A Comparison of *Z* Variograms Obtained by Transformation Using Hermite Polynomials and Monte Carlo Simulation

Brandon J. Wilde, Chad Neufeld and Clayton V. Deutsch

Centre for Computational Geostatistics
Department of Civil & Environmental Engineering
University of Alberta

A previous paper introduced the concept of using Monte Carlo Simulation to 'back-transform' a normal scores variogram to a true values variogram. This note compares the results obtained using this method with those obtained using Hermite polynomials. It is shown that the transformation results are the same.

Introduction

It is well established that variograms calculated from untransformed data can be unstable and erratic due to the nature of the data distribution. Outliers in the data are the main cause of this instability. Transforming the data to a standard normal distribution substantially reduces the effect of these outliers. Variograms calculated from this transformed data are therefore much more consistent and stable. Two methods have been devised for transforming the variogram calculated from standard normal data to a variogram that is in the units of the original data. One idea is to use Monte Carlo Simulation (MCS) to sample a bivariate distribution with its correlation taken from the normal scores variogram (Wilde and Deutsch, 2005). This procedure is repeated at each lag. The other method is to use Hermite Polynomials as suggested by John Vann and Henry Sans (Vann, 2005). Each is remarkably robust and provides a consistent and stable original data variogram. The details of how each transform is performed are given. The purpose of this paper is to compare the results of each transform method and to determine if one is more applicable than the other.

Methods

Common to both methods is the need for an experimentally calculated normal scores variogram. To obtain this variogram, the data must be transformed to a standard normal distribution. It may be necessary to weight the data by declustering prior to performing this transformation. If declustering weights are used in the transformation, the variance of the data must also be calculated by using these weights. The variance is an important input parameter for the MCS method. The transformation table is an important input for both methods and must therefore be saved for future use. Once the data has been transformed, the experimental variogram can be calculated. This variogram will have a sill of 1.0 as that is equivalent to the variance for a standard normal distribution. It should be a good deal more stable and consistent than the variogram calculated from the normal scores. The variogram transformations can now be performed using the normal scores variogram and the transformation table.

MCS/Mapping

The MCS method has been modified somewhat compared to the original idea presented in Wilde and Deutsch, 2005. It was seen that the step of modeling the normal variogram was not necessary and thus, this time consuming step was removed from the process. The process of modeling is now reserved solely for the original data variogram.

The MCS method is performed one lag at a time. Therefore, the following procedure is repeated for each lag. The first step is to determine the correlation coefficient at that lag, **h.** The relation between the variogram and the correlation coefficient is as follows:

$$\rho = 1 - \gamma_y$$

where γ_y is the normal scores variogram value at h. The correlation coefficient is used to generate a bivariate Gaussian distribution characterized by ρ. This is done by generating a large number of pairs of random numbers in [0,1]. These pairs are then transformed into standard normal space as Y_1 and Y_2. They are then given the appropriate correlation by the following equation:

$$Y_h = \rho \cdot Y_1 + \sqrt{1-\rho^2} \cdot Y_2$$

The new correctly correlated pair of Y_1 and Y_h are then back transformed into units of the original data as Z_1 and Z_2 according to the initial normal scores transformation performed on the original variable. The squared difference of Z1 and Z2 is calculated for each of the paired samples. The average of these squared differences is taken to be the value of the original variable variogram for the specific **h**. This process is then repeated for each **h.** The variogram can be normalized by dividing by the variance of the original data. This variogram can then be fit and used for modeling and estimation purposes. It is unbiased and much more stable than the variogram calculated directly from the original data.

The original mapping method was identical in all aspects to this except the derivation of ρ. The variogram value used in obtaining the correlation was taken from an explicit model of the variogram. This made the mapped original data variogram much smoother. It also allowed any number of lags to be mapped at any value of **h**. But it also meant that the time consuming modeling process had to be repeated twice for any original data variogram: both the normal scores and the original data variograms had to be fit. This could become extremely time consuming and tedious where there are numerous original data variograms needed. Eliminating the step of modeling the normal scores variogram has effectively cut in half the time needed to arrive at a variogram model for the original data.

Hermite Polynomials

The method using Hermite polynomials employs an anamorphosis function to calculate the transformed variogram. The anamorphosis is a function defined by a polynomial expansion that is fit to the data. Once the polynomials have been fit, the function provides a mapping of the point variable *Z* to the Gaussian variable *Y* and vice versa:

$$\begin{aligned} z(\mathbf{u}) &= \Phi(y(\mathbf{u})) \\ &\approx \sum_{p=0}^{\infty} \phi_p H_p[y(\mathbf{u})] \end{aligned}$$

where ϕ_p is a fitted coefficient for each term of the polynomial expansion, and $H_p[y(\mathbf{u})]$ is the Hermite polynomial value defined by the term of the expansion and the *y* value. The previous equation is referred to as the Gaussian anamorphosis. The number of terms in the polynomial expansion is usually limited to a number less than 100. The more terms used, the better the polynomial fitting will be, but the programs will run slower.

Hermite polynomials are a family of orthogonal polynomials that are related to the normal distribution. They are defined by Rodrigues' Formula:

$$H_p(y) = \frac{1}{\sqrt{p!} \cdot g(y)} \cdot \frac{d^p g(y)}{dy^p}$$

where *g(y)* is the probability of the value *y* for a standard normal distribution. The first two Hermite polynomials, order 0 and 1, are given by:

$$H_0(y) = 1 \qquad H_1(y) = -y$$

Higher order polynomials can be calculated with the following recursive formula when $p \geq 2$:

$$H_{p+1}(y) = -\frac{1}{\sqrt{p+1}} y H_p(y) - \sqrt{\frac{p}{p+1}} H_{p-1}(y)$$

Hence, it is easy to calculate the Hermite polynomials for a given order *p*, starting from a normal value *y*.

We now need to fit the anamorphosis function to the data by calculating the ϕ coefficients. The first order coefficient is:

$$\begin{aligned}\phi_0 &= E\{\Phi(Y(\mathbf{u}))\} \\ &= E\{Z(\mathbf{u})\}\end{aligned}$$

or the expected value of $Z(\mathbf{u})$. Higher order ϕ coefficients are

$$\begin{aligned}\phi_p &= E\{Z(\mathbf{u}) \cdot H_p(Y(\mathbf{u}))\} \\ &= \int \Phi(y(\mathbf{u})) \cdot H_p(y(\mathbf{u})) \cdot g(y(\mathbf{u})) \cdot dy(\mathbf{u})\end{aligned}$$

This integral can be approximated with the sample data as a finite summation:

$$\phi_p \approx \sum_{\alpha=2}^{n} (z(\mathbf{u}_{\alpha-1}) - z(\mathbf{u}_\alpha)) \cdot \frac{1}{\sqrt{p}} H_{p-1}(y(\mathbf{u}_\varepsilon)) \cdot g(y(\mathbf{u}_\alpha))$$

An example showing how these coefficients are used to fit the data is included herein. Since there is no correlation between the different polynomials, the variance of the polynomial expansion is:

$$\begin{aligned}Var\{\Phi(Y(\mathbf{u}))\} &= Var\{Z(\mathbf{u})\} \\ &= \sum_{p=1}^{\infty}\sum_{q=1}^{\infty} \phi_p \phi_q \operatorname{cov}\{H_p(Y(\mathbf{u})), H_q(Y(\mathbf{u}))\} \\ &= \sum_{p=1}^{\infty} \phi_p^2\end{aligned}$$

where $Var\{Z(\mathbf{u})\}$ is the variance of Z at the point support (Neufeld, 2005).

Once the coefficients are calculated, the variogram transformation can be performed using the relation given by Vann and Sans (Vann, 1995):

$$\gamma_z(h) = \sum_{n=1}^{N} \frac{\psi_n^{\,2}}{n!} (1 - (1 - \gamma_y(h))^n)$$

where $\frac{\psi_n^{\,2}}{n!}$ are the coefficients of the Hermite polynomials which we have chosen to represent as ϕ_p^2. We have also chosen to standardize the variogram to have a sill of 1.0 by dividing by the variance. The standardized original data variogram is calculated as follows:

$$\gamma_z(h) = \frac{\sum_{n=1}^{N} \phi_n^{\,2} (1 - (1 - \gamma_y(h))^n)}{\sum_{n=1}^{N} \phi_n^{\,2}}$$

where γ_y is the normal variogram value and γ_z is the original data variogram value. With transformed values calculated for each value of γ_y, the resulting variogram can be modeled for use in estimation etc. As with the MCS method, this variogram is unbiased and a good deal more stable than that calculated from the original data.

Example of Distribution Fitting with Hermite Polynomials

This section details how Hermite polynomials are used to fit a distribution. It also shows the effect that using only a few polynomials will have. The first step is to calculate the Hermite polynomials. These polynomials are well defined. The first six Hermite polynomials are shown below:

$$H_0(y) = 1 \qquad H_3(y) = -\frac{1}{\sqrt{6}}(y^3 - 3y)$$
$$H_1(y) = -y \qquad H_4(y) = \frac{1}{2\sqrt{6}}(y^4 - 6y^2 + 3)$$
$$H_2(y) = \frac{1}{\sqrt{2}}(y^2 - 1) \qquad H_5(y) = -\frac{1}{2\sqrt{30}}(y^5 - 10y^3 + 15y)$$

The recursive formula previously mentioned gives the higher order polynomials. We now calculate the coefficients using the summation approximation given previously and recalling that $\phi_0 = E\{Z(\mathbf{u})\}$. We calculate ϕ_p and H_p for different values of p to allow us to compute $Z(\mathbf{u})$.

We will use Hermite polynomials to fit the typical log-normal distribution shown in Figure . As the figure shows, the distribution has a mean of 98.625 and ranges from 5.448 to 882.042. For p=0, ϕ_0=98.625 and $H_0(y)$=1. Recall that $Z(\mathbf{u})$ is defined as the sum of the product of ϕ_p and $H_p(y)$ for all p. We will start by using only one value for p, p=0.

$$\begin{aligned} z(\mathbf{u}) &= \phi_0 \bullet H_0(y) \\ &= 98.625 \cdot 1 \\ &= 98.625 \end{aligned}$$

We see that $z(\mathbf{u})$ = 98.625 when just one polynomial is used. Obviously this is a very poor fit as shown in Figure 2a.

We will now use two Hermite polynomials, that is, two values for p: p=0,1. For p=1, ϕ_1=-81.833 and $H_1(y)$= -y. Summing the product of these values with the product for p=0, we obtain the following:

$$\begin{aligned} z(\mathbf{u}) &= \phi_0 \cdot H_0(y) + \phi_1 \cdot H_1(y) \\ &= (98.625 \cdot 1) + (-81.833 \cdot -y(\mathbf{u})) \\ &= 98.625 + 81.833y(\mathbf{u}) \end{aligned}$$

This defines a normal distribution with mean 98.625 and standard deviation 81.833 as shown in Figure 2b. This is also a poor fit so we continue, utilizing more polynomials by increasing p.

We will now use three Hermite polynomials: p=0,1,2. ϕ_2=45.810 and $H_2(y)$ is as defined above. The three polynomials give the following:

$$\begin{aligned} z(\mathbf{u}) &= \phi_0 \cdot H_0(y) + \phi_1 \cdot H_1(y) + \phi_2 \cdot H_2(y) \\ &= (98.625 \cdot 1) + (-81.833 \cdot -y(\mathbf{u})) + (45.810 \cdot \frac{1}{\sqrt{2}}(y^2(\mathbf{u}) - 1)) \\ &= 98.625 + 81.833y(\mathbf{u}) + 32.392y^2(\mathbf{u}) - 32.392 \\ &= 66.232 + 81.833y(\mathbf{u}) + 32.392y^2(\mathbf{u}) \end{aligned}$$

This defines a lognormal distribution with mean 97.946 and standard deviation 55.339 as shown in Figure 2c.

The results for additional Hermite polynomials are also shown in Figure 2. Distributions were constructed for p-values of 1, 2, 3, 4, 5, 10, 20, 30, 50, and 100. It is easily seen that using more polynomials creates a better fit of the distribution.

Variogram Comparison Examples

We will now examine the results of the variogram comparison. Ten variograms were calculated from the normal scores of ten variables from four different data sets.

The *4133* dataset comes from the bench of a copper mine in Peru. Variograms are calculated from the normal scores of three variables: bismuth, copper and zinc. The results of the variogram transformation for the three variables are shown in Figure 3. The two transformed variograms lie nearly on top of each other and in all three cases the correlation between them is perfect.

The *Amoco* dataset contains three variables from which we generate normal score variograms. These are permeability, porosity, and seismic. The spatial variability of these variables is such that even with a normal transform, the variogram remains noisy and erratic. This is good for showing the robustness and similarity of each transform. The results of the variogram transformation are shown in Figure 4. Again, the two transformed variograms follow each other exactly and for two of the three variables the correlation is perfect.

The *dallas* dataset has only one variable, lead. The variogram calculated from the normal transformed lead data is shown in Figure 5 along with the variograms generated by the two transformation methods. The two transformed variograms are identical and exhibit perfect correlation.

Finally, the variables used from the *red* dataset are silver, copper, and zinc. The variograms generated by this dataset by both calculation and transformation are shown in Figure 6. The two transformed variograms are again identical with perfect correlation in all three cases.

Comparison

Each method of transformation is remarkably simple to implement. The MCS method requires only the transformation table and the calculated normal scores variogram as input. The Hermite method requires the transformation table, the calculated normal scores variogram, and the number of polynomials the user wishes to employ. Too few polynomials will create a poor fit of the data and too many take a long time to calculate. It is recommended that 30-100 polynomials be used in the transformation. For each of the examples presented in this note, 100 polynomials were used and CPU time was no issue at all. Choosing the number of polynomials to employ is trivial and should not deter from the simplicity of the method. Each method is equally easy to employ and both create the same results.

Conclusions

The benefit of doing this transformation is that we end up with variograms in the units of the original data without having to calculate them from the original data. They are calculated from the normally transformed data which creates a much more stable and consistent variogram. The results of each variogram transformation method are essentially the same. Both rely on the transformation table created during the initial data transformation, although each utilize this table in a different manner.

References

Neufeld, C.T., 2005. *Guide to Recoverable Reserves with Uniform Conditioning.* CCG, Edmonton, AB.

Vann, J. and Sans, H., 1995. *Global Estimation and Change of Support at the Enterprise Gold Mine, Pine Creek, Northern Territory – Application of the Geostatistical Discrete Gaussian Model*. APCOM XXV 1995 Conference.

Wilde, B.J. and Deutsch, C.V., 2005. *A New Approach to Calculate a Robust Variogram for Volume Variance Calculations and Kriging.* CCG Annual Report, 2005.

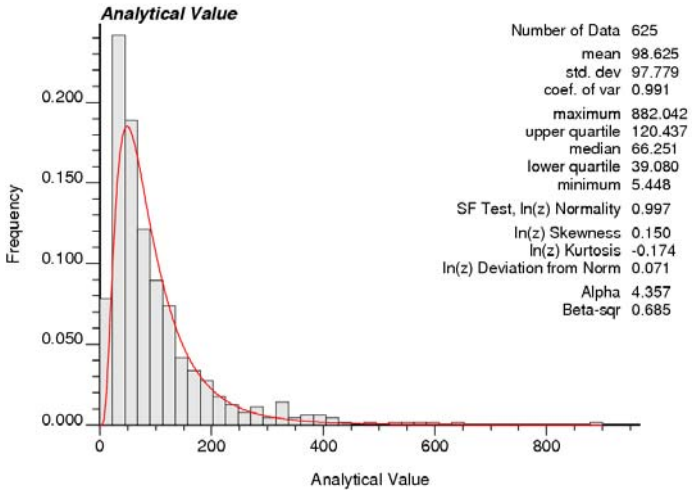

Figure 1: A typical log-normal distribution. This distribution is fit using Hermite polynomials.

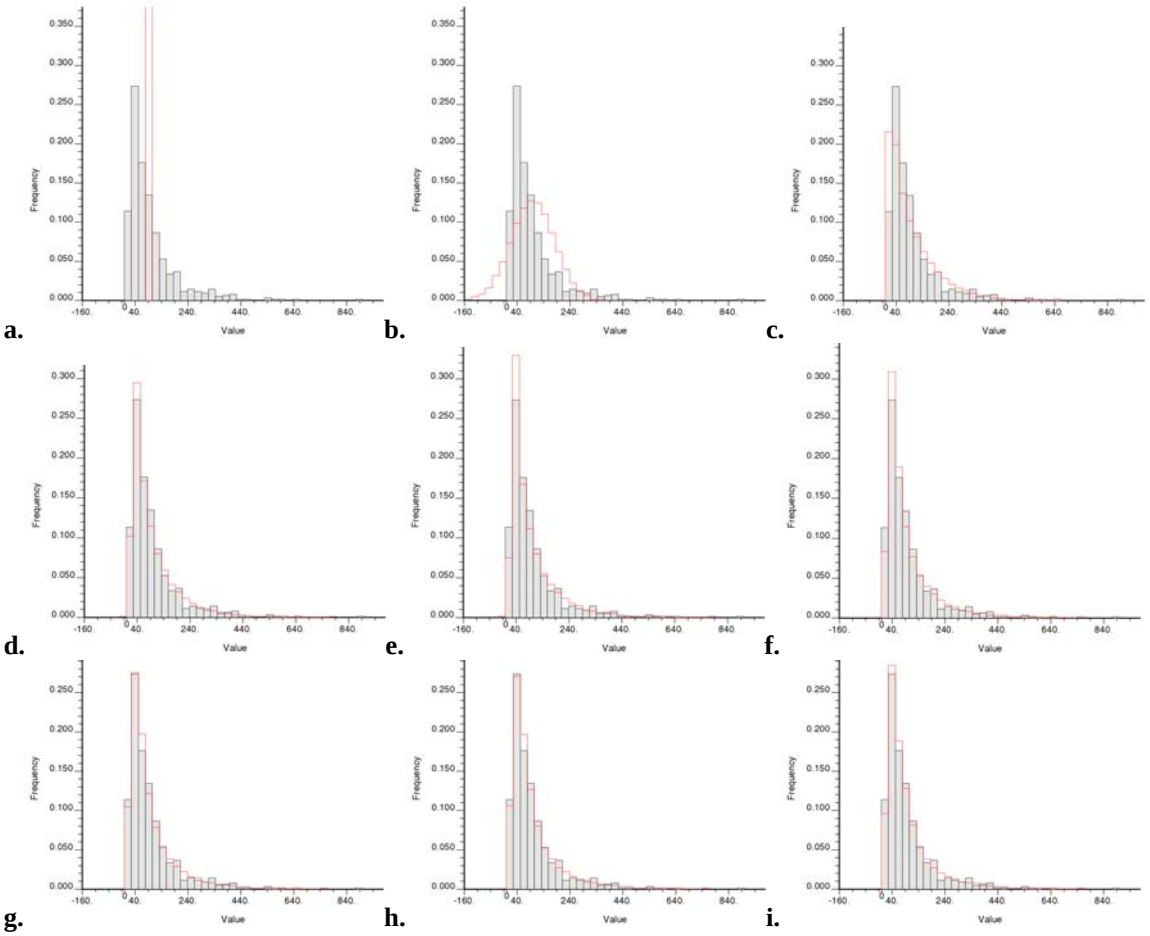

Figure 2: Distribution fitting using different numbers of Hermite polynomials. The number of polynomials used are 1, 2, 3, 4, 5, 10, 15, and 20 for plots a to i, respectively.

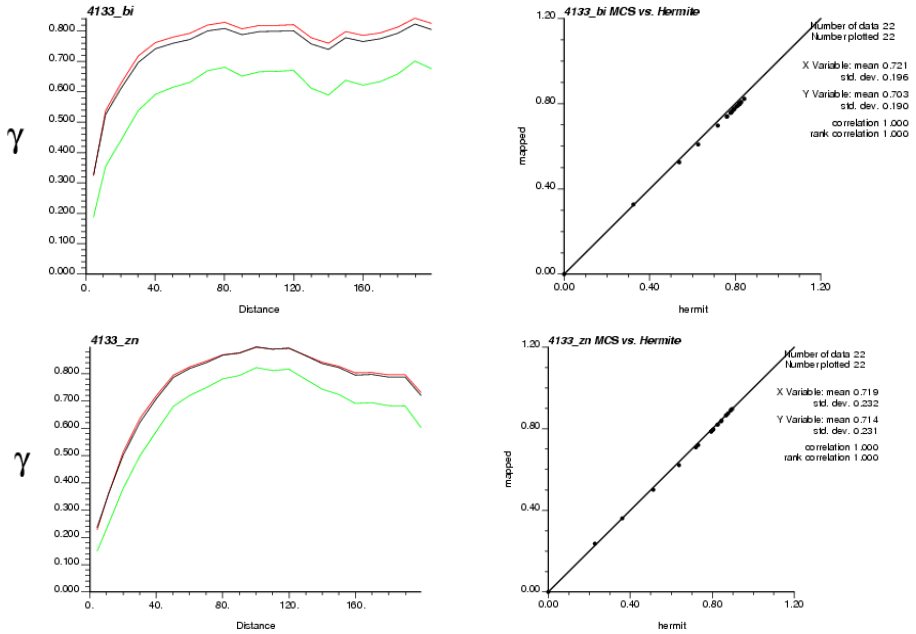

Figure 3: Variogram transformation results for bismuth and zinc for one dataset. The normal scores variogram is shown in green, the variogram transformed by the Hermite method is shown in red while the variogram transformed by the MCS method is shown in black.

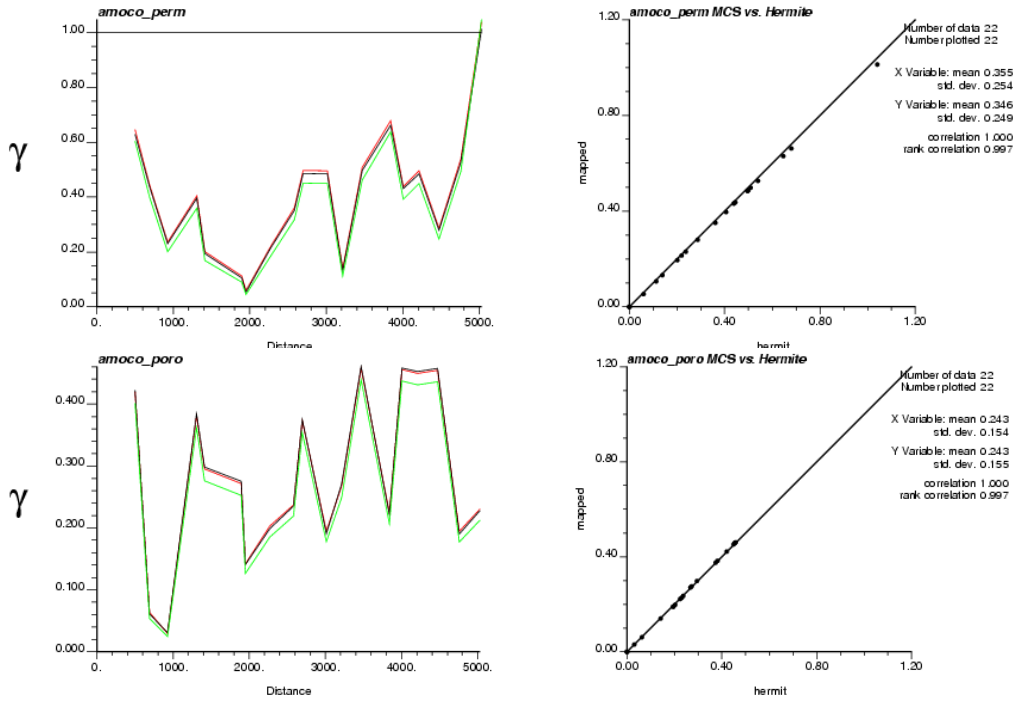

Figure 4: Variogram transformation results for permeability and porosity from Amoco.dat. The normal scores variogram is shown in green, the variogram transformed by the Hermite method is shown in red while the variogram transformed by the MCS method is shown in black.

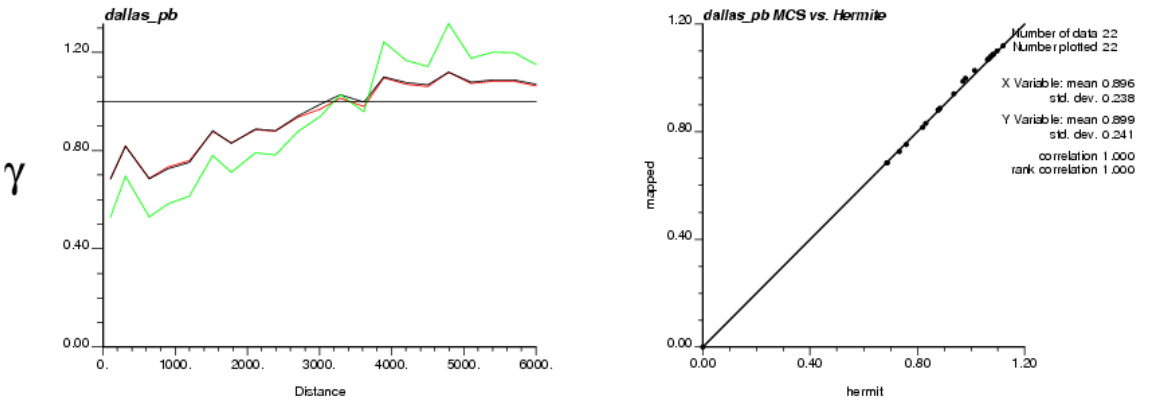

Figure 5: Variogram transformation results for lead from dallas.dat. The normal scores variogram is shown in green, the variogram generated by the MCS method of transformation is black, and the Hermite variogram is red. The scatterplot shows the similarity between the MCS and Hermite methods.

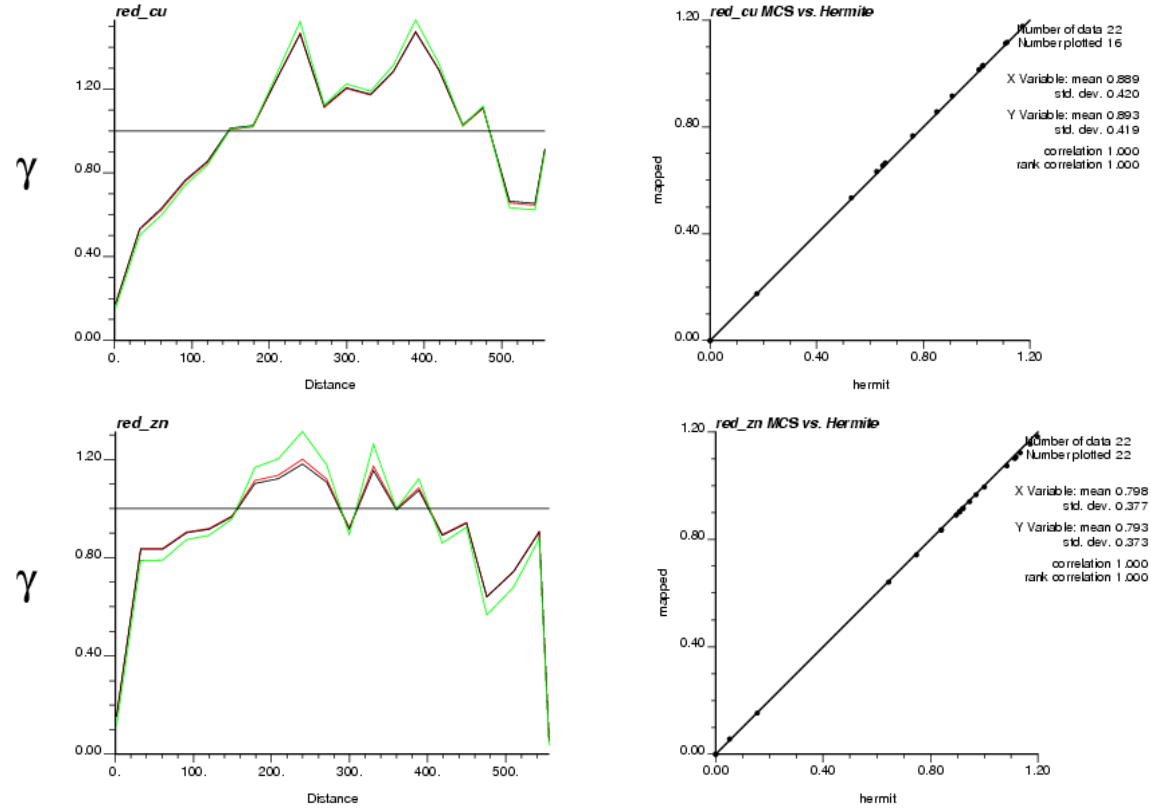

Figure 6: Variogram transformation results for copper and zinc from red.dat. The variogram shown in green is the normal score transformed data. The black variogram is the one calculated by the MCS transform. The red variogram is the one calculated by the Hermite transform.