An Intrinsic Model of Coregionalization that Solves Variance Inflation in Colocated Cokriging

Olena Babak and Clayton V. Deutsch

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

The longstanding problem of variance inflation in Sequential Gaussian Simulation with Collocated Cokriging causes the input statistics not to be reproduced. In particular, there is often a systematic bias in the mean and variance of simulated realizations. An alternative approach is presented here that is equally simple, that is, collocated secondary are used and a model of coregionalization is constructed with the correlation coefficient between primary and secondary data. The secondary data is used at the location being estimated and at all data locations. An intrinsic model of coregionalization is assumed. The resulting technique can be referred to as Intrinsic Collocated Cokriging (ICCK). *The resulting estimates are checked carefully and no variance inflation is observed. SGSIM has been modified to include this option and a number of examples are presented. This implementation should systematically replace all versions of the Markov model and collocated cokriging.*

Introduction

Stochastic simulation is a powerful tool for modeling of phenomena that cannot be described deterministically due to their complexity. A number of methods have been developed for joint simulation of dependent random variables. The most popular and simplest method for modeling primary variable based on extensively sampled secondary information is the Sequential Gaussian simulation with Collocated Simple Cokriging. The Collocated Simple Cokriging is build on a Markov-type hypothesis by which collocated secondary information is assumed to screen further away data of the same type.

An unfortunate feature of collocated cokriging is that the kriging variance may be slightly too high. This variance inflation compounds over the sequential path of Sequential Gaussian simulation leading to potentially serious problems with histogram reproduction. An ad hoc method of variance correction has been proposed for dealing this problem; however, the correction is case dependent and requires manual tuning. This tuning is often not performed leading to biased resource estimates.

In this paper we investigate Collocated Simple Cokriging and the reason underlying variance inflation. As a result of this study we propose a new improved approach for cosimulation of dependent random functions without the inference and modeling of a full cross-covariance matrix. The proposed method is based on the intrinsic model of correlation between primary and secondary random variables: Intrinsic Collocated Cokriging (ICCK).

The Sequential Gaussian simulation with full Simple Cokriging based on intrinsic model is tested with a number of small examples to illustrate the correction of variance inflation, the reproduction of the correlation between primary and secondary data, and improved reproduction of the variogram in the new approach as opposed to the conventional Sequential Gaussian simulation with Collocated Simple Cokriging.

Simple Cokriging

Simple Cokriging (CSK) is a natural extension of the Simple Kriging to the case when multivariate data is available. The Simple Cokriging method allows estimating an unknown value at the location of interest not only from the data itself, but also based on the auxiliary variables in the neighborhood. Specifically, the Simple Cokriging estimator is the following weighted linear combination of the mean of the variable of interest (m_*) and the data from different variables located at sample points in the neighborhood of the estimation location *u**

$$Z*_{CSK}\,(u*) = m_* + \sum_{i=1}^{N}\sum_{\alpha=1}^{n_i}\lambda_\alpha^i(Z_i(u_\alpha)-m_i),. \tag{1}$$

where the CSK weights $[\lambda_1^T,\ldots,\lambda_N^T]^T$ are found from a Simple Cokriging system given by

$$\begin{pmatrix} C_{11} & \cdots & C_{1j} & \cdots & C_{1N} \\ \vdots & \ddots & & & \vdots \\ C_{i1} & \cdots & C_{ii} & & C_{iN} \\ \vdots & & & \ddots & \vdots \\ C_{N1} & \cdots & C_{Nj} & \cdots & C_{NN} \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_i \\ \vdots \\ \lambda_N \end{pmatrix} = \begin{pmatrix} c_{1*} \\ \vdots \\ c_{i*} \\ \vdots \\ c_{N*} \end{pmatrix}, \tag{2}$$

where the left hand side covariance matrix is built up with square symmetric n_i by n_i blocks C_{ii} on the diagonal and with rectangular n_i by n_j blocs C_{ij} off the diagonal, with

$$C_{ij} = C_{ji}^T.$$

The blocks C_{ij} contain either direct ($i = j$) or cross ($i \neq j$) covariances between sample points. The vectors c_{i*} contain the covariances with the variable of interest, for a specific variable of the set, between sample points and the estimation location. The vectors λ_i represent the weight attached to the sample of the *i*-th variable.

It follows from system (1)-(2) that in order to perform Simple Cokriging we require a joint model for the matrix of covariance functions. Thus, when *K* different variables are considered, the covariance matrix in the left hand side of Equation (2) requires K^2 covariance functions. Such inference is very demanding in terms of data and subsequent joint modeling, therefore more simple estimation techniques (like Collocated Simple Cokriging) handling multiple data variables are frequently employed instead.

Collocated Simple Cokriging

Collocated Simple Cokriging is a strategy in which the neighborhood of the auxiliary variable is reduced to only one point at the estimation location. This value of the auxiliary variable $Y(u*)$ is said to be collocated with the variable of interest $Z(u)$ at the estimation location *u**. The Collocated Simple Cokriging estimator is given by

$$Z*_{CCSK}\,(u*) = m_Z + \lambda_0(Y(u*)-m_Y) + \sum_{\alpha=1}^{N}\lambda_\alpha(Z(u_\alpha)-m_Z), \tag{3}$$

where Collocated Simple Cokriging weights $[\lambda_Z^T\;\lambda_Y]^T$ are found from the following system of equations

$$\begin{pmatrix} C_{ZZ} & c_{YZ} \\ c_{YZ}^T & \sigma_{YY} \end{pmatrix}\begin{pmatrix} \lambda_Z \\ \lambda_Y \end{pmatrix} = \begin{pmatrix} c_{ZZ} \\ \sigma_{YZ} \end{pmatrix}, \tag{4}$$

where C_{ZZ} is the left hand matrix of the Simple Kriging system of $Z(u)$ and c_{ZZ} is the corresponding right hand side covariance vector. The vector c_{YZ} contains the cross covariances between the *n* sample points of $Z(u)$ and the estimation location *u** with its collocated value $Y(u^*)$. The cross covariance c_{YZ} in system (4) is usually calculated using the following Markov or intrinsic correlation model:

$$c_{YZ} = \sigma_{YZ} r_Z, \tag{5}$$

where σ_{YZ} denotes the covariance between *Z* and *Y*; and r_Z is the vector of spatial correlations $\rho(u_\alpha - u_0), \quad \alpha = 1,\ldots,n.$

Using the Markov model for the cross covariance c_{YZ}, we can rewrite system (4) for the Collocated Simple Cokriging weights $[\lambda_Z^T \ \lambda_Y]^T$ as

$$\begin{pmatrix} \sigma_{ZZ} R_Z & \sigma_{YZ} r_Z \\ \sigma_{YZ} r_Z^T & \sigma_{YY} \end{pmatrix} \begin{pmatrix} \lambda_Z \\ \lambda_Y \end{pmatrix} = \begin{pmatrix} \sigma_{ZZ} r_Z \\ \sigma_{YZ} \end{pmatrix}, \tag{6}$$

where R_Z is the matrix of spatial correlations $\rho(u_\alpha - u_\beta), \quad \alpha, \beta = 1,\ldots,n.$

Note that in order to to perform the collocated cokriging with Markov model, we only need to know the covariance function

$$C_{ZZ}(h) = \sigma_{ZZ}\rho(h),$$

the variance σ_{YY} of the auxiliary variable and the correlation coefficient $\rho_{YZ} = \rho_{YZ}(0)$. Retaining only the collocated secondary data, in general, does not affect the resulting estimate, since the close neighborhood data are usually very similar in values. However, it may affect the Cokriging estimation variance. Cokriging variances are overestimated, oftentimes significantly. This causes serious problem in sequential simulation.

In order to understand the reason underlying the variance inflation in the Collocated Simple Cokriging, in the next section we will investigate the question whether the Simple Cokriging reduces to Collocated Simple Cokriging with the intrinsic correlation model.

Simple Collocated Cokriging is *not* at Intrinsic Model

It is quite interesting to learn that the Simple Cokriging does not reduce to Collocated Simple Cokriging with the intrinsic correlation model. Let us review the proof of this fact (see also Wackernagel, 1995).

Assume that $Z(u)$ and $Y(u)$ are intrinsically correlated with unit variances. Consider Simple Cokriging to find the unknown value of the variable of $Z(u)$ at location *u** based on the neighbor data $Z(u_\alpha)$ at *n* sample locations $u_\alpha, \alpha = 1,\ldots,n,$ the corresponding values $Y(u_\alpha)$ at the same locations and the value of the auxiliary variable $Y(u)$ at the estimation location *u**. Then the value of the variable of $Z(u)$ at location *u** is given by the Simple Cokriging approach as

$$Z^*_{CSK}(u^*) = m_Z + \lambda_0(Y(u^*) - m_Y) + \sum_{\alpha=1}^{n} \lambda_{Y,\alpha}(Y(u_\alpha) - m_Y) + \sum_{\alpha=1}^{n} \lambda_{Z,\alpha}(Z(u_\alpha) - m_Z), \tag{7}$$

where the CSK weights $[\lambda_Z^T \ \lambda_Y^T \ \lambda_0]^T$ are given by

$$\begin{pmatrix} R_Z & \rho_{YZ} R_Z & \rho_{YZ} r_Z \\ \rho_{YZ} R_Z & R_Z & r_Z \\ \rho_{YZ} r_Z & r_Z^T & 1 \end{pmatrix} \begin{pmatrix} \lambda_Z \\ \lambda_Y \\ \lambda_0 \end{pmatrix} = \begin{pmatrix} r_Z \\ \rho_{YZ} r_Z \\ \rho_{YZ} \end{pmatrix}, \tag{8}$$

where r_Z is the vector of spatial correlations $\rho(u_\alpha - u_0), \quad \alpha = 1, \ldots, n,$ and R_Z is the matrix of spatial correlations $\rho(u_\alpha - u_\beta), \quad \alpha, \beta = 1, \ldots, n.$

Now note that in order for Simple Cokriging to be reduced to Collocated Simple Cokriging (see eq. (3)-(6)), the vector of weights $[\lambda_Z^T \; 0^T \; \lambda_0]^T$ must be solution of the system (8). Let us check this. Specifically, let us substitute $[\lambda_Z^T \; 0^T \; \lambda_0]^T$ in system (8), then we will obtain

$$\begin{pmatrix} R_Z & \rho_{YZ} R_Z & \rho_{YZ} r_Z \\ \rho_{YZ} R_Z & R_Z & r_Z \\ \rho_{YZ} r_Z & r_Z^T & 1 \end{pmatrix} \begin{pmatrix} \lambda_Z \\ 0 \\ \lambda_0 \end{pmatrix} = \begin{pmatrix} r_Z \\ \rho_{YZ} r_Z \\ \rho_{YZ} \end{pmatrix},$$

or, in equation format

With respect to the value of ρ_{YZ}, we can consider now 3 cases:

1. $\rho_{YZ} = 0$, then we have reduction to simple kriging and $\lambda_0 = 0.$ However, $\lambda_0 = 0$ can not be a solution of the Collocated Cokriging.

2. $\rho_{YZ} = \pm 1.$ Consider first $\rho_{YZ} = -1,$, then system (9) can be rewritten as

$$\begin{cases} \lambda_Z R_Z - \lambda_0 r_Z = r_Z \\ -\lambda_Z R_Z + \lambda_0 r_Z = -r_Z . \\ -\lambda_Z r_Z^T + \lambda_0 = -1 \end{cases} \tag{10}$$

Looking at system (10), it becomes clear that the 1st and 2nd equation in this system are the same, thus system (10) can be reduced to the Collocated Simple Cokriging system

$$\begin{cases} \lambda_Z R_Z - \lambda_0 r_Z = r_Z \\ -\lambda_Z r_Z^T + \lambda_0 = -1 \end{cases}. \tag{11}$$

with solution $\lambda_Z = 0, \;\; \lambda_0 = -1.$ Proof that for $\rho_{YZ} = 1,$ system (9) reduces to the trivial Collocated Simple Cokriging system with solution $\lambda_Z = 0, \;\; \lambda_0 = 1,$ can be obtained following the same approach as outlined above.

3. $\rho_{YZ} \neq 0, \pm 1,$ then if we multiply the first equation of matrix (9) by ρ_{YZ}, and subtract the result from the first equation in (9), we will obtain

$$\lambda_Z \rho_{YZ} R_Z + \lambda_0 r_Z - \rho_{YZ}(\lambda_Z R_Z + \lambda_0 \rho_{YZ} r_Z) = \rho_{YZ} r_Z - \rho_{YZ} r_Z ,$$

or

$$\lambda_0 r_Z - \lambda_0 \rho_{YZ}^2 r_Z = 0 .$$

And, thus,

$$\lambda_0 (1 - \rho_{YZ}^2) r_Z = 0 . \tag{12}$$

Due to the fact that $\rho_{YZ} \neq \pm 1,$ and there exist non-zero spatial correlations $\rho(u_\alpha - u_\beta), \alpha, \beta = 1, \ldots, n,$ we can conclude that necessarily $\lambda_0 = 0.$ However, $\lambda_0 = 0$ can not be a solution of the Collocated Cokriging.

Since neither of the values for ρ_{YZ} yielded reduction of the Simple Cokriging solution to the Collocated Simple Cokriging solution, we can conclude that Simple Cokriging can not be reduced to Collocated Simple Cokriging, thus *there is no theoretical justification for selecting only one auxiliary sample (collocated) for Cokriging even in the case of intrinsic correlation model.*

Source of Variance Inflation in Collocated Cokriging

Sequential Gaussian Simulation is adapted to model local conditional distributions under a multivariate Gaussian model and collocated cokriging. Simulation is performed by drawing from such conditional distributions. Newly simulated data are used as conditional data in simulation of the new nodes. Multiple equally-probable realizations of the property of interest are created.

In simulation based on Collocated Simple Cokriging, as was mentioned earlier, a strange problem of variance inflation is observed. The aim of this Section is to determine the reason for this variance inflation.

Let us consider estimation of the value of the unit variance normally distributed primary variable *Z* at location u_{2*} using two conditioning (original) neighbor data $Z(u_1), Z(u_2),$ simulated value of the same type at location u_{1*}, $Z*_{CCSK}\,(u_{1*})$ given by

$$Z*_{CCSK}\,(u_{1*}) = m_Z + \lambda_0(Y(u_{1*}) - m_Y) + \sum_{\alpha=1}^{2}\lambda_\alpha(Z(u_\alpha) - m_Z) + R(u_{1*}), \tag{13}$$

where $Y(u_{1*})$ is collocated value to $Z(u_{1*})$ of unit variance normally distributed auxiliary variable, the CCSK weights $[\lambda_1\ \lambda_2\ \lambda_0]^T$ are given by

$$\begin{pmatrix} 1 & \rho_{ZZ}(u_1-u_2) & \rho_{YZ}(u_1-u_{1*}) \\ \rho_{ZZ}(u_2-u_1) & 1 & \rho_{YZ}(u_2-u_{1*}) \\ \rho_{YZ}(u_{1*}-u_1) & \rho_{YZ}(u_{1*}-u_2) & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_0 \end{pmatrix} = \begin{pmatrix} \rho_{ZZ}(u_1-u_{1*}) \\ \rho_{ZZ}(u_2-u_{1*}) \\ \rho_{YZ}(0) \end{pmatrix}, \tag{14}$$

or, in system format as

$$\begin{cases} \lambda_1 + \lambda_2\rho_{ZZ}(u_1-u_2) + \lambda_0\rho_{YZ}(u_1-u_{1*}) = \rho_{ZZ}(u_1-u_{1*}) \\ \lambda_1\rho_{ZZ}(u_2-u_1) + \lambda_2 + \lambda_0\rho_{YZ}(u_2-u_{1*}) = \rho_{ZZ}(u_2-u_{1*}). \\ \lambda_1\rho_{YZ}(u_{1*}-u_1) + \lambda_2\rho_{YZ}(u_{1*}-u_2) + \lambda_0 = \rho_{YZ}(0) \end{cases} \tag{15}$$

and $R(u_{1*})$ is local independent of the data normal random error with mean of zero and variance of $Var(R(u_{1*})) = 1 - \lambda_1\rho_{ZZ}(u_1-u_{1*}) - \lambda_2\rho_{ZZ}(u_2-u_{1*}) - \lambda_0\rho_{YZ}(0);$ and using one collocated data $Y(u_{2*}).$ Then the Collocated Simple Cokriging estimate of $Z_{CCSK}(u_{2*})$ is given by

$$Z*_{CCSK}\,(u_{2*}) = m_Z + \lambda_0(Y(u_{2*}) - m_Y) + \sum_{\alpha=1}^{2}\lambda_\alpha(Z(u_\alpha) - m_Z) + \lambda_3(Z*_{CCSK}\,(u_{1*}) - m_Z), \tag{16}$$

where the CCSK weights $[\lambda_1\ \lambda_2\ \lambda_3\ \lambda_0]^T$ are given by

$$\begin{cases} \lambda_1 + \lambda_2\rho_{ZZ}(u_1-u_2) + \lambda_3\rho_{ZZ}(u_1-u_{1*}) + \lambda_0\rho_{YZ}(u_1-u_{2*}) = \rho_{ZZ}(u_1-u_{2*}) \\ \lambda_1\rho_{ZZ}(u_2-u_1) + \lambda_2 + \lambda_3\rho_{ZZ}(u_2-u_{1*}) + \lambda_0\rho_{YZ}(u_2-u_{2*}) = \rho_{ZZ}(u_2-u_{2*})\ . \\ \lambda_1\rho_{ZZ}(u_{1*}-u_1) + \lambda_2\rho_{ZZ}(u_{1*}-u_2) + \lambda_3 + \lambda_0\rho_{YZ}(u_{1*}-u_{2*}) = \rho_{ZZ}(u_{1*}-u_{2*}) \\ \lambda_1\rho_{YZ}(u_{2*}-u_1) + \lambda_2\rho_{YZ}(u_{2*}-u_2) + \lambda_3\rho_{YZ}(u_{2*}-u_{1*}) + \lambda_0 = \rho_{YZ}(0) \end{cases} \tag{17}$$

Let us now check the following

1) Is the covariance (correlation) between the new estimate and the conditioning data values correct?

For any *i* =1, 2,

$$
\begin{aligned}
&Cov(Z(u_i), Z*_{CCSK}\ (u_{2*}))\\
&= Cov\left(m_Z + \lambda_0 (Y(u_{2*}) - m_Y) + \sum_{\alpha=1}^{2} \lambda_\alpha (Z(u_\alpha) - m_Z) + \lambda_3 (Z*_{CCSK}\ (u_{1*}) - m_Z), Z(u_i) \right)\\
&= \lambda_0 Cov(Y(u_{2*}), Z(u_i)) + \lambda_1 Cov(Z(u_1), Z(u_i)) + \lambda_2 Cov(Z(u_2), Z(u_i)) + \lambda_3 Cov(Z*_{CCSK}\ (u_{1*}), Z(u_i))\\
&= \lambda_0 \rho_{YZ}(u_2 - u_{2*}) + \lambda_1 \rho_{ZZ}(u_1 - u_i) + \lambda_2 \rho_{ZZ}(u_2 - u_i) + \lambda_3 \rho_{ZZ}(u_2 - u_{1*}) = \rho_{ZZ}(u_i - u_{2*}).
\end{aligned}
$$

Thus, it is correct! Note that the last two substitutions followed from systems (15) and (17), respectively.

2) Is the covariance between the new estimate and the previously calculated estimates correct?

$$
\begin{aligned}
&Cov(Z*_{CCSK}\ (u_{1*}), Z*_{CCSK}\ (u_{2*}))\\
&= Cov\left(Z*_{CCSK}\ (u_{1*}), m_Z + \lambda_0 (Y(u_{2*}) - m_Y) + \sum_{\alpha=1}^{2} \lambda_\alpha (Z(u_\alpha) - m_Z) + \lambda_3 (Z*_{CCSK}\ (u_{1*}) - m_Z) \right)\\
&= \lambda_0 Cov(Y(u_{2*}), Z*_{CCSK}\ (u_{1*})) + \lambda_1 Cov(Z(u_1), Z*_{CCSK}\ (u_{1*})) + \lambda_2 Cov(Z(u_2), Z*_{CCSK}\ (u_{1*})) + \lambda_3 Var(Z*_{CCSK}\ (u_{1*}))\\
&= \lambda_0 Cov(Y(u_{2*}), Z*_{CCSK}\ (u_{1*})) + \lambda_1 \rho_{ZZ}(u_{1*} - u_1) + \lambda_2 \rho_{ZZ}(u_{1*} - u_2) + \lambda_3.
\end{aligned}
$$

Now note that if $Cov(Y(u_{2*}), Z*_{CCSK}\ (u_{1*})) = \rho_{YZ}(u_{2*} - u_{1*}),$ then the last substitution from equation (15) would result in correct covariance between the new estimate and the previously calculated estimates. But does $Cov(Y(u_{2*}), Z*_{CCSK}\ (u_{1*})) = \rho_{YZ}(u_{2*} - u_{1*})$?

Let us assume that it is true, that is, $Cov(Y(u_{2*}), Z*_{CCSK}\ (u_{1*})) = \rho_{YZ}(u_{2*} - u_{1*}),$ then the Simple Cokriging problem would be necessarily reduced to the Collocated Simple Cokriging problem (which inherently ensures this equality), however, as shown in the section above such reduction is impossible even in the case of intrinsic correlation model. Thus, we can conclude that due to the fact that Collocated Cokriging system has no control over the cross covariance $Cov(Y(u_{2*}), Z*_{CCSK}\ (u_{1*})),$ _*the correct covariance between the new estimate and the previously calculated estimates cannot be ensured*! When Collocated Simple Cokriging is put into sequential simulation mode, incorrect covariance of the simulated data is transferred to the new simulated data, and, as result, unavoidable variance inflation is observed.

Proposed Solution to the Variance Inflation Problem

$$
E(Z_2(u) \mid Z_1(u) = z, Z_1(u+h) = z') = E(Z_2(u) \mid Z_1(u) = z) \quad \forall h, \forall z'
$$

Case study I

Let us compare the variance of the Simple Cokriging estimator with the variance of the Collocated Simple Cokriging estimator. To assess how much larger (smaller) is the variance predicted by the Collocated Simple Cokriging than the variance of the Simple Cokriging, we consider the following scale measure:

$$
\textbf{Relative increase in variance} = \frac{\textbf{(Variance of the CCSK estimator - Variance of the CSK estimator)}}{\textbf{Variance of the CSK estimator}} \bullet 100\%
$$

Based on the above scale measure we will assess the following:

For a data configuration given in Figure 1, we will determine the relative increase in the variance with respect to the following factors:

1. Correlation between primary and secondary variable;

2. Variogram model (and its range).

When performing Simple Cokriging, secondary data will be considered available at the locations closest to the estimation location (part A) and at the same locations as the data on the primary variable (part B).

Case Study I Results: Part A

Figure 2 illustrates the two configurations considered for the secondary data information. Each of the secondary data locations are considered on the grid of 10 by 10 units (size of the domain is 100 by 100 units). Figures 3 and 4 illustrate the relative increase in the variance in the Collocated Simple Cokriging over Simple Cogriging obtained based on Spherical variogram model with the range of correlation equal to the size of the domain for the correlation coefficients between primary and secondary data equal to $\rho = 0.1$, 0.3, 0.6, 0.9 for the configurations a) and b) of Figure 2, respectively. Note that Figures 3 and 4 also show for comparison the Collocated Simple Cokriging variances for estimates in the study domain.

First what we must note from Figures 3 and 4 is that when the correlation between primary and secondary variable is small, the variance predicted for the estimates by the Collocated Simple Cokriging estimator is virtually the same as the variance predicted for the estimates by the Simple Cokriging estimator. This is what we expected, since in the case of insignificant primary-secondary data correlation, secondary attribute contains little or no information of use for estimation of the primary variable. With increase in the correlation between primary and secondary variable, the variances predicted by Collocated Simple Cokriging estimator become increasingly higher than the variance predicted for the estimates by the Simple Cokriging estimator. The increase in the variance is not that significant, however, we believe that due to the sequential mode of the geostatistical simulation, even this increase in the variance can result in significant variance inflation. The second factor which would, of course, contribute to the variance inflation in the simulation is not appropriateness of the intrinsic correlation model. This factor will be discussed later in this work.

Also note from Figures 3 and 4 the non-smoothness of the maps for the relative increase in the variance. This non-smoothness is connected to the configuration of the secondary data information used in the full Simple Cokriging. To avoid the non-smoothness of the map we can either use either larger secondary data configuration (note the increase in the smoothness when changing from 4 point secondary data configuration (Figure 3) to 12 point secondary data configuration (Figure 4)) or pick secondary data randomly. Due to the fact that using more secondary data can result in singularity matrix problems, instability, etc (see Wackernagel, 1995), it is better to use a second option. Namely, when performing Simple Cokriging we should consider secondary data information at the same locations as the data on the primary variable. This approach is illustrated below.

Case Study I Results: Part B

Figure 5 illustrates the relative increase in the variance obtained based on the Spherical variogram model with the range of correlation equal to the size of the domain, that is 100 units, for the correlation coefficients between primary and secondary data equal to $\rho = 0.1$, 0.3, 0.6, 0.9. In estimation via Simple Cokriging secondary data is considered to be available at the same locations as the data on the primary variable. For better visual comparison Figure 5 also shows the Collocated Simple Cokriging variances for estimates in the domain of interest.

From Figure 5 we can come to the same conclusions as from Figures 3 and 4. That is, with increase in the correlation between primary and secondary random variables, variance inflation in primary random variable model increases. However, when we use in the Simple Cokriging secondary data at the same locations as the primary data, we also gain the smoothness of the resulting relative variance inflation maps.

Now let us consider results shown in Figure 5 more precisely. Figures 7-9 show the comparison of the full Simple Cokriging (CSK) estimation variances and estimation variances obtained based on Collocated Simple Cokriging (CCSK) for the three slices, that is slice at X = 20, slice at X = 40 and slice at X = 80, respectively, in the domain of interest, see Figure 6. Estimation variance in each case were calculated based on the isotropic Spherical variogram aas before and for the same four values of the correlation coefficient (0.1, 0.3, 0.6, 0.9) between primary and secondary information were considered. Note that the three chosen slices correspond to the following tree scenarios: 1) conditioning data location is one of the points being estimated (slice at X = 20); 2) estimation is performed close to the conditioning data (slice at X = 40); 3) estimation is performed quite far from the conditioning data (slice at X = 80).

It is easy to see from Figures 7-9 that the CSK and CCSK estimation variances obtained for the points on the three considered slices are only close to each other when the correlation between primary and auxiliary variable is very small. With increase in the correlation between primary and auxiliary information, the difference between the CSK estimation variance and the CCSK estimation variance becomes larger and larger. The CSK estimation variance is becoming significantly smaller than the CCSK estimation variance with increase in the correlation between the primary and secondary information.

To understand how the CSK and CCSK estimation variances change with increase/decrease in the primary variable range of the correlation, let us consider estimation (Figures 10-12) of the points in the same three slices as before based on the isotropic Spherical variogram with the range of correlation equal to a) 25% of the size of the study domain; b) 50% of the size of the study domain; c) equal to the size of the study domain and d) 250% of the size of the study domain and the correlation between primary and secondary information equal to 0.9. Figures 10-12 clearly show that with increase in the range of correlation, the difference between the CSK estimation variance and the CCSK estimation variance becomes larger and larger. The CSK estimation variance becomes significantly smaller (relatively) than the CCSK estimation variance. Moreover, if the estimation location is located outside of the range of correlation from the conditioning data, than both the CSK and CCSK estimation variances are the same.

Impact of the variogram model on the CSK and CCSK estimation variances is shown in Figures 13-15. Results for the slices at X = 20, X = 40 and X = 80, respectively, were obtained based on the Spherical, Exponential and Gaussian variogram models with the range of correlation equal to the size of the study domain and the correlation between primary and secondary information equal to 0.9. One can clearly note from Figures 13-15 that the variance inflation associated with CCSK estimation is significant for all of the variogram models. The most variance inflation is observed when Spherical model is used as the variogram model for the data.

Case study II: Impact of the Secondary Data on the Simple Cokriging Estimate with Intrinsic Model

Let us now investigate how the weights given to the primary and secondary data are distributed when the full Simple Cokriging is performed based on the intrinsic model. The design of this case study is the same as in the case study I. That is, we consider estimation of the domain 100 by 100 units based on 8 data points. When performing Simple Cokriging, secondary data information is considered at the same locations as the data on the primary variable.

The weights received by the primary data in the Collocated Simple Cokriging and the full Simple Cokriging based on the intrinsic model when estimating locations: a) (20, 20); b) (40, 70); c) (80, 40) are shown in Figures 16-18, respectively. Note that Figures 16-18 show the weights obtained based on the Spherical variogram model with the range of correlation equal to the size of the study domain using the correlation coefficient between primary and secondary information equal to 0.3 and 0.9. The resulting CSK and CCSK estimation variances obtained when estimating these three locations of interest are also tabulated in these figures.

It is apparent from Figures 16-18 that when the range of correlation is small the primary data weights obtained based on both models are virtually the same, however, with increase in the range of correlation, the accumulated weight received by the primary data in the full Simple Cokriging based on the intrinsic model is much higher than the accumulated weight received by the primary data in the Collocated Simple Cokriging. The weight received by the collocated data in the Simple Cokriging is also higher than the respective weight obtained in the Collocated Simple Cokriging. However, the accumulated weight received by the secondary data in the Simple Cokriging based on the intrinsic model is quite small.

Proposed Solution to the Variance Inflation Problem

In the view of the above results a natural solution for the problem of the variance inflation in the Sequential simulation of the primary variable based on the auxiliary data information is the following. Instead of just considering Collocated Simple Cokriging based on the Markov correlation model, we should consider the full Simple Cokriging based on the intrinsic model. Due to the fact that Simple Cokriging estimates have larger variance (smaller missing variance), less variability will be added to the estimates to obtain stationary variance. As result, less variability will be observed in the generated primary variable realizations.

In order to assess the reduction in the variance inflation and appropriateness of the above outlined proposal, we will consider the following small example.

Example

Let us consider the following Linear Model of Coregionalization (LMC) for the primary unit variance, zero mean random variable *Z* and secondary unit variance, zero mean random variable *Y* (see Deutsch, 2002)

$$\begin{aligned}\gamma_{YY}(h) &= 0.1\cdot Sph_{16}(h) + 0.9\cdot Gaus_{32}(h)\\ \gamma_{YZ}(h) &= 0.25\cdot Sph_{16}(h) + 0.25\cdot Gaus_{32}(h)\ ,\\ \gamma_{ZZ}(h) &= 0.9\cdot Sph_{16}(h) + 0.1\cdot Gaus_{32}(h)\end{aligned} \tag{18}$$

where $Sph_{16}(h), Gaus_{32}(h)$ denote the Spherical variogram model with the range of 16 and Gaussian variogram model with the range of 32. Note that system (18) is a valid LMC, since

$$0.1\cdot 0.9 = 0.09 \geq 0.25\cdot 0.25 = 0.0626 \text{ and } 0.9\cdot 0.1 = 0.09 \geq 0.25\cdot 0.25 = 0.0625\,.$$

Then the correlation at lag 0 between primary and secondary random variables can be calculated under stationarity as

$$\rho_{YZ} = 1-\gamma_{YZ}(0) = 1-[0.25\cdot Sph_{16}(0) + 0.25\cdot Gaus_{32}(0)] = 1-0.25-0.25 = 0.5\,.$$

Now let us consider unconditional sequential simulation (SGS) based on the Simple Collocated Cokriging for the primary variable *Z* (continuity of Z is given by $\gamma_{ZZ}(h)$ in system (18)) using exhaustive secondary random variable *Y* and coefficient of correlation $\rho_{YZ} = 0.5$. The exhaustive secondary information for *Y* was obtained by unconditional Sequential Gaussian Simulation (SGS) with variogram model $\gamma_{YY}(h)$ given in (18). Figure 19 shows the distributions of the means and variances of the secondary random variable for 100 SGS realizations. In sequential simulation for both primary and secondary random variables maximum number of simulated nodes to use was set to 12, maximum search radii were set to largest variogram range (that is, 32).

Summary of the results for the primary random variable *Z* for 100 SGS realization on the area of 256 by 256 cells are shown in Figure 20. This figure shows the distributions of the means and variances of the primary random variable *Y* obtained based on Sequential Gaussian Simulation with Simple Collocated Cokriging. Note that despite the expected mean of the distribution of the primary random variable Z modeled by Sequential Gaussian Simulation with Simple Collocated Cokriging is virtually zero, there is a dramatic deviation of the observed expected variance from the target variance of one. The deviation of the variance of the primary random variable Z modeled by Sequential Gaussian Simulation with Simple Collocated Cokriging from the target is, on average, around 28%. One can argue that this strong deviation is largely impacted by the mismatch in the continuity of the primary and secondary random variable (Markov model is inappropriate). In order to understand this better, let us, first, consider modeling of the primary random variable Z using Sequential Gaussian Simulation with full Simple Cokriging based on the intrinsic model of correlation (see (8)). [Secondary information in Sequential Gaussian Simulation with full Simple Cokriging is selected at the same locations as the primary data]. Figure 21 shows the distributions of the means and variances of the primary random variable *Z* obtained in this case. Note that both the expected mean and expected variance are both virtually the same as target mean of 0 and target variance of 1. Thus, we can see that indeed the main factor which triggers the variance inflation in Sequential Gaussian Simulation with Simple Collocated Cokriging is not the assumption of Markov model, but using only one collocated data when performing Cokriging. Therefore, our recommendation is to *always* use full Simple Cokriging in the sequential simulation mode. Note also that Figure 23 shows results for the means and variances of the realizations of the primary random variable *Z* obtained using Sequential Gaussian Simulation with full Simple Cokriging based on the intrinsic model of correlation when secondary information is taken at the nearby locations to the estimation location, see Figure 22 for the data configuration. Clearly, reproduction of the target statistics shown in Figure 23 is much better than the respective results of Sequential Gaussian Simulation with Simple Collocated Cokriging. However, as we can see the results depend somewhat on the configuration. Therefore, ones again we conclude that the Sequential Gaussian Simulation with full Simple Cokriging using secondary data at the same locations as the primary data is better choice (no dependence on the data configuration, etc.)

Another advantage of using the Sequential Gaussian Simulation with full Simple Cokriging based on intrinsic model over Sequential Gaussian Simulation with Simple Collocated Cokriging is the improved primary variable variogram reproduction. Figures 24-26 show variogram reproduction of the secondary random variable *Y* by the unconditional Sequential Guassian Simulation, variogram reproduction of the primary random variable *Z* by the unconditional Sequential Guassian Simulation with Simple Collocated Cokriging and variogram reproduction of the primary random variable *Z* by the unconditional Sequential Guassian Simulation with full Simple Cokriging (secondary information selected at the same locations as primary variable). It is apparent from Figures 25 and 26 that mismatch between target semivariogram for the primary variable Z is significantly reduced by applying full Simple Cokriging based on intrinsic model instead of Collocated Simple Cokriging. Note that amount of mismatch could also depends on such parameters as maximum number of nodes used in simulation, search radii, etc. In the case study considered in this paper this parameters were set to 12 nodes and maximum variogram range, respectively. Note also that if the continuity of the primary and secondary random variables would be the same the mismatch between target semivariogram and semivariogram reproduced in the unconditional Sequential Gaussian Simulation with full Simple Cokriging would be removed entirely.

It is also important to note that modeling of the primary variable based on the secondary random variable in the full Simple Cokging based on the intrinsic model framework also ensures reproduction of the correlation between primary and secondary random variable. This point is illustrated in Figure 27. Figure 27 shows 100 correlation coefficients obtained for each of the 100 SGS realizations with full Simple Cokging of the primary random variable. The observed mean correlation coefficient of 0.4640 is very close to the target correlation coefficient of 0.5 used in simulation.

FORTRAN code

The Sequential Gaussian Simulation with full Simple Cokriging based on intrinsic model was incorporated into sgsim program. The new program is called cck_sgsim. This program uses the same parameter file as sgsim, see Deutsch and Journel (1998) for reference. The only difference is that it has 6 options for the type of Kriging, instead of five. Specifically, the first five options are the same as in sgsim: 0 = Simple Kriging, 1 = Ordinary Kriging, 2 = Locally Varying Mean, 3 = External Drift, 4 = Collocated Simple Cokriging, and the last one sixth option (new) is 5 = Collocated Cokriging. Format of results is the same as that of original sgsim.

Discussion

In this paper a new approach for dealing with variance inflation in the sequential simulation using secondary data information was proposed. The proposed approach employs the full Simple Cokriging based on the intrinsic variogram model to calculate local distributions. It was shown via small examples that the new methodology removes entirely the variance inflation, insures reproduction of the correlation between primary and secondary data and improves the reproduction of the variogram even when the primary and secondary variables differ significantly in continuity.

References:

Deutsch, C.V., *Geostatistical Reservoir Modeling,* 2002.

Deutsch C.V., and Journel A.G., *GSLIB: Geostatistical Software Library and User’s Guide*, 1998.

Wackernagel, H., *Multivariate Geostatistics,* 1995.