out that the real potential lies with hybrid approaches that bring the strengths of GPR and SOMs together with more established geostatistical methods such as OK.

This thesis explores the combined application of OK, GPR, and SOM to develop an integrated framework for spatial data analysis. This research aims to enhance the predictive accuracy, improve uncertainty quantification, and deepen the interpretation of spatial patterns by effectively leveraging the complementary strengths of the methods. Concretely, this work investigates the OK's effectiveness for spatial interpolation, the GPR capacity in capturing nonlinear spatial dependencies, and the role of SOM in guiding kriging estimations when data present clustering. The integrated approach not only overcomes the intrinsic drawbacks of traditional geostatistical methods but also provides a wide toolset for geospatial analysis, supporting applications like sustainable resource management, environmental monitoring, and spatial planning.

Furthermore, this study contributes to the development of one generalized framework that goes beyond zinc concentration datasets, showing that such hybrid methods can be extended to scalability and adaptability in different geospatial contexts. This thesis will illustrate the symbiotic benefits of combining OK, GPR, and SOM to provide valuable insights for optimizing spatial data workflows with enhanced resource allocation strategies and effective environmental impacts.The analysis and modeling presented in this research were conducted using MATLAB, leveraging its advanced capabilities in geostatistical analysis, machine learning, and data visualization.The outcome of this study provides a robust foundation for new research in which further development with machine learning models can elevate methodologies in geostatistics in the evolving landscape of data.

# 2 Theoretical Geostatistic Analysis

This chapter provides an in-depth examination of geostatistical methods, beginning with essential visualization techniques and advancing to more sophisticated approaches for spatial analysis.

## 2.1 Random Fields

The concept of random fields extends the mathematical properties of random variables to those cases in which variables are distributed over space. A random field, in this case, is a set of random variables describing the variation in space of some given property over a region [27]. In such a context, a random field can be seen as a multi-dimensional random variable. Since physical quantities at different locations are interdependent, random fields possess different mathematical properties compared with sets of independent random variables. In general, a random field may be denoted as $Z(x)$, where $x$ is the vector of spatial location. One is often interested in the prediction of the value at an unobserved location $x_0$ using the values at the known points in the field.

### *2.1.1 Definitions*

- If the field takes values from a discrete set of numbers, it is called a discrete-valued field. As an example, the pixel values in a black-and-white image are a discrete-valued field since the values have only two possible values.
- If the field values are from a continuous range of real numbers, the field is called a continuous-valued field. A typical example can be the temperature variation over a geographic area or sea level variations due to tides.
- If the variation is defined at discrete points on a grid or lattice, we have a lattice field. The use of lattice fields is common in computational models since they allow numerical methods and comparisons among different geostatistical techniques to be more easily carried out.
- If the variation is defined over a continuous space, we have a continuous space field. Most physical phenomena such as temperature or elevation are modeled as continuous space fields. However, in practice, discretization is required for measurement and computational modeling purposes.

It is important to note that the nature of the field (continuous or discrete) is independent of whether the space where the field is defined is continuous or discrete. For example, a field defined over a continuous space can have discrete values, and conversely, a field defined on a lattice can have continuous values.

### *2.1.2 Basic Concepts of Random Fields*

Random fields describe a variety of spatial dependencies with distinct mathematical properties. The probability distribution of a random field and its parameters extend the concepts that apply to individual random variables. In the context of random fields, $Z(x)$ represents the set of possible values of the field, while a specific realization of the field at a particular location is denoted by $Z(x)$. The variance of the field at location $x$ is represented as $\mathrm{Var}(Z(x))$.

Visualization of random fields plays a crucial role in understanding their properties, as it provides a visual perspective on the field's structure. In the case of a random field that varies in two dimensions (e.g., surface temperature), a three-dimensional plot can be used where the horizontal axes represent spatial coordinates, and the vertical axis represents the field values $Z(x)$. Alternatively, a two-dimensional contour plot can be used, where color or grayscale shading is applied to represent different field values.

The primary objective in spatial interpolation problems is to estimate the value $Z(x_0)$ at an unsampled location $x_0$, based on the known values of the field at neighboring locations $Z(x_1), Z(x_2), \ldots, Z(x_n)$.

## 2.2 Covariance Function

The covariance function is a fundamental tool for measuring the spatial correlation between two points in a random field, $Z(\mathbf{x})$ and $Z(\mathbf{x}+\mathbf{h})$, where $\mathbf{x}$ represents the location and $\mathbf{h}$ is the separation or lag vector [27]. Mathematically, the covariance function $c(\mathbf{x}, \mathbf{x}+\mathbf{h})$ is defined as:

$$c(\mathbf{x}, \mathbf{x}+\mathbf{h}) = E\left[(Z(\mathbf{x}) - E(Z(\mathbf{x})))(Z(\mathbf{x}+\mathbf{h}) - E(Z(\mathbf{x}+\mathbf{h})))\right], \tag{1}$$

where $E$ denotes the expectation operator, which calculates the average (or expected) value of a random variable. The covariance measures how much two random variables co-vary, capturing the strength and direction of their spatial relationship.

Alternatively, the covariance function can be expressed in a simplified form:

$$c(\mathbf{x}, \mathbf{x}+\mathbf{h}) = E[Z(\mathbf{x})Z(\mathbf{x}+\mathbf{h})] - E[Z(\mathbf{x})]E[Z(\mathbf{x}+\mathbf{h})]. \tag{2}$$

This formulation emphasizes the relationship between the expected product of values at two points and their individual expectations, providing insights into the spatial structure of the data [24, 26].

The empirical covariance function, denoted by $c_\delta(\mathbf{h})$, estimates the spatial covariance from a sample of data points. It is calculated as:

$$c_\delta(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} \left[ Z(\mathbf{x}_i) Z(\mathbf{x}_i + \mathbf{h}) - m_\delta(\mathbf{x}_i) m_\delta(\mathbf{h}) \right], \tag{3}$$

where:

- $N(\mathbf{h})$ is the number of data pairs separated by the distance $\mathbf{h}$,
- $Z(\mathbf{x}_i)$ and $Z(\mathbf{h})$ are the observed values at locations $\mathbf{x}_i$ and $\mathbf{x}_i + \mathbf{h}$,
- $m_\delta(\mathbf{x}_i)$ and $m_\delta(\mathbf{x}_i + \mathbf{h})$ are the sample means at these positions.

This empirical form provides a practical way to estimate the covariance for a specific lag distance based on observed data.

### *2.2.1 Spectral Density and Covariance Function*

A valid theoretical covariance model must satisfy the conditions outlined by Bochner's theorem, which ensures the physical plausibility and mathematical consistency of the covariance function [27]. The key conditions are:

1. The power spectral density $\tilde{c}_Z(\mathbf{k})$, which is the Fourier transform of the covariance function, must exist. It is computed as:

$$\tilde{c}_Z(\mathbf{k}) = \int e^{-i\mathbf{k}\cdot\mathbf{r}} c_Z(\mathbf{r})\, d\mathbf{r}, \tag{4}$$

   where $\mathbf{r}$ is the separation vector and $\mathbf{k}$ is the wavenumber vector.

2. The spectral density must be non-negative for all values of $\mathbf{k}$:

$$\tilde{c}_Z(\mathbf{k}) \geq 0, \tag{5}$$

   ensuring that the variance remains non-negative.

3. The integral of the spectral density across the frequency domain must be finite:

$$\int \tilde{c}_Z(\mathbf{k})\, d\mathbf{k} < \infty, \tag{6}$$

   which guarantees that the variance of the random field is bounded.

The covariance function can be recovered from the spectral density using the inverse Fourier transform:

$$c_Z(\mathbf{r}) = \frac{1}{(2\pi)^d} \int e^{-i\mathbf{k}\cdot\mathbf{r}} \tilde{c}_Z(\mathbf{k})\, d\mathbf{k}, \tag{7}$$

where $d$ represents the dimensionality of the space and $\mathbf{k}$ is the wavenumber. Conversely, the spectral density $\tilde{c}_Z(\mathbf{k})$ is obtained from the covariance function using the forward Fourier transform:

$$\tilde{c}_Z(\mathbf{k}) = \int e^{i\mathbf{k}\cdot\mathbf{r}} c_Z(\mathbf{r})\, d\mathbf{r}. \tag{8}$$

These transforms reveal the equivalence between the covariance function in the spatial domain and its spectral representation in the frequency domain. The conditions of Bochner's theorem, as outlined above, ensure that both representations are mathematically valid and physically meaningful.

### *2.2.2 Acceptable Covariance Functions*

The following are some widely used covariance models in geostatistical analysis, where $\mathbf{r}$ represents the vector between two spatial locations [28, 29]:

1. **Exponential Covariance Function:**

$$c_Z(\mathbf{r}) = \sigma_Z^2 \exp\left(-\frac{\mathbf{r}}{\xi}\right), \tag{9}$$

   where:

   - $\sigma_Z^2$ is the variance of the random field,
   - $\xi$ is the range parameter, indicating the distance at which the covariance approaches zero,
   - $\mathbf{r}$ represents the distance between two points in the field.

2. **Gaussian Covariance Function:**

$$c_Z(\mathbf{r}) = \sigma_Z^2 \exp\left(-\frac{\mathbf{r}^2}{\xi^2}\right), \tag{10}$$

   where the parameters are the same as in the exponential covariance function. The key difference is that the Gaussian model decays more rapidly than the exponential model as $\mathbf{r}$ increases.

3. **Spherical Covariance Function:**

$$c_Z(\mathbf{r}) = \sigma_Z^2 \left(1 - 1.5\frac{\mathbf{r}}{\xi} + 0.5\left(\frac{\mathbf{r}}{\xi}\right)^3\right) \theta(\xi - \mathbf{r}), \tag{11}$$

   where:

- $\theta(\xi - \mathbf{r})$ is the Heaviside step function, which ensures that the covariance becomes zero when $\mathbf{r} > \xi$,
- $\sigma_Z^2$ is the variance of the field,
- $\xi$ is the range beyond which the covariance becomes zero.

In these models, the parameter $\xi$ represents the range of spatial influence, defining how far spatial correlation extends between two points in the field. The parameter $\sigma_Z^2$ represents the variance associated with the random field.

### *2.2.3 Key Attributes of Covariance Functions*

- **Smoothness:** The smoothness of the functions produced by a covariance function is a function of its form. For instance, the RBF covariance gives infinitely differentiable functions that are very smooth, while other covariance functions may give functions with different smoothness properties, including the Matérn covariance or Exponential covariance that can produce rough or jagged behavior.
- **Stationarity:** A covariance function is stationary if the covariance between two points depends only on the difference between their positions and not on their absolute positions. More generally, this means that the statistical properties of the covariance function are constant over the input space. The RBF covariance is stationary, while the Linear covariance is a non-stationary covariance function as it does depend on the absolute position of the input points.
- **Periodicity:** Some covariance functions can introduce periodicity and therefore allow the modeling of periodic patterns such as seasonal effects. One covariance function specifically designed for modeling periodic behavior is the Periodic covariance.
- **Scale (Length Scale):** Many covariance functions have hyperparameters that usually control the scale of the covariance. An important one is the length scale, which basically tells how fast the covariance between points decays when increasing the distance between them. A small length scale gives rapidly varying functions, and a large length scale gives smoother, slowly varying functions.
- **Isotropy vs. Anisotropy:** An isotropic covariance depends only on the distance between points and treats all input dimensions equally. In contrast, an anisotropic covariance allows for different length scales in different dimensions, enabling the modeling of input spaces where variability differs across dimensions.

- **Expressiveness:** Complex expressiveness can be gained when more than one covariance function is combined. For example, the sum and product of simpler covariance functions allow modeling of much more complex behaviors. Linear covariances combined with periodic covariance models not only linear trends but also capture periodic fluctuations, making them more adaptable for various and complex data.

## 2.3 Histogram

A histogram is a graphical representation of the distribution of a dataset, commonly used in geostatistical analysis to display how data points are distributed across different value ranges [11]. By dividing the data into intervals, or bins, and counting the number of data points within each bin, an histogram provides an intuitive way to understand the underlying characteristics of the data.

In the context of geostatistics, histograms are frequently used to visualize the distribution of an attribute, such as the concentration of an element in soil samples collected from different locations. Each bar in the histogram corresponds to the frequency of data points falling within a given bin, with the height of the bar representing the count of occurrences.

There are valuable insights into the dataset that are important to consider. Firstly, the histogram helps identify the central value, or range, where most of the data points cluster. In geostatistics, this central tendency may represent the typical concentration of an element in a study area. In addition, the spread of the distribution shows the variability in the dataset. A histogram with widely spread bars indicates high variability, while narrow distributions suggest low variability.The overall shape of the histogram can reveal whether the data follows a normal (bell-shaped) distribution, is skewed to the left or right, or has multiple peaks . Understanding the shape is critical for determining the appropriate geostatistical models to apply.Histograms also help identify outliers—data points that lie far outside the normal range. In geostatistics, outliers may represent anomalies or areas with unusual values that require further investigation.

### *2.3.1 Frequency Calculation*

The frequency $f_i$ of each bin $i$ is calculated as the number of data points where the attribute value falls within the range of that bin. Mathematically, it can be expressed as:

$$f_i = \sum_{j=1}^{n} I(Z(x_j) \in B_i) \tag{12}$$

where $Z(x_j)$ is the value of the attribute at the spatial location $x_j$, $B_i$ is the $i$-th bin, and $I(\cdot)$ is the indicator function that equals 1 if $Z(x_j)$ falls within the bin $B_i$ and 0 if otherwise.

### *2.3.2 Definition*

A histogram, as shown in Figure 1, is a bar chart representing the frequency distribution of data [12]. The x-axis typically represents the variable of interest, divided into intervals or bins, while the y-axis represents the frequency or count of observations falling within each bin.

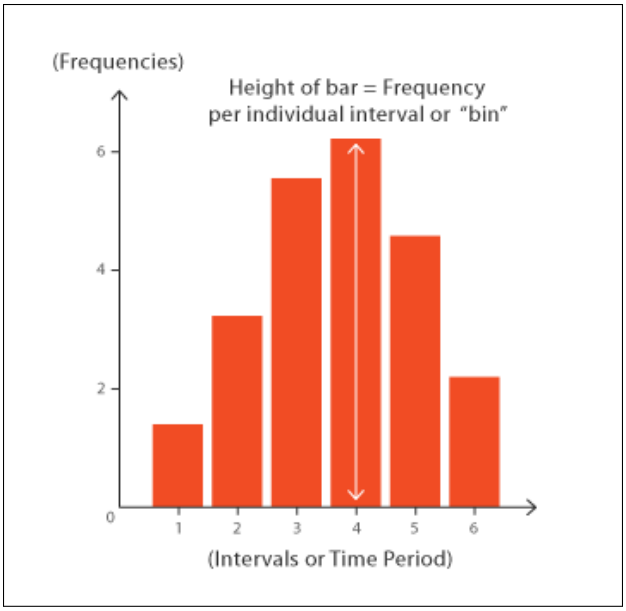


Figure 1: A histogram illustrating the frequency distribution of data points across different bins. The x-axis represents the attribute of interest, divided into distinct intervals (bins), while the y-axis shows the frequency (or count) of data points that fall within each bin. The height of each bar reflects how many observations lie within the corresponding interval, providing a visual summary of the data distribution.

### *2.3.3 Mathematical Representation*

The mathematical representation of a histogram $H(Z(x))$ can be expressed as a piecewise function:

$$H(Z(x)) = \begin{cases} f_i & \text{if } Z(x) \in B_i \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

where $f_i$ is the frequency of observations, when $Z(x_j)$ falls within bin $B_i$.

## 2.4 Cumulative Distribution Function (CDF)

In probability theory, the **cumulative distribution function (CDF)** of a real-valued random variable $Z$ is defined as the function:

$$F_Z(x) = P(Z \leq x) \tag{14}$$

This represents the probability that the random variable $Z$ takes on a value less than or equal to $x$ [13].

For any two real numbers $a$ and $b$, where $a < b$, the probability that $Z$ lies in the interval $(a, b]$ is given by:

$$P(a < Z \leq b) = F_Z(b) - F_Z(a) \tag{15}$$

This formula provides the probability that the random variable $Z$ falls within a specific range. The "less than or equal to" sign in the definition of the CDF is a convention, and it is particularly important when working with discrete random variables. This convention is crucial for the proper interpretation of tables of binomial and Poisson distributions, and for applying formulas such as Paul Lévy's inversion formula for the characteristic function [15].

- For a general CDF, a capital $F$ is used, while a lower-case $f$ is used for probability density functions (PDF) or probability mass functions (PMF).
- Specific distributions may have their own notation, such as the normal distribution, which uses $\Phi$ for the CDF and $\phi$ for the PDF.

### *2.4.1 Relationship Between CDF and PDF*

For continuous random variables, the probability density function $f_Z(x)$ can be derived from the cumulative distribution function by differentiation, using the Fundamental Theorem of Calculus. If the CDF $F_Z(x)$ is differentiable, the PDF is given by:

$$f_Z(x) = \frac{d}{dx} F_Z(x) \tag{16}$$

Conversely, the CDF can be obtained by integrating the PDF:

$$F_Z(x) = \int_{-\infty}^{x} f_Z(t)\, dt \tag{17}$$

### *2.4.2 Discrete Components in the CDF*

If the random variable $Z$ has a discrete component at a specific value $b$, the probability that $Z = b$ is given by the difference between the value of the CDF at $b$ and the left-hand limit of the

CDF at $b$:

$$P(Z = b) = F_Z(b) - \lim_{x \to b^-} F_Z(x) \tag{18}$$

If the CDF $F_Z(x)$ is continuous at $b$, this difference is zero, indicating that there is no discrete component at $b$ [14].

## 2.5 Empirical Cumulative Distribution Function (ECDF)

The Empirical Cumulative Distribution Function (ECDF) is a non-parametric estimator that approximates the CDF using observed data [16]. For a dataset $Z = \{z_1, z_2, \ldots, z_N\}$, the ECDF $F_N(x)$ is defined as:

$$F_N(x) = \frac{1}{N} \sum_{i=1}^{N} 1(z_i \leq x), \tag{19}$$

where $1(z_i \leq x)$ is the indicator function, which takes the value 1 if $z_i \leq x$, and 0 otherwise. The ECDF provides a stepwise approximation of the CDF based on the observed data points, showing the proportion of values less than or equal to a given value $x$.

### *2.5.1 Deriving the PDF from the ECDF*

Since the ECDF is a stepwise function, the PDF cannot be derived directly as a smooth function. However, the PDF can be approximated using finite differences between consecutive values of the ECDF. For two adjacent points $x_i$ and $x_{i+1}$, the PDF can be estimated as:

$$p(x) \approx \frac{F_N(x_{i+1}) - F_N(x_i)}{x_{i+1} - x_i} \tag{20}$$

This finite difference provides an estimate of how the probability is distributed between these two values. To ensure that the PDF represents a valid probability distribution, it must be normalized so that the total area under the curve integrates to 1:

$$\int_{-\infty}^{\infty} p(x)\, dx = 1 \tag{21}$$

In summary, the ECDF is a useful tool for estimating the CDF based on observed data, and the PDF can be approximated by differentiating the ECDF using finite differences. This approach is valuable for analyzing the distribution of geospatial data and understanding its underlying structure.

## 2.6 Normal Probability Plot

The normal probability plot is a graphical tool used to assess whether a set of data is approximately normally distributed [17]. The primary purpose of the normal probability plot is to

visually identify any departures from normality, such as outliers, skewness, kurtosis, or mixtures of distributions. This method is frequently used in residual analysis, model diagnostics, and goodness-of-fit testing.

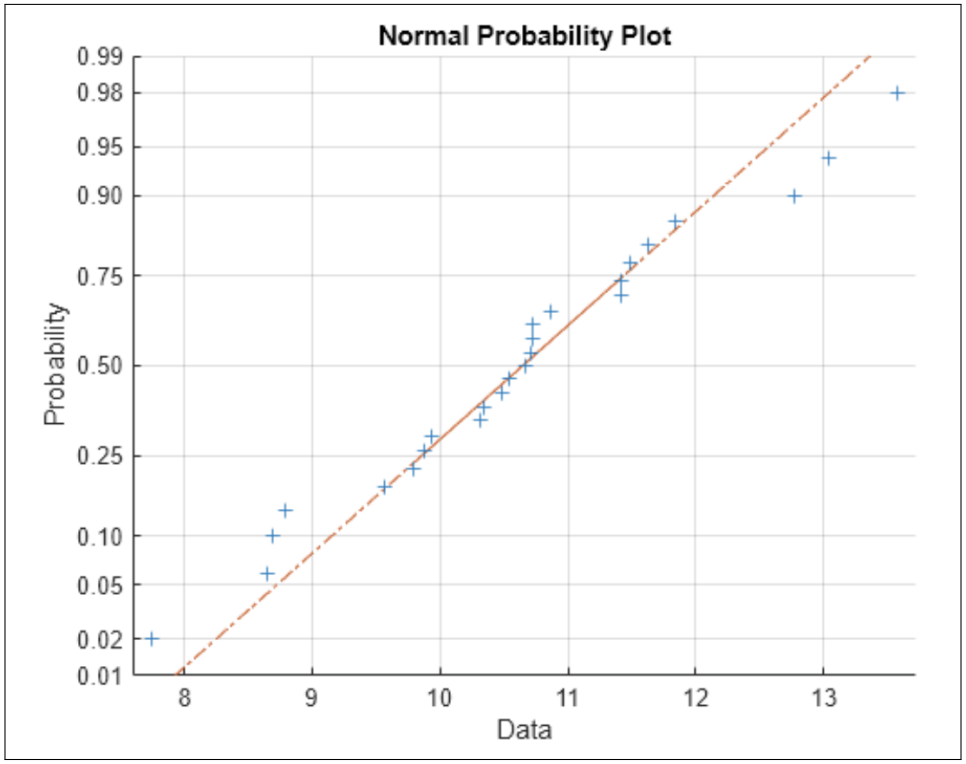


Figure 2: Normal Probability Plot: This plot visualizes the comparison of sorted data against the theoretical quantiles of a normal distribution. The closer the points align with the reference line, the closer the data follow a normal distribution. Deviations from the straight line highlight departures from normality, indicating potential skewness or outliers [20].

In Figure 2, the normal probability plot displays the relationship between the sorted data (on the x-axis) and the probability (on the y-axis). The points that closely follow the reference line indicate that the data approximates a normal distribution, while any deviations suggest departures from normality.

### *2.6.1 Construction of the Normal Probability Plot*

A normal probability plot is constructed by plotting the sorted sample data against the theoretical quantiles of the standard normal distribution. If the data follow a normal distribution, the plotted points will form a roughly straight line [19]. Deviations from this straight line suggest departures from normality.

To construct a normal probability plot, the raw data is first arranged in ascending order. The next step is to calculate the theoretical quantiles by applying the inverse of the standard normal cumulative distribution function (CDF) to equally spaced points in the interval (0, 1). These

quantiles correspond to the means or medians of the order statistics from the normal distribution. Once calculated, the sorted data values are plotted against the theoretical quantiles. If the data are normally distributed, the points will align closely along a straight line. A reference line can be added to assess the degree of normality, which may be fitted using least squares or by setting the slope to 1 and using the mean of the data as the intercept.

### *2.6.2 Interpretation of the Normal Probability Plot*

The extent to which the points deviate from a straight line in a normal probability plot indicates the degree of departure from normality. Points that closely follow a straight line suggest the data are approximately normally distributed. A concave curve suggests right skewness, while a convex curve indicates left skewness. An S-shape may indicate kurtosis issues, where a steep S-shape suggests heavy tails (leptokurtic) and a flattened S-shape suggests light tails (platykurtic). Points far from the pattern represent outliers.

### *2.6.3 Rankit and Quantile Calculation*

The vertical or horizontal axis (depending on the orientation) of the normal probability plot is based on rankits, which are approximations of the means or medians of the order statistics from the normal distribution [18]. The rankit for the $i$-th order statistic, $z_i$, can be calculated using the following formula:

$$z_i = F^{-1}\left(\frac{i-a}{n+1-2a}\right), \tag{22}$$

where $F^{-1}$ is the inverse of the standard normal cumulative distribution function (CDF), $i$ is the rank of the data point, $n$ is the total number of points, and $a$ is a constant determined as follows:

$$a = \begin{cases} \frac{3}{8} & \text{if } n \leq 10, \\ 0.5 & \text{if } n > 10. \end{cases} \tag{23}$$

The rankits correspond to the quantiles of the normal distribution and are used to assess whether the data points deviate significantly from normality. If the data are consistent with a sample from a normal distribution, the points should lie close to a straight line. Any deviations from this line indicate departures from normality.

## 2.7 Voronoi Polygons-Spatial Foundations in Geostatistics

In the realm of geostatistics, Voronoi polygons emerge as fundamental spatial constructs, providing a geometric framework to partition spatial domains into contiguous, non-overlapping regions . Each polygon represents the territorial extent associated with a specific data point, delineating its spatial influence within the dataset [21]. Mathematically, the Voronoi polygon $V_i$ linked to the $i$-th data point $(x_i, y_i)$ encompasses all points $(x, y)$ in the domain that are closer to $(x_i, y_i)$ than to any other data point. This partitioning yields a tessellation of the spatial domain into discrete regions, as seen in Figure 3, offering a nuanced representation of spatial relationships and influences.

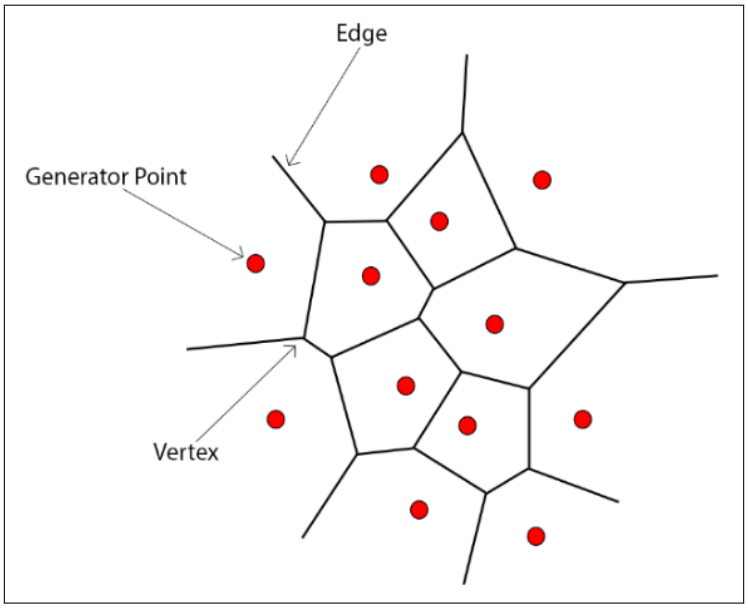


Figure 3: Voronoi polygons with the edges of the polygons defining the boundaries of regions that are closest to their respective generator points [22].

Beyond their role in spatial partitioning, Voronoi polygons find diverse applications in geostatistical analysis [23] .

1. **Spatial Interpolation**: Voronoi-based interpolation techniques leverage the known values at data points within each polygon to estimate values at unsampled locations. By inferring spatial patterns from neighboring data points, these methods enable the creation of continuous surfaces, facilitating spatial analysis and visualization.

2. **Area Weighting**: The area of each Voronoi polygon can be considered a measure of the spatial influence, with larger polygons showing wider influence and vice versa. In this respect, incorporating polygon areas as weights allows geostatistical analyses to take into account spatial variability in the accuracy of spatial predictions and assessments.

3. **Spatial Connectivity and Proximity Analysis**: Voronoi polygons provide valuable information about spatial connectivity and proximity of the dataset. The visualization of adjacency of polygons and clustering helps in the identification of spatial patterns, gradients, and trends, thus providing analysts with the ability to decipher geologic structures and environmental dynamics.

Applications range from spatial weighting, interpolation, and pattern recognition, and are thus very important in spatial datasets when understanding and interpretation are necessary in environmental and geological sciences.

## 2.8 Variography

Variography is the fundamental tool of geostatistics, which aims to measure and understand the spatial variability of a phenomenon in space over a geographic area. The main concepts, methods, and mathematical developments of variography are presented in this chapter
citechiles.

Variography is the study of spatial dependence or correlation between data values across a study area. It tries to show how the values of a variable at different locations relate to each other, which forms the basis for understanding key spatial patterns and structures.

### *2.8.1 The Variogram Function*

The variogram function, denoted $\gamma(\mathbf{h})$, where $\mathbf{h}$ represents the lag distance, measures the variability between pairs of data points as a function of their separation. It quantifies spatial correlation, with smaller $\gamma(\mathbf{h})$ values indicating stronger correlation and larger values reflecting weaker correlation. The variogram provides essential information for spatial interpolation methods like kriging by describing how spatial relationships change with distance [31]. The variogram function is calculated using the formula:

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} (Z(\mathbf{x}_i + \mathbf{h}) - Z(\mathbf{x}_i))^2, \tag{24}$$

where:

- $N(\mathbf{h})$ is the number of data point pairs at lag distance $\mathbf{h}$,
- $Z(\mathbf{x}_i)$ and $Z(\mathbf{x}_i + \mathbf{h})$ are the values of the variable at locations $\mathbf{x}_i$ and $\mathbf{x}_i + \mathbf{h}$ respectively.

If the random field $Z(\mathbf{x})$ is statistically homogeneous, the variogram is directly related to the covariance function by the following equation:

$$\gamma_Z(\mathbf{h}) = \sigma_Z^2 - c_Z(\mathbf{h}). \tag{25}$$

From the above equation, it follows that the variogram asymptotically approaches the value of the variance as $\mathbf{h}$ increases. Additionally, for statistically homogeneous fields, the variogram contains the same information as the covariance function.

### *2.8.2 Lag Distance and Directional Variography*

In Figure 4, the variogram illustrates the relationship between the **semivariance** (vertical axis) and the **lag distance** (horizontal axis) between paired data points in a geospatial dataset [32]. This specific variogram follows an **exponential model**, which is commonly used in geostatistics to describe how spatial correlation decreases with increasing distance.

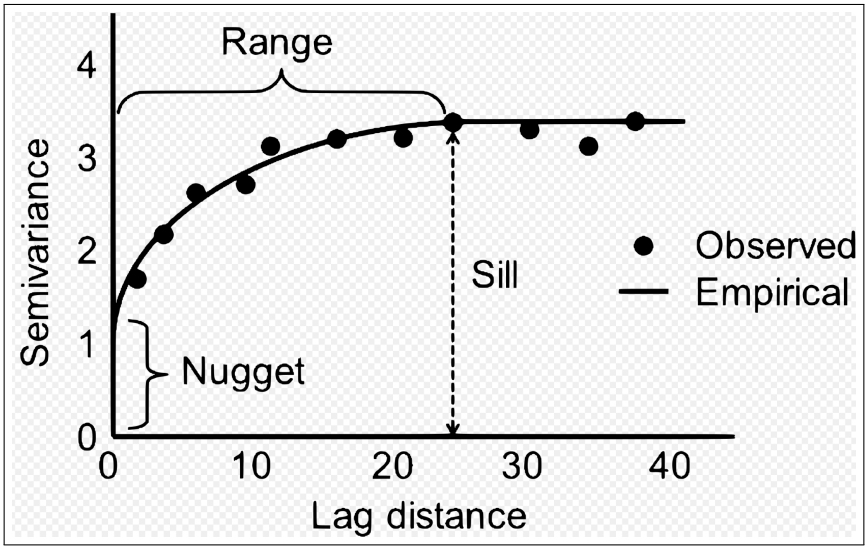


Figure 4: The variogram illustrates the relationship between semivariance and lag distance in a geospatial dataset. The black dots represent the observed semivariance at various lag distances, while the smooth curve represents the fitted exponential model. The curve shows how semivariance increases with distance, indicating a reduction in spatial correlation as the lag increases. Beyond a certain lag distance (the range), the semivariance levels off, reaching the sill, which signifies the limit of spatial correlation. The nugget effect can also be observed as the initial rise in semivariance at small lag distances due to measurement error or micro-scale variability [33].

- **Semivariance:** The semivariance measures the **spatial dependence** between data points, quantifying how much two points differ in value as a function of the distance between them. As the distance (lag) increases, the semivariance typically increases, indicating that data points farther apart are less similar.

- **Nugget:** The nugget effect represents the semivariance at very small lag distances, where measurement error and small-scale spatial variation can cause an initial jump in semivariance. In an exponential variogram, this jump occurs rapidly as the semivariance rises from the nugget.
- **Range:** The range is the distance over which spatial correlation between data points exists. In the exponential model, the spatial correlation decreases exponentially with distance. After the range, the semivariance levels off, meaning that points separated by distances larger than the range are no longer spatially correlated.
- **Sill:** The sill represents the point at which the semivariance reaches its maximum value, after which no further increase occurs. Beyond the sill, data points are considered independent of each other in terms of spatial correlation.

**Exponential Model of the Variogram:** In this example, we understand that the variogram follows an **exponential model** because the semivariance increases sharply at small lag distances and then gradually approaches the sill. This behavior is characteristic of the exponential variogram, where spatial correlation diminishes rapidly over short distances but persists for some range before leveling off. The **black dots** in Figure 4 represent the observed semivariance at different lag distances, while the fitted curve is the empirical model capturing how semivariance changes with increasing distance.

In an exponential variogram, the semivariance does not reach the sill as abruptly as in other models (e.g., the spherical model), but instead follows a smooth, gradual approach, reflecting how spatial dependence decays exponentially with distance.

## 2.9 Isotropic and Anisotropic Variogram Models

In geostatistics, variograms are essential tools for quantifying spatial correlation and variability in data [26, 27]. The variogram measures how spatial similarity decreases with increasing distance between sample locations. In isotropic cases, spatial correlation is assumed to be uniform in all directions. However, anisotropy occurs when spatial dependence varies with direction, and anisotropic variograms can capture this behavior.

### *2.9.1 Isotropic Variogram Models*

The following are commonly used isotropic models, where spatial variability is assumed to be uniform across all directions [26]:

1. **Exponential Model**: The exponential model describes a situation where spatial dependence decreases rapidly at first and then approaches the sill asymptotically. The model is given by:

$$\gamma_X(\mathbf{r}) = \sigma_X^2 \left[1 - \exp\left(-\frac{\|\mathbf{r}\|}{\xi}\right)\right], \tag{26}$$

where $\sigma_X^2$ represents the sill, $\|\mathbf{r}\|$ is the distance (magnitude of the vector $\mathbf{r}$) between points, and $\xi$ is the range parameter, controlling the distance over which spatial correlation persists.

2. **Gaussian Model**: The Gaussian model is characterized by a smooth approach to the sill, making it suitable for processes where spatial continuity is strong and variability increases more gradually with distance. It is expressed as:

$$\gamma_X(\mathbf{r}) = \sigma_X^2 \left[1 - \exp\left(-\frac{\|\mathbf{r}\|^2}{\xi^2}\right)\right]. \tag{27}$$

This model is used in situations where the decay in spatial correlation occurs more slowly over distance.

3. **Spherical Model**: The spherical model reaches the sill at a finite distance and is frequently used in practice for spatial modeling. It is described by:

$$\gamma_X(\mathbf{r}) = \begin{cases} \sigma_X^2 \left[1.5\left(\frac{\|\mathbf{r}\|}{\xi}\right) - 0.5\left(\frac{\|\mathbf{r}\|}{\xi}\right)^3\right], & \|\mathbf{r}\| < \xi \\ \sigma_X^2, & \|\mathbf{r}\| \geq \xi, \end{cases} \tag{28}$$

where $\xi$ is the range, beyond which no further spatial correlation is observed.

4. **Power (Algebraic) Model**: The power model is used for cases where variability continues to increase with distance without reaching a sill. It is typically applied to model long-range dependence, and is given by:

$$\gamma_X(\mathbf{r}) = \alpha \|\mathbf{r}\|^{2H}, \quad 0 < H < 1,\ \alpha > 0, \tag{29}$$

where the parameter $H$ governs the rate of increase, while $\alpha$ controls the magnitude of variability.

5. **Nugget Effect**: The nugget effect accounts for small-scale variability or measurement errors that are not captured by the model at the sample intervals. The nugget variogram is expressed as:

$$\gamma_X(\mathbf{r}) = \begin{cases} 0, & \mathbf{r} = 0 \\ c_0, & \mathbf{r} \neq 0, \end{cases} \tag{30}$$

where $c_0$ represents the jump in the variogram for small distances greater than zero.

### *2.9.2 Anisotropy in Variography*

Many real-world spatial datasets exhibit **anisotropy**, where the spatial correlation depends on the direction of the measurements [24]. Anisotropy can be captured by calculating directional variograms, which provide insights into how spatial variability changes with direction. Three main types of anisotropy are commonly recognized [25]:

- **Geometric Anisotropy**: the range of the variogram, $\xi$, changes depending on the direction. This means that in some directions, the spatial correlation extends over longer distances, while in others, the correlation decays more quickly. This type of anisotropy is common in many natural systems, such as geological formations, where certain features create directional dependencies in the data. In this case, different ranges are assigned to different directions, indicating greater spatial continuity in some directions over others.
- **Zonal Anisotropy**: occurs when the sill, $\sigma_X^2$, changes depending on the direction. In this case, the variogram reaches different sill values depending on the direction of the measurements. Zonal anisotropy suggests that the magnitude of spatial variability differs across directions. For example, there may be higher variability in one direction due to environmental or geological factors, while another direction may exhibit less variability. This directional variation is reflected in the different sill values for each direction.
- **Isopleths (Constant Level Curves)**: represents a set of points in space where the variogram value is constant. Isopleths provide a visual representation of the spatial structure described by the variogram. In isotropic conditions, isopleths are circular, as the spatial correlation decays uniformly in all directions. However, in anisotropic cases, the isopleths take on elliptical or more complex shapes, reflecting the directional dependence of spatial variability. Analyzing isopleths is essential for identifying and modeling anisotropy in geostatistical data.

#### *Geometric Anisotropic Variogram Models in 2D*

The following models are frequently used to describe geometric anisotropy in two-dimensional space, where spatial correlation decays differently along distinct axes:

- **Exponential Model with Anisotropy**: Assumes exponential decay of spatial correlation with direction-specific ranges $(\xi_x, \xi_y)$:

$$\gamma_X(r) = \sigma_X^2 \left[1 - \exp\left(-\frac{r_x^2}{\xi_x^2} - \frac{r_y^2}{\xi_y^2}\right)\right]. \tag{31}$$

- **Gaussian Model with Anisotropy**: Assumes smoother decay of spatial correlation with direction-specific ranges:

$$\gamma_X(r) = \sigma_X^2 \left[1 - \exp\left(-\frac{r_x^2}{\xi_x^2} - \frac{r_y^2}{\xi_y^2}\right)\right]. \tag{32}$$

By using both isotropic and anisotropic models, geostatistical analyses can effectively capture the underlying spatial structure, leading to improved accuracy in spatial predictions. Accounting for anisotropy through geometric and zonal models ensures that spatial variability is adequately represented, supporting better decision-making in areas such as environmental monitoring, geological exploration, and resource management.

## 2.10 Kriging

Kriging is a statistical method used for spatial interpolation, where linear combinations of observed values at neighboring locations are used to estimate values at unsampled locations within a random field. The core principle of Kriging relies on spatial dependence, which assumes that the value at a given location is influenced by the values at surrounding locations [34]. To apply Kriging effectively, a spatial dependence model must be defined, typically represented using a variogram.

In the Kriging framework, the value at an unsampled location $x_0$ is estimated using a linear combination of neighboring observed values $Z(x_i)$, with the weights determined by the variogram model. The neighborhood of points, denoted $\omega(x_0)$, is defined around the point $x_0$, where the extent of the neighborhood is influenced by the spatial correlation of the data.

The estimated value at point $x_0$ is given by:

$$\hat{Z}(x_0) = m_Z(x_0) + \sum_{i=1}^{n(x_0)} \lambda_i \left[Z(x_i) - m_Z(x_i)\right], \tag{33}$$

where:

- $m_Z(x_0)$ is the mean of the random field at the location $x_0$,
- $Z(x_i)$ represents the observed value at the neighboring location $x_i$,
- $m_Z(x_i)$ is the mean of the random field at the neighboring location $x_i$,
- $\lambda_i$ are the weights assigned to each observation, and
- $n(x_0)$ denotes the number of observations within the neighborhood around $x_0$.

The estimation error $\epsilon(x_0)$ is defined as the difference between the actual value of the random field at point $x_0$ and the estimated value:

$$\epsilon(x_0) = Z(x_0) - \hat{Z}(x_0). \tag{34}$$

In Kriging, the weights $\lambda_i$ are optimized to minimize the variance of the estimation error. This leads to an optimal estimate $\hat{Z}(x_0)$, which represents the best linear unbiased prediction (BLUP) for the value of the random field at $x_0$.

### *2.10.1 Ordinary Kriging*

Ordinary Kriging (OK) is one of the most common geostatistical methods for interpolating values at unsampled locations, based on the spatial relationship of the observed points [35]. It ensures minimum prediction variance along with unbiasedness of the estimates. OK expresses the prediction at an unknown location $x_0$ as a weighted sum of the observed values at neighboring locations. The sum of weights is constrained to equal one, ensuring that the estimate is unbiased:

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i), \tag{35}$$

where:

- $\lambda_i$ are the weights assigned to each observed value $Z(x_i)$,
- $n$ is the number of neighboring points used in the estimation.

The weights $\lambda_i$ must satisfy the constraint:

$$\sum_{i=1}^{n} \lambda_i = 1.$$

### *2.10.2 Ordinary Kriging Linear System*

The system of equations for determining the weights $\lambda_i$ is derived by minimizing the Kriging variance, subject to the unbiasedness constraint. This results in the following system of equations for Ordinary Kriging:

$$\begin{pmatrix} \gamma(x_1 - x_1) & \cdots & \gamma(x_1 - x_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(x_n - x_1) & \cdots & \gamma(x_n - x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(x_1 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \\ 1 \end{pmatrix}, \tag{36}$$

where:

- $\gamma(x_i - x_j)$ is the variogram value between locations $x_i$ and $x_j$,
- $\mu$ is the Lagrange multiplier ensuring that the weights sum to one.

### *2.10.3 Covariance-Based Representation of the Linear System*

The linear system used to calculate the optimal weights in Ordinary Kriging can also be expressed in terms of covariance:

$$\begin{pmatrix} \sigma_X^2 & C(x_1 - x_2) & \cdots & C(x_1 - x_n) & 1 \\ C(x_2 - x_1) & \sigma_X^2 & \cdots & C(x_2 - x_n) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C(x_n - x_1) & C(x_n - x_2) & \cdots & \sigma_X^2 & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} C(x_1 - x_0) \\ C(x_2 - x_0) \\ \vdots \\ C(x_n - x_0) \\ 1 \end{pmatrix} \tag{37}$$

where:

- $\sigma_X^2$ is the variance of the random field at the observed points.
- $C(x_i - x_j)$ represents the covariance between two observed points $x_i$ and $x_j$.
- $C(x_i - x_0)$ denotes the covariance between each observed point $x_i$ and the unsampled point $x_0$.
- $\lambda_i$ are the optimal weights assigned to each observation $Z(x_i)$.
- $\mu$ is the Lagrange multiplier, which ensures that the weights sum to one.

**Matrix Formulation**

This system can also be written in matrix form:

$$\mathbf{C}\boldsymbol{\Lambda} = \mathbf{C_u} \tag{38}$$

where:

- $\mathbf{C}$ is the covariance matrix of the observed values, with elements $C(x_i - x_j)$.
- $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n, \mu)^T$ is the vector of weights and the Lagrange multiplier.
- $\mathbf{C_u} = (C(x_1 - x_0), C(x_2 - x_0), \ldots, C(x_n - x_0), 1)^T$ is the vector of covariances between the observed points and the unsampled location $x_0$.

The optimal weights $\lambda_i$ can be obtained by solving the equation:

$$\boldsymbol{\Lambda} = \mathbf{C}^{-1}\mathbf{C_u} \tag{39}$$

### *2.10.4 Kriging Variance*

The Kriging variance, which is used to measure the uncertainty of the prediction at the unsampled location $x_0$, is given by:

$$\sigma^2(Z(x_0)) = -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(x_i - x_j) + 2\sum_{i=1}^{n}\lambda_i\gamma(x_i - x_0) \tag{40}$$

or, alternatively, using covariance:

$$\sigma^2(x_0) = \sigma_X^2 - \sum_{i=1}^{n}\lambda_i C(x_i - x_0) - \mu \tag{41}$$

where:

- $\sigma_X^2$ is the variance of the observed values.
- $\lambda_i$ are the weights obtained by solving the system of equations.
- $C(x_i - x_0)$ is the covariance between each observed point $x_i$ and the unsampled location $x_0$.
- $\mu$ is the Lagrange multiplier.

The Lagrange multiplier $\mu$, which is typically negative, adjusts the Kriging variance to account for any bias caused by the spatial correlation between the observed sample values. In Ordinary Kriging, this adjustment leads to smaller uncertainty compared to Simple Kriging, as the Lagrange multiplier compensates for the correlation among the observations.

Ordinary Kriging provides the Best Linear Unbiased Prediction (BLUP) by minimizing the prediction variance while ensuring unbiasedness in the estimates. It is widely used in environmental modeling, resource estimation, and other geostatistical applications, relying on stationarity assumptions for the random field.

## 2.11 Trending and Detrending

A trend is a large-scale pattern or regularity in the data that is not related to the local variations of interest in a geostatistical analysis. For example, if zinc concentration data systematically increases from south to north, this gradual change represents a trend.

Detrending is an essential step in geostatistical analysis to isolate the spatial structure of residuals, which capture small-scale variations in the data. The presence of a trend can obscure the true spatial correlation, making it difficult to compute and analyze the experimental variogram. Below is an explanation of the detrending process and its importance [37].

### 2.11.1 Detrending Process

The detrending process involves estimating the trend and subtracting it from the original data to focus on the residuals. Linear regression is commonly used to estimate the trend.

1. **Trend Estimation:** A linear regression is performed to estimate the trend in the data, which describes the relationship between the spatial coordinates (e.g., Easting and Northing) and the zinc concentrations (in logarithmic scale).
2. **Trend Removal:** Once the trend is estimated, it is subtracted from the original zinc concentration values, leaving the residuals, which represent local variations without the large-scale trend.

Let the zinc concentration $Z(\mathbf{x})$ depend on the spatial coordinates, where $\mathbf{x} = (X, Y)$ is a vector representing the spatial location. In this context, $X$ and $Y$ refer to the individual spatial coordinates (e.g., east-west and north-south). The trend in zinc concentration, denoted as $m(\mathbf{x})$, can be modeled using a linear regression as follows:

$$m(\mathbf{x}) = \beta_0 + \beta_1 X + \beta_2 Y, \tag{42}$$

where:

- $\beta_0$ is the intercept (baseline concentration),
- $\beta_1$ is the coefficient associated with the spatial coordinate $X$,
- $\beta_2$ is the coefficient associated with the spatial coordinate $Y$.

### 2.11.2 Detrending and Reintroduction of Trend

To isolate the local spatial variability, the large-scale trend $m(x)$ is removed from the observed zinc concentrations $Z(x)$. This process, called detrending, results in residuals $\hat{R}(x)$, which represent the deviations from the overall trend and capture the small-scale spatial variations:

$$\hat{R}(x) = Z(x) - m(x), \tag{43}$$

where $Z(x)$ is the observed zinc concentration at the location $x = (X, Y)$, and $\hat{R}(x)$ represents the data fluctuation. These residuals are then used for spatial interpolation methods such as Ordinary Kriging (OK) or Gaussian Process Regression (GPR) to model the local fluctuations in zinc concentration.

After the spatial interpolation is completed, the trend $m(x)$ is reintroduced to the interpolated residuals $\hat{R}(x)$ to obtain the final zinc concentration predictions:

$$Z(x) = \hat{R}(x) + m(x). \tag{44}$$

This ensures that both the global trend and local variations are incorporated into the final predictions.

The trend can affect the calculation of the experimental variogram, leading to incorrect estimates of spatial correlation. By removing the trend, we can examine the local variations more clearly. In addition, kriging is a spatial interpolation method that assumes the data is stationary, meaning it does not contain large-scale trends. So, detrending makes the data more suitable for kriging, and the reintroduction of the trend ensures that both local and large-scale variations are preserved. Finally, detrending helps isolate the small-scale variations, improving the accuracy of predictions. Reintroducing the trend after interpolation ensures that predictions are comprehensive and account for both local variability and large-scale patterns in the data.

## 2.12 Error Metrics

Error metrics are essential tools for evaluating the accuracy of spatial models and predictions [38].

### *2.12.1 Mean Error (ME)*

The Mean Error (ME) quantifies the bias in the predictions, representing the average difference between the predicted values and the actual observed values [40]. It is mathematically expressed as:

$$ME = \frac{1}{N} \sum_{i=1}^{N} \left[ \hat{Z}(\mathbf{x}_i) - Z(\mathbf{x}_i) \right], \tag{45}$$

where:

- $Z(\mathbf{x}_i)$ is the observed value at location $\mathbf{x}_i$,
- $\hat{Z}(\mathbf{x}_i)$ is the predicted value at location $\mathbf{x}_i$,
- $N$ is the total number of observations.

The closer the Mean Error is to zero, the less biased the predictions.

### 2.12.2 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) measures the average magnitude of the absolute difference between the predicted values and the observed values, irrespective of the direction of the errors [36]. It is calculated as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{Z}(\mathbf{x}_i) - Z(\mathbf{x}_i)\right|. \tag{46}$$

Lower MAE values indicate higher predictive accuracy.

### 2.12.3 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) provides a measure of the model's accuracy, with a higher emphasis on larger errors [39]. RMSE is computed using the following formula:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{Z}(\mathbf{x}_i) - Z(\mathbf{x}_i)\right)^2} \tag{47}$$

RMSE gives more weight to large errors compared to MAE, making it more sensitive to outliers.

### 2.12.4 Pearson Correlation Coefficient

The Pearson Correlation Coefficient, denoted as $\rho$, measures the strength and direction of the linear relationship between the predicted and observed values [41]. It is calculated as:

$$\rho = \frac{\sum_{i=1}^{N}\left(Z(\mathbf{x}_i) - \bar{Z}\right)\left(\hat{Z}(\mathbf{x}_i) - \bar{\hat{Z}}\right)}{\sqrt{\sum_{i=1}^{N}\left(Z(\mathbf{x}_i) - \bar{Z}\right)^2 \sum_{i=1}^{N}\left(\hat{Z}(\mathbf{x}_i) - \bar{\hat{Z}}\right)^2}}, \tag{48}$$

where:

- $Z(\mathbf{x}_i)$ represents the observed value at location $\mathbf{x}_i$,
- $\hat{Z}(\mathbf{x}_i)$ is the predicted value at location $\mathbf{x}_i$,
- $\bar{Z}$ is the mean of the observed values,
- $\bar{\hat{Z}}$ is the mean of the predicted values.

The Pearson Correlation Coefficient $\rho$ ranges between -1 and 1, providing insight into the direction and strength of the linear relationship:

- A value of $\rho = 1$ indicates a perfect positive linear relationship, where increases in one variable correspond to proportional increases in the other.

- A value of $\rho = -1$ indicates a perfect negative linear relationship, where increases in one variable correspond to proportional decreases in the other.
- A value of $\rho = 0$ indicates no linear relationship between the predicted and observed values.

Values of $\rho$ closer to 1 or -1 indicate a strong linear relationship, whereas values close to 0 indicate a weak or no linear association. In practice, a strong linear relationship is usually found within the range of 0.5 to 1 or -0.5 to -1.

In the context of geostatistics, the Pearson correlation coefficient indicates how well the model captures spatial variability by comparing the predicted spatial trends with the observed data. A high correlation suggests that the model effectively captures the spatial patterns, while a low correlation may indicate discrepancies between the predicted and observed values.

# 3 Theoretical Geostatistical Analysis Using Machine Learning Methods

This chapter explores the theoretical basis of geostatistical methods and their enhancement by incorporating machine learning techniques. The integration of geostatistical approaches, such as Ordinary Kriging, with advanced machine learning models like Gaussian Process Regression and Self-Organizing Maps, develops a more robust framework for spatial data analysis. This integration will, therefore, help in addressing the increasing complexity of spatial datasets for better predictive performance and comprehensive insights into spatial relationships.

## 3.1 Gaussian Process Regression

### *3.1.1 Definition of a Gaussian Process*

A **Gaussian Process (GP)** is a collection of random variables, any finite subset of which follows a joint multivariate Gaussian distribution. Formally, a GP $\{f(\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$, indexed by an input space $\mathcal{X}$, is characterized by its mean function $m(\mathbf{x})$ and covariance function $c(\mathbf{x}, \mathbf{x}')$, and can be written as:

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), c(\mathbf{x}, \mathbf{x}')\right), \tag{49}$$

where:

- $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ is the mean function, which gives the expected value of the process at input $\mathbf{x}$,
- $c(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$ is the covariance function, which defines the relationship (similarity) between function values at different input points $\mathbf{x}$ and $\mathbf{x}'$.

A Gaussian Process is fully specified by these two components—the mean function and the covariance function. GPs provide a flexible, non-parametric framework for modeling functions, as they do not assume a specific form of the function but instead define a distribution over all possible functions that could explain the observed data.

Because of its analytical tractability, the GP is commonly used to model real-valued functions, especially in cases where uncertainty needs to be quantified. In practice, one can think of a GP as a distribution over functions, where any collection of function values has a multivariate Gaussian distribution.

### *3.1.2 The Covariance Function (Kernel)*

The covariance function, also known as the kernel function, is a fundamental component of Gaussian Process Regression (GPR). It defines the covariance between function values at different input points, controlling key properties of the resulting function such as smoothness, periodicity, and behavior over different scales. Formally, the covariance function $c(\mathbf{x}, \mathbf{x}')$ takes two input points $\mathbf{x}$ and $\mathbf{x}'$ and returns a scalar that represents the covariance between the corresponding function values, $f(\mathbf{x})$ and $f(\mathbf{x}')$.

A covariance (kernel) function $c(\mathbf{x}, \mathbf{x}')$ is a non-negative, real-valued, integrable function that represents a mapping $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Let $\mathbf{x}_n$ and $\mathbf{x}_m$ be two points in the $\mathbb{R}^d$ Euclidean space, and let $u = \|\mathbf{x}_n - \mathbf{x}_m\|$ represent the Euclidean distance between them [31]. A covariance function satisfies the following properties:

1. **Non-negativity:** $c(u) \geq 0$ for all $u$.
2. **Symmetry:** $c(u) = c(-u)$ for all $u$.
3. **Normalization:** $\int_{-\infty}^{\infty} du\, c(u) = 1$.
4. **Finite second-order moment:** The integral $\int_{-\infty}^{\infty} du\, u^2 c(u)$ exists.
5. **Mode at the origin:** $c(u)$ takes its maximum value at $u = 0$.
6. **Continuity:** $c(u)$ is a continuous function of $u$.
7. **Scaling:** If $c(u)$ is a covariance function and $h > 0$ is a non-negative number, then the function $c_h(u) = \frac{1}{h} c\left(\frac{u}{h}\right)$ is also a covariance function.

**Models of Covariance Functions:** Commonly used models of covariance functions are listed below. These functions depend on the normalized distance $u = \frac{\|\mathbf{x}-\mathbf{x}'\|}{h}$, where $h > 0$ represents the kernel bandwidth.

For radial kernels, the covariance functions depend only on the magnitude $u = \|\mathbf{x} - \mathbf{x}'\|$, not the direction of the Euclidean distance between two points.

1. **Quadratic:**
$$c(u) = \frac{3}{4}(1 - u^2)\mathbb{1}_{|u|\leq 1}(u). \tag{50}$$
2. **Uniform:**
$$c(u) = \frac{1}{2}\mathbb{1}_{|u|\leq 1}(u). \tag{51}$$
3. **Triangular:**
$$c(u) = (1 - |u|)\mathbb{1}_{|u|\leq 1}(u). \tag{52}$$

4. **Quartic (biweight):**

$$c(u) = \frac{15}{16}\left(1-u^2\right)^2 \mathbb{1}_{|u|\leq 1}(u). \tag{53}$$

5. **Tricubic:**

$$c(u) = \frac{70}{81}\left(1-|u|^3\right)^3 \mathbb{1}_{|u|\leq 1}(u). \tag{54}$$

6. **Epanechnikov:**

$$c(u) = \frac{3}{4}(1-u^2)\mathbb{1}_{|u|\leq 1}(u). \tag{55}$$

7. **Spherical:**

$$c(u) = \frac{4}{3}\left(1-1.5u+0.5u^3\right)\mathbb{1}_{|u|\leq 1}(u). \tag{56}$$

8. **Gaussian:**

$$c(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2). \tag{57}$$

9. **Exponential:**

$$c(u) = \frac{1}{2}\exp(-|u|). \tag{58}$$

10. **Cauchy:**

$$c(u) = \frac{1}{\pi}\frac{1}{1+u^2}. \tag{59}$$

In the above, $\mathbb{1}_{|u|\leq 1}(u)$ denotes the indicator function, which satisfies:

$$\mathbb{1}_{|u|\leq 1}(u) = \begin{cases} 1 & \text{if } |u| \leq 1, \\ 0 & \text{if } |u| > 1. \end{cases} \tag{60}$$

The first seven covariance functions are compactly supported, meaning they are zero outside the interval $|u| \leq 1$, while the last three have unbounded support.

### *3.1.3 Maximum Likelihood Estimation (MLE)*

Maximum Likelihood Estimation (MLE) is a commonly used method for selecting the parameters $\boldsymbol{\theta} \in \Theta$ of a parameterized covariance kernel $c_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ [43]. Under the Gaussian Process model $\mathcal{GP}(m, c_{\boldsymbol{\theta}})$, the probability density function of the observed data $\mathbf{y}$ given $\boldsymbol{\theta}$ is expressed as:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{\sqrt{\det(2\pi\mathbf{C})}}\exp\left(-\frac{1}{2}\mathbf{y}^\top\mathbf{C}^{-1}\mathbf{y}\right), \tag{61}$$

where:

- $\mathbf{y}$: The observed target values,
- $\mathbf{C}$: The covariance matrix computed using the kernel $c_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ for the observed inputs $\mathbf{x}$,

- $\boldsymbol{\theta}$: Parameters of the kernel (e.g., length scale, variance, nugget).

Maximizing this likelihood function is equivalent to minimizing the negative log-likelihood function. The log-likelihood function is given by:

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top \mathbf{C}^{-1}\mathbf{y} - \frac{1}{2}\log\det\mathbf{C} - \frac{n}{2}\log(2\pi), \tag{62}$$

where $n$ is the number of observed data points.

For computational convenience, we often minimize the **modified log-likelihood function**, which removes constant terms:

$$l(\boldsymbol{\theta}|\mathbf{y}) = \mathbf{y}^\top \mathbf{C}^{-1}\mathbf{y} + \log\det\mathbf{C}. \tag{63}$$

### *3.1.4 Optimization of Kernel Parameters*

The parameters $\boldsymbol{\theta}$ are selected by solving:

$$\boldsymbol{\theta}_{\mathrm{ML}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} l(\boldsymbol{\theta}|\mathbf{y}), \tag{64}$$

where:

- The first term $\mathbf{y}^\top \mathbf{C}^{-1}\mathbf{y}$ measures the goodness-of-fit of the model to the data,
- The second term $\log\det\mathbf{C}$ penalizes model complexity to prevent overfitting.

By minimizing this objective function, MLE balances the trade-off between model fit and complexity, ensuring that the selected kernel parameters $\boldsymbol{\theta}$ generalize well to unseen data. This process is critical for accurately modeling the covariance structure $c(\mathbf{x}, \mathbf{x}')$ between observed data $\mathbf{x}$ and unobserved test points $\mathbf{x}'$.

### *3.1.5 Constructing the Gaussian Process*

To model a set of data $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we can employ a **Gaussian Process (GP)**. This provides a flexible, non-linear alternative to simple linear regression models, such as $Y = \beta_0 + \beta_1 X + \epsilon$, which may not fully capture the complexity of the data [43]. Instead of assuming a linear relationship, we hypothesize a non-linear function $\mathbf{f}$ that maps inputs to outputs via the model:

$$y_i = \mathbf{f}(\mathbf{x}_i) + \epsilon_i, \tag{65}$$

where $\epsilon_i$ represents noise or observation error, typically assumed to be independent and identically distributed (i.i.d.) according to:

$$\epsilon_i \sim \mathcal{N}(0, \sigma_n^2), \tag{66}$$

with $\sigma_n^2$ being the noise variance.

In GPs, the function values $\mathbf{f}(\mathbf{x})$ and $\mathbf{f}(\mathbf{x}')$ for any two distinct known inputs $\mathbf{x}$ and $\mathbf{x}'$ are treated as random variables that follow a joint Gaussian distribution. This assumption allows us to model the correlation between function values at different input locations, as illustrated in Figure 5.

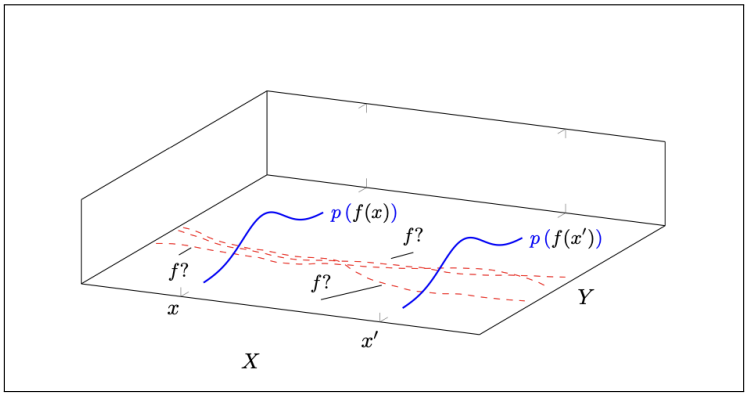


Figure 5: The unknown function $\mathbf{f}$, with possible realizations shown by red dashed lines. The solid blue lines illustrate the assumption that $\mathbf{f}(\mathbf{x})$ and $\mathbf{f}(\mathbf{x}')$ are jointly Gaussian distributed, with only marginal distributions shown for simplicity.

The joint distribution of the function values at known inputs $\mathbf{x}$ and $\mathbf{x}'$ can be written as:

$$\begin{bmatrix} \mathbf{f}(\mathbf{x}) \\ \mathbf{f}(\mathbf{x}') \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} \mathbf{m}(\mathbf{x}) \\ \mathbf{m}(\mathbf{x}') \end{bmatrix}, \begin{bmatrix} \mathbf{c}(\mathbf{x}, \mathbf{x}) & \mathbf{c}(\mathbf{x}, \mathbf{x}') \\ \mathbf{c}(\mathbf{x}', \mathbf{x}) & \mathbf{c}(\mathbf{x}', \mathbf{x}') \end{bmatrix} \right), \tag{67}$$

where $\mathbf{m}(\mathbf{x})$ and $\mathbf{m}(\mathbf{x}')$ are the mean function values at the known inputs $\mathbf{x}$ and $\mathbf{x}'$, and $\mathbf{c}(\mathbf{x}, \mathbf{x}')$ is the covariance between the function values at $\mathbf{x}$ and $\mathbf{x}'$.

A **Gaussian Process (GP)** is formally defined as a collection of random variables, any finite subset of which follows a joint multivariate Gaussian distribution. Thus, for a finite set of known inputs $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\}$, the corresponding function values $\{\mathbf{f}(\mathbf{x}^{(1)}), \mathbf{f}(\mathbf{x}^{(2)}), \ldots, \mathbf{f}(\mathbf{x}^{(n)})\}$ are jointly distributed as:

$$\mathbf{f} = \begin{bmatrix} \mathbf{f}(\mathbf{x}^{(1)}) \\ \mathbf{f}(\mathbf{x}^{(2)}) \\ \vdots \\ \mathbf{f}(\mathbf{x}^{(n)}) \end{bmatrix} \sim \mathcal{N}(\mathbf{m}, \mathbf{C}), \tag{68}$$

where $\mathbf{m}$ is the mean vector and $\mathbf{C}$ is the covariance matrix derived from the covariance function $\mathbf{c}(\mathbf{x}, \mathbf{x}')$, also known as the kernel. This covariance function governs the relationships between the function values at different known inputs.

If the target function $f(x)$ is estimated as a Gaussian process, then one needs to check if the consistency requirement of the estimator is fulfilled [42]. One can easily do this by the marginalization property of multivariate Gaussian distributions. The marginalization property tells us that if

$$(\mathbf{f}_1, \mathbf{f}_2) \sim \mathcal{N}(\mathbf{m}, \mathbf{C}), \tag{69}$$

then we also have

$$\mathbf{f}_1 \sim \mathcal{N}(\mathbf{m}_1, \mathbf{C}_{11}), \quad \mathbf{f}_2 \sim \mathcal{N}(\mathbf{m}_2, \mathbf{C}_{22}), \tag{70}$$

where $\mathbf{C}_{11}$ and $\mathbf{C}_{22}$ are sub-matrices of $\mathbf{C}$. The proof is as follows:

**Proof.** Prove that the marginal density $p(\mathbf{f}_1)$ follows a Gaussian distribution. First, by definition of marginalization, we have

$$p(\mathbf{f}_1) = \int p(\mathbf{f}_1, \mathbf{f}_2)\, d\mathbf{f}_2, \tag{71}$$

where

$$p(\mathbf{f}_1, \mathbf{f}_2) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \mathbf{C}}} \exp(E), \tag{72}$$

and $E$ is given by

$$E = -\frac{1}{2}\left(\mathbf{f}_2 - \left(\mathbf{m}_2 - \mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21}(\mathbf{f}_1 - \mathbf{m}_1)\right)\right)^T \mathbf{\Lambda}_{22} \left(\mathbf{f}_2 - \left(\mathbf{m}_2 - \mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21}(\mathbf{f}_1 - \mathbf{m}_1)\right)\right) \tag{73}$$

$$+\frac{1}{2}\left(-\mathbf{f}_1^T \mathbf{\Lambda}_{11}\mathbf{f}_1 + 2\mathbf{f}_1^T \mathbf{\Lambda}_{11}\mathbf{m}_1 - \mathbf{m}_1^T \mathbf{\Lambda}_{11}\mathbf{m}_1\right). \tag{74}$$

Here, $\mathbf{\Lambda}$ is the information matrix:

$$\mathbf{\Lambda} = \mathbf{C}^{-1} = \begin{pmatrix} \mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12} \\ \mathbf{\Lambda}_{21} & \mathbf{\Lambda}_{22} \end{pmatrix}. \tag{75}$$

Using the matrix inversion lemma (see Appendix), we have

$$\mathbf{C}^{-1} = \mathbf{\Lambda}_{11} - \mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21}. \tag{76}$$

By substituting this into $p(\mathbf{f}_1, \mathbf{f}_2)$, we can write

$$p(\mathbf{f}_1, \mathbf{f}_2) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \mathbf{C}}} \exp(E_1) \exp(E_2), \tag{77}$$

where

$$E_1 = -\frac{1}{2}\left(\mathbf{f}_2 - \left(\mathbf{m}_2 - \mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21}(\mathbf{f}_1 - \mathbf{m}_1)\right)\right)^T \mathbf{\Lambda}_{22} \left(\mathbf{f}_2 - \left(\mathbf{m}_2 - \mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21}(\mathbf{f}_1 - \mathbf{m}_1)\right)\right), \tag{78}$$

$$E_2 = -\frac{1}{2}(\mathbf{f}_1 - \mathbf{m}_1)^T \left(\mathbf{\Lambda}_{11} - \mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21}\right)(\mathbf{f}_1 - \mathbf{m}_1). \tag{79}$$

Since $E_2$ is independent of $\mathbf{f}_2$, the marginalization becomes

$$p(\mathbf{f}_1) = \frac{1}{(2\pi)^{n_1/2}\sqrt{\det \mathbf{C}_{11}}} \exp\left(-\frac{1}{2}(\mathbf{f}_1 - \mathbf{m}_1)^T \mathbf{C}_{11}^{-1}(\mathbf{f}_1 - \mathbf{m}_1)\right). \tag{80}$$

Again, by the matrix inversion lemma, we have

$$\det \mathbf{C} = \det \mathbf{C}_{11} \cdot \det(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}), \tag{81}$$

and

$$\mathbf{\Lambda}_{22}^{-1} = \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}. \tag{82}$$

Thus, the final expression is

$$p(\mathbf{f}_1) = \frac{1}{(2\pi)^{n_1/2}\sqrt{\det \mathbf{C}_{11}}} \exp\left(-\frac{1}{2}(\mathbf{f}_1 - \mathbf{m}_1)^T \mathbf{C}_{11}^{-1}(\mathbf{f}_1 - \mathbf{m}_1)\right), \tag{83}$$

which proves that $\mathbf{f}_1 \sim \mathcal{N}(\mathbf{m}_1, \mathbf{C}_{11})$.

Through consistency of the Gaussian process, one knows that if $(\mathbf{f}_1, \mathbf{f}_2) \sim \mathcal{N}(\mathbf{m}, \mathbf{C})$, then $\mathbf{f}_1 \sim \mathcal{N}(\mathbf{m}_1, \mathbf{C}_{11})$ and $\mathbf{f}_2 \sim \mathcal{N}(\mathbf{m}_2, \mathbf{C}_{22})$. Now, one can use Gaussian processes to model the target function that one wants to estimate.

### *3.1.6 The Prior in Gaussian Process Regression*

In **Gaussian Process Regression (GPR)**, the **prior** encodes assumptions about the underlying function before any data is observed [43]. The prior is specified by two key components:

- A **mean function**, $m(\mathbf{x})$, typically indicating no prior knowledge about the function's value before data is observed. Often, it is assumed to be zero, i.e., $m(\mathbf{x}) = 0$.
- A **covariance function** (or kernel), $c(\mathbf{x}, \mathbf{x}')$, which defines the correlation between function values at different input points.

The function we wish to model is assumed to be drawn from a Gaussian process:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}) = 0, c(\mathbf{x}, \mathbf{x}')), \tag{84}$$

where $f(\mathbf{x})$ represents the function, $m(\mathbf{x}) = 0$ is the mean function (often assumed to be zero), and $c(\mathbf{x}, \mathbf{x}')$ is the covariance function, which controls the structure and smoothness of the function.

The GP *prior* describes a distribution over possible functions, before any observations are made. Specifically, the prior assumes that any finite set of function values follows a **joint Gaussian distribution**, meaning that the values of $f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)$ for different input points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ are jointly distributed according to a multivariate Gaussian.

The **covariance function** $c(\mathbf{x}, \mathbf{x}')$ plays a critical role in the prior, as it governs the variability and structure of the functions that can be drawn from the Gaussian process. It defines how correlated the function values $f(\mathbf{x})$ and $f(\mathbf{x}')$ are at different input points $\mathbf{x}$ and $\mathbf{x}'$. The properties of the covariance function, such as smoothness, periodicity, or length scale, directly shape the set of plausible functions, capturing the assumptions about the relationships between inputs.

In summary, the GP prior encapsulates our beliefs about the range of possible functions the model can represent, with the covariance function dictating how function values are related across the input space.

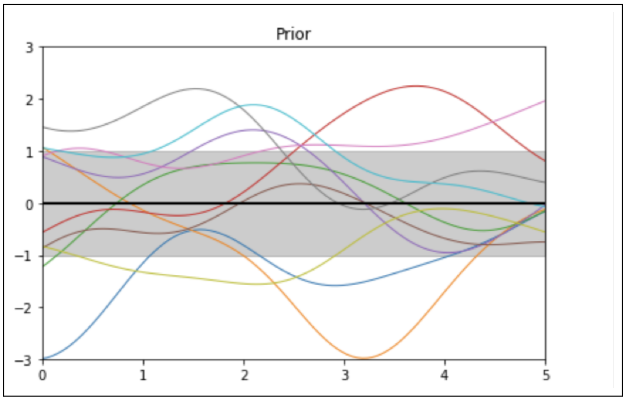


Figure 6: The prior of a Gaussian Process with its parameters set at initial values. The bold black line represents the mean function, which in this case is assumed to be zero. The colored lines are sample functions drawn from this Gaussian Process. The grey shaded region shows one standard deviation from the mean at each input point, reflecting the uncertainty in the prior [44].

As seen in Figure 6, the prior encapsulates a range of plausible functions that the GP can represent before any data is observed. The shaded region highlights the uncertainty in the prior, and the sample functions drawn from this distribution show how the GP can capture different patterns depending on the data observed.

### *3.1.7 Constructing the Gaussian Process Regression*

Gaussian Process Regression (GPR) assumes that the function we wish to model, denoted $f(\cdot)$, is drawn from a Gaussian process [45]. Given a set of training inputs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and corresponding observed function values $\mathbf{f} = \{f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)\}$, the goal is to predict the function value $f(\mathbf{x}_0)$ at a new test input $\mathbf{x}_0$.

To achieve this, we model the joint distribution of the function values at the training points

and the test point. This distribution is Gaussian and described by a mean function $m(\cdot)$ and a covariance function $c(\cdot, \cdot)$ (also known as the kernel function).

The joint distribution of the function values $f(\mathbf{x})$ at the training points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and the test point $\mathbf{x_0}$ can be expressed as:

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_N) \\ f(\mathbf{x_0}) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_N) \\ m(\mathbf{x_0}) \end{bmatrix}, \begin{bmatrix} c(\mathbf{x}_1, \mathbf{x}_1) & \cdots & c(\mathbf{x}_1, \mathbf{x}_N) & c(\mathbf{x}_1, \mathbf{x_0}) \\ \vdots & \ddots & \vdots & \vdots \\ c(\mathbf{x}_N, \mathbf{x}_1) & \cdots & c(\mathbf{x}_N, \mathbf{x}_N) & c(\mathbf{x}_N, \mathbf{x_0}) \\ c(\mathbf{x_0}, \mathbf{x}_1) & \cdots & c(\mathbf{x_0}, \mathbf{x}_N) & c(\mathbf{x_0}, \mathbf{x_0}) \end{bmatrix} \right), \tag{85}$$

where:

- $m(\mathbf{x}_i)$ is the mean of the Gaussian process at the point $\mathbf{x}_i$,
- $c(\mathbf{x}_i, \mathbf{x}_j)$ is the covariance between the function values $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$, and it is determined by the chosen kernel function,
- $\mathcal{N}(\mathbf{m}, \mathbf{C})$ represents the multivariate normal distribution with mean vector $\mathbf{m}$ and covariance matrix $\mathbf{C}$.

### *3.1.8 Posterior Distribution of $f(\mathbf{x_0})$*

To predict the function value $f(\mathbf{x_0})$ at the test point $\mathbf{x_0}$, we compute the posterior distribution by conditioning the joint distribution of the training data and test point on the observed values at the training points [46]. The posterior distribution of $f(\mathbf{x_0})$, given the observed training data $f(X)$, is a Gaussian distribution with the following mean and variance:

**Posterior Mean:**

$$\mathbb{E}[f(\mathbf{x_0})|f(X)] = m(\mathbf{x_0}) + C(\mathbf{x_0}, X)C(X, X)^{-1}(f(X) - m(X)), \tag{86}$$

where:

- $m(\mathbf{x_0})$ is the prior mean of the function value at the test point $\mathbf{x_0}$,
- $C(\mathbf{x_0}, X)$ is the vector of covariances between the test point $\mathbf{x_0}$ and the training points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, i.e., $C(\mathbf{x_0}, X) = [c(\mathbf{x_0}, \mathbf{x}_1), \ldots, c(\mathbf{x_0}, \mathbf{x}_N)]$,
- $C(X, X)$ is the covariance matrix of the function values at the training points, with $(i, j)$-th entry given by $c(\mathbf{x}_i, \mathbf{x}_j)$,
- $f(X) = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)]^T$ is the vector of observed function values at the training points,

- $m(X) = [m(\mathbf{x}_1), \ldots, m(\mathbf{x}_N)]^T$ is the mean vector of the function values at the training points.

The posterior mean $\mathbb{E}[f(\mathbf{x_0})|f(X)]$ represents the best estimate of the function value at the test point $\mathbf{x_0}$ based on the training data. It is a linear combination of the observed values $f(X)$, where the weights depend on the kernel function and the relative distances between the test point and the training points.

**Posterior Variance:**

$$\text{Var}[f(\mathbf{x_0})|f(X)] = c(\mathbf{x_0}, \mathbf{x_0}) - C(\mathbf{x_0}, X)C(X, X)^{-1}C(X, \mathbf{x_0}), \tag{87}$$

where:

- $c(\mathbf{x_0}, \mathbf{x_0})$ is the prior variance (i.e., the variance without observing any data) of the function value at the test point $\mathbf{x_0}$,
- $C(\mathbf{x_0}, X)C(X, X)^{-1}C(X, \mathbf{x_0})$ is the reduction in uncertainty due to the information provided by the observed data.

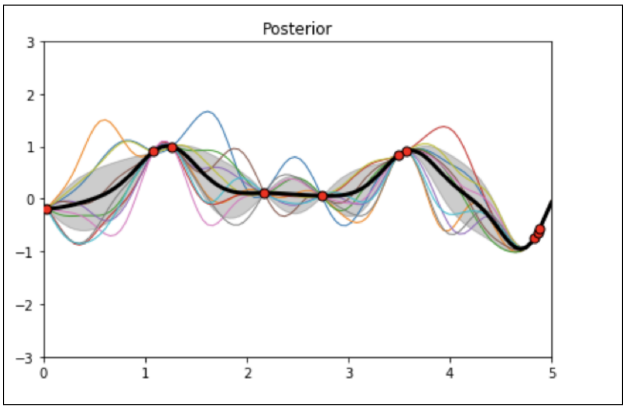


Figure 7: The posterior of a Gaussian Process after observing data points (shown as red dots on the mean line). The colored lines are sample functions drawn from the posterior distribution. The grey area is one standard deviation away from the mean at each input point. After observing data, the uncertainty decreases around observed points, as illustrated by the shrinking grey area [44].

As seen in Figure 7, the posterior distribution incorporates information from the observed training data. Near the observed data points (red dots), the model has high confidence (low variance), and the uncertainty in the predictions decreases. In contrast, for regions far from the observed data, the variance remains high, reflecting greater uncertainty in the model's predictions.

The posterior variance quantifies the uncertainty in the prediction at the test point $\mathbf{x_0}$. If $\mathbf{x_0}$ is close to one or more training points (as determined by the kernel function), the variance will be reduced, indicating greater confidence in the prediction. Conversely, if $\mathbf{x_0}$ is far from the training points, the variance will remain close to the prior variance, reflecting high uncertainty.

### *3.1.9 Prediction*

The posterior distribution enables us to make predictions for new test inputs $\mathbf{x_0}$ based on the observed training data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. The posterior mean $\mathbb{E}[f(\mathbf{x_0})|f(X)]$ provides a point estimate of the predicted function value, while the posterior variance $\mathrm{Var}[f(\mathbf{x_0})|f(X)]$ quantifies the uncertainty associated with this prediction.

By employing a Gaussian process with a suitable covariance function (e.g., the squared exponential kernel), the predicted function values are guaranteed to be smooth and continuous. Moreover, GPR provides not only a prediction but also a measure of uncertainty, making it a powerful tool for regression tasks that require both accuracy and uncertainty quantification.

This ability to estimate uncertainty is particularly useful in applications such as active learning, where one can decide whether to trust a prediction or whether more data should be collected in regions of high uncertainty.

### *3.1.10 Advantages and Challenges*

- **Advantages:**
  - Provides uncertainty quantification, allowing for informed decision-making.
  - Flexible model that can handle complex data relationships through kernel functions.
- **Challenges:**
  - Scalability issues with large datasets, as GPR has computational complexity of $O(n^3)$ due to the inversion of the kernel matrix.
  - Requires careful tuning of kernel hyperparameters for optimal performance.

## 3.2 Self-Organizing Map (SOM) Algorithm

The Self-Organizing Map (SOM) was introduced by Teuvo Kohonen. It is an unsupervised learning algorithm, mainly designed for dimensionality reduction and clustering of high-dimensional data. The SOM maps input vectors from a continuous, high-dimensional input space to a lower-dimensional discrete grid of neurons while preserving the topological structure of the data. This makes SOMs especially useful for visualizing patterns and structures in complex datasets [50].

### *3.2.1 Architecture of SOM*

The architecture of a Self-Organizing Map (SOM) consists of two layers: an input layer that represents the input vectors and a feature map where each neuron is fully connected to all the input vectors through a set of weight vectors [52, 53]. These neurons in the feature map are arranged in a two-dimensional grid, and their weight vectors are iteratively updated during the training process. This adjustment allows the map to output a spatially sorted list that retains topological contiguity for the source data.

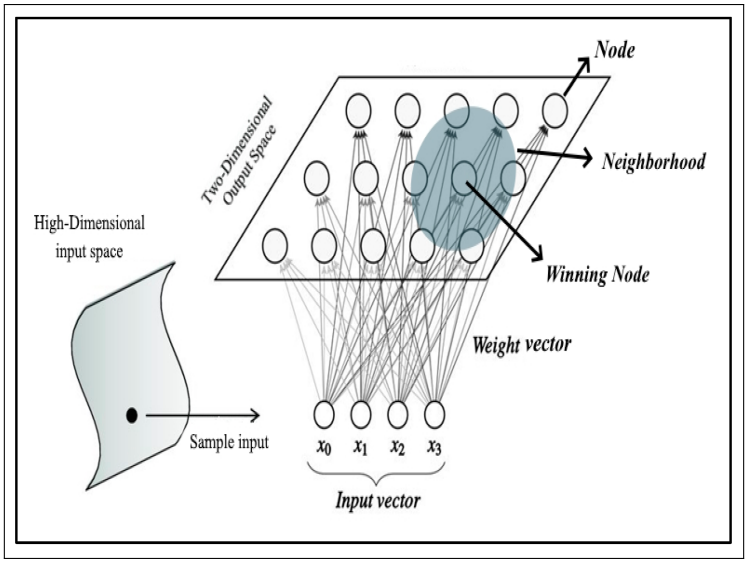


Figure 8: The Self-Organizing Map (SOM) architecture and learning process. The figure demonstrates how a high-dimensional input space is mapped onto a two-dimensional feature map. The Best Matching Unit (BMU) is identified for each input vector, and its weight vector, as well as the weights of its neighboring neurons, is updated. Over iterations, the SOM converges, and the feature map represents a topologically ordered approximation of the input space [51].

In the SOM method, the transformation of the high-dimensional input space to two-dimensional output space is explained using Figure. The left of Figure shows the high-dimensional input space, in which one can view the input vector as a point in the multidimensional space. Let the input layer consists of multi-dimensional vectors, $x_0, x_1, x_2, x_3$, as shown in Figure. These input vectors are processed by the feature map layer, which consists of a grid of neurons. Each neuron in the feature map connects to all the input dimensions through a weight vector, with the same dimensionality as the input vector. The SOM picks out the neuron whose weight vector is most similar to the given input vector and defines this neuron as the *Best Matching Unit (BMU)*. This similarity is usually calculated by using a distance metric, such as the Euclidean distance. After determining the BMU, the weights of the BMU and its neighbors are updated to move their weights closer to the input vector. The magnitude of the weight update decreases with the increase in distance from the BMU in the feature map. This iterative process continues for successive input vectors, where the weight vectors of the neurons are gradually refined to allow

the feature map to self-organize. In time, the neurons converge to weights that reflect the distribution of the input data, while similar input vectors are mapped to neighboring neurons, thus preserving the topological structure of the input space.

The shaded area in the feature map shows the neighborhood of the BMU, which is also affected during the process of updating the weights. This neighborhood decreases with time, allowing the SOM to fine-tune the weight vectors and converge to a stable configuration. The final SOM gives a representation of the input data where the spatial arrangement of neurons reflects both the topological relationships and the density distribution of the input space.

The SOM represents a powerful tool for clustering, visualization, and dimensionality reduction because of its ability to map high-dimensional data onto a lower-dimensional space. In contrast to linear methods like Principal Component Analysis, the SOM uses a non-linear mapping that enables it to capture more complex patterns and structures in the data. The feature map is a visual representation of the clustering process, where similar input vectors are grouped into adjacent regions on the map. This makes SOMs especially useful in the analysis and interpretation of high-dimensional data.

### *3.2.2 Overview of the Algorithm*

The goal of the SOM algorithm is to map similar input data vectors to neighboring neurons in the grid, thereby creating a spatial organization of the input space [47]. The learning process is iterative, where neurons compete to represent input patterns, and weight updates are performed to adjust both the winning neuron and its neighbors. This allows the network to progressively adapt to the input data and organize the grid accordingly.

The algorithm follows several steps:

**1: Initialization**. Each neuron $j$ in the grid is initialized with a weight vector $\mathbf{w}_j \in \mathbb{R}^d$, where $d$ is the dimensionality of the input space [48]. These weight vectors are typically initialized randomly, often chosen from a uniform distribution over the input space:

$$\mathbf{w}_j(0) \in [0,1]^d. \tag{88}$$

The weights of all neurons are initialized before any training begins, allowing the algorithm to start from a diverse range of weight values.

**2: Sampling**. During each iteration $k$, a sample input vector $\mathbf{x}(k) \in \mathbb{R}^d$ is drawn randomly from the input dataset [47]. Here, $k$ represents the iteration index, denoting the number of steps

elapsed during the training process. The iterative process ensures that the SOM adapts to various patterns within the dataset over time.

**3: Best Matching Unit (BMU) Identification**. For each sampled input vector $\mathbf{x}(k)$, the algorithm identifies the Best Matching Unit (BMU), which is the neuron whose weight vector is closest to the input vector in terms of Euclidean distance. The BMU is determined by minimizing the distance between the input vector and the weight vectors of all neurons in the grid:

$$D_j(k) = \|\mathbf{x}(k) - \mathbf{w}_j(k)\| = \sqrt{\sum_{i=1}^{d}(x_i(k) - w_{ji}(k))^2}, \quad \text{for each } j, \tag{89}$$

where $x_i(k)$ and $w_{ji}(k)$ represent the $i$-th components of the input vector and the weight vector, respectively. The neuron $j$ with the smallest distance $D_j(k)$ is identified as the BMU:

$$\mathbf{w}_{\text{BMU}}(k) = \min_j D_j(k). \tag{90}$$

The BMU is the neuron that most closely resembles the input vector, and it will be the focal point for weight updates.

**4: Weight Vector Update**. After identifying the BMU, the algorithm proceeds to update the weight vectors of the BMU and its neighboring neurons to move them closer to the input vector $\mathbf{x}(k)$ [52]. This update process is essential for adapting the SOM to the input data over time. The weight update rule is as follows:

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) + \theta(j, \text{BMU}, k) \cdot \alpha(k) \cdot (\mathbf{x}(k) - \mathbf{w}_j(k)), \tag{91}$$

where:

- $\alpha(k)$ is the learning rate, which decreases over time,
- $\theta(j, \text{BMU}, k)$ is the neighborhood function, which controls how much neighboring neurons are updated based on their distance from the BMU,
- $\mathbf{w}_j(k)$ is the weight vector of neuron $j$ at iteration $k$,
- $\mathbf{x}(k)$ is the input vector at iteration $k$.

The neighborhood function $\theta(j, \text{BMU}, k)$ ensures that neurons close to the BMU in the grid are updated more significantly, while neurons farther away experience smaller updates [48]. A commonly used neighborhood function is Gaussian:

$$\theta(j, \text{BMU}, k) = \exp\left(-\frac{\|\mathbf{r}_j - \mathbf{r}_{\text{BMU}}\|^2}{2\sigma^2(k)}\right), \tag{92}$$

where $\mathbf{r}_j$ and $\mathbf{r}_{\text{BMU}}$ are the positions of neuron $j$ and the BMU in the grid, and $\sigma(k)$ is the neighborhood radius, which also decreases over time.

**Step 5: Continuation**. The SOM algorithm repeats steps 2 through 4 for each input vector in the dataset, continually adjusting the weight vectors over time. As the algorithm progresses, both the learning rate $\alpha(k)$ and the neighborhood radius $\sigma(k)$ decrease, allowing the SOM to refine the map and focus on local adjustments:

$$\alpha(k) = \alpha_0 \cdot \exp\left(-\frac{k}{\tau_\alpha}\right), \quad \sigma(k) = \sigma_0 \cdot \exp\left(-\frac{k}{\tau_\sigma}\right), \tag{93}$$

where $\alpha_0$ and $\sigma_0$ are the initial values of the learning rate and neighborhood radius, and $\tau_\alpha$ and $\tau_\sigma$ are time constants that control the rate of decay. The algorithm continues until the weight updates become negligible, signaling that the map has converged and the neurons have stabilized to represent the input space effectively [49].

Once the SOM has converged, the feature map that emerges is a spatially organized representation of the input space, reflecting its structure, density, and patterns. The key properties of the SOM, including topological ordering, density matching, and feature selection, ensure that the map accurately represents the input data, making it useful for various data-driven applications.

# 4 Initial Data Examination

The dataset utilized in this analysis is described in great detail, along with the preparation made on the data toward further geostatistical and machine learning modeling. Precisely, this chapter outlines the characteristics of the dataset, examining linear trends within the data, a very important step toward further understanding the relationships between different variables. This will help ensure accuracy in the subsequent spatial models for developing better reliability in predictions by giving a necessary look into those foundational data aspects.

## 4.1 Data Acquisition and Preprocessing Techniques

The dataset consists of 636 spatial data points from a mining site carrying spatial coordinates, each with Easting and Northing, along with zinc concentration measurements in ppm, ranging between 30 ppm and 7000 ppm. For effective utilization of this data in further analysis, some preprocessing is required to be performed on the data in various ways, as mentioned below.

- **Converting Spatial Coordinates:** The Easting and Northing of the spatial coordinates are in meter units using the Ireland Transverse Mercator (ITM) system. These could be converted to kilometers to make them easier to interpret and keep consistency during spatial analysis.

Table 1: Sample of spatial coordinates and zinc concentrations.

| **Easting (ITM)** | **Northing (ITM)** | **Zn (ppm)** |
|---|---|---|
| 614485.63 | 798189.58 | 70 |
| 614514.36 | 798188.45 | 60 |
| 614547.59 | 798186.76 | 70 |
| 614576.89 | 798185.06 | 70 |
| 614603.36 | 798183.93 | 80 |

- **Handling Missing Zinc Concentration Values:** Any missing values (NaN) in Zinc concentration are appropriately handled by removing them.
- **Logarithmic Transformation:** The large range and skewness of zinc concentrations necessitate a logarithmic transformation, particularly for geostatistical approaches such as Kriging and Gaussian Process Regression. This transformation normalizes the distribution, reduces variance, and minimizes significant outliers, thereby improving spatial predictions.

The following table presents the descriptive statistics for the zinc concentration ($Zn$) values, which were used as the primary variable in the spatial analysis and model development:

Table 2: Descriptive Statistics for Zinc Concentration ($Zn$)

| **Statistic** | **Value** |
|---|---|
| Minimum Zinc Concentration (*Znmin*) | 30.00 |
| Maximum Zinc Concentration (*Znmax*) | 7000.00 |
| Mean Zinc Concentration (*Znmean*) | 860.45 |
| Median Zinc Concentration (*Znmedian*) | 405.62 |
| Standard Deviation (*Znstd*) | 1420.38 |
| Skewness (*Znskewness*) | 3.27 |
| Kurtosis (*Znkurtosis*) | 13.12 |

The above table gives the summary descriptive statistics for the zinc concentration data. The original skewness of 3.27 indicates a highly skewed distribution with a long tail of extreme values, suggesting significant outliers. The degree of this skew suggests that a logarithmic transformation is necessary to normalize the data. After log transformation, the skewness is reduced to 1.19, resulting in a relatively symmetric distribution. This transformation produces a distribution closer to normal, making it more appropriate for geostatistical analysis.

The following table shows descriptive statistics for the log-transformed Zn concentration:

Table 3: Descriptive Statistics for Log-transformed Zinc Concentration

| **Statistic** | **Value** |
|---|---|
| Minimum Zinc Concentration (*Znmin*) | 3.40 |
| Maximum Zinc Concentration (*Znmax*) | 8.85 |
| Mean Zinc Concentration (*Znmean*) | 5.24 |
| Median Zinc Concentration (*Znmedian*) | 5.07 |
| Standard Deviation (*Znstd*) | 0.98 |
| Skewness (*Znskewness*) | 1.19 |
| Kurtosis (*Znkurtosis*) | 4.31 |

### *4.1.1 Proportion of Data Across Selected Concentration Ranges*

To better understand the distribution of zinc concentrations, the percentage of values falling within specific concentration ranges was examined. This breakdown reveals how the data is distributed across different levels of zinc concentration.

Table 4: Proportion of Zinc Concentration Values Across Selected Ranges

| **Zinc Concentration Range (ppm)** | **Proportion of Data (%)** |
|---|---|
| 30 - 300 | 79.15% |
| 300 - 500 | 5.53% |
| 500 - 1500 | 10.42% |
| 1500 - 5000 | 4.43% |
| 5000 - 7000 | 0.47% |

Most of the data reflects low zinc values, with **79.15%** falling between 30 and 300 ppm. A smaller proportion, **5.53%**, falls between 300 and 500 ppm, indicating moderate values. Higher concentrations between 500 and 1500 ppm account for **10.42%** of the data. Only **4.43%** of the data lies between 1500 and 5000 ppm, and concentrations between 5000 and 7000 ppm account for just **0.47%**.

## 4.2 Linear Trend Model

Let the zinc concentration $Z(x)$ at a spatial location $x = (X, Y)$ depend on the spatial coordinates. The large-scale trend in zinc concentration can be modeled as:

$$m(x) = \beta_0 + \beta_1 X + \beta_2 Y, \tag{94}$$

where $\beta_0$ is the intercept, and $\beta_1$ and $\beta_2$ are coefficients representing the contributions of spatial coordinates $X$ (Easting) and $Y$ (Northing), respectively, to the overall trend.

Table 5: Linear Trend Model Coefficients

| **Parameter** | **Value** |
|---|---|
| $\beta_0$ (Intercept) | $-284.825$ |
| $\beta_1$ (Coefficient for $X$) | 0.55 |
| $\beta_2$ (Coefficient for $Y$) | $-0.06$ |

The positive coefficient $\beta_1 = 0.55$ indicates that zinc increases with $X$ (Easting), while the negative coefficient $\beta_2 = -0.06$ suggests a slight decrease with $Y$ (Northing). This provides insight into the large-scale spatial behavior of zinc concentration in the study area.

### *4.2.1 Importance of Trend Removal and Reintroduction*

In spatial modeling, detrending and reintroducing the linear trend are crucial. By removing the large-scale trend, the interpolation focuses on local spatial correlations, ensuring that the residuals $\hat{R}(x)$ represent small-scale variations. Reintroducing the trend afterward ensures that final predictions account for both the overall trend and local fluctuations, providing a comprehensive spatial estimate across the study area.

# 5 Data Exploration and Preliminary Data Distribution Analysis

This chapter covers a comprehensive analysis of the dataset on zinc concentration using a set of exploratory data analysis techniques and geostatistical methods. The main goal is to gain a preliminary view of the distribution, spatial structure, and relationships in the data, which will aid in the further stages of modeling. Using a range of visualization techniques—scatter plots, histograms, and empirical CDFs—alongside spatial interpolation, this chapter aims to present patterns of zinc concentration against existing spatial variability. Such analysis helps in selecting appropriate models and methods for spatial prediction.

## 5.1 Scatter Plot of Sampling Locations and Zinc Concentrations

The figure below shows a scatter plot of the spatial distribution of sampling locations where measurements of zinc concentration were collected in the study area.

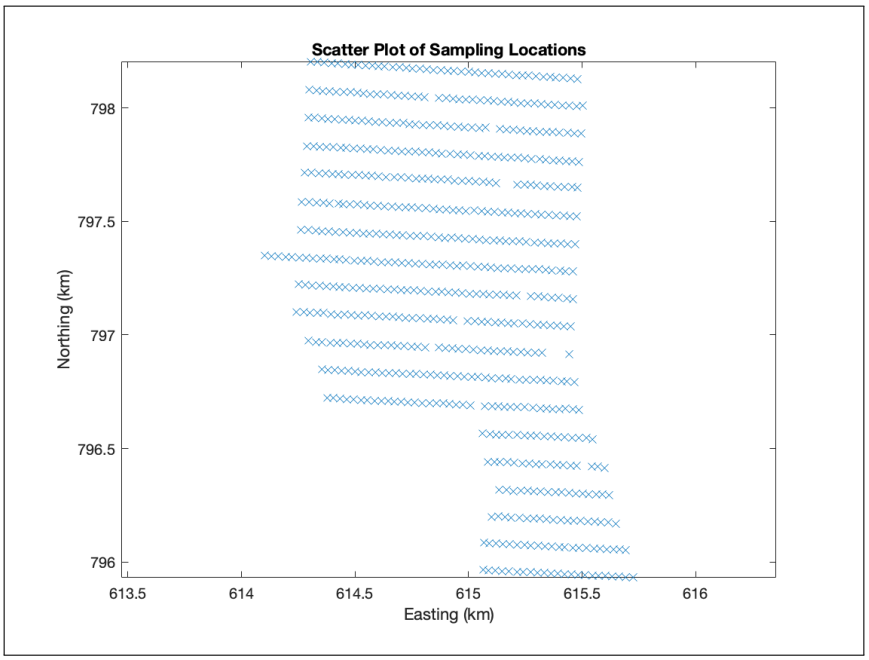


Figure 9: Scatter plot showing the spatial distribution of sampling locations across the study area. Each "x" represents a site where the concentration of zinc was measured. The Easting and Northing coordinates are in kilometers. The grid-like nature of the sampling points indicates a systematic sampling approach to ensure adequate spatial coverage of the area for geostatistical analysis.

The Easting values range approximately from 613.5 km to 616 km, representing the horizontal positions of the sampling locations. The Northing values span from 796 km to 798 km, corre-

sponding to the vertical positioning of the sites. Each cross on the scatter plot marks a location where zinc concentration (`Zn`) data was collected.

The systematic and grid-like placement of sampling locations reflects structured sampling to achieve an even spatial distribution. This scatter plot serves as a visual framework for analyzing the spatial distribution of zinc concentrations and is an essential input to subsequent geostatistical interpolation.

## 5.2 Spatial Interpolation of Zinc Concentrations

Spatial interpolation provides a continuous visualization of zinc concentration across the study area. It estimates concentrations in unsampled areas using values measured at sampled locations. Linear interpolation fits a surface between neighboring points to provide a smooth approximation of zinc concentrations. The key variables are the spatial coordinates, $X$ and $Y$, and the measured zinc concentrations $Zn$ at the sampled locations.

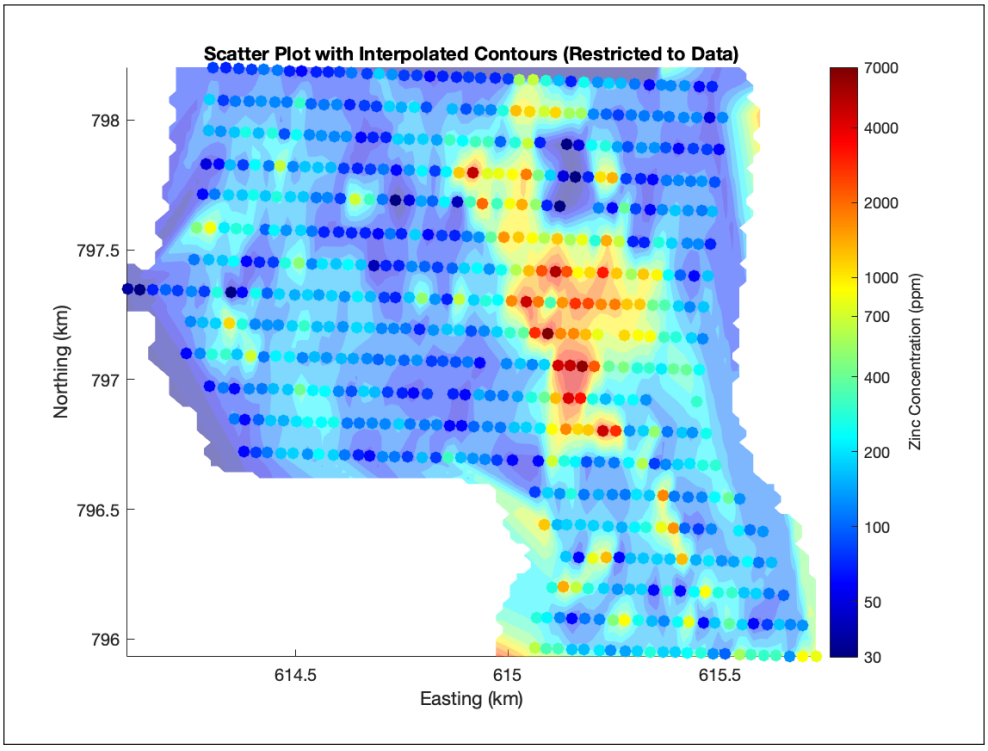


Figure 10: Scatter plot of sampling locations overlaid with interpolated contour lines, showing zinc concentrations across the study area. Scatter points are color-coded based on measured concentrations, while contour lines represent interpolated estimates for unsampled areas.

This interpolation assumes that areas with no direct data in between are likely to have similar concentrations to nearby points. However, interpolation is restricted to areas where actual data exists, ensuring reliability and consistency with known data.

- **Hotspot Identification:** Areas with closer spacing of contour lines, particularly near the

center, may indicate regions where zinc concentrations vary rapidly, suggesting potential hotspots.

- **Gradients and Spatial Patterns:** More widely spaced contour lines indicate gradual variations in zinc concentration, while closely spaced lines reflect sharp gradients.
- **Clustering Patterns:** The scatter plot combined with contour lines helps verify that the clustering patterns of high and low zinc concentrations in the raw data are preserved after interpolation.

This approach provides a detailed spatial overview of zinc concentration, filling in unsampled areas to reveal trends and patterns across the region. Such visualizations are vital for identifying regions of high zinc concentrations or unexpected spatial anomalies.

## 5.3 Normal Probability Plot and Histogram Analysis

Figure 11 presents two essential visualizations used to assess the distributional characteristics of the zinc concentration data: a histogram (top) and a normal probability plot (bottom). Each plot provides valuable insights into the shape and behavior of the zinc concentrations. The values are presented in parts per million (ppm) after back-transforming from the logarithmic scale.

### *5.3.1 Histogram*

The histogram (Figure 11, top) represents the frequency distribution of zinc concentration data. The horizontal axis displays zinc concentrations, while the vertical axis indicates the frequency of observations within each bin. The distribution shows a right-skewed pattern, with most values concentrated between approximately 4 and 6 ppm. This indicates that lower zinc concentrations are more frequent, while higher concentrations, represented by the extended right tail, are relatively rare. The histogram's asymmetry suggests that, while the logarithmic transformation reduced skewness, the data is not perfectly normalized. The presence of a right tail indicates potential outliers or extreme values.

### *5.3.2 Normal Probability Plot*

The normal probability plot (Figure 11, bottom) compares empirical cumulative probabilities with the theoretical quantiles of a normal distribution. If the data were perfectly normal, the points would align along the reference line. Deviations from the line, particularly in the tails,

indicate departures from normality.

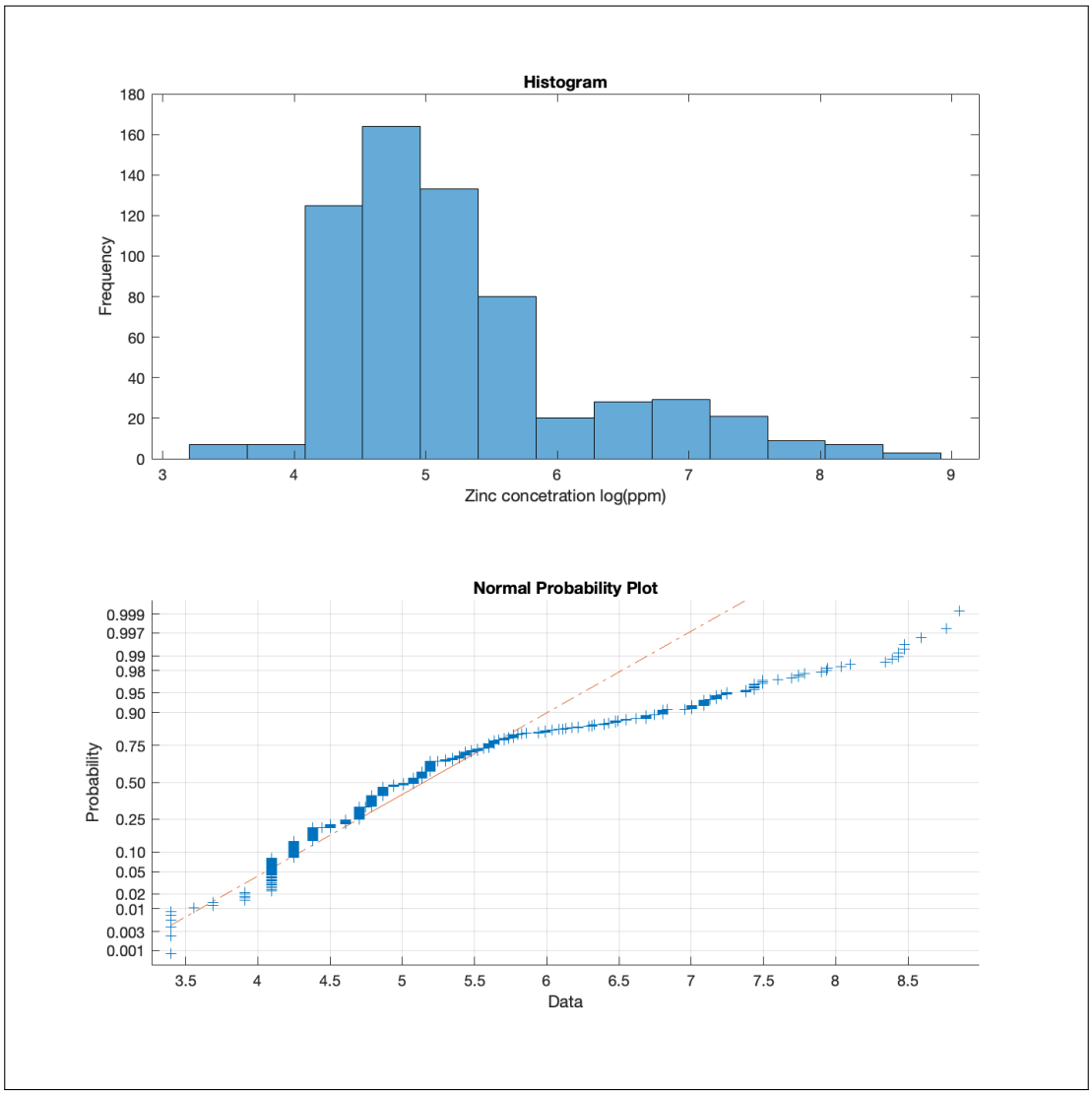


Figure 11: Top: Histogram of the zinc concentration data (`Zn`) depicting the frequency distribution within the range of observed values. Bottom: Normal probability plot of `Zn` values against a theoretical normal distribution. Curvature, especially in the tails, indicates deviations from normality.

## 5.4 Empirical Cumulative Distribution Function (ECDF)

The ECDF evaluates the empirical distribution of log-transformed data against a theoretical Gaussian distribution. This comparison highlights deviations from normality.

In Figure 12, the black curve represents the ECDF, while the red curve corresponds to the cumulative distribution function (CDF) of a fitted Gaussian distribution. Deviations between the two curves, particularly in the tails, indicate that the data is not strictly normal.

Given the robustness of geostatistical methods to deviations from strict normality, the dataset is treated as approximately normal for the purpose of geostatistical modeling. This assumption

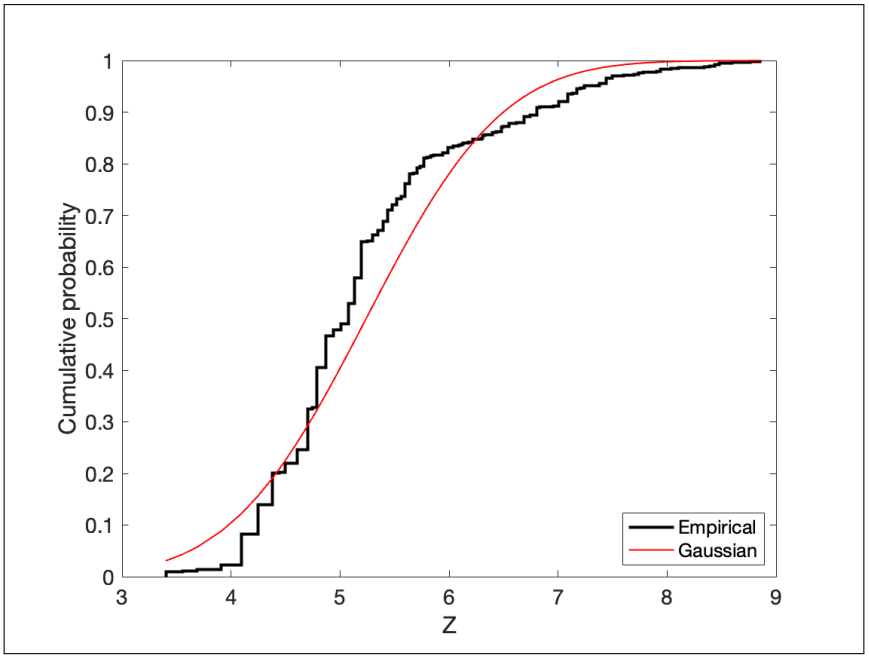


Figure 12: Comparison between the empirical cumulative distribution function (black line) and the fitted Gaussian distribution (red line) for zinc concentration values in parts per million (ppm).

enables the application of techniques such as Kriging while acknowledging the presence of extreme values or localized anomalies.

## 5.5 Voronoi Diagram

The Voronoi diagram presented in Figure 13 illustrates the partitioning of the study area into regions based on the spatial influence of individual data points, referred to as seeds.

The area in this Voronoi diagram corresponds to the respective area of the study region that is closer to any particular data point-seed. The boundaries of the areas are defined by an equidistant point from surrounding seeds and represent a geometric interpretation of spatial association and proximity among the point occurrences of each sample. This method aids well in showing the influence a sample exerts on an area clearly. The color code of the polygons shows the value of the concentration of zinc measured at every seed point:

- **Blue areas** denote the lowest concentration values, indicating regions of low intensity.
- **Green and Yellow shades** represent moderate concentration values, corresponding to intermediate levels of the measured variable.
- **Red areas** highlight the highest concentration values, indicating regions of high intensity.

The **Voronoi diagram serves as a valuable tool for elucidating proximity-based relationships in spatial data**, enabling the identification of patterns, clusters, and areas of

interest, thereby guiding further investigation into the spatial variability within the dataset.

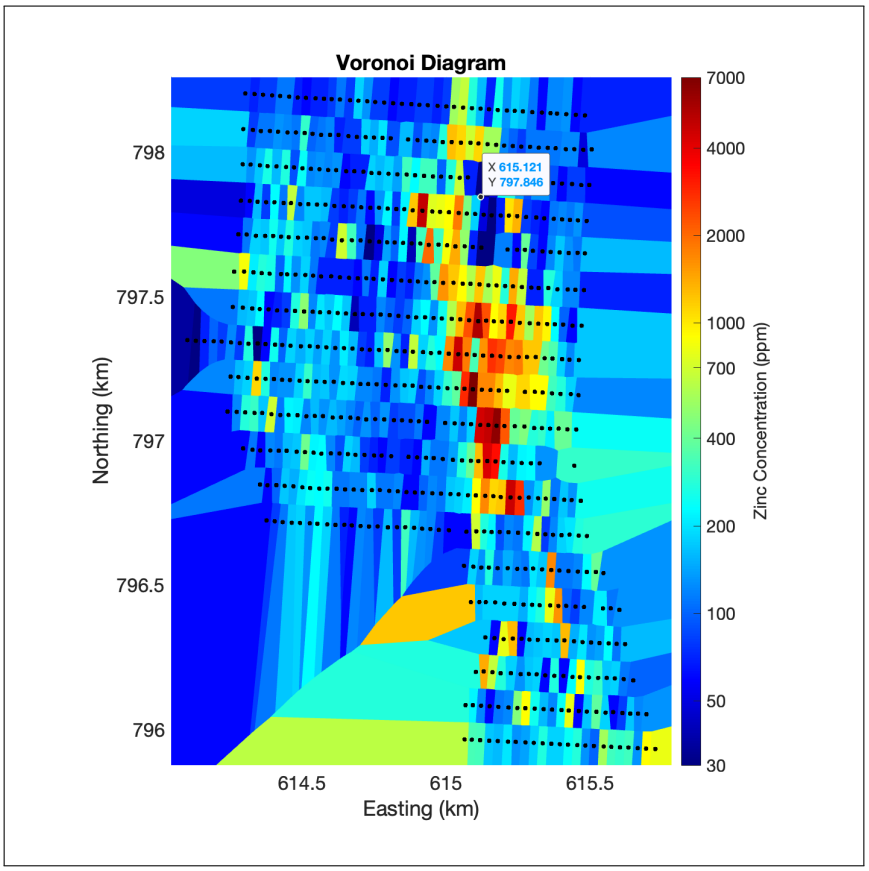


Figure 13: Voronoi diagram partitioning the study area based on proximity to each generator (seed) point, utilizing the concentration data for zinc. Each cell represents the region that is closest to a particular data point in the dataset, with boundaries equidistant between neighboring points. The diagram visually illustrates the influence area of each sampled location across the spatial domain, providing a comprehensive view of the spatial relationships among the data points. The values are represented in parts per million (ppm) after back-transforming from the logarithmic scale.

## 5.6 Variography

Variography is an essential method in geostatistics for analyzing spatial relationships within a dataset. It allows for the measurement of spatial dependencies between data points and helps in understanding the spatial structure of the dataset.

### *5.6.1 Experimental Variogram Calculation*

The first step in variography is calculation of the experimental variogram, which describes the spatial dependency of pairs of data points as a function of distance between them. The variogram is developed by calculating the average squared differences of pairs of samples and plotting the values versus their separation distance (the lag). This indicates the breakdown in spatial similarity

with increased separation.

Several maximum distances, 0.2 km, 0.3 km, and 0.4 km were tested to see their performances in capturing the spatial structure of the data. These correspond approximately to lag distances of 25 m, 40 m, and 55 m, respectively, after further investigation. After careful consideration, a distance of 0.3 km corresponding to a lag distance of 40 m with 21 lags was decided to be appropriate. This choice provided an optimal balance between spatial resolution and the number of pairs within each lag bin, ensuring a robust and detailed representation of spatial variability.

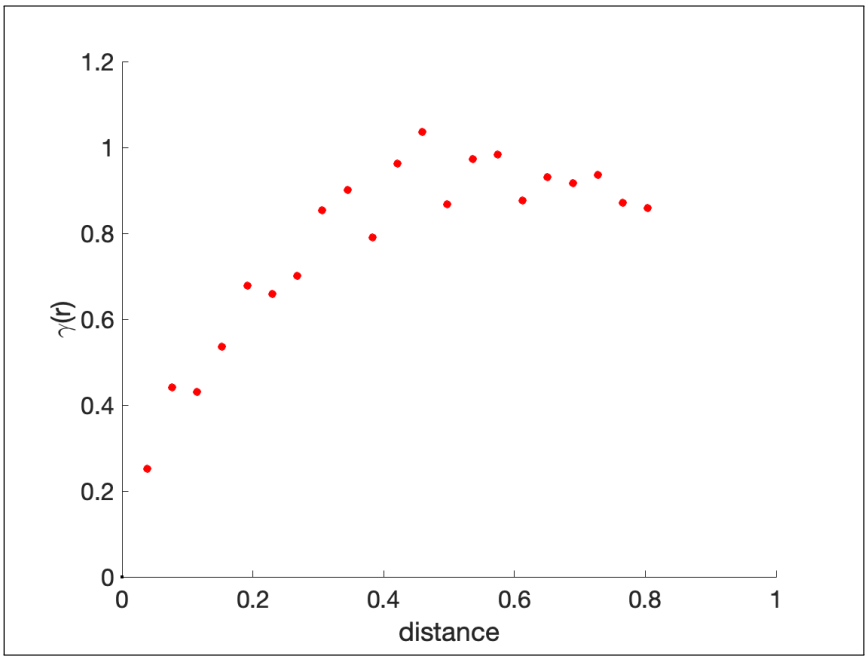


Figure 14: Experimental variogram based on the zinc concentration data, showing the relationship between lag distance and semi-variance. At short distances, the semi-variance is low, indicating strong spatial correlation. As the lag distance increases, the semi-variance rises, reflecting a weakening of spatial correlation. The curve eventually levels off, indicating the sill, beyond which no further spatial dependence is observed. The values are presented in parts per million (ppm) after back-transforming from the logarithmic scale; the estimations are predicted in the logarithmic scale and then converted to ppm, as illustrated in the figure.

As depicted in Figure 14, the x-axis represents the lag distance, while the y-axis denotes the semi-variance, a measure of spatial variability.

The variogram curve reveals key properties, including the **range**, which defines the maximum distance over which spatial correlation is observed, and the **sill**, which represents the point at which the semi-variance stabilizes, signifying the absence of further spatial dependence.

The procedure for calculating the **empirical variogram** involves several key steps:

- Computing pairwise distances between all sampling locations,
- Categorizing these distances into lag classes, which represent distinct ranges of separation distances,

- Calculating the semi-variance for each lag class, defined as the average squared difference between the data values within each class.

### *5.6.2 Variogram Model Selection and Fitting*

Once the **experimental variogram** has been computed, the next step is to fit a theoretical model that characterizes how spatial correlation decays as a function of increasing distance. Commonly utilized models include the **spherical**, **exponential**, and **Gaussian variogram models**, each offering a unique mathematical representation of spatial continuity. In this particular analysis, the **exponential model** was selected as the most appropriate fit. This model works even better when the dataset has sharp spatial autocorrelation over shorter distances, with a gradual leveling to the sill. Unlike the spherical model, the exponential model does not have an abrupt range at which correlation effectively ends; instead, it displays a continuous decay as the distance of separation increases. Since directionality could also be incorporated directly with anisotropy through various weighting functions, directional filtering applies when limiting analysis to particular tolerances in angle.

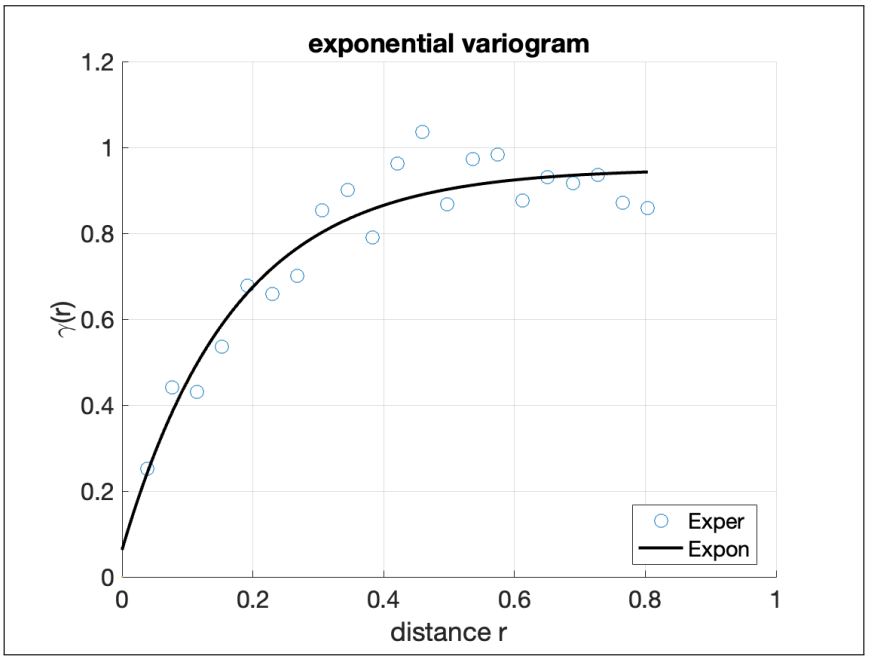


Figure 15: Fitted exponential variogram model based on the log-transformed zinc concentration data (`log(Zn)`). The x-axis represents the lag distance (in kilometers), and the y-axis represents the semi-variance. The curve illustrates how spatial correlation decreases with distance, with the sill and range parameters defining the points where spatial dependence becomes negligible.

Figure 15 shows the fitted **exponential variogram**, which provides critical insights into the spatial structure of the dataset. The nugget, sill, and range parameters have been adjusted to ensure that the model fits the empirical data precisely, which is required for accurate spatial

interpolation using techniques such as **kriging**. The fitted model provides complete information on the scale of spatial dependence, thus helping make more reliable predictions at unsampled locations.

## 5.7 Kriging Model Evaluation and Prediction Analysis

This section carries out in detail the performance of the kriging model for predicting the spatial zinc concentrations by using both cross-validation and residual error analyses. First, cross-validation is carried out by systematically withholding subsets of data, predicting the values of those withheld points, and comparing them to the observed values-a process that gives some insight into the model's generalization ability and accuracy, along with its robustness in cases when it is applied to unseen data.

### *5.7.1 Cross-Validation Process for Kriging Model*

Cross-validation is used to assess the kriging model's generalization capability by withholding subsets of data, predicting values at those locations, and comparing predictions with the observed values. The process involves the following steps:

- Prediction errors are computed by comparing the predicted zinc concentrations with the actual values for the withheld data, providing a direct measure of model accuracy.
- Cross-validation is repeated across different subsets, ensuring that every data point is used for both training and testing, leading to a comprehensive evaluation.
- Key metrics such as **RMSE** (Root Mean Square Error), **MAE** (Mean Absolute Error), and the **correlation coefficient** are calculated to assess prediction accuracy.
- A histogram is generated to visualize the distribution of prediction errors, offering insights into the regions where the model performs well or encounters difficulties.

This cross-validation process ensures a robust evaluation of the kriging model, providing insights into its predictive performance and generalization capability.

### *5.7.2 Residual Error Analysis*

The histogram in Figure 16 shows the distribution of residual errors from the cross-validation of the kriging model. These residuals represent the difference between the predicted and observed Zinc concentrations (in ppm). The histogram provides a visual summary of how well the

model's predictions align with the actual measurements, revealing areas of underestimation or overestimation by the model.

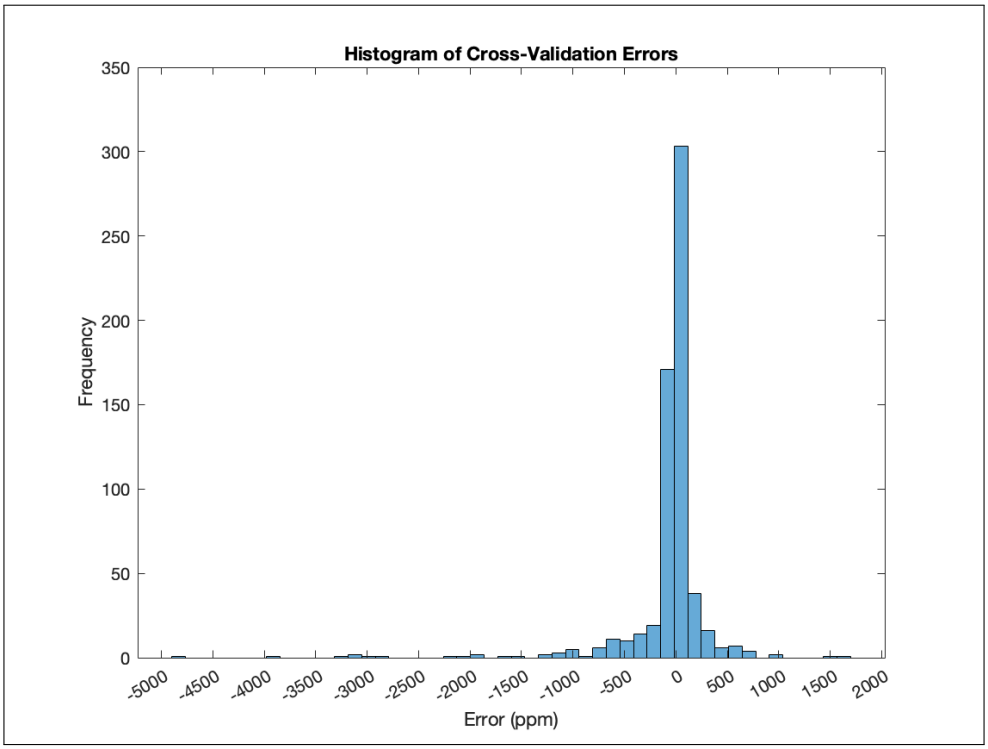


Figure 16: Histogram of residuals from kriging, showing the distribution of prediction errors. Most errors are concentrated around zero, but some large negative residuals (down to -5000 ppm) indicate underestimation in certain areas.

Most of the residuals are in the vicinity of zero; more precisely, in between -500 to 500 ppm. This means that with most data points, the kriging model makes relatively well-accurate predictions. Thus, the concentration of errors around zero could support the hypothesis that the model has been well-calibrated. Nevertheless, there is a significant spread of negative residuals, as low as -5000 ppm. These negative residuals simply show that the model has underestimated the Zinc concentrations in some places. The more extreme underestimations fall between -1000 and -5000 ppm, showing regions of high spatial variability or areas with too sparse data coverage that result in large prediction errors. On the positive side, the residuals do not exceed 2000 ppm. This suggests that overestimations by the model are less frequent and, when they do occur, are typically less extreme than the underestimations. The extreme negative residuals, especially those at approximately -5000 ppm, highlight the particular regions where the kriging model is unable to capture the local variation of Zinc concentrations. These **outliers** could be due to sharp local variations in concentration that the model cannot account for or because of the insufficiency of data points in those areas.

In general, the histogram shows that, for most of the study area, the kriging model performs well, with a majority of the errors around zero. However, large negative residuals indicate that

there is room for improvement, especially in regions of more complicated spatial patterns. Refining the variogram model or increasing data density in areas of high uncertainty may help reduce these large underestimations and improve the overall predictive accuracy of the model.

### *5.7.3 Ordinary Kriging for Spatial Prediction*

After the development of the empirical variogram and the fitting of the **exponential variogram model**, OK has been applied for the estimation in locations where no observations were available.

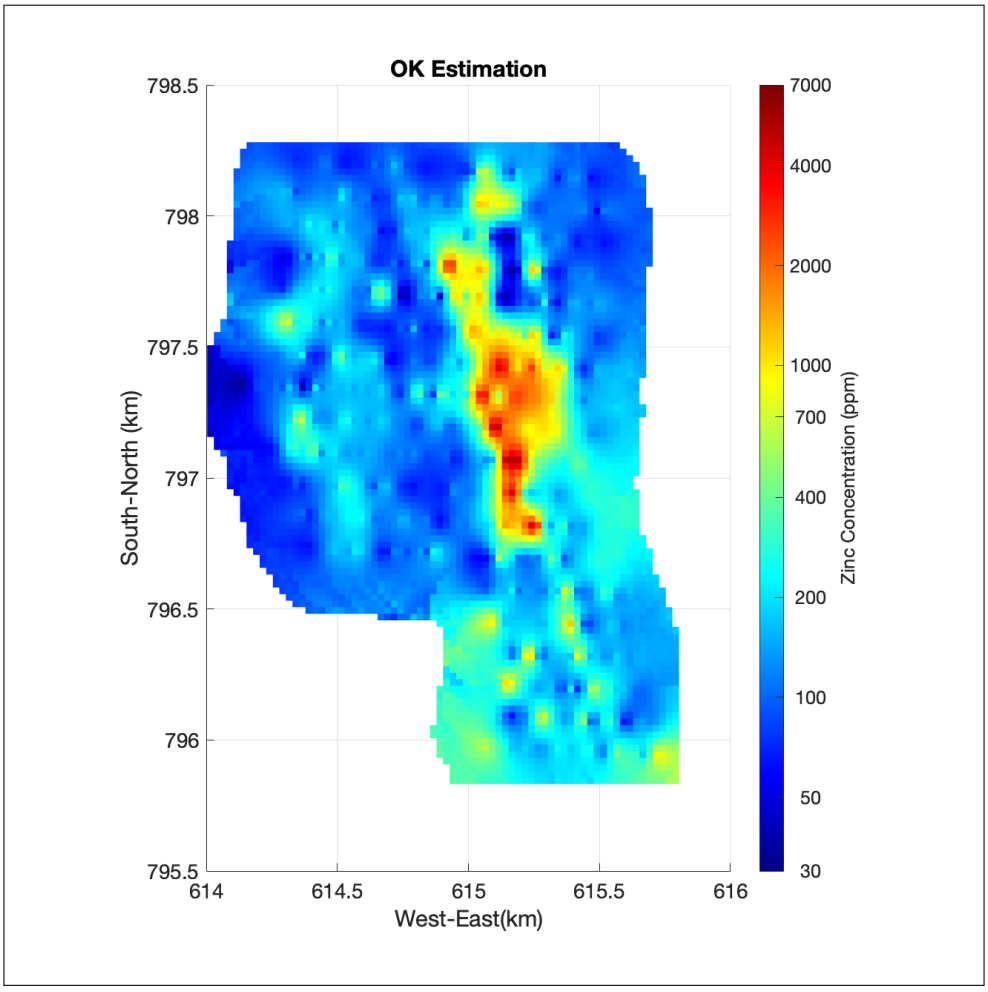


Figure 17: Ordinary Kriging (OK) spatial prediction map of zinc concentrations, based on transformed data. The predictions are derived from the spatial correlation modeled by the fitted exponential variogram, ensuring smooth transitions between low and high concentration zones. The values are presented in parts per million (ppm) after back-transforming from the logarithmic scale; specifically, the estimations are predicted in the logarithmic scale and then converted to ppm.

In this geostatistical interpolation technique, spatial dependence, captured by the variogram is used to make unbiased predictions with the minimum possible estimation variance.

Ordinary Kriging works within the framework of the **Best Linear Unbiased Estimator (BLUE)** so that the predictions are statistically optimal and unbiased. The fitted variogram

model, described by the parameters **nugget**, **sill**, and **range**, forms the basis of quantifying spatial correlation. These parameters are crucial in determining the weights of neighboring data points to achieve spatially coherent and reliable estimates.

The accuracy of the kriging model is validated through cross-validation, and diagnostic plots are analyzed to assess the model's performance. Figure 17 presents a spatial prediction map generated using ordinary kriging, illustrating the spatial distribution of the predicted zinc concentrations across the study area. This map visually distinguishes areas of high and low concentration.

- The **red areas** at the center, corresponding to the highest predicted values, indicate potential **"hotspots"** or zones of high zinc concentration that may warrant further investigation.
- Conversely, **blue areas** towards the borders correspond to regions with lower predicted values, suggesting areas of less significant concentrations.

The map ensures smooth transitions across high and low concentration zones, enabling both local variations and the broader spatial trends to be captured. Such continuity is essential for correct mapping of the spatial distribution of zinc concentrations and makes the map an efficient tool for understanding and visualization of zinc distribution across the area. In general, the ordinary kriging prediction map enables the understanding of the spatial distribution of zinc concentrations, which could be useful in the identification of patterns, potential hotspots, and areas of concern within the study area.

### *5.7.4 Kriging Prediction Uncertainty*

The heatmap shown in Figure 18 visualizes the prediction uncertainty from the kriging model. The colors represent the standard deviation of the kriging errors, which provides a measure of uncertainty for each predicted zinc concentration value. The regions with higher uncertainties suggest the need for additional sampling or refinements in the variogram model in order to reduce the prediction uncertainty and improve confidence in those regions. This heatmap is based on log-transformed zinc concentrations, meaning that both the predictions and their uncertainties are calculated in the log scale. The transformation was done to normalize the distribution of the zinc concentrations, which is otherwise skewed. This is typically done so that the kriging model can provide more accurate predictions and estimates of uncertainty. Because the model is operating in the log-transformed space, the uncertainty values in the heatmap represent variability in the log scale, not in the original concentration values.

- **Low Uncertainty (Blue Areas):** Most of the study area is displayed in blue, indicating a low standard deviation of around 0.3 to 0.5 ppm in the log-transformed zinc concentrations.

These areas likely have dense sampling or a strong spatial correlation, making the kriging predictions more reliable.

- **High Uncertainty (Red and Yellow Areas):** The red and yellow areas, particularly along the edges, show higher prediction uncertainties, with standard deviations reaching up to 1.1 ppm. These areas either have sparse sampling points or high variability in zinc concentrations, leading to less confident predictions.

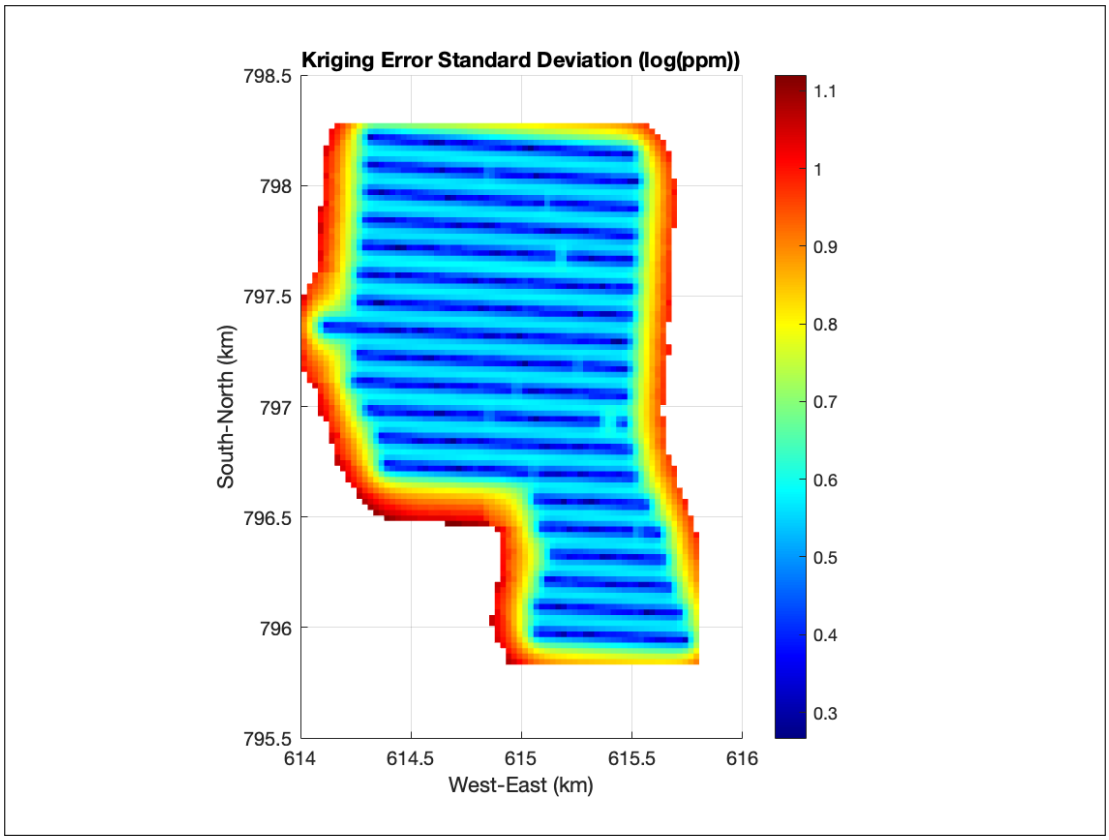


Figure 18: Heatmap of kriging prediction standard deviations, representing uncertainty in the predictions. Lower uncertainties are represented by blue, while higher uncertainties are represented by yellow and red.

The log transformation has an important impact on the uncertainty calculations. Specifically, after applying the logarithmic transformation to the predicted values, the uncertainty fluctuates depending on how the log transformation is handled. The addition or subtraction of uncertainty in the log-transformed space remains in the logarithmic scale, and when exponentiating back to the original scale, this uncertainty can either increase or decrease. The logarithm transforms multiplication and division operations into addition and subtraction, respectively, which affects how uncertainties propagate through the kriging model.

For example, if $\hat{Z}^*(x)$ is the predicted log-transformed zinc concentration at location $x$, and

$\sigma_{\mathrm{OK}}(x)$ is the kriging standard deviation (uncertainty), the uncertainty is calculated as follows:

$$\exp\left[\hat{Z}^*(x) + \sigma_{\mathrm{OK}}(x)\right] = \exp\left[\hat{Z}^*(x)\right] \cdot \exp\left[\sigma_{\mathrm{OK}}(x)\right]$$

This equation shows how the uncertainty is transformed when exponentiating from the log scale back to the original scale. The uncertainty in the log-transformed predictions becomes multiplicative when converting back to the original scale, which explains why it is challenging to directly interpret uncertainties in terms of original zinc concentrations.

# 6 Gaussian Process Regression (GPR)

This chapter investigates Gaussian Process Regression (GPR), a statistical method that bears conceptual similarities to Ordinary Kriging (OK) in predicting unsampled values based on spatial dependencies. Both methodologies incorporate uncertainty into their predictive frameworks.

The chapter presents the findings derived from applying GPR to the zinc concentration dataset and conducts a comparative analysis with OK. Emphasis is placed on evaluating predictive accuracy and quantifying uncertainty to critically assess the strengths and limitations inherent in each approach.

## 6.1 Model Training using Gaussian Process Regression (GPR)

GPR was applied to model the spatial relationships between zinc concentrations and their spatial coordinates. GPR is a flexible, non-parametric regression method that has been very effective in the modeling of complex patterns in spatial data by modeling correlations between data points using kernel functions. GPR assumes that the data points are drawn from a multivariate normal distribution and models the spatial structure of the data through a covariance function. The kernel selection is an important step, since it determines how the zinc concentrations at different spatial locations will be related based on their distance.

In this case, the **ARD Exponential Kernel** was chosen since it suitably models the spatial behavior manifested in these data. The ARD Exponential Kernel models anisotropic spatial dependencies, by allowing different length scales, hence range parameters, for each of the spatial dimensions, hence being similar to the exponential variogram utilized in geostatistics. It allows the correlation between two points to decrease as a function of their separation distance in an exponential form, thus modeling spatial relations on the zinc data quite effectively.

The kernel parameters are signal variance, length scales, and noise variance, which were optimized by using **Maximum Likelihood Estimation (MLE)**. This method gives the best fit to the observed data as it maximizes the likelihood of the data under the Gaussian Process model.

GPR employs the ARD Exponential Kernel to accurately capture the spatial correlation structure and provides robust predictions of zinc concentrations at unsampled locations. It also estimates the uncertainty associated with each prediction, thus providing a comprehensive understanding of the reliability of the results.

### *6.1.1 Gaussian Process Regression (GPR) Plot Analysis*

Figure 19 presents a GPR output for spatial predictions of zinc concentrations. Large concentrations can be observed on the map, mainly on the central area of the region, with hot spots having a greater concentration in this particular hotspot. This high central value thus points to some substantial accumulation, which would probably have important reasons for such concentration and could therefore potentially be studied more deeply. Surrounding this hotspot are areas of lower concentration, showing spatial variability across the landscape.

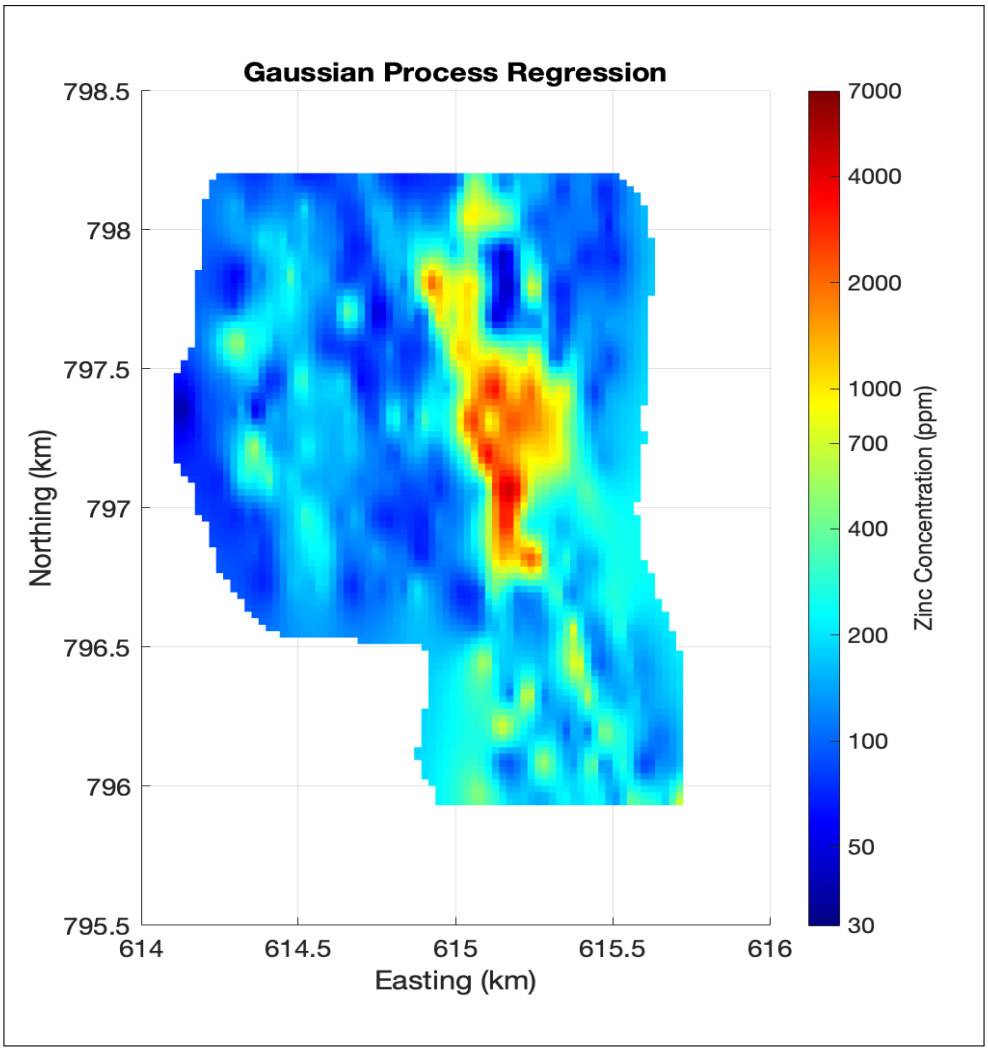


Figure 19: Spatial prediction map of zinc concentrations using Gaussian Process Regression (GPR). The plot use a logarithmic scale for clearer differentiation across concentrations, while the colorbar is displayed on the original concentration scale, ranging from 30 ppm to 7000 ppm. This visualization highlights spatial patterns, showing high-concentration "hotspots" and low-concentration "coldspots", enhancing the interpretation of spatial variability.

The map shows high values of zinc, especially in the center of the study area where high values are realized. This central hotspot could signal a significant accumulation of zinc and that the underlying factors contributing to such a concentration may be of further investigation interest.

Surrounding this central high value area, lower concentration areas are realized, indicating spatial variability across the landscape. In order to perform the rigorous evaluation of predictive accuracy for the GPR model in the determination of zinc concentrations within the study area, LOOCV was employed. This method involves the systematic omission of each data point in turn from the training set, allowing the GPR model to be trained on the remaining data and then used to predict the value at the excluded location. Such a process enables an overall and unbiased evaluation of prediction error across individual data points, thus increasing confidence in the model's ability to generalize well. The integration of LOOCV within GPR is particularly useful in spatial analysis to ascertain that the identified spatial patterns, such as regions of high concentration of zinc ("hotspots"), are representative of real trends and not possibly overfitting. Moreover, LOOCV offers insight into the model's spatial parameterization and guides refinements toward a more accurate and reliable representation of spatial variability in zinc concentrations.

## 6.2 Spatial Difference Visualization: Comparison of Ordinary Kriging and Gaussian Process Regression

Figure 20 illustrates the spatial distribution of differences in zinc concentration predictions (in ppm) between the models generated by Ordinary Kriging (OK) and Gaussian Process Regression (GPR).

The differences in predicted zinc concentrations speak volumes about the performance of each model. GPR yields considerably higher predictions than OK for the central parts of the study area where dark red values are dominant. This pattern may indicate that GPR is very efficient at capturing local variation in zinc concentration, which thus reflects the responsiveness of this model to small-scale spatial trends.

On the contrary, peripheral areas in blue show minimal differences, where OK's predictions are slightly higher than those from GPR. This effect is enhanced in areas with fewer data points, where OK's fixed variogram model yields high estimates compared to the more adaptive approach of GPR.

The contrasting predictions between OK and GPR underline the differences in underlying assumptions and methodologies. OK makes use of a fixed variogram, implying stable spatial autocorrelation, which results in smoothed predictions. On the contrary, GPR uses a nonparametric, adaptive covariance structure that allows it to make variable spatial characteristics within the data.

These results show that the differences between the models do not only reflect their comparative strengths and weaknesses but are partly due to limitations in available data density and

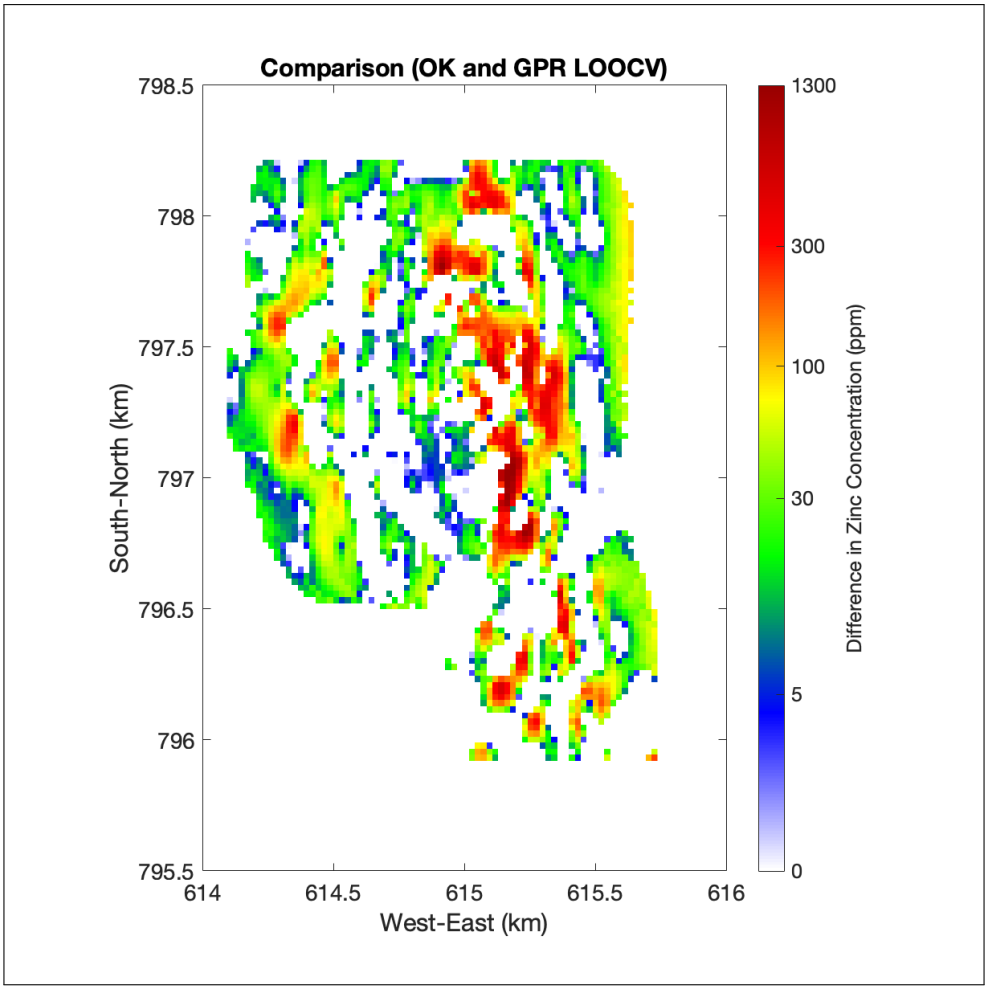


Figure 20: Spatial differences in zinc concentration predictions (ppm) between Ordinary Kriging (OK) and Gaussian Process Regression (GPR). Warmer colors indicate regions with greater differences in predictions, where GPR tends to predict higher zinc concentrations than OK. Cooler colors reflect minimal differences, suggesting that OK predictions are slightly elevated. The color bar illustrates that differences range from approximately 0 ppm in the blue areas to values close to 1300 ppm in the dark red areas.

partly due to the potential discrepancy between predicted and true values. This further underlines the relevance of integrating true concentration values into the evaluation process in order to fully understand the reliability and accuracy of spatial predictions.

## 6.3 Scatter Plot Comparison of Predictions with Observed Zinc Concentrations

The scatter plot compares the predictive performance of OK and GPR against the observed zinc concentrations measured in ppm. The $y = x$ line represents perfect prediction and has been included as a benchmark. OK performs strongly at lower to moderate concentrations, $< 2000$ ppm, with predictions closely lying on the observed values, a reflection of its ability to produce stable and smooth estimates. However, OK overestimates at higher concentrations (>3000 ppm),

as seen by upward deviations from the ideal line. In contrast, GPR gives a higher variability for all concentrations while it captures the localized spatial trend and finer heterogeneity better. However, GPR has wide deviations at extreme concentrations (>3000 ppm) and hence is not good at generalizing from sparse or highly variable data. Overall, the scatter plot illustrates how OK tends to capture broader spatial trends with consistency, while GPR captures more detailed variability but perhaps at a higher uncertainty for the extreme values.

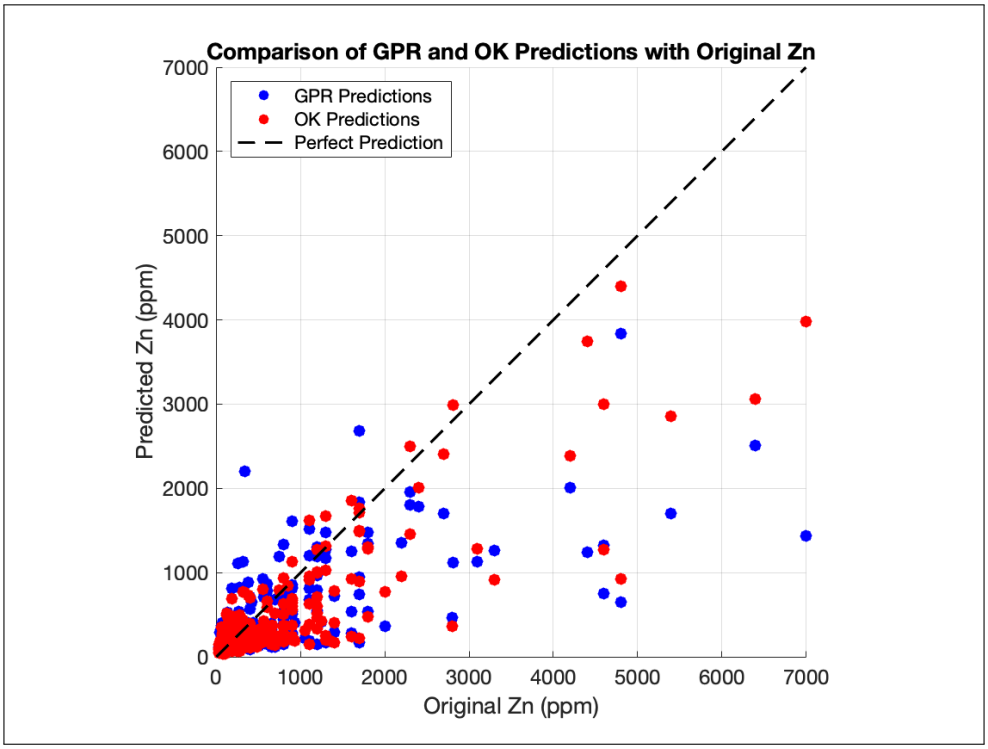


Figure 21: Scatter plot comparing original zinc concentrations (ppm) with predictions from OK and GPR. Blue points represent GPR predictions, red points represent OK predictions, and the black dashed line represents perfect prediction ($y = x$).

## 6.4 Model Performance Comparison: Ordinary Kriging vs. Gaussian Process Regression

The results in Table 6 reveal key distinctions between Ordinary Kriging and Gaussian Process Regression (GPR) in terms of predictive accuracy, average error, and error correlation. The findings are summarized as follows:

- **Root Mean Square Error (RMSE):** Ordinary Kriging demonstrates a lower RMSE (485.79) compared to GPR (532.14), indicating superior performance in minimizing squared errors across the dataset.
- **Mean Absolute Error (MAE):** Similarly, Ordinary Kriging achieves a slightly lower

Table 6: Performance Metrics of Ordinary Kriging and Gaussian Process Regression (GPR)

| **Metric** | **Ordinary Kriging** | **Gaussian Process Regression (GPR)** |
|---|---|---|
| RMSE (Root Mean Square Error)(ppm) | 485.79 | 532.14 |
| MAE (Mean Absolute Error)(ppm) | 186.97 | 190.46 |
| Correlation Coefficient | 0.76 | 0.73 |
| E20 (Error at 20th Percentile) | 31.59% | 32.54% |
| E80 (Error at 80th Percentile) | 84.04% | 85.31% |

MAE (186.97) than GPR (190.46), suggesting that Ordinary Kriging yields smaller average prediction errors overall.

- **Correlation Coefficient:** Ordinary Kriging exhibits a marginally stronger correlation with observed values, with a coefficient of 0.76 as compared to GPR's 0.73. This suggests that Ordinary Kriging better captures the linear relationship between predicted and actual values.

- **Percentile Errors (E20 and E80):** Both Ordinary Kriging and GPR exhibit similar performance in terms of percentile errors. GPR shows an E20 of 32.54% and an E80 of 85.31%, while Ordinary Kriging achieves an E20 of 31.59% and an E80 of 84.04%. This indicates that both methods perform comparably in predicting values within these specified percentile ranges.

### *6.4.1 Summary of the comparison OK-GPR*

OK and GPR are complementary methods of spatial prediction, with each method having different strengths and areas to which it can be applied. OK is a modified version of the Best Linear Unbiased Estimator (BLUE), thus having strengths in capturing large-scale spatial trends by yielding smooth and stable predictions. The fixed variogram model is stable, as demonstrated by the lower RMSE and MAE of OK and a slightly higher correlation coefficient of OK compared to GPR. However, OK tends to overestimate the concentrations for higher values (>3000 ppm).

In contrast, GPR with its Bayesian framework and adaptive covariance structure is particularly effective in capturing localized spatial heterogeneity and fine-scale variability. While GPR performs well when there is a high density of data, it introduces greater variability in the results, especially in poorly sampled areas or when one attempts to predict extreme values. It is this complementary nature-OK provides robust and generalized predictions over the study area while GPR captures localized patterns of variability-that is highlighted from the scatter plots and spatial difference maps.

Both methods behave similarly in terms of the percentile errors E20 and E80, showing their

effectiveness for general prediction. Together, these methods can offer a balanced approach to understanding spatial distribution patterns by combining precision and generalization to meet varied analytical objectives. Principles and tailored adjustments are applied to the dataset.

# 7 Self-Organizing Map (SOM) Method

The SOM algorithm is a family of unsupervised neural network techniques that projects high-dimensional input data nonlinearly onto a usually two-dimensional, structured grid. In particular, SOMs are well suited for visualization of complex datasets and clustering of spatially correlated data since the mapping preserves topological relationships in the data while capturing interesting underlying patterns. In this paper, a three-dimensional SOM is used for the analysis of the spatial distribution of zinc concentrations across the study area. Spatial coordinates are combined with zinc concentration as input features to ensure that both spatial and attribute-based patterns are effectively captured in the clustering process.

## 7.1 Evaluating and Selecting the Optimal Map Size

Various SOM grid configurations were evaluated to ensure robust clustering and good representation of the dataset. The goal was to find the map size that best preserved the spatial relationships within the dataset while minimizing error. Each map size was evaluated based on two primary performance metrics:

- **Quantization Error (QE):** QE is calculated by the average Euclidean distance of each data point to the closest neuron or the Best Matching Unit, which means the lower the QE is, the more the neurons in this grid effectively catch and represent the intrinsic features of the data for the purpose of clustering.
- **Topographic Error (TE):** Topographic Error represents the measure of how well SOM preserves the topological structure of the dataset. It is the proportion of data points for which the first and second closest neurons (Best Matching Units) are not direct neighbors. A smaller TE would suggest that the map preserves the spatial continuity of the data, which is important in correctly carrying out spatial pattern recognition.

To identify an optimal balance between QE and TE, SOM map sizes ranging from $4 \times 5$ to $12 \times 14$ were tested. Table 7 summarizes the results for each tested configuration.

The following trends were observed from the map size evaluations:

- **Smaller Map Sizes**: Configurations such as $4 \times 5$ yielded relatively high QE (0.3285) and moderate TE (0.6256), suggesting that smaller grids lack the capacity to represent

Table 7: Quantization Error (QE) and Topographic Error (TE) for Different SOM Map Sizes.

| Map Size | Quantization Error (QE) | Topographic Error (TE) |
|---|---|---|
| $4 \times 5$ | 0.3285 | 0.6256 |
| $5 \times 6$ | 0.2703 | 0.6714 |
| $5 \times 7$ | 0.2515 | 0.7425 |
| $5 \times 8$ | 0.2394 | 0.7030 |
| $10 \times 10$ | 0.1489 | 0.7283 |
| $10 \times 12$ | 0.1316 | 0.7409 |
| $12 \times 14$ | 0.1068 | 0.7709 |

the dataset's complexity effectively. Such grid sizes oversimplify the data, leading to poor clustering quality.

- **Medium Map Sizes**: The $5 \times 6$ grid achieved a Quantization Error of 0.2703 and a Topographic Error of 0.6714. This configuration demonstrated an effective balance between minimizing quantization error and preserving the spatial topology. It captured both localized and large-scale concentration patterns in the study area without excessive computational overhead.
- **Larger Map Sizes**: Grids such as $10 \times 10$ and $12 \times 14$ achieved lower QE (0.1489 and 0.1068, respectively), indicating improved data representation. However, their higher TE values (0.7283 and 0.7709) suggest a decline in topological consistency. Additionally, larger map sizes demand more computational resources, making them less practical for this study.

These results suggest that the $5 \times 6$ grid size is the best compromise between the accuracy of data representation and topological preservation. This provides an adequate resolution for the spatial analysis without superfluous computational overhead.

## 7.2 Training Progress and Performance

With the optimal map size determined, the SOM was trained on the spatial coordinates of zinc concentration data. During training, each data point was matched to its Best Matching Unit (BMU) on the grid. The BMU's weights, as well as those of its neighboring neurons, were iteratively updated to approximate the input data. This adaptation process allowed the SOM to gradually organize similar spatial coordinates into clusters, representing distinct zinc concentration patterns.

The training progress of the SOM, as shown in Figure 22, summarizes the performance of the model during the learning process applied to the zinc concentration dataset.

- **Epoch Progress:** The training was conducted over a specified number of epochs (200), during which the SOM progressively adapted to the spatial zinc concentration data. The

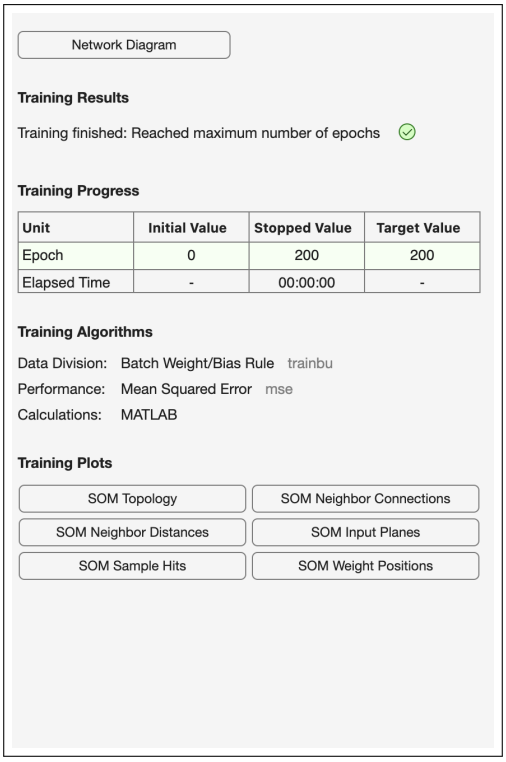


Figure 22: SOM training progress and key performance metrics for zinc concentration data, indicating successful convergence of the model.

green checkmark in the "Training Results" section indicates that the model has converged and successfully organized the input data into clusters.

- **Performance Metric:** The Mean Squared Error (MSE) is used as a performance metric, reflecting the difference between the observed zinc concentration data and the weights of the neurons after training. A low MSE indicates that the SOM has effectively learned the spatial structure of the data.
- **Training Algorithm:** The Batch Weight/Bias Rule (*trainbu*) was employed to update the weights and biases of the neurons. This algorithm is well-suited for adjusting large-scale weight vectors, making it efficient for handling the spatial patterns in the zinc concentration data.

**A series of plots** are used to evaluate the SOM's performance, illustrating how well the model organizes the zinc concentration data:

## SOM Topology

The Self-Organizing Map (SOM) topology plot is a visual representation that provides insights into the structure and arrangement of the SOM. This plot creates a visual display of the map's neurons and their connections.

The SOM topology plot in Figure 23 provides a visual representation of the structure of the SOM:

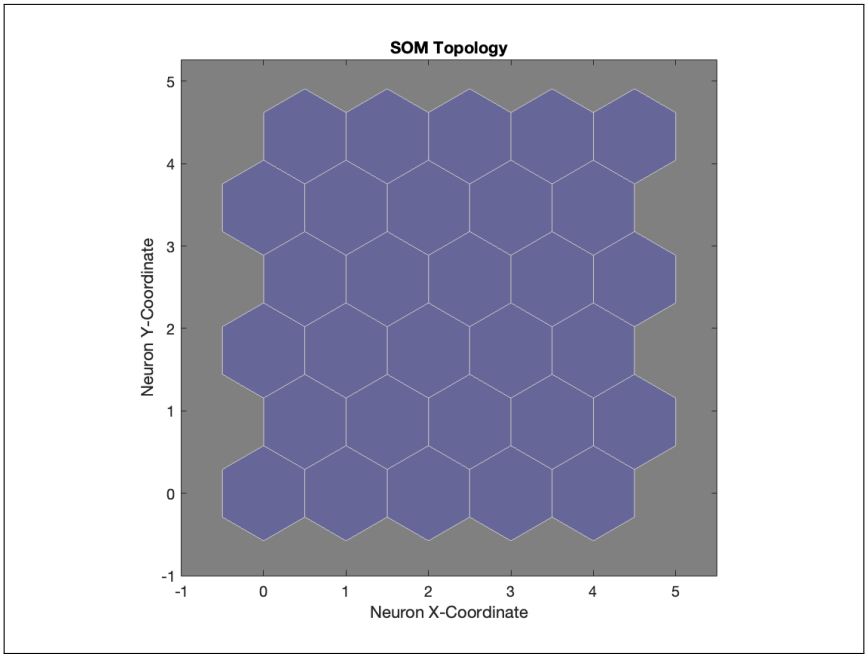


Figure 23: SOM Topology plot showing the arrangement of neurons and their relationships within the grid. Neurons that represent similar data points are closer together, forming clusters that reflect patterns in the input data.

- **Hexagonal Grid:** The neurons in the SOM are laid out on a hexagonal grid. This is usually preferred because, for the same number of neighbors, it offers more natural and flexible neighborhood structures than a rectangular grid. Each hexagon represents a neuron in the SOM, and the hexagonal layout helps to maintain continuity and relationships between neighboring neurons.

- **Neuron Arrangement:**Each hexagon in the grid corresponds to a neuron. This is very important for preserving the topological properties of the input data. More similar input data points are mapped to nearby neurons in the grid, which means that spatial relationships and patterns in the data are preserved.

- **Topological Relationships:** The plot helps visualize how neurons are connected and how the map is organized to reflect the spatial relationships in the input data. This structure aids in the clustering process, ensuring that similar data points are grouped together. The

neurons' positions relative to each other indicate how the input space is partitioned into regions of similar data points.

- **Color Coding:** In this specific plot, the uniform color of the hexagons indicates that the focus is on the structural arrangement rather than the data values associated with each neuron. In other contexts, color coding might be used to represent different data values or clusters associated with the neurons.

The proximity of neurons in the grid corresponds to the similarity of input vectors, making the plot valuable for cluster identification and spatial pattern recognition. The SOM topology serves as a validation tool by enabling the evaluation of how well the map has captured the structural properties of the dataset. Moreover, it reflects the underlying training dynamics of the SOM, providing insight into how the neurons have adjusted during the learning process.

## SOM Neighbor Connections

The SOM neighbor connections plot visualizes the topological relationships between neurons in the Self-Organizing Map (SOM) after training.

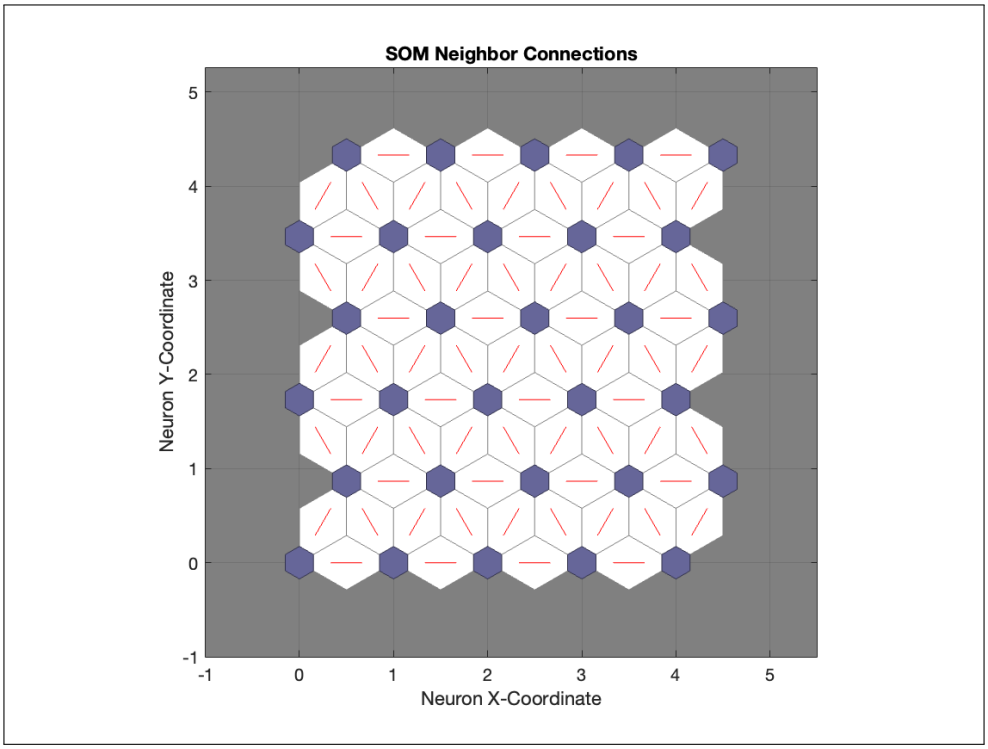


Figure 24: SOM Neighbor Connections plot, illustrating the topological relationships between neurons in the SOM grid. The red lines represent the connectivity between neurons, reflecting the similarity between their weight vectors and the proximity of the input data they represent.

Each blue circle of any given neuron represents one cell of the cluster in an input data space,

according to the neurons representing different dispersions of the points onto the grid. The red lines between neurons represent the connections between neighbour neurons based on their vicinity in the data space. Neurons with closer connections will have more similar weight vectors, which means they represent similar regions of the input space. The hexagonal grid does not distort the topological properties of the input data, ensuring that similar data points will be mapped to neurons that are close in the grid. Areas with fewer connections or neurons indicate regions in the input space where the data is less dense. These sparse regions suggest lower data concentration in the corresponding parts of the feature space.

The closeness of neurons in a grid, as well as their strength in being connected, offer insight into how well the SOM has grasped relationships within data. It visually aids the process of identification of clustering and regions with similarity in their respective data points by preserving much of the overall input space.

## SOM Neighbor Distances

The SOM neighbor distances plot visualizes the distances between neighboring neurons, which aids in understanding the map's structure and how clusters of data points are formed. The plot shown in Figure 25, the distances between neurons are color-coded, providing a clear view of the spatial relationships.

The plot represents the pairwise distances between neurons:

- **Larger distances**, represented by darker shades, indicate regions where the data mapped to neighboring neurons are more dissimilar. These regions typically correspond to cluster boundaries or transitions between different groups of data points.
- **Smaller distances**, shown by lighter colors, indicate a high degree of similarity between neighboring neurons. This reflects areas where the input data points are homogeneous or continuous.

By examing at the distances between neighboring neurons, one can see how well the SOM has managed to represent the input space. Large distances highlight well-separated clusters, which is important in the process of identifying regions of dissimilarity in data. Smaller distances, on the other hand, represent continuous or homogeneous regions where data points exhibit a high degree of local similarity.

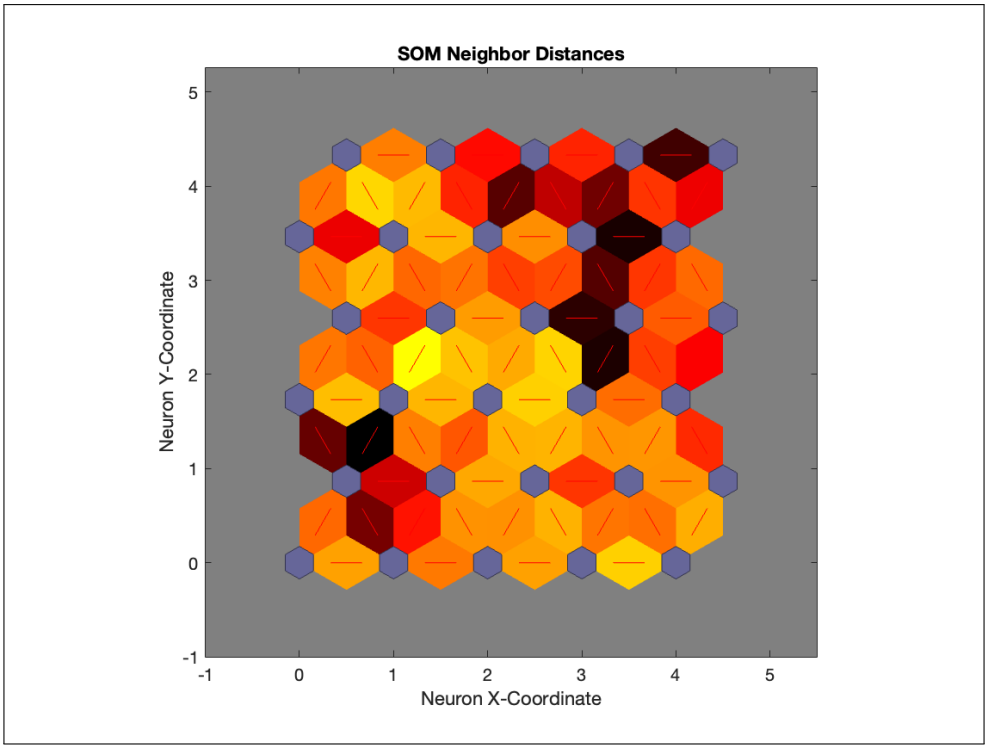


Figure 25: SOM Neighbor Distances plot, where darker colors represent larger distances between neighboring neurons. This plot helps in identifying clusters and boundaries within the data, with closer neurons indicating similar data points.

## SOM Input Planes

The Self-Organizing Map (SOM) input planes provide a visual representation of how each input feature contributes to the neurons in the SOM grid. Each input plane corresponds to a single input variable and illustrates how the values of that variable are distributed across the neurons. The color of each hexagonal cell represents the weight associated with a specific input feature for that neuron. Darker colors typically indicate higher weights, meaning the corresponding input feature has a greater influence in those regions of the SOM grid.

### *Interpreting Individual Input Planes*

In Figure 26, each input plane represents one input feature:

- **Input Plane 1:** The first input plane visualizes the influence of the first feature (spatial coordinate $X$) across the SOM grid. Neurons with darker colors (e.g., red or black) indicate regions where this feature has higher values, representing areas strongly influenced by this input variable.
- **Input Plane 2:** The second input plane shows the distribution of the second input feature (spatial coordinate $Y$). Neurons with darker colors in this plane highlight areas where the

second input feature plays a significant role.

- **Input Plane 3: Zinc Concentration:** The third input plane demonstrates the contribution of zinc concentration values. Darker regions correspond to higher zinc concentrations, emphasizing areas with higher zinc levels in the dataset.

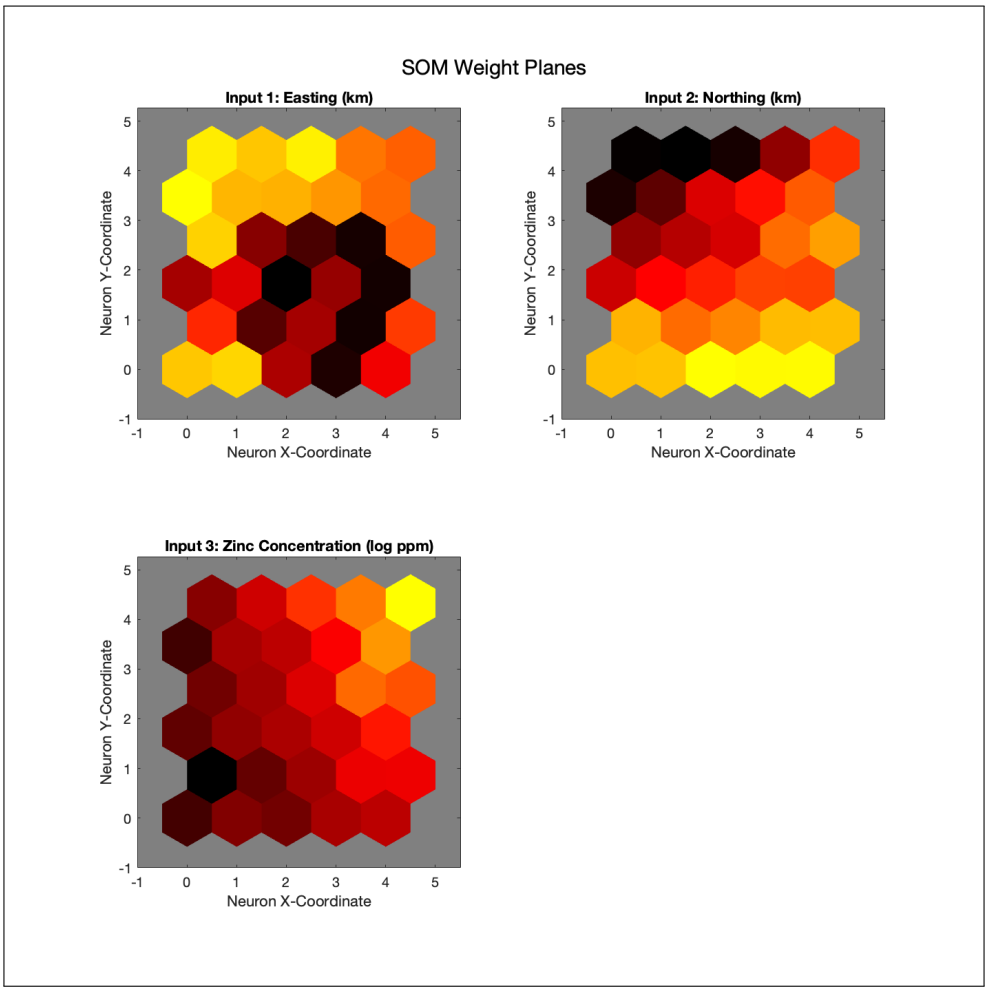


Figure 26: SOM Input Planes showing the mapping of different input variables to the SOM neurons. Each plane corresponds to an individual input variable, helping to visualize how the variable's values are distributed across the SOM grid. The third input plane corresponds to the zinc concentration.

### *Comparing Input Planes*

By comparing the input planes, we can gain insights into the relationship between features. Regions where several planes all show similar patterns-for example, darker colors in the same neurons-indicate that these features are strongly correlated, driving the same regions of the SOM grid. Areas where one plane shows dark regions while another shows light regions suggest that the features drive different neurons, highlighting their independence in the dataset.

### *Purpose of SOM Input Planes*

The input planes are a powerful tool for understanding and validating the SOM model. Input planes highlight which features are most influential in shaping the SOM. Variability across an input plane suggests the corresponding feature plays a significant role in clustering the data.Patterns in the input planes often correspond to clusters, where certain neurons respond strongly to specific feature values. Comparing input planes reveals correlations or contrasts between features, offering insights into the interplay between input variables.

## SOM Sample Hits Plot

The Self-Organizing Map (SOM) Sample Hits Plot, shown in Figure 27, illustrates the frequency with which each neuron in the SOM grid was selected as the Best Matching Unit (BMU) during the training phase. Each hexagonal cell represents a neuron, and the number inside each hexagon indicates how many input data points were mapped to that neuron as their closest match.

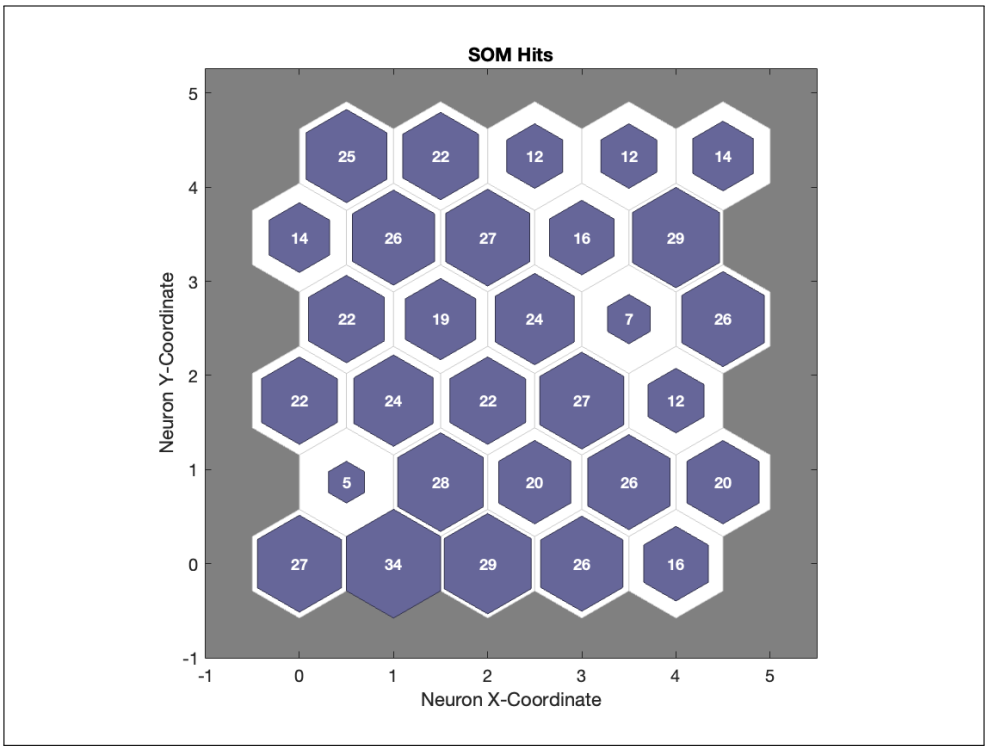


Figure 27: SOM Sample Hits Plot showing the number of times each neuron was selected as the Best Matching Unit (BMU) during training. Neurons with higher hit counts represent regions with dense data, while neurons with lower hit counts correspond to sparsely populated regions in the data space.

- **High Hit Counts:** Neurons with high hit counts, such as those labeled 34, 32, and 29, are observed in the lower and central areas of the SOM grid. These neurons correspond to densely populated regions of the input data space, indicating areas where many input data

points share similar characteristics.

- **Low Hit Counts:** Neurons with lower hit counts, such as 7, 12, and 14, are observed in the upper areas of the SOM grid. These regions represent sparsely populated areas in the input data space, where fewer input data points have similar features.
- **Cluster Distribution:** The distribution of hit counts across the SOM grid provides insights into the clustering of input data. Areas with concentrated high hits correspond to dense clusters in the data, while regions with lower hit counts indicate sparsely distributed data points or outliers in the dataset.

### *Importance of the SOM Sample Hits Plot*

The Sample Hits Plot serves several purposes in analyzing and validating the SOM. It verifies whether the SOM accurately reflects the density and distribution of the input data. Dense data regions in the input space should correspond to neurons with high hit counts in the plot. Moreover, the plot highlights areas of the SOM grid where clusters of similar data points are formed. Dense clusters are indicated by neurons with consistently high hit counts in neighboring regions. Neurons with low hit counts indicate areas where the SOM grid represents sparsely populated or less significant regions of the input data space. This can help identify potential outliers or underrepresented areas in the dataset.
A well-trained SOM will show an even distribution of hits in its neurons, with the correct representation of actual density of input data. Those regions which show high counts correspond to a very well-defined cluster in the input data, while neurons showing lesser hits correspond to those parts of data space which are poorly populated. The implication is that this distribution implies the capture of the pattern and structure within the dataset.

## SOM Weight Positions and Their Significance

The SOM Weight Positions plot in Figure 28 is a critical tool for evaluating how well the Self-Organizing Map (SOM) has adapted to the input data. It visualizes the final positions of the SOM neurons' weight vectors in the input space and their relationship to the original data points. This plot provides insights into how effectively the SOM represents the data's structure and density, as well as how well it preserves topological relationships.

This plot helps in several ways. First, the proximity of the blue circles (neurons) to the green points (input data) indicates how accurately the SOM has captured the distribution of the dataset. Neurons located close to clusters of green points demonstrate that the SOM has

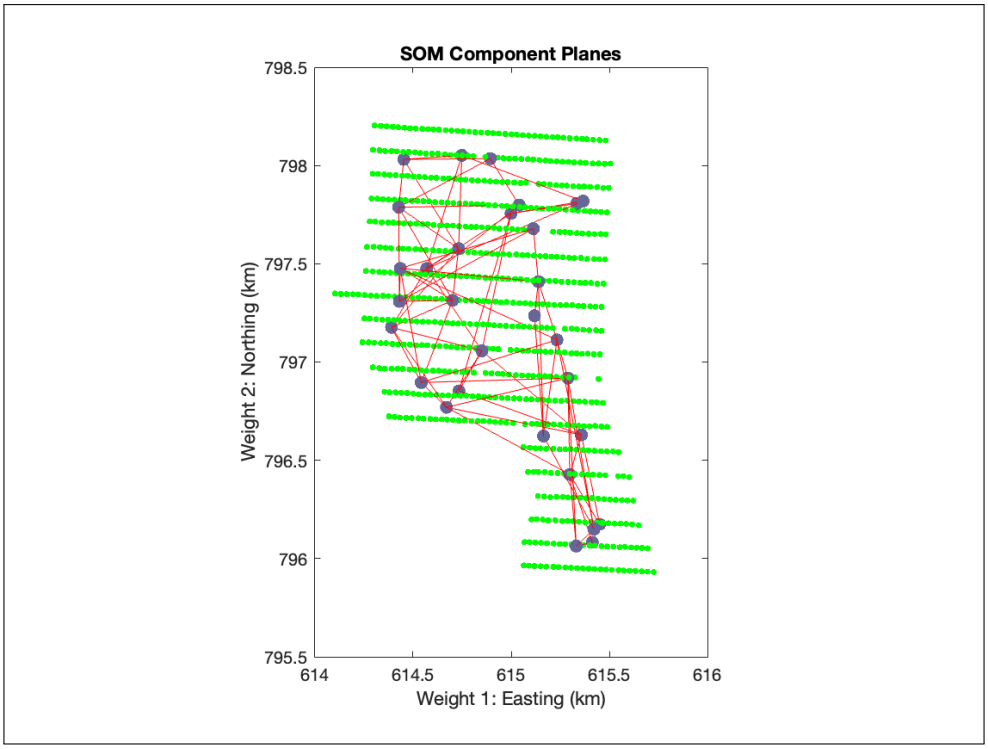


Figure 28: SOM Weight Positions plot showing the final locations of the neurons' weight vectors in the input space. The green points represent the input data, blue circles represent the SOM neurons (weight vectors), and red lines indicate the connections between neighboring neurons.

successfully learned and represented the dense regions of the input data. Conversely, sparse regions of green points highlight areas where fewer neurons are needed, confirming that the SOM adapts to data density.

Second, the red lines connecting neighboring neurons provide a visual representation of the SOM's topological structure. These lines illustrate the relationships between neurons, showing how similar or different the regions they represent are. Shorter red lines indicate areas where neurons are closely related, corresponding to homogenous or clustered regions of the input data. In contrast, longer red lines may highlight transitions between different clusters or regions of sparse data, revealing areas where the input data changes significantly.

This plot is also very informative for detecting under- or over-representation of the input space. For instance, if blue circles are too far from green points, this may indicate that the SOM did not catch those regions of the input space. On the other hand, regions with excessive overlap of blue circles may indicate redundant neurons and thus overfitting to some parts of the dataset.

In summary, the SOM Weight Positions plot verifies the strength of the SOM in learning structural and density aspects of the input while maintaining topological relationships. In fact, the arrangement of neurons should reveal whether the SOM was well-trained, give more details about the distribution of a dataset, the relations between cluster data, and what may be the weakest point to be improved. This thus acts like a bridge between input data and the SOM

model by providing an overall insight into the performance of SOM.

## 7.3 OK Estimation with SOM Neuron Grouping

The **OK Estimation with SOM Neuron Grouping** (Figure 29) visualizes the application of Ordinary Kriging (OK) to predict zinc concentration across the study area, utilizing a Self-Organizing Map (SOM) to group the data into neurons.

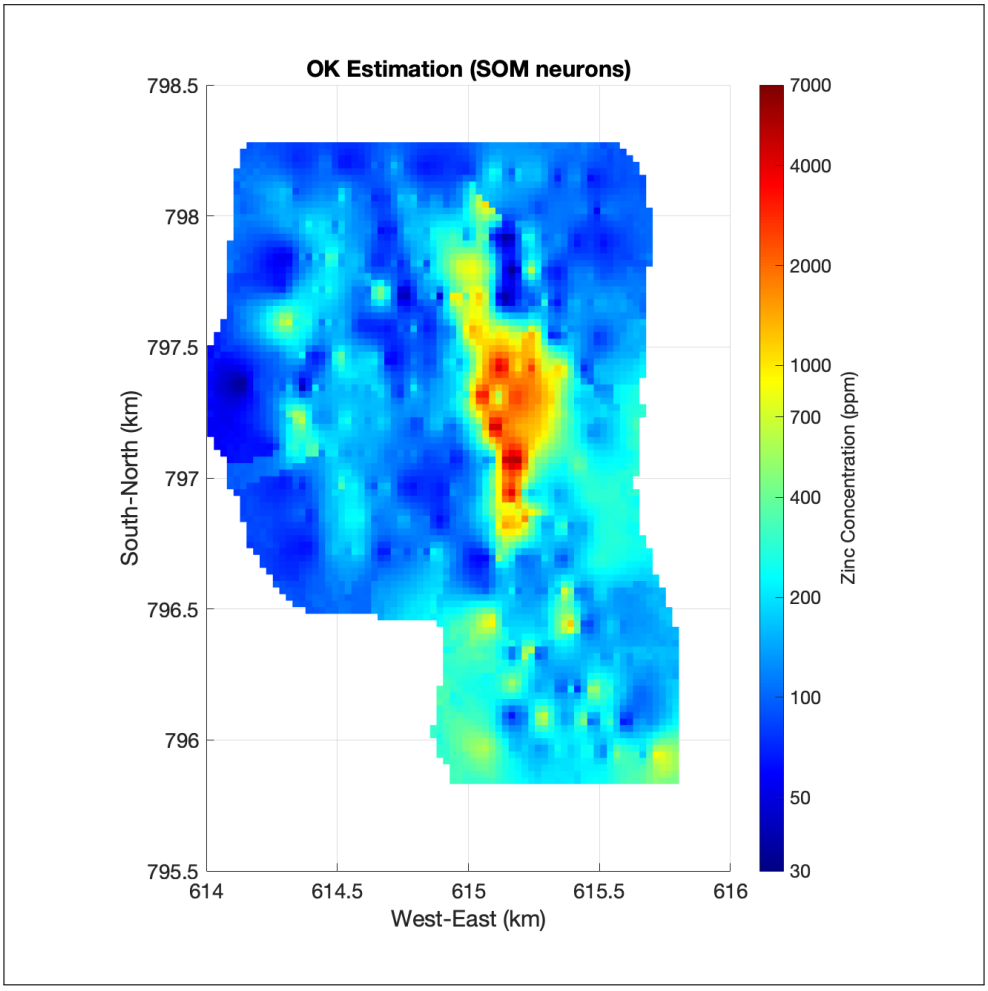


Figure 29: OK Estimation with SOM Neuron Grouping showing the spatial distribution of zinc concentration (ppm) across the study area. The plot visualizes estimated zinc concentration, with areas of higher concentration indicating potential mineralization hotspots. Values are presented in parts per million (ppm) after back-transforming from the logarithm; specifically, the estimations are predicted in the logarithmic scale and then converted to ppm, as shown in the figure.

This method allows for spatial predictions of zinc concentration in parts per million (ppm) after back-transforming from the logarithmic scale. That is, estimations are predicted in the logarithmic scale and then converted to ppm, as shown in this figure.

This plot shows the kriging estimation, where regions of high zinc concentration are well reflected, especially in the central part, indicating the probable mineralization hotspots that

deserve further investigation. These reflect considerable accumulations of zinc, which imply favorable geological conditions for mineral deposits. On the other hand, the peripheral parts have lower concentrations, and hence, the chances of considerable mineralization are low.

The Ordinary Kriging method interpolates the zinc concentration at unsampled locations based on the spatial autocorrelation of the measured data points; this allows the production of continuous predictions over the whole study area, including in areas with no direct measurement. Moreover, the grid structure shown in this plot was also influenced by the SOM neuron grouping that divided the study area into distinct clusters-neurons-before applying kriging. This grouping ensures that the estimation considers local characteristics of each neuron, thus increasing the accuracy and reliability of the predicted zinc concentrations throughout the region.

By providing a continuous spatial prediction, the plot facilitates the identification of areas with elevated or reduced zinc concentration. The plot assists in pinpointing specific areas of interest for further exploration, particularly regions with high estimated zinc concentrations that may warrant closer investigation.The integration of OK with SOM neuron grouping reveals spatial variability in zinc concentration, highlighting potential local anomalies and broader regional trends.

### *7.3.1 Methodology of OK Estimation with SOM Neurons*

The methodology for OK estimation using SOM neurons involves:

1. **Data Grouping with SOM:** Apply SOM to cluster data points into neurons based on similarity, organizing spatial data into a grid for local pattern capture.

2. **Selecting Nearest Neurons:** For each prediction location $(X_p, Y_p)$, identify the nearest neurons. Use data from these neurons to inform the kriging estimation, ensuring local influence.

3. **Ordinary Kriging with Localized Data:** Apply OK using data from selected neurons. Model spatial correlation with a variogram to weight data points according to their distance from $(X_p, Y_p)$, yielding localized predictions.

4. **Generating the Spatial Prediction Map:** Create a continuous zinc concentration prediction map by applying kriging across all target locations. Visualize as a heatmap with values expressed in parts per million (ppm) for clarity.

### *7.3.2 Aggregate Data from Nearest Neurons*

The decision to aggregate data from the 5 nearest neurons when a neuron has no data. Without this step, the kriging process would fail at locations where the corresponding neuron lacks data. Aggregating data from nearby neurons ensures that predictions are continuous across the study area, avoiding gaps in the spatial prediction map. By incorporating data from nearby neurons, the method captures local spatial patterns more effectively. This is particularly important in sparsely populated areas, where relying on a single neuron may not adequately reflect the local variability in zinc concentration.Sparse regions in the dataset can result in empty neurons, leading to unreliable or incomplete predictions. Aggregating data from the 5 nearest neurons mitigates this issue by leveraging information from neighboring regions, ensuring robust predictions even in areas with fewer data points.

This integrated approach leverages kriging's ability for smooth spatial predictions and SOM's capacity for localized pattern recognition, ensuring accurate predictions that are sensitive to local data variations.

## 7.4 Spatial Difference Visualization: Ordinary Kriging vs. SOM-Based Grouping

Figure 30 presents the spatial distribution of prediction differences in zinc concentrations (ppm) between **Ordinary Kriging (OK)** and **SOM-Based Grouping**. The color scale ranges from small differences (near 0 ppm) to substantial discrepancies (up to 1300 ppm). Warmer colors (yellow to red) highlight areas where SOM-based predictions exceed OK, while cooler colors (blue) indicate the reverse.

- **Regions with Higher SOM Predictions:** Peripheral zones, particularly along the eastern and northern edges, display warmer hues, signifying higher SOM predictions. This reflects SOM's strength in capturing localized variations and spatial heterogeneity, emphasizing finer details that OK tends to smooth.
- **Regions with Higher OK Predictions:** Central areas and some isolated zones exhibit cooler hues, indicating higher OK predictions. This aligns with OK's characteristic of producing smoother, conservative estimates that average out the localized details emphasized by SOM.
- **Overall Observation:** The visualization underscores the predictive tendencies of the two methods. OK provides smoother, continuous predictions suited for broader regional analy-

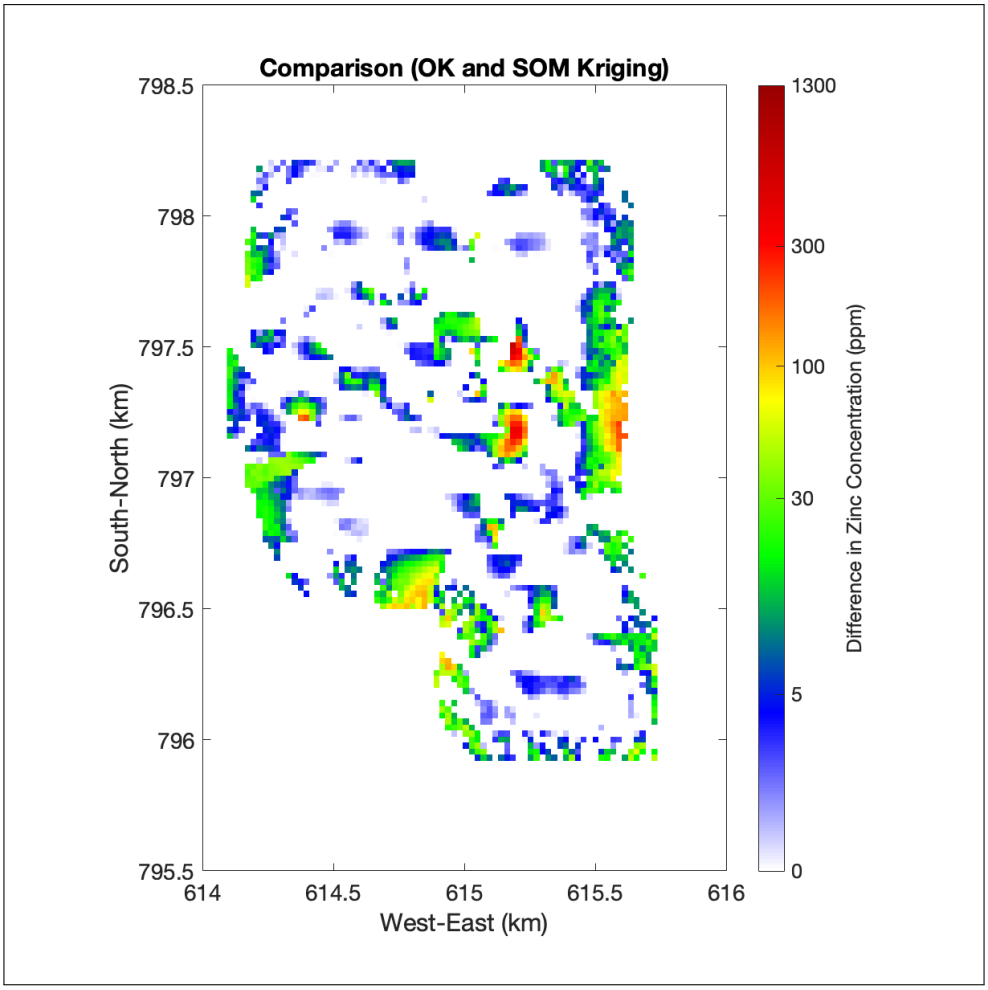


Figure 30: Spatial differences in zinc concentration (ppm) between Ordinary Kriging (OK) and SOM-Based Grouping. Warmer colors (yellow to red) indicate areas where SOM predictions exceed OK, while cooler colors (blue) represent regions where OK predictions are higher. The color scale quantifies the differences in ppm, enabling a direct comparison.

ses, whereas SOM highlights localized anomalies and spatial clusters.

### *Methodological Comparison*

The analysis reveals the complementary nature of the two approaches:

- **Ordinary Kriging (OK):** Ideal for generating broad, continuous spatial predictions, OK smooths variability and ensures regional consistency. However, it may overlook fine-scale variations and anomalies.
- **SOM-Based Grouping:** SOM excels in detecting localized patterns, clusters, and spatial heterogeneity. It captures small-scale anomalies but may overemphasize local variability in homogenous areas.

Combining these methods leverages OK's strength in regional continuity and SOM's ability

to identify localized details, offering a more comprehensive understanding of zinc concentration patterns.

## 7.5 Scatter Plot Comparison of Predictions with Observed Zinc Concentrations

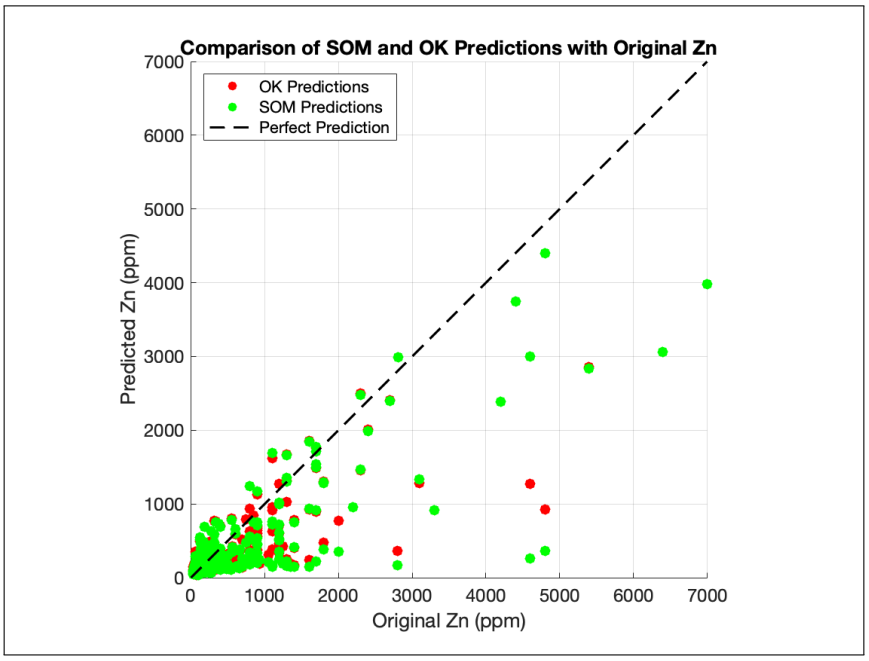


Figure 31: Scatter plot comparing original zinc concentrations (ppm) with predictions from OK and SOM-Based Grouping. Red points represent OK predictions, green points represent SOM predictions, and the black dashed line corresponds to perfect predictions ($y = x$).

Figure 31 shows the scatter plot of the OK and SOM-based grouping predictions against the observed zinc concentrations. OK predictions mostly fall along the perfect prediction line y = x, suggesting that OK performs quite consistently over a wide range of concentrations. However, there is a slight overestimation in the OK predictions for higher ranges of observed concentrations greater than 3000 ppm.

On the contrary, the SOM-based predictions have higher variability, reflecting more information in the data. SOM has a better performance at low to moderate concentrations ($<$ 3000 ppm), where its values are closer to the observed ones and with finer spatial resolution. This variability reflects the ability of the SOM model to capture localized spatial heterogeneity. However, at higher concentrations ($>$ 4000 ppm), the SOM predictions showed larger deviations from the perfect prediction line, which might be due to the effect of sparse data or the presence of some outliers.

Overall, the results emphasize the complementary strengths of the two methods. While OK is smoother and more generalized in predictions that apply at larger spatial scales, SOM captures

local variability and finer-scale data patterning. This duality suggests the reasons for combining these approaches, with SOM providing detailed predictions over areas with high data density, while OK ensures robust estimates over extensive areas.

## 7.6 Performance Evaluation of Ordinary Kriging vs. Self-Organizing Map (SOM) Based Grouping

Beyond the visual comparison of spatial distribution, it is essential to quantitatively assess the accuracy of each model's predictions.

Table 8: Performance Comparison of Ordinary Kriging and SOM-Based Grouping

| **Metric** | **Ordinary Kriging** | **Kriging (SOM Neurons)** |
|---|---|---|
| RMSE (Root Mean Square Error)(ppm) | 485.79 | 490.93 |
| MAE (Mean Absolute Error)(ppm) | 186.97 | 153.26 |
| Correlation Coefficient | 0.76 | 0.91 |
| E20 (Error at 20th Percentile) | 31.59% | 45.50% |
| E80 (Error at 80th Percentile) | 84.04% | 92.89% |

Table 8 presents the performance evaluation that shows the differences in predictive performance between OK and SOM-based Kriging. The Kriging approach using SOM proves to be the best for spatial pattern and localized variability. With its higher correlation coefficient and better performance in error evaluation metrics (E20 and E80), it is more suitable for identifying finer-scale anomalies and trends in the dataset. What is more, OK has a slightly lower RMSE, reflecting the ability to provide smooth, generalized predictions over larger spatial extents. However, its relatively low correlation coefficient and poor performances in E20 and E80 suggest that it may lose localized patterns or small-scale variability. SOM-based Kriging offers higher predictive accuracy on localized patterns but may result in higher RMSE due to its sensitivity to local variability. In contrast, OK provides smoother and more conservative estimates, turning out to be more fitted for regional-scale assessments where small-scale details are less critical.

### *7.6.1 Summary of the comparison OK-SOM-based Kriging*

The comparative analysis of OK versus the SOM-based grouping demonstrates clearly the different strengths each technique has in modeling zinc concentrations. OK is highly effective at producing smooth, consistent spatial predictions for larger areas, making it well-suited for broad-scale assessments where fine-scale variations are less critical. It is strong because it provides reliable and generalized estimates, although this may sometimes come at the cost of a deficiency in monitoring spatial heterogeneity at a more local scale. By contrast, SOM-based Grouping

excels at capturing smaller-scale anomalies and finer details that occur in data-rich regions by identifying localized trends and spatial clusters. Its sensitivity to local variability provides greater deviations in sparse data conditions. Performance metrics emphasize the same differences: OK has a smaller RMSE, while the correlation coefficient and thus goodness of fit for capturing localized variability was better for SOM-based Grouping. This complementarity indicates that a combination of regional continuity obtained from OK with the fine-scale pattern detection capability of SOM provides a fuller and more detailed understanding of the spatial variability in zinc concentration.

# 8 Conclusion

This thesis underlines the importance of integrating traditional geostatistical approaches with advanced machine learning techniques to solve complicated spatial modeling problems. Each of these approaches Ordinary Kriging, Gaussian Process Regression, and Self-Organizing Maps offers certain advantages that, when combined, constitute a robust framework for the analysis of spatial data.

Ordinary Kriging remains a reliable and interpretable tool for general spatial interpolation, providing stable predictions based on well-established geostatistical principles. However, the inability to cope with nonlinear relationships and heterogeneous spatial structures points to the necessity of complementary methods. Gaussian Process Regression addresses these limitations by introducing flexibility in modeling complex spatial dependencies and is a powerful tool for capturing nonlinear trends. While GPR requires careful tuning of parameters to manage variability, its adaptability greatly enhances spatial modeling. Self-Organizing Maps further enhance this framework by clustering data into localized groupings that improve kriging predictions in areas of high spatial variability and allow the enhanced detection of fine-scale anomalies.

Taking all approaches into consideration, the integrated application of these methods presents promising improvements that balance predictive accuracy, uncertainty quantification, and localized pattern analyses. However, the results also point out some potential paths for further improvement. For example, SOM-guided kriging promised improvements in localized enhancements, though without significant reduction of overall prediction errors. Similar concerns about the optimization of kernel functions and fine-tuning of parameters were raised regarding GPR, given its large variability in predictions.

This research provides the foundation that can be used to make further improvements within spatial data modeling, hence showing how geostatistics and machine learning can answer current challenges and open the way to new applications.

### *Future Work and Scenarios*

Future research can focus on improving the SOM-Kriging framework to address specific challenges in the field. Developing dynamic SOM structures that adapt neuron configurations based on local spatial density and heterogeneity could enhance data selection for geostatistical predictions. Expanding the method to spatiotemporal applications, such as groundwater quality monitoring and

climate change forecasting, would allow for the integration of data with both spatial and temporal dimensions. Additionally, advancing uncertainty quantification techniques could provide more reliable insights into the accuracy of predictions, particularly for geological mapping and environmental datasets. The integration of multimodal datasets, including geological, hydrological, and satellite data, can further support the prediction of resource availability and agricultural land management. Finally, the creation of user-friendly tools that combine SOM, Kriging, and GPR methodologies could make these advanced techniques more accessible to professionals in environmental science, agriculture and medicine, bridging the gap between complex modeling and practical applications.

# References


[1] Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

[2] Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press.

[3] Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

[4] Kohonen, T. (2001). *Self-Organizing Maps* (3rd ed.). Springer.

[5] Liu, Y., & Weisberg, R. H. (2011). A review of self-organizing map applications in meteorology and oceanography. In *Self-Organizing Maps: Applications and Novel Algorithm Design*, 253–272.

[6] Chiles, J.-P., & Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty* (2nd ed.). Wiley.

[7] Bação, F., Lobo, V., & Painho, M. (2005). Self-organizing maps as substitutes for K-means clustering. *Computers & Geosciences*, 31(5), 545–553.

[8] Hengl, T., Heuvelink, G. B. M., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120(1-2), 75–93.

[9] Liu, Y., & Weisberg, R. H. (2005). Patterns of ocean current variability on the West Florida Shelf using the self-organizing map. *Journal of Geophysical Research: Oceans*, 110(C6).

[10] Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3), 586–600.

[11] Howitt, D., & Cramer, D. (2008). *Introduction to Statistics in Psychology* (4th ed.). Prentice Hall. ISBN 978-0-13-205161-3.

[12] Dekking, M. (2005). *A Modern Introduction to Probability and Statistics: Understanding Why and How*. Springer. ISBN 978-1-85233-896-1.

[13] Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for Machine Learning*. Cambridge University Press. ISBN 978-1108455145.

[14] Park, K. I. (2018). *Fundamentals of Probability and Stochastic Processes with Applications to Communications*. Springer. ISBN 978-3-319-68074-3.

[15] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley & Sons, Inc., 2003, p. 104. ISBN 0-471-20454-4.

[16] D. Zwillinger and S. Kokoska, *CRC Standard Probability and Statistics Tables and Formulae*, CRC Press, 2010, p. 49. ISBN 978-1-58488-059-2.

[17] Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis* (ch. 6. Assessing distributional assumptions about data, p. 194). Duxbury Press.

[18] Box, G. E. P., & Draper, N. (2007). *Response Surfaces, Mixtures, and Ridge Analysis* (2nd ed.). Wiley. ISBN 978-0-470-05357-7.

[19] Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions* (4th ed.). Wiley. ISBN 0-471-90763-4.

[20] MathWorks. (2024). *Normal Probability Plots (normplot)*. Retrieved December 4, 2024, from `https://www.mathworks.com/help/stats/normplot.html`.

[21] Austin, D. (2006). *Voronoi Diagrams and a Day at the Beach*. Retrieved from `http://www.ams.org/publicoutreach/feature-column/fcarc-voronoi`.

[22] Bagheri-Moghaddam, F., Banihashemi, S., Bakhshoodeh, R., Fort, J., & Navarro, I. (2023). *Biomimicry Green Façade: Integrating Nature into Building Façades for Enhanced Building Envelope Efficiency*.

[23] Klee, V. (1980). *On the Complexity of d-Dimensional Voronoi Diagrams. Archiv. Math.*, 34, 75–80.

[24] Wackernagel, H. (2003). *Multivariate Geostatistics*. Springer.

[25] Chiles, J.-P., & Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty* (2nd ed.). John Wiley & Sons, New York.

[26] Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley-Interscience.

[27] Christopoulos, D. (2014). *Applied Geostatistics*. Department of Mineral Resources Engineering, Technical University of Crete.

[28] Diggle, P. J., & Ribeiro, P. J. (2007). *Model-Based Geostatistics*. Springer Series in Statistics, Springer, pp. 51–56.

[29] Eddelbuettel, D., & Sanderson, C. (2014). *RcppArmadillo: Accelerating R with High-Performance C++ Linear Algebra. Computational Statistics and Data Analysis*, 71, 1054–1063.

[30] Chilès, J.-P., & Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons Inc.

[31] Hristopoulos, D. T. (2020). *Random Fields for Spatial Data Modeling: A Primer for Scientists and Engineers*. Advances in Geographic Information Science. Springer Netherlands.

[32] Bivand, R., Pebesma, E., & Gomez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. Springer.

[33] Al-Taee, K. (2018). *Geospatial Technology for Ground Water Quality Parameters Assessment in DHI-QAR Governorate-Iraq by Using (GIS). International Journal of Civil Engineering and Technology*, 9, 358–370.

[34] Matheron, G. (1963). *Principles of Geostatistics. Economic Geology*, 58, 1246–1266.

[35] Cressie, N. (2015). *Statistics for Spatial Data*. John Wiley & Sons.

[36] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30, 79–82. doi:10.3354/cr030079.

[37] Vieira, S. R. (2000). *Geoestatística em Estudos de Variabilidade Espacial do Solo*. In Novais, R. F., Alvarez, V. H., & Schaefer, G. R. (Eds.), *Tópicos em Ciência do Solo* (Vol. 1, pp. 1–54). Viçosa: Sociedade Brasileira de Ciência do Solo.

[38] Zafeiropoulos, K., & Mylonas, N. (2017). *Statistics with SPSS: Includes Probability Theory*. Tziolas, Thessaloniki.

[39] "Evaluating Forecast Accuracy." (2016). *OTexts*. Retrieved from `https://www.otexts.org/`.

[40] Hyndman, R., & Koehler, A. (2005). Another look at measures of forecast accuracy. [Online]. Available: `https://robjhyndman.com/`.

[41] "Basic Concepts of Correlation." (2015). *Real Statistics Using Excel*. Retrieved from `https://www.real-statistics.com/`.

[42] Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes* (3rd ed.). McGraw-Hill, New York.

[43] Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

[44] Reeves, S. (2020). *Gaussian Process Modeling for Upsampling Algorithms with Applications in Computer Vision and Computational Fluid Dynamics*.

[45] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

[46] Duvenaud, D. (2014). *Automatic Model Construction with Gaussian Processes*. Ph.D. dissertation, University of Cambridge.

[47] Haykin, S. (2009). *Neural Networks and Learning Machines* (3rd ed.). Prentice Hall.

[48] R. Beale and T. Jackson, *Neural Computing: An Introduction*, Institute of Physics Publishing, 1990.

[49] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*, Perseus Books, 1991.

[50] Villmann, Th., & Bauer, H.-U. (1998). *Applications of the Growing Self-Organizing Map*.

[51] Karimi, E., Haghighi, F., Sheykhfard, A., Azmoodeh, M., & Shaaban, K. (2023). *Self-Organized Neural Network Method to Identify Crash Hotspots*. *Future Transportation*, 3(1), 286–295.

[52] Kohonen, T. (2005). *Intro to SOM. SOM Toolbox*. Retrieved June 18, 2006.

[53] Kohonen, T., & Honkela, T. (2011). Kohonen network. *Scholarpedia*, 2(1), 1568. doi:10.4249/scholarpedia.1568.

[54] Vrieze, O. J. (1995). Kohonen network. In *Artificial Neural Networks*, *Lecture Notes in Computer Science* (Vol. 931, pp. 83–100). University of Limburg, Maastricht. doi:10.1007/BFb0027024.