Entropy of Gaussian Random Functions and Consequences in Geostatistics

Paula Larrondo (larrondo@ualberta.ca)
Department of Civil & Environmental Engineering
University of Alberta

Abstract

Sequential Gaussian simulation algorithm is based on two-points statistics to characterize the spatial distribution. Since Gaussian distribution maximizes the entropy beyond the covariance, the simulated realizations will tend to have less spatial structure than reality. Entropy has been used as a measure of spatial disorder. This paper shows how entropy measurements change as we increase the number of points used to estimate this statistic and after performing simulation. Further, the use of entropy as a measure of connectivity between values from a certain range of a continuous variable is also shown.

Introduction

Sequential Gaussian Simulation (SGS) is one of the most common simulation algorithms used for modeling continuous variables. The great advantage of this technique is that it reproduces the correct spatial variability. Also an assessment of local and global uncertainty is possible due to multiple realizations obtained from simulation are possible do. The statistical advantage of the Gaussian distribution is that in multivariate Gaussian space, all conditional distributions are also Gaussian. Further, the conditional expectation of a Gaussian variable is a linear combination of conditioning data, so the correct estimate corresponds to the one obtained from Kriging.

On the other hand, the great disadvantage of choosing a Gaussian distribution is the characteristic of maximum entropy (Journel and Deutsch, 1993), that is, beyond any second order statistics imposed by the model the spatial disorder will be maximized. Indeed, the unbounded probability density function with finite variance that maximizes the entropy is the Gaussian distribution. SGS realizations will have less spatial structure than reality; extreme values are highly disconnected, while median values have greatest connectivity. Most geological processes result in spatial structures that cannot be fully characterized by two-point statistics, such as the covariance or the variogram. Entropy can be see as a summary statistic of a probability distribution, but like any summary statistic, such as the variance for example, it does not give information on the shape of the distribution.

This “feature” of SGS may have great consequences when looking at specific properties, both for petroleum and mining applications. For example, in an SGS realization of permeability, high values and low values will not have simulated connected paths which could lead to erroneous estimation of fluid flow rates. The design of short-term mine selection plans could be incorrect because of a bias in the variability.

Other simulation algorithms like Sequential Indicator Simulation (Journel, 1983; Journel and Isaaks, 1984), Truncated Gaussian Simulation (Galli *et. al.*, 1994) or Simulated Annealing (Deutsch, 1992) may have different characteristics of the connectivity of extreme values is an issue.

The visual difference between low and high entropy can easily be appreciated; however, it is important to show how it changes from a reference image to a simulated realization.

This paper shows (1) how entropy varies with conditioning data that have different spatial structures, and how certain spatial structures can influence the entropy measurements as we consider more points in the template to obtain the entropy measurement and (2) how entropy can be use as measurement of connectivity of extreme values.

Methodology

Entropy provides a measured of uncertainty associated with the probability density function (pdf) of a random variable. For a univariate continuous distribution with pdf *f(z)*, of a random variable *Z*, the entropy is defined as:

$$H = -\int_{-\infty}^{+\infty} \left[\ln f(z)\right] f(z)dz \tag{1}$$

If *Z* is a discrete random variable, that can take *K* outcomes values with probability p_k, k=1, …, *K*, where $\sum_{k=1}^{K} p_k = 1$, then the entropy can be calculated as,

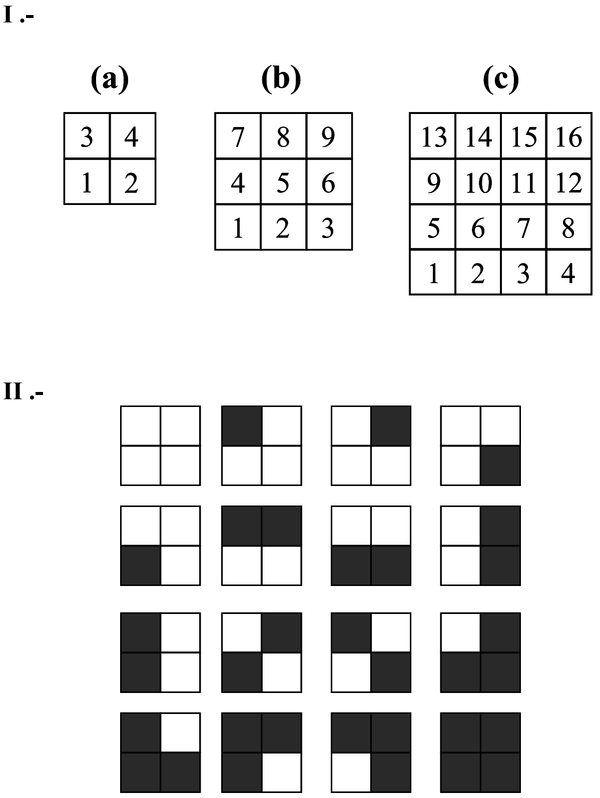
$$H = -\sum_{k=1}^{K} \left[\ln p_k\right] p_k \tag{2}$$

In the case of *n* random variables Z_j, *j* =1, ..., *n*, Equations 1 and 2 , can be rewritten to account for the *n*-variate probability density distribution (Journel and Deutsch, 1993).

To determine the entropy of an image we need to calculate some set of probabilities p_k for Equation 2. Lets consider three different templates (Fig. 1), where each position corresponds to an indicator variable,

$$\mathrm{I}(z(u_\alpha); z_c) = \begin{cases} 1 & \text{if} \quad z(u_\alpha) \geq z_c \\ 0 & \text{if} \quad z(u_\alpha) < z_c \end{cases} \tag{3}$$

where z_c corresponds to a certain threshold in the *Z*-variable distribution, α=1, ..., N, and N=4,9 or 16 depending on the template.



I .-

(a)

3	4
1	2

(b)

7	8	9
4	5	6
1	2	3

(c)

13	14	15	16
9	10	11	12
5	6	7	8
1	2	3	4

II .-

Fig. 1: (I.-) Different templates consider for entropy measurement. Each position corresponds to an indicator variable, i.e., is one or zero. (II.-) The sixteen possible configurations for the template with four positions, white squares are ones, blacks are zeros.

Given that each location can take only two values, each template will have 2^N possible configurations. For example the template with 4 locations has 16 different configurations, the template with 9 locations generates 512 possible configurations, and the pattern of 16 points has 65536 possible configurations.

Then for each template, p_k, k=1, …, 2^N can be defined as the probability of a particular configuration to be found within a certain domain. This probability corresponds to the pdf of a 2^N-multivariate random function and the entropy measured from these probabilities corresponds to a 2^N-multipoint statistic.

Continuous variables can be discretized into a finite number of classes *C*, then the number of configurations for each template will be C^N. In this paper we will consider only two possible outcome values and its implementation is straightforward. As *C* or *N* increases the probabilities of the different configurations becomes smaller and likely at the same time more uniform so the entropy increases (Deutsch, 1992 and Tupin *et al.*, 2000). Maximum entropy occurs only when all categories are equally probable, while minimum entropy is achieved when only one configuration can occur.

A subroutine in FORTRAN was created to determine the entropy of an image, basically the program scans the image, with a step of 1 pixel, and at each node it determines the index of the template (Deutsch, 1992) centered at that location. As the image is scanned the indices of the different templates are saved as an array of size 2^N, from this array, the probabilities and entropy are calculated using Equation 2.

Application

Four *exhaustively* sampled reference images were chosen to show how entropy of the reference image changes with respect to its corresponding simulated realization. These four images (Fig. 2) represent different stratigraphic textures, two of rippling deposition, the first (*ripl01*) with relatively thinner layering than the second (*ripl02*), the third image correspond to an example of cross-bedding deposition (*cross-bed01*), finally the forth image represent the layering of turbiditic deposition (*turb01)*. All of them show a clear spatial structure, and connectivity at certain value ranges. These images correspond to scanned photographs; the grayscale was coded to numbers between 0 and 1. The histograms of the references images are shown in figure 3.

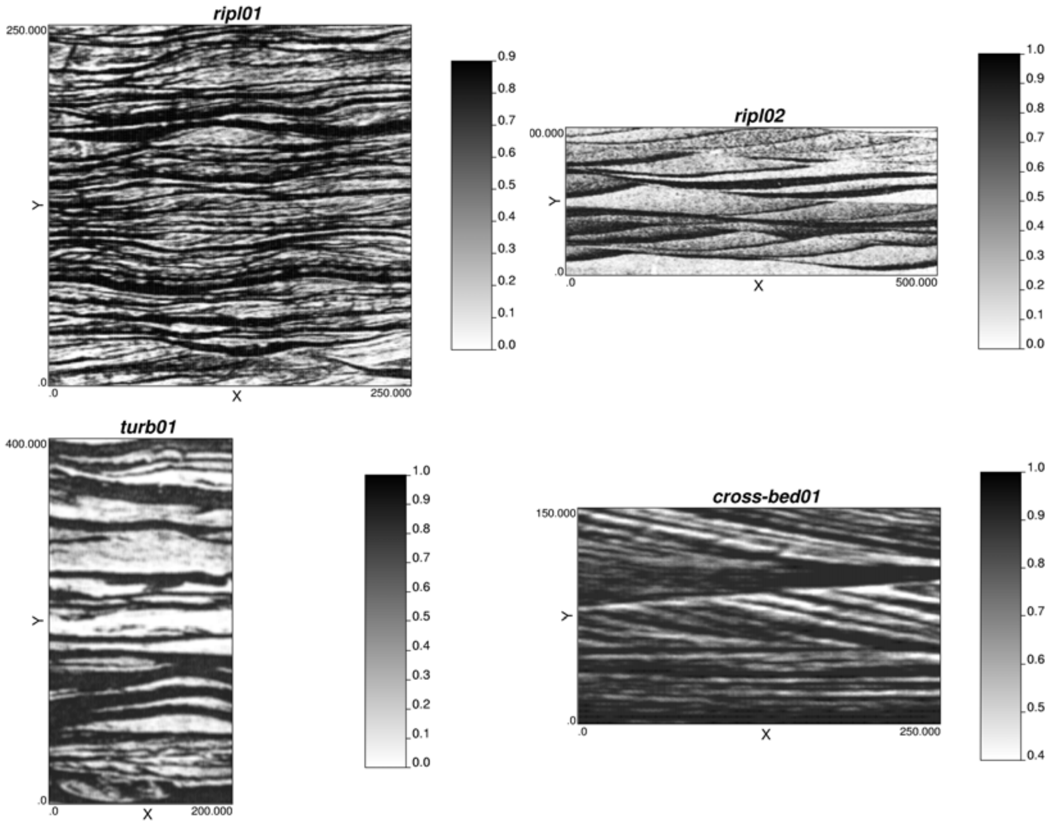


Fig. 2: Exhaustively sampled reference images. The top left and right images correspond to rippling deposition. The bottom left is an example of turbiditic layering and the bottom right of cross bedding layering.

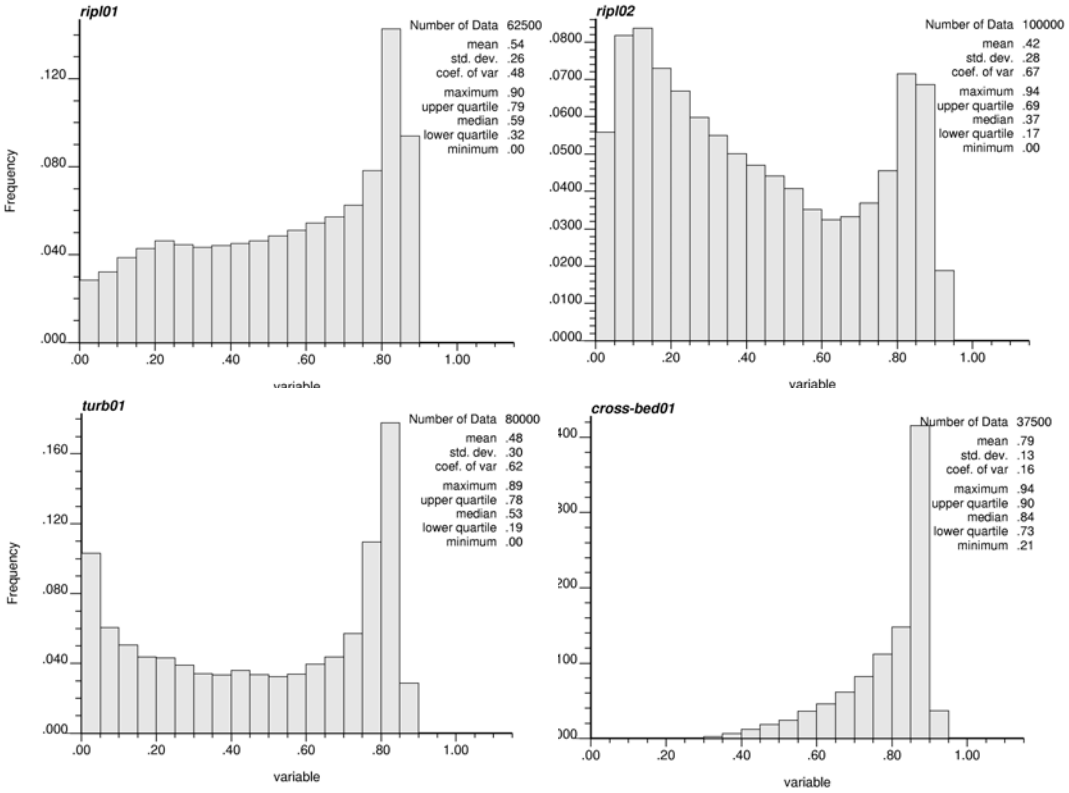


Fig. 3: Histograms of the references images.

The entropy of the four references images was calculated using the three different templates (Table 1 and Fig. 4). As previously noted, the entropy increases as the number of configuration increases. In this case the number of possible configurations increases as more points are consider in the template. The rate in which entropy increases with an increased number of configurations is a function of the complexity of the image, with consideration for the thickness of the layers and the linearity of the contacts between these layers.

The entropy of the four references images was also calculated for template (b) using three different thresholds z_c in the indicator function for each location; the mean and the upper and lower quartiles (Table 2). The trend that follows the entropy for each image is significantly different and corresponds to the different features of their spatial structure. For example, the image *ripl01* shows higher entropy when the threshold is the mean, reflecting the connectivity of the values around the. The entropy for the image *cross-bed01* with the mean as the threshold is still a little lower than the entropy using the upper-quartile threshold. This result is consistent with the high connectivity of the high values and the skewness of its distribution. Another interesting feature is the results of image *turb01*, the entropy varies more smoothly between the different possible thresholds, as one would expect given that the stratigraphic banding is thicker.

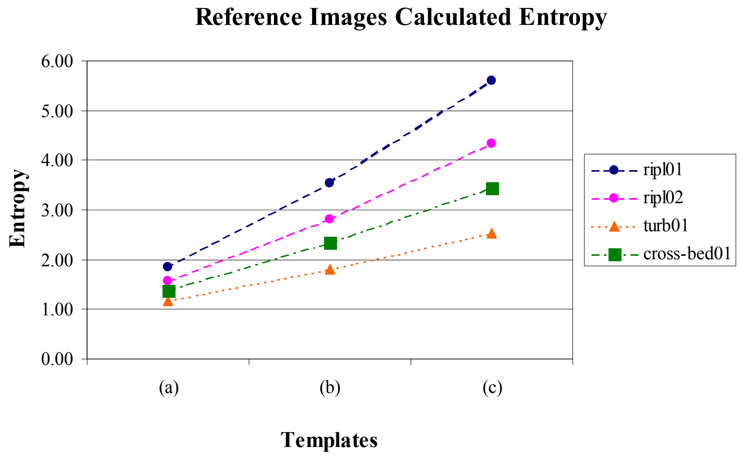


Fig. 4: Calculated Entropy for the four reference images, using different templates.

Image	Entropy Values for different Templates		
	(a) N=4	(b) N=9	(c) N=16
ripl01	1.86	2.27	5.61
ripl02	1.57	2.62	4.33
turb01	1.18	1.51	2.54
cross-bed01	1.38	1.18	3.44

Table 1: Entropy values for the references images, using different templates.

Image	Entropy Values for different z_c		
	Lower quartile	Mean	Upper quartile
ripl01	2.27	3.53	2.90
ripl02	2.62	2.81	1.88
turb01	1.51	1.80	2.13
cross-bed01	1.18	2.35	2.60

Table 2: Entropy values for the references images, using different thresholds: the mean and the lower and upper quartiles.

The variograms in two directions (X and Y of the image) were calculated for each image, and they were fitted using VARFIT program (Larrondo *et. al.*, 1993).

For each one of the references images, 1% of the number of points contained in the images was randomly sampled to perform the simulations. For each set of samples, 100 realizations were simulated using SGSIM (Deutsch and Journel, 1998); a realization for each scenario is shown in figure 6.

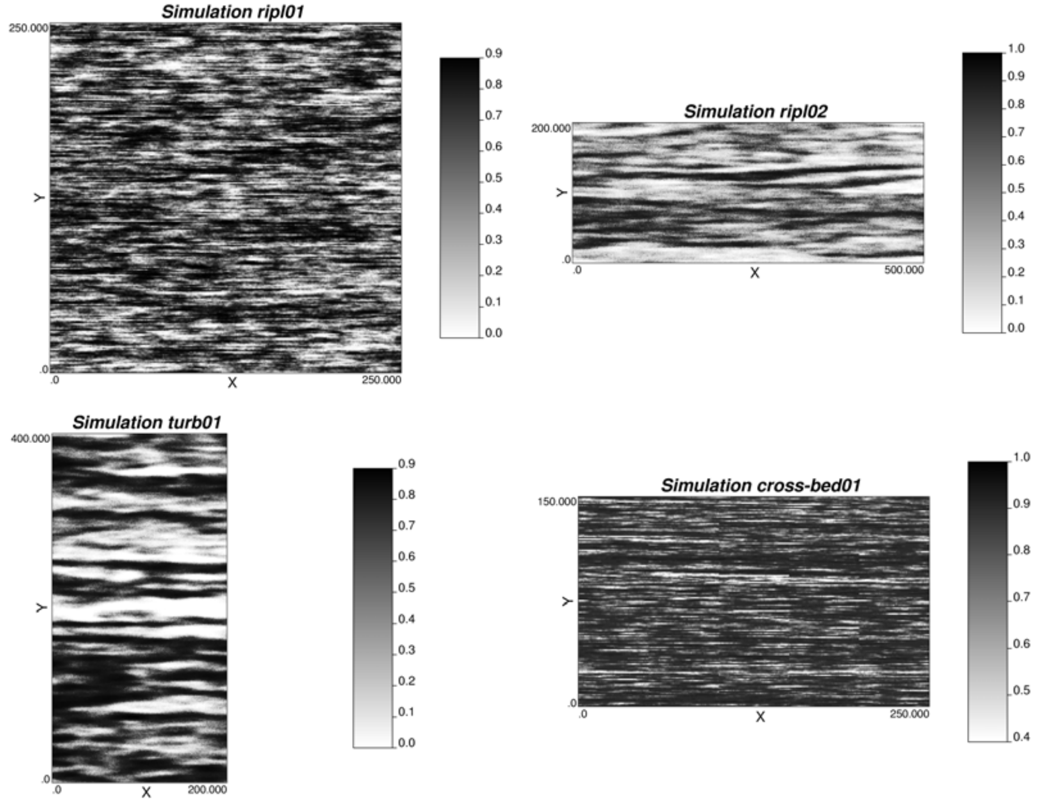


Fig. 6: Selected realizations of SGS for each image.

Using the mean as a threshold, the entropy was calculated for each of the one hundred realizations. A histogram of the results is shown in figure 7, where they are compared with the entropy measured for the reference image. The variance of the entropy distribution is very small for all images, and there seems to be almost no difference in the variance of the entropy distribution when comparing different images with different original entropy values. The characteristic of maximum entropy of the Gaussian distribution used for the simulation does not translate this property to the results (Journel and Deutsch, 1993). The mean value of the entropy distribution for the realizations is consistently higher than the entropy value of the corresponding reference image (Fig. 7). The increment in entropy due to the simulation process is higher for those images with high connectivity of extreme values. In the case with image *cross-bed01* may be a little different, since the cross-bedding stratification is more difficult to reproduce by the simulation process, and therefore the realizations are more unstructured compared with the other images.

Comparison between the reference and the simulated realization entropies, it is done considering the lower or upper quartiles threshold, two results became notorious. The first one is that, for the lower quartile threshold in the *ripl01*image, the entropy calculated from the realizations is less than the entropy measured in the reference with the same threshold. This also happens for the *turb01* image using the upper quartile as the threshold for comparison.

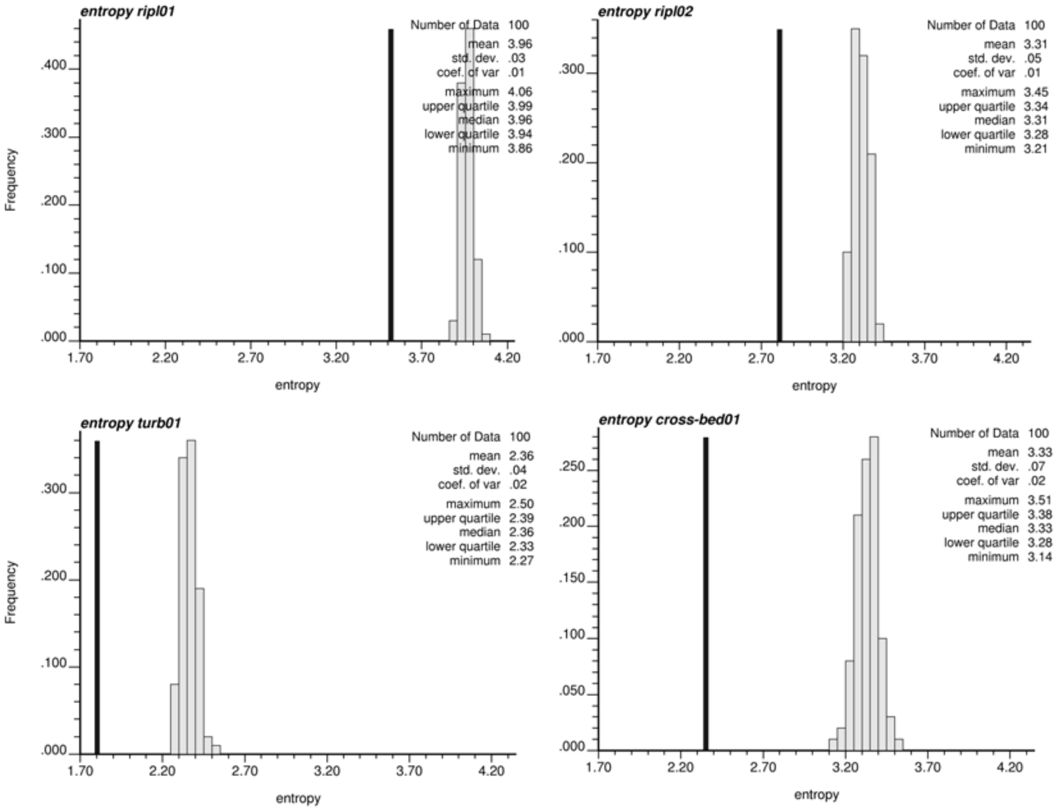


Fig. 7: Histograms of the one hundred realizations of each realization. The black bar represents the entropy measured for the reference image.

Entropy was also compared between the simulated realization of image *ripl02,* generated with different amounts of conditioning data; 5% and 0.25%, and yet the results are remarkably close to the ones resulting from using 1% of conditioning data (3.31; Table 3), 3.30 for the realization done with 5% samples and 3.26 with 0.25% samples

Image	**Entropy Values of Simulated Realization**		
	Lower quartile	**Mean**	**Upper quartile**
ripl01	2.06	3.96	3.05
ripl02	2.86	3.31	2.13
turb01	1.98	2.36	1.82
cross-bed01	1.83	3.33	3.63

Table 3: Mean entropy values for the hundred realizations, using different thresholds: the mean and the lower and upper quartiles.

Conclusions

Most simulation algorithms are based in two-points statistics to characterize the spatial distribution of the variable to model. Most geological processes lead to spatial structures that are more complex than fully described by a variogram.

Sequential Gaussian simulation uses the Gaussian distribution to obtain the simulated values. Its simplicity is balanced against the property of maximum entropy beyond the two-point statistics. As a result SGS realizations could have less spatial structure than reality. Entropy can be seen a measure of spatial disorder.

Entropy was measured in four reference images, using three different templates with an increasing number of possible configurations. The results confirm that as the number of configurations increases the entropy increases as well. The rate in which entropy increases with more possible configurations depends on the spatial complexity of the reference image.

Entropy was also measured at different thresholds for the indicator variable in the template. Results show that entropy will be higher when the threshold belongs to that range of the distribution that shows more spatial connectivity.

Entropy was also measured for one hundred realizations, obtained by SGS, the results show that entropy increases in the simulated images when the threshold chosen is the mean. When the chosen thresholds are the lower or upper quartile, entropy of the simulations can be lower than the entropy of the reference if there is higher connectivity in one of the extreme values. Of course, if the multiple point distribution was used in the geostatistical simulation, then the entropy would be the same as the original reference image.

The two serious problems faced by geostatistics is (1) inference of reliable statistics, perhaps multiple point, from data or structures we deem relevant to the location being studied, and (2) estimation or simulation using all deemed-relevant statistics. Future research will continue in these directions.

References

Deutsch, C.V., 1992. *Annealing Techniques Applied to Reservoir Modeling and the Integration of Geological and Engineering (Well Test) Data.* PhD. Thesis, Stanford University, Stanford, CA.

Deutsch, C.V. and Journel A.G., 1998. *GSLIB: Geostatistical Software Library: and User's Guide*, Oxford University Press, New York, 2nd Ed.

Galli, A., Beucher, H., Le Loc'h G. and Doligez, B. and Heresim Group, 1994. *The Pros and Cons of the Truncated Gaussian Method*, in Geostatistical Simulations, edited by M. Armstrong and P.A. Dowd, Kluwer Academic Publisher, pp. 217-233.

Journel, A.G. and Deutsch, C.V., 1993. *Entropy and Spatial Disorder.* Mathematical Geology, Vol. 25, No. 3, pp. 329-355.

Journel, A.G., 1893. *Nonparametric estimation of spatial distribution*. Mathematical Geology, Vol. 15, No. 3, pp. 445-468.

Journel, A.G. and Isaaks, E.H., 1984. *Conditional indicator simulation: Application to a Saskatchewan uranium deposit.* Mathematical Geology, Vol. 16, No. 7, pp. 685-718.

Larrondo P.F., Neufeld C.T., Deutsch, C.V., 2003. *VARFIT: A Program for Semi-Automatic Variogram Modeling*. Technical Report, Centre for Computational Geostatistics, University of Alberta. September 2003.

Tupin, F., Sigelle, M and Maître, H., 2000 *Definition of a spatial entropy and its use for SAR texture discrimination*. International Conference on Image Processing, Vancouver, September 2000.