ASSIGNMENT 3

Study 1: Simple Kriging Prediction for Cobalt Values

1 INTRODUCTION

This study aims to analyze the pattern of ***spatial dependencies*** among cobalt values measured in parts per million (ppm), based on sample data from Vancouver Island in British Columbia collected by the Geological Survey of Canada (geochemical prospecting survey). The data set contains $n = 286$ sampling locations ($s_i : i = 1,\ldots,n$), $s \in R$ and corresponding cobalt measurements $y_i = y(s_i)$, extending over the area at the northern tip of the island, as outlined in red in Figure 1 below. These locations are better shown in the enlarged Figure 2 below, bounded by a ***minimum enclosing rectangle*** around all selected sites, where the rectangle boundary outlined in red is the region of interest for the following analysis.

Figure 1. Vancouver Sample Area

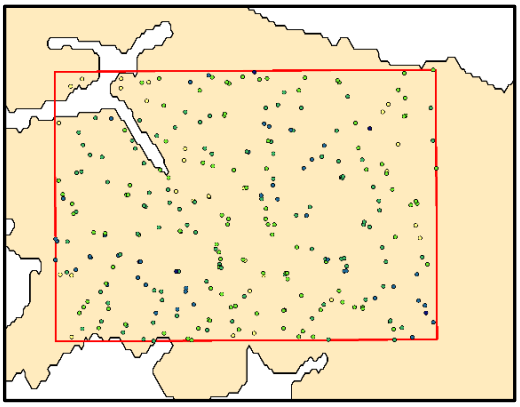

Figure 2. Cobalt Data

Then we will use estimated spatial dependencies from finite sampling point data, in the form of a function named ***empirical variogram*** (and hence its derived ***covariogram*** function), to evaluate cobalt values at ***each*** location of the entire study area, by carrying out the ***Kriging interpolation*** model (i.e., using discrete data to predict the continuous trend surface).

As Horsnail introduces in his article [Geochemical Prospecting], geological factors likely to influence mineral deposit and cause variations are basic considerations must be addressed in a geochemical prospecting survey. Hence, note (in Figure 2) these sample points exhibit some degree of curvilinear patterns, because minerals deposits are more likely to be found along stream beds and lake shores, and smaples are mainly taken there. Moreover, there appears to be some diagonal "waves" of high and low values rippling through the site, which are roughly parallel to the Pacific coastline, and would seem to reflect the history of continental drift in this region.

2 METHODS & RESULTS

The first objective of this analysis is to capture the structure of spatial dependencies measured in variogram and covariogram among the cobalt data. But before launching into more elaborate explanation of these two functions, it's necessary to determine the basic modeling framework throughout the following analysis.

2.1 Spatial Dependencies

2.1.1 Basic Modeling Framework

First of all, we postulate that cobalt values, regarded as random variable $Y(s)$, at each corresponding location s, defined over the study region R, can be modeled by a ***spatial stochastic process*** (or ***random field***). We also assume that the existence of this given set of $n = 286$ observations (data points), $\{y_i = y(s_i): i = 1,\ldots,n\}$ in R, is a realization of the corresponding random variable. Essentially, the statistic variation of random variables, $Y(s)$, can be decomposed into two parts as the formula (2.1.1) defines:

$$(2.1.1) \qquad Y(s) = \mu(s) + \varepsilon(s) \quad , \quad s \in R$$

where $\mu(s)$ is usually taken to be the mean of $Y(s)$ as a ***spatial trend*** term. Hence, $\varepsilon(s)$ is a ***stochastic residual*** term with mean identically zero since:

$$(2.1.2) \qquad \varepsilon(s) = Y(s) - \mu(s) \Rightarrow E[\varepsilon(s)] = E[Y(s)] - \mu(s)$$

$$\Rightarrow E\left[\varepsilon(s)\right] = 0 \quad , \quad s \in R$$

Within such framework, the former part $\mu(s)$ refers to the ***global structure*** of the stochastic Y-process. Likewise, the latter term $\varepsilon(s)$ can be regarded as a ***spatial residual process*** representing ***local structure*** about Y-process, which consists of all unobserved spatial variables influencing $Y(s)$ that are not captured by the global trend.

2.1.2 Spatial Modeling Strategy

Hence, if we find a spatial trend function $\mu(\cdot)$ that well fits the Y-process, the rest residual function $\varepsilon(\cdot)$ will be not statistically distinguishable from "random noise". But it's too restrictive to expect these residuals to be independent as in usual statistic model. Like most to some extent spatially continuous variables that can be expected to present similar values in vicinity, we can also expect that associated residuals $\varepsilon(s)$ and $\varepsilon(v)$ at sufficiently close locations s and v, will exhibit ***positive spatial dependence***. The degree of such dependency between spatial residuals and even ***any*** random variables X and Y can be measured as ***covariance***:

$$(2.1.3) \qquad \mathrm{cov}(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right]$$

Here, ***positive*** $\mathrm{cov}(X, Y)$ indicates that the deviations $X - \mu_X$ and $Y - \mu_Y$ share the ***same*** signs, while ***negative*** $\mathrm{cov}(X, Y)$ represents ***opposite*** signs. And if the $\mathrm{cov}(X, Y)$ is ***zero***, the deviations $X - \mu_X$ and $Y - \mu_Y$ are expected

to be ***not statistically related***. In particular, it's expected to observe positive dependencies among spatial residuals that can be reflected by positive covariance.

2.1.3 Covariance Stationarity

A spatial stochastic process, $\{Y(s): s \in R\}$, is said to be ***covariance stationary*** if and only if it holds such two conditions for any pairs of two locations $s, v \in R$, separated by the distance, $\|s - v\| = h$:

(2.1.4) $E\left[Y(s)\right] = \mu$ (***common mean value***)

(2.1.5) $\|s - v\| = h \Rightarrow \text{cov}[Y(s), Y(v)] = C(h)$ (***common covariance value*** for each distance h)

Since the global trend $\mu(s)$ in have a constant mean μ as implied in (2.1.4) (i.e., $\mu(s) \equiv \mu$). Hence, this process can be written as:

(2.1.6) $Y(s) = \mu + \varepsilon(s)$

where a unique ***residual process*** $\{\varepsilon(s): s \in R\}$, is related to such process. Notice that such spatial stochastic with ***constant mean*** under ***covariance stationarity*** will ***hold*** throughout the following analysis as an important assumption.

Moreover, note that from (2.1.5) that since $\|s - s\| = 0$ and $\text{var}[Y(s)] = \text{cov}[Y(s), Y(v)]$ by definition, it's concluded that these random variables have a ***common variance*** σ^2, given by:

(2.1.7) $\text{var}[Y(s)] = C(0) = \sigma^2, \quad s \in R$

Moreover, since $\text{cov}[Y(s), Y(v)] = E([\mu + \varepsilon(s) - \mu]\,[\mu + \varepsilon(v) - \mu]) = E(\varepsilon(s) \cdot \varepsilon(v))$, such process must satisfy the conditions that:

(2.1.8) $E[\varepsilon(s)] = 0$

(2.1.9) $\|s - v\| = h \Rightarrow E[\varepsilon(s)\varepsilon(v)] = C(h)$

2.2 Variogram and Covariogram

2.2.1 Covariogram and Correlogram

Notice that since the covariance values $C(h)$ discussed above, are unique for each given distance h, in region R, for the covariance stationary process, we can define a function, C, of these distances as ***covariogram***:

(2.2.1) $C(h) = \text{cov}\left[Y(s), Y(v)\right] \quad , \quad s, v \in R \quad \text{and} \quad \|s - v\| = h$

However, the normalized form of covariogram, the ***correlogram***, is more convenient for analyzing spatial

dependencies, because the magnitude of covariogram values in itself is not informative without specific units in which variables are measured. Moreover, it's difficult to interpret these values considering that covariogram are in the form of squared units. The correlogram is related to the ***correlation coefficients***, which are normalized covariance of two random variables dividing by the production of their standard deviations:

$$(2.2.2) \qquad \rho\left[Y(s), Y(v)\right] = \frac{\text{cov}[Y(s), Y(v)]}{\sqrt{\text{var}[Y(s)]}\sqrt{\text{var}[Y(v)]}} = \frac{C(h)}{\sqrt{C(0)}\sqrt{C(0)}} = \frac{C(h)}{C(0)}$$

Hence, the correlogram is defined for each distance, h, for the covariance process:

$$(2.2.3) \qquad \rho(h) = \frac{C(h)}{C(0)}$$

2.2.2 Variogram

Though it's intuitive to interpret the structure spatial dependencies using covariogram and its normalized form, the correlogram, such functions are more difficult to be estimated mathematically than ***variogram***. Essentially, the variogram is a function for describing the degree of spatial dependencies of a spatial stochastic process. Taking the cobalt geochemical prospecting as an example, a variogram measures how much cobalt percentages of two given samples taken from the study area will vary depending on the distance between respective locations of these samples. Here, variogram is defined as the variance of the difference between values at two points, that is, the ***expected squared difference*** of any pair of component variables $Y_s = Y(s)$ and $Y_v = Y(v)$ from a covariance stationary process:

$$(2.2.4) \qquad E\left[\left(Y_s - Y_v\right)^2\right] = E\left[Y_s^2 - 2Y_sY_v + Y_v^2\right] = E\left(Y_s^2\right) - 2E\left(Y_sY_v\right) + E\left(Y_v^2\right)$$

Then we have:

$$(2.2.5) \qquad C(h) = \text{cov}(Y_s, Y_v) = E[(Y_s - \mu)(Y_v - \mu)] = E[Y_sY_v - Y_s\mu - \mu Y_v + \mu^2$$

$$= E(Y_sY_v) - E(Y_s)\mu - \mu E(Y_v) + \mu^2$$

$$= E(Y_sY_v) - \mu^2 - \mu^2 + \mu^2 = E(Y_sY_v) - \mu^2$$

$$(2.2.6) \qquad \Rightarrow E(Y_sY_v) = C(h) + \mu^2$$

If s and v both refer to the same location, that is, $s = v$ and $\| s - v \| = 0$, the formula in (2.2.6) then is:

$$(2.2.7) \qquad E(Y_s^2) = C(0) + \mu^2 = E(Y_v^2)$$

Hence, by substituting (2.2.6) and (2.2.7) into (2.2.5), we see that expected squared differences for all $s, v \in R$ with distance h can be written in the form of the covariogram, C, as:

(2.2.8) $$E\left[\left(Y_s - Y_v\right)^2\right] = 2 \cdot \left[C(0) - C(h)\right]$$

For the purpose of simplicity and convenience, a "scaled" version by suppressing the factor "2" is introduced:

(2.2.9) $$\gamma(h) = \frac{1}{2} E[(Y_s - Y_v)^2] = C(0) - C(h) = \sigma^2 - C(h) \quad , \quad \|s - v\| = h$$

which is often called as "semivariogram". [Note: But since this scaled version is the ***only*** usually used form in practice, it's natural to use the term "variogram" to refer to "semivariogram" in the following analysis of this report.] Moreover, the function formula in (2.2.9) also implies that

(2.2.10) $$C(h) = C(0) - \gamma(h) = \sigma^2 - \gamma(h)$$

which is exactly the derived covariogram we have mentioned above from original variogram that is more easily estimated.

2.2.3 Standard Model

Recall that these two functions we have defined are intended to capture and describe the structure of spatial dependencies. As the ***First Law of Geography*** says, "Everything is related to everything else, but near things are more related than distant things", so that samples taken in far distances will vary more than samples taken close to each other for covariance stationary. Specifically, in terms of covariogram, correlations are high for pairwise values at small distances and will become less as distances increase. Hence, covariograms fall monotonously from $C(0) = \sigma^2$ to zero as distance increases for the ***standard covariogram model***, as shown in Figure 3. The right end of this curve will either be ***exactly zero*** at all greater distances or ***approach zero asymptotically*** (i.e., covariance values will become arbitrarily small yet above zero). The Figure 4 on the right is the plot of associated variogram, indicating that similarly, the standard variogram must either reach the dashed line called ***sill***, or approach this sill asymptotically. It indicates that the expected squared differences, variogram values, will increase as the distance between two variables increases, that is, variables become less similar.

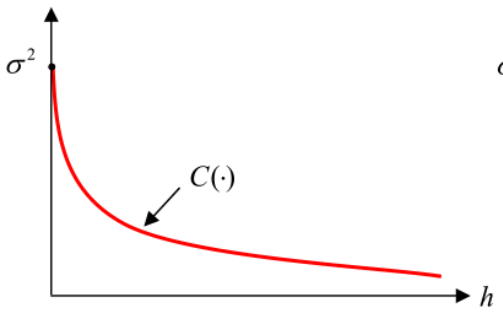

Figure 3. Standard Covariogram

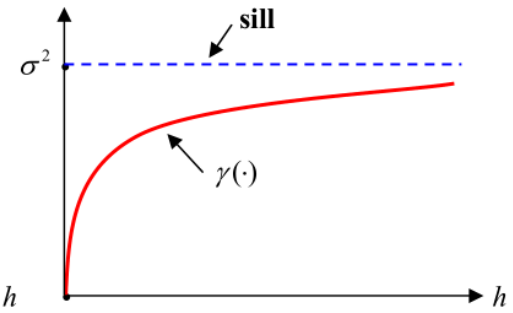

Figure 4. Standard Variogram

2.2.4 Combined Model

In the context of this cobalt study, it's expected that both variograms and covariograms of cobalt values will exhibit such variance trends as the standard model (shown in Figure 3 and 4) implies that the degree of spatial dependencies gradually declines towards zero while the differences correspondingly increases towards the common covariance value. However, such model is exactly an ***extreme*** situation that seldom occurs in practice. Actually, most practical processes are a ***mixture*** of such standard model and another mathematically idealized extreme called ***pure spatial independence model***, in which distinct random components, $Y(s)$ and $Y(v)$, have no relation to each other no matter how close or far they are in space. Mathematically, $\text{cov}[Y(s), Y(v)] = 0$ for all s and v. As the Figure 5 presents, covariogram, C, is consistently equal to 0 except at the origin where it exhibits a ***discontinuity*** because $\text{cov}[Y(s), Y(s)] = \sigma^2 > 0$ for all s. Correspondingly, for pure spatial independence, variogram γ, must also exhibit a discontinuity at the origin as shown on the right in Figure 5.

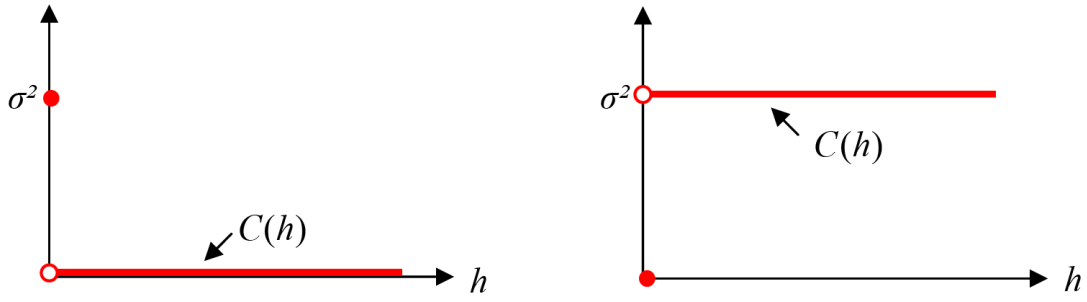

Figure 5. Pure Spatial Independence

Hence, the most common representation in actual processes is the ***combined model*** of these two extremes. Suppose that there are two types of measurement errors of cobalt values at location s, denoted respectively by $\varepsilon_1(s)$ and $\varepsilon_2(s)$. The former refers to the ***spatially dependent errors*** as in standard model (e.g., a mild fluctuation of sediment values in geography). Whiel the latter represents ***spatially independent errors*** as in pure independent model (e.g., the inherent ***system error*** casued by measuring instruments or methods). Then it's natural to postulate that total measurement erros $\varepsilon(s)$ is the sum of these two components:

$$\varepsilon(s) = \varepsilon_1(s) + \varepsilon_2(s) \tag{2.2.11}$$

Moreover, it may be assumed that $\varepsilon_1(s)$ and $\varepsilon_2(s)$ are independent random variables for every pair of locations, $s, v \in R$. Then it follows that the sum of the separate covariograms, C_1 and C_2, for component process ε_1 and ε_2, will be the covariogram C, of error process for the combined model:

$$C(h) = C_1(h) + C_2(h) \quad , \quad h \geq 0 \tag{2.2.12}$$

In addition, the formula in (2.2.12) also implies that:

$$\sigma^2 = C(0) = C_1(0) + C_2(0) = \sigma_1^{\,2} + \sigma_2^{\,2} \tag{2.2.13}$$

where σ_1^2 and σ_2^2 are the corresponding variances for the spatially dependent and independent components, respectively, represented better by the Figure 6 of covariogram for the mixture process below.

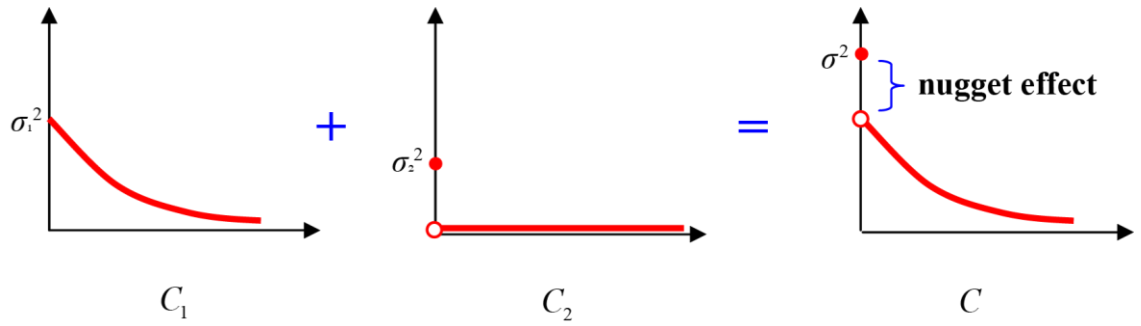

Figure 6. Covariogram for Combined Model

Obviously, the covariogram for the combined model is almost the same as that of the standard model, except for the discontinuity called ***nugget effect*** at the origin, resulting from the pure independent model. The magnitude of this effect is also simply the variance, σ_2^2, of the pure independent component called the ***nugget***. Since it's caused by the spatial independence in the actual process, the ratio named relative nugget effect, σ_2^2/σ^2, represents the relative spatial independence.

Then when it comes to the associated combined variogram, it follows the similar form as covariogram and random errors:

(2.2.14) $$\gamma(h) = \sigma^2 - C(h) = (\sigma_1^2 + \sigma_2^2) - [C_1(h) + C_2(h)]$$

$$= [\sigma_1^2 - C_1(h)] + [\sigma_2^2 - C_2(h)]$$

(2.2.15) $$\Rightarrow \quad \gamma(h) = \gamma_1(h) + \gamma_2(h),$$

where γ_1^2 and γ_2^2 are the variograms for the spatially dependent and independent components, respectively. The combined variogram is shown as the Figure 7 below.

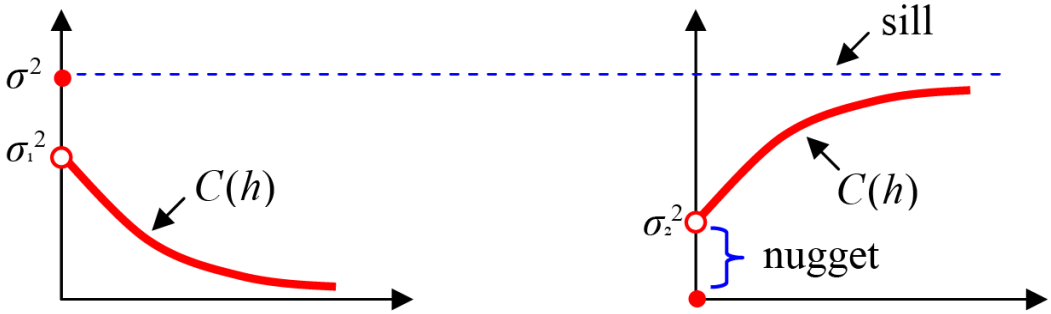

Figure 7. Summary of the Combined Model

2.2.5 Empirical Variogram

Then we will apply the defined empirical variogram and its derived covariogram to our cobalt data set to. Notice that both of these functions describe relation of cobalt values at corresponding distances for all pairwise locations. However, one problem of estimated $\gamma(h)$ with distances is that in any practical finite sample there will generally be at most one pair of locations separated by a specific distance h, which will result in a large number of point pairs if we identify each respective distance, as shown in the Figure 9 of variogram cloud below.

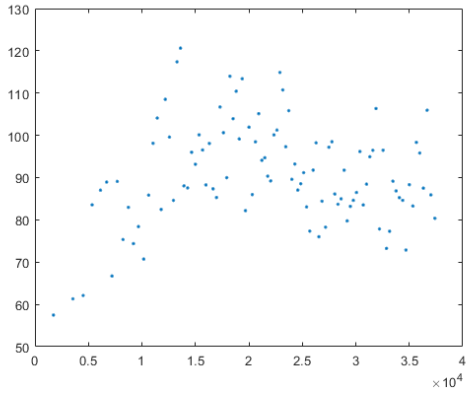

Figure 8. Empirical Variogram

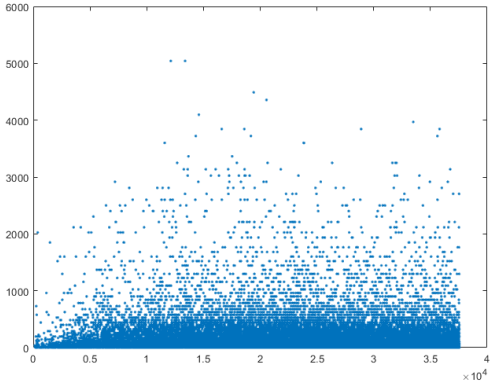

Figure 9. Variogram Cloud

The simplest way to solve this problem is to aggregate point pairs with similar distances and hence estimate $\gamma(h)$ at only a small number of representative distances for each group. For comparison, the Figure 8 is an empirical variogram plot by using such aggregation method. Specifically, to acquire these aggregated representative empirical variograms for all point pairs, we can divide distances into intervals, called ***bins***, and use the average distance, h_k, in each bin k to represent each of aggregated distances, called ***lag distances***, as shown in Figure 10 below.

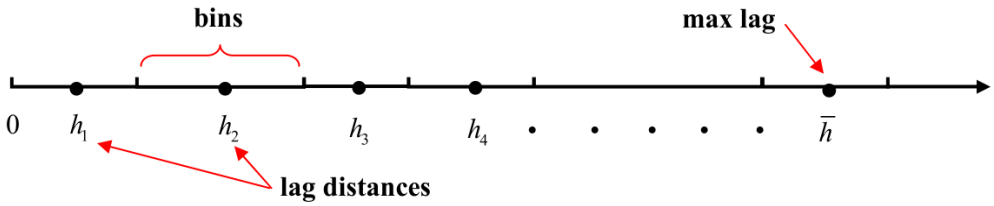

Figure 10. Lag Distances and Bins

In addition, if we denote the set of distance pairs, (s_i, s_j), in each bin k, as N_k, the size (number of point pairs) as $|N_k|$ and the distance between each such pair as $h_{ij} = ||s_i - s_j||$. Then the lag distance h_k, for bin k is defined as:

$$(2.2.16) \quad h_k = \frac{1}{|N_k|} \sum_{(s_i, s_j) \in N_k} h_{ij}$$

To ensure a uniform approximation of lag distances, all bins are expected to be same size (same number of point pairs in each bin). Also, to balance the approximation of lag distances and the number of pairs used to estimate the average variogram at each lag distance, the standard rule of thumb is that the size of each bin should not be smaller than (30) point pairs:

(2.2.17) $|N_k| \geq 30$

Then we denote the ***maximum lag distance*** of all point pairs in the data set as $h_{\max}$. From practical experience, a proper $\bar{h}$ partitioned should be restricted to half of $h_{\max}$:

(2.2.18) $\bar{h} \leq \frac{h_{max}}{2}$

It then follows that for any given set of sample points, $\{s_i: i = 1,\ldots,n\} \subset R$, with associated data, $\{y(s_i): i = 1,\ldots,n\}$, an appropriate estimate of the variogram value, $\gamma(h_k)$, at each lag distance, $h_k \leq \bar{h}$, is given by half the average squared differences $(y(s_i) - y(s_j))^2$ over all point pairs (s_i,s_j) in N_k:

(2.2.19) $\gamma(h_k) = \frac{1}{2|N_k|} \sum_{(s_i,s_j)\in N_k} \left(y(s_i) - y(s_j)\right)^2$

A schematic example of this empirical variogram construction is given in Figure 11 below. Here, the ***blue dots*** represent squared-difference pairs $(y(s_i) - y(s_j))^2$, against distances $h_{ij} = \|s_i - s_j\|$, for each point pair, (s_i,s_j). The bins are separated by the vertical lines. In each bin k, the ***red dot*** in the middle denotes the estimation $\gamma(h_k)$ at average distance h_k, as a representation of all points in this bin.

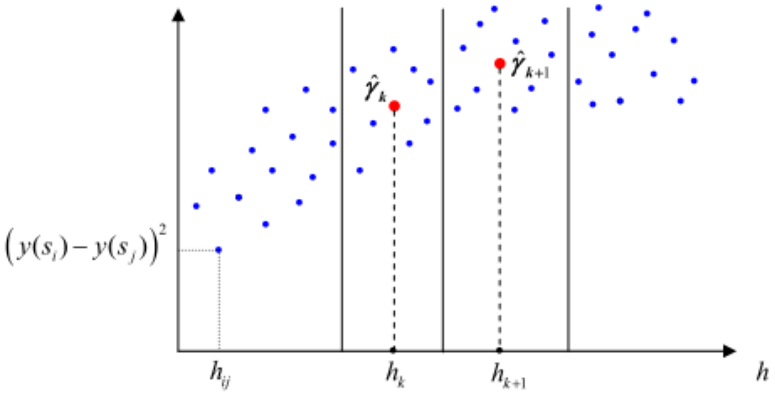

Figure 11. Empirical Variogram Construction

Estimation 1 (Max-Distance = 20,000 Meters):

For this cobalt study, the default number of bins is 100. Firstly, we try a variogram with this maximum distance set as 20000 meters. The plot of estimated variogram and covariogram points at corresponding distances are displayed in Figure 12 below.

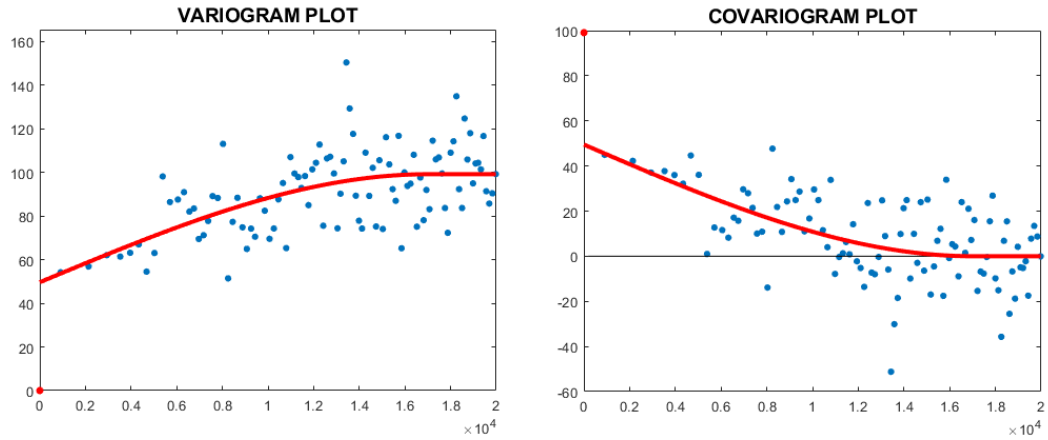

Figure 12. Empirical Variogram Construction (with 20,000-Meter Maximum Distance)

2.2.6 The Fitted Spherical Model

While the combined model provides a useful conceptual framework for understanding and describing variograms and covariograms, what's more, sufficiently explicit mathematical models are required to be estimated statistically. Notice that there is a ***red line*** with a discontinuity at the origin in the variogram plot of Figure 12, which is exactly the most widely used model to fit these empirical variograms, named ***spherical variogram***, defined for all $h \geqslant 0$ by:

$$(2.2.20) \quad \gamma(h;r,s,a) = \begin{cases} 0 & , \quad h = 0 \\ a + (s-a)\left(\frac{3h}{2r} - \frac{h^3}{2r^3}\right) & , \quad 0 < h \leq r \\ s & , \quad h > r \end{cases}$$

Here, parameters (r,s,a) of γ are assumed to satisfy $r,s > 0$, $a \geq 0$ with $s > a$. Recall that covariogram is more difficult to be estimated. Hence, the associated spherical covariogram to expression (2.2.20) is given by:

$$(2.2.21) \quad C(h;r,s,a) = \begin{cases} s & , \quad h = 0 \\ (s-a)\left(1 - \frac{3h}{2r} + \frac{h^3}{2r^3}\right) & , \quad 0 < h \leq r \\ 0 & , \quad h > r \end{cases}$$

Such specific spherical variogram and covariogram models are qualitatively consistent with the combined model, with a small number of parameters that can be estimated. These parameters' mathematical meanings can be directly delivered by observing the Figure 13 and 14. It's clear to find out that s and a respectively refer to the ***sill*** and ***nugget*** we have mentioned above. And r refers to ***range***, the distance at which this curve starts to level off, that is, data values are considered not statistically related any more beyond this distance.

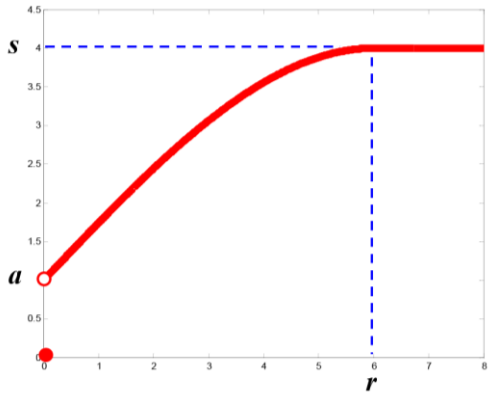

Figure 13. Spherical Variogram

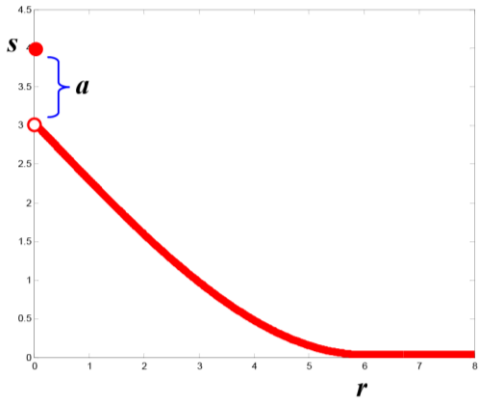

Figure 14. Spherical Variogram Variogram

Given the framework of spherical models at hand, the remaining task is to seek appropriate parameter values, $(\hat{r}, \hat{s}, a)$ to fit the empirical variogram data, where the squared differences can be minimized, known as ***least-squares*** problem:

(2.2.22) $$\min_{(r,s,a)} \sum_{k=1}^{\bar{k}} \left(\gamma_k - \gamma(h_k; r, s, 0)\right)^2$$

Recall we have mentioned that $a \geq 0$ above, but nonnegativity of the nugget a, will fail to hold in some cases, as illustrated on the left of the schematic Figure 15 below.

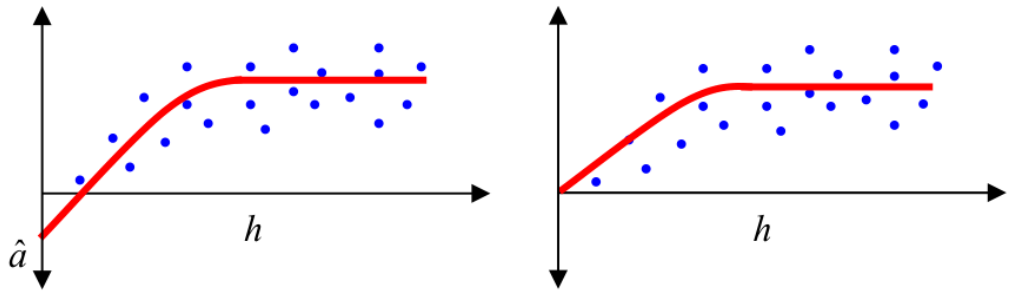

Figure 15. Negative Nugget Problem

Hence, it is natural to constrain $a = 0$, so that the square-minimization problem used to find the "best fit" spherical models in (2.2.22) becomes:

(2.2.23) $$\min_{(r,s,a)} \sum_{k=1}^{\bar{k}} \left(\gamma_k - \gamma(h_k; r, s, a)\right)^2$$

The results of this estimation procedure with 20,000-meter maximum distance are shown in the Table 1.

Table 1. Parameters of Empirical Variogram (with 20,000-Meter Maximum Lag Distance)

MAX-DIST	RANGE	SILL	NUGGET	ITERATIONS
20000	17009.836	99.104	49.626	117

In particular, the RANGE = 17009.836 meters denotes the distance at which variogram "reaches the sill" and starts to level off, that is, beyond this distance there is estimated to be no statistical correlation between cobalt values. However, as we have mentioned above that it's difficult to interpret variogram, so that it's more intuitive to interpret in the form of the derived covariogram. It's this distance = 17009.836 meters that denotes the outset from which covariance and hence correlation ***first falls to zero***. Turning to the other estimated parameters, the SILL = 99.104 represents the estimated variance of individual cobalt values (i.e., the estimated covariance at "zero distance"). Additionally, the NUGGET = 49.626 denotes the magnitude of spatial independence. Hence, in this case, the relative nugget effect here is calculated as 49.626 / 99.104 = 0.501, indicating that a considerable degree of local spatial dependence exists among cobalt values.

Estimation 2 (Max-Distance = 40,000 Meters):

Again, we try another estimation of variogram by setting maximum distance as 40000 meters. The fitted spherical models for empirical variogram and covariogram are shown in the Figure 16, together with estimated parameters shown in the Table 2.

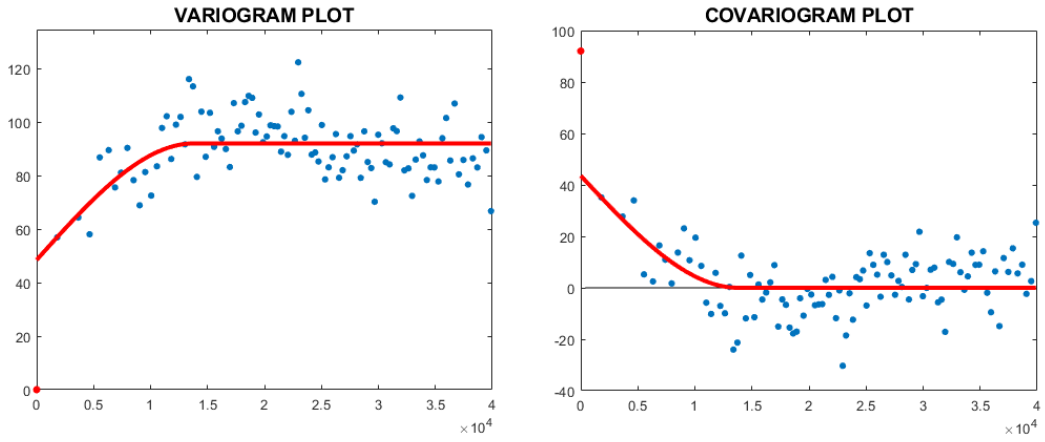

Figure 16. Empirical Variogram Construction (with 40,000-Meter Maximum Lag Distance)

Table 2. Parameters of Empirical Variogram (with 40,000-Meter Maximum Distance)

MAX-DIST	RANGE	SILL	NUGGET	ITERATIONS
40000	13767.093	92.024	48.488	134

In particular, covariance and hence correlation begin to fall to zero at the distance (RANGE) = 13767.093 meters, well smaller than that when maximum distance equals to 20000 meters. The estimated variance at

distance = 0 is 92.024 denoted as SILL. Here, with NUGGET = 48.488, spatial independence among cobalt values is estimated as 48.488 / 92.024 = 0.527, slightly higher than last try.

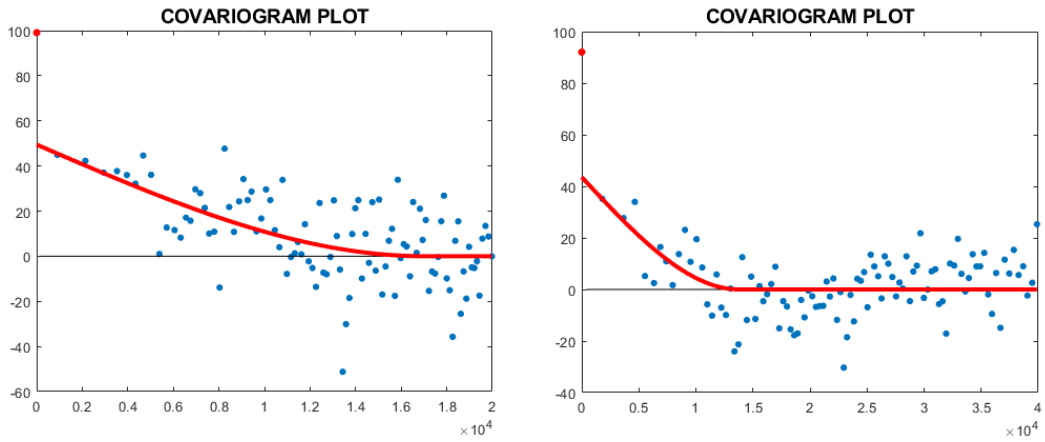

Figure 17. Comparison of Empirical Covariogram (with Different Maximum Distance)

For convenience, a comparison of the estimated covariogram plots using different maximum distances are displayed as the Figure 17 above. The substantial difference of estimated ranges when applying different maximum distances can be explained by the distribution pattern of cobalt values displayed in Figure 18 below.

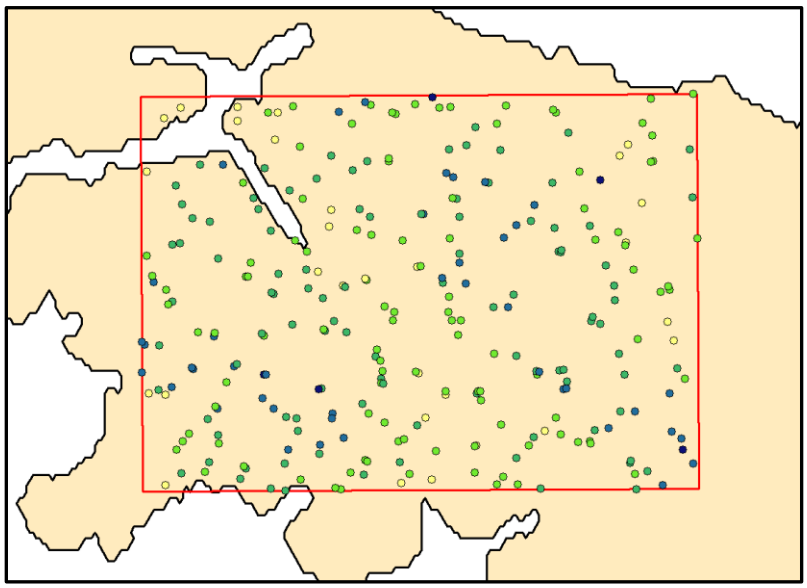

Figure 18. Distribution Pattern of Cobalt

Visually, it turns out that these sampling points exhibit some degree of clustering, where some points even "overlay" each other (i.e., close to each other). The Average Nearest Neighbor Tool in ArcMap yields that the actual mean nearest-neighbor distance of this cobalt data is 1395.42 meters, compared to expected mean distance = 1584.55 meters under CSR, which again supports the clustering of data points quantitatively. Hence, in this case, the quarter of observed maximum pairwise distance = 20,000 meters is sufficient to capture enough points in each bin to estimate average variogram (and hence covariogram). By contrast,

practical experience of using "h_{max} / 2" value can't work satisfyingly. Hence, 40,000 meters is too large to include enough points for bins of large lag distances, since clustered points exhibit less pairs at large distances. The results of 20,000-meter max-distance will be applied into kriging prediction at a specific point using MATLAB in the following section 2.3.3.

2.3 Kriging Interpolation

Given the model of stationary random spatial effects above, we can apply these concepts to spatial stochastic models of the form, $Y(s) = \mu(s) + \varepsilon(s)$, $s \in R$. In continuous spatial data analysis, we focus on ***spatial prediction*** models, where the known values at given locations are used to estimate a continuous surface (i.e., predict values at other locations). Hence, we can "fill in the gaps" between known data points. In particular, such models are referred to as ***spatial interpolation*** (or ***smoothing***) models. There are several types of spatial interpolation, including ***inverse distance weighting*** (***IDW***), ***spline***, and ***Kriging***. It's the ***Kriging interpolation*** model that we will apply to this cobalt data in the second part of this study.

2.3.1 Overview of Kriging Models

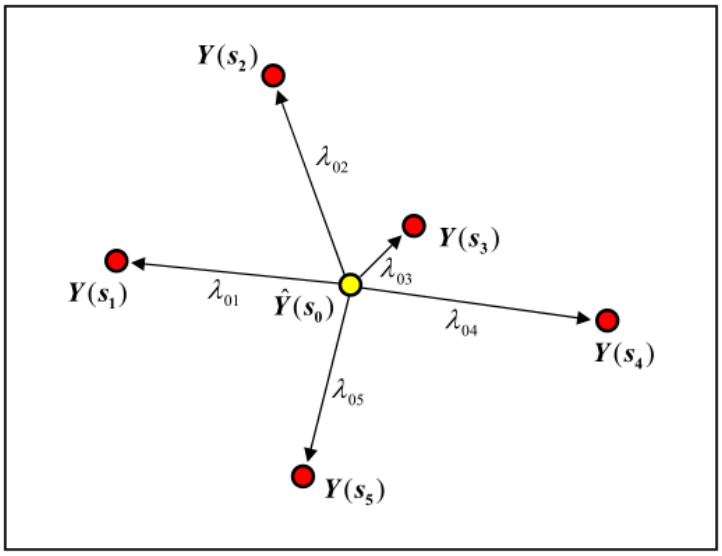

Figure 19. Basic Kriging Framework

As the Figure 19 shows, the fundamental strategy of Kriging models is to predict unknown values based on local information., by treating the observed data as a finite sample from a spatial stochastic process. Hence, we assume that some appropriate subset of sample locations,

(2.3.1) $S(s_0) \subseteq \{s_i : i = 1, ..., n\}$

is chosen for prediction. Through the choice of ***prediction set*** at s_0 is substantially deterministic in the final prediction, the best way to determine a "best prediction set" can start with finding a "best prediction" for any given set and then compare these predictions.

Given prediction set, $S(s_0) = \{s_{1,\ldots,}s_{n_0}\}$, we can treat the prediction, $\hat{y}(s_0)$, based on the sample data, $\{y(s_1),...,y(s_{n_0})\}$, as a ***random variable***, $\hat{Y}(s_0)$, which can be assumed as some ***linear combination*** of subset random variables, $\{Y(s_1),...,Y(s_{n_0})\}$:

(2.3.2) $\hat{Y}(s_0) = \sum_{i=1}^{n_0} \lambda_{0i} Y(s_i)$

2.3.2 Best Linear Unbiased Predictor

Then the key strategy of kriging models is to choose "statistically optimal" weights, λ_{0i}, in an appropriate way that focuses on the ***prediction error***,

(2.3.3) $e(s_0) = Y(s_0) - \hat{Y}(s_0)$

For the prediction error, it should follow three criteria that:

(2.3.4) (i) $E[e(s_0)] = E[Y(s_0) - \hat{Y}(s_0)] = 0$ (***unbiasedness criterion***);

(2.3.5) (ii) ***mean squared error*** $E[e(s_0)^2]$ is as small as possible;

(2.3.6) (iii) $\hat{Y}(s_0)$ have ***minimum mean squared error among all linear unbiased predictors*.**

where the third criterion has another name of ***minimum mean squared error***, usually known as ***MMSE predictors***. Additionally, since $E[e(s_0) = 0]$ implies:

(2.3.7) $\text{var}[e(s_0)] = E[e(s_0)^2] - (E[e(s_0)])^2 = E[e(s_0)^2]$,

Such predictors are also instances of ***minimum variance predictors***, which are also be designated as ***best linear unbiased predictors***, or ***BLU predictors***.

2.3.3 Simple Kriging at One Single Location

We will start with predicting the cobalt value and standard error of prediction at the specific point location, s = (615000, 586500) using a Kriging ***bandwidth*** of 5,000 meters, defined as the radius of a circle around this location. Here, we focus more on the general procedure and output of simple kriging than explanation of how it works. More details will be explained in section 2.3.4. The location of s (marked as a purple triangle) in the study region and other cobalt sampling points are given in Figure 20, where it turns out that such large a bandwidth is sufficient to capture the most important 6 neighbors of point s, as the prediction set to estimate cobalt value at s.

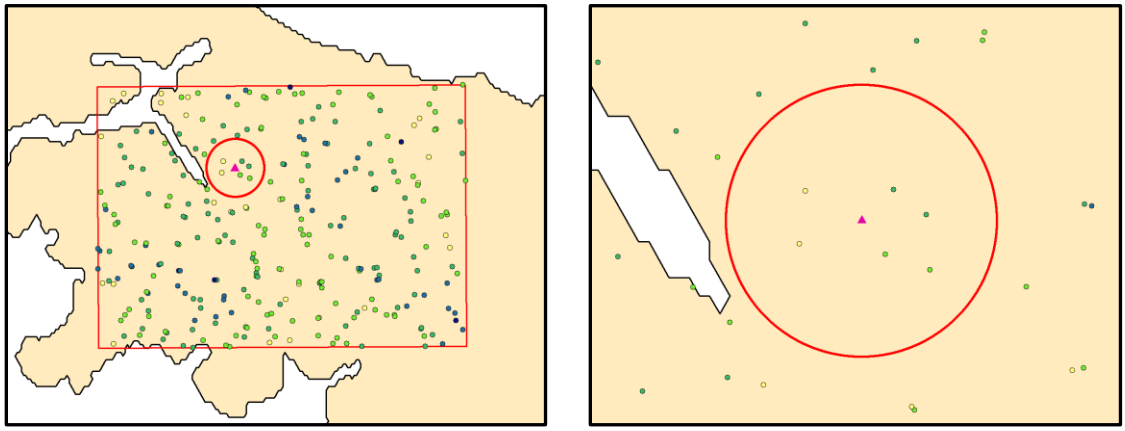

Figure 20. Point s and its Prediction Set

By telling MATLAB the whole cobalt data set, maximum lag distance = 20,000 meters and default number of bins = 100 (as used in MATLAB above), the estimated variograms and covariograms of cobalt values will easily calculated. Then given the location of s and bandwidth = 5,000 meters, it's easy to obtain the ***sample mean***, desired ***simple kriging prediction*** and an associated estimate of the ***standard error of prediction*** at s, which are respectively denoted as MEAN, PREDICTION and SD in Table 3.

Table 3. Prediction at Point s

MEAN	PRIDICTION	SD
25.2727	17.7308	8.035

Since in terms of confidence levels, we will be 100(1 - α)% confident that kriging prediction Y_0 lies in the prediction interval,

(2.3.8) $$[\hat{Y}_0 - \sigma_0 z_{\alpha/2}, \hat{Y}_0 + \sigma_0 z_{\alpha/2},] = [\hat{Y}_0 \pm \sigma_0 z_{\alpha/2}],$$

and the most common instance used for the case is $\alpha = 0.05$, with corresponding critical value $z_{\alpha/2} = z_{.025} = 1.96$, we can determine a 95% prediction interval for this prediction = 17.7308 as $17.7308 \pm 1.96 \times 8.035 = [1.9822, 33.4794]$.

2.3.4 Simple Kriging for Cobalt Data

Finally, we will use the spherical variogram parameters for 20000-meter maximum lag distance case above to apply Simple Kriging model to cobalt data using the Geostatistical Analyst extension (GA) in ARCMAP.

Step 1. Estimation of Mean:

In Simple Kriging model, mean μ, is assumed known outside the model and an estimate of this mean can be acquired just from the sample mean of this data directly:

$$(2.3.9) \qquad \hat{\mu} = \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y(s_i) = \frac{1}{n}\sum_{i=1}^{n} y_i \ ,$$

Which is always ***unbiased*** since

$$(2.3.10) \qquad E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Notice that GA yields a Mean Value of cobalt data used for μ in the Simple Kriging model as 25.27273 as displayed in Figure 21, which is precisely the sample mean given in MATLAB as shown in Table 3 (denoted as MEAN) above. Both of these mean values are directly evaluated from the same cobalt data set.

Dataset #1	
Transformation type	None
Order of trend removal	None
Mean Value	25.27273

Figure 21. Mean Value of Cobalt Data

Step 2. Estimation of Covariances

Also, Simple Kriging assumes that the covariances cov[$\varepsilon(s)$, $\varepsilon(s')$], are known and given for all locations, s, $s' \in R$. Specifically, under covariance stationarity, all covariances can be summarized by a covariogram $C(h)$. As emphasized on the difficulty of estimating covariogram, covariances can be best estimated by first estimating a variogram, $\gamma(h;r,s,a)$ with parameters, r = range, s = sill, and a = nugget. Since the common variance $C(0) = \sigma^2$, is precisely the sill parameter s, We can then obtain the desired covariogram from:

$$(2.3.11) \qquad C(h) = \sigma^2 - \gamma(h) = s - \gamma(h;r,s,a) \quad \text{and hence} \quad \hat{C}(h) = \hat{s} - \gamma(h;\hat{r},\hat{s},\hat{a})$$

Recall that for any pair of point, s, $s' \in R$, separated by distance $\| s - s' \| = h$, the quantity $\hat{C}(h)$ in (2.3.11) yields an estimate of cov[$\varepsilon(s)$, $\varepsilon(s')$]:

$$(2.3.12) \qquad \text{cov}[\varepsilon(s), \varepsilon(s')] = \hat{C}(\|s - s'\|)$$

Using this identity, we can then estimate the full covariance matrix, C_0, relevant for prediction at any location

s_0. In particular, if we let $d_{ij} = \|s_i - s_j\|$ for each pairwise points, $s_i, s_j \in \{s_1, s_2, \ldots, s_{n_0}\}$, and set $\hat{\sigma}_{ij} = \hat{C}(d_{ij})$,

then we can obtain the following estimate of C_0 as:

$$\hat{C}_0 = \left(\begin{array}{c|ccc} \hat{\sigma}^2 & \hat{c}_{01} & \cdots & \hat{c}_{0n_0} \\ \hline \hat{c}_{10} & \hat{c}_{11} & \cdots & \hat{c}_{1n_0} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{c}_{n_0 0} & \hat{c}_{n_0 1} & \cdots & \hat{c}_{n_0 n_0} \end{array}\right) = \begin{pmatrix} \hat{\sigma}^2 & \hat{c}_0' \\ \hat{c}_0 & \hat{V}_0 \end{pmatrix} \tag{2.3.13}$$

To obtain a fit that is roughly comparable to that in MATLAB, we set the number of lags to 10 with a lag size of 2000 meters (yielding a maximum distance of 15 × 2000 = 20,000 meters) as shown in Figure 22. Here, the spherical models we have discussed will fit expression (2.3.11), then the estimated range of 16987.11 meters, denoted as Major Range, is remarkably close to the MATLAB value of 17009.84 meters in Table 2 above. Similarly, the estimated ***nugget*** value, 49.68, and ***sill*** value, (49.68 + 49.38 = 99.06), are also very close to those in Table 1. Based on these parameters, the relative nugget ratio of cobalt measurements is 49.68 / 99.06 = 0.502, which indicates there is a substantial local spatial independence among cobalt values. The slightly different outputs of MATLAB and GA can be explained by that GA takes direction into account when evaluating experimental variograms, while MATLAB regards distances of equal values as completely same.

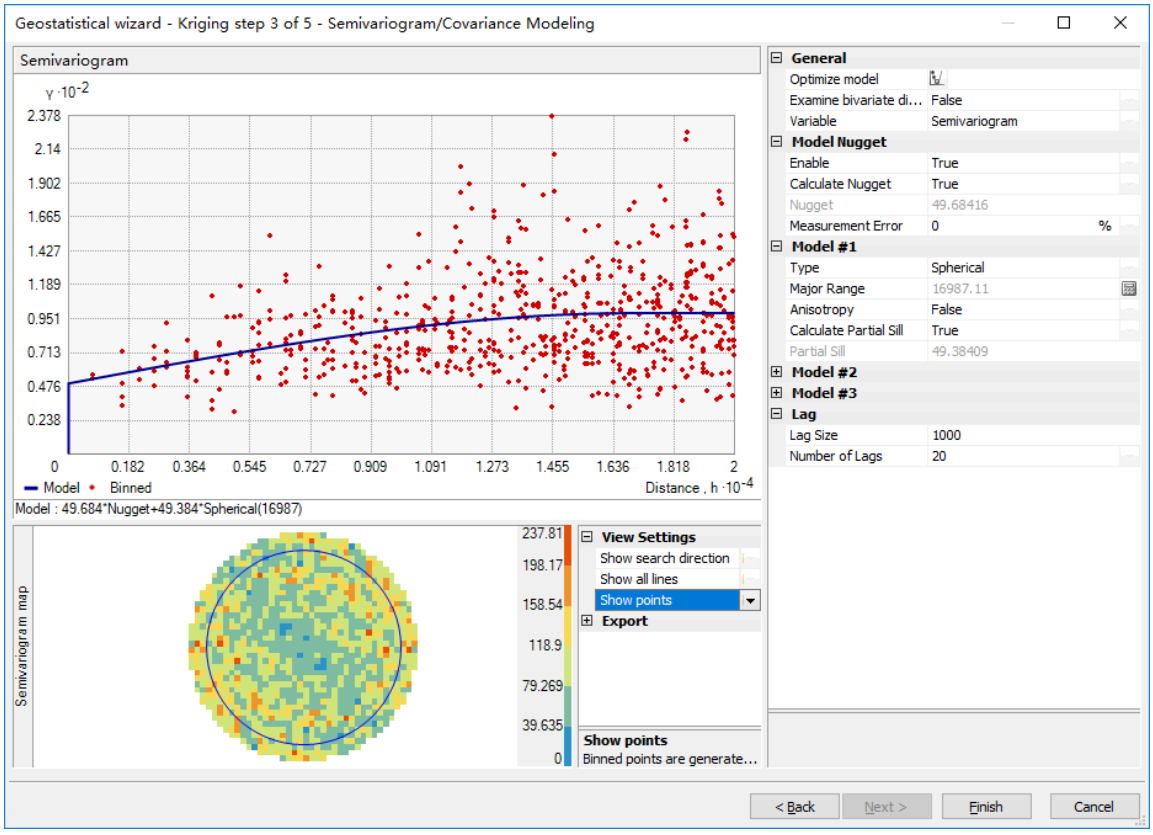

Figure 22. Variogram for Cobalt Data in GA

Step 3. Estimation of Kriging Predictions

Given these parameter estimates, we are ready to estimate the Simple Kriging prediction $\hat{Y}(s_0)$ at each location s_0 as:

(2.3.15) $\hat{\varepsilon}_0 = y_0 - \hat{\mu}$

Simple Kriging prediction of ε_0 at s_0 is given by figuring out the mean squared error (MSE) function:

(2.3.16) $MSE(\lambda_0) = \sigma^2 - 2c_0'\lambda_0 + \lambda_0' V_0 \lambda_0$,

where λ_0 is the weight vector $\lambda_0 = (\lambda_{01}, \lambda_{02}, ..., \lambda_{n_0})'$, of prediction set $S(s_0) = \{s_i : i = 1, ..., n_0\}$, for location s_0.

Then the expression (2.3.16) leads to:

(2.3.17) $\hat{\varepsilon}_0 = \hat{\lambda}_0' \varepsilon = \hat{c}_0' \hat{V}_0^{-1} \hat{\varepsilon}$

where $\hat{\varepsilon}$ is designated as the estimate of the deviation predictors $\hat{\varepsilon} = [\varepsilon_i : s_i \in S(s_0)]' = (\hat{\varepsilon}_1, ..., \hat{\varepsilon}_{n_0})'$.

Then the kriged values for each location of the study area from observed data (shown in Figure 23 below) are given as:

(2.3.18) $\hat{Y}_0 = \hat{Y}(s_0) = \hat{\mu} + \hat{c}_0' \hat{V}_0^{-1} \hat{\varepsilon}$

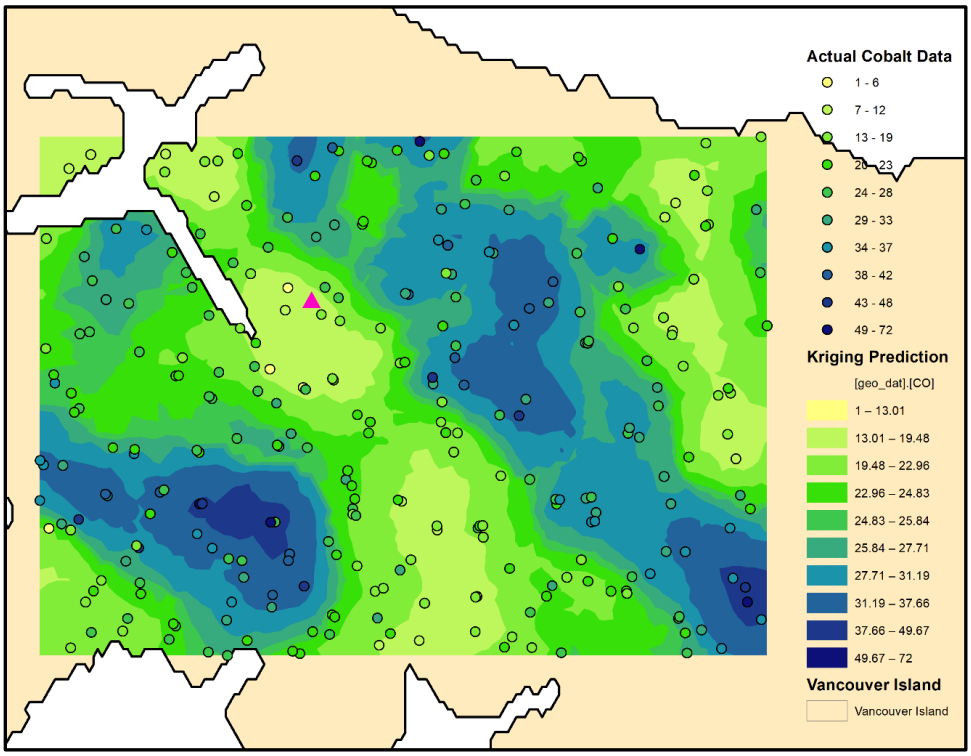

Figure 23. Simple Kringing Predictions

Then we will zoom in on the same point location in section 2.3.3, s = (615000, 586500), denoted as a purple triangle in Figure 23, and determine the Kriged value of Cobalt at this location. One method to approximately identify the prediction at s is to determine values of points just above and below point s from ArcMap, and averaging these. It turns out that the kriging predictions above and below this point in space are respectively 18.219 and 17.204. Then an approximate prediction at s is (18.219 + 17.024) / 2 = 17.6215 that is very close to predicted value = 17.7308 obtained in MATLAB above.

Also, a more direct and specific method to identify the prediction at s is to obtain it from GA. As shown in Figure 24 below, the sample point coordinates have been set to X = 615000 and Y = 586500 to agree with the point s. Similarly, to produce a circular neighborhood of 5000 meters, the "Sector type" is set to the simple ellipse form shown, and the axes are both set to 5000 to yield a circle.

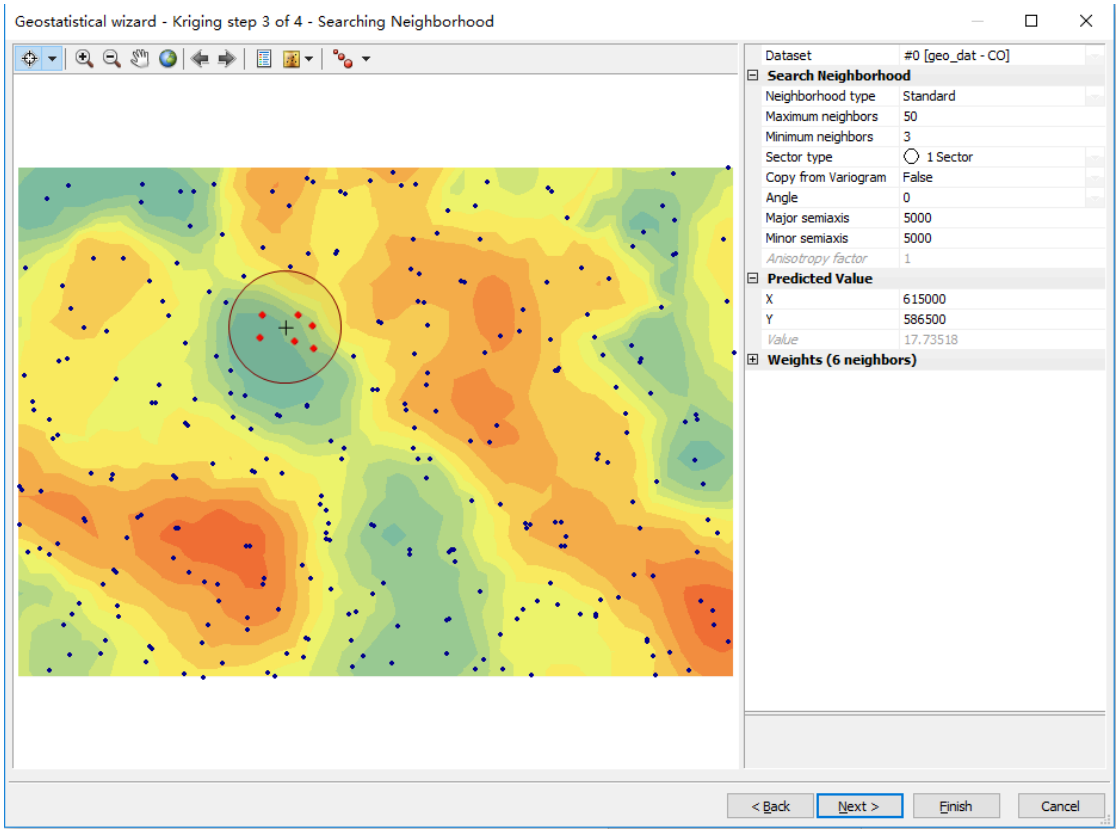

Figure 24. Kriging Prediction at Point s = (615000,586500)

Then the kriging prediction for cobalt value at this point is also displayed in Figure 24 above as "Value = 17.173518" located below the (X,Y) coordinate values. Here, exactly the 6 points inside the circle have been used for this kriging prediction. As expected, this value is seen to be quite close to the MATLAB prediction above. In addition, by examining the statistics of these 6 used points inside the circle in ArcMap, as shown in Figure 25 below, it turns out that the Kriged value = 17.173518 seem reasonable in relation to these data points, since it's well similar to the mean = 16.5 of these 6 neighbor points within 5,000 meters in space.

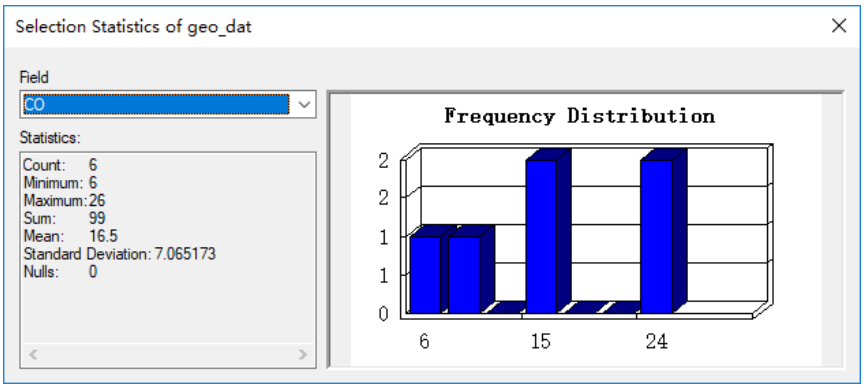

Figure 25. Statistics of Neighbor Points Around Point s = (615000,586500)

Finally, also in GA, we can determine the ***prediction error variance*** (or ***Kriging variance***) at s, as "Value = 8.038742" in Figure 26, which is close to that given in MATLAB. Here, the corresponding standard error is given by:

(2.3.19) $\hat{\sigma} = \sqrt{s - \hat{c}_0' \hat{V}_0^{-1} \hat{c}_0}$ with the (default) 95% prediction interval $[\hat{Y}_0 \pm (1.96)\hat{\sigma}_0]$

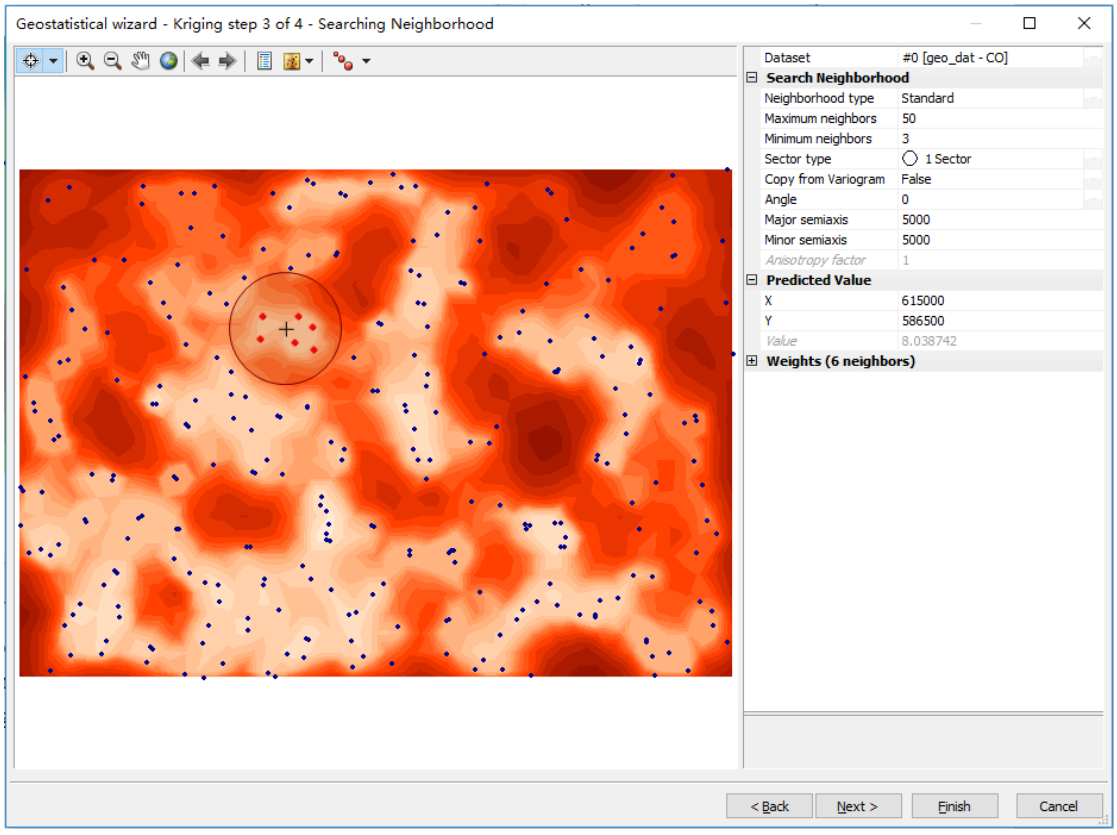

Figure 26. Standard Errors of Simple Kringing

For the purpose of comparison, all predicted result for both whole cobalt data and single point s in MATLAB

and GA are given as the Table below. Generally, MATLAB and GA yield similar kriging predictions in this case.

Table 4. Predictions Comparison

	WHOLE COBALT DATA				AT SINGLE POINT		
	MAX-DIST	RANGE	SILL	NUGGET	MEAN	PRIDICTION	SD
MATLAB	20000	17009.836	99.104	49.626	25.2727	17.7308	8.035
GA	20000	16987.11	99.06	49.68	25.27273	17.173518	8.038742

3 DISCUSSION

In this study, we have discussed necessary conceptual framework, spatial strategy and assumptions as the base of carrying out Kriging interpolation prediction (particularly for Simple Kriging). Given such very useful tool, Kriging, we have successfully and specifically predicted unknown cobalt measurements of the region of interest according to the spatial correlations among observed cobalt data.

4 REFERENCE

R. F. Horsnail, "Geochemical prospecting", in AccessScience@McGraw-Hill, http://www.accessscience.com, DOI 10.1036/1097-8542.285700, last modified: March 29, 2001.