Enhancing Spatial Predictions: An Exploration of Cokriging and Semivariogram Models

Noel Howley - Institute of Mathematical Innovation

August 2023

1 Introduction

Cokriging is an advanced geostatistical interpolation technique that uses multiple correlated variables to make predictions. It operates on the principle that two or more variables sampled at the same locations can provide more information than just one variable alone.

At the heart of cokriging is the relationship between the primary variable (the one we're interested in predicting) and the secondary variable(s) (additional variables that provide auxiliary information). This relationship is quantified using cross-variograms, which measure how the difference in one variable changes as the difference in another variable changes across space.

In the following file we cover the work completed in the associated jupyter notebook where different variogram models have affected the interpolation predictions of cokriging interpolation methods.

As mentioned before, cokriging plays a pivotal role in soil moisture prediction, offering a more refined and accurate approach by integrating multiple sources of correlated spatial data. Soil moisture is a critical parameter in climate studies, and its precise estimation can lead to better drought forecasting or potentially life threatening landslides and floods. Furthermore, when applied to threat detection, cokriging can enhance the identification of anomalous patterns or regions in the soil, such as contaminants or unexplored ordnance. By leveraging correlations between primary data (like soil moisture) and secondary data sets (such as conductivity), cokriging provides a more comprehensive and reliable picture, ensuring timely interventions and informed decision-making in threat mitigation.

1.1 Defining Our Problem

As mentioned, we seek to perform cokriging interpolation methods on multiple variables relating to soil moisture, which in this case is our Primary variable.

To start with we shall be working with the understanding that our secondary data is hydraulic conductivity.

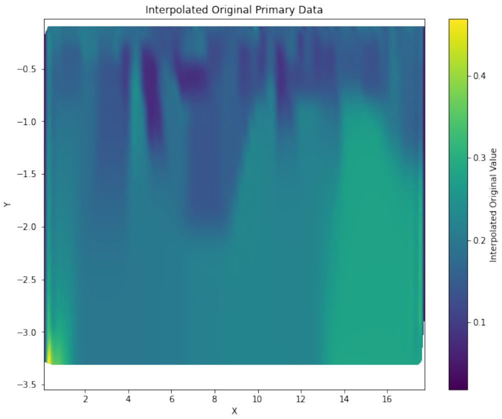

Figure 1: Primary Data Before CoKriging

In the context of our problem we can say our Primary and Secondary data represent $Z_1(\mathbf{u})$ and $Z_2(\mathbf{u})$ respectfully in (1). Without further inspection this implies that $Z(\mathbf{u})$ is our cokriging solution we seek to find, and λ_{1i} $\forall i$ and λ_{2j} $\forall j$ are the weights for the particular data set based off the variogram model.

$$Z(\mathbf{u}) = \sum_{i=1}^{n_1} \lambda_{1i} Z_1(\mathbf{u}_i) + \sum_{j=1}^{n_2} \lambda_{2j} Z_2(\mathbf{u}_j) \tag{1}$$

Traditionally, a variogram is a fundamental tool in geostatistics, quantifying the spatial variability and correlation structure of a spatial variable. Mathematically, the semi-variogram, $\gamma(h)$, is defined as half the expected squared difference between values separated by a specific lag distance h:

$$\gamma(h) = \frac{1}{2} E\left[\left(Z(\mathbf{u}+h) - Z(\mathbf{u})\right)^2 \right] \tag{2}$$

Where $\gamma(h)$ is the semi-variogram at lag distance h, and E is the expectation operator. Due to the quantity of Secondary Data, we cannot calculate the cross-variogram between the data sets but we can approximate them with distributions.

Lastly, we should consider the Primary data before cokriging interpolation, see Figure 1.

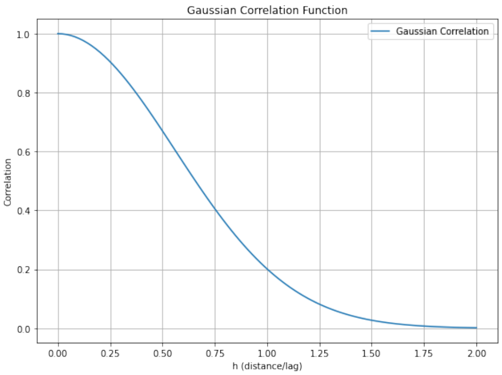

Figure 2: Gaussian Correlation Distribution with $\theta = 0.79$

1.2 Assumtions

1.2.1 Gaussian

The Cross-Variogram may take many distributions and hence we shall be using multiple distributions to approximate this.

The Gaussian distribution, when applied to semivariograms, is especially suited for datasets that display continuous and smooth spatial variations. This model captures long-range dependencies effectively, reflecting a smoother representation of spatial continuity. The bell-shaped curve of the Gaussian distribution aligns well with many empirical semivariograms, making it an excellent choice for datasets where spatial correlations are evident over extended distances. Its ability to represent persistent spatial relationships makes the Gaussian model a valuable tool in geostatistical analyses. The Gaussian correlation distribution is defined as follows:

$$\gamma(h) = e^{\frac{-h^2}{\theta^2}}$$

From Figure 3 we can spot a similarity with Figure 1, which can be examined further by looking at the differences between them with Figure 4. With all the following distributions we will be considering a particular value of the range parameter, θ. This parameter can be seen to cause very large difference in the predictive performance of the cokriging. In the case of the Gaussian distribution a small enough value of θ gives a fairly sensible looking result but should the value of θ be too large the cokriging is nonsense. This therefore implies that the distribution of soil moisture may not be spatially dependent over extended distances and we should be seek variograms models such that the semivariogram supports stronger weights for shorter distances (smaller h) and smaller weights for greater distances (larger h), or tweak the value of θ accordingly.

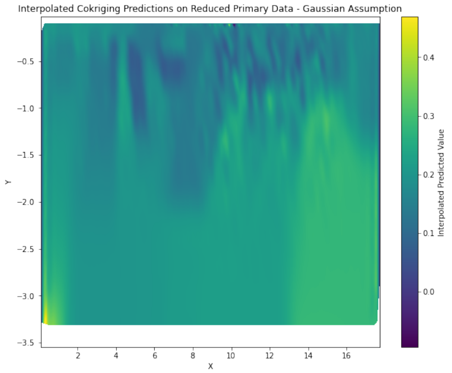

Figure 3: Cokriging with Gaussian assumption

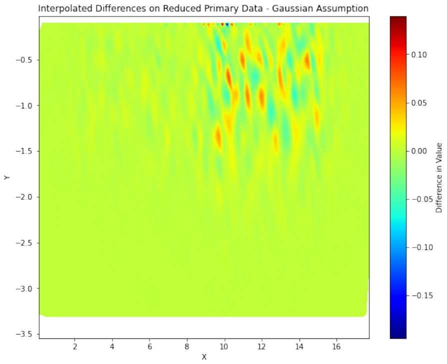

Figure 4: Differnce between the Cokriging Prediction

1.2.2 Spherical

Spherical models are widely employed in semivariogram modeling due to their intuitive and practical characteristics. A primary advantage of the spherical model is that it reaches a sill (a plateau or maximum variance) at a certain range, implying that beyond this range, there is no longer a spatial correlation between observations. This behavior mirrors many natural processes, where the influence of one observation on another diminishes with increasing distance until a threshold is reached, after which they are effectively uncorrelated. Additionally, spherical models are mathematically tractable, ensuring straightforward computation and implementation in geostatistical analyses. Their flexible shape allows for a good fit to a variety of empirical semivariograms, making them a versatile choice for diverse datasets. The Model is defined as follows:

$$\gamma(h) = \frac{3h}{2\theta - \frac{h^3}{2\theta^3}} \tag{3}$$

Unsurprisingly a much more respectable result can be seen in Figures 6 and 7, with errors accumulating around the center of the interpolated field and of magnitude 10^{-15}, this assumption has shown to be much closer to the true value as seen in Figure 1. An optimal value of θ should be considered but the magnitude of the error for many values of $\theta > 1$ have been shown to produced very similar results.

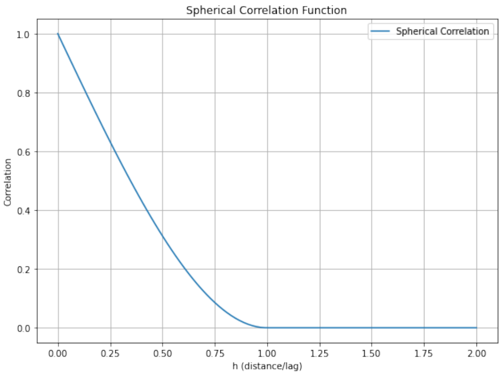

Figure 5: Spherical Distribution with $\theta = 1$

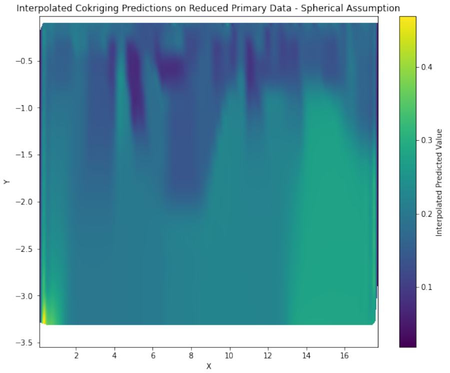

Figure 6: Cokriging with Spherical assumption

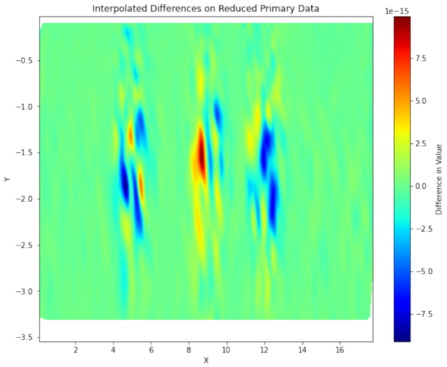

Figure 7: Difference between the Cokriging Prediction and Primary Data

1.2.3 Log-Normal

$$\gamma(h) = \frac{-h^2}{\theta^2} \tag{4}$$

The LogNormal model, when applied to semivariograms, addresses data sets where values are positively skewed and show multiplicative effects. One of the key characteristics of LogNormal distribution is that the logarithm of the values follows a normal distribution. This transformation can stabilize the variance and make the data more amenable to geostatistical techniques. When applied to semivariograms, the LogNormal model can capture and represent the exponential increase in variance often observed in natural phenomena, such as concentrations of minerals or pollutants. The model's ability to handle skewed data and its fit to many empirical semivariograms make it a valuable tool in geostatistical modeling, especially when the data exhibits multiplicative behavior or when dealing with variables, that cannot take negative values, such as

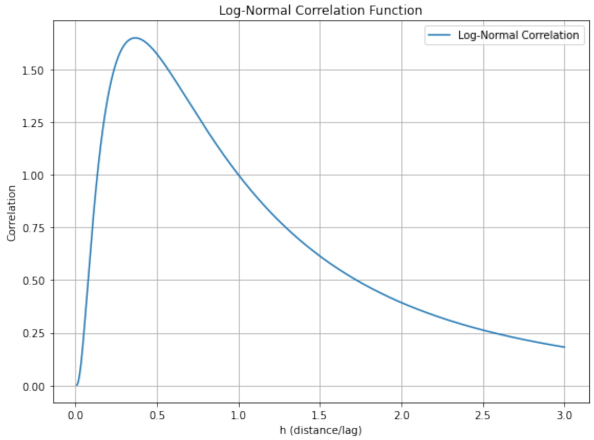

Figure 8: Log-Normal Distribution

concentrations or resistivity.

As can be seen in Figure 9 for the cokriging interpolation of the LogNormal assumption and the error in the cokriging assumption shown in 10, we can see that the model once again has errors of magnitude 10^{-14} implying that the cokriging approximation is close to the Primary Data.

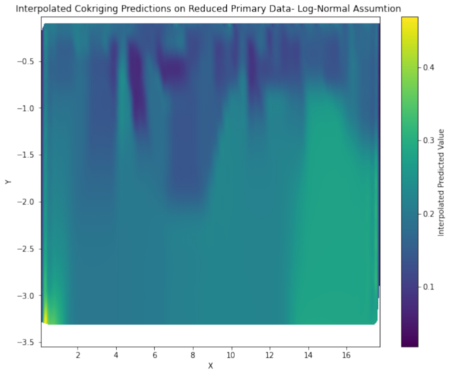

Figure 9: Cokriging LogNormal

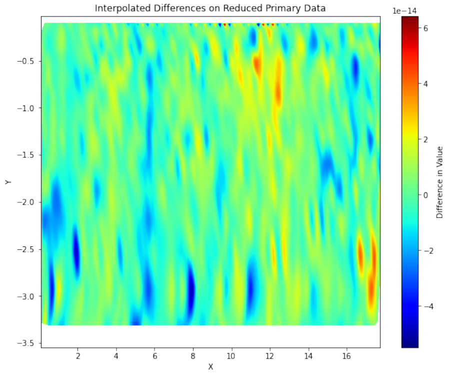

Figure 10: Difference between the Cokriging Prediction and Primary Data

1.3 Summary

During this paper, we have not come to any conclusion about which model is more suitable for the semivariogram, however we have observed how the different assumptions about the semivariogram changes the result of the cokriging interpolation results. Without knowing the true distribution of the semivariogram and the range parameter θ we will not be able to come to a definitive answer on which model is the most suitable in use of cokriging predictions. Most of the distributions would produce predictions that are very similar to the original data if the semivariances include larger weighting for when h is small and small weighting for when h is large, which is unsurprising that we see the predictions looking like the Primary data as in Figure 1. More generally we may say that as the distribution of the semivariance model tends to the Kronecker delta function we will see the predictions tend to the primary data. Unless the optimal semivariogram model is found, assessing other Secondary Datasets may prove to be similarly inconclusive as the results may be biased to the Primary Data as with the results shown in this review.