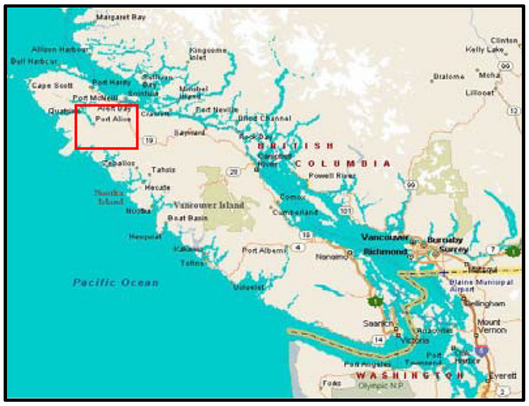
ASSIGNMENT 4

Study 1: Ordinary Kriging Prediction for Manganese Values

1 INTRODUCTION

This study is an extension of Study 1 in previous Assignment 3 [A3], in which we applied ***Simple Kriging*** interpolation model, to cobalt value data collected from Vancouver Island in British Columbia (shown in Figure 1). Recall that the ultimate objective of this kind of spatial analysis in the context of geochemical processing is to find the most ideal mining locations considered to contain more abundant mineral content. Since it is generally very costly and even infeasible to examine deposit values at ***each*** location, in the case of cobalt data, we alternatively switched to predict unknown cobalt measurements of the entire study area, according to observed cobalt values at 286 respective sampling sites. Likewise, this time we will examine concentration of Manganese (Mn) and predict where Manganese deposits are more likely to be found in the same region using ***Ordinary Kriging***, which is ***almost identical*** to Simple Kriging, except for that the constant mean, μ , is now assumed to be ***unknown***, and hence must be estimated within the model.

The data set used here is still from the Geological Survey of Canada, while we shift to Manganese content. As Figure 2 shown below, the same 286 sampling locations is bounded by the minimum enclosing rectangle outlined in red as we have displayed in [A3] that refers to the region of interest, where lower Manganese values are denoted as yellow dots while higher values are denoted as dark blue ones.



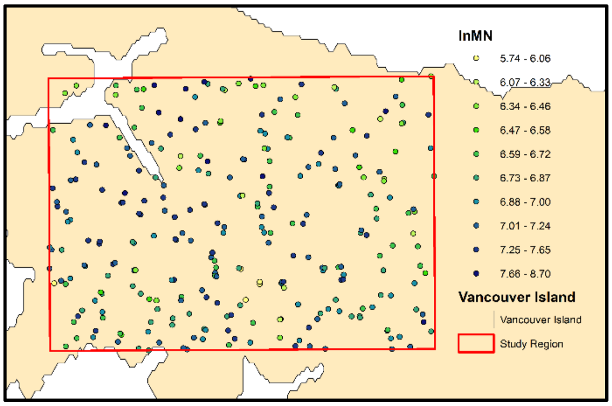


Figure 1. Vancouver Sample Area **Figure 2. Manganese Data**

2 METHODS & RESULTS

Similar to Simple Kriging we have used in [A3], Ordinary Kriging also starts with a spatial ***stochastic process*** (or ***random field***), $\{Y(s) = \mu + \varepsilon(s) : s \in R\}$, where μ and $\varepsilon(s)$ still refer to a constant mean and stochastic residual term respectively. The given data set of 286 observations is again assumed to be a realization (finite sample) of this process and multi-normally distributed with ***known covariances***, $\mathrm{cov}[\varepsilon(s_i), \varepsilon(s_j)]$. Hence, in order to make prediction about Manganese concentrations, $Y(s_0)$, at any location s_0 based on the relevant prediction set, $S(s_0) = \{s_1, \ldots, s_{n_0}\}$, that has been identified within this set of sample locations and corresponding observed Manganese values, $\{Y(s_1), \ldots, Y(s_{n_0})\}$, the remaining task is to seek a best linear unbiased (BLU) predictor, $\hat{Y}(s_0) = \sum_{i=1}^{n_0} \lambda_{i0} Y(s_i)$ of $Y(s_0)$.

Step 1. Estimation of Covariances

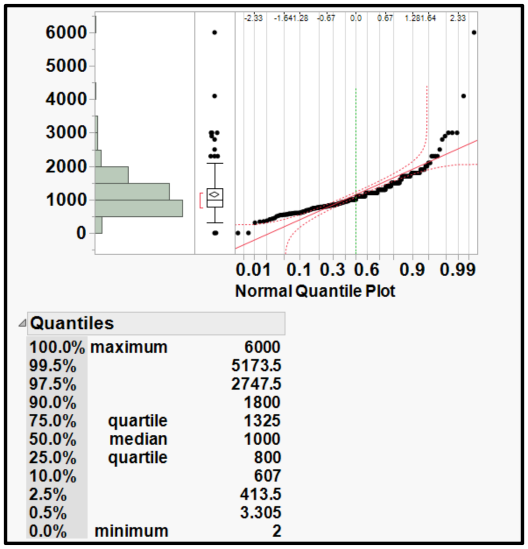
The first step of applying Ordinary Kriging is to estimate the covariances through variogram and hence the derived covariogram, where we yet focus on *Y*-process rather than the ε-process in Simple Kriging. Here, since the Ordinary Kriging also assumes a constant mean, these two processes are identical. Recall that direct estimation of the covariogram is ***biased***, we again start with variogram estimation, which is designated as the ***expected squared difference*** between a pairwise points separated by distance at *h*. In practice, we use an aggregated distance and respective average variogram value to represent similar distances in each ***bin*** (i.e., distance interval), which can be estimated from the sample data. Hence, such representative estimations named ***empirical variogram*** will be fitted using a parametric function that minimizes the squared differences of estimations and actual empirical variograms. In this study, we again use the widely used spherical models to estimate variogram as for each any pair of point separated by distance *h*:

$$(2.1.1) \quad \hat{\gamma}(h; \hat{r}, \hat{s}, \hat{a}) = \begin{cases} 0 & , \quad h = 0 \\ \hat{a} + (\hat{s} - \hat{a})(\frac{3h}{2\hat{r}} - \frac{h^3}{2\hat{r}^3}) & , \quad 0 < h \le \hat{r} \\ \hat{s} & , \quad h > \hat{r} \end{cases} ,$$

and hence the estimation of derived covariogram is given by

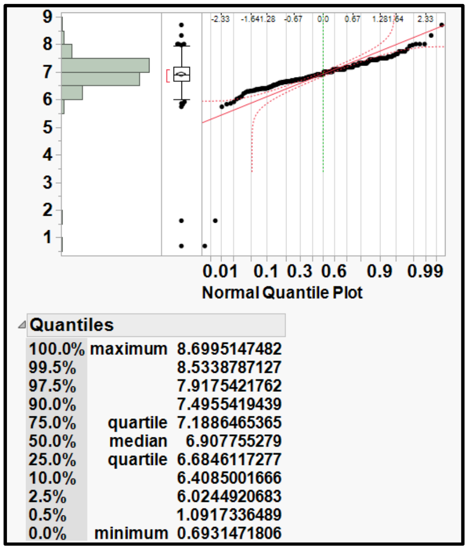
$$(2.1.2) \qquad C\left(h;\hat{r},\hat{s},\hat{a}\right)=\begin{cases}\hat{s} & , \quad h=0\\ (\hat{s}-\hat{a})(1-\dfrac{3h}{2\hat{r}}+\dfrac{h^3}{2\hat{r}^3}) & , \quad 0<h\le\hat{r}\\ 0 & , \quad h>\hat{r}\end{cases}$$

Before applying the spherical models to our Manganese data set, we start with observing its frequency histogram as shown in Figure 3 below. It's natural to find out that the distribution of original Manganese values is ***not very normal*** but ***positive-skew*** with a long tail in the positive direction on the number line (i.e., at larger values), because the observed Manganese data is ***nonnegative*** (i.e., truncated at zero). Note that it is difficult to krige this data directly because the normality assumption is violated that would yield corresponding prediction intervals with ***little validity***. To solve this problem, this data is transformed to ***natural logs*** as shown in Figure 4, which appears to be more "bell-shaped" (normal) than before.



```
Quantiles
100.0% maximum        6000
99.5%               5173.5
97.5%               2747.5
90.0%                 1800
75.0%    quartile     1325
50.0%    median       1000
25.0%    quartile      800
10.0%                  607
2.5%                 413.5
0.5%                 3.305
0.0%     minimum         2
```

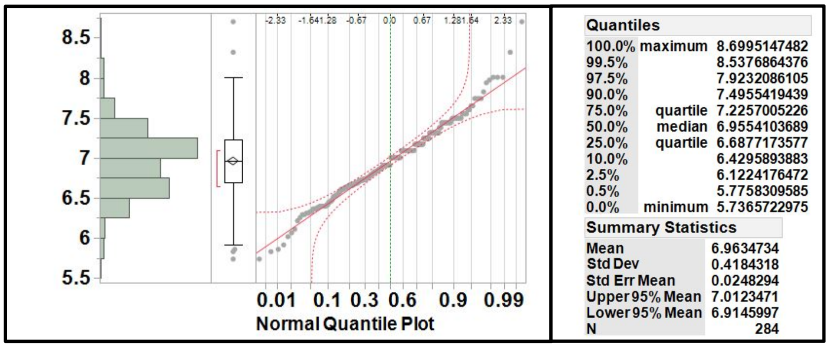
Figure 3. Original Manganese Data



```
Quantiles
100.0% maximum  8.6995147482
99.5%           8.5338787127
97.5%           7.9175421762
90.0%           7.4955419439
75.0%  quartile 7.1886465365
50.0%  median    6.907755279
25.0%  quartile 6.6846117277
10.0%           6.4085001666
2.5%            6.0244920683
0.5%            1.0917336489
0.0%   minimum  0.6931471806
```

Figure 4. Log-Manganese Data

However, there still exists a problem as the two ***outliers*** at the left lower corner that are far away from the "clustering" dot-curve in the Normal Quantile Plot, which refer to the relatively smaller values at 0.6931 and 1.0917 shown in the Quantiles table. Since the least-squares procedure is sensitive to outliers, we will remove them from the data set. Then the normal quantile plot of the new sample data is shown in Figure 5 below.



Quantiles

100.0%	maximum	8.6995147482
99.5%		8.5376864376
97.5%		7.9232086105
90.0%		7.4955419439
75.0%	quartile	7.2257005226
50.0%	median	6.9554103689
25.0%	quartile	6.6877173577
10.0%		6.4295893883
2.5%		6.1224176472
0.5%		5.7758309585
0.0%	minimum	5.7365722975

Summary Statistics

Mean	6.9634734
Std Dev	0.4184318
Std Err Mean	0.0248294
Upper 95% Mean	7.0123471
Lower 95% Mean	6.9145997
N	284

Figure 5. Log-Manganese Data (Removing Outliers)

Given the spherical models and modified data set, we try to construct emprical variogarm and hence the derived covariogram in MATLAB using two different maximum lag distances: one is the default half of maximum pairwise distance = 37,559 meters; the other is a smoewhat larger value = 46,000 meters. The covariomgram plots we are concerned and corresponding parameters (r,s,a) used in spherical models are shown below:

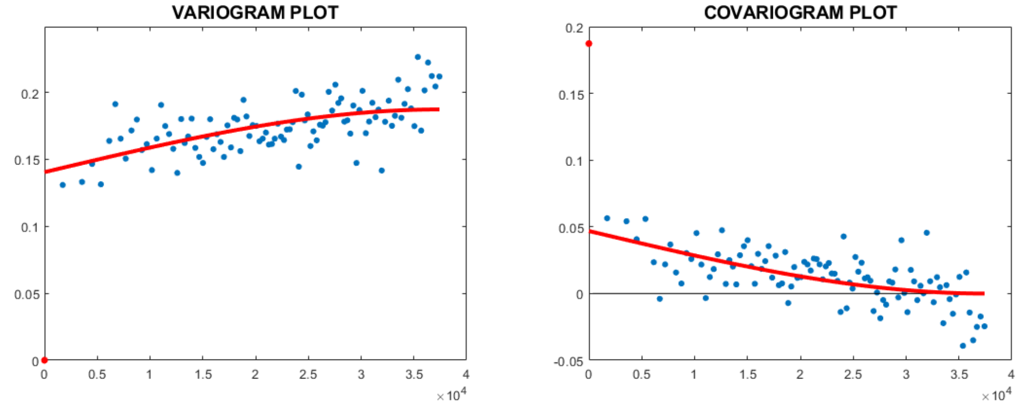


Figure 6. Empirical Variogram Construction (with 37,559-Meter Maximum Lag Distance)

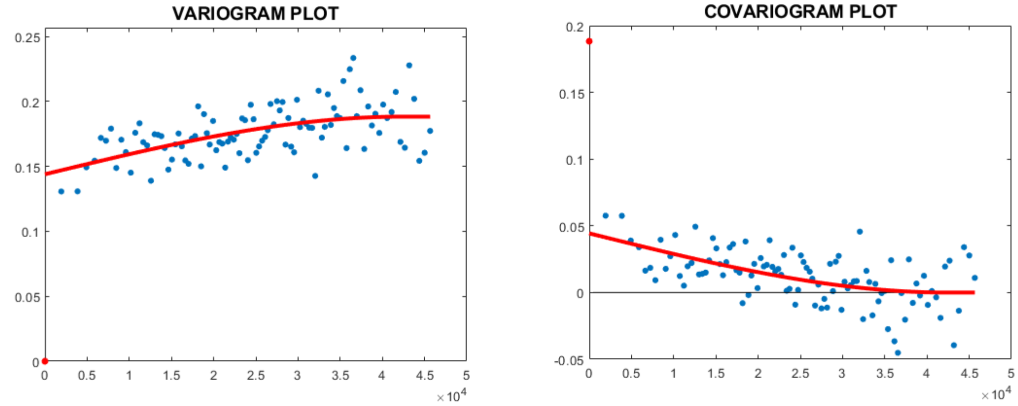


Figure 7. Empirical Variogram Construction (with 46,000-Meter Maximum Lag Distance)

Here, parameters (*r*,*s*,*a*) in spherical models for different maximum lag distances are respectively denoted as RANGE, SILL and NUGGET in Table 1 below.

Table 1. Parameters Comparison of Empirical Variogram

MAX-DIST	RANGE	SILL	NUGGET
37559	37415	0.187	0.140
46000	42331	0.188	0.144

In particular, the RANGE = 42331 meters denotes the distance at which variogram "reaches the sill" and starts to level off. It's more intuitive to interpret this in the form of the derived covariogram that beyond this distance there is estimated to be no statistical correlation between Manganese values. Turning to the other estimated parameters, the SILL = 0.188 represents the estimated variance of individual Manganese values (i.e., the estimated covariance at "zero distance"). Additionally, the NUGGET = 0.144 denotes the magnitude of spatial independence. Hence, in this case, the ***relative nugget effect*** here is calculated as 0.144 / 0.188 = 0.766. Such a well high value indicates the existence of very weak ***spatial dependence*** among Manganese values.

Given the estimated covariogram, $\hat{C}(h)$ above, now we can estimate covariances for each pairwise points, $s_i, s_j \in R$, separated by distance $h = \|s_i - s_j\|$ in the form of *Y* values rather than *ε* values, which distinguishes a difference from Simple Kriging:

$$(2.1.3) \qquad \hat{\sigma}_{ij} = \text{cov}[Y(s_i), Y(s_j)] = \hat{C}(\|s_i - s_j\|)$$

The quantity in expression (2.1.3) then yields the estimate of the full-sample covariance matrix, $V = \text{cov}(Y_n)$:

$$(2.1.4) \qquad \hat{V} = \begin{pmatrix} \hat{\sigma}^2 & \cdots & \hat{\sigma}_{1n} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{n1} & \cdots & \hat{\sigma}^2 \end{pmatrix}$$

Similar to (2.1.3), then we can also obtain estimates for all covariances between the predicted values Y_0 at s_0 based on the given prediction set from sample data, $\{Y(s_1), \ldots, Y(s_{n_0})\}$:

$$(2.1.5) \qquad \hat{\sigma}_{0j} = \text{cov}[Y(s_0), Y(s_j)] = \hat{C}(\|s_0 - s_j\|) \quad , \quad j = 1, \ldots, n_0$$

Finally, the full covariance matrix, $\hat{C}_0$, relevant for prediction at s_0 is obtained as:

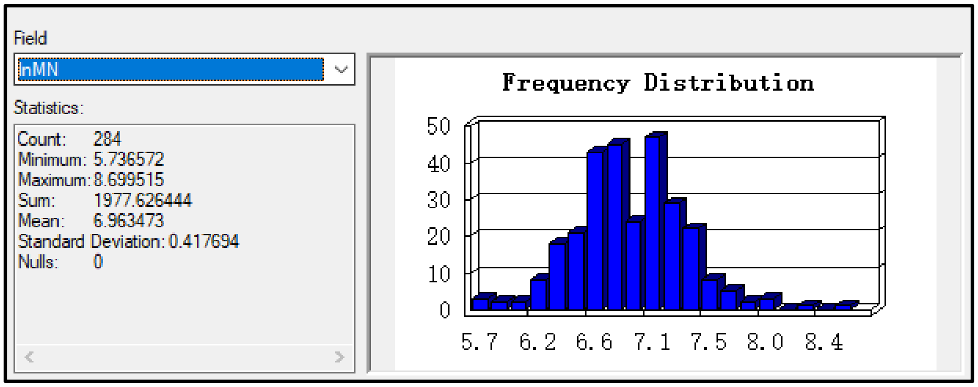
$$(2.1.6)\qquad \hat{C}_0 = \begin{pmatrix} \hat{\sigma}^2 & \hat{c}_0' \\ \hat{c}_0 & \hat{V}_0 \end{pmatrix} \quad ,$$

where $\hat{\sigma}^2$ is the estimation of common covariance among Manganese values, and $\hat{c}_0$ is again the column vector of estimated covariances of Y_0 (not ε_0) at s_0 with each predictor variable, $\{Y(s_1),\ldots,Y(s_{n_0})\}$, and is denoted by:

$$(2.1.7)\qquad \hat{c}_0 = \begin{pmatrix} \hat{\sigma}_{01} \\ \vdots \\ \hat{\sigma}_{0n_0} \end{pmatrix} \quad ,$$

which can be calculated from the empirical covariogram.

Step 2. Estimation of the Mean



Field: lnMN

Statistics:

Count: 284
Minimum: 5.736572
Maximum: 8.699515
Sum: 1977.626444
Mean: 6.963473
Standard Deviation: 0.417694
Nulls: 0

Figure 8. Statistics of Log-Manganese Data

This step involves the main departure from Simple Kriging. Recall that the only difference between Ordinary Kriging model and Simple Kriging is that the constant mean, μ, is not assumed to be known now. Instead of using the sample-mean estimator, $\overline{Y}_n$ (i.e., adding up all observed values in the sample data set and averaging them, denoted as Mean in Figure 8 above), we use the BLU estimator, $\hat{\mu}_n$, which hence must be estimated within the model as:

$$(2.2.1)\qquad \hat{\mu}_n = \frac{1_n' \hat{V}^{-1} y}{1_n' \hat{V}^{-1} 1_n} \quad ,$$

where $y = (y_1,\ldots,y_n)'$ is the sample data vector, and 1_n refers to the unit vector of length n, $1_n = (1,\ldots,1)'$.

Step 3. Estimation of Kriging Predictions

The final step to evaluate Ordinary Kriging predictions here is identical to that in the Simple Kriging procedure, except for replacing the sample-mean estimate, $\hat{\mu}$, with the BLU estimate, $\hat{\mu}_n$. Hence, we can use the BLU estimator of mean, $\hat{\mu}_n$ (from Step 2), and full covariance matrix, $\hat{C}_0$ (from Step 1), to predict Manganese deposit at any location in the study region by the following formula:

(2.3.1) $\hat{Y}_0 = \hat{Y}(s_0) = \hat{\mu}_n + c_0'\hat{V}_0^{-1}\hat{\varepsilon}$

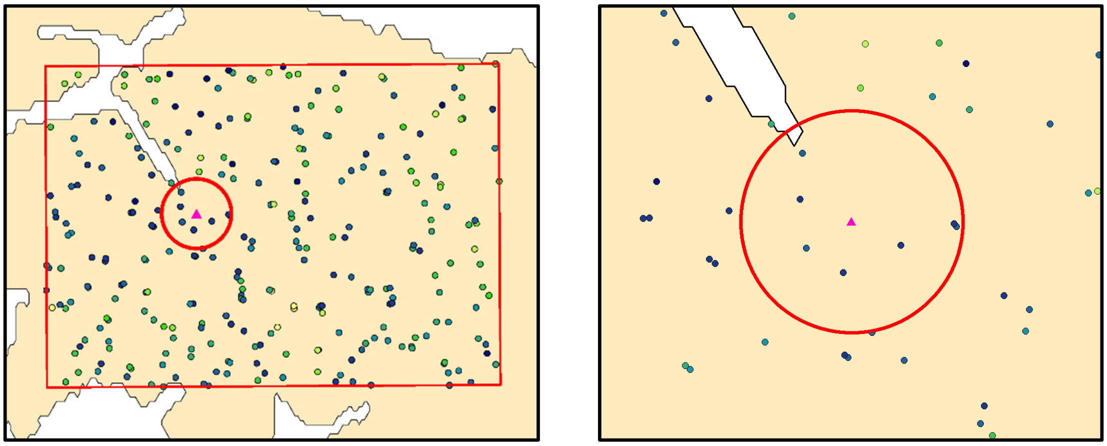


Figure 9. Point *L* and Its Prediction Set

In particular, it's easy to use MATLAB to apply a bandwidth of 4,900 meters to predict the Manganese value, $\hat{Y}_L$, and standard error, $\hat{\sigma}_L$, at a specific point location, L = (612300,579700), marked as a purple triangle in Figure 9 above. It's observed that 7 neighborhood points are included in the circle of 4,900-meter bandwidth as the relevant prediction set, $\{S(s_{Lj}) : j = 1,...,7\}$, and their corresponding values will be used to make this prediction. Here, if we denote the desired vector of prediction weights of these 7 predictors by $\lambda_L = (\lambda_{L1},...,\lambda_{L7})'$, a ***linear combination*** of these predictors' Manganese measurements, denoted as a column vector, $Y = (Y_{L1},...,Y_{L7})'$, can be written as:

(2.3.2) $\hat{Y}_L = \lambda_L Y$,

which is exactly the BLU estimator at this point. Then, the best solution to minimize residual ***mean squared error***:

(2.3.3) $E[(Y_L - \hat{Y}_L)^2] = E[(Y_L - \lambda_L{}'Y)^2] = E[(\varepsilon_L - \lambda_L{}'\varepsilon)^2] = MSE(\lambda_L)$,

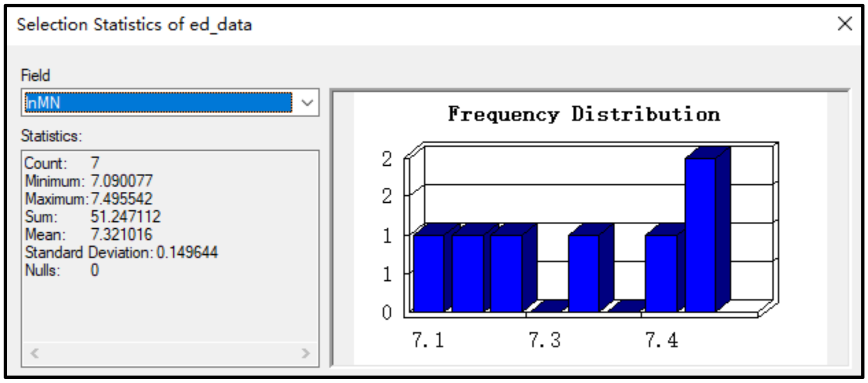
would lead to expression (2.3.1) mentioned above (more specific and detailed procedure to solve the problem can be seen in section 6.3 of Notebook). The results of Ordinary Kriging at this point are displayed in Table 2 below:

Table 2. Prediction at Point *L*

MEAN	PRIDICTION	SD
6.9111	7.186	0.4104

Notice that the BLU mean, $\hat{\mu}_n$, denoted as MEAN = 6.9111 in the Table is different from the rounded sample-mean = 6.9635 above. Here, Ordinary Kriging value at *L* = (612300,579700) is designated as PRIDICTION = 7.186, together with standard error, denoted as SD = 0.41035. Again, the next process is to determine a default (common) 95% prediction interval. Recall that the data set we use here is log-transformed to obtain a better normal approximation for statistically valid prediction intervals, so it's only meaningful to interpret these results if we transform them back to original numbers by exponentiating. Hence, the 95% prediction interval for the true Manganese value at *L* is given by [exp(7.186 - 1.96 × 0.4104), exp(7.186 + 1.96 × 0.4104)] = [590.88, 2952.43]. In a word, compared to the estimated global mean Manganese value = exp(6.9111) = 1003.3503 ppm, the predicted value at this point is exp(7.186) = 1320.8094 ppm, and we are 95% confident that the value would be larger than 590.88 ppm and smaller than 2951.43 ppm.

To examine whether this prediction at *L* is reasonable, a quick checking way is to see the distribution of these 7 neighbors of prediction set, along with the mean and standard deviation, as shown in Figure 10 below:



Selection Statistics of ed_data

Field

lnMN

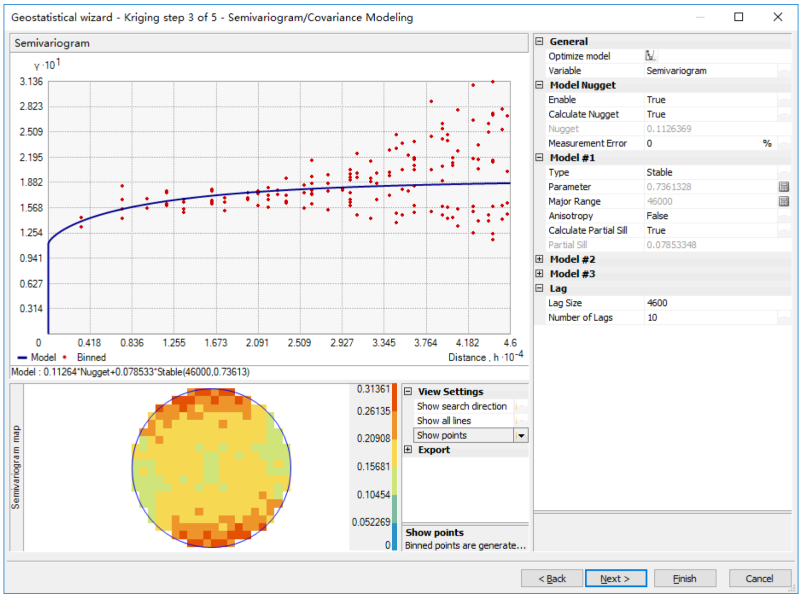
Statistics:

```
Count:     7
Minimum:   7.090077
Maximum:   7.495542
Sum:       51.247112
Mean:      7.321016
Standard Deviation: 0.149644
Nulls:     0
```

Figure 10. Statistics of Prediction Set

It’s observed that the mean Manganese of these 7 points is 7.321, which is very close to the predicted kriging value = 7.186, while the standard deviation designated as 0.1496 is well smaller than the associated value = 0.4104 given by Ordinary Kriging.

Finally, as we have done in [A3], for purpose of comparison, we will repeat Ordinary Kriging analysis for Manganese data in ArcMap using GA extension. Here, we set Lag Size = 4,600 and Lags = 10 that would yield the same maximum lag distance of 46,000 meters as used in MATLAB, as shown in Figure 11 below:

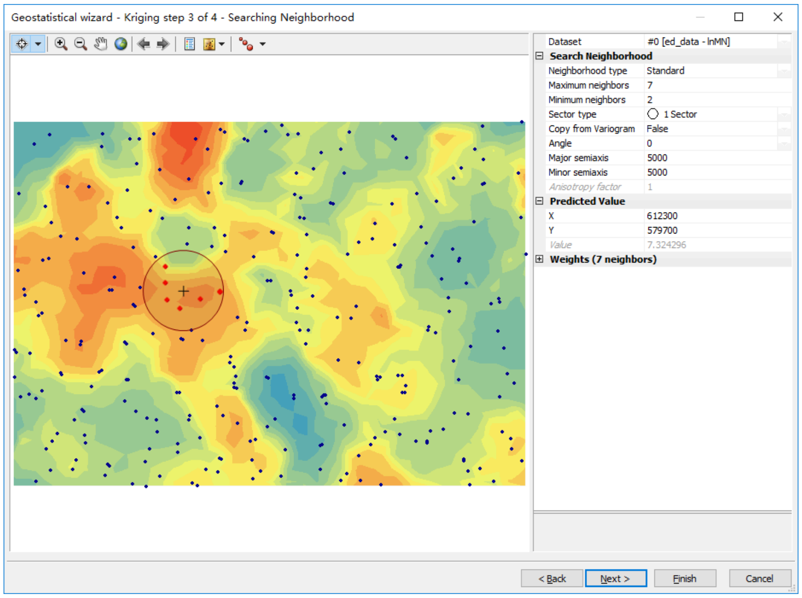


Model : 0.11264*Nugget+0.078533*Stable(46000,0.73613)

General	
Optimize model	
Variable	Semivariogram
Model Nugget	
Enable	True
Calculate Nugget	True
Nugget	0.1126369
Measurement Error	0 %
Model #1	
Type	Stable
Parameter	0.7361328
Major Range	46000
Anisotropy	False
Calculate Partial Sill	True
Partial Sill	0.07853348
Model #2	
Model #3	
Lag	
Lag Size	4600
Number of Lags	10

Figure 11. Variogram for Manganese Data in GA

Then by setting Maximum Neighbors = 7, bandwidth = 5,000 meters for Major and Minor semiaxes and giving the coordinate of L = (612300,579700), the predicted Manganese at this point is evaluated as 7.324, as shown in Figure 12 below.


Geostatistical wizard - Kriging step 3 of 4 - Searching Neighborhood
< Back
Next >
Finish
Cancel

Dataset	#0 [ed_data - lnMN]
⊟ **Search Neighborhood**	
Neighborhood type	Standard
Maximum neighbors	7
Minimum neighbors	2
Sector type	○ 1 Sector
Copy from Variogram	False
Angle	0
Major semiaxis	5000
Minor semiaxis	5000
Anisotropy factor	1
⊟ **Predicted Value**	
X	612300
Y	579700
Value	7.324296
⊞ **Weights (7 neighbors)**	

Figure 12. Kriging Prediction at Point *L* = (612300,579700)

Thus, a contour prediction map based on the Ordinary Kriging procedure can be created, shown in Figure 13. By examining Krige predictions around the point *L* and averaging them, an approximate estimate of 7.258 is given. This approximation and specific Kriging value = 7.324 estimated in GA above both agree with MATLAB-based prediction = 7.186.

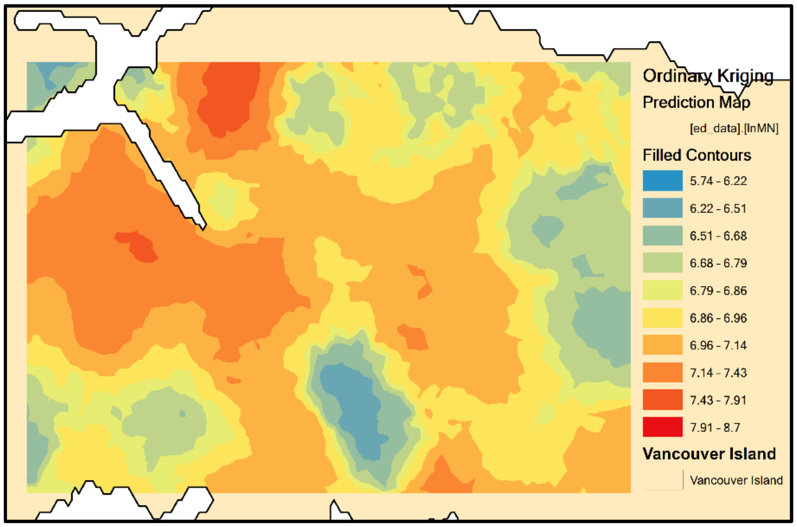
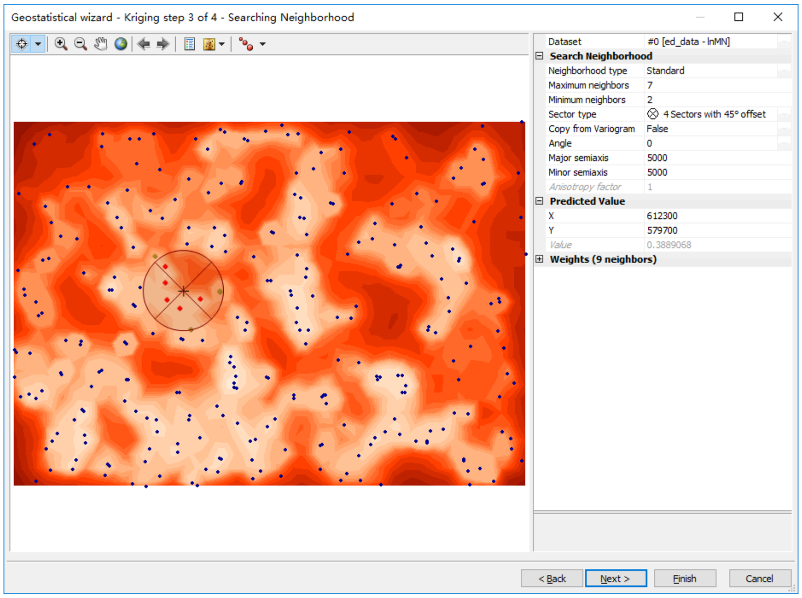


Figure 13. Oridinary Kriging Predictions

Next we examine the associated standard error of Kriging prediction at location *L*. The GA yields a value of 0.3889, which is also almost identical to that in MATLAB.


Geostatistical wizard - Kriging step 3 of 4 - Searching Neighborhood

Dataset	#0 [ed_data - lnMN]
⊟ **Search Neighborhood**	
Neighborhood type	Standard
Maximum neighbors	7
Minimum neighbors	2
Sector type	⊗ 4 Sectors with 45° offset
Copy from Variogram	False
Angle	0
Major semiaxis	5000
Minor semiaxis	5000
Anisotropy factor	1
⊟ **Predicted Value**	
X	612300
Y	579700
Value	0.3889068
⊞ **Weights (9 neighbors)**	

Figure 14. Standard Errors of Simple Kriging

3 DISCUSSION

In this study, we have extended another more usual Kriging model named Ordinary Kriging that assumes an unknown constant mean. Based on the Manganese value data, we have demonstrated the powerful utility of Ordinary Kriging in predicting unknown values and associated confidence intervals at any location without exact observations there but sample data of other observations. For the predicted Kriging value at the specific location $L = (612300,579700)$, since MATLAB and GA use different mathematical methods to krige the sample data, it' s natural to obtain slightly different results correspondingly. But the mild deviation lies within the allowable range, indicating that the predictions are both reasonable.

4 REFERENCE

[A3] J.Wu (2019), "Assignment 3: Cobalt Study for ESE 502"