Experimental Variogram Cleaning

Miguel Cuba and Oy Leuangthong

In practice, domains with no trends in the mean or variance are assumed as stationary. However, variations in the variogram could make the domain unsuitable for modeling using conventional geostatistics. The presence of highly variable regions within the domain may mask and reduce the spatial continuity of the rest of the domain. The variogram is very sensitive to extreme values. This document proposes a methodology for accounting for the extreme increments in the experimental variogram under assumption of multivariate Gaussianity. A one dimensional study case is presented; however, the proposed methodology can be expanded to high dimensional problems.

Introduction

Variograms are used to characterize the spatial continuity of a variable of interest. In mining, this variable is assumed to be a realization of a SRF, since there is only one value for each sampled location (e.g. metal grade, contaminant concentration, oxidation ratio). In order to transfer the spatial continuity information into a geostatistical model, the experimental variogram is fitted by a licit variogram model. However, variogram modeling is a subjective task; even when semi-automatic fitting algorithms are used there is still a certain degree of subjectivity. The variogram of a SRF is modeled by comparing and combining a set of available licit variogram models. While this technique does not guarantee an exact representation of the variogram, it provides a good approximation.

In practice, the stationary assumption is less valid in situations where there is a dense sampling pattern over the domain. If the variable of interest is known at all the locations of the domain, the influence of the geologic structures is dominant. However, geologic processes are non-stationary. Therefore, even when the local means and variances are considered constant along the domain, the intrinsic stationary assumption becomes less valid. This is sufficient to make the domain non-stationary.
The variogram model does not characterize the spatial continuity of a non-stationary domain in the same way as a stationary domain does. The spatial continuity of the domain represented in the variogram model, which should be characteristic at all the locations, is now an averaged representation of it. Therefore, the spatial variability characterized by the variogram model is misrepresented locally. It is underestimated in some regions and overestimated in others.

A moment-of-inertia based approach is introduced to calculate experimental variograms. The ideal conditions and characteristics of real data for SGS are discussed. The impact of non-stationary environments on the variogram are described and explained under such conditions. An iterative approach for dealing with patterns in a domain is developed. It involves removing the influence of outlier data pairs and recalculating the variogram model until it describes the spatial continuity of the majority of the dataset. A case study is provided.

Fitting Experimental Variograms

This section describes the impact of experimental variograms on estimation/simulation. In order to avoid any external non-stationary influence, a stationary environment is assumed.

The conventional process of variogram modeling consists of fitting an experimental variogram using licit variogram models. In a stationary environment, the available dataset is assumed to be a realization of a SRF. Based on this assumption, the experimental variogram calculated from one realization is a fair representation of the spatial continuity of the SRF. The use of a set of licit models makes the variogram model an approximation and not necessarily an exact representation of the spatial continuity of the SRF. It is exact when the true variogram of the SRF is the same as the model used for fitting. Alternatively, it is still an approximation when the true variogram is not present as any combination of the set of licit models used for fitting (see Figure 8).

Even under optimal stationary conditions, the process of variogram modeling is subjective and dependent on the user's experience and knowledge. To limit the resultant variability in outcome, several semi-automatic algorithms have been developed for use during the modeling process (Larrondo &

Neufeld, 2003), (Sinclair & Blackwell, 2004). Even then, the way a user inputs parameters may have a discernable effect on the outcome. Because of this inherent subjectivity, the resulting variogram model may be classified as follows (see Figure 1):

- *Pessimistic:* when the model is above the experimental variogram. As a result, the ensuing conditional variances at the estimated locations are larger than expected. The uncertainty assessed in the estimated domain is inflated; therefore, simulated maps are much more variable.
- *Fair*: when the model closely follows the experimental variogram. The estimated conditional variances at unsampled locations are a good approximation of the theoretical variances of the corresponding SRF.
- *Optimistic*: when the model is below the experimental variogram. In this case, the spatial continuity is exaggerated; data locations that should not have a significant influence are forced to contribute information. The conditional variances are smaller than they should be and the uncertainty associated with the model is underestimated.

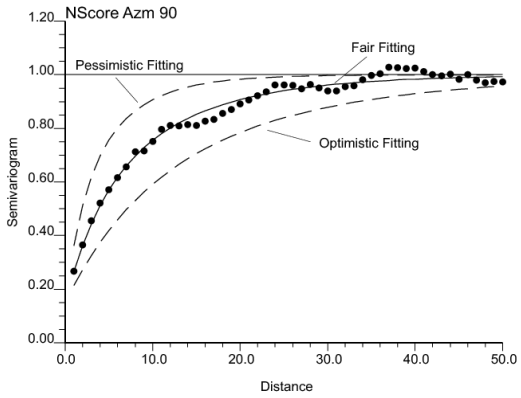

Figure 1: Experimental variogram (black dots), and three cases of variogram modeling (black dots)

With a simple configuration of experimental variogram points, the variogram model should be easy to fit. However, the experimental variogram could describe structures that would be difficult to fit using a common set of variogram models (see Figure 8). In such a situation, the variogram modeling is used to fit a representative portion rather than the whole experimental variogram. By doing this, various regions of the experimental variogram become: optimistic, pessimistic and fair. The variogram model is expected to be globally representative. Since the simulated maps tend to reproduce the variogram models, such features of the modeling process are transferred to the simulated model.

Aspects on the Moment-of-Inertia Experimental Variogram Calculation

The experimental variogram points can be represented as h-scattergrams. In the h-scattergram the data values at the two extremes of the separation vector are presented in the form of a scatterplot. The experimental variogram can be interpreted as the moment of inertia of the h-scattergram about its first bisector. The moment of inertia is the average of the squared orthogonal distances (1) of the data pairs to the first bisector (2) (Goovaerts, 1997). Under the assumption of multi-gaussianity, the bivariate distribution of the data pairs of any h-scattergram is bivariate normal. A multi-gaussian RF is assumed for conventional implementation of SGS.

During the exploratory data analysis stage, the **h**-scattergrams can be used to verify particular features that could correspond to sub-populations. The decision to separate them into small domains is based on the availability of sample points that support such a decision (Goovaerts, 1997). According to (Genton, 1998) this practice is informal and cannot supersede a robust estimation technique of the experimental variogram. However, from an engineering perspective, identifying patterns in the dataset improves the knowledge of the natural phenomena under study. The variogram is a tool that works well under stationary conditions but not as satisfactorily in real-case scenarios.

$$d_i(\mathbf{h}) = \frac{\sqrt{2}}{2}\left|z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h})\right| = \cos 45° \left|z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h})\right| \quad (1)$$

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} d_i^2(\mathbf{h}) \quad (2)$$

Real versus Ideal Environments

In conventional practice of SGS, the attribute of interest is transformed from original units to normal score units and assumed to be multi-gaussian. This transformation is referred to as a normal score transformation and, by construction, will transform any distribution to a univariate standard normal distribution (Deutsch & Journel, 1998). In practice, it is not a requirement for a dataset that is assumed to be sampled from a multi-gaussian process to follow exactly a univariate normal distribution.

Mineral deposits are the result of the interaction of many different previous geologic events over time. The dataset collected from it captures such geologic features, based on processes that obey physical and chemical laws. Ideally, the variable of interest should be modeled in such a manner that accounts for these laws, which in many cases are extremely complex to define. The normal score transformation of a dataset sampled from a mineral deposit preserves the geologic features captured by the original scale data values (see Figure 2). The geologic structures are represented in the datasets as patterns that are analyzed and interpreted by earth scientists. This condition makes the assumption of multi-gaussianity of the domain inappropriate in the presence of dense sampling over the domain.

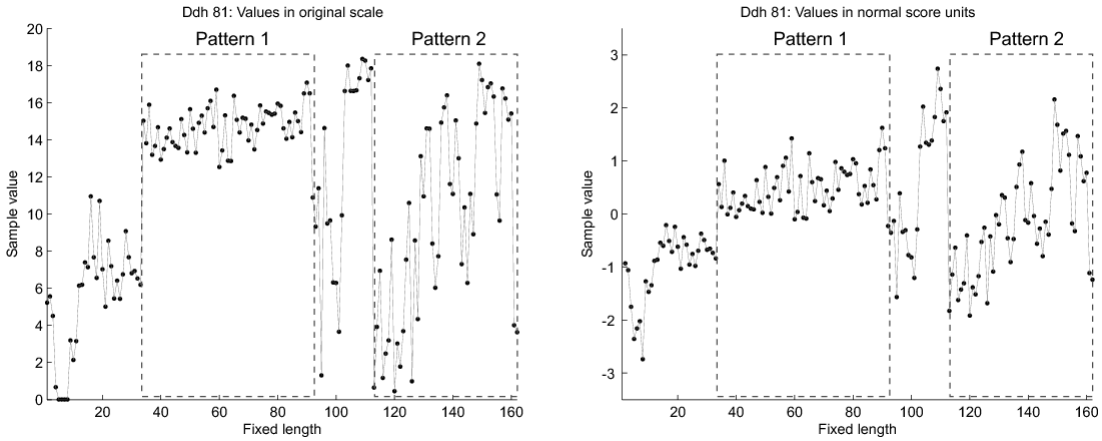

Figure 2: Ddh-81sample values for 162 regularly spaced locations in original units (left) and normal score scale (right), two patterns are highlighted in the dataset that are present both in original and normal score scale units.

Conventional geostatistics uses linear estimation to infer a value for variable of interest at an unsampled location based on the same variable or secondary variables. One necessary condition is the assumption of stationarity of the domain. Geologic features cannot be represented by a SRF. Discrepancies in the use of such models arise when estimated/simulated maps are compared against real information, such as geologic mapping of mined faces and mined production. Two sets of 162 data values from an unconditional realization are shown in Figure 3. When comparing these two datasets with the normal score values of the real case (Figure 2 right) it can be seen the local geologic patterns cannot be reproduced. The unconditional realization was simulated using a variogram model fitted of the real dataset.

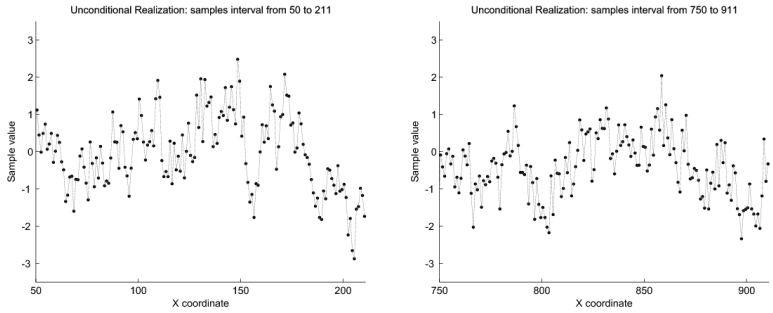

Figure 3: Two sub-datasets of 162 data points from an unconditional realization of 1000 data points

In a non-stationary environment (assuming the local mean and local variance are constant), the variogram is very sensitive to the structural geologic patterns. If a small portion of the domain presents a different variability than the rest of the domain the variogram captures it as a summary of all the structures (Journel & Huijbregts, 1978). Because of the stationary assumption, the information of spatial continuity is generalized to the entire domain.

An exercise is presented that mimics the interaction between two geologic processes in a domain. This is present not because domaining was poorly performed, but spatial structures are inherent to natural processes. Consider one dataset that is the combination of two unconditional realizations, A (80%) and B (20%) with different variograms (see Figure 4). The variogram model of dataset A is $\gamma_A(\mathbf{h}) = Exp_{30}(\mathbf{h})$ and of dataset B is $\gamma_B(\mathbf{h}) = 0.4 + 0.6Exp_{15}(\mathbf{h})$. The resulting combined dataset is non-stationary, despite the local mean and local variance are constant. The absence of the intrinsic hypothesis condition is what makes the dataset non-stationary. As a consequence, the experimental variogram is no longer representative of the dataset. Some features of the small region are transferred to the entire domain. The nugget effect of the small sub-dataset B is scaled and assumed present in the entire dataset. A better description of the spatial continuity could be obtained by: (1) separating the domain into two parts, that is, the sub-dataset A and B, (2) calculating and fitting an experimental variogram that accounts for the majority of the domain and correct the conditional distributions of the rest of the domain accordingly. The next section proposes an approach to identify minor structures that affect the experimental variogram. The cost of working directly in a SRF environment is the generalization of the spatial variability, thereby making it difficult to reproduce local features.

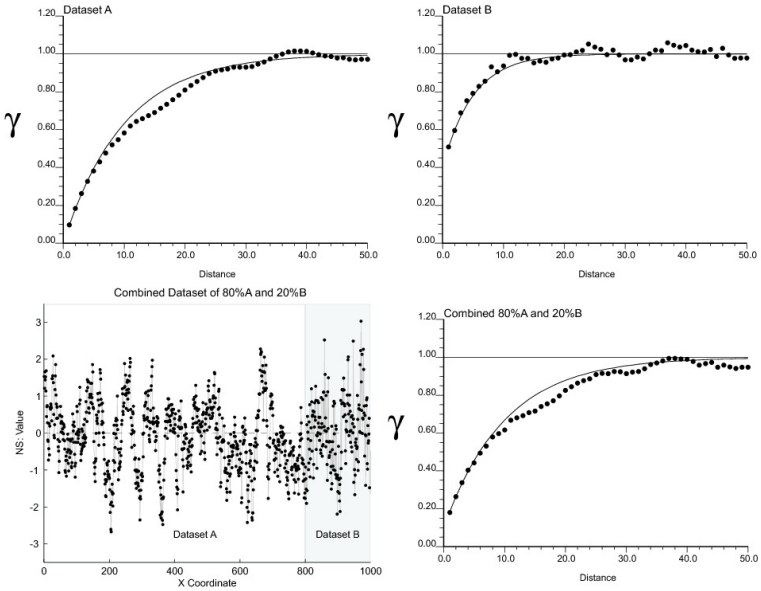

Figure 4: Experimental variogram of a unconditional realization using an exponential variogram model with range 30u and C_0=0 (top left), with range 15u and C_0=0.4 (top right), combined dataset of 80% of dataset A and 20% of dataset B (bottom left), and its experimental variogram (bottom right)

Experimental Variogram Outlier Data Pairs

Outliers are unusual observations that do not appear to belong to a nearby pattern of variability. However, not all the outlier values are wrong numbers; they can be considered as information that could lead to a better understanding of the phenomena being studied (Johnson & Wichern, 2007). The experimental variogram of a non-stationary domain presents anomalies when compared to the analytical distributions in a multi-gaussian environment. Such anomalies or outliers can be used to identify patterns in the domain that affect the non-stationary features of the experimental variogram.

This approach aims to identify outliers in the experimental variogram by defining outlier limits based on the proposed variogram model and a probability interval. The outliers are considered to belong to another type of spatial behavior that affects the experimental variogram of the majority of the domain. Once identified, the decision to model them separately or to correct their variability can be made. Two approaches for identifying outliers are discussed in this section.

Control Limit Ellipses

In a multivariate gaussian framework, control limits are used to verify the stability of processes and identify occurrences of special cases of variation that are unlikely part of the process. They make the special variations visible and allow distinguishing them from the common ones in the process (Johnson & Wichern, 2007). Because of dimensionality, control limits are used in the form of charts for multivariate datasets. However, in a two dimensional case a control limit ellipse can be used. The control limit ellipse is defined as a contour of constant density of the bivariate standard normal distribution for a given confidence control value. It can be expressed using the generalized form of the ellipsoid of constant density of *p*-dimensions (3) (Johnson & Wichern, 2007).

$$\left(\mathbf{x}-\boldsymbol{\mu}\right)^{T}\mathbf{C}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)=\chi_{p}^{2}\left(\alpha\right) \tag{3}$$

where $\mathbf{x}$ is a vector of observations on different variables, $\boldsymbol{\mu}$ is the vector of means of $\mathbf{x}$, $\mathbf{C}$ is the covariance matrix, p is the number of dimensions or elements in $\mathbf{x}$, and $\chi_p^2(\alpha)$ is the upper (100α)-th percentile of a chi-square distribution with p degrees of freedom. Expression (3) gives the contour for the multivariate case that represents $(1-\alpha)100\%$ of the probability.

The experimental variogram is analyzed on an **h**-scattergram basis. Analytically the distribution of data pairs on each **h**-scattergram is bivariate normal. Therefore, the control limits are set up to the two dimensional case. Making x_1 be the data values at location $\mathbf{u}$ and x_2 the data values at location $\mathbf{u}+\mathbf{h}$, and for simplicity, assuming both distributions are standard normal. Then the expression of the control limit ellipse is (4).

$$\frac{x_1^2-2\rho x_1x_2+x_2^2}{1-\rho^2}=\chi_2^2\left(\alpha\right) \tag{4}$$

However, an analytical model is required that provides the parameters of such bivariate distributions. That information is supplied by the variogram model fitted to the experimental variogram. The control limit ellipses are used to identify outlier data pair(s) according to the fitted variogram model on an **h**-scattergram basis. For a probability of $(1-\alpha)100\%$, the proportion of data pairs that are expected to fall outside each control limit ellipse of the h-scattergrams is $(\alpha)100\%$. In the same way, the proportion of the outlier data pairs for the entire variogram model is also $(\alpha)100\%$.

For the case presented, the geometry of the control limit ellipse is a function of the correlation coefficient and the probability confidence limit. The control limit ellipse is rotated so that the major axis is in the direction of the first bisector of the h-scattergram. The length of the major and minor ratios is (5), (6) (Johnson & Wichern, 2007).

$$rt_{mj}=\chi_2^2\left(\alpha\right)\sqrt{1+\rho^2} \tag{5}$$

$$rt_{mn}=\chi_2^2\left(\alpha\right)\sqrt{1-\rho^2} \tag{6}$$

Notice that for the same probability interval, the length of the major ratio is proportional to the correlation coefficient (see Figure 5). This is a problem for analyzing the spatial continuity, since data pairs placed close to the first bisector that are considered as good observations for high correlation coefficients

would become outliers for small correlation coefficients. It is contradictory to the definition of the variogram which is a measure of dissimilarity. The smaller the correlation coefficient, the more tolerant the variogram to outlier values is expected to be. In Figure 5, consider the data pair P with $x_1 = x_2 = 2.2$. Notice that the data pair P lies within the acceptable region of three out of four control limit ellipses shown in the example. Even when the values of the data pair P are equal some control limit ellipses will identify it as an outlier. An alternative approach is required that identifies outlier limits based on the increments of the data pair values $z(\mathbf{u}) - z(\mathbf{u}+\mathbf{h})$ rather than its position on the **h**-scattergram.

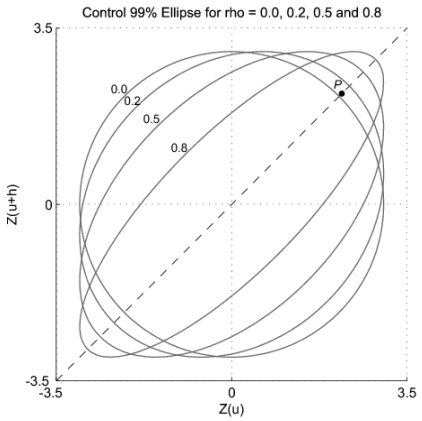

Figure 5: Control limit ellipses for 99% of bivariate standard normal distribution for four different values of correlation coefficient; the data point P is placed over the first bisector and is evaluated in the four cases.

Confidence Limits of the Distribution of Experimental Variogram Data Pairs

Under the assumption of multi-gaussianity the data pair increments, $z(\mathbf{u}) - z(\mathbf{u}+\mathbf{h})$, follow a univariate normal distribution and the square of the increments $[z(\mathbf{u}) - z(\mathbf{u}+\mathbf{h})]^2$ a scaled chi-square distribution with one degree of freedom $2\gamma(\mathbf{h})\chi_1^2$ (Cressie & Hawkins, 1980). The previous statement can be re-expressed as $[z(\mathbf{u}) - z(\mathbf{u}+\mathbf{h})]^2/2 \sim \gamma(\mathbf{h})\chi_1^2$ or in terms of the orthogonal distance of the data pair to the first bisector $d^2(\mathbf{h}) \sim \gamma(\mathbf{h})\chi_1^2$ (see Figure 6). Recall the mean of a chi-square distribution is the number of degrees of freedom, in this case 1, so the expected value of the orthogonal distances is the variogram $E\{d^2(\mathbf{h})\} = \gamma(\mathbf{h})$. Similar to the control limit ellipses approach, extreme values of $d^2(\mathbf{h})$ can be discriminated using confidence limits. This way the increments of the data pairs are accounted for directly.

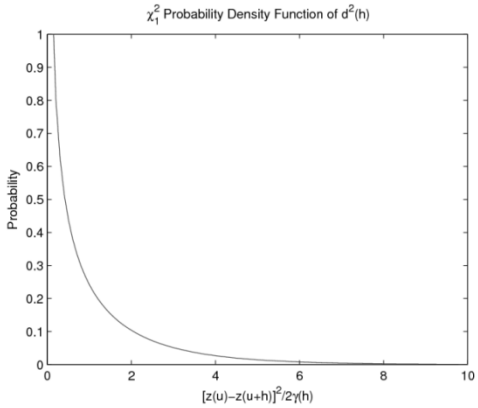

Figure 6: χ_1^2 Distribution of the orthogonal distances from data pairs to the first bisector

Experimental Variogram Calculation with Confidence Limits

The experimental variogram is analyzed in the form of a cloud variogram using confidence limits. The variogram model defines the ideal conditions of how the conditioning data should be. By comparing the

ideal conditions to the dataset that is sampled from the real domain, the patterns that are produced by the geologic features can be identified.

Each data pair consists of two data values separated by a vector. Therefore, there is a reference value of variogram model. Then, each data pair can be evaluated independently by its corresponding outlier limit regardless of the tolerances used to calculate the experimental variogram. The outlier limit is the maximum squared orthogonal distance for a confidence limit (7), where the probability of the chi-square distribution α defines the $(1-\alpha)100\%$ probability of occurrence of the data pairs:

$$d_{MAX}^2(\mathbf{h}) = \chi_1^2(\alpha)\gamma(\mathbf{h}) \tag{7}$$

The data pairs that are outside of the limit, that is $d_i^2(\mathbf{h}) > d_{MAX}^2(\mathbf{h})$, are marked as outliers for the given parameters of confidence limits α and variogram model $\gamma(\mathbf{h})$. Even when a data pair is marked as an outlier, it does not imply that the corresponding head and tail values are outliers too. Particular patterns of data pairs in the cloud variogram provide evidence of the geologic structures present in the available dataset due to the nature of the domain. This makes it possible to assess whether additional sub-domaining is required or find a way to reproduce these local patterns in the geostatistical model. For instance, an abrupt change in metal grades at a short distance could be an indicative of the presence of veins or some other type of small structures in the domain. In mining such structures are of special importance because they can define whether some regions are of economic interest or not. Therefore, they impact directly the economic potential of the mineral deposit.

Once the outlier data pairs are identified, the experimental variograms with and without the outlier data pairs should be compared to verify their impact in the variogram analysis. If there is a notorious impact in the experimental variogram after removing the outlier data pairs, it is recommended to be fitted by a new variogram model; otherwise, that would mean the initial proposed variogram model is fairly representative of the dataset.

The continuous form of the outlier limits can easily be shown for a 1D dataset in a cloud variogram plot. Recall that each data pair is plotted as $0.5[Z(\mathbf{u}_i) - Z(\mathbf{u}_i + \mathbf{h})]^2$, that is $d_i^2(\mathbf{h})$. Since $d_{MAX}^2(\mathbf{h})$ is a continuous function it can be plotted with respect to the variogram model for the different confidence limits and for any lag distance (see Figure 7). For a multi-gaussian dataset the density of data pairs outside each control limit is uniformly distributed. However, for real dataset particular patterns or clusters of data pairs are present.

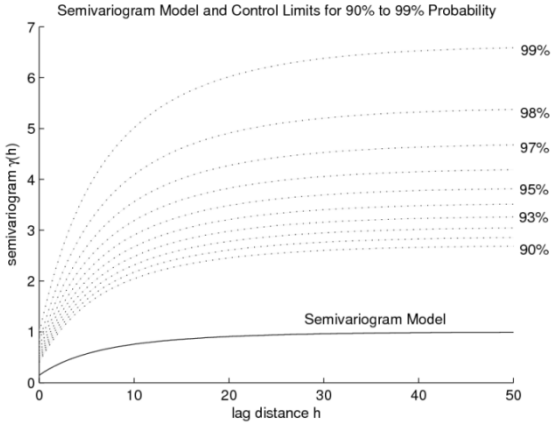

Figure 7: Control limits for different probabilities, from 90% to 99% (dotted line) with respective variogram model (solid line)

The style of variogram fitting (see Section 4.1) has an impact on the process of cleaning the experimental variogram. In practice, the style of fitting a variogram model is subjective. The style of fitting is not applied to the entire experimental variogram, but to segments of it. Overall, the experimental variogram is recommended to be fitted to the entire experimental variogram as close as possible. The consequences of the styles of fitting are:

- *Pessimistic*: The correlation coefficients of the **h**-scattergram are larger than expected, so outliers are more difficult to identify. The theoretical variability is larger than the experimental variogram. The limit $d_{MAX}(\mathbf{h})$ tends to be larger.
- *Fair*: The correlation coefficients of the **h**-scattergram try to reproduce the variability of the dataset. This is a good condition to identify outlier data pairs, since it directly compares the analytical to the experimental form of the spatial variability. The limit $d_{MAX}(\mathbf{h})$ is representative.
- *Optimistic:* The correlation coefficients of the **h**-scattergram are smaller. As a consequence, many data pairs are identified as outliers. The limit $d_{MAX}(\mathbf{h})$ tends to be smaller.

Case Study

Consider the real case presented in the first part of this document. The dataset corresponds to a borehole of 162 samples regularly spaced in normal score units (see Figure 2 right). Therefore, there is no influence of tolerances or clusters in the experimental variogram. The only source of influence comes from the structural patterns in the dataset. The variogram model used for the exercise is given in (8) and shown in Figure 8.

$$\gamma(\mathbf{h}) = 0.15 + 0.35Exp_{12.5}(\mathbf{h}) + 0.50Exp_{35.0}(\mathbf{h}) \tag{8}$$

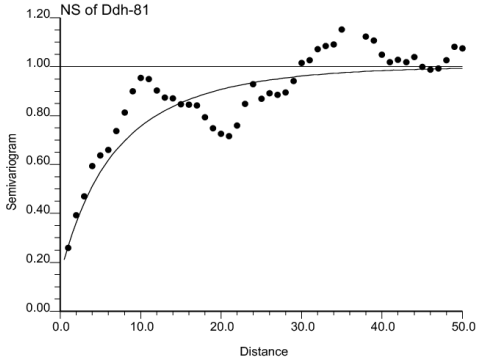

Figure 8: Experimental variogram (black dots) and variogram model (solid line) of Ddh-81 dataset

The selection of the probability limit is based on identifying the smallest number of data pairs as possible that significantly impact the experimental variogram. Several control limits were applied to the experimental variogram of the dataset Ddh-81 at different ranges in order to identify special patterns in the data pairs (see Figure 9). Figure 9 (A) considers a range of probabilities from 90% to 99%; based on this range, two patterns can be seen: the first one from lag distances 0 to 18 and the second one from 35 to 38. The range of outliers is reduced to an interval from 97% to 99.5% in Figure 9 (B). The first part from the lag distance 0 to 20 becomes more pronounced while the second part recedes. Reducing the range of outliers even more (Figure 9 (C)) from 99.5% to 99.9%, the number of data pairs in the first part is still dominant, while the second part virtually disappears. Finally, the limit that identifies the outliers of the first part is chosen (Figure 9 (D)) that corresponds to a 99.7% probability. This removes 0.6% of data pairs (or one data pair) of the experimental variogram which is not a significant proportion of the information provided by the dataset.

The data pairs removed from the cloud variogram clearly makes a particular pattern (see Figure 10 top left). Including the influence of this sub-region in the variogram model causes an inflation of the conditional variance to the rest of the domain. On the other hand ignoring such variability means the conditional variances of the small sub-region are unaccounted. Depending on the spatial configuration and size of the sub-region two decisions can be made: (1) split the domain so that two domains can better account for different patterns of variability, or (2) focus attention on the major part of the domain and ignore the variability of the sub-region and correct the pattern of variability locally.

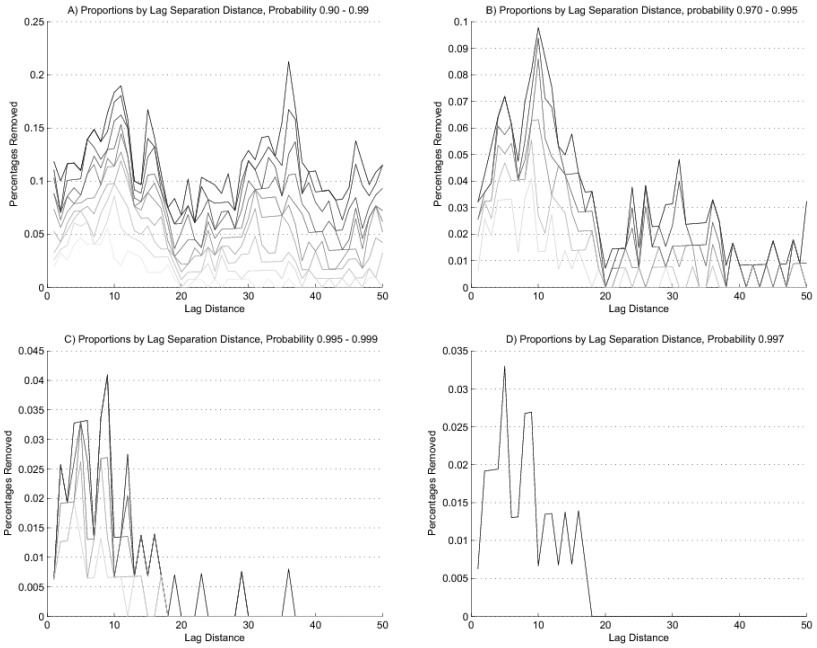

Figure 9: Proportions of data pairs identified as outliers for different ranges of control limits for the experimental variogram of the dataset Ddh-81 in normal score units.

Notice that after removing the 0.6% of the outlier data pairs there are significant changes in the experimental variogram. The fixed experimental variogram is more continuous with fewer jumps (see Figure 10 bottom). Therefore, the fitting of the experimental variogram tends to be more straightforward (9). This can be seen in the reduction of the mean squared error (MSE) (see Figure 10 top right). This, of course, does not change the subjectivity of the fitting process; it merely reduces the trade-offs that the modeler must consider while fitting a variogram model.

$$\gamma(\mathbf{h}) = 0.12 + 0.37 Exp_{20}(\mathbf{h}) + 0.51 Exp_{38}(\mathbf{h}) \quad (9)$$

The second part of the analysis consists of verifying whether the outlier data points are grouped in certain specific parts or are spread around the entire domain. If they are placed over particular regions or follow clear patterns, then this will increase our knowledge of the nature of the domain. A decision of sub-domaining can be considered here. On the other hand, if the data points are dispersed throughout the domain, then this could mean the identified variability is part of the domain and it can be modeled conventionally. For the data set Ddh-81, the data points that produce extra variability are located in a specific sub-region. They can be considered for sub-domaining in order to improve the estimation/simulation of the domain (see Figure 11).

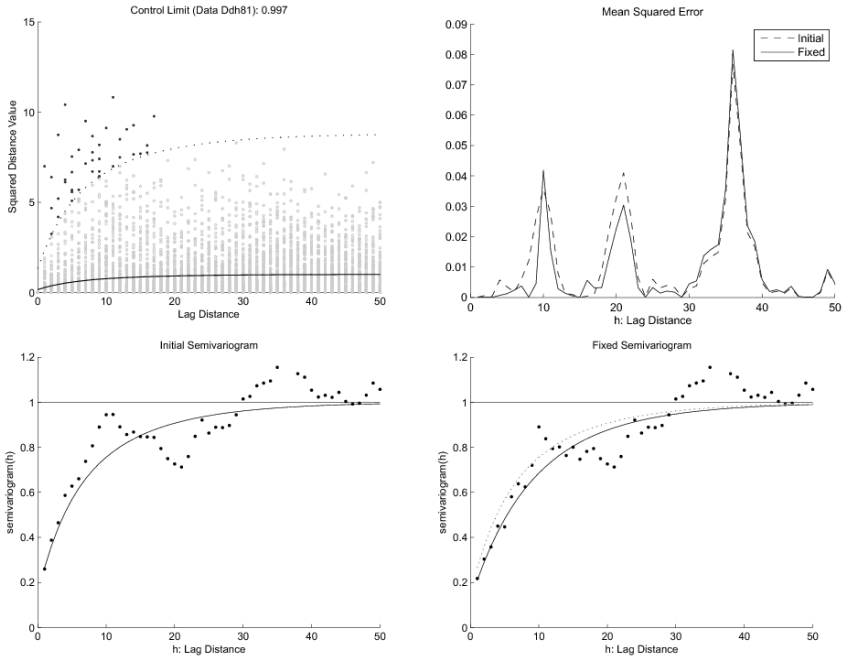

Figure 10: Cloud variogram with control limit at 99.7% (top left), variogram model fitting MSE for initial (dashed line) and fixed (solid line) experimental variograms (top right), initial experimental variogram (black dots) and variogram model (solid line) (bottom left) and fixed experimental variogram (black dots) and its respective variogram model (solid line), and initial variogram model (dotted line) (bottom right)

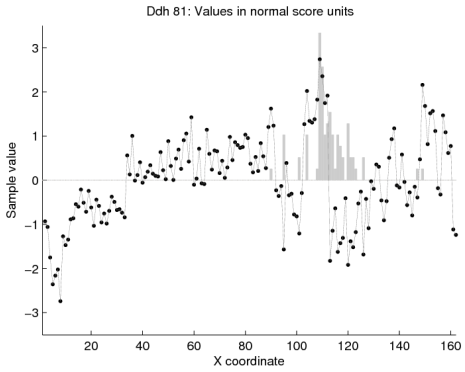

Figure 11: Occurrences of data points of outlier data pairs (gray bars) compared with the input dataset Ddh-81 (black dots) in normal score units.

References

Cressie, N., & Hawkins, D. (1980). Robust Estimation of the Variogram I. *Mathematical Geology* , 115-125.

Deutsch, C. V., & Journel, A. (1998). *GSLIB Geostatistical Software Library and User's Guide.* New York: Oxford Press.

Genton, M. (1998). Highly Robust Variogram Estimation. *Mathematical Geology* , 213-221.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* New York: Oxford Press.

Johnson, R., & Wichern, D. (2007). *Applied Multivariate Statistical Analysis.* New Jersey: Pearson Prentice Hall.

Journel, A., & Huijbregts, C. J. (1978). *Mining Geostatistics.* New York, USA: The Blackburn Press.

Larrondo, P., & Neufeld, C. (2003). *VARFIT: A Program for Semi-Automatic Variogram Modelling.* Edmonton: Centre of Computational Geostatistics.

Sinclair, A., & Blackwell, G. (2004). *Applied Mineral Inventory Estimation.* Cambridge University Press.