

Predicting Leukemia Diseases using Co-Kriging Technique

D. A. H. Alsharabi ^(*), G. M. Dhaher ⁽²⁾

^(1*,2) Department of Mathematics, College of Education for Pure Science, University of Mosul, Mosul, Iraq

Article information

Article history:

Received: February 20, 2026

Revised: May 27, 2026

Accepted: June 01, 2026

Available online: October 01, 2026

Keywords:

Co-Kriging

Leukemia Types

Cross-Variogram

Error Validity Criteria

Correspondence:

Diana Ali Hussain Alsharabi

diana.24esp16@student.uomosul.edu.iq

Abstract

This research paper addresses the issue of predicting leukemia types using the co-kriging technique. The study aims to identify factors influencing leukemia risk, determine the age-standardized overall incidence rate, and monitor any temporal increase in incidence rates among patients by applying a cross-variogram function based on primary variables of leukemia types and a secondary (auxiliary) variable, namely lymphadenopathy, which plays a role in leukemia pathogenesis. The data adopted in this research from real spatial data was collected in Mosul city, Iraq. The study sample included real data from the medical records of Mosul Nuclear Hospital during 2020, namely: age, sex, leukemia type, clinical symptoms, and lymphadenopathy (a marker of cancer). The study included the three most common types of leukemia: acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), and chronic myeloid leukemia (CML). Varying prevalence rates were recorded for the different leukemia types. By cross validation in the combined spatial prediction and Error validity criteria and Error validity criteria, small values were obtained that support the prediction process. It was found that the cancer incidence rate is higher in the cities than in the neighborhoods and villages of Mosul due to population density and numerous sources of pollution, in addition to radioactive contaminants resulting from military operations during the wars. Acute lymphoblastic leukemia (ALL) is classified as the most dangerous type of leukemia due to its unique biological characteristics, which make it highly aggressive and prone to rapid spread. Lymphadenopathy is a symptom of (ALL), in addition to genetic predisposition.

DOI: [10.33899/jes.v35i4.61650](https://doi.org/10.33899/jes.v35i4.61650), ©Authors, 2026, College of Education for Pure Science, University of Mosul.

This is an open access article under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Accurate predictions of disease occurrence are a key issue in modern medical research that focuses on studying epidemic diseases that spread very rapidly, in order to understand the causes and how to treat them. There are several types of leukemia. In our study, we address firstly acute lymphoblastic leukemia, secondly chronic myeloid leukemia, and thirdly acute myeloid leukemia. There are other types that we did not address because the data is limited to these three types. To improve prediction accuracy, these characteristics require advanced statistical models capable of exploiting spatial dependence. By considering observed values as realizations of spatial random fields, geostatistical methods provide a robust framework for analyzing spatially correlated data. Among these methods, the kriging technique has been widely and successfully applied in environmental and health-related fields due to its ability to produce spatial distribution, particularly for complex and life-threatening diseases such as leukemia. Medical and epidemiological data are often characterized by small sample sizes, spatial heterogeneity, and strong spatial correlations between environmental and clinical indicators. Unbiased predictions with minimal variance in estimation (one of the prediction conditions), in addition to data homogeneity.[1], [2], When additional variables exhibit strong spatial correlation with the phenomenon under study, reliance on a single variable is no longer sufficient. To address this limitation, the co-kriging technique was developed as an extension of classical kriging, offering the capability to

integrate one or more auxiliary variables to improve the spatial estimation of the primary variable. Previous studies have demonstrated that incorporating auxiliary information significantly enhances prediction accuracy, particularly when the primary variable data are sparsely or irregularly distributed [3]. In practical applications, however, the co-kriging technique faces several challenges, most notably the assumption of spatial stationarity, the need for accurate modeling of covariance and cross-covariance structures, and sensitivity to data scarcity. Moreover, recent research has highlighted the importance of spatial support, showing that differences in the spatial scale at which data are measured can substantially influence estimation results and should be carefully considered when applying co-kriging models [7]. Despite these challenges, advanced spatial interpolation techniques, including co-kriging, have proven effective in numerous scientific applications. Recent studies have also shown that combining geostatistical methods with modern analytical approaches can lead to notable improvements in spatial prediction accuracy, even in complex and data-limited environments. Based on these considerations, this study aims to apply the co-kriging technique to predict patterns of leukemia distribution by integrating primary disease indicators with spatially correlated auxiliary variables. The proposed approach seeks to enhance prediction accuracy while providing a deeper spatial understanding of leukemia distribution, thereby supporting early detection strategies and informed decision-making in the healthcare domain [8].

2. Methodology

2.1 Regionalized variables

The concept of the spatial random variable represents a statistical extension of the conventional random variable to describe a phenomenon or property distributed in space, such that each spatial location is associated with a value that has a probabilistic interpretation. The presence of spatial dependence can be analyzed through the covariance function and the variogram; the former describes the statistical dependence between the values of the variable at two locations separated by a given distance, whereas the variogram expresses half of the variance of the difference between these two values at that distance, thereby reflecting the degree of spatial similarity or dissimilarity. The kriging technique is based on these models, as it uses weights derived from the spatial covariance model to produce the best linear unbiased estimate of the value at unsampled locations [5], [13], [18].

2.2 Multivariate technique

The difference between the kriging technique and the multivariate technique is that the kriging technique describes a single random variable, while the multivariate technique relies on primary and secondary variables within the prediction process. This technique is known as the co-kriging technique, and it has been studied more extensively for regionally correlated variables such as health data, mineral data, and groundwater level. Multivariate geostatistics is a specialized branch of statistical analysis concerned with the study of more than one variable spatially at the same time, taking into account that these variables are not randomly distributed but are influenced by spatial distance and the mutual relationships among them. Its importance appears in cases such as the analysis of soil properties or groundwater, where it is not possible to treat each variable independently, because spatial correlation and inter-variable correlation play a decisive role in analysis and prediction. Based on this aspect, shared spatial models can be built that allow the estimation of variables at unobserved locations with greater accuracy than if each variable were analyzed separately. These models also provide the ability to determine the degree of uncertainty in predictions and to exploit the mutual relationships between variables in order to use some of them to improve the predictions of others. Thus, multivariate geostatistical analysis becomes a fundamental tool in the fields of environment, geological engineering, and water resources, where spatial variables are interconnected and complex at the same time [4].

2.3 Co-regionalized variables

As mentioned previously, spatial variables represent a set of ordinary variables, where each spatial variable has a specific location expressed by coordinates in a plane or in space (three-dimensional). If we assume that represents a random variable at location within the region, then a sample of size can be represented as follows: $Z(x_1), \dots, Z(x_n)$. This is referred to as a sample of the random function (stochastic process), where is considered a random function defined over the domain $\{Z(x), x \in D \subset R^p\}$ where $p=2$ two dimensions, $p=3$ in space. The spatial variables are separated by the spatial lag, which represents the distance between locations. The lag is calculated using the Euclidean distance [4].

$$|h| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

Joint Modeling (Linear Coregionalization) is considered one of the fundamental concepts in multivariate Geostatistics, as it provides a mathematical framework aimed at describing the spatial correlation among several variables measured within the same geographical domain, the main idea of this concept lies in constructing a single model capable of representing both the individual variances of the variables and the cross-variances between different variables, thereby ensuring the creation of a spatial structure that is valid for geostatistical analysis, the concept of coregionalization is based on the assumption that all spatial variables can be expressed as a linear combination of a limited number of independent basic fields, each of which shares a distinct spatial structure[4]-[21]. This model is known as the Linear Model of Coregionalization (LMC) and is expressed by the relationship:

$$C_{ij}(h) = \sum_k \beta_{ij}^k p_k(h) \quad (2)$$

the cross-covariance function between the two variables. C_{ij} Where:

the basic spatial function (spatial structure) of the k-th component. p_k

β_{ij}^k : the matrix coefficients that represent the degree of contribution of the component to the relationship between variables i and j [23].

This model provides a means to ensure that all resulting covariance and cross-covariance functions are positive definite, which is a fundamental requirement for the validity of any spatial model used in geostatistical estimation or prediction. Moreover, recent studies have demonstrated that the LMC model is one of the most efficient models in describing multivariate spatial relationships, especially when analyzing data from different sources such as spatial chemical measurements and spatial physical measurements in the process of co-kriging, which leads to improved estimation accuracy at unsampled locations [19].

2.4 Cross - variance

Cross-variance represents the primary basis of the statistical relationship between two spatial variables at the same location and is considered the first indicator that shows the degree of local dependence in the behavior of the two variables before examining how they change with distance. This value is regarded as a direct measure of the existence of correlation between the two variables and clarifies its nature and strength. When the two variables tend to take similar values or vary in the same direction at the same location, this indicates that the cross-variance value is positive and relatively high. In contrast, a negative sign indicates opposite behavior, where the value of one variable increases while the value of the other decreases. If the value is close to zero, this reflects the absence or weakness of a direct relationship between the two variables. The cross-variance is defined when the lag is zero as follows:

$$C_{ij}(0) = cov(Z_i(x), Z_j(x)) \quad (3)$$

Its practical importance lies in the fact that it represents the fundamental basis for constructing variance and cross-variance models, as it includes the sill value and the cross-variogram. It is also used as a basis for estimating the spatial relationship between variables at different distances. For this reason, cross-variance is considered an essential first step before proceeding to the modeling of the cross-variance or the cross-variogram within the framework of multivariate geostatistical analysis [3].

2.5 Cross – covariance

The cross-covariance function is considered one of the most important fundamental tools in multivariate geostatistics, as it measures the change in spatial correlation between two variables with increasing distance separating the measurement points, and this function plays an important role in revealing the nature of the spatial relationship between the two variables. As the distance between two points increases, the correlation often decreases, and accordingly, the cross-covariance is mathematically expressed as follows:

$$C_{ij}(h) = cov(Z_i(x), Z_j(x+h)) \quad (4)$$

where this quantity $C_{ij}(h)$ represents the covariance between variables at locations. Due to the fact that the predictive power of this function depends on the degree of spatial correlation between the primary and secondary variables, it is therefore directly used in Co-kriging applications. In general, when the spatial relationship is strong in a way that allows it to be exploited to improve the accuracy of spatial estimates, this indicates that cross-covariance values are positive and high at short distances. In contrast, when the values are negative or close to zero, this indicates weak correlation and the presence of an inverse relationship, which reduces the possibility of incorporating secondary variables in the prediction process [3], [21].

2.6 Relationship between cross- variance and cross covariance

In the context of multivariate geostatistical analysis, it is often necessary to deal with two spatial variables obtained through two different sampling methods. One of these variables, denoted by the symbol $Z(x)$, represents exploration data collected during the initial stages of the study, while the other variable, denoted by the symbol $Y(x)$, represents production data. Both data sources are jointly utilized to describe the spatial variation and internal structure of the ore body. Each of the two variables is subjected to a standardization process in order to enable direct comparability between them, whereby the mean is adjusted to be equal to zero, and the variance is standardized to take a value of one. This procedure allows for a more accurate and objective analysis of the spatial relationship between the two variables. To ensure comparability between the two variables, each is standardized such that its mean becomes zero and its variance becomes one. This process is carried out by subtracting the sample mean from each value and then dividing the result by the standard deviation. This standardization is applied within regions that have been predetermined as stationary, and the standardized variables are expressed as follows:

$$y_{std} = \frac{y(x)-m_y}{\sigma_y}, Z_{std} = \frac{Z(x)-m_z}{\sigma_z} \quad (5)$$

After completing the geostatistical estimation process, the results can be returned to their original units by multiplying the estimated value by the standard deviation of the primary variable and then adding its mean. Therefore, calculating the mean

and standard deviation accurately is important to avoid biases in the results [6]. The spatial continuity of each of the two variables is expressed through the direct data according to the formulas:

$$\gamma_{zz}(h) = E \left[(Z(x) - Z(x+h))^2 \right] \quad (6)$$

$$\gamma_{yy}(h) = E \left[(Y(x) - Y(x+h))^2 \right] \quad (7)$$

As for the spatial relationship between the two variables, it is measured using the cross-variogram and the cross-covariance, defined as follows

$$\gamma_{zy}(h) = E \{ [Z(x) - Z(x+h)][Y(x) - Y(x+h)] \} \quad (8)$$

$$C_{zy}(h) = E [Z(x)Y(x+h)] - E[Z(x)]E[Y(x+h)] \quad (9)$$

With the assumption of first- and second-order stationarity, the two functions of cross-variogram and cross-covariance can be linked by the relationship:

$$\gamma_{zy}(h) = C_{zy}(0) - C_{zy}(h) \quad (10)$$

This relationship gains particular importance when dealing with data that are not spatially coincident, since the joint variogram can be calculated directly only when measurements of both variables are available at the same locations. Therefore, the cross-covariance is often calculated first using interpolation methods, and then the cross-variogram is derived from it. This procedure was discussed in detail [14]. Additionally, calculating the joint variogram requires the presence of values for both variables at the two ends of the lag vector, whereas the cross-covariance requires having one variable at the first end and the other variable at the second end. It should also be noted that differences in the sampling methods between the two variables may lead to differences in sample size or support, for example, differences in the spacing between drilling measurements and production measurements, which must be taken into account when calculating these spatial functions [16].

2.7 Stationarity of cross -covariance function

Stationarity of cross-covariance represents one of the theoretical foundations required when constructing covariance and variogram models, whether in univariate or multivariate analysis. Second-order stationarity is considered the most widely used type of stationarity in geostatistics because it provides a simplified and suitable mathematical environment for the estimation of covariance functions, under these conditions, it is assumed that the mean of each variable remains constant across the study area, and it is also assumed that the cross-covariance depends only on the displacement between the two points and not on their absolute locations [7]. This assumption allows the estimation of the cross-covariance function using real field data and constitutes the theoretical basis for applications, as it results in the use of standard covariance models that can be fitted to experimental data. Without this condition, the process of describing the spatial relationship between variables becomes more complex, and the covariance functions may lose their mathematical properties [14], [10]. However, it should be noted that the assumption of stationarity may not always be valid in complex multivariate and spatio-temporal systems. In such cases, classical stationary cross-covariance models may fail to adequately represent the spatial dependence structure, motivating the development of nonstationary cross-covariance functions [12].

2.8 Variogram and cross -variogram

The variogram function is considered one of the fundamental tools in spatial statistics, as it is used to describe the nature of spatial variation of random variables based on the separation distance between observation points. Krige proposed the use of the variogram instead of traditional correlation coefficients (ρ)

$$\rho = \frac{\text{cov}[Z(x), Z(x+h)]}{\sigma_{Z(x)}(\sigma_{Z(x+h)})} \quad (11)$$

Where ρ is the spatial correlation coefficient, $Z(x)$ and $Z(x+h)$ are the variable values at locations x and $x+h$, h is the lag distance, denotes covariance, and σ represents the standard deviation.

Due to the potential for large variance values in spatial data to lead to inaccurate interpretations when relying solely on correlation measures. The variogram function enables a more precise representation of the spatial structure of the data, particularly in cases where spatial variance is non-uniform.

The semi-variogram $\gamma(h)$ of a spatial random variable $Z(x)$, under the assumption of second-order stationarity, is defined as:

$$\gamma(h) = \frac{1}{2} E [Z(x+h) - Z(x)]^2 \quad (12)$$

where h represents the displacement vector (distance or spatial lag) between two locations. In practice, this function is estimated by forming pairs of points separated by a specified distance, then calculating the arithmetic mean of the squared differences of the values within each distance class, which is known as the experimental variogram, which is represented mathematically by the known equation:

$$2\gamma(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} [Z(s_i) - Z(s_i+h)]^2 \quad (13)$$

The semi-variogram $\gamma(h)$ is directly related to the covariance function $C(h)$ and this relationship can be expressed as:

$$\gamma(h) = C(0) - C(h) \quad (14)$$

where $C(0)$ represents the total variance of the variable, while $C(h)$ denotes the covariance between values separated by a distance h . This relationship represents the mathematical foundation that integrates variance and covariance structure, and it is considered the basis upon which various kriging models are built.

The variogram function is characterized by three main parameters, which can be explained as follows. The first is the range, denoted by (a) , which represents the distance on the horizontal axis or the displacement value at which the variogram values cease to increase, indicating the end of spatial correlation. The second parameter is the sill, which represents the total variance of the variable and equals the sum of $(C + C_0)$, The third parameter is the nugget effect, denoted by (C_0) , which reflects the amount of random errors associated with measurement processes or very fine spatial variations, and is interpreted mathematically as the discontinuity of the variogram function at displacement $(h = 0)$ (i.e., at the origin); hence it is also known as the discontinuity.

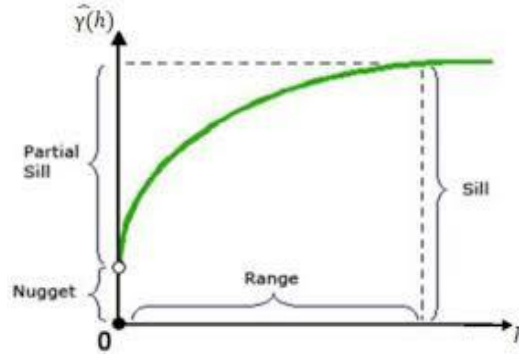


Figure (1). Parameters of the variogram function

Figure (1) illustrates the characteristics of the variogram function, where denotes the different values specific to the variogram function (range/ sill/ nugget effect). When the semi-variogram is unaffected by the direction of displacement, the phenomenon is described as isotropic, and this condition is expressed by the following relationship:

$$\gamma(h) = \gamma(|h|) \quad (15)$$

In this case, the spatial structure is symmetric in all directions, which facilitates variogram modeling and enhances the accuracy of spatial interpretation. Conversely, if the shape of the variogram differs with direction, this indicates the presence of anisotropy, reflecting non-uniform spatial behavior of the studied phenomenon. Applying the variogram or semi-variogram function is considered an essential step for determining weights, due to the difficulty of choosing appropriate covariance models for the observations or data in geostatistics:

$$\lim_{h \rightarrow 0^+} \gamma(h) = C_0 \neq 0 \quad (16)$$

As the displacement variable h approaches zero from the positive side, the semi-variogram function does not equal zero but takes a finite value denoted by C_0 . This value represents the degree of discontinuity of the phenomenon or the function at $h = 0$, and this phenomenon is known in geostatistics as the nugget effect. It represents the random errors associated with different measurement units when converting between distance scales, such as from centimeters to meters or from meters to kilometers. If $\lim_{h \rightarrow 0} \gamma(h) = 0$ this indicates the absence of the nugget effect, i.e., that the phenomenon or function is completely continuous at $(h = 0)$. The cross-semi-variogram function $2\gamma(h)$ is calculated based on a set of observations taken in more than one direction, whether along the same measurement direction or along perpendicular directions. This function shows variation in values as a result of changing direction, reflecting the effect of anisotropy in spatial variance. To describe the spatial relationship between two different variables, the cross-variogram function is employed, representing the multivariate extension of the variogram concept. The ordinary variogram for a single variable is defined as follows:

$$\gamma_1(h) = \frac{1}{2} E(Z_1(x_i) - Z_1(x_i + h))^2 \quad (17)$$

$$\gamma_2(h) = \frac{1}{2} E(Z_2(x_i) - Z_2(x_i + h))^2 \quad (18)$$

and the cross-variogram is defined as follows:

$$\gamma_{12}(h) = \frac{1}{2} E(Z_1(x_i) - Z_1(x_i + h))(Z_2(x_i) - Z_2(x_i + h)) \quad (19)$$

The cross-variogram between two variables $Z_{i(x)}$ be a random function (random field) of a primary variable and $Z_j(x)$ a secondary variable is defined as:

$$\gamma_{ij}(h) = \frac{1}{2} E[(Z_i(x+h) - Z_i(x))(Z_j(x+h) - Z_j(x))] \quad (20)$$

Where $\gamma_{ij}(h)$ it is known as the auto-variogram (variogram of a single variable) when $(i = j)$ and called the cross-variogram when $(i \neq j)$. Here h is the separation distance (lag) between the two locations. The relationship between the variogram and the covariance function can be written as shown:

$$\gamma_{ij}(h) = C_{ij}(0) - C_{ij}(h) \quad (21)$$

Modeling the cross-variogram is a fundamental step in applying the Co-kriging method, as both the individual variogram models and the cross-variogram models must be mathematically consistent to ensure the positive definiteness of the joint covariance matrix, which is a necessary condition for successful multivariate spatial estimation. Accordingly, the variogram and cross-variogram functions are among the most important statistical tools for analyzing spatial structure, due to their essential role in understanding joint spatial variability [9].

3. Kriging Theory

The initial seeds of the kriging idea crystallized in the mid-twentieth century when the South African engineer Danie Krige sought to develop a practical method that would allow the estimation of gold concentrations in areas where only limited data were available. With the growing interest in spatial analysis, researchers began to view Krige's work as a foundational step toward a methodology capable of providing more accurate estimates in unstructured environments. Over time, the researcher Georges Matheron reformulated this idea within a comprehensive mathematical framework, transforming it from a local applied attempt into a clearly defined scientific method. He later gave it the formal name and established the theoretical foundations that kriging is one of the central pillars of geostatistics.

3.1 Types of kriging

Kriging methods are classified according to four main pillars: the mean, the number of variables (mono/multi), the objective estimation (point/block in isotropic), and the structural form (isotropic/anisotropic).

- Simple kriging assumes that the mean is known and constant
- Ordinary kriging assumes that the mean is locally constant but unknown
- Universal kriging is used when the mean varies according to a specific spatial trend, and the trend model is incorporated into the equations.
- Indicator kriging transforms the values into binary indicators to estimate probabilities.
- Co-kriging combines a primary variable with one or more auxiliary variables that are spatially correlated, to improve the estimation. [6], [15], [16]

3.2 Co-kriging Technique

3.2.1 Theoretical Background:

Co-kriging is considered one of the advanced geostatistical methods that aims to improve the accuracy of spatial prediction by integrating information from more than one variable. In contrast to traditional kriging, which relies on only a single variable in the estimation process, co-kriging benefits from spatial correlations and cross-correlations among multiple variables to increase the reliability of the results. The importance of this method stems from its ability to utilize auxiliary variables that are spatially correlated with the primary variable, which allows the production of more accurate spatial maps at lower cost, especially in cases where direct measurements of the variable of interest are limited or expensive to obtain for example, measurements of a certain subsurface mineral, which may be scarce or difficult to acquire, can be combined with data from another low-cost variable that has a strong spatial correlation with it, such as electrical soil resistivity, to improve the accuracy of predicting the distribution of that mineral, theoretically co-kriging is expected to provide better performance than univariate kriging when there is a clear and influential correlation between the input variables. However, in the absence of such correlation or when it is weak, the benefit becomes limited, and traditional kriging may be a more scientifically appropriate option due to its simplicity. Co-kriging also inherently requires more complex models to represent covariance and cross-covariance relationships among variables, which increases the demands of modeling and statistical rigor. Nevertheless, this method has proven effective in many geological, hydrological, and environmental applications and has become increasingly adopted in multivariate spatial data analysis, owing to its ability to exploit the available informational richness and transform it into high-quality spatial estimation [5], [4].

3.2.2 Mathematical Formulation of Co-kriging:

The co-kriging estimator $\hat{Z}(x_0)$ for the primary variable $Z(x)$ at an unsampled location x_0 is defined as a linear combination of observations of the primary and secondary variables as follows:

$$\hat{Z}(x_0) = \sum_{i=1}^n \lambda_i^0 Z(x_i) + \sum_{j=1}^k \delta_j^0 Y(x_j) \quad (22)$$

Where $(i = 1 \dots, n), (j = 1 \dots, k)$

Unbiasedness Conditions

Assuming an unknown but spatially constant mean, the unbiasedness conditions are given by

$$\sum_{i=1}^n \lambda_i^0 = 1 \quad (23)$$

$$\sum_{j=1}^k \delta_j^0 = 0 \quad (24)$$

or equivalently, using a single unbiasedness constraint:

$$\sum_{i=1}^n \lambda_i^0 + \sum_{j=1}^k \delta_j^0 = 1 \quad (25)$$

Variance Minimization : The optimal weights are obtained by minimizing the estimation error variance subject to the unbiasedness constraints. The estimation error variance is expressed as:

$$\begin{aligned} \text{Var}\{Z^*(x_0) - Z(x_0)\} = & \sum_{i=1}^n \sum_{j=1}^n \lambda_i^0 \lambda_j^0 C_Z(h_{ij}) + \sum_{i=1}^k \sum_{j=1}^k \delta_i^0 \delta_j^0 C_Y(h_{ij}) + \\ & \sum_{i=1}^n \sum_{j=1}^k \lambda_i^0 \delta_j^0 C_{ZY}(h_{ij}) + \sum_{j=1}^k \sum_{l=1}^n \delta_j^0 \lambda_l^0 C_{YZ}(h_{l0}) - 2 \sum_{i=1}^n \lambda_i^0 C_Z(h_{i0}) \\ & - 2 \sum_{j=1}^k \delta_j^0 C_Y(h_{j0}) + C_Z(h_{00}) \end{aligned} \quad (26)$$

Covariance and Cross-Covariance Definitions: $C_Z(h_{ij})$: denotes the covariance function of the primary variable. $C_Y(h_{ij})$ denotes the covariance function of the auxiliary variable.

The cross-covariance function can be expressed by:

$$C_{ZY}(h_{ij}) = E\{Z(u_i)Y(u_j)\} - m_Z m_Y \quad (27)$$

Matrix Formulation and Solution

The co-kriging system can be written in compact matrix form

$$C \cdot \lambda = W \quad (28)$$

Where:

C : the covariance and cross-covariance matrix: the vector of unknown weights and Lagrange multipliers λ

W : the covariance vector between the observations and the estimation location u_0

Solution of the System:

The co-kriging system is defined by:

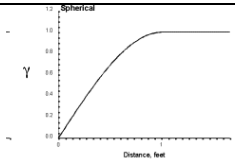
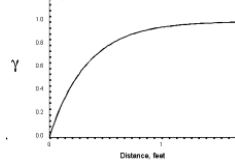
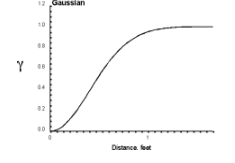
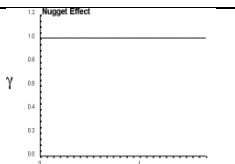
$$\lambda = C^{-1} W \quad (29)$$

In the co-kriging method, the resulting weights are directly incorporated into the defined estimator while the estimation variance provides a quantitative measure of accuracy and prediction uncertainty [17].

3.5 Model of variogram and covariance function

The variogram and covariance functions are considered fundamental tools in geostatistics, as they are used to describe the degree of spatial correlation between data-available points [4]. Since this correlation reflects the model's ability to exploit the available information from a spatially distributed variable across different distances, it is therefore considered very important in prediction processes, which makes the model suitable for spatial prediction. These models require the fulfillment of a set of conditions, the most important of which is that the covariance function must be positive definite, in order to ensure that the values produced by the model are consistent and not contradictory to the nature of the data. The fulfillment of these conditions is considered necessary to guarantee the stability of the spatial model and to provide correct estimates at different distances. The basic descriptive factors on which these models depend, such as the range, the sill, and the way covariance changes with increasing distance, help determine the appropriate model form and its degree of compatibility with the spatial domain [20].

Table 1. Model of variogram functions

N o.	Model	Equation	The drawing
1	Spherical	$y_{a,c+c_0}^{sph} = \begin{cases} (c + c_0) \left(\frac{3 h }{2a} - \frac{1}{2} \left(\frac{ h }{a} \right)^3 \right) & \text{if } 0 \leq h \leq a \\ (c + c_0) & h > 0 \end{cases}$	
2	Exponential	$y_{a,c+c_0}^{exp} = (c + c_0) \left(1 - \exp \left(\frac{- h }{a} \right) \right), \quad h \neq 0$	
3	Gaussian	$y_{a,c+c_0}^{gau} = (c + c_0) \left(1 - \exp \left(\frac{- h ^2}{a^2} \right) \right) \text{ for } h \geq 0$	
4	Nugget effect	$y_{a,c+c_0}^{nug} = \begin{cases} 0 & , \text{if } h = 0 \\ (c + c_0) & , \text{other wise} \end{cases}$	

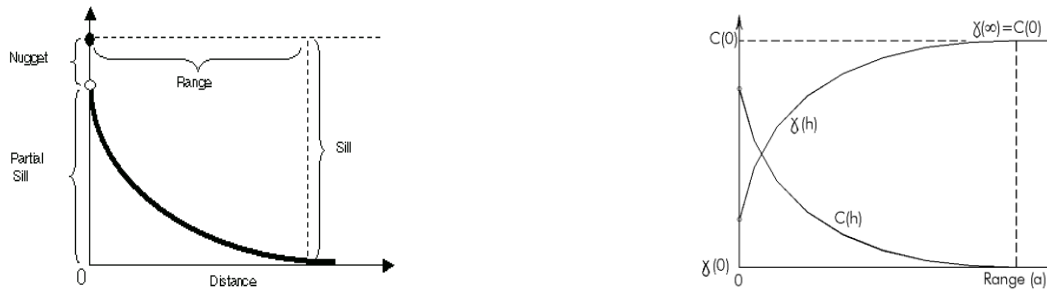


Figure (2). left is the covariance, and right is the variogram and covariance

Figure (2), the left curve illustrates the covariance function, where the curve is decreasing. The nugget effect represents the highest point on the y-axis, and the curve decreases until it approaches zero in the case of infinity(∞). Meanwhile, Figure (2) right curve illustrates both the variogram and covariance functions on the same axes, as shown, the variogram curve increases from the nugget effect until it reaches the $\gamma(\infty) = C(0)$, which represents the variance, whereas the covariance curve decreases from the nugget effect until it approaches zero in the case of infinity(∞). In both cases (variogram and covariance), the range determines whether the curve exhibits continuity or not [22].

3.6 Performance of the model

A set of statistical error-validation metrics was adopted in order to estimate the efficiency of kriging models and the accuracy of the resulting spatial predictions. These metrics provide a comprehensive understanding of model performance in terms of bias, variability of the estimates, and error magnitude when compared with the observed values.

Mean square error (MSE):

Mean Squared Error is used as a measure of prediction accuracy, as it is obtained by calculating the arithmetic mean of the squared differences between the estimated values and the observed values. It is expressed by the following equation:

$$MSE = \frac{1}{N} \sum_{i=1}^N [\hat{z}(x_i) - z(x_i)]^2 \quad (30)$$

Due to the presence of the squaring operation, this metric is characterized by high sensitivity to errors, which makes it more suitable for evaluating models in which large deviations are undesirable.

Root mean square error (RMSE):

The RMSE metric is used to measure the overall dispersion of prediction errors around the observed values, and it is calculated using the following equation:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{z}(x_i) - z(x_i))^2} \quad (31)$$

RMSE is considered a widely used metric in spatial modeling, as it combines the sensitivity of MSE to large errors with ease of interpretation, due to the consistency of its units with those of the original data.

While KRMSE kriging of Root Mean Squared Error is used to quantify the overall dispersion of prediction errors around the observed values, and is calculated as:

$$KRMSE = \frac{1}{N} \sum_{i=1}^n \left\{ \frac{(\hat{z}(x_i) - z(x_i))^2}{\sigma_{c0,i}^2} \right\} \quad (32)$$

Coefficient of Determination (R^2):

The coefficient of determination represents the proportion of variance in the observed values that can be explained by the prediction model. It is expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^N (z(x_i) - \hat{z}(x_i))^2}{\sum_{i=1}^N (z(x_i) - \bar{z}(x_i))^2} \quad (33)$$

Values of R^2 close to unity indicate a strong ability of the model to reproduce the observed data, whereas lower values suggest poor model performance [11].

4. Results and Data Analysis

4.1 Study data

The data adopted was collected from Mosul city in Iraq. The study sample included real data from the medical records of Mosul Nuclear Hospital during 2020 (from January to December), namely: age, sex, leukemia type, clinical symptoms, and lymphadenopathy (a marker of cancer). The sample that was taken of this research contains (192) cases of leukemia types divided into three types, each type has (64).

Table 2. Types of leukemia

	Types of leukemia	No.	%	Mean age (months)	Males	Females
1.	Acute lymphoblastic leukemia (ALL)	99	43.4	66	63	36
2.	Acute myelogenous leukemia (AML)	67	27.8	138	36	31
3.	Chronic myeloid leukemia (CML)	64	27.8	119	40	24

In Table 2, the distribution of leukemia types. We selected (64) cases of each leukemia type as a sample study of this research, was recorded in mean age (months), with the gender (males and females).

Table 3. Signs and Symptoms of Patients

Signs and Symptoms	No. of Patients
Fever	61
Paller	45
Bleeding tendency	33
Bone Pain	35
Lymphadenopathy	30
Fatigue - weight loss	18

Table 3 illustrates the symptoms and signs of patients (Fever, Pallor, Bleeding, Bone Pain, Lymphadenopathy, and Fatigue - weight loss), also, there are many common signs between patients.

Table 4. Statistics of leukemia

	Leukemia	N	Mean	Median	Min-max	Standard deviation
1.	Acute lymphoblastic leukemia (ALL)	64	18.64	18.8	13.9 - 24.7	3.33
2.	Acute myelogenous leukemia (AML)	64	21.3	24.35	20 - 25.7	1.03
3.	Chronic myeloid leukemia (CML)	64	18	18.75	12.3 - 22.9	2.87
4.	Lymphadenopathy	64	14.8	43.94	14.8 – 59.1	14.08

Table 4 shows the statistics of types of leukemia and Lymphadenopathy (mean, median, min-max, and standard deviation) with the select (64) as sample of each type.

Table 5. distribution of leukemia types according to Age group

Age group		ALL	AML	CML	Total
One year	Count	2	0	0	2
	% within Age group	100.0%	0.0%	0.0%	100.0%
Two years	Count	11	2	0	13
	% within Age group	84.6%	15.4%	0.0%	100.0%
Three years	Count	13	5	0	18
	% within Age group	72.2%	27.8%	0.0%	100.0%
Four years	Count	8	5	0	13
	% within Age group	61.5%	38.5%	0.0%	100.0%
Five years	Count	11	3	0	14
	% within Age group	72.2%	27.8%	0.0%	100.0%
Total	Count	45	15	0	60

Table 5 shows the age distribution by type of leukemia. The highest incidence of acute lymphoblastic leukemia (ALL), followed by acute myeloid leukemia (AML), was recorded in children aged two to five years. No cases of chronic myeloid leukemia (CML) were recorded during this period (two to five).

Table 6. distribution of type of leukemia according to gender.

Gender		ALL	AML	CML
Male	Count	26	10	0
	% within Age group	72.2%	27.8%	0.0%
Female	Count	19	5	0
	% within Age group	79.2%	20.8%	0.0%
Total	Count	45	15	0

Table 6 illustrates the distribution by sex and subtypes of leukemia. The current study showed that males have a higher prevalence of all leukemia subtypes.

Table 7. distribution of type of leukemia according to residence

Residence		ALL	AML	Total
Rural	Count	26	8	34
	% within Age group	76.5%	23.5%	100 %
Urban	Count	19	7	26
	% within Age group	73.1%	26.9%	100%
Total	Count	45	15	60

Table 7 illustrates the distribution of types of leukemia according to residence. These results show the Rural is more than urban, with Acute lymphoblastic leukemia (ALL), followed by Acute myelogenous leukemia (AML), and nothing of chronic myeloid leukemia.

Table 8. Causes of death of patients

	Disease	Bleeding	Sepsis	Tumor lysis Syndrome	Total
1.	Acute lymphoblastic leukemia (ALL)	11 (39%)	16 (57%)	1 (14%)	28
2.	Acute myelogenous leukemia (AML)	2(18.18%)	7(63.6%)	2(18.18%)	11
3.	Chronic myeloid leukemia (CML)	2 (40%)	3 (60%)	0	5

Table 8 shows the causes of death for patients in the cases (bleeding, sepsis, and tumor lysis syndrome). With notes, the Acute lymphoblastic leukemia (ALL) has a high total and follows Acute myelogenous leukemia (AML), compared with Chronic myeloid leukemia (CML).

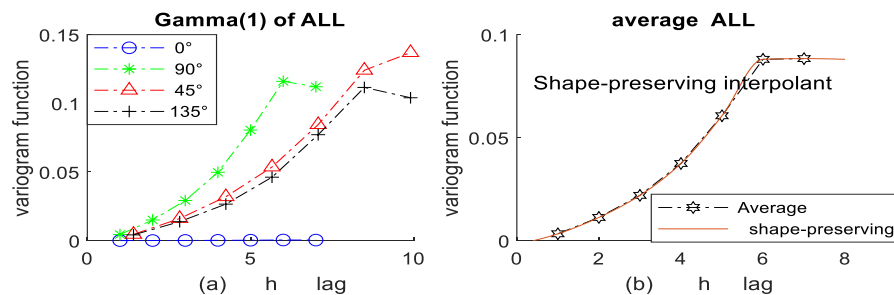
**Figure 3.** Curves of Acute Lymphoblastic Leukemia (ALL)

Figure 3 illustrates one of the subtypes of leukemia in Mosul, Acute lymphoblastic leukemia (ALL), showing the curves of the variogram function in Figure (a) of the standard directions of compass (E-W, N-S, NE-SW, NW-SE) corresponding to the angles (0°, 90°, 45°, 135°), respectively. In Figure (b), show the average of the variogram function of all directions with shape interpolant and data statistics (Nugget effect (c_0) = 0.003328, sill($c_0 + c$) = 0.08835, and range (a) = 6).

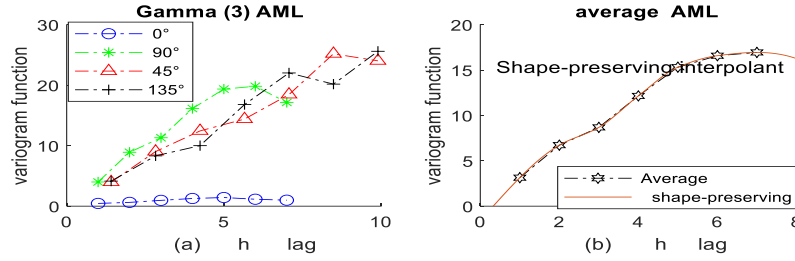


Figure 4. Curves of Acute Myelogenous Leukemia (AML)

Figure 4 shows the subtypes of leukemia, Acute myelogenous leukemia (AML). In the same way shown the curves of the variogram function are shown in Figure (a) of the standard directions corresponding to the angles (0°, 90°, 45°, 135°) respectively. In Figure (b), show the average of the variogram function, with data statistics (Nugget (c_0) = 0.456, max ($c_0 + c$) = 1.432, range (a) = 6).

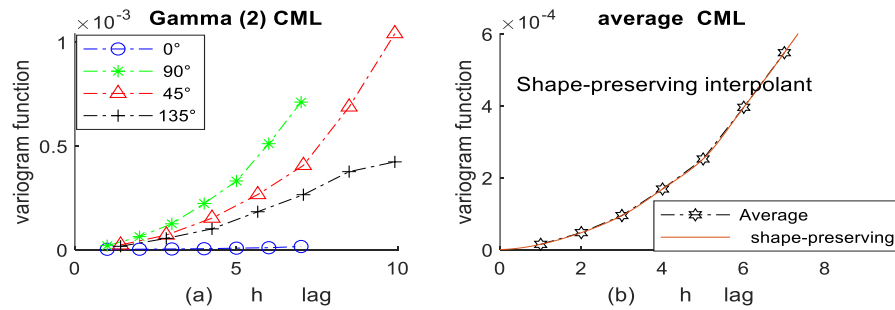


Figure 5. Curves of Chronic Myeloid Leukemia (CML)

Figure 5 shows one of the subtypes of leukemia, Chronic myeloid leukemia (CML), in the same way as the curves of the variogram function in Figure (a) of the standard directions of compass (E-W, N-S, NE-SW, NW-SE) corresponding to the angles (0°, 90°, 45°, 135°), respectively. In Figure (b), show the average of the variogram function of all directions, with data statistics (Nugget effect (c_0) = 1.454e-05, max ($c_0 + c$) = 0.000547, and range (a) = 6).

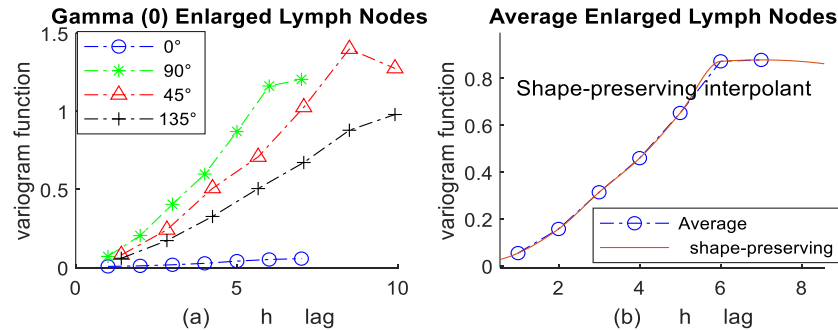


Figure 6. Curves of Enlarged Lymph Nodes

Figure 6 shows the curves of enlarged lymph nodes, where Enlarged Lymph nodes are considered the most common cause of leukemia. leukemia, shown the curves of the variogram function in Figure (6) (a) of the standard directions. In Figure (6) (b), show the average of the variogram function blue curve, and the shape of the interpolant with data statistics (Nugget effect (c_0) = 0.05454, max ($c_0 + c$) = 0.8789, and range (a) = 6).

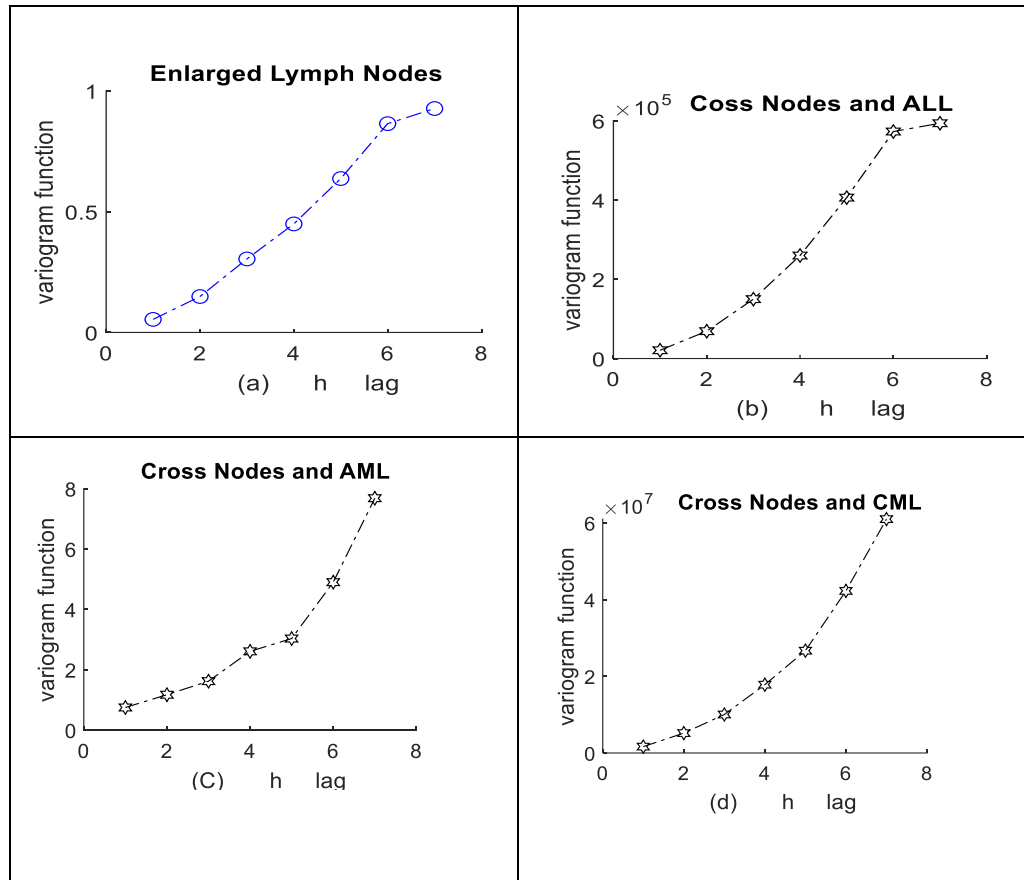


Figure 7. Curves of the cross-variogram function

Figure 7 shows (a) the curves of enlarged lymph nodes, where Enlarged Lymph nodes cause leukemia, showing the average of the variogram function in Figure (7) (b) illustrates the cross variogram between the Nodes and ALL. In Figure (7) (c), show the cross-average of the variogram function Between Nodes and AML. Figure (7) (d) shows the curve of cross Nodes and CML. These curves show that the behaviors of cross with types of leukemia in Figures (a,b) have the same curves, while in Figures (c,d) are the same curves. The behavior of curves (a) and (b) is similar and resembles the spherical model, while the behavior of curves (c) and (d) is that of the exponential model curve. Figure a: nugget = 0.03328, range = 6, sill = 0.8835. Figure b: nugget = 5.607×10^5 , range = 6, sill = 6×10^6 . Figure c: nugget = 0.05637, range = 7, no sill. Figure d: nugget = 4.503×10^6 , range = 7, no sill.

Table 9. Cross-validation of leukemia cancer data estimator

$\hat{z}(x_0)$	MSE	RMSE	KRMSE	R ²
14.7	0.225	0.3479	0.230	0.887
8.9	0.375	0.5244	0.332	0.954
20.5	0.344	0.5486	0.402	0.899
23.8	0.322	0.3302	0.367	0.999

Table 9 shows four true values for predicting cancer types data with cross-validation of the estimator. The table contains the estimator of, and evaluates it using the mean squared error (MSE) We note the small results in the scales (MSE, RMSE, R², KRMSE). The results in this Table show the smallest values, and this indicates the accuracy of the prediction and the validity of the validation

5. Conclusions

This research addressed the issue of predicting the incidence of various types of leukemia using the cokriging technique to identify and study the factors influencing the incidence of different leukemias and their underlying causes. Leukemia does not form a single mass or tumor; therefore, the stage of the disease does not depend on the size or extent of the tumor. Instead, leukemia classification is based on the number of abnormal cells in the blood or bone marrow. The spread of cancer cells (including platelets and red blood cells) lymphocytosis with enlarged lymph nodes are characteristic and the most common type, acute lymphoblastic leukemia (ALL), and Chronic myeloid leukemia (CML) is divided into the chronic stage (few symptoms, usually responding well to targeted therapy), the accelerated stage (leukemia cells begin to grow rapidly, and new chromosomal changes may appear more quickly), and the blast stage (a highly aggressive stage similar to acute leukemia, characterized by a large number of immature cells). This information on primary variables and secondary variables (enlarged lymph nodes) helps classify patients into low-, intermediate-, or high-risk groups, guiding decisions regarding chemotherapy intensity, bone marrow transplant eligibility, and clinical trial options. Four potential predictive indicators—including comorbidity, disease biology, treatment logistics, and by using error-correction criteria. The predictive values were found to be very close to the true values for various leukemia subtypes. The Nineveh Health Department reported a significant increase in cancer rates, particularly in the years following the wars in Mosul, especially after 2003, amid warnings of the potential for genetic mutations in newborns, which could lead to an increase in birth defects.

6. Acknowledgements

At the end of our work, we can only thank the University of Mosul / College of Education for Pure Sciences / Department of Mathematics for providing moral support to complete this research.

7. Declarations

7.1 Ethics approval and consent to participate

Not applicable.

7.2 Consent for publication

Not applicable.

7.3 Availability of Data and Materials

Data will be provided upon receiving a valid request.

7.4 Conflicts of interest

The authors declare that there is no conflict of interest.

7.5 Funding

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. References

- [1] A. G. Journel and C. J. Huijbregts, Mining Geostatistics. London, U.K. and New York, Academic Press, 1978.
- [2] N. A. C. Cressie, Statistics for Spatial Data. New York, NY, USA: Wiley, 1993.
- [3] G. Matheron, "Principles of geostatistics," Economic Geology, vol. 58, no. 8, pp. 1246–1266, 1963, doi: 10.2113/gsecongeo.58.8.1246. [Online]. Available: <https://doi.org/10.2113/gsecongeo.58.8.1246>
- [4] H. Wackernagel, Multivariate Geostatistics: An Introduction with Applications. Berlin–Heidelberg, Germany: Springer, 2003.
- [5] J.-P. Chilès and P. Delfiner, Geostatistics: Modeling Spatial Uncertainty. Hoboken, NJ, USA: John Wiley & Sons, 2012.
- [6] P. A. Dowd and E. Pardo-Igúzquiza, "The many forms of co-kriging: A diversity of multivariate spatial estimators," Mathematical Geosciences, 2023. [Online]. Available: <https://link.springer.com/>
- [7] M. A. Oliver and R. Webster, Geostatistics for Environmental Scientists. Chichester, U.K.: John Wiley & Sons, 2015.
- [8] H. Rakotonirina, P. Hénine, O. Atteia, and A. Van Exem, "A generative deep neural network as an alternative to co-kriging," Applied Computing and Geosciences, 2024, doi: 10.1016/j.acags.2024.100198. [Online]. Available: <https://doi.org/10.1016/j.acags.2024.100198>
- [9] N. Cressie and C. K. Wikle, Statistics for Spatio-Temporal Data. Hoboken, NJ, USA: John Wiley & Sons, 2011.
- [10] M. L. Stein, Handbook of Spatial Statistics. Boca Raton, FL, USA: CRC Press / Chapman & Hall, 2020.
- [11] T. Gneiting et al., "Covariance functions for space–time data," Journal of the American Statistical Association, vol. 105, no. 491, pp. 1022–1040, 2010. [Online]. Available: <https://doi.org/10.1198/jasa.2010.tm09420>
- [12] M. L. O. Salvaña and M. G. Genton, "Nonstationary cross-covariance functions for multivariate spatio-temporal random fields," Spatial Statistics, vol. 37, Art. no. 100411, 2020. [Online]. Available: <https://doi.org/10.1016/j.spasta.2020.100411>
- [13] P. Goovaerts, Geostatistics for Natural Resources Evaluation. Oxford, U.K.: Oxford University Press, 1997.
- [14] C. V. Deutsch, Geostatistics Course Notes. Edmonton, AB, Canada: Geostatistics Center, University of Alberta, 2015.

- [15] C. V. Deutsch, "The decision of stationarity," Geostatistics Lessons, Jun. 8, 2022. [Online]. Available: <https://geostatisticslessons.com/the-decision-of-stationarity/>
- [16] L. Dávila and C. V. Deutsch, "Co-kriging with inequality-sampled data," Dept. of Civil and Environmental Engineering, University of Alberta, Edmonton, AB, Canada, 2022.
- [17] X. Ma, Multivariate Geostatistical Modeling. Edmonton, AB, Canada: University of Alberta, 2009.
- [18] T. Hengl and R. A. MacMillan, Predictive Soil Mapping with R. Wageningen, Netherlands: OpenGeoHub Foundation, 2019. [Online]. Available: <https://soilmapper.org/>
- [19] T. S. Gonçalves, H. Klammler, L. R. B. Leal, and L. de Queiroz Salles, "Multivariate geostatistics for mapping of transmissivity and uncertainty in karst aquifers," Water, vol. 16, no. 17, Art. no. 2430, 2024. [Online]. Available: <https://www.mdpi.com/2073-4441/16/17/2430>
- [20] M. L. Krock, W. Kleiber, D. Hammerling, and S. Becker, "Modeling massive highly multivariate nonstationary spatial data with the basis graphical lasso," Journal of Computational and Graphical Statistics, pp. 1–16, 2023, doi: 10.1080/10618600.2023.2174126. [Online]. Available: <https://doi.org/10.1080/10618600.2023.2174126>
- [21] K. Alie, D. A. Stephens, and A. M. Schmidt, "Computational considerations for the linear model of coregionalization," arXiv preprint, 2024. [Online]. Available: <https://arxiv.org/>
- [22] J. M. Ver Hoef and N. Cressie, "Multivariable spatial prediction," Mathematical Geology, vol. 25, no. 2, pp. 219–240, 1993, doi: 10.1007/BF00893273. [Online]. Available: <https://doi.org/10.1007/BF00893273>
- [23] J. M. Mohammed, "Multivariate spatial statistical data point estimation with application," Journal of Administrative and Economic Sciences, Univ. of Kirkuk, vol. 21, no. 85, 2015.

التنبؤ بأمراض اللوكيميا باستخدام تقنية الكريكنك المشترك (Co-Kriging)

ديانا علي حسين الشرايبي⁽¹⁾، غانم محمود ظاهر⁽²⁾

^(1,2) قسم الرياضيات، كلية التربية للعلوم الصرفة، جامعة الموصل، الموصل، العراق

المستخلص:

يتناول هذا البحث مشكلة التنبؤ بأنواع سرطان الدم (اللوكيميا) باستخدام تقنية الكريكنك المشترك (Co-Kriging). وتهدف الدراسة الى تحديد العوامل المؤثرة في خطر الإصابة بسرطان الدم، وحساب معدل الإصابة الكلي المعياري حسب العمر ورصد أي زيادة زمنية في معدلات الإصابة بين المرضى وذلك من خلال تطبيق دالة التغيرات المتقاطع (Cross-Variogram) المعتمدة على المتغيرات الرئيسية لأنواع سرطان الدم ومتغير ثانوي (مساعد)، وهو تضخم العقد اللمفاوية، الذي يلعب دوراً في امراض سرطان الدم. أجريت دراسة مقطعية في مدينة الموصل _ العراق وشملت عينة الدراسة بيانات حقيقية مأخوذة من السجلات الطبية لمستشفى الموصل للأورام خلال عام 2020 وتمثلت هذه البيانات في: العمر، الجنس، نوع سرطان الدم، الاعراض السريرية وتضخم العقد اللمفاوية (بوصفه مؤشراً على الإصابة بسرطان). شملت الدراسة الأنواع الثلاثة الأكثر شيوعاً من سرطان الدم وهي: ابيضاض الدم اللمفاوي الحاد (ALL)، وابيضاض الدم النخاعي الحاد (AML)، وابيضاض الدم النخاعي المزمن (CML)، حيث سجلت معدلات انتشار متفاوتة بين هذه الأنواع. ومن خلال تطبيق تحليل تبين الخطأ في التنبؤ المكاني المشترك ومعايير التحقق من صحة الخطأ، تم الحصول على قيم صغيرة تدعم كفاءة عملية التنبؤ.

كما أظهرت النتائج ان معدل الإصابة بسرطان اعلى في المدن مقارنة بالأحياء والقرى التابعة لمدينة الموصل، ويعزى ذلك الى الكثافة السكانية المرتفعة وتعدد مصادر التلوث، إضافة الى الملوثات الاشعاعية الناتجة عن العمليات العسكرية خلال الحروب. وقد تبين ان ابيضاض الدم اللمفاوي الحاد (ALL) يصنف بوصفه احطر أنواع سرطان الدم، نظراً لخصائصه البيولوجية الفريدة التي تجعله شديدة العدوانية وسريع الانتشار. ويعد تضخم العقد اللمفاوية السبب الرئيس في الإصابة ب (ALL)، إضافة الى الاستعداد الوراثي.