Conditional Distribution Fitting of High Dimensional Stationary Data

Miguel Cuba and Oy Leuangthong

The second order stationary assumption implies the spatial variability defined by the variogram is constant along the domain. Simulated models have this characteristic by construction. On the other hand, real geologic domains have internal structures as a result of geologic processes that formed the mineral deposit. The spatial variability of highly variable regions within a domain is not accounted by the variogram model because it is an averaged representation of the spatial variability of the domain. In those regions the conditional distributions are often underestimated, that is, locally the simulated model is less variable than in reality. Even when such situation can be verified, the conventional simulation approach (SGS) cannot account for local variable patterns. The multi-gaussian assumption that is a requisite for simulating with SGS is very restrictive to local modifications. The present document proposes an approach for standardizing the local variability of a domain in order to make the variogram model do not under estimate conditional distributions in the domain. This is done by moving the domain from its original dimensional space to a higher dimensional space. The conditional distributions are tested using the cross-validation methodology.

Introduction

The variogram can be used as a tool for identifying the non-stationary features in the domain related to the intrinsic assumption of the SRF. Two scenarios remain from the variogram analysis: 1) the problematic locations are grouped in sub-regions, so that they can be considered for sub-domaining and 2) the problematic locations are in small groups dispersed throughout the domain such that sub-domaining cannot be performed. This document focuses on the latter scenario and proposes an alternative method to account for the influence of these data in the domain so that the local uncertainty calculated using estimation/simulation can be used for a detailed mine plan strategy.

Sequential Gaussian simulation (SGS) is recommended to be used for simulating the variable of interest in a domain due to its simplicity in the implementation and availability in many commercial mining software packages (e.g., Pangeos, Vulcan, MineSight, and Datamine, among others). The implementation of SGS is very similar to simple kriging; both use the same parameters such as the variogram model and search specifications. For implementation it is necessary to assume a multi-gaussian environment for SGS since all simulation is performed in Gaussian space. Although SGS generates multiple realizations of the attribute, it can be verified against an SK model since the local average from SGS, taken over many realizations, will tend to SK estimates. The problem with SGS and SK is that it assumes that the conditioning data is part of a SRF, that is, it assumes there are no outlier increments or the variogram model adequately defines both local and global uncertainty. For this reason and in these highly variable regions, SGS does not account for the local features of the domain and is not used for simulating medium/short term models. The goal of this reserach is to provide a methodology to prepare the dataset for performing SGS, so that it reproduces the non-stationary features required for mine planning.

In general, only high values are considered problematic because a more pessimistic estimation is preferred over an optimistic one. In some cases, outliers are the result of an erroneous characterization of the geology (e.g., presence of samples that belong to different geologic processes other than their tagged category). In Figure 1 a sketch of reality is compared to a tentative geologic interpretation; due to the scale of the interpretation some samples of the high metal grade region are classified as part of the low grade metal domain. Estimating such domains using kriging in a strict manner would smear the high grade value over a considerable region in the low grade domain. As a consequence, the geostatistical model of the low grade domain is neither realistic nor appropriate for proposing a detailed mine plan. In a more general context, the presence of sub-patterns in the domain makes the available dataset behave as a SRF with a spatial continuity cannot be adequately defined by one variogram model.

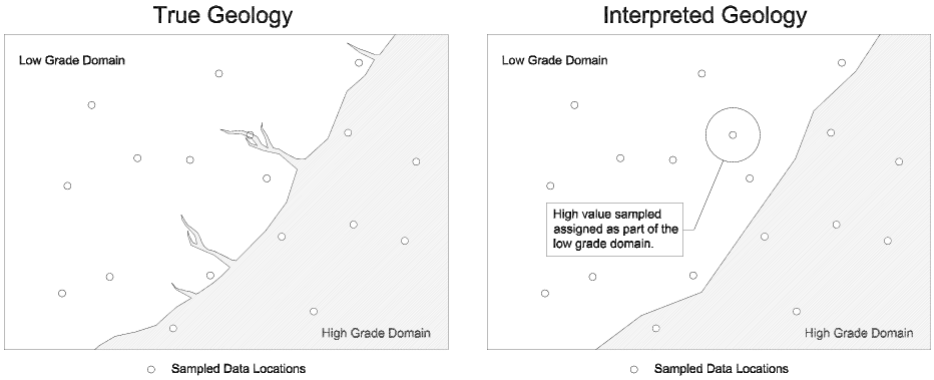

Figure 1: Impact of generalization of geology due to the scale of geologic interpretation; reality (left) is not fully characterized when models are built (right).

In this document the influences of trends in the mean and in the variance are considered to be absent. However, non-stationarity in the domain occurs when the intrinsic hypothesis[1] of the SRF is not satisfied. Because of this, the estimated model appears locally unrealistic. Internal structures of the domain are present as patterns in the dataset which are not necessarily removed by normal score transformation. The presence of such patterns in the domain makes the decision of stationarity less appropriate.

Proper estimation of the conditional distribution is important for simulation. The farther the true value falls from the confidence limits, the more difficult it is for the simulation to pick a realization that reproduces the true value at such a location and globally reproduces the sub-patterns in the domain. Ideally, a mine plan is based on the analysis of the geologic region to be mined; the impact of the geologic characteristics in the variable of interest is an important input in the process of decision making in mine planning.

Depending on the variogram model fitted, the degree of accuracy of the conditional distributions changes. A good variogram model is one that accounts for the spatial variability of the major part of the domain. The remaining spatial variability that is not properly accounted for by the variogram model is under-estimated in some regions and over-estimated in others. In real situations, it is unrealistic to expect that any variogram can account for the geologic features of a domain like it may under theoretical conditions. The approach presented in this document aims to account for the variability of this minority of the domain, where the spatial variability is under-estimated by tuning the distances between the samples, so that the conditional distributions are consistent with the conditioning information. The distances are modified by adding an extra dimension to the dimensional space of the conditioning dataset. In this way, the influence of the variogram model on the estimated parameters of the conditional distributions is approximated until the parameters are consistent with the dataset within some confidence limits. The goodness of the conditional distributions is verified using cross validation and confidence limit parameters.

Measure of Accuracy

Cross validation is used as a technique to test the quality of the estimated parameters of the conditional distribution with respect to the true values (Armstrong & Jabin, 1981). In this document, it is assumed that if the verified conditional distributions using cross validation properly account for their corresponding true values the conditional distributions of the rest of the unsampled locations will also do the same.

For all the data locations, confidence limits are considered as a measure of whether the true value with respect to its estimated conditional distribution is predicted within some tolerance intervals or not (see Figure 2). If the true value falls outside the confidence limits, then the proposed variogram model is considered to be inadequate and this location is flagged for pre-processing. All the conditional

[1] The expected value of the variable of interest exists $E\{Z(\mathbf{u})\} = m$, and the variance of the increments is assumed not to be a function of the position vector $\mathbf{u}$, but of $\mathbf{h}$, $Var\{Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\} = 2\gamma(\mathbf{h})$ (Journel & Huijbregts, 1978).

distributions where true values are within the confidence limits are assumed to be consistent with the surrounding information. Theoretically, if 90% probability interval is chosen then 10% of the dataset is expected to fall outside the confidence limits. For this, one basic condition is the dataset is truly part of a SRF realization, recall that a real dataset is non-stationary. Contrary to the theoretical conditions, the proposed approach forces all the data point values to fall within the confidence interval in order to ensure the realizations reproduce the sub-patterns in the domain.

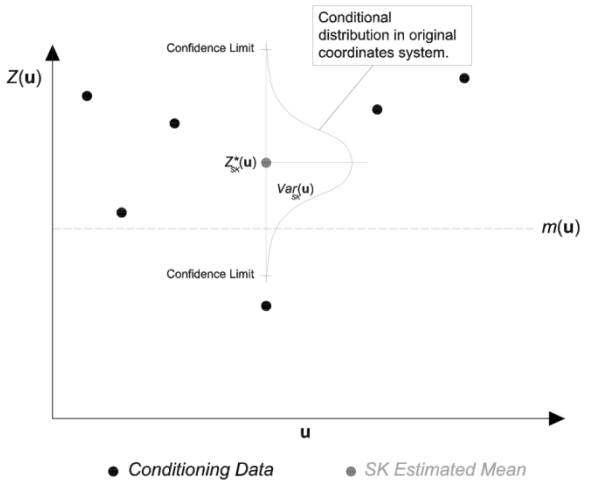

Figure 2: Sketch of cross validation, the true value (black dot) falls outside of the confidence limits of the conditional distribution (gray lines) calculated using the rest of the information and the proposed variogram model.

The accuracy of the estimates is measured by standardizing to one the distance from the estimated mean to any of the confidence limit values, therefore, if the true value falls outside the confidence limits the standardized distance is greater than one, and the conditional distribution is considered as improperly accounting for the data. Assuming the global univariate distribution of the domain is standard normal, then for the data values that are outside the confidence limits of the standard normal distribution, the standardized accuracy is re-scaled making the true values be the new confidence limits. This is done to ensure that for all values in the dataset there is a conditional distribution that includes the true values within reasonable confidence limits.

Dimensional Conditional Distribution Fitting

The variogram model $\gamma(\mathbf{h})$ is a function of the separation vector $\mathbf{h}$ that fully defines the spatial continuity of a SRF. The kriging estimated parameters of the conditional distribution at the unsampled location are functions of this variogram model. The variogram model is used to calculate the linear dependence between data values and the unsampled location in the kriging system. The smaller the distance from the conditioning data to the unsampled location, the smaller the conditional variance. Regardless of the values of the conditioning data the surrounding spatial configuration of the unsampled location is what really matters for calculating the conditional variance. The estimated mean tends to be similar in value to the closer surrounding data values.

Since the spatial covariance $C(\mathbf{h})$ is a function of the separation vector $\mathbf{h}$ the estimated parameters of the conditional distribution can be manipulated by modifying the separation distances from the unsampled location to the conditioning data. The influence of the conditioning data for estimating the parameters of the conditional distribution at the unsampled location decreases proportionally as the unsampled location is separated further from the data. The unsampled location becomes gradually more uncertain until the parameters of the conditional distribution are equal to the global distribution parameters, which occur when the influence of the conditioning data is negligible. When there are no samples within the effective range of the variogram model the estimated mean will equal the global mean $Y_{SK}^{*} \cong m$ and the estimated variance will equal global variance $\sigma_{SK}^{2} \cong \sigma^{2}$ this is the state of local maximum uncertainty.

A small example is shown in Figure 3 to illustrate the sensitivity of the conditional distribution with respect to the position of the unsampled location. Consider a four samples dataset $\sim N(0,1)$ and a location to estimate in 1D (see Figure 3-bottom). When the unsampled location is separated gradually from its original position by adding a new dimension the conditional distribution changes as the influence of the conditioning data diminishes (see Figure 3-top right). The change is gradual until the conditional distribution equals to the global distribution of the dataset, in this case $\sim N(0,1)$. For the example presented a nested exponential model is used (see Figure 3 top left, equation (1)). In Figure 3-top right notice the SK mean reaches 0 for the effective variogram range and the SK variance to the sill value of 1.0 for the effective range. Both parameters approach the global values asymptotically because an exponential model is used.

$$\gamma(\mathbf{h}) = 0.25Exp_{1.5}(\mathbf{h}) + 0.75Exp_{3.0}(\mathbf{h}) \quad (1)$$

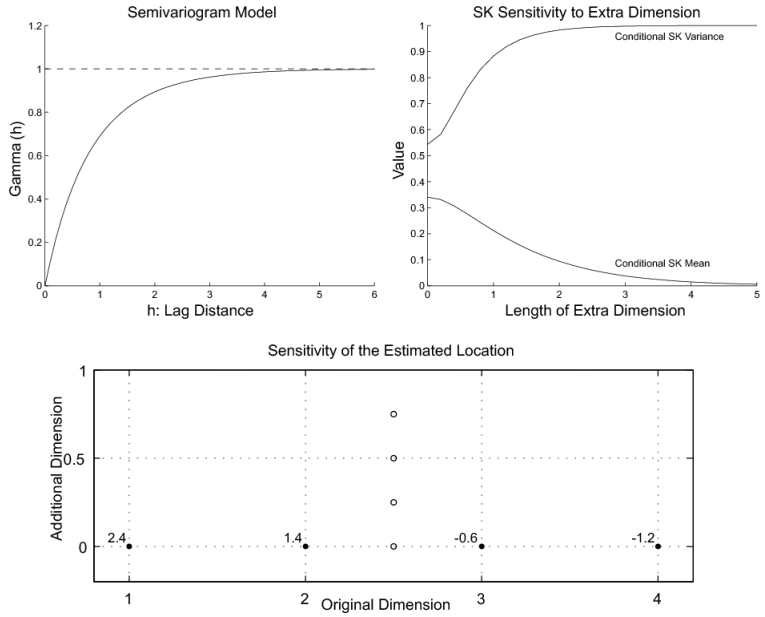

Figure 3: Nested exponential variogram model (top left), sensitivity of SK variance and SK mean to the inclusion of additional dimension to the original position of the unsampled location (top right) and a sketch of spatial configuration of conditioning data (black dots) and locations of the unsampled location (empty dots) to different lengths of the additional dimension (bottom).

The condition that the separation distances $d_{u-u'}$ between the unsampled location $\mathbf{u}' = [x'_1, x'_2, \ldots, x'_n, y'_k]$ and the conditioning data $\mathbf{u}_j = [x_{1j}, x_{2j}, \ldots, x_{nj}, 0]$ do not decrease is guaranteed by the additional dimension $y'_k \; \forall \; y'_k \neq 0$. This additional Cartesian component y'_k at the unsampled location is always equal or greater than zero, so it is additive when the separation distance is calculated, $d'_{u_j-u'} = \left[\sum_{i=1}^{n}(x_{ij} - x'_i)^2 + (y'_k)^2\right]^{1/2}$ therefore $d'_{uj-u'} \geq d_{u-u'}$.

The SK variance is said that it cannot be used as a measure of local variability or accuracy since it is based on a variogram model which is a global approximation of the spatial continuity (Journel, 1986) or because SK variances are independent of the data values and only provide a comparison of an alternative data configuration (Deutsch & Journel, 1998). The uncertainty assessment of SK fully relies on the assumption that the conditioning data is a realization of a SRF, with two requisites: (1) the distribution of errors is gaussian, and (2) the variance of errors can be predicted (Isaaks & Srivastava, 1989). Let us consider 1000 data points of an unconditional simulation where each cross validation conditional distribution is evaluated using confidence limits with respect to the corresponding true values for different probability confidence intervals. The proportions of samples within the confidence limits are similar to the theoretical expected proportions since the dataset is a realization of a multi-gaussian SRF (see Figure 4). Under correct conditions the SK conditional distributions account for local uncertainty properly. These two

conditions are (1) the variogram is known and (2) the conditional error distribution is calculated with no additional influence of any source of error, (Chilés & Delfiner, 1999). This is not the case of geologic processes where there is no true variogram. In fact, in practice the variogram fitting is based on the experience of the person in charge and on the objective of the study (Goovaerts, 1997).

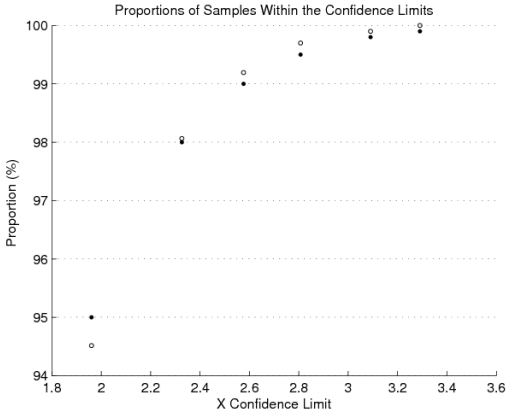

Figure 4: Proportions of true values within the confidence intervals (empty dots) compared to the theoretical proportions (black dots) of their respective cross validation conditional distributions. The true values are from an unconditional realization.

Under the assumption of multi-gaussianity the conditional distributions estimated by SK are univariate Gaussian $\sim N(Y_{SK}^{*}, \sigma_{SK}^{2})$. All values are plausible outcomes to occur depending on their probability, even when they are very extreme values. Let us consider a value at a certain location has a probability of 1×10^{-1000} to occur in a conditional distribution calculated using cross validation. Statistically the true value is a valid outcome of the conditional distribution. For a mineral deposit model used for economic decision, such an estimate could have serious consequences if it is considered as a valid result. For the small case presented the true value is smaller than the SK mean and when a realization is drawn at such location due to its very small probability the true value is very unlikely to be simulated. If the true value is unknown there is no way to verify the validity of the estimated conditional distribution. However, problems arise when the true value is known and its cross-validation conditional distribution does not account for it properly. Then when the surrounding locations are estimated, even though the SK means tend to average the big difference in values of the conditioning data, the conditional variances still remain smaller because the variogram model does not account for the true value properly.

The proposed approach tunes the conditional distributions to the conditioning data within some confidence limits under the assumption the conditioning dataset is representative of the domain. Using cross validation, once the conditional distributions account properly for their respective true values the resulting models is assumed to account for local and global uncertainty properly. Since the local conditional distributions are tuned to the dataset using additional dimensions the new spatial configuration of the conditioning information is called *proper stationary state* of the dataset. Some geologic features present in the dataset that were not captured by the variogram model initially are now explained by the additional dimensions. The additional geological information is added to the dataset in the form of position vectors.

Cost and Benefit of the Conditional Distribution Fitting

As mentioned before, the main goal of this reserach is to make the conditional distributions account properly for the local uncertainty of the realizations. Modeling conditional distributions is considered an ambitious goal and sometimes unrealistic to achieve without a specified theoretical model (Chilés & Delfiner, 1999). By adding extra dimensions there are many possible solutions to this problem. Many different configurations could give different degrees of fitting of the conditional distributions with respect to their true values. There would be some negative impact on the accuracy of the surrounding data locations of the fixed data location. This problem can be solved considering additional restrictions in the

algorithm such as reducing the negative impact on the accuracy of the rest of the samples, reducing the negative impact on the SK means or by trying to use a small number of additional dimensions, etc.

Using small probability intervals could be very restrictive for this approach. The smaller the probability interval the more extra dimensions are necessary to fit the conditional distributions. By making all the data values fall within the confidence limits the sense of probability loses meaning because it does not comply with the theoretical conditions. Theoretically for a 95% confidence interval 5% of the true values is expected to fall outside their respective conditional distributions (see Figure 4), while the approach tries to eliminate such proportion of data values. Setting up the confidence interval is a subjective part of this approach. In mining a 95% probability is commonly used (Journel & Huijbregts, 1978). A real example of a Chilean copper mine (Chuquicamata) is presented in (Journel & Huijbregts, 1978) where 96% of the observed errors of mean block grades fall within the 95% interval. On other types of deposits such as skarn type where the grade variability is high and more geologic structures are present such result would be very difficult to obtain because of the complexity of the geologic environment. When the distribution of errors is non-gaussian but continuous and unimodal a confidence interval of $\pm 3\sigma_{SK}^2$ which correspond to 99.73% probability interval is preferable, see discussion in (Chilés & Delfiner, 1999). The distribution of errors is assumed gaussian for this approach.

There is a set of widely used variogram models which are present in many mining commercial packages, such as spherical, exponential, gaussian, etc. which are licit models considering dimensional spaces up to $\mathbb{R}^3$. In mining it is very unlikely to deal with data in higher dimensions than $\mathbb{R}^3$. By adding extra dimensions to the conditioning dataset it is highly probable the dimensional space increases to $\mathbb{R}^n$ with $n > 3$ and therefore some of the variogram models valid in $\mathbb{R}^3$ would end up not being licit models in such higher dimensions. The problem of using a non licit model is the possibility of get negative conditional variances. The conditional variance is a linear combination of the covariances $Var\{Y^*(\mathbf{u}_0)\} = \sum_{\alpha=0}^{n}\sum_{\beta=0}^{n}\lambda_\alpha\lambda_\beta C(\mathbf{u}_\alpha - \mathbf{u}_\beta) \geq 0$ and must be non negative (Goovaerts, 1997). To ensure this the covariance must be positive semi-definite and/or the variogram negative semi-definite (Goovaerts, 1997). The number of dimensions of the space is important for choosing a variogram model, a positive definite covariance function in $\mathbb{R}^m$ is also positive definite in $\mathbb{R}^n$ if $m \geq n$, however it is not necessarily valid for $m < n$ (Chilés & Delfiner, 1999). Two examples are presented in (Armstrong & Jabin, 1981) where it is shown that a variogram model would lead to negative conditional variances. The set of covariance models that can be used for the proposed approach are reduced to the ones that are positive-definite in any dimensions. There is a variety of variogram models that are proved are licit in any dimensions, some of them are:

- Spherical models based on a sphere of $\mathbb{R}^n$. The spherical variogram model without any specification of the dimension n is commonly referred to as the spherical model in $\mathbb{R}^3$, also there are other well known covariance models such as spherical in $\mathbb{R}^2$ is known as the circular model, and in $\mathbb{R}^1$ the triangle model (Chilés & Delfiner, 1999).
- Exponential models are positive definite in $\mathbb{R}^n$. Also Radon transforms of the exponential covariance provide differentiable covariances that are also valid in (Chilés & Delfiner, 1999).
- Gaussian model with a scale parameter greater than zero (Chilés & Delfiner, 1999).
- Generalized Cauchi model (Chilés & Delfiner, 1999).
- K-Bessel model (Chilés & Delfiner, 1999).
- Logarithmic or de Wijsian model (Chilés & Delfiner, 1999).
- Stable model. The exponential and gaussian models belong to this family (Chilés & Delfiner, 1999).
- Matérn model (Rasmussen & Williams, 2008).

Algorithm for Conditional Distribution Fitting

The only additional parameter different from conventional practice is the definition of a confidence interval. However, the intrinsic hypothesis is not assumed in this approach even when a variogram model is proposed. The variogram model has to account for the spatial continuity of the major part of the domain, not an average as in conventional practice. The part of the dataset which spatial variability is under-estimated by the variogram model is corrected by the conditional distribution fitting so that the

variogram model accounts for the entire domain in the fitted higher dimension. However, the region where the spatial variability is over-estimated still remains the same. This is equivalent to a pessimistic fitting of the variogram model that makes the conditional distributions be properly accounted but overestimated. And a consequence is the geostatistical model is more uncertain than it should be.

The proposed algorithm can be considered as a prototype of a non-conventional geostatistical modeling which is aimed for the requirements of mining industry. Many different additional strategies of fitting the conditional distributions can be added according to other requirements in the model, such as maximizing the accuracies or maximizing the fitting of the conditional means, etc. The algorithm is presented as a workflow and each of the steps are discussed:

1) **Cross validation and verification of the input parameters.** The goodness of the reproduction of the conditional distributions is verified when compared to their respective true values. The data locations where the true values fall outside the confidence limit parameters are marked for conditional distribution fitting. If the variogram model is fitted in a pessimistic manner a very small amount of the conditional distributions will require to be fitted; conversely, when the variogram model is fitted in an optimistic manner a very large amount of conditional distributions will require fitting. This is why it is important for the variogram model to account for the spatial continuity of the major part of the domain, so that, only an optimal number of locations require fitting. Also, the selection of the confidence interval influences in the proportion of samples to be fitted, it has to account fairly for the data values of the conditioning data.
2) **Verification of spatial relationship among the marked samples.** It would the case a group of the marked samples are part of secondary populations. The verification is based on a cross validation analysis using only the marked samples. The mutual samples that estimate the parameters of conditional distributions that account for the true values within the given confidence intervals are grouped. Finally for each group of samples and independent samples different dimensions are assigned. It can be interpreted as each identified pattern is assigned to a particular dimension. This is the number of required dimensions.
3) **Tuning the extra dimensions.** The distances on each the additional dimensions are calibrated until the conditional distributions at the marked locations account properly for their respective true values. The calibration process is:
    a. The samples at the marked locations are separated and the linear dependences between them are calculated using cross-validation. Samples that are mutually dependent are grouped and the number of groups becomes the number of extra dimensions to solve the dataset. Each group of samples shares only one extra dimension. This is done for simplicity, otherwise the problem might become intractable to solve.
    b. Small lengths are added to each respective extra dimension at the marked locations. Using cross-validation it is verified if the conditional distribution accounts for the true value within the confidence limits. The same dimension length is added to the non-marked locations where the accuracy of the prediction was affected negatively.
    c. Got to step b and repeat until all the marked locations account for the true values within the specified confidence limits.
    d. Once the marked locations account for the true values the state of the extra dimensions is saved as a solution of the problem. It is worth to mention that there is a negative impact in the some of the surrounding conditional distributions that are minimized in step b.
4) **Save Results.** Store the conditioning dataset including the information of the additional fitted dimensions as the new conditioning dataset.

The goal in the use of extra dimensions is to find sub stationary sub-regions in the domain that are suitable for modeling using conventional techniques. Each sub-region is assigned to a particular extra dimension so that the behavior of the conditioning data on each dimension is more stable in terms of increments, $|z(\mathbf{u}) - z(\mathbf{u}+\mathbf{h})|$, in the original dimensional space (see Figure 5). These sub-regions vanish in the high dimensional space and the whole dataset is considered as stationary and any conventional

technique can be used for modeling uncertainty both in local and global terms. In the proposed algorithm step 3-a calculates the number of extra dimensions required for solving the problem according to the confidence interval parameters.

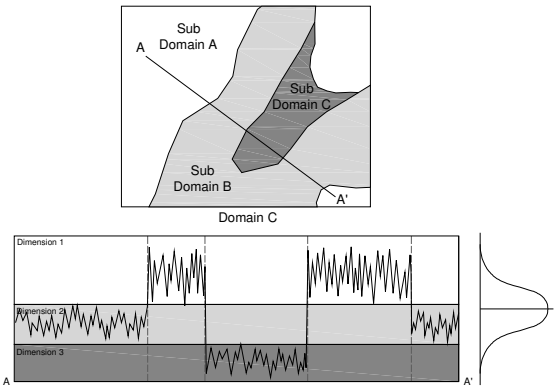

Figure 5: Sketch of classification of sub-domains by using extra dimensions

Case Study

Consider the borehole Ddh-81 dataset in normal score units and the cleaned variogram model (4.9) for 99.7% probability as input information for building a geostatistical model. The confidence interval probability parameter chosen for this example is 95%.

The algorithm solves the dataset using three additional dimensions at twenty data point locations. From the three additional extra dimensions the first one consists of eleven data locations, the second one of five data locations and the third one of four data locations. The presence of outlier data points in the variogram cleaning may show that some data pairs are not accounted for by the variogram model (4.9) due to their large increments in the data pairs. The same locations are also identified in this approach. The initial accuracy of the conditional distributions using the conventional approach show some locations which true values fall outside their confidence limits (see Figure 6-top left). Even when the algorithm approaches the conditional distributions to the accuracy targets, there might be some locations that cannot solved completely. However, they will tend to be close to the fitting conditions and are accepted as solutions using small tolerances in the approximation. That is not the case in this example, the accuracy of all the locations are solved successfully (see Figure 6-top right). The improvement in the calculation of the conditional distributions is evident, however the accuracy of a small proportion of locations is slightly negatively affected at very few locations but they are still within the fitting target limits (see Figure 6-bottom).

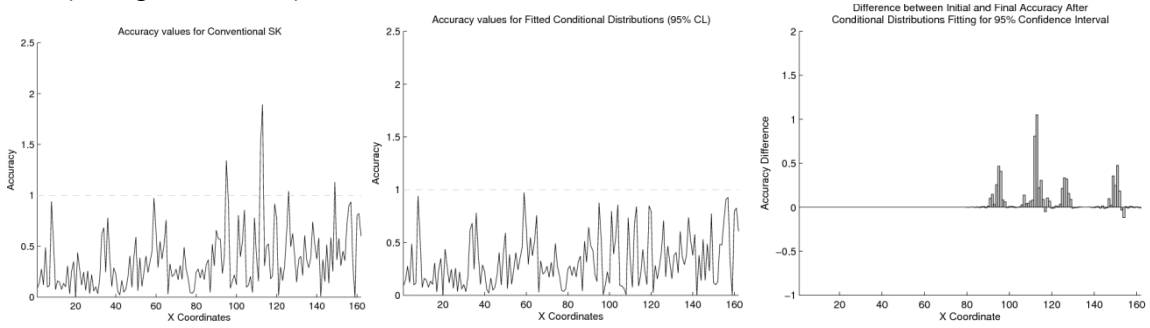

Figure 6: Initial status of the accuracy of the conditional distributions calculated using the conventional approach (top left), conditional distribution fitting (top –right) and comparison of them (initial - fitted) (bottom)

The fitting of the conditional distributions have an inevitable impact in the estimation of the SK means. For the locations where the conditional distributions are fitted there are negligible improvements in the prediction of the SK means. Because, what the approach does is to make them locally more uncertain. However, the influence of fitted data values (outliers) on the surrounding data locations is reduced. This

makes the surrounding samples to be influenced mostly by the non-outlier data values when their conditional distributions are calculated. The improvement of the SK means can be seen when the conventional approach is compared to the conditional distribution fitting approach (see Figure 7). The correlation coefficients of the SK means compared to their true values show a small improvement, that is, 0.8253 for the conventional approach and 0.8671 for the proposed approach. Graphically the SK means of the proposed approach (right side) show a better local fitting than in the conventional case (see Figure 7-left side).

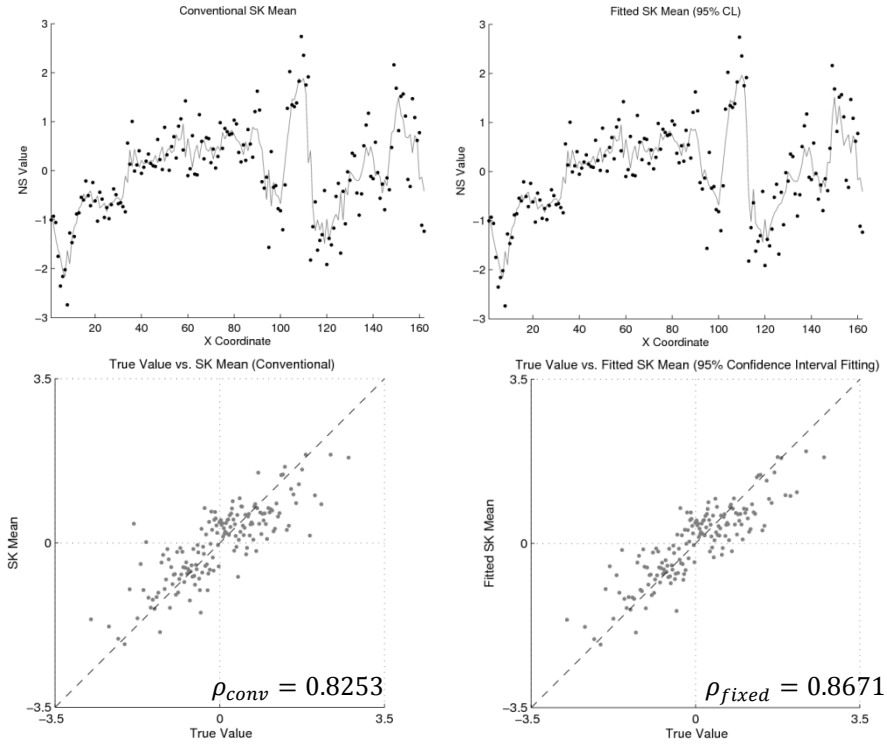

Figure 7: Cross validation SK means calculated using conventional approach (top –left), conditional distributions fitting (top right), and scatter plots of cross validation SK mean versus true values using conventional approach (bottom left), conditional distributions fitting (bottom right)

As well as the SK means the conditional variances are also affected. At the fitted locations the conditional variances tend to increase until the true values fit within the confidence limits of the conditional distribution. The conditional variances can be considered as data dependent in the original dimensional space, since the tuning of the conditional distribution is based on the occurrence of the true values within their respective conditional distributions. While they still remain stationary and configuration dependent in the fitted higher dimensional space.

The number of additional dimensions and locations that require conditional distribution fitting tend to decrease as the confidence limits of the conditional distributions increases, that is because fitting to larger probability intervals is less demanding for the proposed algorithm. For the probability intervals of 97% and 99% the number of required dimension are three and two respectively and the outlier sample locations fifteen and twelve.

Discussion

The extra dimensions capture the information of patterns in the domain that are unaccounted for by the proposed variogram model. Patterns in the domain are present due to the different geologic processes that deposit the concentrations of metal grades or any other element of interest in the domain. Modeling the domain without taking into account such patterns may result in a non-realistic representation of the domain. Consider the same domain as in Figure 5 which consists of three stationary sub-domains A, B and C (see Figure 8-top). The resulting domain becomes non-stationary even when the local mean and

variance are assumed to be constant. This can be seen in a cross section of the values of the domain. Notice that in Figure 8-bottom, for calculating the conditional distributions in the regions of sub-domain A the SK weights of the data points of the sub-domains B and C should be less relevant than of the sub-domain A. In the proposed approach, the degree of relevance in the estimation of the conditional distributions at any location of the domain is tuned by the extra dimensions. When estimating at any location of the sub-domain A the samples of sub-domain B and C become less relevant because the extra dimensions tend to move the samples from domain B and C away the domain A or in another words make other sub-domains samples much more different. The effect of the additional dimensions for fitting the conditional distributions along the domain can be considered as an enhanced form of anisotropy since the directions of preferential continuity are defined more precisely and are shaped by the existing data, that is, data dependent.

In this proposed approach, cross validation is used in the process of fitting the conditional distributions. For each data location the best condition to analyze the estimation of the parameters of its conditional distribution is by using as much information as possible, because for modeling the unsampled locations in the domain the entire dataset used. The larger is the dataset the better is the understanding of the domain is. Other testing techniques such as jackknife are not considered because the analysis has to be done locally, that is, sample by sample. Jackknife tends to be more global and that is not the purpose of this approach.

The over-estimation of the spatial variability in some regions of the domain cannot be identified by using the variogram alone. The solution of these sub-regions would require the reduction of the distances between the existing data locations and perhaps the definition of a new more continuous variogram model. There is no easy way to find out the conditions when the conditional distributions have to be narrowed. Finally, for identifying the conditional distributions which are over-estimated the analysis has to be made in groups of data locations rather than individual samples.

In this document the high-dimensional configuration of the conditional dataset is assumed to behave more stationary in the sense that the conditional distributions calculated via cross-validation account properly for the true data. This condition is extrapolated to the rest of the domain, that is, the estimated conditional distributions of the domain at the unsampled locations in the same dimensional space also accounts for reality as the conditional data does, both local and global.

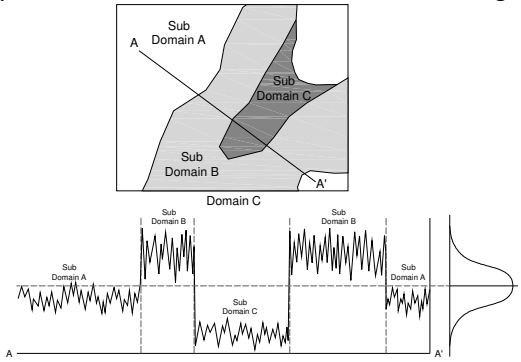

Figure 8: Sketch of combination o two stationary domains A and B into C that mimic a geologic process (top), section of the resulting non-stationary domain C which shows the patterns in data values (bottom)

References

Armstrong, M., & Jabin, R. (1981). Variogram Models Must Be Positive Definite. *Mathematical Geology* , 455-459.

Chilés, J. P., & Delfiner, P. (1999). *Geostatistics, Modeling Spatial Uncertainty.* New York: Wiley-Interscience Publication.

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

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

Isaaks, E., & Srivastava, M. (1989). *An Introduction to Applied Geostatistics.* New York: Oxford Press.

Journel, A. (1986). Models and Tools for Earth Sciences. *Mathematical Geology* , 119-140.

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

Rasmussen, C., & Williams, C. (2008). *Gaussian Processes for Machine Learning.* The MIT Press.