The Structure of Nodes Visited in Sequential Simulation

Yupeng Li and Clayton V. Deutsch

Sequential simulation is commonly employed in the context of Gaussian, indicator and multiple point statistics methods. Detailed analysis of sequential simulation is important. This short note addresses the configurations of previously simulated nodes encountered as sequential simulation proceeds. This is viewed as a homogenous point process.

Introduction

Sequential simulation is a powerful and widely used stochastic simulation technique to sample multivariate probability spaces. For a given multivariate model, sequential simulation can be used to determine the conditional probability of a single random variable given any number of conditioning values. From a practical point of view, it is the basis for sequential indicator simulation (SISIM) (Deutsch and Journel, 1998) or single normal sequential simulation (SNESIM) (Strebelle, 2002) for discrete variable and sequential Gaussian simulation (SGSIM) (Deutsch and Journel, 1998) for continuous variables.

The sequential simulation methodology is very straightforward. Most theory researches are focused on how to model the multivariate probability space and determine the conditional probabilities. Some practical details such as the search radius and number of conditioning data have been investigated (McLennan, 2002; Zanon and Leuangthong, 2004). The data pattern changes continuously along the random path in the sequential simulation process. In this paper, the random path is investigated to find the most important and representative data configurations.

Random Path

Sequential simulation consists of the following four steps: (1) choose the research domain and discrete it into grids, (2) define a random path to visit every grid in the research domain, (3) at each grid node: a) search to find nearby data and previous simulated nodes; b) calculate the conditional distribution for the current grid node; c) perform Monte Carlo simulation to obtain a single value from the distribution, and (4) repeat step 3 until every grid node has been visited.

A set of equally probability realizations are simulated with different random paths and different random drawings from the conditional distributions. A random path is used to avoid artifacts. Since the simulation path is random, the number of previous simulated nodes within a certain area will be statistically independent. The random path picking process in sequential simulation can be viewed as a homogenous point process (complete spatial randomness) (Bar-Hen and Picard, 2006).

Homogeneous Point Process

The Poisson process is mainly used where data are correlated over time. A homogeneous Poisson process is characterized by a rate parameter λ, also known as intensity, such that the number of events in time interval $(t, t+\tau]$ follows a Poisson distribution with associated parameter $\lambda\tau$:

$$P[(N(t+\tau)-N(t))=k]=\frac{e^{-\lambda\tau}(-\lambda\tau)}{k!},\ k=0,1,2,.... \tag{1}$$

However, for data in two or more spatial dimensions, no such ordering is generally present. This is the primary difference between a Poisson process in time series and spatial data. In spatial statistics, a homogeneous Poisson process is a point process such that the random variable $N(\mathbf{B})$ counting the number of events within $\mathbf{B}$ follows a Poisson law with parameter: $\lambda(v(\mathbf{B}))$, where λ is the intensity of the process.

$$\mathbf{p}(N(\mathbf{B})=n)=\frac{(\lambda v(\mathbf{B}))^n}{n!}e^{-\lambda\cdot v(\mathbf{B})}\ with\ n=0,1,2,.... \tag{2}$$

This shows that the number of points in any bounded set ***B*** follows a Poisson distribution with mean $\lambda v(\mathbf{B})$, which can be written as:

$$E(N(\mathbf{B}))=\lambda v(\mathbf{B}) \tag{3}$$

In (2)and(3), the parameter λ is the called the intensity or point density of the homogeneous Poisson process. It describes the mean number of points to be found in a unit spatial volume. A specific type of the distribution in (2) is the void probabilities describing the probability that there are no points in a specific subset ***B***, that is $\mathbf{p}(N(\mathbf{B})=0)$. For example, if $\mathbf{B}=b(x,r)$ is the sphere or disc of a radius r centred at x, and then $\mathbf{p}(N(\mathbf{B})=0)$ is the probability that this disc does not contain any points. Based on this probability, we can calculate the probability:

$$H(r)=1-\mathbf{p}(N(\mathbf{B})=0) \tag{4}$$

The so called 'typical' point of a point process is frequently considered. This is a point that has been chosen by a selection procedure in which every point of the process has the same chance of being selected. In sequential simulation, usually many realizations are simulated for uncertainty evaluation. Although, in one realization, the node picking along the random path is determined, when looking at all those realizations, along the random path, every point has the same chance of being selected. Thus, this 'typical' point equivalence to the next picked unsampled location in sequential simulation.

The probability in (4) is a distribution function of r and can be interpreted as the probability that the distance between the nearest point in the Poisson process to the 'typical' point. The distribution is called spherical contact distribution function or location-to-nearest-point distribution function (Illian, 2007, p.68, p.75). If the point process is a homogeneous Poisson process, the distribution will be:

$$H(r)=1-e^{-\lambda\bullet\nu(\mathbf{B})} \tag{5}$$

In Equation(5), $\nu(\mathbf{B})$ is the volume of this subset volume of the whole spatial volume. For example, in 2d case, it will be πr^2, the spherical contact distance distribution will be:

$$H(r)=1-e^{-\lambda\pi r^2} \tag{6}$$

Generally, in 2D case, the distribution function $H_k(r)$, which could be the distances to the 1st, 2nd, 3rd,... nearest neighbours, is

$$H_k(r)=1-\sum_{j=0}^{k-1}\exp(-\lambda\pi r^2)\frac{(\lambda\pi r^2)^j}{j!} \tag{7}$$

and the corresponding probability density functions are

$$h_k(r)=\frac{2(\lambda\pi r^2)^k}{r(k-1)!}\exp(-\lambda\pi r^2) \tag{8}$$

For 3D case, the subset of the volume $\nu(\mathbf{B})$ will be $4\pi r^3/3$.

As stated above, given a certain previous simulated nodes, the distribution of them in the spatial space can be viewed as a homogeneous point process. Thus, in sequential simulation, the intensity of the point process is defined as the number of previous simulated nodes per unit area and can be approximated by the data density calculated from the previous simulated nodes and hard data. For example, assuming there is no hard data, an unconditional simulation will be applied in the domain composed by *nxyz* grids, the data density will be $1/nxyz$ at the beginning. As the simulation goes on, more and more previous simulated data will be added into conditioning data. Thus, the density will reach to 1 when all the nodes are simulated. Although each step is a homogeneous point process, in the whole process, the intensity for each homogeneous point process increases consistently. Figure 1 is an example of point distribution at four different simulation stages.

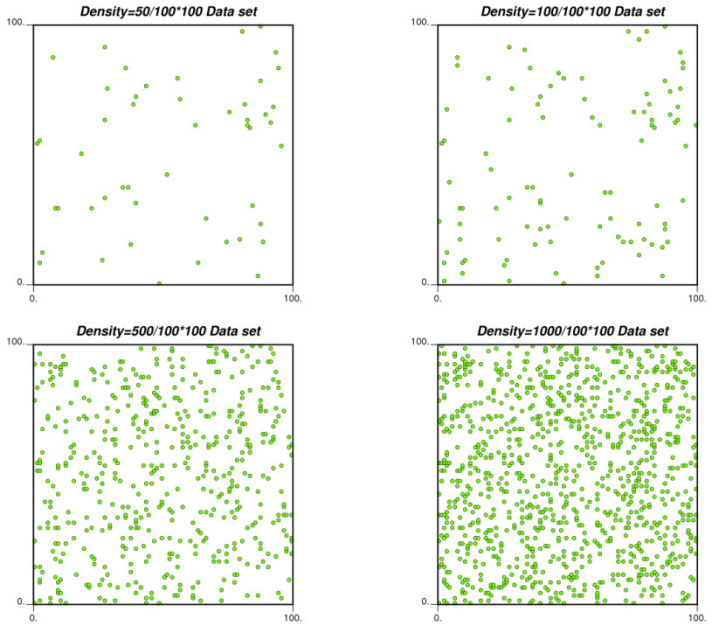

Figure 1 four homogeneous point process with different data density in sequential simulation

Random Path as a Homogeneous Point Process

Firstly, what we calculate the mean number in certain search radius for each simulation step that is looked at as a homogeneous point process. Based on Equation(3), using different search radius ($R = n \bullet gridsize$), the mean number (in 2D and 3D) can be calculated as listed in Table 1 and Table 2. The number increase quickly as the search radius increases.

Table 1: theoretical mean data number in a certain research area given different data density (2D case)

		data density										
		0.05	**0.06**	**0.07**	**0.08**	**0.09**	**0.1**	**0.15**	**0.2**	**0.3**	**0.4**	**0.5**
search radius	**5*gridsize**	3	4	5	6	7	7	11	15	23	31	39
	6*gridsize	5	6	7	9	10	11	16	22	33	45	56
	7*gridsize	7	9	10	12	13	15	23	30	46	61	76
	8*gridsize	10	12	14	16	18	20	30	40	60	80	100
	9*gridsize	12	15	17	20	22	25	38	50	76	101	127
	10*gridsize	15	18	21	25	28	31	47	62	94	125	157
	11*gridsize	19	22	26	30	34	38	57	76	114	152	190
	12*gridsize	22	27	31	36	40	45	67	90	135	180	226
	13*gridsize	26	31	37	42	47	53	79	106	159	212	265
	14*gridsize	30	36	43	49	55	61	92	123	184	246	307
	15*gridsize	35	42	49	56	63	70	106	141	212	282	314

Table 2: theoretical mean data number in a certain research area given different data density (3D case)

		data density													
		0.01	0.015	0.016	0.017	0.018	0.019	0.02	0.04	0.06	0.07	0.08	0.09	0.1	0.2
search radius	**4*gridsize**	2	3	3	3	3	3	4	8	12	14	16	18	20	40
	5*gridsize	3	5	6	6	7	7	7	15	23	27	31	35	39	78
	6*gridsize	6	10	10	11	12	12	13	27	40	47	54	61	67	136
	7*gridsize	10	16	17	18	19	20	21	43	64	75	86	96	107	216
	8*gridsize	16	24	25	27	28	30	32	64	96	112	128	144	160	323
	9*gridsize	22	34	36	38	41	43	45	91	137	160	183	206	229	460
	10*gridsize	31	47	50	53	56	59	62	125	188	219	251	282	314	631

The distribution of the minimum distance (to the first closest previous node) is of interest. For each data density, the mean distance to the first closest previous nodes can be calculated from Equation (7) or Equation(8). Figure 2 shows 8 different probability distributions with different data densities in a 2D case. As the simulation goes on, the mean distance of the first close data decreases to a very small distance.

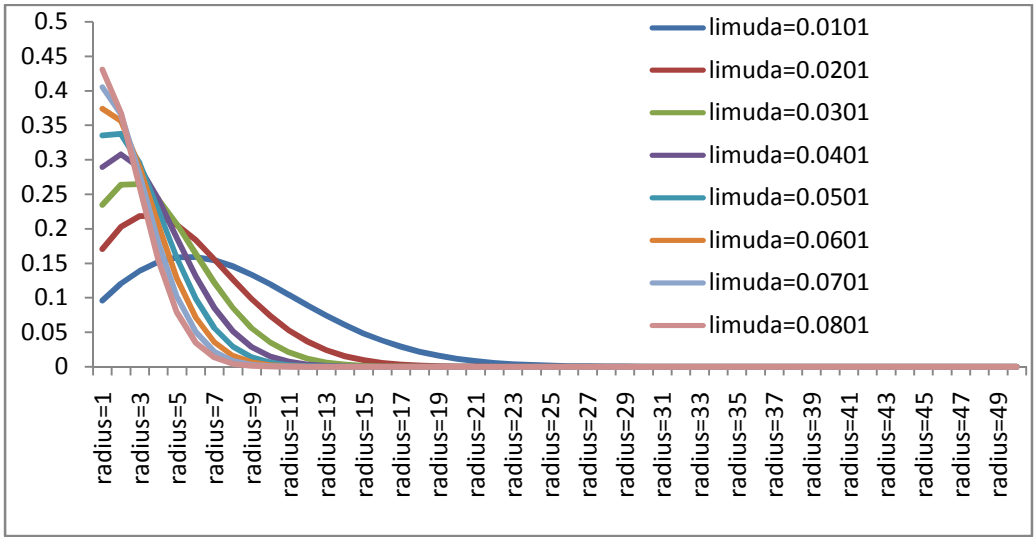

Figure 2 the probability distribution of the first closest distance in eight different data density in the sequential simulation process stage

From the distribution in Figure 2, two statistics for each homogeneous point process can be obtained: the mean distances and its probability. Because all the nodes in the simulated area are independently to each other during the picking process, the mean distances to the first close previous node will also be a function of the data density in the whole sequential simulation process which can be written as:

$$r = f(\lambda) \tag{9}$$

The mean distances and the maximum probabilities for all the mean distance are plotted in Figure 3.

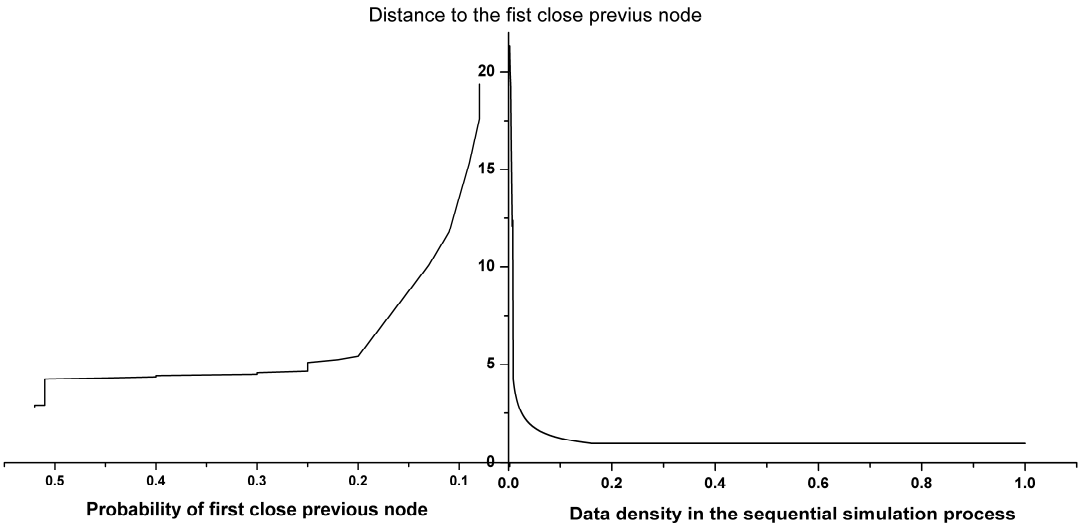

Figure 3 the mean distance of the minimum distance to the previous simulated nodes and its probability in the simulation process

As shown in Figure 3, during the whole simulation process, the very short distance to the first close previous node has a higher probability to exist than the larger mean distance. Clearly, those simulated nodes in the early stages will have a higher probability to be used in the later stages of simulation. In other words, the influence area of the first close previous nodes is decreasing as the simulation goes on. Based on this point, the importance of the first close node can be defined as the influence area or volume in 2D and 3D and denoted as:

$$S = r^d \Big/ r_{\max}^d \quad , d = 2 \; or \; 3 \tag{10}$$

In Equation(10), r will be the mean distance to the first close previous node, $r_{\max}$ is the maximum distance in them. This importance (the influence area) will decrease with the data density increase as the simulation goes on as shown in Figure 4.

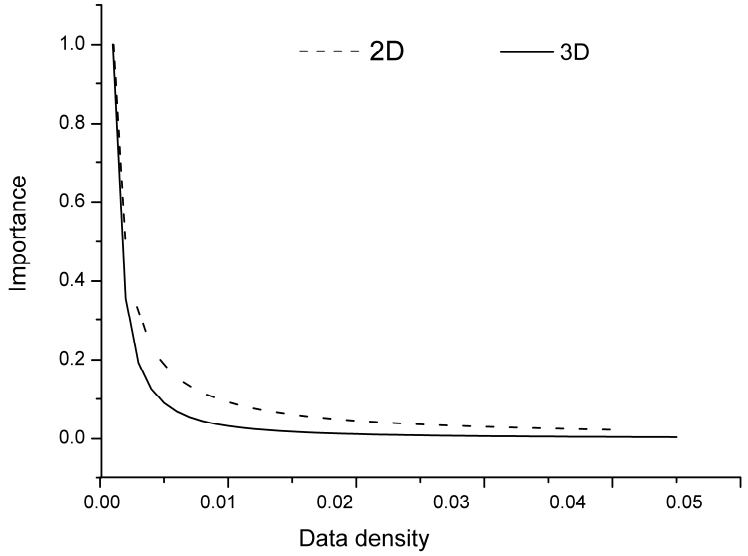

Figure 4 the importance of the first close previous node in 2D and 3D case

Conclusions

For the 2D case, the mean distance has a very low probability in the sequential simulation process when data density is smaller than 0.05, their importance is more critical than those in later stage when data density is larger than 0.05. While for 3D case, the critical data density is about 0.01 where the most importance nodes will be picked for later simulation using. After the data density is larger than 0.01, the importance of the nodes changes little along the simulation path. Thus, we should pay more attention to the simulation stage when data density is lower than 0.05 for 2D case and 0.01 in 3D case.

Given the importance of data density in 2D and 3D, the maximum mean probability distribution shows that those data configuration in search area less than $15 \bullet gridsize$ for 2D case and $10 \bullet gridsize$ for 3D case will have the probability larger than 0.15 and 0.2 respectively which is far more larger than those distances that larger than $15 \bullet gridsize$ for 2D case and $10 \bullet gridsize$ for 3D case. So in 2D case, the representative and most importance data configuration data will be those that data density is about 0.05, the mean distance of the first close previous node is about $15 \bullet gridsize$. In 3D case, the representative and important data configuration will be those that data density is about 0.01, the mean distance of the first close previous node is about $10 \bullet gridsize$.

Finally, given the most important and representative first close previous node, the mean number in the influenced area will be 35 and 30 in 2D and 3D case respectively. This research assumes the point process is in an infinite spatial space which is not exactly correct; nevertheless, the results of this research can be used as a basic guide when designing a detailed algorithm comparison using sequential simulation.

References

BAR-HEN, A. and PICARD, N., 2006. Simulation study of dissimilarity between point process. *Computational Statistics*, **21,** 487-507.

DEUTSCH, C.V. and JOURNEL, A.G., 1998. GSLIB: Geostatistical software Library and User's Guide. New York, Oxford University Press.

MCLENNAN, J., 2002. The Effect of the simulation path in sequential Gaussian simulation In: annual conference of Centre for Computational Geostatstics.

STREBELLE, S., 2002. Conditional Simulation of Complex Geological Structures Using Multiple-Point Statistics. *Mathematical Geology*, **34,** 1-21.

ZANON, S. and LEUANGTHONG, O., 2004. Implementation Aspects of Sequential Simulation. *Book Series: Quantitative Geology and Geostatistics*, **14,** 543-548.