

A Recall of Expected Ergodic Fluctuations in Gaussian Simulation

Martha E. Villalba and Clayton V. Deutsch

The simulation of standard normal values should reproduce the Gaussian distribution with mean zero and variance one; however, ergodic fluctuations cause these statistics to depart from their expected values. Statistical fluctuations are a part of the global uncertainty in the variable of interest. A non ergodic domain is the one where the range of correlation is large with respect to the domain size and where statistical fluctuations in the variance of the spatial average are expected. The research proposes to quantify the magnitude of expected statistical fluctuations by the use of analytical equation. These results are validated by the outcomes of non conditional realizations. A program LUSIM is modified to calculate the variance of spatial averages analytically.

1. Introduction

Geostatistical techniques for resource evaluation, such as Kriging and Simulation, require two assumptions. The first assumption of stationarity states that all multivariate distributions are invariant by translation over the study domain. Multivariate distributions are summarized by the mean vector and covariance matrix for all locations. The second assumption of ergodicity states that the spatial average (1) of a random stationary function (RF) $Z(\mathbf{u})$ over a domain A converges to the expected value $m=E\{Z(\mathbf{u})\}$ when A tends to infinity (2) (Chilès & Delfinier, 1999).

$$\bar{Z}_A = \frac{1}{|A|} \int_A Z(\mathbf{u}) du \quad (1)$$

$$\lim_{A \rightarrow \infty} \bar{Z}_A = m \quad (2)$$

When the domain size tends to infinity, the variance of the spatial average is expected to be zero. In practice A is finite and the spatial average Z_A will be variable when A is finite. Figure 1 shows the change of the spatial average variance as a function of the size of A .

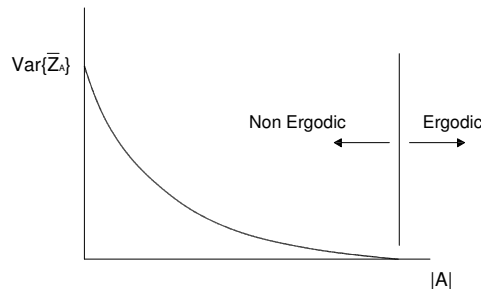


Figure 1: The variance of spatial average versus A . When this variance is greater than zero, the domain is called non ergodic.

Simulation algorithms are based on the multivariate Gaussian RF model. This parametric model is the most widely used with extremely congenial properties (Goovaerts, 1997). The simulation of standard normal values should reproduce the Gaussian distribution with mean zero and variance one; however, ergodic fluctuations make the results different from zero and one. A study on acceptable ergodic fluctuations (Leuangthong, McLennan, & Deutsch, 2005) shows significant statistical fluctuations for three examples with variogram range of 20%, 50%, and 100% of the domain. Even when the domain becomes relatively large compared to the range of correlation, these statistical fluctuations are considerable and are a part of the global uncertainty. The magnitude of the statistical fluctuations can be quantified by performing non conditional simulation. The expected fluctuations in the mean are derived below in presence of conditioning and verified by numerical examples.

2. Expected Fluctuations in the Mean

The variance of the spatial average in the domain A is a measure of the expected fluctuations. The domain could be discretized by N nodes, these are defined by the variable function $Z(\mathbf{u}^{(i)})$, where the location of each node is $\mathbf{u}^{(i)}$, i

$= 1, \dots, N$. The available data are defined by $z(\mathbf{u}_k)$, where the location of each data value is \mathbf{u}_k $k = 1, \dots, n$. These n available data values and N nodes define the domain A . Values at each node are estimated conditioned to the available n values.

The covariance of the RF $Z(\mathbf{u})$ should be constant over the domain, however. A non stationary covariance is observed in the presence of conditioning data. The covariance near the conditioning data is a function of the input “ergodic” covariance model and the location of conditioning data. Figure 2 shows a domain A that has $z(\mathbf{u}_k)$, $k = 1, \dots, 6$ conditioning data. The discretization of the domain is with 100 nodes. As expected, the covariance between adjacent node location \mathbf{u}^{39} and \mathbf{u}^{40} will be different than the covariance between the node locations \mathbf{u}^{71} and \mathbf{u}^{72} . This difference is because $C(\mathbf{u}^{71}, \mathbf{u}^{72})$ has node locations that are near the conditioning data; $C(\mathbf{u}^{39}, \mathbf{u}^{40})$ has node locations that are far from the conditioning data. That is, the distance to the conditioning data matters in the evaluation of covariances.

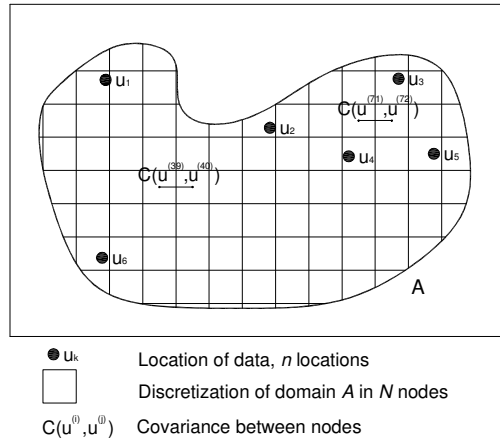


Figure 2: The covariance $C(\mathbf{u}^{39}, \mathbf{u}^{40})$ that is far from the conditioning data is different to the covariance $C(\mathbf{u}^{71}, \mathbf{u}^{72})$ that is near from the conditioning data.

This non stationary covariance is correctly reproduced in sequential Gaussian simulation because the previous simulated nodes are used in the estimation of subsequent nodes (Neufeld, Ortiz, & Deutsch, 2005). These conditional covariances are required to compute the variance of the spatial average, which is given by:

$$\text{Var}\{\bar{Z}_A\} = \frac{1}{N^2} \sum_i^N \sum_j^N E\{Z(\mathbf{u}^i)Z(\mathbf{u}^j)\} - [E\{\bar{Z}_A\}]^2 \quad (3)$$

The variance of the spatial average is expanded below. The first term is equivalent to the expected value of the conditional covariance between nodes plus the quadratic of the expected value of the spatial average, and the second term is the quadratic of the expected value of the spatial average.

$$\text{Var}\{\bar{Z}_A\} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N [C\{Z(\mathbf{u}^i)Z(\mathbf{u}^j)\} + (E\{\bar{Z}_A\})^2] - [E\{\bar{Z}_A\}]^2$$

The previous equation is simplified and the quadratic of the expected value of the spatial averages are canceled out. Where, $\text{Cov}\{Z(\mathbf{u}^i)Z(\mathbf{u}^j)\}$ corresponds to the covariance between two random variables conditioned to the available data.

$$\text{Var}\{\bar{Z}_A\} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N C\{Z(\mathbf{u}^i)Z(\mathbf{u}^j)\} \quad (4)$$

The conditional covariance requires the following steps: Define an $n \times n$ covariance matrix between available n data as C_{11} .

$$C_{11} = \begin{bmatrix} C(\mathbf{u}_1 - \mathbf{u}_1) & \dots & C(\mathbf{u}_1 - \mathbf{u}_n) \\ \vdots & \ddots & \vdots \\ C(\mathbf{u}_n - \mathbf{u}_1) & \dots & C(\mathbf{u}_n - \mathbf{u}_n) \end{bmatrix} \quad (5)$$

Define the covariance matrix between n data and N locations of the discretized domain as C_{12} . Also the notation $Z(\mathbf{u})$ will be simplified in the expressions by just vector \mathbf{u} .

$$C_{12} = \begin{bmatrix} C(\mathbf{u}_1 - \mathbf{u}^{(1)}) & \dots & C(\mathbf{u}_1 - \mathbf{u}^{(N)}) \\ \vdots & \ddots & \vdots \\ C(\mathbf{u}_n - \mathbf{u}^{(1)}) & \dots & C(\mathbf{u}_n - \mathbf{u}^{(N)}) \end{bmatrix} \quad (6)$$

Define the covariance matrix between N locations of the discretized domain in nodes as C_{22} .

$$C_{22} = \begin{bmatrix} C(\mathbf{u}^{(1)} - \mathbf{u}^{(1)}) & \dots & C(\mathbf{u}^{(1)} - \mathbf{u}^{(N)}) \\ \vdots & \ddots & \vdots \\ C(\mathbf{u}^{(N)} - \mathbf{u}^{(1)}) & \dots & C(\mathbf{u}^{(N)} - \mathbf{u}^{(N)}) \end{bmatrix} \quad (7)$$

The expression for the conditional covariance matrix of N node locations given n conditioning values is given after combining covariances matrices C_{11} , C_{12} and C_{22} . The calculations of all the covariances use an input model covariance.

$$C_{(u^1, \dots, u^N | u_1, \dots, u_n)} = C_{22} - C_{12}^T C_{11}^{-1} C_{12} \quad (8)$$

The kriging system is given in Equation (9). This term is observed in the conditional covariance equation.

$$[\lambda] = C_{11}^{-1} C_{12} \quad (9)$$

The covariance matrix between the n data and the N nodes is transposed and multiplied by the kriging weights.

$$C_{12}^T [\lambda] = \begin{bmatrix} \sum_{k=1}^n \lambda_k^{(1)} C(\mathbf{u}_k - \mathbf{u}^{(1)}) & \dots & \sum_{k=1}^n \lambda_k^{(N)} C(\mathbf{u}_k - \mathbf{u}^{(1)}) \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^n \lambda_k^{(1)} C(\mathbf{u}_k - \mathbf{u}^{(N)}) & \dots & \sum_{k=1}^n \lambda_k^{(N)} C(\mathbf{u}_k - \mathbf{u}^{(N)}) \end{bmatrix} \quad (10)$$

The previous matrix is substituted by the outcome of minimizing the kriging variance Equation (11). The covariance between random variables in the presence of conditioning data is deduced (12) (Neufeld, Ortiz, & Deutsch, 2005)

$$\sum_{k'=1}^n \lambda_{k'} C_{kk'} = C_{k0} \text{ where } k = 1, \dots, n \quad (11)$$

$$C_{(u^1, \dots, u^N | u_1, \dots, u_n)} = C_{22} - \begin{bmatrix} \sum_{k=1}^n \sum_{k'=1}^n \lambda_k^{(1)} \lambda_{k'}^{(1)} C(\mathbf{u}_k - \mathbf{u}_{k'}) & \dots & \sum_{k=1}^n \sum_{k'=1}^n \lambda_k^{(N)} \lambda_{k'}^{(1)} C(\mathbf{u}_k - \mathbf{u}_{k'}) \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^n \sum_{k'=1}^n \lambda_k^{(1)} \lambda_{k'}^{(N)} C(\mathbf{u}_k - \mathbf{u}_{k'}) & \dots & \sum_{k=1}^n \sum_{k'=1}^n \lambda_k^{(N)} \lambda_{k'}^{(N)} C(\mathbf{u}_k - \mathbf{u}_{k'}) \end{bmatrix} \quad (12)$$

The simplified equation of the previous matrix is a function of the covariances between node locations, the set of weights and the covariances between conditioning data. Where no conditioning data are present, the covariances are identical to the input covariance model.

$$C_{(\mathbf{u}^{(i)}, \mathbf{u}^{(j)} | u_1, \dots, u_n)} = C_{ij} - \sum_{k=1}^n \sum_{k'=1}^n \lambda_k^{(j)} \lambda_{k'}^{(i)} C_{kk'} \quad (13)$$

The equation of the conditional covariance between random variables is replaced in Equation (4) to obtain the conditional variance of the spatial average.

$$Var\{\bar{Z}_A\} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left(C_{ij} - \sum_{k=1}^n \sum_{k'=1}^n \lambda_k^{(j)} \lambda_{k'}^{(i)} C_{kk'} \right) \quad (14)$$

The variance of the spatial average is expanded. The two terms show their influence on the total $Var\{\bar{Z}_A\}$. That equation accounts for the covariances of all the nodes that are inside the domain A. The non stationary covariance is reproduced in the presence of conditioning data. The first term is the average of the $N \times N$ covariances between nodes that belong to domain A, and the second term is the average of $N \times N$ nodes combinations of redundancy measures of the available data regard to the nodes.

$$Var\{\bar{Z}_A\} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (C_{ij}) - \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left(\sum_{k=1}^n \sum_{k'=1}^n \lambda_k^{(j)} \lambda_{k'}^{(i)} C_{kk'} \right) \quad (15)$$

The expected value of the spatial average is represented by:

$$E\{\bar{Z}_A\} = E\left\{ \frac{1}{N} \sum_{i=1}^N z_i^* \right\} = \frac{1}{N} \sum_{i=1}^N E\{z_i^*\} = \frac{1}{N} \sum_{i=1}^N E\left\{ \sum_{k=1}^n \lambda_k^{(i)} z_k \right\} \quad (16)$$

assume $E\{z_k\} = z_k \implies E\{\bar{Z}_A\} = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^n \lambda_k^{(i)} z_k$

The previous two equations will be useful to evaluate non ergodic fluctuations. In an ideal ergodic domain the variance is zero because the mean of a realization of standard Gaussian distribution is zero, then, the variance of the means of many realizations should be zero. Equation (15) show statistics fluctuations, these statistical fluctuations are part of the uncertainty in the mean for the domain A. These variations depend of the relation size of the domain and the range of correlation. When the size of the domain becomes in the order of 10 times the range of correlation the fluctuations converge to zero. By the other hand, the uncertainty in the input parameters is an additional uncertainty that must be accounted for.

3. Application

A simple scenario is used to show the influence of the conditioning data in the evaluation of the covariance. Three samples are located in an area of 150 meters \times 150 meters. These samples are standard normal Gaussian. The variogram model is spherical and isotropic.

$$\gamma(\mathbf{h}) = 0.2 + 0.8 \cdot sph_{a=150}(\mathbf{h})$$

The area of study is discretized by nine nodes $\mathbf{u}^{(i)}$, $i = 1, \dots, 9$. The covariance between node $\mathbf{u}^{(2)}$ and node $\mathbf{u}^{(3)}$ given three conditioning data \mathbf{u}_k , $k = 1, \dots, 3$ requires the kriging weights for each node σ_k , $k=1, \dots, 3$, the covariance between samples locations $C_{kk'}$ and the covariance between $\mathbf{u}^{(2)}$ and $\mathbf{u}^{(3)}$. For the node $\mathbf{u}^{(2)}$ the kriging weights result -0.063, 0.109 and 0.508 and for the node $\mathbf{u}^{(3)}$ the kriging weights result -0.094, 0.208 and 0.192. As expected, the kriging weights are proportional to the distance between node and samples. Furthermore, the covariance matrix between the samples locations is as follow:

$$C_{kk'} = \begin{bmatrix} 1 & 0.373 & 0.250 \\ 0.373 & 1 & 0.276 \\ 0.250 & 0.276 & 1 \end{bmatrix}$$

From Figure 3, the size of each node is 50 meters \times 50 meters, then, the covariance $C(\mathbf{u}^{(2)}, \mathbf{u}^{(3)})$ between adjacent evaluated nodes has the lag distance \mathbf{h} equal to 50 meters. The covariance between $\mathbf{u}^{(2)}$ and $\mathbf{u}^{(3)}$ given three conditioning samples is solved in the next equation:

$$\begin{aligned} C_{(\mathbf{u}^{(2)}, \mathbf{u}^{(3)}) | \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3} &= C_{23} - \sum_{k=1}^3 \sum_{k'=1}^3 \lambda_k^{(3)} \lambda_{k'}^{(2)} C_{kk'} \\ &= 0.41 - (-0.094 \cdot -0.063 \cdot 1 + -0.094 \cdot 0.109 \cdot 0.373 + -0.094 \cdot 0.508 \cdot 0.250; \\ &\quad + 0.208 \cdot -0.063 \cdot 0.373 + 0.208 \cdot 0.109 \cdot 1 + 0.208 \cdot 0.508 \cdot 0.276; \\ &\quad + 0.192 \cdot -0.063 \cdot 0.250 + 0.192 \cdot 0.109 \cdot 0.276 + 0.192 \cdot 0.508 \cdot 1) \\ &= 0.41 - 0.138 = 0.272 \end{aligned}$$

To verify the non stationary covariance in the presence of conditioning data, two other nodes with the same vector lag distance \mathbf{h} are evaluated, namely the covariance between $\mathbf{u}^{(4)}$ and $\mathbf{u}^{(5)}$ that are near to the conditioning data.

Where, the kriging weights for the node $\mathbf{u}^{(4)}$ result 0.258, 0.004 and 0.457; for the node $\mathbf{u}^{(5)}$ result 0.141, 0.363 and 0.387. Like the previous evaluation, the equation of the covariance conditioning to the data is as follow:

$$\begin{aligned}
 C_{(\mathbf{u}^{(4)}, \mathbf{u}^{(5)}) | \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3} &= C_{45} - \sum_{k=1}^3 \sum_{k'=1}^3 \lambda_k^{(5)} \lambda_{k'}^{(4)} C_{kk'} \\
 &= 0.41 - (0.141 \cdot 0.258 \cdot 1 + 0.141 \cdot 0.004 \cdot 0.373 + 0.141 \cdot 0.457 \cdot 0.250; \\
 &\quad + 0.363 \cdot 0.258 \cdot 0.373 + 0.363 \cdot 0.004 \cdot 1 + 0.363 \cdot 0.457 \cdot 0.276; \\
 &\quad + 0.387 \cdot 0.258 \cdot 0.250 + 0.387 \cdot 0.004 \cdot 0.276 + 0.387 \cdot 0.457 \cdot 1) \\
 &= 0.41 - 0.337 = 0.073
 \end{aligned}$$

The data of the example are illustrated in Figure 3. The covariance C_{23} and C_{45} without conditioning nada are equal to 0.41 because the vector distances in both cases are the same; however, in the presence of conditioning nada, the conditional covariances over the domain become non stationary and depend on the distance to the conditioning data. The conditional covariance $C_{23|1,2,3}$ is farther from the condition data than the other conditioning covariance $C_{45|1,2,3}$, then, the example shows that $C_{23|1,2,3}$ is greater than $C_{45|1,2,3}$ because the second term depend on kriging weights and is subtracted from the constant value 0.41 to get the conditional covariance of these nodes distant in 50 meters. For instance, the nodes $\mathbf{u}^{(4)}$ and $\mathbf{u}^{(5)}$ obtain greater kriging weights than nodes $\mathbf{u}^{(2)}$ and $\mathbf{u}^{(3)}$ because they are located close to the conditioning data. As a result, the second term is 0.337 for nodes $\mathbf{u}^{(4)}$ and $\mathbf{u}^{(5)}$ greater than 0.138 for nodes $\mathbf{u}^{(2)}$ and $\mathbf{u}^{(3)}$.

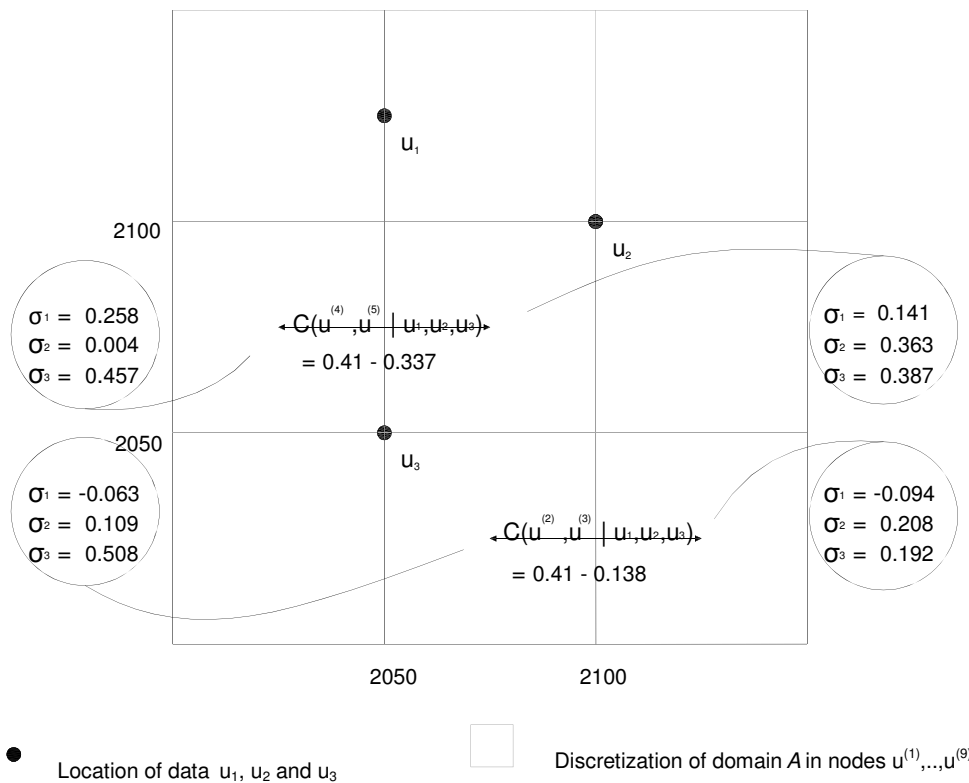


Figure 3: Graphic of the non stationary covariance in the presence of conditioning data.

The example shows that conditional covariance $C_{45|1,2,3}$ near the data results in 18 % of the covariance model. Otherwise, the conditional covariance $C_{23|1,2,3}$ located a little far from the data results in 66 % of the covariance model (0.41). Those results are verified by simulating 1000 realizations; the $C_{45|1,2,3}$ gives 0.081 and $C_{23|1,2,3}$ gives 0.337. That is, the covariance given n conditioning data increases and becomes close to covariance model as the evaluated nodes are far from the conditioning data.

3.1. Verification of Analytical Variance of the Spatial Average

The data and covariance model from the previous example is used to demonstrate the drop of the variance as the domain, A , increases. The next equation shows the variance of the spatial average of the domain that is discretized by nine nodes with three conditioning data. The first term corresponds to the covariance average between nodes equal to 0.3204, this value does not depend on the conditioning data; and the second term equal to 0.2022 corresponds to the term that accounts the location of the conditioning data. The expected value of the spatial average 0.0964 accounts the values of the input data.

$$\begin{aligned} Var\{\bar{Z}_A\} &= \frac{1}{9^2} \sum_{i=1}^9 \sum_{j=1}^9 (C_{ij}) - \frac{1}{9^2} \sum_{i=1}^9 \sum_{j=1}^9 \left(\sum_{k=1}^3 \sum_{k'=1}^3 \lambda_k^{(j)} \lambda_{k'}^{(i)} C_{kk'} \right) \\ &= 0.3204 - 0.2022 \\ &= 0.1182 \\ E\{\bar{Z}_A\} &= \frac{1}{9} \sum_{i=1}^9 \sum_{k=1}^3 \lambda_k^{(i)} z_k \\ &= 0.0964 \end{aligned}$$

These values are shown in Table 1, where the parameters of the input covariance model are kept constant as the size of the domain A is increased from 150 meters to 1550 meters. The first term becomes smaller as the size of the domain increases because the covariances between distant nodes are less; the second term becomes smaller because the conditioning data are located farther from the nodes as the size of the domain increases.

$A(Xd \times Yd)$	First term	Second term	Analytical Model
150×150	0.3204	0.2022	0.1182
350×350	0.0802	0.0158	0.0644
550×550	0.0351	0.0026	0.0325
750×750	0.0195	0.0008	0.0187
950×950	0.0124	0.0003	0.0121
1150×1150	0.0086	0.0001	0.0085
1350×1350	0.0063	0.0001	0.0062
1550×1550	0.0048	0	0.0048

Table 1: Change of the variance of the spatial average with different size of domains.

As expected, the variance becomes smaller as the size of the domain is increased, the variance approaches zero asymptotically, but it is practically zero when the ratio of the domain size and range of correlation is around 10. These analytical values are compared to the numerical model in Figure 4.

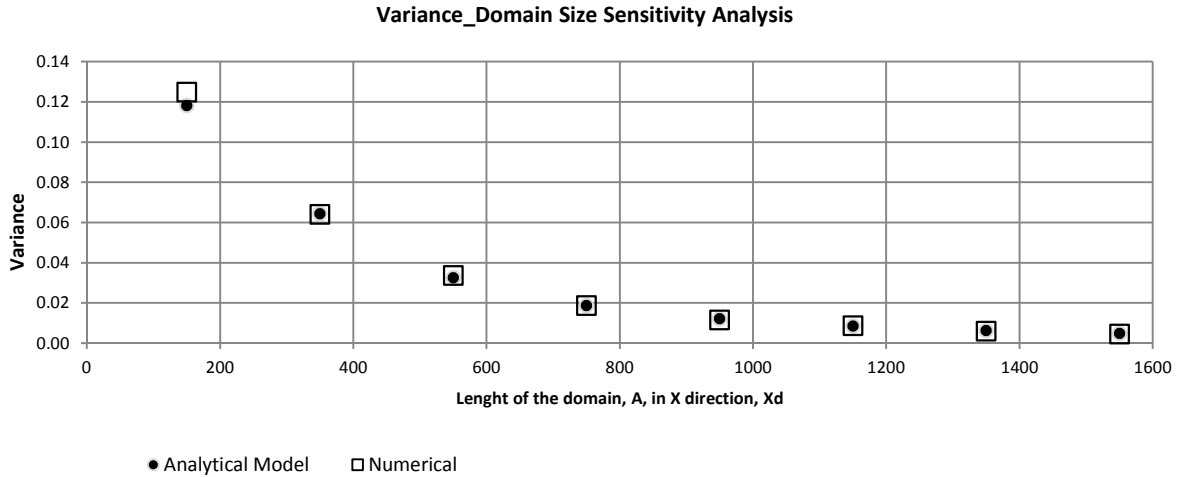


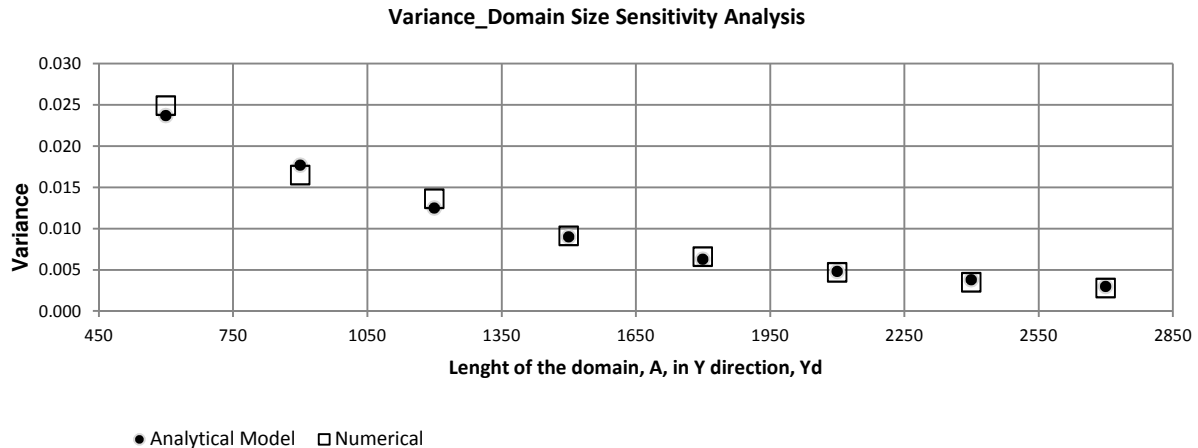
Figure 4: Synthetic data, non-ergodic variance of spatial average with different domain size.

As expected, both analytical model and numerical show the decrease of the variance of the spatial average as the domain increases; however, slight differences are observed due to the limit number of samples. The ratio of the domain size and the range of correlation in the domain 150×150 is 1 and in the domain 350×350 is 2.3. Those ratios correspond to non ergodic domain because the ratios are less than 10. The numerical approach show slight variations of the variance of the spatial average. For instance 200 realizations of the domain 150×150 show variance 0.13 and 2000 realizations show variance 0.12.

A second example is used to compare values of variance from 200 simulations (numerical) and analytical model. The “red” data contain 60 values of gold that are located in area of 500 meters \times 600 meters. These values are transformed to normal Gaussian score and their anisotropic variogram is defined by:

$$\gamma(\mathbf{h}) = 0.2 + 0.8 \cdot sph_{ah1=250, ah2=150}(\mathbf{h})$$

The size of nodes is 50 meters \times 50 meters, the domain size ($Xd \times Yd$) 500 meters \times 600 meters is increased eight times proportionally until the domain size reach ($Xd \times Yd$) 2250 meters \times 2700 meters. The largest domain is equivalent to 10 times the range of correlation. The previous example used synthetic data that contained 3 sample locations. Meanwhile, the current example shows a real scenario of 60 values. More samples and real scenario evaluate fairly the analytical model against the classical result of many simulations.



4. **Figure 5:** Variance of spatial average with different domain sizes for the Red data.

The variance of the analytical model and numerical are similar, the slight difference is due to the numerical model being sensitive to the random generator of realizations. The decrease of expected fluctuations as the size of the domain increase is reproduced as the previous example. The examples validate the analytical model.

5. Conclusions

The concept of ergodicity states that the spatial average of a random stationary function (RF) $Z(\mathbf{u})$ over a domain A converges to the expected value $m=E\{Z(\mathbf{u})\}$ when A tends to infinity. The expected value in normal score Gaussian is zero, then, the statistical fluctuations between realizations is projected equal to zero.

Statistical fluctuations of realizations are reduced as the ratio between domain size and range of correlation increase. The examples show that fluctuations statistical practically reach zero at a ratio of 10.

The analytical model is validated with numerical results of many realizations in two examples; it is observed that the results of analytical model are congruent with the statistical fluctuations of many simulations.

6. References

- Chilès, J.-P., & Delfinier, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. New York: John Wiley & Sons.
- Deutsch, C. V., & Journel, A. (1998). *GSLIB: Geostatistical Software Library and User's Guide* (2nd Edition ed.). New York: Oxford University Press.
- Deutsch, C. V., Leuangthong, O., & Ortiz C., J. (2006). A Case for Geometric Criteria in Resources and Reserves Classification. *Eight Annual Report of the Centre for Computational Geostatistics* (p. 21). Edmonton: Department of Civil & Environmental Engineering-University of Alberta; Department of Mining Engineering-University of Chile.
- Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press.
- Leuangthong, O., McLennan, J., & Deutsch, C. V. (2005). *Acceptable Ergodic Fluctuations and Simulation of Skewed Distributions*. Edmonton: Centre for Computational Geostatistics, Department of Civil&Environmental Engineering University of Alberta.
- Neufeld, C. T., Ortiz, J. M., & Deutsch, C. V. (2005). A Non Stationary Correction of the Probability Field Covariance Bias. *Seventh Annual Report of the Centre for Computational Geostatistics*. Edmonton: University of Alberta; Department of Mining Engineering-University of Chile.
- Pyrzcz, M., & Deutsch, C. (2002). Two artifacts of probability field simulation. *Mathematical Geology*, 33, 775-800.