Uncertainty Assessment and Computational Cost of Conditional Sequential Simulation in 3D Modeling

Dr.eng. Mohammad Saleh Al-Abdalla

Faculty of Civil Engineering – Damascus University

Abstract

Conditional Sequential Simulation Processes takes relatively long computational time in 3D modeling problems depending on many relevant factors like: type of the conditional method used, model of the Variogram function, size of the spatial framework (grid) and obviously number of the repeated simulations. On the other hand, the uncertainty of the simulation depends on many factors; like simulation method, Variogram model, the nature of data, its distribution, the spatial grid framework etc. The present paper study both subjects, (1) the uncertainty analysis and assessment and (2) computational cost analysis and performance.

Through this study, two methods well known in geostatistics were used, namely Conditional Sequential Gaussian Simulation (SGSim) and Conditional Sequential Indicator Simulation (SISim). In addition, two Variogram models were applied, the Spherical Variogram and the Gaussian Variogram. The theoretical background for each methods has been explained briefly as well as their algorithmic steps have been specified. On the other hand variogram models were not discussed and one can find much information on this in the relevant literature.

For the purpose of this research, many tests were applied using real geo-referenced data freely available on the web. In more than 200 tests that performed, some factors were fixed as they have no much effect on the final accuracy and speed, and three factors only were changed, namely; the size and structure of the 3D grid, the Variogram function and number of simulations each time.

Those tests showed that the uncertainty of results is improved when increasing the size of the grid and number of simulations, but this demands more computational time. Still, we need an answer the most relevant questions: What is the appropriate size of grid? How many simulations required? Which Variogram model should we use?, in order to obtain the best accurate results with a minimum computational cost?.

After many tests and the detailed statistical analysis of the results, the study extracted significant information for optimization the Conditional Sequential Simulation in 3D modeling and has given clear, precise answers to the questions proposed in this research.

Keywords: *Conditional Sequential Simulation, 3D Modeling by Simulation, Uncertainty Assessment, Simulation Performance and Computational Cost.*

تقييم الدقة والكلفة الحسابية للمحاكاة الشرطية المتتالية في النمذجة ثلاثية الأبعاد

د. محمد صالح العبدالله

كلية الهندسة المدنية – جامعة دمشق

قسم الهندسة الطبوغرافية

الملخص

تستغرق عمليات المحاكاة الشرطية المتتالية في النمذجة ثلاثية الأبعاد للمسائل الكبيرة الحجم وقتاً حسابياً طويلاً نسبياً ويتغير هذا الوقت تبعاً لعوامل أساسية عديدة مثل نوع الطريقة الشرطية المستخدمة، نموذج تابع التغيرية (الفاريوغرام)، حجم الإطار المكاني وبالطبع عدد تكرارات المحاكاة. من جهة أخرى دقة نتائج المحاكاة بدورها تتغير تبعاً لعدد كبير من العوامل منها الطريقة المستخدمة في المحاكاة، نموذج التغيرية (الفاريوغرام)، طبيعة البيانات وتوزعها، الإطار المكاني المصفوفي ..الخ. تركز هذه المقالة على دراسة هذين الموضوعين وهما (1) تحليل وتقييم الأخطاء وكذلك على (2) تحليل زمن الحساب والأداء.

في إطار هذه الدراسة تم استخدام الطريقتين الأكثر استخداماً في الجيوستاتستيك وهما المحاكاة الشرطية المتتابعة بموجب توزيع غاوص (SGSim) والمحاكاة الشرطية المتتابعة التصنيفية (SISim)، كذلك تم استخدام نموذجين فقط لتابع التغيرية (الفاريوغرام) وهما التابع الكروي والتابع الغاوصي. تم شرح الأساس النظري لكلا الطريقتين مع وضع الخوارزميات لكل منهما، لكنه لم يتم التطرق إلى شرح نماذج التغيرية المستخدمة فهي مشروحة في أغلب المراجع ذات الصلة.

بقصد إنجاز هذه الدراسة تم إجراء اختبارات عديدة باستخدام بيانات مكانية بمرجعية جغرافية وهذه البيانات متوفرة على الشبكة العنكبوتية مجاناً. في كل الإختبارت التي عددها تجاوز 200 إختبار تم تثبيت بعض العوامل التي ليس لها تأثير كبير على الدقة أو سرعة الحساب في حين تم تغيير حجم المصفوفة 3d للإطار المكاني، تابع التغيرية وعدد سيناريوهات المحاكاة في كل مرة.

أثبتت التجارب بأن دقة النتائج تتحسن عند زيادة حجم الإطار المصفوفي وكذلك عند زيادة عدد مرات المحاكاة ولكن هذا كله يكون على حساب زمن الحساب. وهنا نحتاج إلى إجابة على الأسئلة التالية: ما هو حجم الإطار المصفوفي؟ كم عدد مرات المحاكاة؟ و ماهو نموذج الفاريوغرام؟ كي نحصل في النهاية على أفضل دقة وبصورة نختصر فيها من زمن الحساب إلى الحد الأدنى؟

بعد التجارب العديدة والتحليل التفصيلي الإحصائي للنتائج استخلصت هذه الدراسة معلومات هامة ومفيدة للحل الأمثل للمحاكاة الشرطية المتتالية في النمذجة ثلاثية الأبعاد وأعطت إجابات واضحة ودقيقة على الأسئلة المطروحة في البحث.

الكلمات المفتاحية: المحاكاة الشرطية المتتالية، النمذجة ثلاثية الأبعاد باستخدام المحاكاة، تقييم الدقة، الأداء والكلفة الحسابية للمحاكاة.

Introduction

The new techniques in global positioning systems (GPS), as well as the recent developments in geographic information systems (GIS) and remote sensing, have been permitted the possibility of collecting a large amount of scientific, geo-referenced data. This developments created an increasing interest in geostatistical spatial data analysis and modeling [Chiles J., Delfiner P. (1999) Møller (2003), Banerjee et al. (2004) and Schabenberger & Gotway (2004)]. As it well known, *Statistics* forms the foundation of many scientific fields and applications, *Geostatistics* on the other hand, forms the basis of those scientific fields that is interested in the analysis and interpretation of geo-referenced data [Cressie (1993), Goovaerts, P. (1997), Bolstad W.M. Curran J.M (2007)]. In combination and cooperation with GIS techniques and data, *Geostatistics* became a respected scientific tool used in many applications: e.g. in remote sensing, one can detect spatial changes in land use and land cover, perform radiometric enhancements in the images or selecting the best image resolution for spatial data [Atkinson and Quattrochi, et al (2000)]. In GIS the generation, simulation or improvement of DEM's can be optimized using geostatistical methods [Brus and Heuvelink, (2007)]. One can find many other applications using those methods in water resources assessments and managements, in environmental sciences, forestry, agriculture, soil sciences, ecology, geology, metrology etc. [Christakos, G. (2005) , Hengl T. (2007), Lantuejoul C. (2002) , Mund Jan-Peter (2013)]. Geostatistical analysis is concerned with studying phenomena that have spatial or spatiotemporal extent and its variability using a collection of deterministic and stochastic tools in order to model the phenomenon by simulation with the Monte Carlo Method [Al-Abdalla M. (1998)]. The essence of 3D modeling continuity by simulation lies in the assumed relations between information and unknowns and between the various elements and characteristics of the available data [Al-Abdalla M. (1998)].

In geostatistics, conditional simulation is used to estimate, by Monte Carlo Methods, complicated nonlinear functions that depend explicitly on multivariate stochastic distributions. When the simulation domain is discrete, a sequential procedure can be considered (Journel 1989). This consists of prescribing an arbitrary ordering of all of the points of the domain, and simulating each point in turn according to a Conditional Gaussian distribution given the generated values of all the previous points. On the other hand, The method of Sequential Indicator Simulation (SIS) is type of conditional simulation that uses the indicator random function models, being binary. This method is ideally suited for simulating categorical variables controlled by two-point statistics.

Research Objectives

In earth sciences engineers face the problem of modeling spatial structures from limited data, especially in 3D. The data is few, sparse, and typically contains varying degrees of noise. Most often questions are raised such as:

(1) W*hat is happening (or existing) in certain unsampled locations*?, (2) *how much we are confident with the results after a simulation process takes place*?, (3) *assuming the Variogram models are known, does the uncertainty associated with those results meet our requirements*?, (4) *Do we have enough computational power to run as many simulations as we need to*?

The main objectives of this study is to present the results of a comparative study designed to evaluate uncertainty and performance of two different geostatistical simulation methods, namely the *Conditional Sequential Gaussian Simulation* (SGS) and the *Conditional Sequential Indicator Simulation* (SIS). The two methods will be presented in later sections of this paper. With ***adequate computing power***, simulation by the Monte Carlo Method is possibly the best way to study the uncertainty associated with 3D modeling using *probabilistic multivariate transfer functions*. The *frequency distribution* (histogram) reflects the uncertainty that can be obtained from a certain number of simulations yielding different equiprobable representations given the spatial structure (or variability) of the data. This structure is known as the *Variogram Model.*

Conditional Simulation Concept

The *Conditional Simulation* generates a *Random Fields* (RF) that simulate the spatial variability of the underlying random process $Z(x)$. Usually *Kriging* interpolation (or prediction) provides a minimum-variance unbiased estimator, while

kriging variance provides a measure of the local estimation uncertainty. The main advantage of *stochastic estimation* using kriging techniques is that the uncertainty (error variance) is estimated together with the prediction value. Unfortunately, unless a parametric distribution of the spatial error is assumed, the kriging approach cannot provide confidence intervals associated with the predicted values. With conditional simulation, the uncertainty estimation or the confidence intervals are guaranteed after performing a certain number of simulations. Generation of more realizations would lead to much precise estimation of the uncertainty. Journel and Huijbregts (1978) showed that the posterior estimation variance of *Conditional Simulation* is as twice as that of *Kriging*, thus one should emphasize that the objective of conditional simulation is not to obtain the best *unbiased estimator* that produced by a *Kriging* predictions. *Conditional Simulation* is useful to obtain information about the amount of variability remaining in the physical process $Z(x)$ conditioning with respect to the observations (Journel 1989), thus *Kriging Predictions* and *Conditional Simulation* address two different problems.

The only three elements, the mean function $\mu(\cdot)$, the covariance function $\mathrm{C}(\cdot)$ and the data vector z_d forms the basic elements of a conditional simulation. The conditionally simulated nodes $Z_{sc}(x)$ must pass through the data z_d, having unconditional mean $\mu(\cdot)$ and variance $\mathrm{C}(\cdot)$. *Kriging* predicator $Z_{ak}(x)$ would satisfy the requirements, because it does interpolate the data exactly and it is unbiased. However *Kriging* has a smoothing tendency, thus it does not possess enough variability in order to give a posterior probability distribution about the uncertainty. With *Conditional Simulation* we are able to generate an infinite number of possible realizations of a *Random Field* $\{Z_s(x), x \in D, s = 1 \to \infty\}$. From among the infinite simulations we choose those that meet certain condition

$$Z_s(x_a) = Z_0(x_a),\ \forall x_a \in D.$$

For example if we want the simulated model honors data values at the actual data locations, we set:

$$Z_{cs}(x_a) = Z_0(x_a),\ \forall x_a \in D,$$

Where x_a represents data locations.

This is known as *Conditional Simulation*, which has the same variability characteristics as the real observed phenomenon. This means that the simulated values $Z_{cs}(x_a)$ have the same first two experimentally found moments (the mean and the variance) representing the histogram of the real values $Z_0(x_a)$. Now consider the decomposition of the process into a kriging predicator and an *unconditional residual* (Journel and Huijbregts 1978).

$$Z_{cs}(x) = Z^*(x) + [Z_{us}(x) - Z^*_{us}(x)] \quad (1)$$

Where $Z_{cs}(x)$ is the conditional simulation, $Z^*(x)$ is the kriging estimators using the real data set (representing the estimated grid), $Z_{us}(x)$ is the unconditional simulation, and $Z_{sk}(x)$ is simple-kriging estimators using the unconditional simulated data. The two components of the right-hand side $Z^*(x)$ and $Z_{us}(x) - Z^*_{us}(x)$ are orthogonal. This orthogonality implies that $Z_{cs}(x)$ has the same unconditional covariance as (x) , that is $C(\cdot)$. The quantity $Z_{us}(x) - Z^*_{us}(x)$ can be obtained by kriging the difference between data values and the unconditionally simulated ones at data locations. Thus the above expression can rewritten as follows [Cressie (1993)]

$$Z_{cs}(x) = Z_{us}(x) + C(x)' \cdot \sum^{-1}(z_d - z_{us}) \quad (2)$$

Where $Z_{cs}(x)$ and $Z_{us}(x)$ are the conditional and unconditional simulations respectively,

$C(x)' \equiv C(x_d, x_g), \forall x_d \in D, \forall x_g \in G$: is the

Covariance vector between data nodes *D* and the simulated grid nodes *G*,

$\sum^{-1}(z_d - z_{us})$ is the *Variance-Covariance* Matrix between the data values and the simulated ones,

z_d *and* z_g are two vectors representing actual data and the simulated ones at the data node locations respectively.

Sequential Gaussian Simulation (SGS)

Simple or Ordinary kriging is used to obtain estimates of the necessary conditional distribution defined by the only the two Gaussian parameters; namely its mean and variance. The simulations are then drawn randomly from this distribution using *inverse transform* method. Finally, the results of the Gaussian simulation are transformed back into the original data space. In general, the principle of *Conditional Sequential Simulation*, is once the new value simulated, it is added to the original set of

conditioning data, and the procedure repeated. Finally all simulated nodes will have the same initial spatial structure provided that all node values at data locations preserved. The principle of *Conditional Sequential Simulations* can be described as follows:

Consider $f(z_1, z_2, .., z_n \mid z_0)$ is the *cpdf*, where z_0 denotes the conditioning data at n_0 locations.

This probability function can be defined as

$$f(z_1, z_2, .., z_n \mid z_0) = f(z_1 \mid z_0) \cdot f(z_2 | z_1 \cup z_0) \ldots \cdot f(z_n | z_1, z_2, .., z_{n-1} \cup z_0) \quad (3)$$

Thus the generation of a realization by *Sequential Simulation* takes the following steps:

Algorithm (1):

1. Draw a value z_1 from the conditional probability distribution f_1 given the set z_0 as conditioning data,
2. Draw a value z_2 from the conditional probability distribution f_2 given $z_0 \cup z_1$ as conditioning data,
3. Draw the next a value z_i from the conditional probability distribution f_i in the same way and repeat the process.
4. Draw the last value z_n from the conditional probability distribution f_n given the set $(z_n | z_1, z_2, .., z_{n-1} \cup z_0)$ as conditioning data.

Remark 1: in case unconditional simulation is needed one should reduce the set of conditioning data to the null set; all simulations would be replaced by drawing from the marginal distribution f_1.

Remark 2: There is no restriction on the spatial locations of the random variables yielding an algorithm that can be equally applied to generate one or more variables on either a regular or irregular grid.

However, it remains the problem of determining the *Cumulative Probability Distribution Function* (*cpdf*) of any single random variable given any set of conditioning data. This problem has been solved for the Gaussian distribution, where the data first are transformed to the standard Gaussian values.

If the continuous phenomenon $\{Z(x), x \in D\}$ is generated by the sum of a number of independent sources $\{y_k(x), x \in D, k = 1, .., K\}$ with similar spatial distributions then the phenomenon can be modeled by a *Multi-Gaussian Random Fields Model*. Multi-Gaussian models are extremely congenial (good-natured), well understood, and they have large record of successful applications. A random function is said to be Gaussian or Multi-Gaussian if any linear combination of its variables follows the Gaussian distribution,

$$Z(x) = \sum_{k=1}^{K} \lambda_k Y_k(x) \equiv Gaussian \quad (4)$$

The *Gaussian Function* is unique for its analytical simplicity and for being the extreme distribution of many analytical theorems globally known as '*Central Limit Theorem*'.

Sequential Gaussian Simulation Algorithm

The general procedure for generating a simulation of a multivariate Gaussian field is provided by the sequential principle described in *Algorithm* (1). Each variable is simulated sequentially according to its Gaussian *Cumulative Distribution Function cdf*. The conditioning data consists of all original data and all previously simulated values found within a predefined neighborhood. The *SGS* algorithm proceeds as follows (*Deutsch and Journel 1992, Deutcsh 2002*):

Algorithm (2):

1. Determine the *cpdf* of the random variable that represents the entire study area (the distribution of z-data).
2. Perform the normal score transform of z-data into y-data with the standard normal *cpdf*, i.e. the conditioning data should be transformed into *Standard Gaussian*.
3. Given the model of the *Semivariogram*, compute the covariance table of the transformed conditioning data.
4. Define a random path through all grid nodes. The path visits each node only once. At each grid node retain a specific number of neighboring data including both original y-data and previously simulated ones.
5. At each grid node to be simulated
   a. determine the conditioning data within the search distance,
   b. use kriging, with the normal score variogram model, to determine the Gaussian parameters

(mean and variance) of the *cpdf* of the random variable at that node,

c. draw a simulated value from that *cpdf*.
d. add this node to the original data set,
e. Go back to step (5.a) and repeat the process until all grid nodes have been visited.

6. Back transform the simulated normal values into simulated values for the original variable.

$$z(x) = G^{-1}\big(y(x)\big), x \in D \qquad (5)$$

Remark 1:

the first condition for *Sequential Gaussian Simulation* SGS is that the conditioning data are multivariate Standard Gaussian with zero-mean and unit variance. Most earth science data do not present symmetric Gaussian histogram. In this case a non-linear transformation should be applied in order to obtain a standard Normal y-data.

Sequential Indicator Simulation (SIS)

Given a set of spatially distributed data and grid node system, SIS is a procedure for estimating a new value at any non-sampled location (grid node) by Indicator kriging. The SIS procedure is achieved through sequential estimation of the Cumulative Probability Distribution Function (*cpdf*) at each grid node. The *cpdf* provides the probability that at a particular location the variable of interest does not exceed a certain threshold. This probability is conditioned to the initial data values and to all previously simulated ones. After the *cpdf* has been estimated, the simulation of the corresponding value is achieved by Monte Carlo, where a random number is drawn from the uniform distribution $U(0,1)$ then the simulated value is computed from the inverse of the *cpdf* by using the *inverse transform method*.

The theory behind the SIS technique is as follows:

Suppose $\{Z(x), x \in D\}$ is the random process defines the conditioning data. Define the indicator random variable at location x for the threshold z_0 by using the binary transformation (Journel 1989),

$$I(x, z_0) = \begin{cases} 1 & if \quad Z(x \le z_0) \\ 0 & otherwise \end{cases} \qquad (6)$$

where the conditional expected value of the indicator random variable $I(x, z_0)$ is defined as

$$E\{I(x, z_0)|Z(x)\} = \Pr\{Z(x) \le z_0\} \qquad (7)$$

Consider A threshold z_0 is defined by $z_0 = G^{-1}(1-p)$, where p is the mean of the indicator value, and *G(z)* is the cumulative Gaussian distribution function. SIS is achieved by simulating values of a standard Gaussian variable and applying the threshold z_0 to the result. Therefore one can estimate the value of the conditional probability defined above by be estimating the corresponding indicator conditional expectation using *Indicator Kriging* from the indicator *inverse transform* of the conditioning data. The SIS algorithm estimates *cpdf* for all classes at the class limits z_0, by using *Simple or Ordinary Kriging.*

Sequential Indicator Simulation Algorithm

Before performing the *Sequential Indicator Simulation*, the range of each category is established along with its variability model (or the indicator covariance function $C_1(h; z_0)$, that is needed for each thresholds z_0. Given an initial data set *z* and *K* indicator covariance functions, the SIS algorithm proceeds as follows (Hernandez & Srivastava 1990, Deutsch C.V. 2006):

Algorithm (3)

1. Transform the initial conditioning data into indicator data sets, so that the range of values taken by the attribute z is classified into *K* categories each associated with a certain threshold z_k. Code each conditioning value into a vector of *K* indicator values.

2. Transform the *K* indicator covariance models into the same number of kriging covariance matrices. Establishing the kriging systems in advance would speed up data search.

3. Define a random path through all grid nodes, so that each node is visited only once.

4. Now for each grid node to be simulated along the random path:
   a) Determine the conditioning data within the search distance.
   b) Retain the closest data points up to a specified maximum number per octant.
   c) Again for each threshold $z_k, k = 1, .., K$. Set up the kriging system using the indicator covariance model and solve the system.

d) Compute the *cpdf* estimate for the threshold as a linear combination of the indicator conditioning data,

$$Z(x) = \sum z_k \cdot I_k(x), \;\; Pr\{Z(x) \leq z_k\} \quad (8)$$

e) Draw a random number from the uniform distribution $(0,1)$, and simulate a value z by reading from *cpdf*, applying the *inverse transform* (*Monte Carlo Method*).
f) Transform the simulated values into a series of K indicator values according to the same K thresholds.
g) Add the simulated node obtained in the previous steps to the set of conditioning values.
h) Go back to step (4). Repeat until all grid nodes have been visited.

Remark 1: The implementation of the SIS is more demanding than other simulation methods. In addition more information is needed to establish the spatial variability structure that has to be reproduced by the simulation.

Remark 2: The simulated nodes with SIS algorithm is not continuous but pre-classified into a number of categories, or in other words the range of variability is split into number of classes, each simulated separately by SIS.

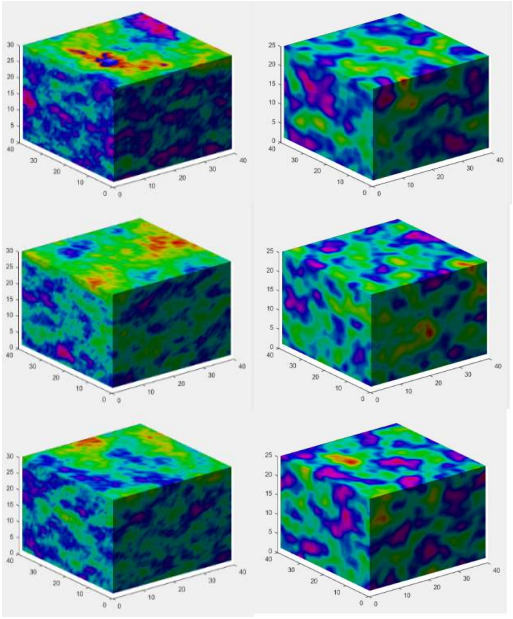
Remark 3: One can use as many classes as needed in order to obtain the required resolution, if only information about the variogram model is available. The classes need not to of equal amplitude, thus one can focus on that part of the range of variability most significant to the simulation. We can also use one model for all categories, which speed the execution time of the algorithm considerably, because in this case only one Kriging system has to be solved.

The advantages of Sequential Indicator Simulation (SIS) algorithm are:

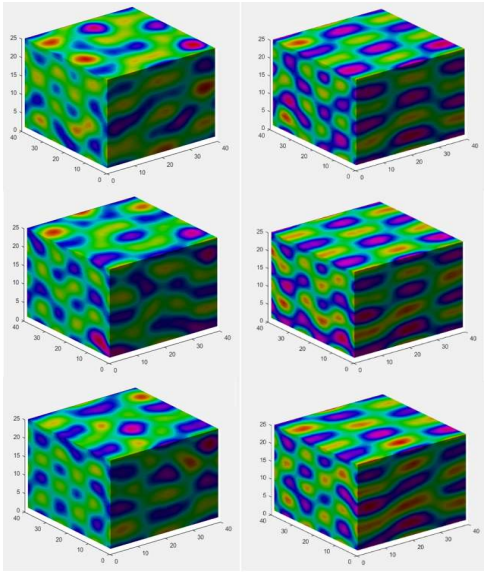
1. Conditioning is done as integral part of the simulation.
2. No assumption about the probability distribution (*cpdf*) is required.
3. It is not restricted to spatial forms of the covariance or variogram functions.
4. Qualitative or quantitative information can be included in the simulation.

Examples of Conditional Sequential Simulation using synthetic data

Some examples are presented here in order to show how the output would appear when applying the conditional SGS or SIS simulation methods on synthetic 3D data. figure(1) presents SGS method, and figure(2) presents SIS method, using Spherical and Gaussian Variograms respectively.



Figure(1) Conditional SGS Simulation, with Spherical Variogram (left) and Gaussian Variogram (right)



Figure(2) Conditional SIS Simulation, with Spherical Variogram (left) and Gaussian Variogram (right)

Implementation the Conditional Sequential Simulations With 3D data

Both conditional SGS and SIS simulations has been implemented using real 3D data related to the Historical Lake Water, Groundwater Levels, and Spring Flows in Central Florida. Maps and data used in the study belong the Lake Area in Central Florida, where data collected mainly by the U.S. Geological Survey over the last 80 years [see Scientific Investigations Report 2014–5032, O'Reilly A.M., and others (2014) for details]. The Study area is located within:

$[-82°07', -80°90']East\ , [27°65', 29°00']\ North.$

The data consists of measured data collected by the east-central Florida transient (ECFT) model [see Scientific Investigations Report 2012–5161, Sepúlveda N, et al. 2012], and by the Central Florida Artificial Neural Network Decision Support System (CFANN DSS). All data is available for download from the USGS site (http://dx.doi.org/ 10.3133/sir20145032).

Figure (3) presents the base map of the Lakes Area by USGS digital data using UTM projection-Zone 17. Groundwater observation locations and boundary districts are shown on the map. Water-level were available for 438 wells in the study area at different depths from 15m to 485m (1450feet). Figure(4) shows a 3D scatter plot of the data points.

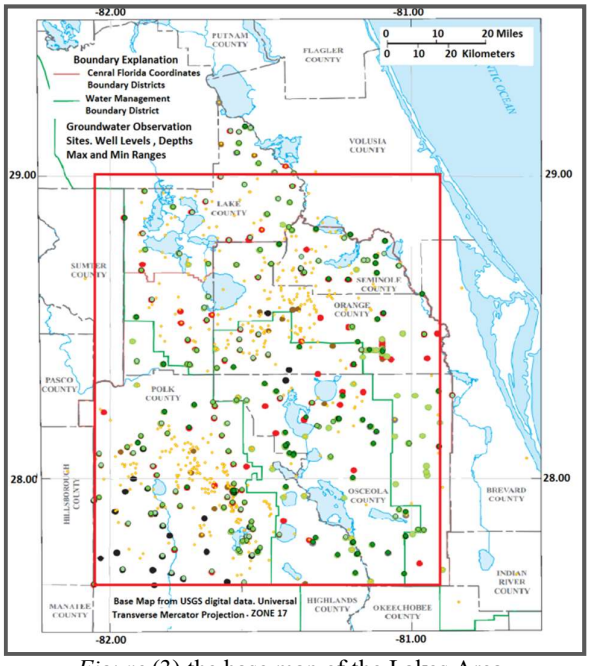


Figure (3) the base map of the Lakes Area

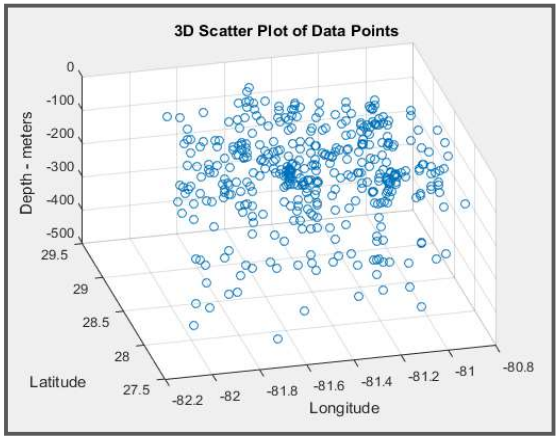


Figure (4) 3D scatter plot of the data points

Only two Variogram models were used; the Spherical Figures (5) and the Gaussian Figure (6). Distances here are represented by Geographic Units (for example 0.15deg≈16.5 km). Variogram Ranges that were fixed in all subsequent tests are: max=0.25, medium=0.20, min=0.15. the data Values were standardized to fit the Gaussian Distribution (1, 0.5), and the sill (maximum Semivariogram value):

$$\max(\gamma_x(h)) = \max(\frac{1}{2N(h)}\sum_{k=1}^{N(h)}(Z(x_i) - Z(x_j))^2) \approx 1.0$$

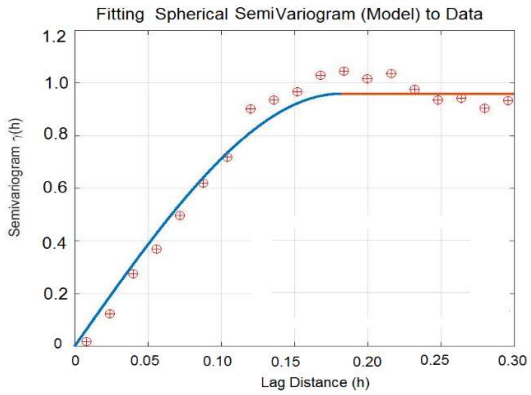


Figure (5) Fitting Spherical Variogram Model

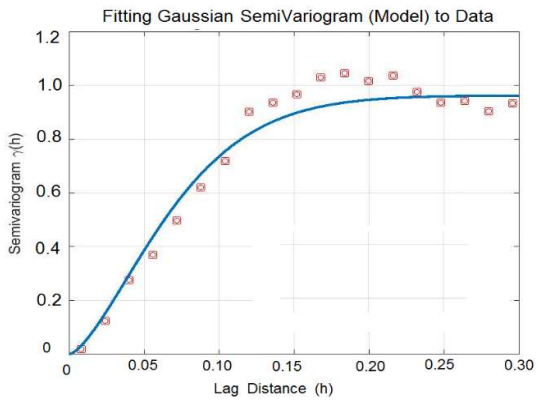


Figure (6) Fitting Gaussian Variogram Model

Implementation of the Conditional Sequential Gaussian Simulation (SGS):

Simulations have been completed using the following Parameters:

- Variogram Model Types: the *Spherical* (Left set of figures below), then Variogram Model Type : *Gaussian* (Right set of figures below)
- Variogram Ranges : max: 0.25, medium: 0.20, min: 0.15
- Simulation Seed Value = 1804910
- Maximum Conditioning Data = 25
- Number of all Simulations = 36
- Number of Data Points = 438
- Final Grid Arrangement:

	X	Y	Z
Grid Cells:	60	60	50
Cell Size :	0.013	0.013	0.028
Start Coord:	-81.80	29.00	0.05 (-485m)
Finish Coord:	-81.20	28.20	1.45 (-15m)
Coord. Units:	deg.	deg.	×1000feet

The output of this implementation is shown in the figure (7). Notice the differences between the left set where Spherical Model was used and the set on the right where the Gaussian Model was used. The second representation exhibits smoother patches.

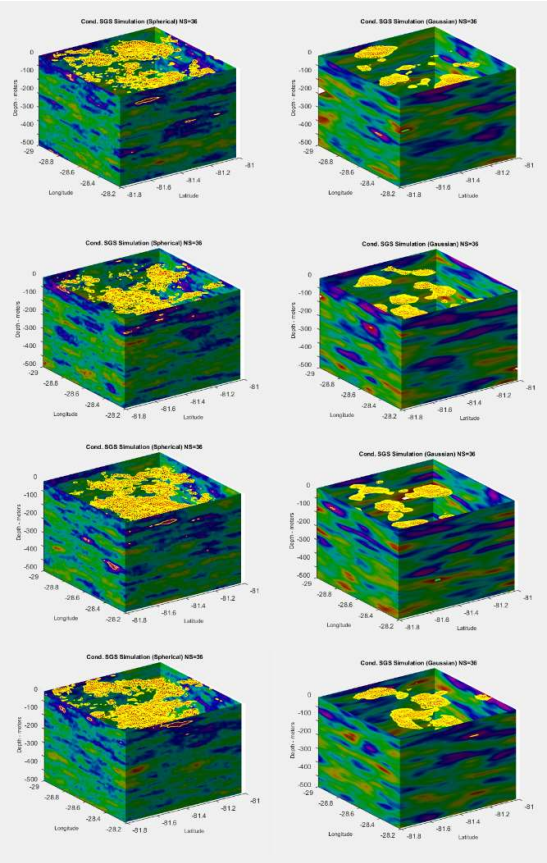


Figure (7) Figures of Sequential Gaussian Simulation (SGS) with Spherical Variogram (left) and Gaussian Variogram (right)

Implementation of the Conditional Sequential Indicator Simulation (SIS):

This type of simulation has been completed using the following Parameters:

- Variogram Model Types: the *Spherical* (Left set of figures below), then Variogram Model Type : *Gaussian* (Right set of figures below)
- Variogram Ranges : max: 0.25, medium: 0.20, min: 0.15
- Simulation Seed Value = 52470184
- Number of Indicators = 3
- Marginal Probabilities values: 0.65, 0.25, 0.10
- Maximum Conditioning Data = 25
- Number of all Simulations = 36
- Number of Data Points = 438
- Final Grid Arrangement:

	X	Y	Z
Grid Cells:	60	60	50
Cell Size :	0.013	0.013	0.028
Start Coord:	-81.80	29.00	0.05 (-485m)
Finish Coord:	-81.20	28.20	1.45 (-15m)
Coord. Units:	deg.	deg.	×1000feet

The output of this implementation is shown in the figure (8). Here the patches for both variograms are the same.

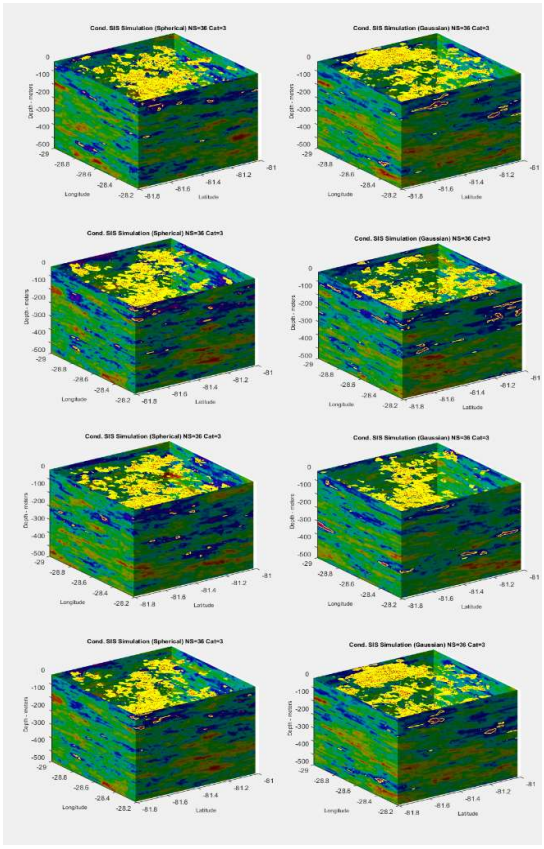
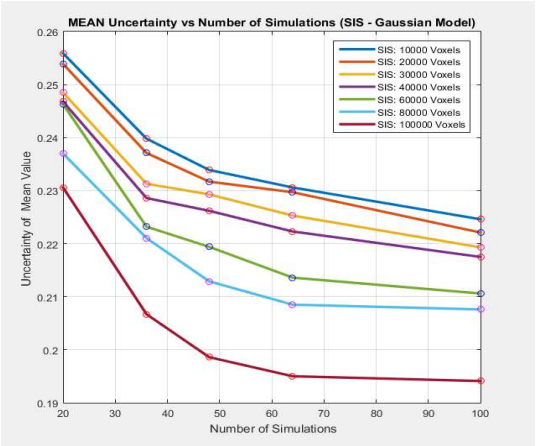


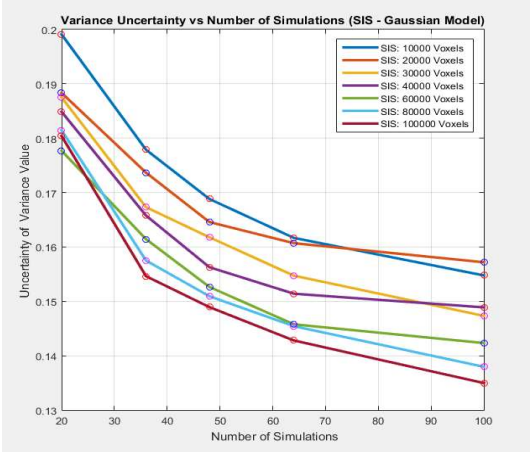
Figure (8) Figures of Sequential Indicator Simulation (SIS) with Spherical Variogram (left) and Gaussian Variogram (right)

Uncertainty Assessment with SIS

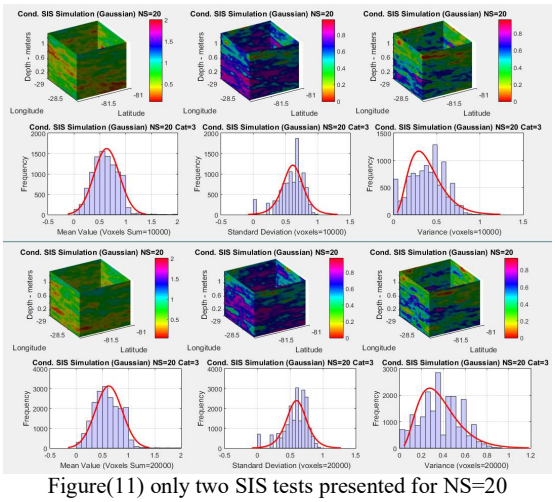
The figures below show the results of simulation tests that performed using *Conditional Sequential Indicator Simulation* (SIS) method with the two models (the Gaussian and the Spherical), changing grid structure (from 10000 total voxels till 100000 total voxels), and for each structure changing number of simulations (NS= 10, 20, 36, 48, 64 and 100 simulation). All test used the marginal probabilities values: 0.65, 0.25, 0.10, assuming that 25% belong to the groundwater data 10% to lakes data and 65% stands for unavailable data. After all simulations for each round are ready, the output of the *Mean value, Standard Deviation* and *Variance* could be computed and presented as shown in the figures (11,12,13). Then uncertainty for each of the three statistical measures was computed and registered. Those tests show that *Mean Uncertainty* decreases in the same way by increasing number of voxels or by increasing total simulations. After ~64 simulations, one can obtain better results only by increasing number of voxels as we see in the figure(9). On the other hand Standard deviation uncertainty or *Variance Uncertainty* (Variance=σ^2; σ *Standard deviation*) do not show stability after 64 simulations and their values continue decreasing beyond that, as we see in the figure (10).



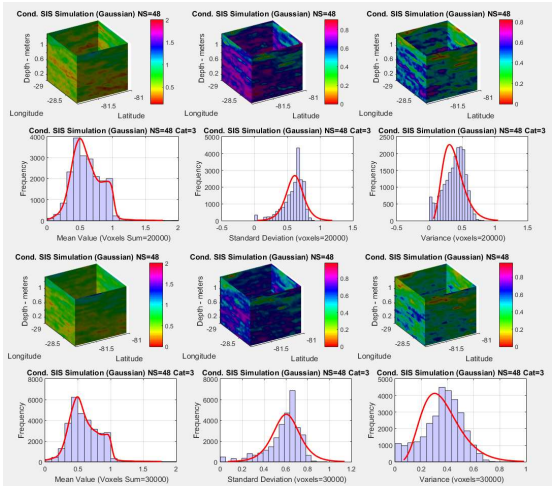
Figure(9) Mean Uncertainty vs. number of simulations and seven Grid Structure (or Voxels) – SIS Method – Gaussian model



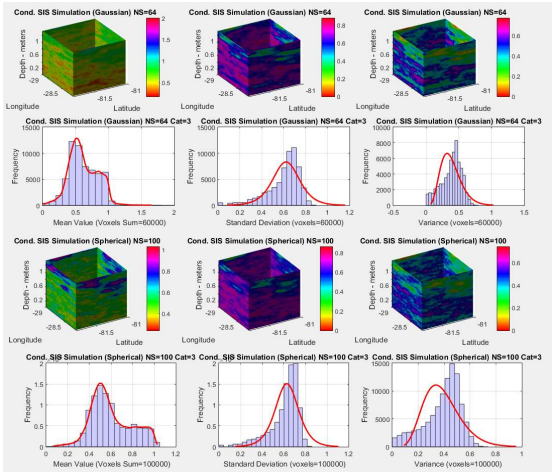
Figure(10) Variance Uncertainty vs. number of simulations and seven Grid Structure (or Voxels) SIS Method – Gaussian model



Figure(11) only two SIS tests presented for NS=20



Figure(12) only two SIS tests presented for NS=48



Figure(13) two SIS tests presented for NS=48 and 100

Uncertainty Assessment with SGS

Similar tests to SIS were performed using *Conditional Sequential Gaussian Simulation* (SGS) method with the two models (the Gaussian and the Spherical), changing grid structure in same way (from 10000 total voxels till 100000 total voxels. The figures below show the results of simulation tests, where for each structure changing number of simulations (NS=12, 24, 36, 48, 64 and 100 simulation). Variogram Ranges: max=0.25, medium=0.20, min=0.15 were fixes for all. Again, after all simulations for each round are ready, the output of the *Mean value, Standard Deviation* and *Variance* could be computed and presented as shown in the figures (16,17,18) . Then uncertainty for each of the three statistical measures was computed and registered. Those tests show that *Mean Uncertainty* decreases in the same way by increasing number of voxels or by increasing total simulations. For SGS method, the mean value of uncertainty decreases slowly by increasing number of simulations NS, or by increasing number of voxels as we see in the figure (14). Standard deviation uncertainty or *Variance Uncertainty* (Variance=σ^2; σ *Standard deviation*) measures do not show any stability after 100 simulations and their values continue decreasing beyond that, [figure (15)].

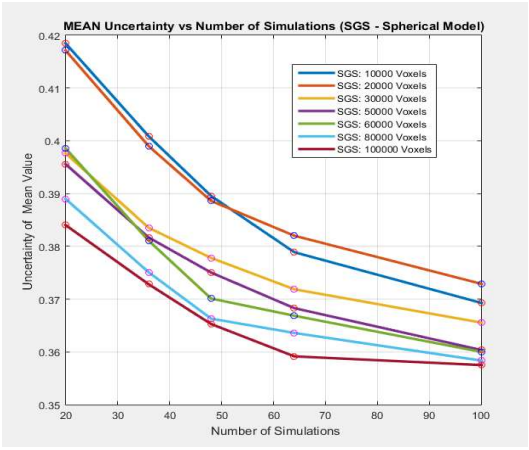


Figure (14) Mean Uncertainty vs. number of simulations and seven Grid Structure (Voxels) SGS Method – Spherical model

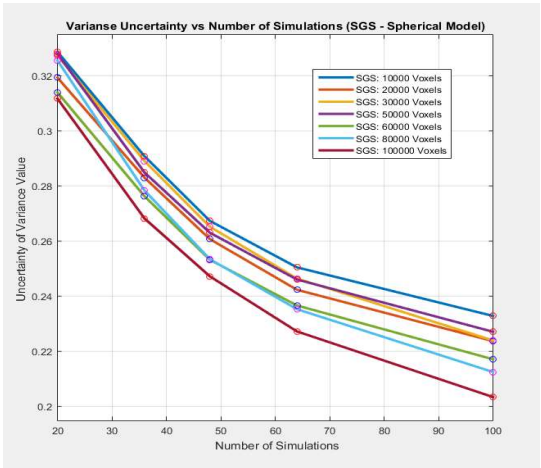
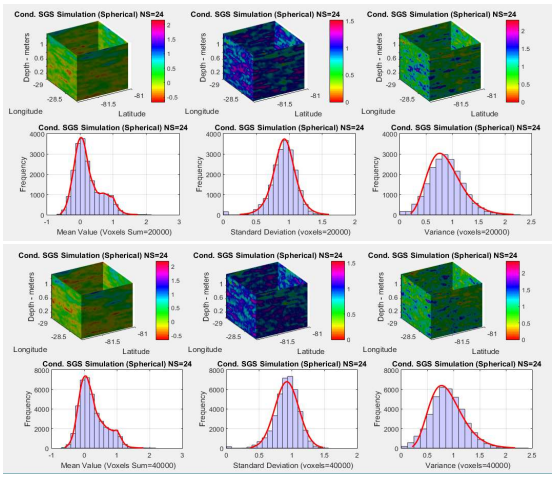


Figure (15) Variance Uncertainty vs. number of simulations and seven Grid Structure (Voxels) SGS Method – Spherical model



Figure(16) only two SGS tests presented for NS=24

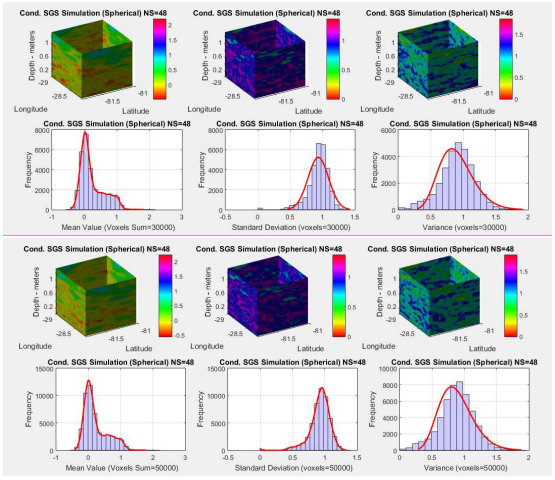


Figure (17) only two SGS tests presented for NS=48

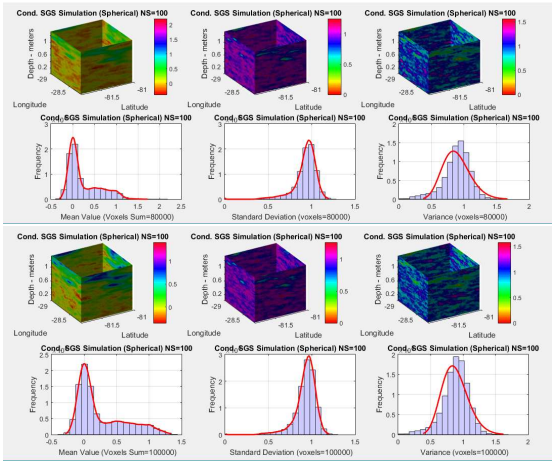


Figure (18) only two SGS tests presented for NS=100

Computational Cost Assessment

Simulation tests have been performed for both SIS Method and SGS Method and using the Spherical and Gaussian Variogram models changing grid structure (17 in total starting from 10,000 voxels till 1000,000 total voxels). Again for each structure changing number of simulations (NS) (for SIS six in total, NS= 6, 20, 36, 48, 64 and 100 , and for SGS six also, NS=12, 24, 36, 48, 64 and 100). This means that for each SIS or SGS Method about 100 test have been executed and for each test the time of execution was measured precisely (with accuracy ±1.0 millisecond). All tests were performed in the way with the same parameters explained in the previous sections. The time that has been measured belongs only to the CPU time for simulations and writing results to disk. There is an extra time is needed for presenting outputs or other arrangements was not included because this is not depend on number of voxels, the NS number or the method used. The characteristics of the CPU processor that has been used is Intel i7 2.20 GB runs by Windows 10 (64bit) operating system. Table (1) and table (2) show the results of the execution time (in seconds) for both SIS and SGS Methods respectively. Those results demonstrated graphically in figure (19) for SIS and in figure (20) for SGS simulations.

Number of Voxels	6 Simulations	20 Simulations	36 Simulations	48 Simulations	64 Simulations	100 Simulations
10000	0.957	3.164	4.246	6.778	8.526	10.940
20000	1.696	6.127	7.986	13.464	17.174	21.426
30000	2.385	9.074	11.826	19.790	26.250	32.050
40000	3.169	12.475	15.865	27.173	35.437	43.990
50000	3.710	15.747	19.468	33.791	44.410	53.912
60000	4.366	19.022	23.627	41.115	54.783	63.871
80000	5.682	25.166	31.281	54.794	72.781	85.304
100000	7.533	31.726	39.262	69.284	91.552	108.385
125000	9.292	39.484	50.377	86.596	113.949	136.375
150000	10.726	47.528	60.839	105.191	136.602	160.893
180000	13.242	57.857	71.710	125.786	164.080	192.063
216000	16.183	69.792	88.416	151.536	199.836	227.631
252000	18.808	80.863	102.950	178.318	234.080	269.031
343000	24.951	111.598	140.986	242.516	328.084	369.040
512000	36.540	167.781	210.476	368.346	483.043	535.505
729000	54.890	241.600	323.816	556.850	711.978	788.527
1000000	73.225	346.064	495.254	805.784	1008.368	1115.895

Table (1) SIS Method execution time in seconds

Number of Voxels	12 Simulations	24 Simulations	36 Simulations	48 Simulations	64 Simulations	100 Simulations
10000	2.014	2.937	5.475	7.540	8.624	10.515
20000	4.055	5.573	10.397	14.667	17.704	22.417
30000	6.228	8.417	15.641	21.706	26.579	33.904
40000	8.527	11.158	21.329	28.786	36.097	46.374
50000	10.798	13.685	26.910	36.047	46.885	57.499
60000	13.115	16.529	32.274	43.063	55.947	69.324
80000	17.401	22.414	43.196	56.995	74.960	92.302
100000	21.626	27.855	54.146	71.489	93.181	115.511
125000	27.100	34.441	68.238	90.002	117.541	144.314
150000	32.818	41.662	82.520	107.098	141.000	175.398
180000	39.692	50.500	99.736	128.775	173.270	216.380
216000	48.043	61.433	119.876	153.065	205.204	260.834
252000	56.096	71.415	140.153	178.815	239.598	310.892
343000	77.444	96.976	190.977	246.474	332.064	424.607
512000	115.075	146.141	293.699	365.988	494.793	648.297
729000	166.619	209.843	425.011	528.138	693.255	934.007
1000000	230.320	292.715	591.122	731.846	958.337	1278.054

Table (2) SGS Method execution time in seconds

The x and y axis of the charts are Log-Log which show more details for smaller grid structures (voxels number). As we see from the figures that the relationship is linear in both cases (SIS and SGS). Those figures also useful for making predictions by interpolation or extrapolation (unless the PC has same or similar parameters).

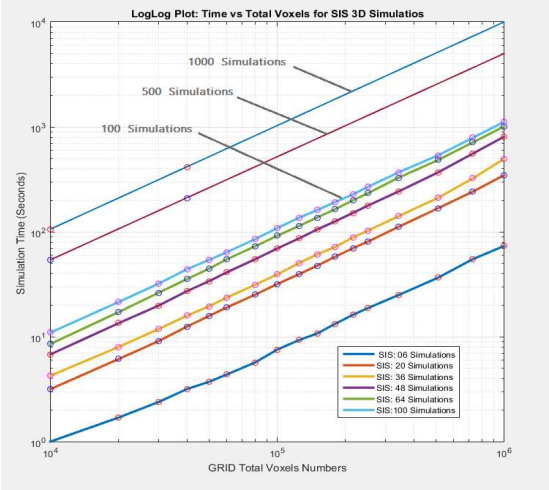


Figure (19) the linear relationship between simulation time (seconds) and Total Number of Voxels (**SIS Method**)

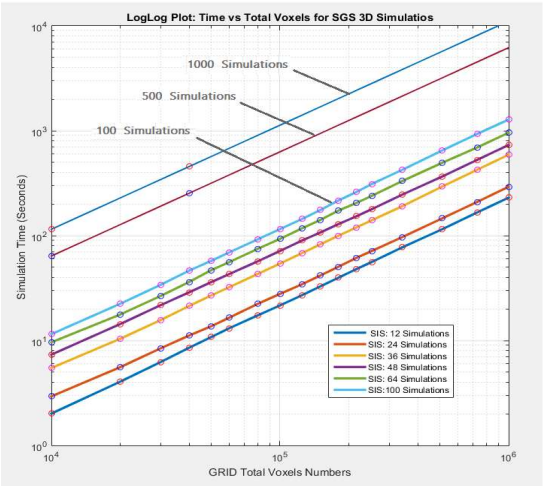


Figure (20) the linear relationship between simulation time (seconds) and Total Number of Voxels (**SGS Method**)

The last results do not show clearly which method is faster SIS or SGS because there are many disturbing values when comparing the two tables above. Also we still do not know whether the relationship is also linear or in other words, whether the total number of voxels processed in one second using SIS Method or using SGS Method will be the same, if all other parameters are the same !. One more question; do the Spherical Variogram model behave in the same way as with the Gaussian in term of computational cost?

For this purpose, another synthetic measure was created which calculates the speed of simulation by the following formula

$$Speed = \frac{NS \cdot TV}{T}$$

Where *Speed* refers to the total simulated voxel that is generated in one second,
NS refers to number of simulations, TV refers to Total Voxels and T refers to the time in seconds.

The values in the following tables show ("*Speed" Computations*) the total number of voxels processed in a second using SIS Method fixed in table (3) and using SGS Method in table (4). The tables values have been illustrated in the figures (21) and (22) respectively.

SIS	6 / Sph	20 / Gau	36 / Sph	48 / Gau	64 / Gau	100 /Sph
10000	62696	63211	84786	70817	75065	91408
20000	70755	65285	90158	71301	74531	93345
30000	75472	66123	91324	72764	73143	93604
40000	75734	64128	90766	70658	72241	90930
50000	80863	63504	92459	71025	72056	92744
60000	82455	63085	91421	70047	70095	93939
80000	84477	63578	92069	70081	70348	93782
100000	79650	63040	91692	69280	69906	92264
125000	80715	63317	89326	69287	70207	91659
150000	83908	63121	88759	68447	70277	93230
180000	81559	62222	90364	68688	70210	93719
216000	80084	61898	87948	68419	69177	94890
252000	80391	62328	88120	67834	68900	93670
343000	82482	61471	87583	67888	66910	92944
512000	84072	61032	87573	66720	67837	95611
729000	79687	60348	81046	62839	65530	92451
1000000	81939	57793	72690	59569	63469	89614
Average	**79232**	**62675**	**88123**	**68569**	**69994**	**92930**

Table (3) Total Number of Voxels processed in a second using SIS Method (*Sph: Spherical model, Gau: Gaussian*)

SGS	12 / Gau	24 / Sph	36 / Gau	48 / Sph	64 / Gau	100 /Sph
10000	59583	81716	65753	75660	74212	95102
20000	59186	86130	69251	77453	72300	89218
30000	57803	85541	69049	78341	72237	88485
40000	56292	86037	67514	78699	70920	86255
50000	55566	87687	66890	78580	68252	86958
60000	54899	87120	66927	78879	68636	86550
80000	55169	85661	66673	79374	68303	86672
100000	55489	86160	66487	79143	68684	86572
125000	55351	87105	65946	78665	68061	86617
150000	54848	86410	65439	79228	68085	85520
180000	54419	85545	64972	79094	66486	83187
216000	53952	84385	64867	79736	67367	82811
252000	53908	84688	64729	79645	67313	81057
343000	53148	84887	64657	78798	66108	80781
512000	53391	84083	62758	79150	66226	78976
729000	52503	83377	61749	78255	67300	78051
1000000	52101	81991	60901	77588	66782	78244
Average	**55153**	**85207**	**65562**	**78605**	**68663**	**84768**

Table (4) Total Number of Voxels processed in a second using SGS Method (*Sph: Spherical model, Gau: Gaussian*)

The answers to all above suggested questions can be deduced from the figures (21) and (22).

- In general, SIS Method is much faster than SGS Method 10-15%, as we see with the Spherical variogram and NS=100, SIS speed is about 93000 voxels/sec. while it is near 85000 voxels/sec. for SGS. When using Gaussian variogram and NS=36, SIS speed is about 88000 voxels/sec. while SGS speed is about 66000 voxels/sec.

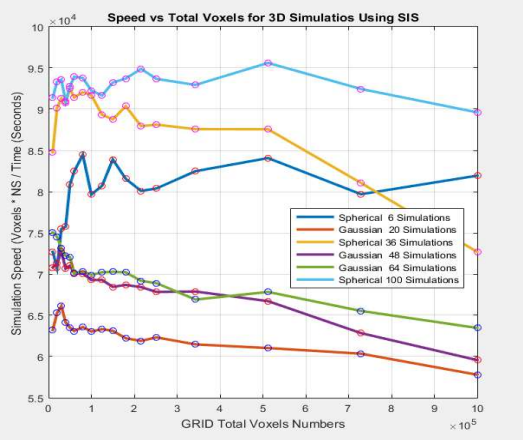


Figure (21) the relationship between simulation "Speed" (Total Voxel/second) and number of voxels for one simulation Using **SIS Method**.

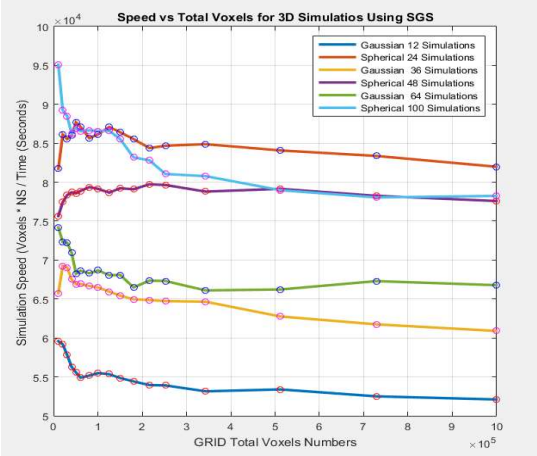


Figure (22) the relationship between simulation "Speed" (Total Voxels/second) and number of voxels for one Simulation using **SGS Method**.

- For both methods (SIS or SGS) the speed is not stable all the time, as we see some distributions in the beginning for grid total voxels number less than 350,000. After that limit SGS speed becomes stable. On the other hand SIS speed has also some stability, but its performance becomes a little slower for larger grid total voxels.

- One can also notice, in general, from tables (3) and (4) or from corresponding figure that when using the Spherical Variogram the speed is nearly 20%-25% faster than its level using the Gaussian variogram, no matter whether the simulation is performed by SIS or SGS method.

Conclusions

Statistical tests have been proved that the *Mean Uncertainty* decreases (with both *SIS* or *SGS Conditional Simulations*) by either increasing number of total voxels (3D grid) or by increasing number of simulations (NS) or both. For SIS method this uncertainty becomes stable after the limit NS=64, while for SGS method the same feature becomes stable after NS=100 limit.

The tests also proved that by either increasing number of total voxels (3D grid) or by increasing number of simulations (NS) *Variance Uncertainty* continue decreasing beyond the limit NS=100, but for SIS method this feature is a little slower and for SGS method the decreasing is much faster.

For both *SGS* and *SIS Conditional Simulation* methods, there is a clear linear relationship between *Computational Cost* (simulation time) and number of Voxels of the 3D grid no matter which CPU processor is used. This conclusion helps to predict precisely the computational cost for large 3d grid structure and/or very large number of simulations (say NS>100). Note that each of SIS or SGS has its own chart and its own speed, thus we should not unify the two charts.

The multiple tests (more than 200) proved that SIS method speed is 10-15% faster than SGS Method. The tests also proved that speed of simulations is faster 20%-25% using Spherical Variogram than when using the Gaussian one.

- *Special Matlab programs have been used in all implementation, and presentations of this research, with support from mGstat and SGeMS libraries for performing simulations only. This software is free online* [see *mGstat:* Hansen T.M (2011)] and [SGeMS; Rémy N., Wu J., Boucher A. (2004)].

References

Al-Abdalla Mohammed (1998) Geostatistical Analysis and Quantification Uncertainty for 3D Modeling by Simulation, MSc thesis (ITC, the Netherlands).

Atkinson P, Quattrochi DA, Goodman HM (2000) Introduction to geostatistics and geospatial techniques in remote sensing. Computers and Geoscience (ISSN 0098-3004) vol.26; 359. Elsevier Science Ltd.

Banerjee S. (2004) On Geodetic Distance Com-putations in Spatial Modeling. Biometrics, Vol.61(2), 617-625.

Bolstad W.M. (2007) Introduction to Bayesian Statistics. 2nd Edition. Wiley J. & Sons.

Brus DJ, Heuvelink GBM (2007) Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138 (2007) 86-95.

Chiles, J.; Delfiner, P. (1999) Geostatistics: Modeling Spatial Uncertainty; Wiley J. & Sons, New York.

Christakos, G. (2005) Simulation of Natural Processes. Random Field Models in Earth Sciences; Dover: New York; pp. 295–336.

Cressie, N.A.C. (1993). Statistics for Spatial Data. Wiley.

Deutsch C.V. & Journel AG. (1992) GSLIB, Geostatistical Software Library and User's Guide.

Deutsch, C.V. 2002, Geostatistical Reservoir Modeling, Oxford University Press.

Deutsch C.V. 2006 A Sequential Indicator Simulation Program for Categorical Variables with Point and Block Data. BlockSIS, Computers & Geoscience 2006 Elsevier. 402,1-22.

Journel, A.G. & Huijbregts, C. (1978) Mining Geostatistics. Academic press.

Journel, A.G (1989) Fundamentals of Geostatistics in Five Lessons. Wiley Online.

Goovaerts, Pierre (1997) Geostatistics for Natural Resources Evaluation; Oxford University Press.

Hansen T.M 2011, mGstat a geostatistical Matlab toolbox.

Hengl T. (2007) A Practical Guide to Geostatistical Mapping of Environmental Variables. EUR 22904 EN.

Hernandez JJ.G & Srivastava RM (1990) An Ansi-C Three-dimensional Multiple Indicator Conditional Simulation Program. Computers & Geoscience Vol.16, issue 4, pages 395-440.

Lantuéjoul, C., 2002. Geostatistical Simulation, Models and Algorithms. Springer, Berlin.

Møller, J. (Ed.) (2003) An introduction to model-based geostatistics. Springer New York, pages 43-86.

Mund Jan-Peter (2013) Geospatial statistics and spatial data interpolation methods. GIS'Em 2013 at Eberswalde.

O'Reilly A.M., Roehl, Jr. E., Conrads P.A., Daamen R.C., and Petkewich M.D. (2014) Simulation of the Effects of Rainfall and Groundwater Use on Historical Lake Water Levels, Groundwater Levels, and Spring Flows in Central Florida. Scientific Investigations Report 2014–5032 (U.S. Geological Survey)

Rémy N. Wu J. Boucher A. (2004) The Stanford Geostatistical Modeling Software (SGeMS) A User's Manual.

Schabenberger P.O., Gotway C.A. (2004) Statistical Methods for Spatial Data Analysis. Simulation of Random Fields. Chapman & Hall/CRC texts in Statistical Science.

Sepúlveda N, Tiedeman C.R., O'Reilly A.M., Davis J.B. and Burger P. (2012) Groundwater Flow and water Budget in the Surficial and Floridan Aquifer Systems in East-Central Florida. Scientific Investigations Report 2012–5161 (U.S. Geological Survey).