Robust Alternatives to the Traditional Variogram

Brandon Wilde and Clayton V. Deutsch

Centre for Computational Geostatistics (CCG)
Department of Civil and Environmental Engineering
University of Alberta

Geostatistical procedures such as kriging and simulation require a stationary covariance or variogram model. The variogram is often computed because it contains equivalent information to the covariance (under a second order decision of stationarity) and it requires a less strict decision of stationarity. The variogram, however, is sensitive to extreme values and clustered data. A number of robust alternatives have been proposed. The most common alternatives are the pairwise relative variogram, the correlogram, the normal scores variogram and the normal scores variogram corrected to the original variable. We develop these different measures and show a number of examples where a large dataset is available. A "D" statistic like the Kolmogorov-Smirnov D statistic used to compare two distributions is used to judge how well the true vaiogram is measured in presence of sparse data. All of the robust measures are better than the experimental variogram in presence of sparse and clustered data. The correlogram is seen to be the most robust.

Introduction

The common goal of all types of variogram calculation is to understand the true variography within a deposit. The stationary covariance or stationary variogram is required for volume variance calculations and estimation of the original variable. The most correct method for acquiring this measure is to calculate a variogram directly from the data gathered from the deposit. But the variogram of original grades is notoriously unstable. Clustered data and outliers are the main source of instability. Relative variograms and correlograms address these issues in the definition of the calculated statistic. Normal-score and median indicator variograms address these issues by a transformation of the data. In the presence of sparse and clustered data, we have no access to the original variable variogram and it is better to have an approximation of the spatial structure than no structure at all. But which of these methods is the most accurate?

The goal of this paper is to compare the different types of variogram calculation to determine which is the most accurate.

Methodology

In order to determine which type of variogram is the best, we use an exhaustive data set (one in which the data are closely spaced throughout). The abundance of data allows us to accurately calculate a well behaved traditional experimental variogram, one that is not noisy or sensitive to clustered data. The exhaustive data is then resampled at various densities. For our purposes, we have considered three different data densities. Each type of variogram is calculated for each data spacing and is compared to the reference variogram calculated from all of the data. The types of variograms considered are:

- Experimental
- Correlogram
- Pairwise Relative
- Normal Scores
- Mapped

Each type of variogram compares the value of a variable at a location (**u**) to the value of this variable at another location (**u+h**) a certain distance (**h**) away.

The experimental variogram is obtained by calculating the average of the squared differences of all the pairs of data separated by a certain distance, **h**. This is defined as follows:

$$2\gamma(\mathbf{h}) = \frac{1}{n}\sum_{i=1}^{n}(z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h}))^2$$

where there are *n* pairs of data. On a cross plot of z(**u**) and z(**u+h**), the experimental variogram is the average of the square of the horizontal or vertical distance (they are identical) between a point on the plot and the 45º line (see Figure 1)

A correlogram is, as it sounds, a graph of the correlations at specific separation distances, **h**. It is typically 'turned over' by subtracting the value of the correlogram from 1.0 (if the distribution is standardized). As such, the 'turned over' correlogram has the following form:

$$\gamma_{corr}(\mathbf{h}) = 1 - \frac{C(\mathbf{h})}{\sigma_{(\mathbf{u})}\sigma_{(\mathbf{u}+\mathbf{h})}}$$

where

$$C(\mathbf{h}) = \frac{1}{n}\sum_{i=1}^{n} z(\mathbf{u}_i)z(\mathbf{u}_i + \mathbf{h}) - \frac{1}{n}\sum_{i=1}^{n} z(\mathbf{u}_i)\frac{1}{n}\sum_{i=1}^{n} z(\mathbf{u}_i + \mathbf{h})$$

$$\sigma(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} z(\mathbf{u}_i)z(\mathbf{u}_i) - \left[\frac{1}{n}\sum_{i=1}^{n} z(\mathbf{u}_i)\right]^2$$

$$\sigma(\mathbf{u}+\mathbf{h}) = \frac{1}{n}\sum_{i=1}^{n} z(\mathbf{u}_i + \mathbf{h})z(\mathbf{u}_i + \mathbf{h}) - \left[\frac{1}{n}\sum_{i=1}^{n} z(\mathbf{u}_i + \mathbf{h})\right]^2$$

The correlation is, in effect, the moment of inertia of the scatterplot about the 45º line (see Figure 2).

A pairwise relative variogram is calculated using a method similar to the experimental variogram, but it accounts for the outliers and clustered data by a weighting scheme. The definition of a pairwise relative variogram is as follows:

$$\gamma_{PR}(\mathbf{h}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h})}{\left(\frac{z(\mathbf{u}_i) + z(\mathbf{u}_i + \mathbf{h})}{2}\right)}\right)^2$$

The weighting is performed by giving the outliers of the data less weight than the other data (see Figure 3).

The normal scores data is calculated in exactly the same manner as the experimental variogram, but the data has been transformed to a standard normal distribution. This removes the effect of outliers and clustered data creating a smooth, well behaved variogram (see Figure 4).

The normal scores variogram can be back transformed to a variogram representing the variography of the original variable. This procedure is discussed in detail herein. See the variogram mapping section for more complete information on this method.

In order to evaluate the accuracy of each type of variogram, a measure is used called the D-value (see Figure 5). This value represents the greatest vertical separation between compared variograms.

Variogram Mapping

Variogram mapping is a new method for calculating a robust variogram. The normal scores (Y) variogram is very robust because the Gaussian transform (1) ensures that there are no outliers, and (2) removes the proportional effect, which mitigates the effect of clustered data. The Y-variogram, however, is biased with respect to the original variable (Z) variogram. The key idea of variogram mapping is to infer the Y-variogram and then back transform it to an unbiased Z-variogram. This back transformation of the variogram is referred to as 'mapping'. The back transformation is non-trivial because we must consider each bivariate distribution and perform the back transformation with Monte Carlo Simulation (see Figure 6).

The first step is to transform the original variable by the normal score transform. Variograms are then calculated using the normal score data. Accurate models of these variograms are then created using either a variogram fitting algorithm or the guess and check method of creating a variogram model. It is important to create as accurate a model as possible. Once the normal scores variogram has been correctly modeled, it is used to 'map' a variogram of the original variable. This mapping is performed one lag distance (**h**) at a time until each lag is mapped. The mapping process begins by first determining the correlation, rho (ρ), at a given '**h**'. The correlation is obtained from the normal scores variogram by the following equation:

$$\rho(\mathbf{h}) = 1 - \gamma_y(\mathbf{h}) \tag{2}$$

A large number of paired samples are then taken from a bivariate Gaussian distribution characterized by ρ. This bivariate Gaussian distribution is sampled using Monte Carlo Simulation. Two values, Y_1 and Y_2, are drawn randomly from the distribution. These two values are then back transformed to Z_1 and Z_2 according to the initial normal scores transformation performed on the original variable. The squared difference of Z_1 and Z_2 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**' until the complete variogram is mapped. This variogram can then be fit and used for modeling purposes. It is unbiased and much more stable than the variogram calculated directly from the original variable.

Examples

Testing of the variogram mapping theory was done using two exhaustive data sets. Exhaustive data sets were used as they provide a means of checking the accuracy of the different types of variograms. This is due to the fact that the 'true' (original variable) variogram is obtainable from an exhaustive data set.

The first data set considered is blasthole data from a copper mine in Peru. The area is 250 x 300 m and contains 955 blastholes. The locations of the blastholes are shown in Figure 7. A standardized variogram is calculated from the copper grades as a reference variogram to compare the other types of variograms to. The data set is resampled at various densities in order to observe the accuracy of variograms calculated from data with varying densities. Three different densities are considered. The densest considers 617 data, the next 370 and the sparsest with 111 data. The locations of the data are shown in Figure 8. The reference variogram calculated from all of the data is shown in Figure 9. The fifteen variograms calculated (5 types, 3 data densities) are shown in Figure 10. The first variogram type is an experimental variogram calculated directly from the copper values, the second is a correlogram, the third is a pairwise relative variogram, the fourth is a variogram calculated from the normal scores of the data, and the fifth is a variogram mapped from the normal scores variogram. These variograms are then compared with the reference variogram as shown in Figure 11. The reference variogram is in red, the black points are from the above calculated variogram, and the black line is the fitting of the calculated variograms. The D-values are compared in Figure 12, Figure 13, and Figure 14. From these figures we see that the correlogram consistently has the lowest D-value and is the most accurate. The mapped variograms are also very accurate. They have a higher D-value, but it occurs at a distance of zero (due to choice of nugget). The rest of the structure is very accurate, at least for the higher data densities.

The second data set is one created by authors Edward H. Isaaks and R. Mohan Srivastava from a digital elevation model from the western United States; the Walker Lake area in Nevada. The original elevation values are not used, but the variable is related to the elevation. The data set locations are shown here in Figure 15. There are 78,000 data. Similar to the previous example, the dataset is resampled at various densities. The locations of these new data sets are shown in Figure 16. The reference variogram calculated from all of the data is shown in Figure 17. The five types of variograms calculated at different data densities are shown in Figure 18. The comparison of calculated variograms to the reference variogram is shown in Figure 19 and the comparisons of D-values are shown in Figure 20, Figure 21, and Figure 22. These results concur with the results of the previous example. The D-values are lowest for the correlograms, while the mapped variograms show an accurate structure at higher data densities.

Conclusions and Recommendations

From these comparisons we see that the correlogram is an accurate and robust method of calculating a measure of spatial dependency. We also see that the mapped variogram is very robust and consistent (unbiased). Different types of variograms are needed for different purposes. For example, when calculating kriged estimates, it is important to have an original variable variogram; but when performing simulation in a Gaussian context, it is important to have a variogram calculated from the normal scores transformation of the data. A correlogram works well for estimation techniques requiring an original variable variogram. The mapping technique can be used as a check for the normal scores variogram when it is needed for simulation.

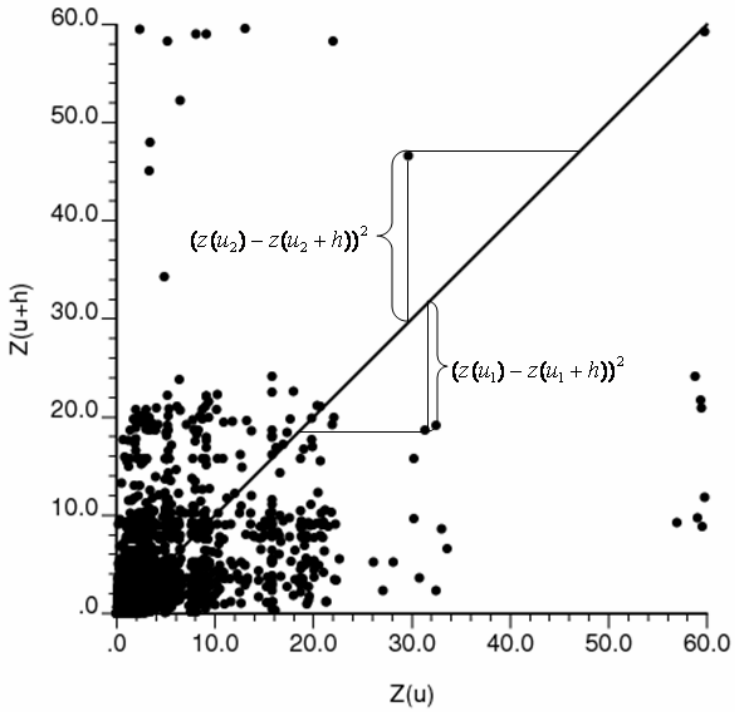

Figure 1: The experimental variogram is the average of all of these distances squared.

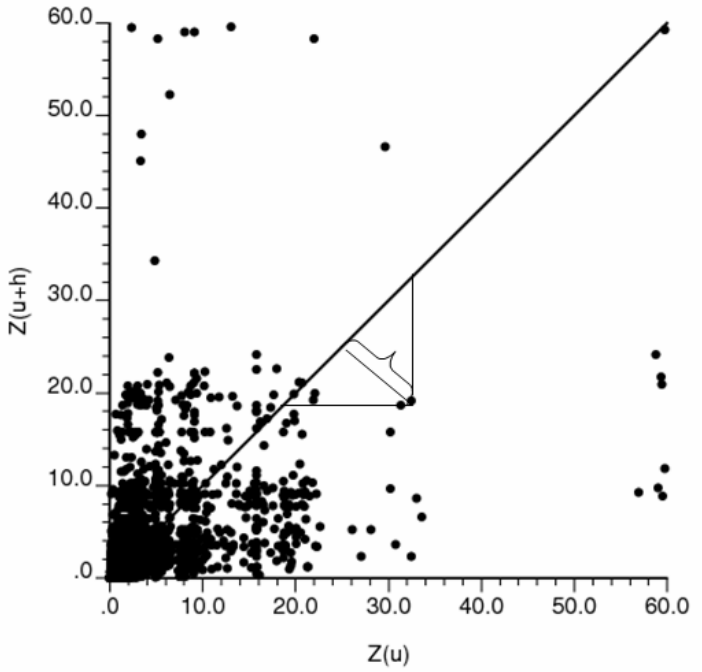

Figure 2: The indicated distance is the one used to calculate the correlation used in the calculation of a correlogram.

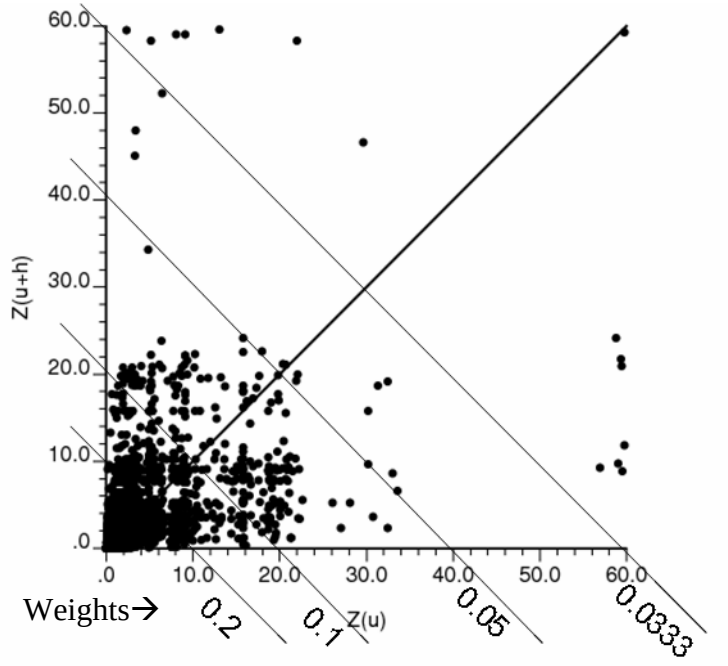

Figure 3: The Pairwise Relative variogram gives outliers less weight than the other values.

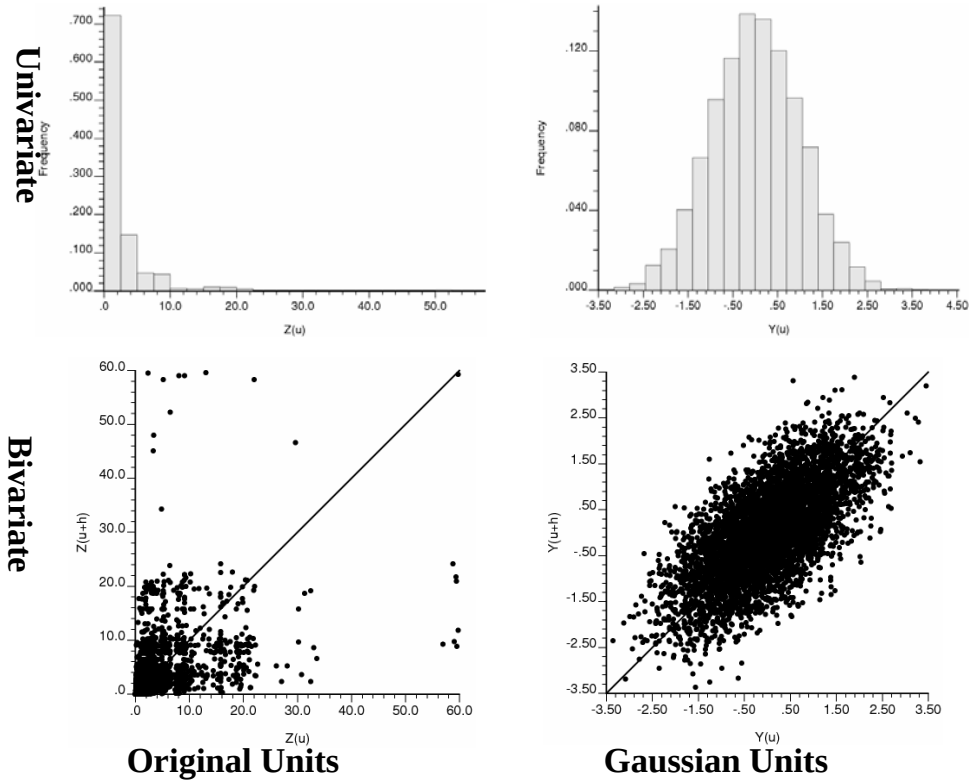

Figure 4: Transforming the data mitigates the effect of outliers, giving a smoother variogram.

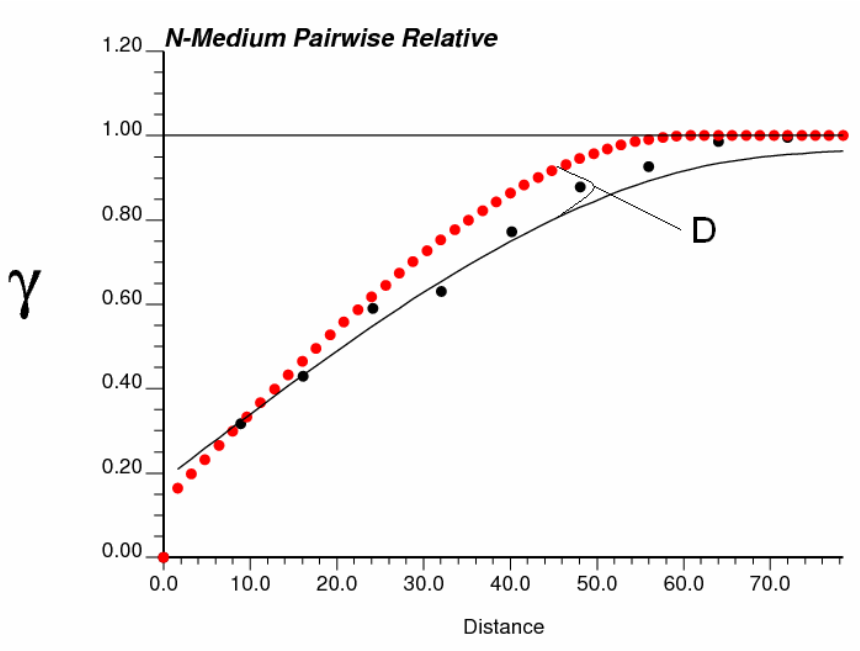

Figure 5: The D-value is used to evaluate the accuracy of each type of variogram.

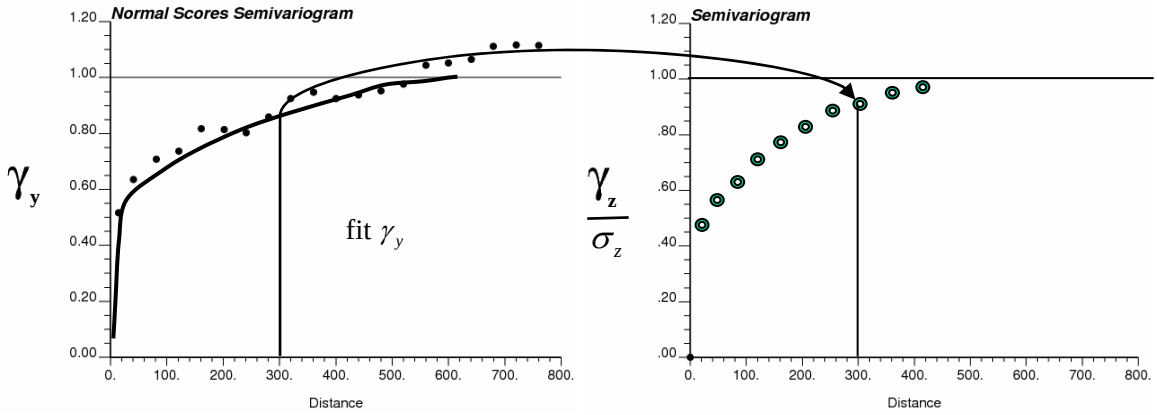

Figure 6: The variography of the original variable can be obtained by back transforming the normal scores variogram.

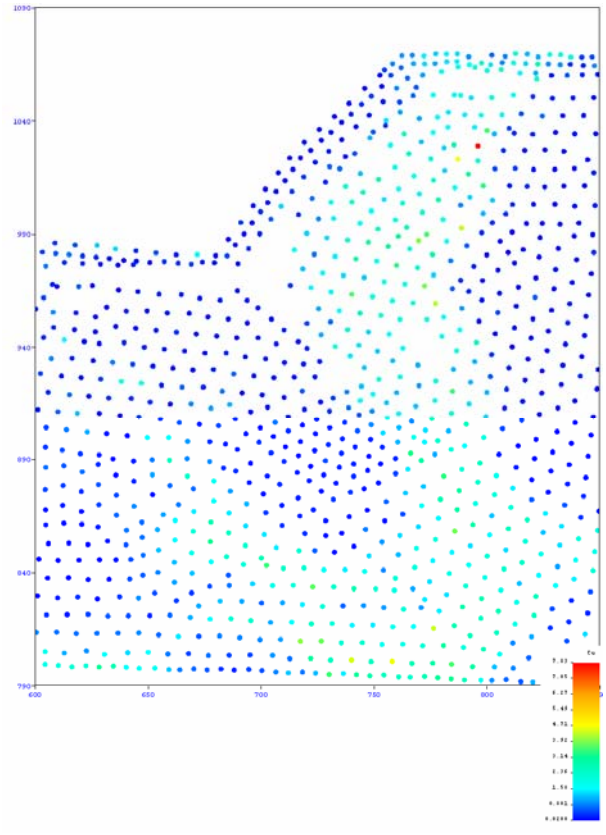

Figure 7: Exhaustive Data Set 1, blastholes from bench in copper mine in Peru.

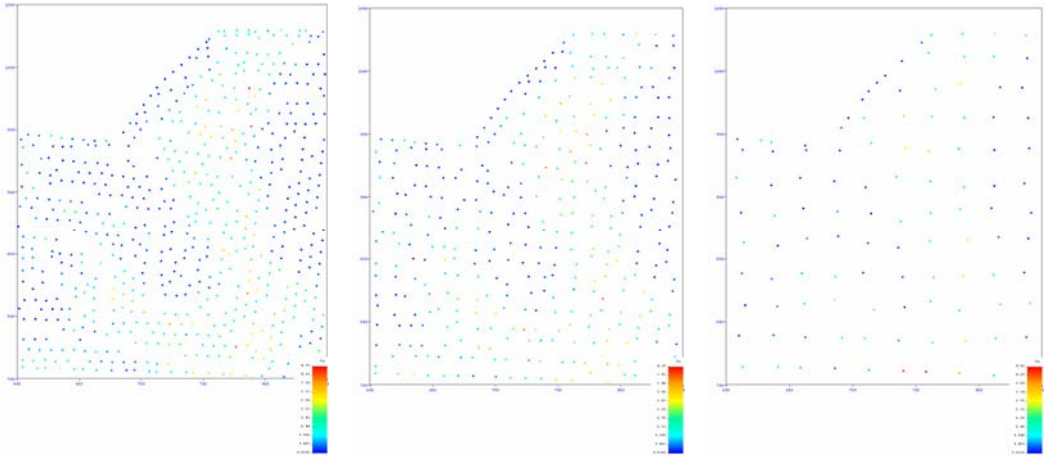

Figure 8: Locations of Resampled Data for varying densities.

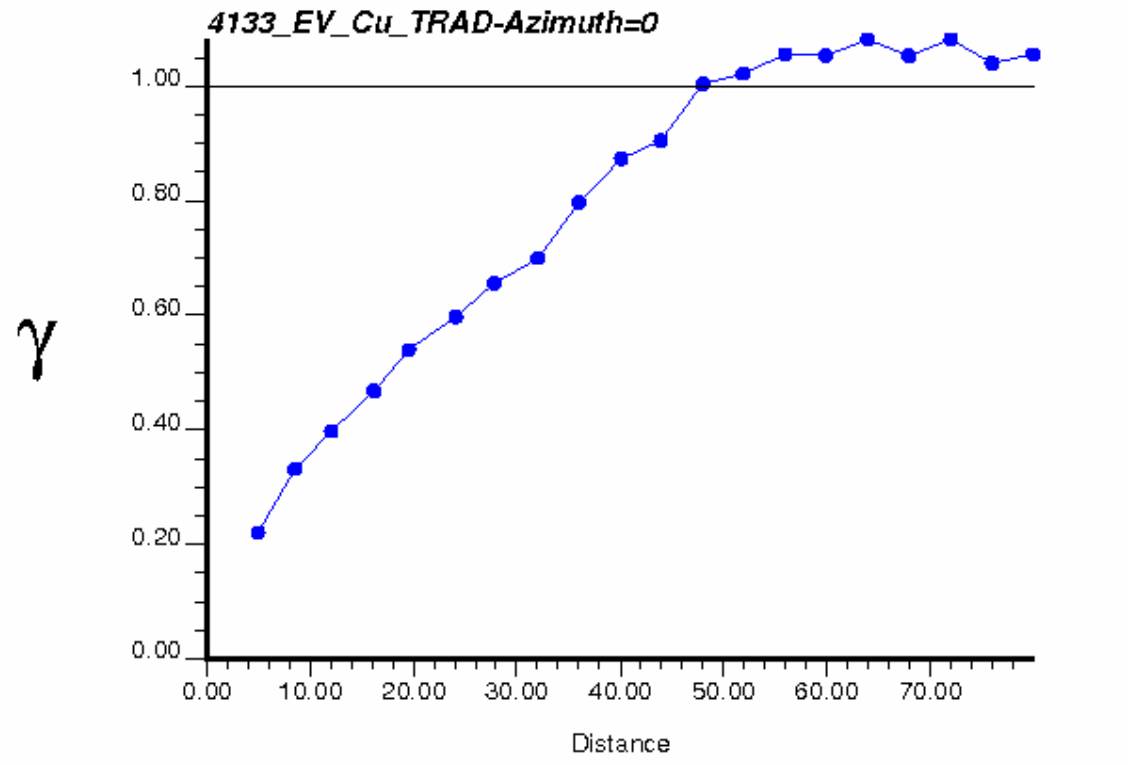

Figure 9: Reference Variogram

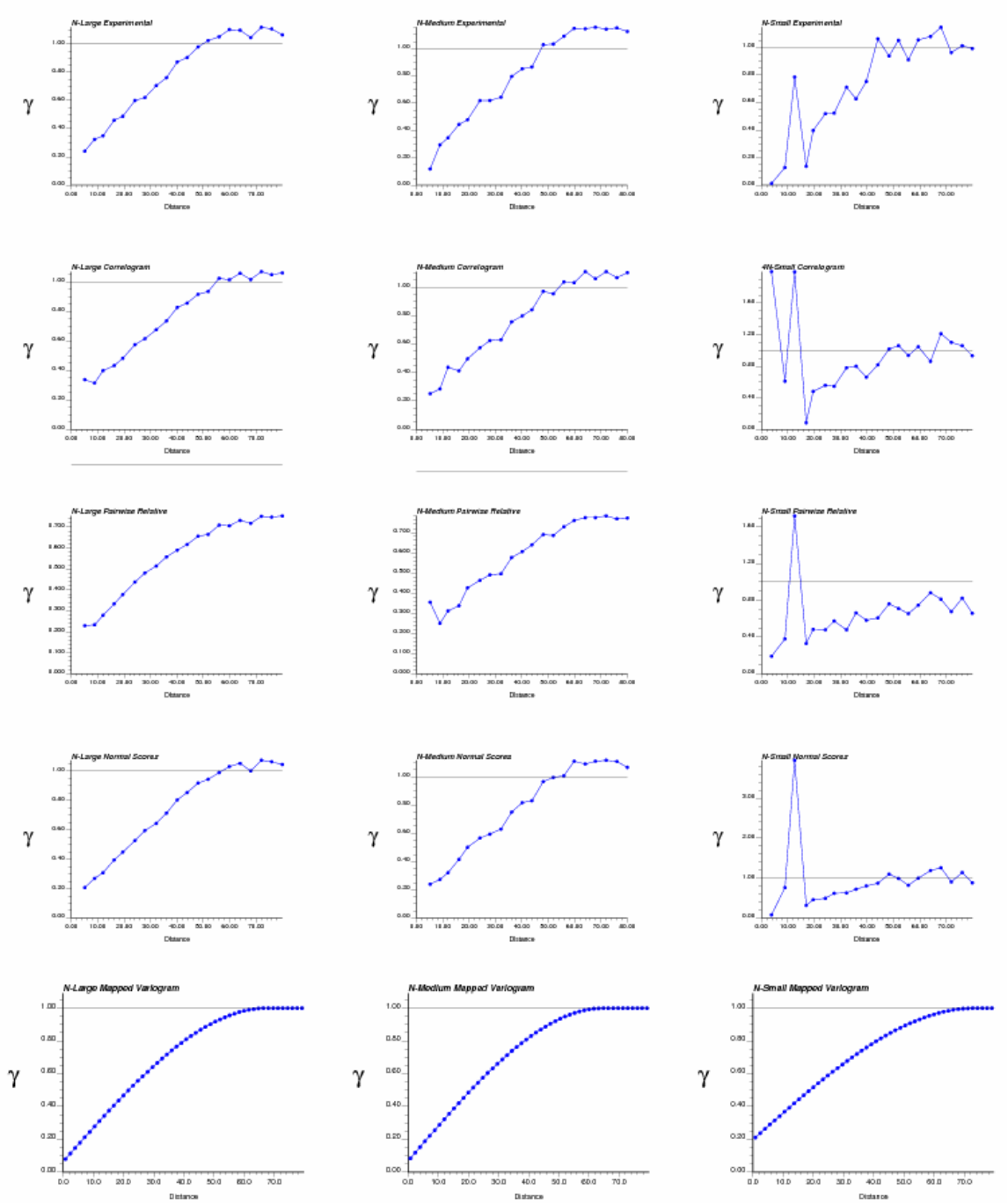

Figure 10: Variogram Types at different densities

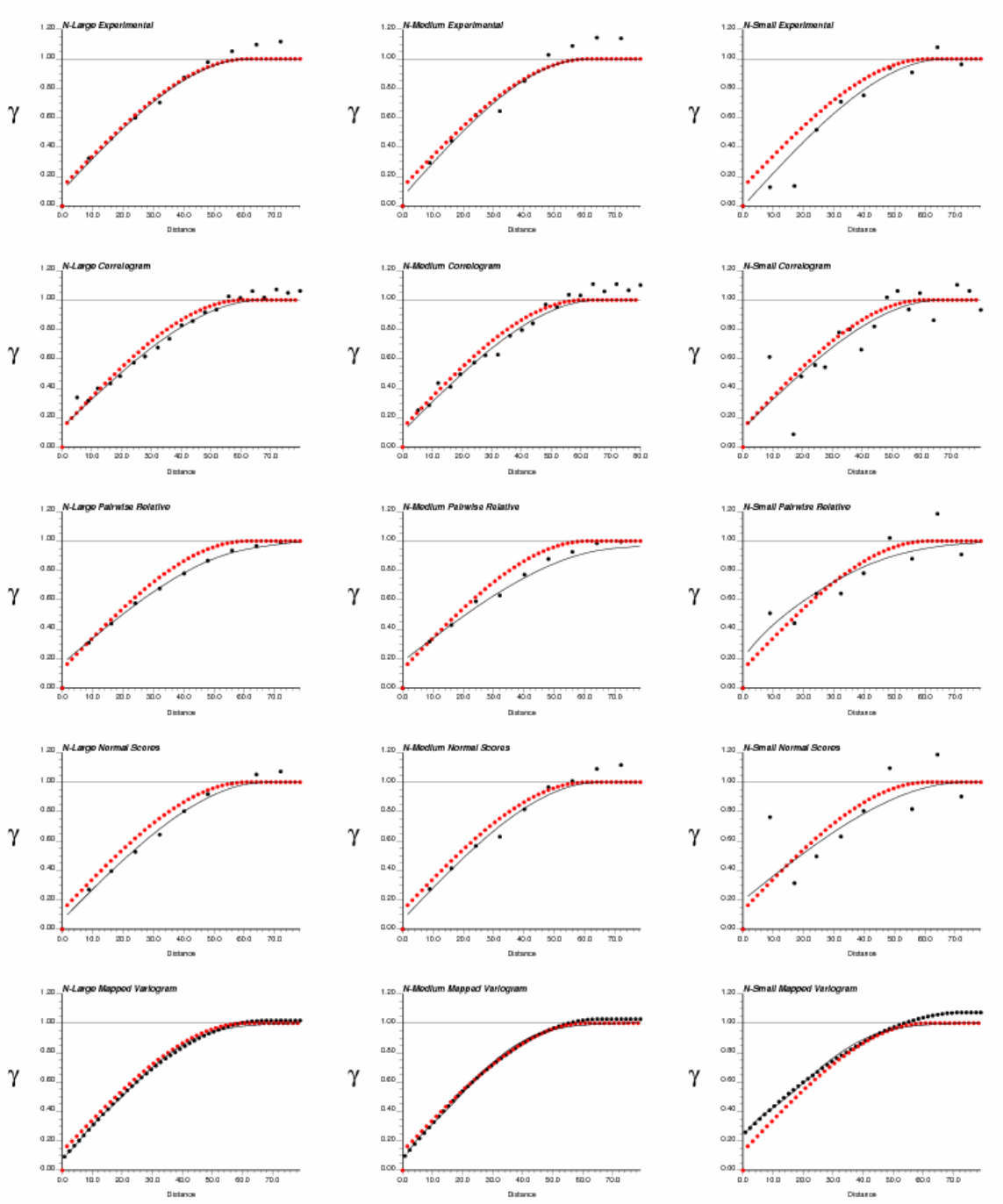

Figure 11: Comparison of variogram types with the reference variogram

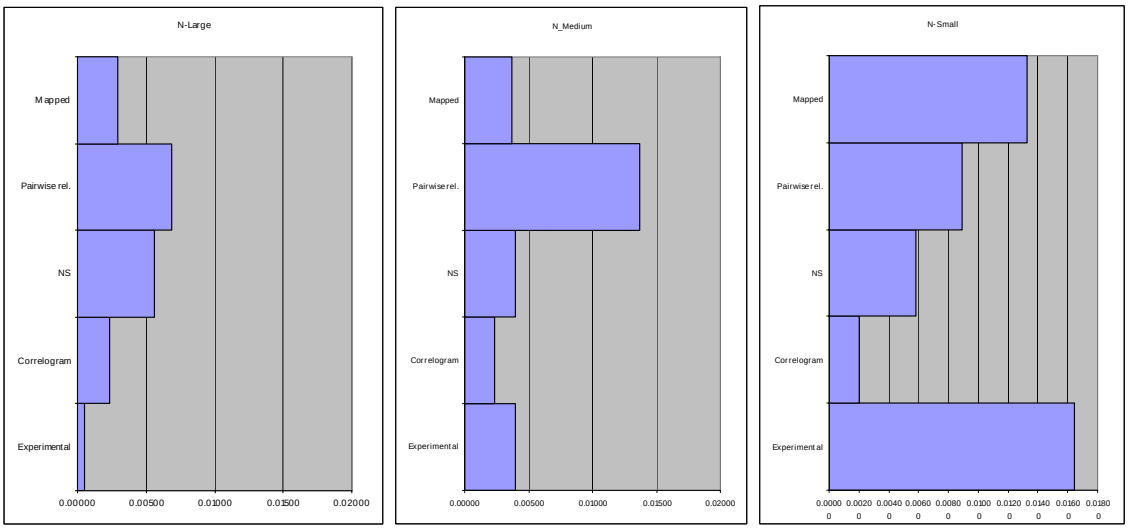

Figure 12: D-values for different types of variograms at different spacings

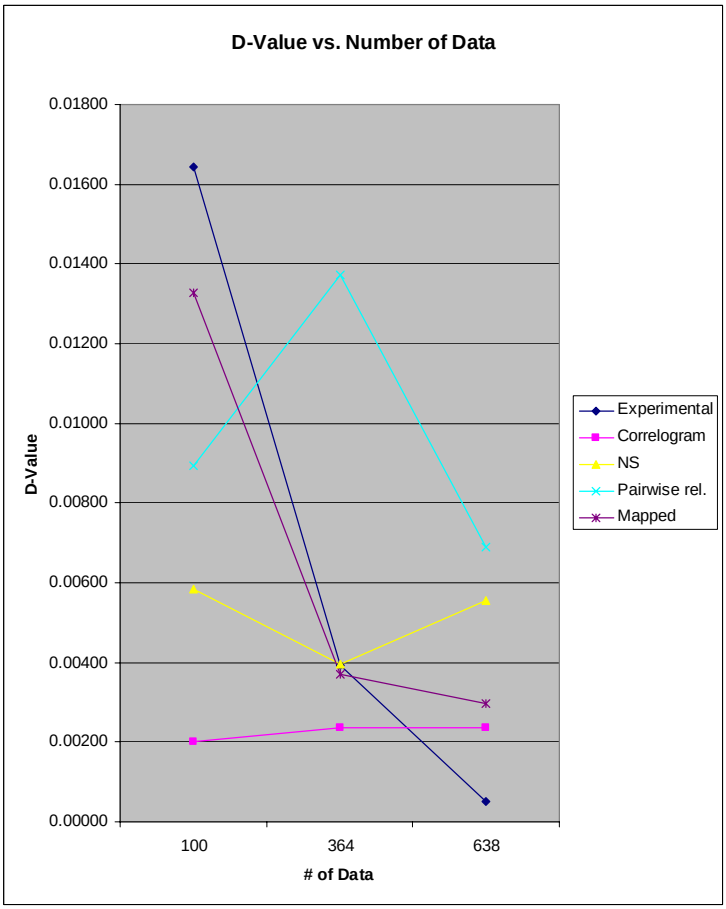

Figure 13: Comparison of D-values

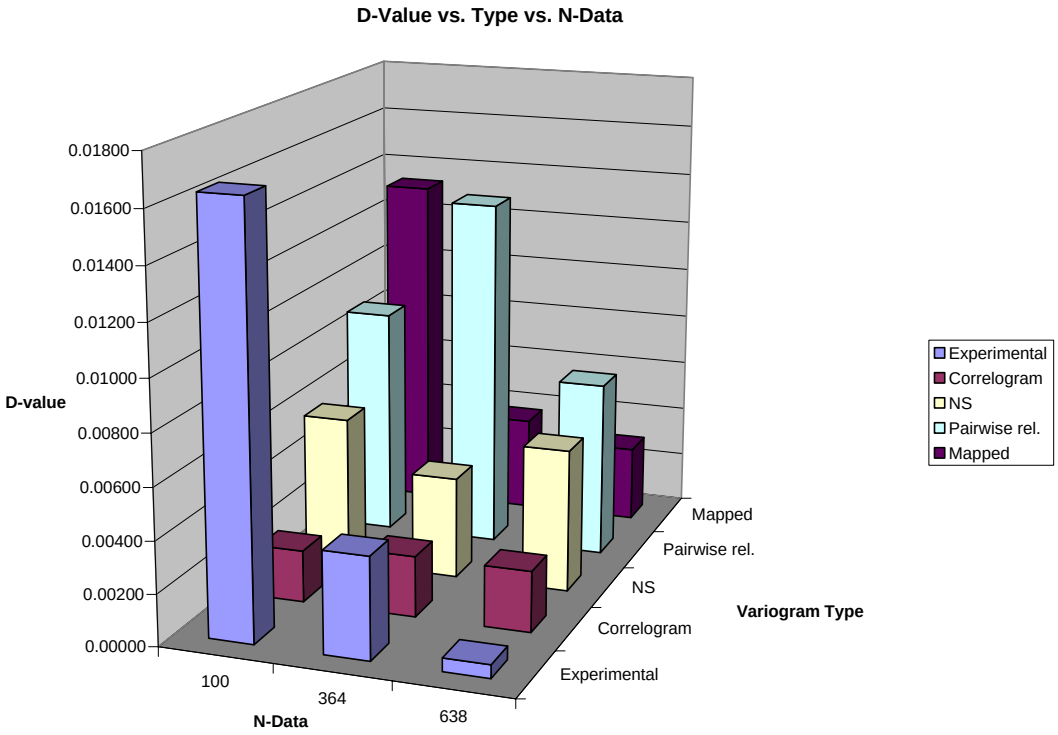

Figure 14: Comparison of D-Values

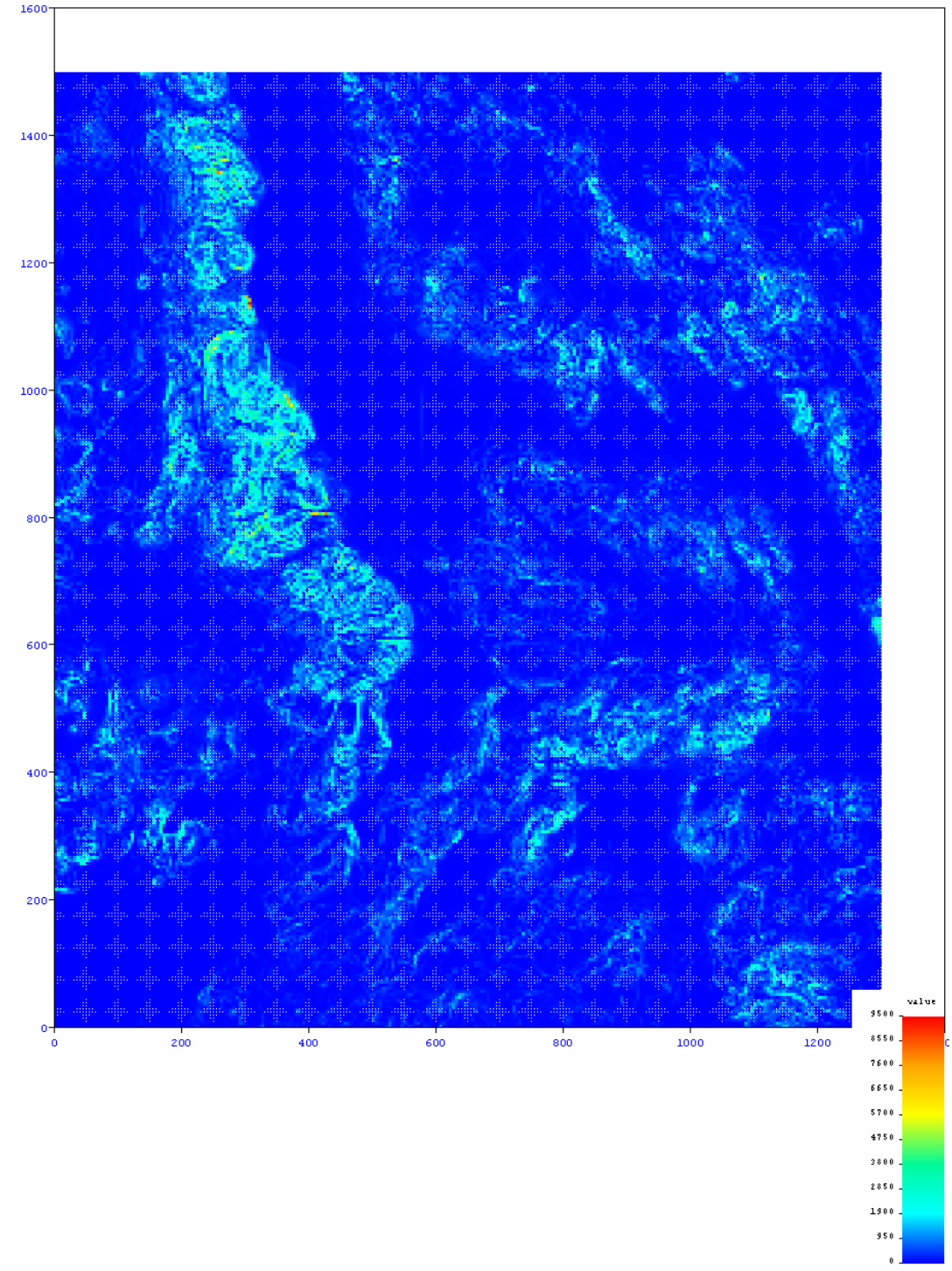

Figure 15: Locations of all Walker Lake data

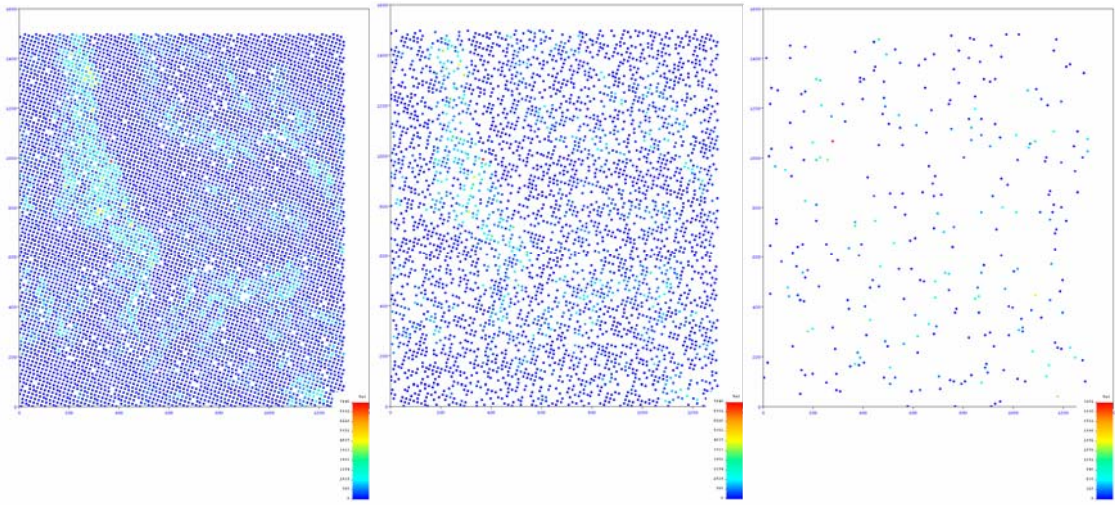

Figure 16: Locations of Resampled Data

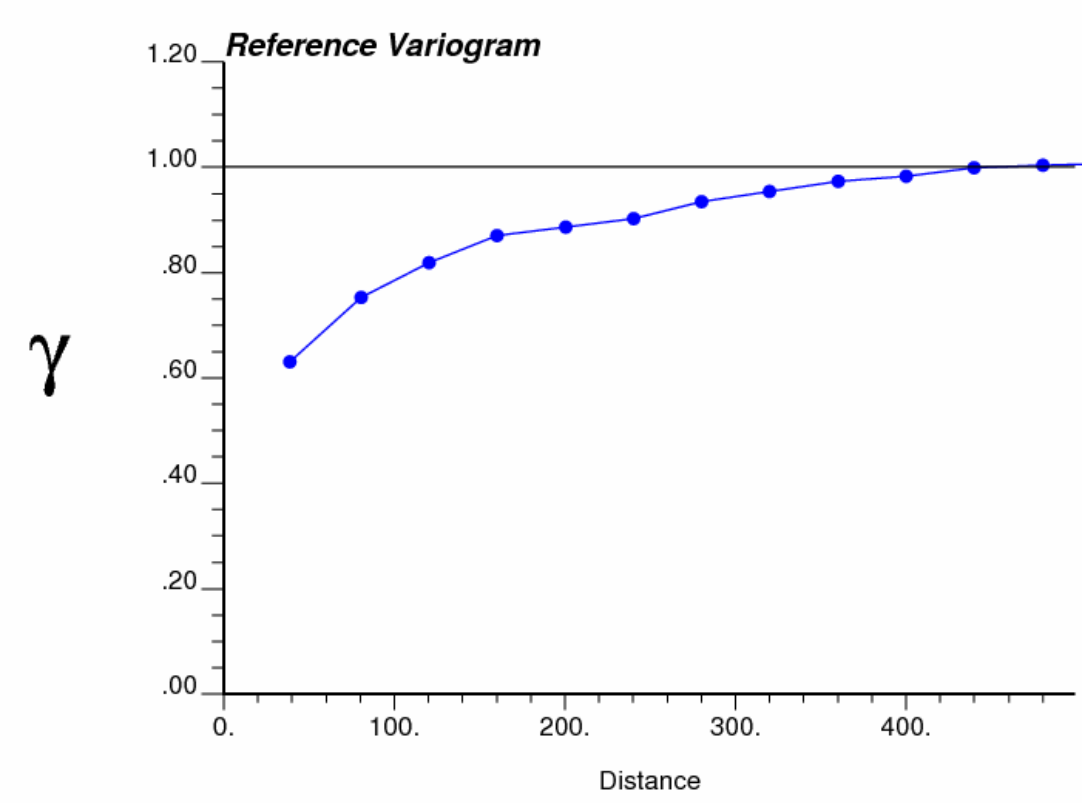

Figure 17: Reference Variogram

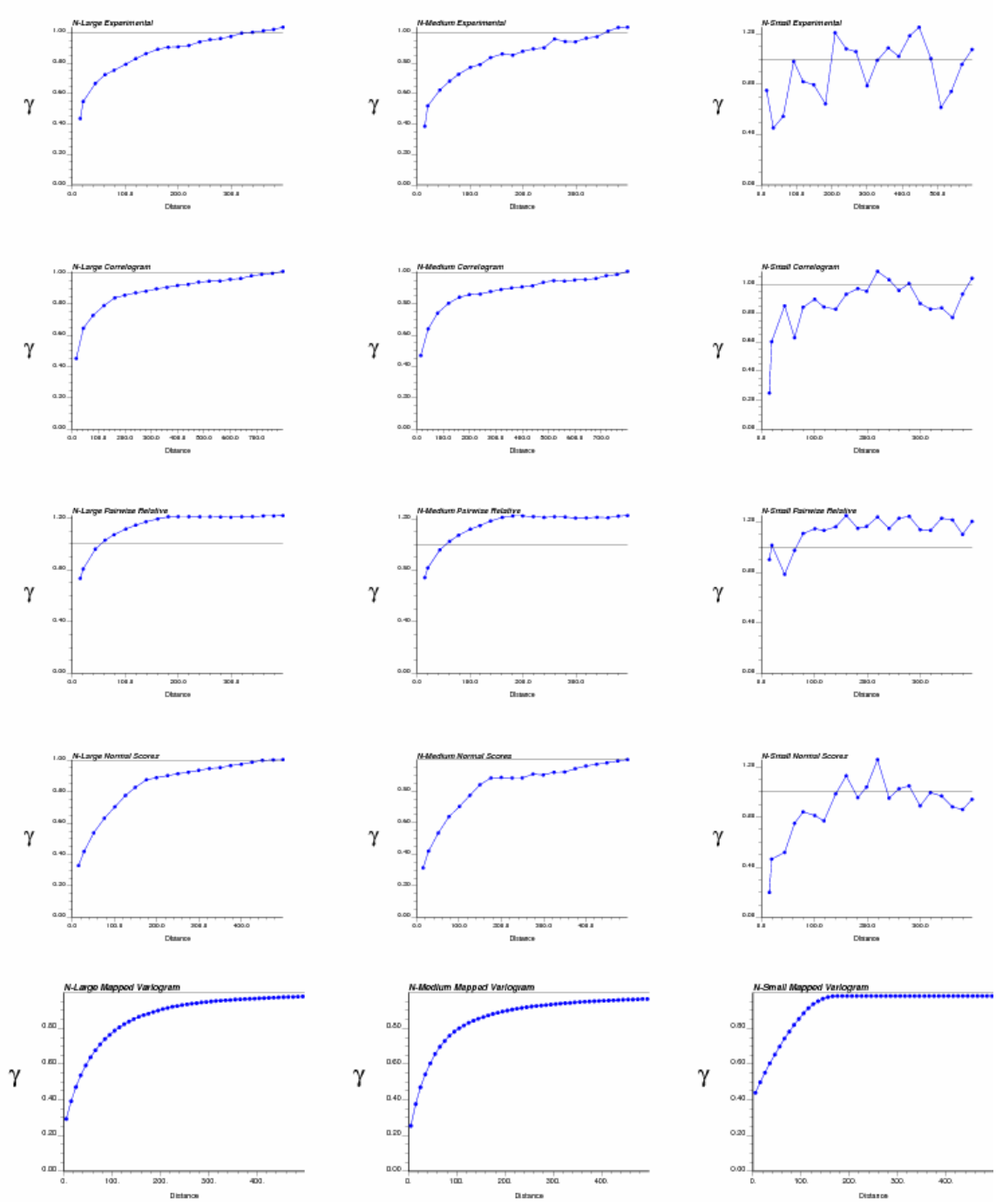

Figure 18: Calculated Variograms at Varying Data Densities

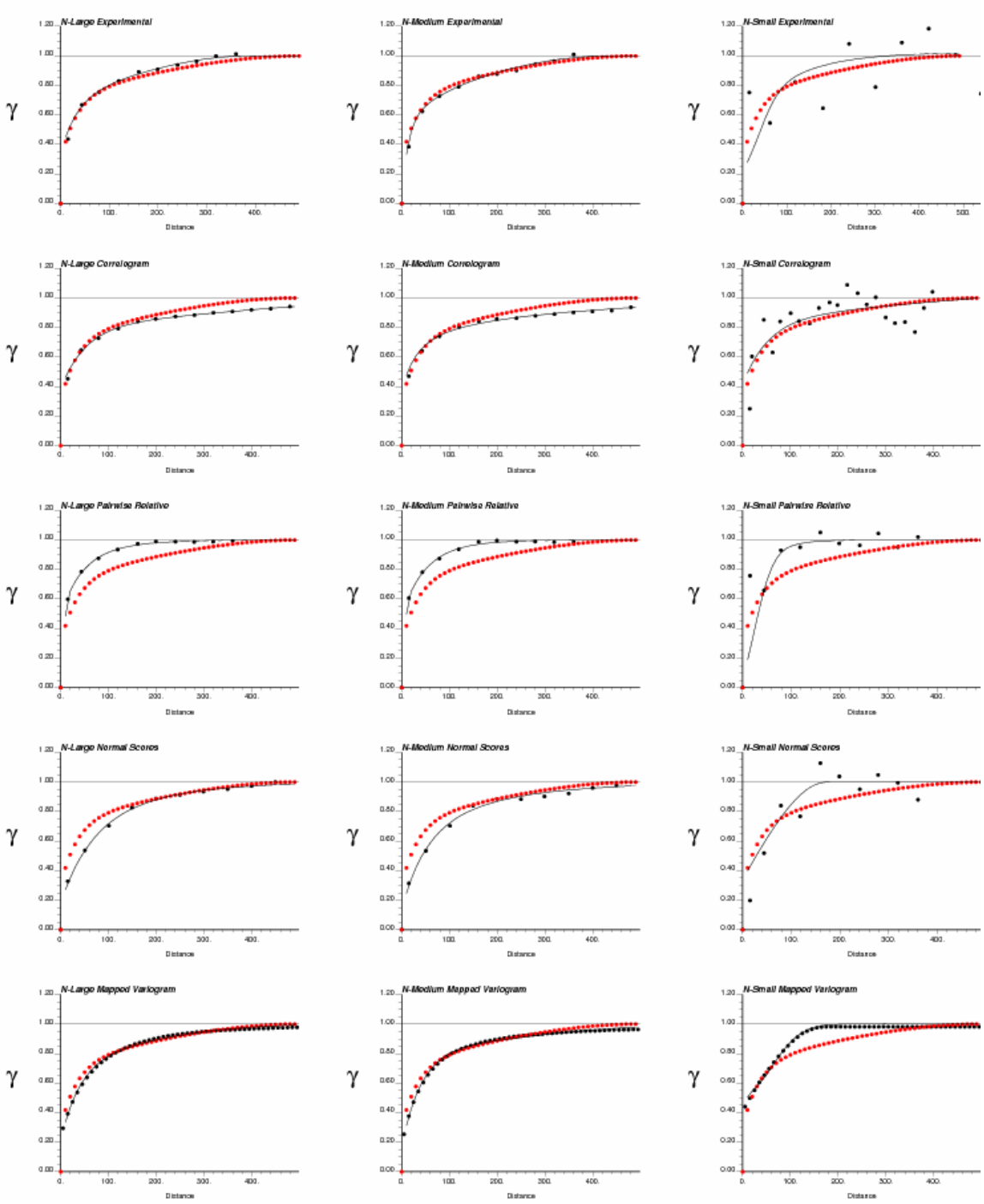

Figure 19: Comparison of variograms with reference variogram

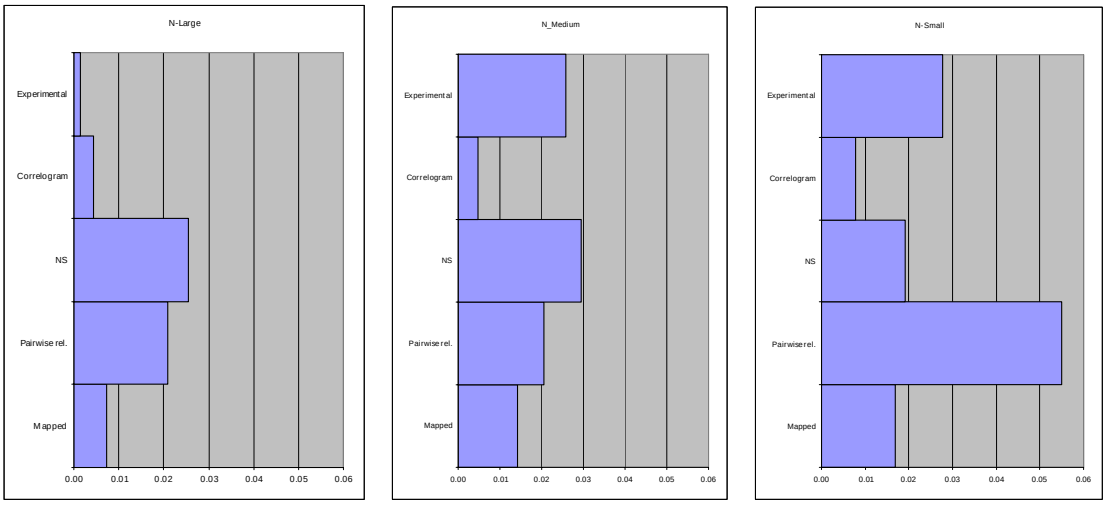

Figure 20: D-values compared

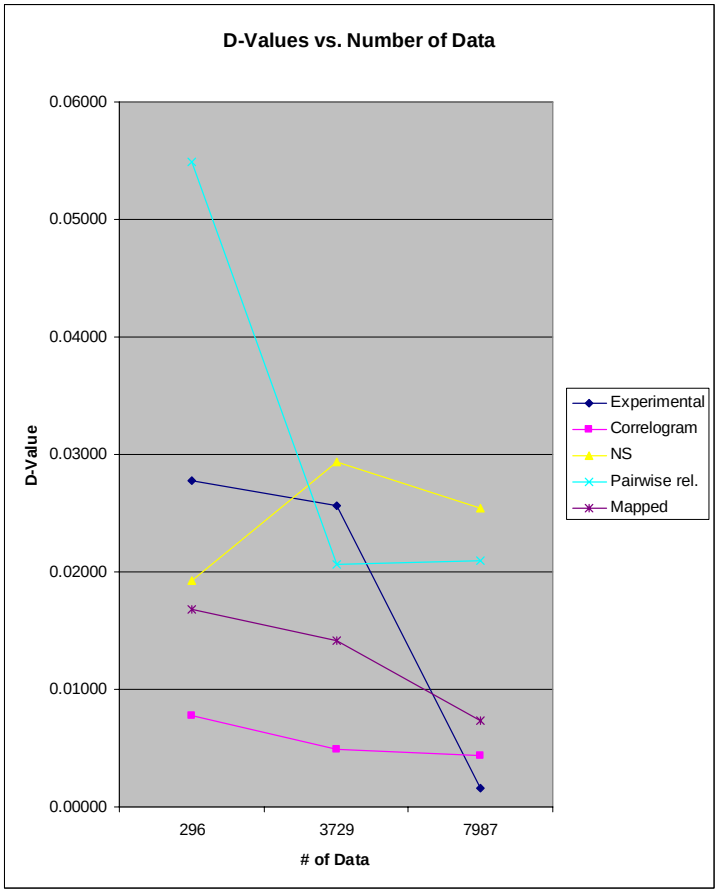

Figure 21: D-values compared

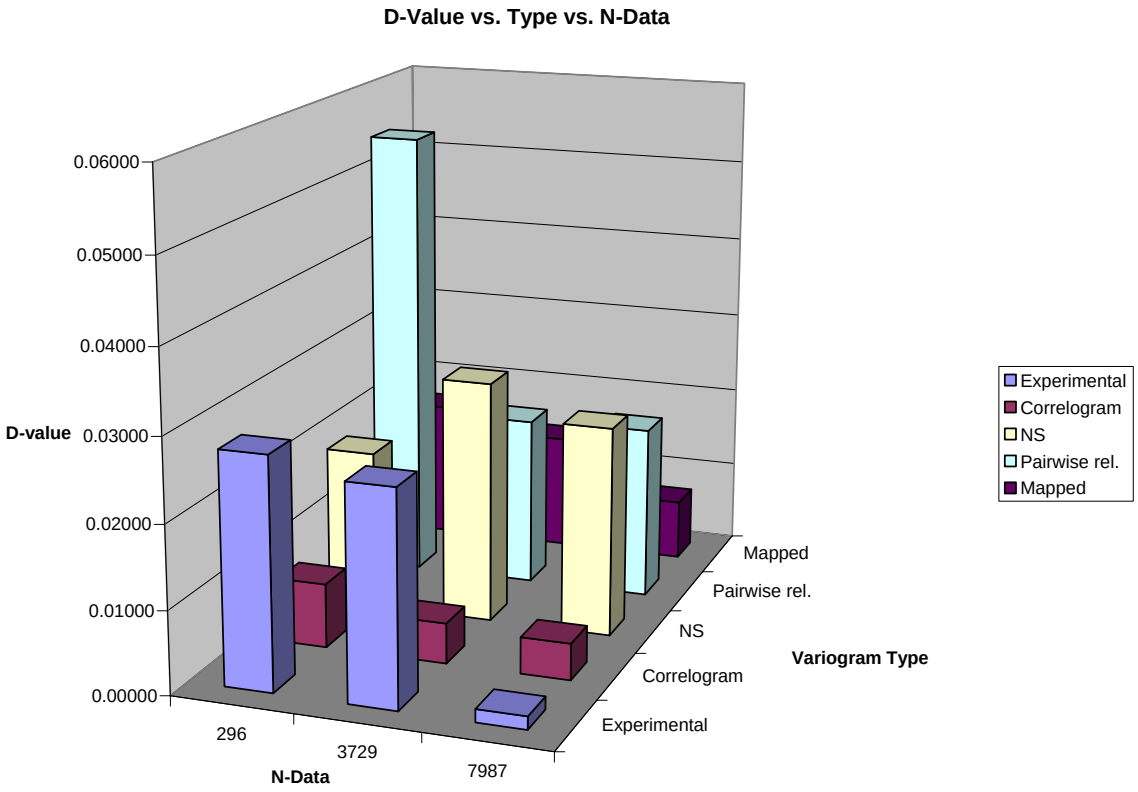

Figure 22: D-values compared