# Hierarchical Simulation of Runs for Improved Geostatistical Models

Julián Ortiz C. (jmo1@ualberta.ca) and Clayton V. Deutsch (cdeutsch@ualberta.ca)
Centre for Computational Geostatistics, University of Alberta

## Abstract

*Conventional geostatistical techniques use only two-point statistics to characterize the continuity of a random variable. Reproduction of complex features, such as large range connectivity and non-linear behavior, calls for the use of multiple-point statistics. We propose the use of runs, that is, strings of adjacent data with a similar indicator value for a given threshold, to improve the numerical models, and hence the decisions made based on them. Runs are easily inferred from drillhole or well data, thus training images are not required. We present a brief discussion on the use of multiple-point statistics in geostatistical simulation, along with the theoretical background to predict the frequency of runs. We introduce a hierarchical method to simulate a random variable that honors the input histograms of runs above and below a given threshold. Implementation details are discussed. Several one dimensional examples are shown. The steps required to develop a full three dimensional implementation are discussed.*

## Introduction

Geostatistical techniques aim to reproduce important features of the data that permit an accurate quantification of the uncertainty after a transfer function. Conventional simulation techniques, such as sequential indicator and Gaussian simulation, rely only on the covariance function to characterize the continuity of the underlying regionalized variable. However, in most practical applications high-order continuity that cannot be captured by two-point statistics exists and considerably changes the resulting uncertainty after the transfer function, since long range continuity will critically affect flow properties and recoverable reserves. The advantage of characterizing differently continuity for low and high values was shown by Journel and Alabert [7]. The data they used showed a higher connectivity than both the indicator and Gaussian realizations. Multiple-point statistics are required to reproduce this n-point connectivity curve.

The idea of obtaining more representative realizations of the phenomenon motivates the incorporation of multiple-point statistics in simulation. Several attempts have been made to represent these features, however, none of them have found wide applicability in practice. Two problems are systematically encountered: the inference of multiple-point statistics and their incorporation in a simulation framework.

Inference of multiple-point statistics is difficult because a stationarity assumption must be made. Data must be pooled together to allow inference of frequencies of different configurations of the points. Some of the configurations of interest may not be available a sufficient number of times to allow a reliable estimation of their frequencies. The use of training images partially overcomes this problem, since there are many replications of configurations in an exhaustive image. However some events may still not be present in the training image. Finally, it is hard to know how well the training image represents the actual study area.

Extracting multiple-point statistics from the data avoids the problem of representivity, although not all desired configurations of multiple-points can be found. In petroleum and mining applications, data come in strings from drillholes or wells. The connectivity of adjacent points in a line or runs can be assessed. Curvilinear features of the phenomenon will probably not be captured by this statistic, however, the use of runs relies on statistical inference *from* data.

Using multiple-point statistics in simulation has been addressed several times. The use of extended normal equations was proposed by Guardiano and Srivastava [5]. It is a generalization of sequential indicator simulation. The implementation of this technique was improved by Strebelle and Journel [11], by using a search tree to find the frequencies of the multiple-point events in the training image. Deutsch [3] applied simulated annealing for reservoir modelling. The difficult setting of the annealing schedule and high computational cost of this technique makes it unappealing to practitioners. Another iterative technique was proposed by Caers [2] that is based on the use of neural networks to model the conditional distribution function in a non-linear fashion.

All implementations proposed assume that multiple-point statistics are available. They consider training images for their inference. Features that belong to the training image but not to the underlying process being simulated has not been addressed. We may want to reproduce the general appearance of the training image but not all its details. Caers [2] uses a technique to avoid overtraining the neural network, however the question of which features should be extracted from the training image is not answered. Furthermore, transferring statistics from the training image to the realization is a problem. The univariate and bivariate statistics of the training image may not be exactly the same than those of the phenomenon. Once again, the use of multiple-point statistics inferred from the data does not have this problem, since the statistics are consistent to a common spatial law.

Implementation of simulation that accounts for runs above and/or below different thresholds requires a hierarchical framework. The runs are by definition nested from one threshold to the next. The indicator coding of the data is used, so that the samples from the drillholes become zeros and ones. The n-point connectivity statistic proposed by Journel and Alabert [7] is equivalent to the definition of runs documented by Mood in the early forties [8].

We present a quick review on the theory of runs and show a method to predict the frequency of runs when the spatial law is known. We show the results for the multi-Gaussian case. Furthermore, we compare the results of a variable without spatial correlation with the limit distributions encountered by Mood. Finally, we introduce the algorithm for simulating runs above and below a threshold and show some examples.

## The Theory of Runs

A. M. Mood [8] derived the distributions of runs of two and $k$ kinds of elements, showing their moments and asymptotic distributions as the number of elements $n$ tends to infinity, from random arrangements of fixed numbers of elements, and for binomial and multinomial populations.

Recall that given a sequence of numbers $\{z(\mathbf{u}_\alpha), \alpha = 1, ..., n\}$ and $K$ thresholds $\{z_k, k = 1, ..., K\}$, $K$ new sequences can be built using an indicator coding:

$$i(\mathbf{u}_\alpha; z_k) = \begin{cases} 1, & \text{if } z(\mathbf{u}_\alpha) \leq z_k \\ 0, & \text{otherwise} \end{cases} \qquad k = 1, ..., K$$

where $z(\mathbf{u}_\alpha)$ is the value at the location $\mathbf{u}_\alpha$.

In this manner, the continuous variable $Z$ is transformed into a discrete binary variable $I$. This could be extended to categorical variables that are coded into one of $K$ classes, although this falls outside the scope of this paper.

Now that a sequence of elements of two kinds has been constructed, we can define the concept of runs.

## Definition

Consider the following sequence of uniform numbers in [0,1] and the three sequences coded for thresholds 0.25, 0.50, and 0.75.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.24 | 0.35 | 0.07 | 0.85 | 0.94 | 0.66 | 0.48 | 0.65 | 0.35 | 0.79 | 0.19 | 0.65 | 0.38 | 0.95 | 0.58 | 0.11 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_1 = 0.25$ : | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| $z_2 = 0.50$ : | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| $z_3 = 0.75$ : | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |

A run of length $L$ above a threshold $z_k$ can be identified as a sequence of $L + 2$ adjacent nodes valued as zeros, except for the first and last nodes, valued as ones. The first run above the threshold 0.25 in the previous example is of length $L = 1$. The second run has a length $L = 7$ and is highlighted.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_1 = 0.25$ : | 1 | 0 | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **1** | 0 | 0 | 0 | 0 | 1 |

Notice also that the runs above the threshold are nested, that is, at low thresholds, the runs above the threshold contain the runs above a higher threshold. This is shown next, where runs above the thresholds 0.25, 0.50, and 0.75 are highlighted, showing the nesting.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_1 = 0.25$ : | 1 | 0 | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **1** | 0 | 0 | 0 | 0 | 1 |
| $z_2 = 0.50$ : | 1 | 1 | **1** | **0** | **0** | **0** | **1** | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| $z_3 = 0.75$ : | 1 | 1 | **1** | **0** | **0** | **1** | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |

The same happens with runs below a threshold.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_1 = 0.25$ : | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| $z_2 = 0.50$ : | 1 | 1 | 1 | 0 | 0 | **0** | **1** | **0** | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| $z_3 = 0.75$ : | 1 | 1 | 1 | 0 | **0** | **1** | **1** | **1** | **1** | **0** | 1 | 1 | 1 | 0 | 1 | 1 |

## Distributions of Runs of Length $i$

Using combinatorial analysis Mood derived the distribution for the number of runs of length $i$ of elements of one kind. The first and second moments are derived from the expression for the probability distribution. Given a sequence with $n_1$ elements of one kind and $n_2$ elements of another kind ($n_1$ and $n_2$ are fixed and $n_1 + n_2 = n$), the expected number of runs of kind 1 of length $i$ can be calculated using the following expression:

$$E(r_{1i}) = \frac{(n_2+1)^{(2)} n_1^{(i)}}{n^{(i+1)}}$$

where the factorial $x^{(a)}$ corresponds to $x^{(a)} = x \cdot (x-1) \cdot (x-2) \cdot \ldots \cdot (x-a+1)$.

The variance and covariance between the number of runs of different lengths $i$ and $j$ for elements of one kind are given by:

$$\begin{aligned} \sigma_{ij} &= \frac{n_2^{(2)}(n_2+1)^{(2)} n_1^{(i+j)}}{n^{(i+j+2)}} - \frac{n_2^2(n_2+1)^2 n_1^{(i)} n_1^{(j)}}{n^{(i+1)} n^{(j+1)}} \\ \sigma_{ii} &= \frac{n_2^{(2)}(n_2+1)^{(2)} n_1^{(2i)}}{n^{(2i+2)}} + \frac{(n_2+1)^{(2)} n_1^{(i)}}{n^{(i+1)}} \left(1 - \frac{(n_2+1)^{(2)} n_1^{(i)}}{n^{(i+1)}}\right) \end{aligned}$$

The expected value for the total number of runs of elements of one kind and its variance are given by:

$$\begin{aligned} E(r_1) &= \frac{(n_2+1)n_1}{n} \\ \sigma_{r_1}^2 &= \frac{(n_2+1)^{(2)} n_1^{(2)}}{n n^{(2)}} \end{aligned}$$

Finally, when $n_1$ and $n_2$ are fixed, the distribution of the total number of runs of elements of one kind is asymptotically normal:

$$r_1 \sim N\left(\frac{n_1 n_2}{n}, \frac{n_1^2 n_2^2}{n^3}\right)$$

When the number of elements are random variables drawn from a binomial population, then the numbers $n_1$ and $n_2$ are not fixed and the results change. The expected number of runs of a given length $i$, its variance, and covariance become:

$$\begin{aligned} E(r_{1i}) &= p_1^i p_2 \{(n-i-1)p_2 + 2\} \\ \sigma_{ij} &= p_1^{i+j} p_2^2 \{(n-i-j)^{(2)} p_2^2 + (n-i-j)p_2(1+5p_1) \\ &\quad +6p_1^2 - ((n-i-1)p_2+2)((n-j-1)p_2+2)\} \\ \sigma_{ii} &= p_1^{2i} p_2^2 \{(n-2i)^{(2)} p_2^2 + (n-2i)p_2(1+5p_1) \\ &\quad +6p_1^2 - ((n-i-1)p_2+2)^2\} + p_1^i p_2((n-i-1)p_2+2) \end{aligned}$$

where $p_1$ and $p_2 = 1 - p_1$ are the probabilities of drawing an element of kind one and two respectively.

And the limit distribution of the total number of runs of both kinds is asymptotically normal with the following mean and variance:

$$r \sim N\left(2np_1p_2, 4np_1p_2(1-3p_1p_2)\right)$$

If the data were coded as indicators then the elements of kind 1 could be the ones and the elements of kind 2, could be the zeros. The probabilities $p_1$ and $p_2$ would then be the probability of the random variable to be below a threshold and its complement, respectively.

# Deriving the Frequency of Runs

## General Case

A run of length $L$ of elements below a threshold can be seen as the event of having a string of nodes of length $L+2$, such that the first and the last values are above the threshold $z_k$ and all other nodes are below the threshold value. This multiple-point event occurs with the following joint probability:

$$Prob\{\text{Run of length L}\} = Prob\{Z_i > z_k, Z_{i+1} \leq z_k, ..., Z_{i+L+1} \leq z_k, Z_{i+L+2} > z_k\}$$

This joint probability can be calculated by a recursive application of Bayes' postulate, that is, if $A$ and $B$ are two events (multiple-point events or not), the probability of both events happening is equal to the probability of the first event conditional to the second multiplied by the probability of the second event occurring:

$$Prob\{A, B\} = Prob\{A|B\} \cdot Prob\{B\}$$

In the case of runs, the multiple-point event "run of length $L$" has a probability of occurring given by:

$$\begin{aligned} Prob\{\text{Run of length L}\} &= Prob\{Z_i > z_k | Z_{i+1} \leq z_k, ..., Z_{i+L+1} \leq z_k, Z_{i+L+2} > z_k\} \cdot \\ &\quad Prob\{Z_{i+1} \leq z_k | Z_{i+2} \leq z_k, ..., Z_{i+L+1} \leq z_k, Z_{i+L+2} > z_k\} \cdot ... \cdot \\ &\quad Prob\{Z_{i+L+1} \leq z_k | Z_{i+L+2} > z_k\} \cdot Prob\{Z_{i+L+2} > z_k\} \end{aligned}$$

## The Multi-Gaussian Case

The conditional probabilities involved in the calculation can only be completely retrieved in a case where the spatial law is fully known. This condition is of course extremely difficult to satisfy. However, in the multi-Gaussian case all conditional distributions are characterized by the one- and two-point statistics, that is, by the mean and covariance matrix.

Lets denote the multi-Gaussian variable $Y$. It could be the normal score transform of $Z$. If we consider a threshold $y_k$ we can code the data as indicators and rewrite the expression for the joint probability as:

$$\begin{aligned} Prob\{\text{Run of length L}\} &= Prob\{I_i = 0 | I_{i+1} = 1, ..., I_{i+L+1} = 1, I_{i+L+2} = 0\} \cdot \\ &\quad Prob\{I_{i+1} = 1 | I_{i+2} = 1, ..., I_{i+L+1} = 1, I_{i+L+2} = 0\} \cdot ... \cdot \\ &\quad Prob\{I_{i+L+1} = 1 | I_{i+L+2} = 0\} \cdot Prob\{I_{i+L+2} = 0\} \end{aligned}$$

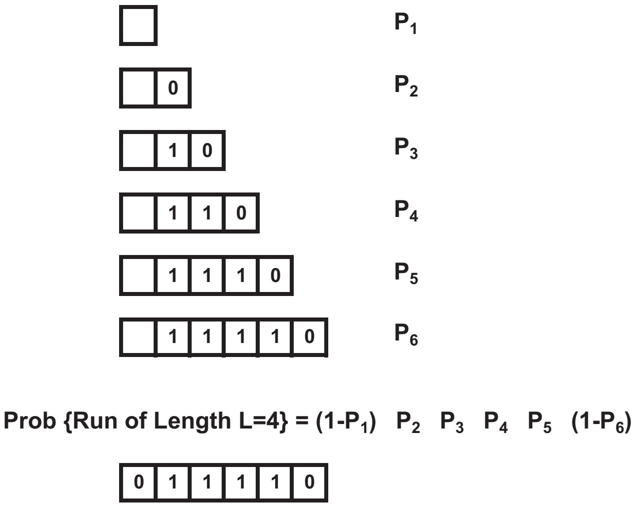


Figure 1: Illustration of the increasing conditioning in the calculation of the joint probability of having a run of length L. L=4 in this example.

Now, the conditional probabilities can be calculated exactly in the multi-Gaussian case, by simple indicator kriging [7]:

$$Prob\{Y_i \leq y_k | (n)\} = Prob\{I_i = 1 | (n)\} = \sum_{j=1}^{n} \lambda_j \cdot I_j + (1 - \sum_{j=1}^{n} \lambda_j) \cdot F(y_k)$$

where the $\lambda_j, j = 1, ..., n$ are the solution of the system:

$$\sum_{j=1}^{n} \lambda_j \cdot C_I(k,j) = C_I(k,i) \quad k = 1, ..., n$$

The increasing conditioning is illustrated in **Figure 1**. Notice that the covariance function $C_I(k,j)$ has to be calculated consistently with the continuous variogram for the Gaussian variable, as shown in [4].

## Example

The previous results can be tested by constructing a realization of a multi-Gaussian variable with a known covariance function. Then the continuous simulated values can be coded as indicators for a given threshold, say the median, and the indicator variogram can be calculated for that threshold. This experimental indicator variogram must be modelled. Using simple indicator kriging, as illustrated in **Figure 1**. A chart showing a comparison between the theoretical calculation and the experimental result is presented in **Figure 2**. An almost perfect match between the two suggests that the n-point connectivity curve calculated by Journel and Alabert [7] could have been predicted analytically beforehand.

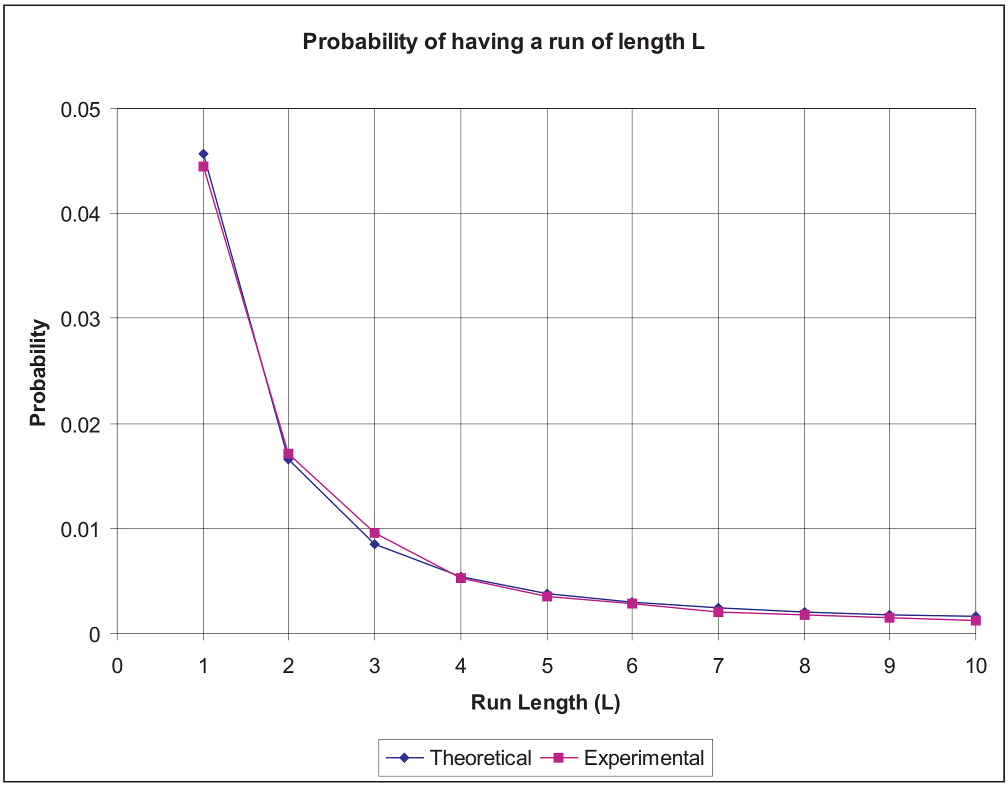


Figure 2: Comparison of theoretical and experimental results for the calculation of the probability of having a run of length L.

### The Random Case: Relation with Mood's Results

Mood's results are calculated in a finite domain of size $n$. The probabilities derived using a recursive application of Bayes' postulate are defined in an infinite (ergodic) domain. When all elements are drawn independently, then Bayes postulate becomes:

$$Prob\{A,B\} = Prob\{A\} \cdot Prob\{B\}$$

Thus, all conditional distributions are reduced to their marginals. The probability of having a run of elements below a threshold of length $L$ is simply:

$$\begin{aligned} Prob\{\text{Run of length L}\} &= Prob\{I_i = 0\} \cdot Prob\{I_{i+1} = 1\} \cdot ... \cdot \\ &\quad\; Prob\{I_{i+L+1} = 1\} \cdot Prob\{I_{i+L+2} = 0\} \\ &= (1-p_1) \cdot p_1 \cdot p_1 \cdot ... \cdot p_1 \cdot (1-p_1) \\ &= (1-p_1)^2 \cdot p_1^L \\ &= p_1^L p_2^2 \end{aligned}$$

Taking Mood's expected values for runs of length $L$, when the elements above and below the threshold are drawn from a binomial population, and considering the total number of events of length $L+2$ in a sequence of $N$ nodes, the proportion of multiple-point events of size $L+2$ of which the nodes have a run configuration can be calculated. Thus, the corresponding probability can be derived by taking the limit as $n$ approaches infinity:

$$\begin{aligned} Prob(r_{1L}) &= \lim_{n\to\infty} \frac{p_1^L p_2\{(n-L-1)p_2+2\}}{n-(L+2)+1} \\ &= \lim_{n\to\infty} \left( p_1^L p_2^2 + \frac{2p_1^L p_2}{n-L-1} \right) \\ &= p_1^L p_2^2 \end{aligned}$$

Thus the result derived by Mood is exactly the same we propose using the analytical derivation proposed above. The total number of runs can be calculated as the infinite sum of the expected number of runs of length $L$:

$$\begin{aligned} E(r) &= E(r_1 + r_2) = E\left( \sum_{L=1}^{\infty} (r_{1L} + r_{2L}) \right) \\ &= \sum_{L=1}^{\infty} n\left(Prob(r_{1L}) + Prob(r_{2L})\right) \\ &= \sum_{L=1}^{\infty} n \left( p_1^L p_2^2 + p_2^L p_1^2 \right) \\ &= \sum_{L=1}^{\infty} n \left( p_1^L (1-p_1)^2 + p_2^L (1-p_2)^2 \right) \end{aligned}$$

Rearranging the values and using the following result for the infinite sum [6]:

$$\lim_{n\to\infty} \sum_{L=1}^{n} p^L = \frac{p}{1-p} \qquad |p| < 1$$

we get:

$$\begin{aligned} E(r) &= n(p_1(1-p_1)+p_2(1-p_2)) \\ &= n(p_1p_2+p_2p_1) \\ &= 2np_1p_2 \end{aligned}$$

which is exactly the value given by Mood.

## RUNS: The Algorithm

An algorithm to simulate a field in one dimension such that the histograms of runs above and below a threshold are reproduced is presented. All the nodes in the domain start above the threshold and then some of them are selected to be below the cutoff. This is done with the help of a "selection function" that will be maximum if the change of a node minimizes the difference between the current and target histograms of runs above and below. Details of the algorithm along with preliminary examples are shown.

### Input Statistics

The program requires two basic input statistics:

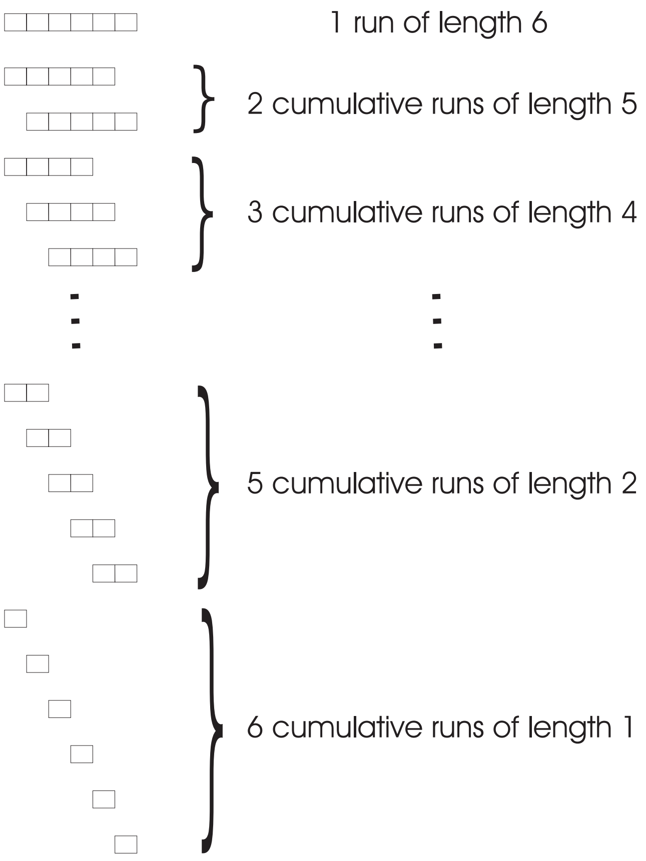
- Proportion of nodes below the threshold: this corresponds to the cdf value of the threshold.
- Histograms of runs: the histograms of runs above and below the threshold are required. They can be inferred from drillhole data or from any source where the samples represent the same support and are linearly located (e.g. samples from trenches).

The calculation of runs is illustrated in **Figure 3** [9]. That is, if we take what classically is called a run of length $L$, we consider it as one run of length $L$ plus two runs of length $L-1$, plus three runs of length $L-2$, and so forth. In general a classically defined run of length $L$ corresponds to $i$ runs of length $L-i+1$, with $i=1,...,L$. Therefore, if in a very simplified case, the target histogram consists of a run of length 3 and two runs of length 2, in the classical sense, the cumulative histogram would be calculated, based on the cumulative concept here presented, as the sum of the three histograms shown in **Figure 4**.

Scaling the histogram of runs from drillhole length to the length of the field being simulated is an issue that must be addressed. The expected number of runs of a given length changes as the field size increases. Further investigation in this topic is required to be able to use the information inferred from the data in simulation. For now, it is assumed that the histograms input in the algorithm are those that effectively represent the runs above and below the threshold at the scale of the simulation domain.

### Implementation

Simulation of runs above and below a given threshold has been implemented for a one dimensional domain. The algorithm can be summarized as:



1 run of length 6
2 cumulative runs of length 5
3 cumulative runs of length 4
5 cumulative runs of length 2
6 cumulative runs of length 1

Figure 3: Example to illustrate the concept of "cumulative runs". One single run of length 6 corresponds to 2 cumulative runs of length 5, 3 cumulative runs of length 4,..., and 6 cumulative runs of length 1.

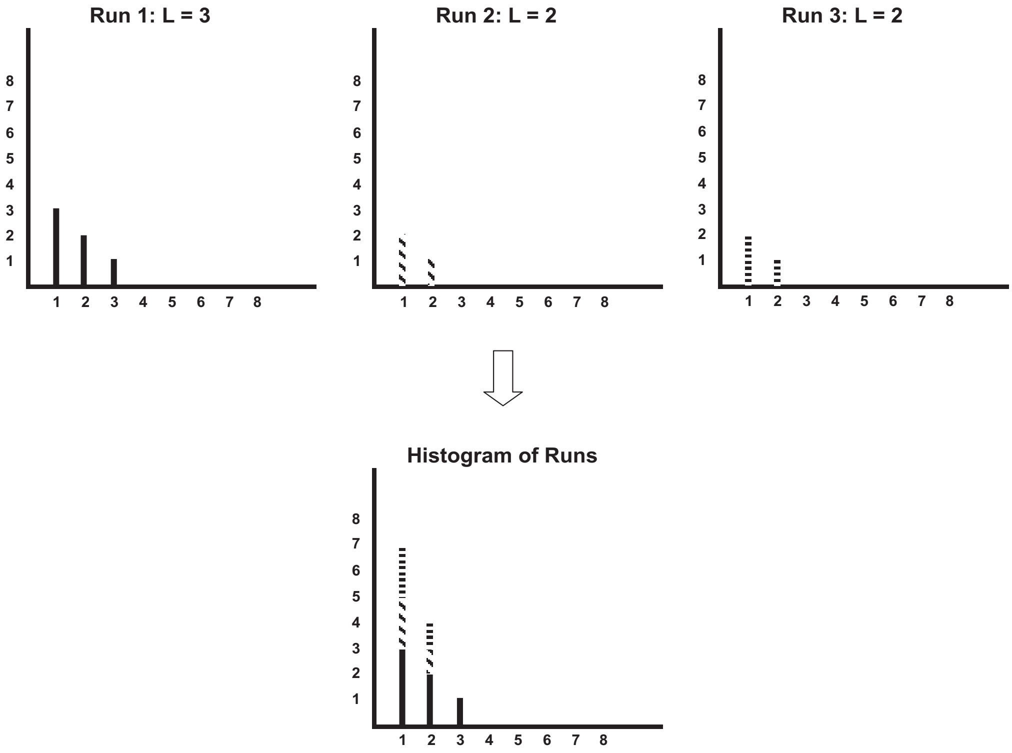


Figure 4: Example to calculate the cumulative histogram of runs given three runs defined classically.

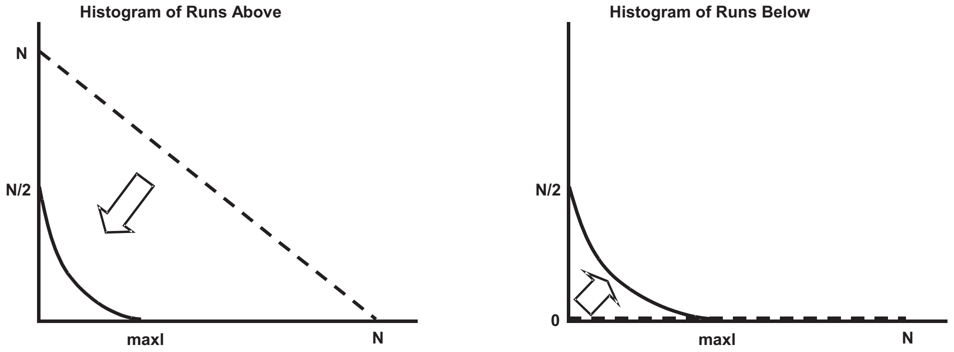


Figure 5: Histograms of runs above and below the threshold at the beginning of the simulation. The current histograms must converge to the targets.

1. Set all nodes above the threshold (indicator value equals zero).
2. Pick randomly a node and assign it a value of one, thus it is assumed to be below the threshold. This node is now unavailable for the remaining of the simulation.
3. Calculate current histograms of runs above and below the threshold (see **Figure 5**).
4. For all available nodes in the grid, that is, all nodes above the threshold, calculate the value of the selection function as:

$$s(\mathbf{u}_i) = \prod_{l=1}^{2 \cdot maxl} f_{above}(l) \cdot f_{below}(l), \qquad i = 1, ..., N$$

where $l$ is the length of a cumulative runs, $maxl$ is the maximum length of runs in the target histograms and

$$f_{above}(l) = \begin{cases} 0.05 + 0.95 \cdot e^{-(\frac{-\Delta}{a})^w} & , \text{ if } \Delta < 0 \\ 0.05 + 0.95 \cdot e^{-(\frac{\Delta}{a})^w} & , \text{ otherwise} \end{cases}$$

with $\Delta$ being the difference between the number of cumulative runs of length $l$ in the current histogram and the number of cumulative runs of the same length in the target histogram:

$$\Delta = freq_{curr}(l) - freq_{targ}(l)$$

$f_{below}(l)$ is defined exactly as $f_{above}(l)$ but with

$$\Delta = freq_{targ}(l) - freq_{curr}(l)$$

The parameters $a$ (scaling) and $w$ (power) give the shape of the function $f(l)$, as illustrated in **Figure 6**.

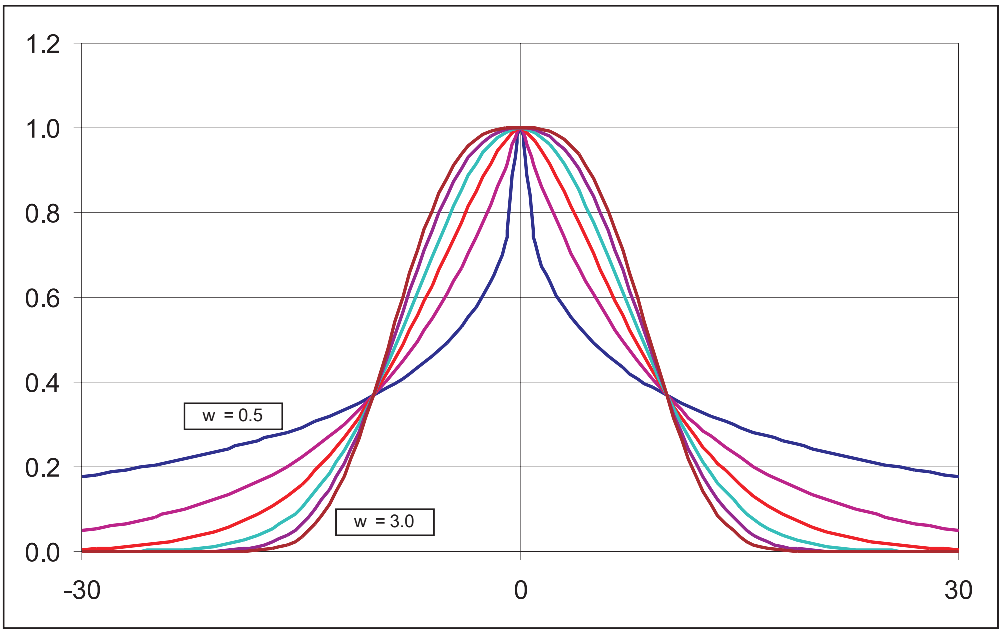


Figure 6: Function $f(l)$ used in the calculation of the selection function value for each candidate node to be switched. In this case, the parameter $a$ has been set to 10 and $w$ takes the values 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0.

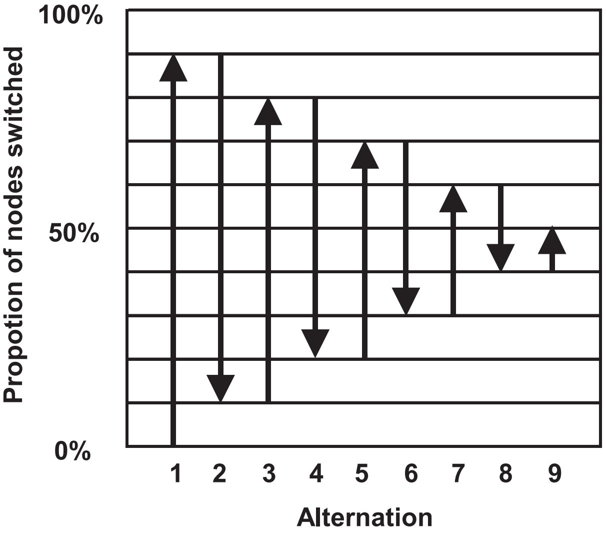


Figure 7: The concept of alternating to converge to the solution is showed in this schematic.

5. Switch the node with maximum value in the selection function. If more than one node have the same maximum value, pick one randomly.
6. Update current histograms of runs above and below the threshold.
7. Goto 4 and repeat until enough nodes have been switched to one.

Several problems found during the implementation process have been solved using rather classical solutions: the grid was wrapped around to avoid edge effects [2, 10, 12]; The algorithm alternates around the desired proportion to ensure convergency (see for example [1]). This last point is done by switching more nodes than required below the threshold, then switching nodes to be above it, then switch nodes to be below and so forth, until the right number of nodes below the threshold is achieved. **Figure 7** shows an example where 9 alternations are used. The target proportion is 50%, i.e. the threshold corresponds to the median. Each alternation goes beyond the required number to be switched.

Precision problems in the calculation of the selection function value were solved by taking an arbitrary number of lengths, in this case $2 \cdot maxl$, where $maxl$ is the maximum length on the target histograms of runs. Also, the function $f$ has a minimum value of 0.05 to avoid values too close to 0.

## Categorical Binary Case

The algorithm was tested for several one dimensional cases, where two categories are present. The parameters $a$ and $w$ are critical to ensure convergency. After a sensitivity analysis, it was found that $a$ should be approximately the length of the simulated array, while $w$ is

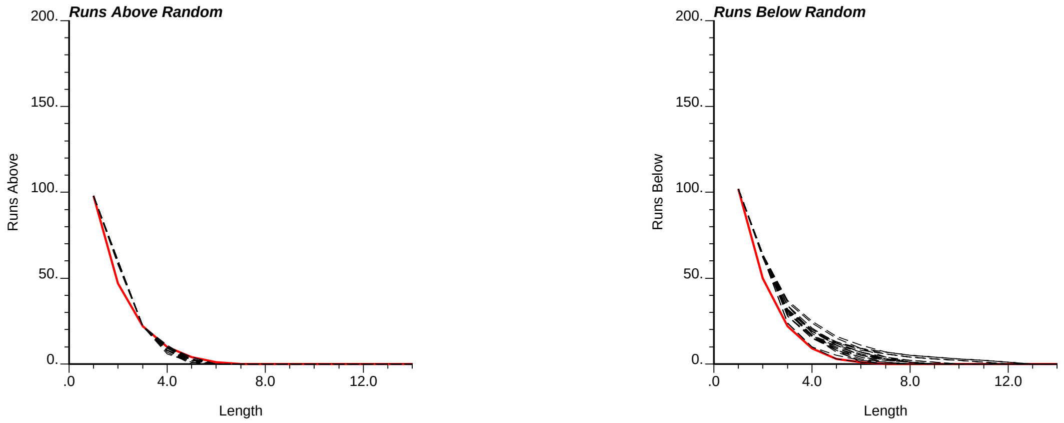


Figure 8: Reproduction of runs above and below the median for a random sequence.

dependent on the complexity of the problem. All of the examples presented worked with a $w = 4.6$.

### Random Case

The first case consists on simulating the runs found on random sequences, when coded as above or below a threshold. A reference sequence of random numbers was generated with the random number generator `acorni`. The numbers generated were then coded as 1 if below or equal 0.5, and 0 if above 0.5. The reproduction of the histograms of runs above and below the threshold is shown in **Figure 8**. Some fluctuations around the target values are observed.

### Regular Case

A regular array with sets of five nodes above and five below the threshold was used as a reference to test the algorithm. The reproduction of the histograms of runs is shown in **Figure 9**. As with the random case, the histogram of runs above the threshold seems to be better reproduced. This could be caused by the alternation sequence that tends to give more importance to the histogram above, since the differences are larger (see **Figure 5**).

### Multi-Gaussian Case

A one dimensional array was simulated using the algorithm `sgsim` of GSLIB [4]. The realization was then truncated at the median to generate a binary array. The histograms of runs extracted from it was simulated using the algorithm `RUNS`. The resulting histograms of runs are presented in **Figure 10**. Notice the good reproduction obtained.

### Real Data 1

A string from an exhaustive data set obtained from scanning a realistic geological image was used to obtain the multiple-point runs statistics. The resulting simulated sequences

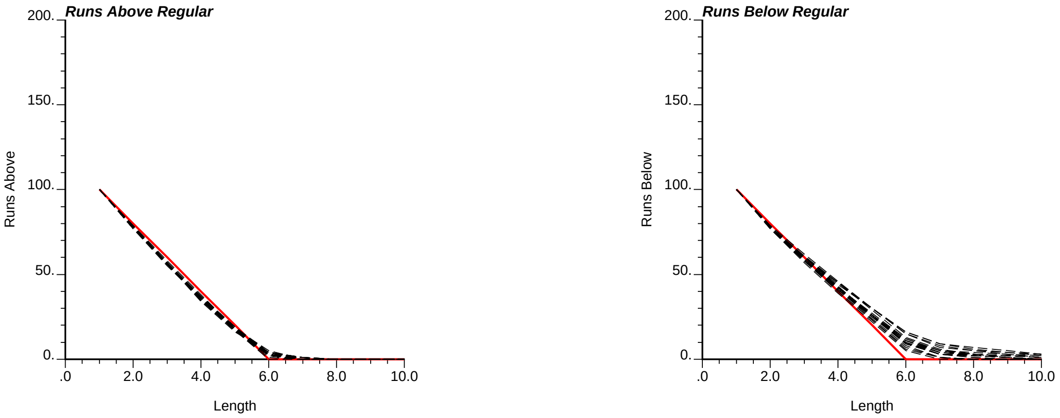


Figure 9: Reproduction of runs above and below the median for a regular sequence.

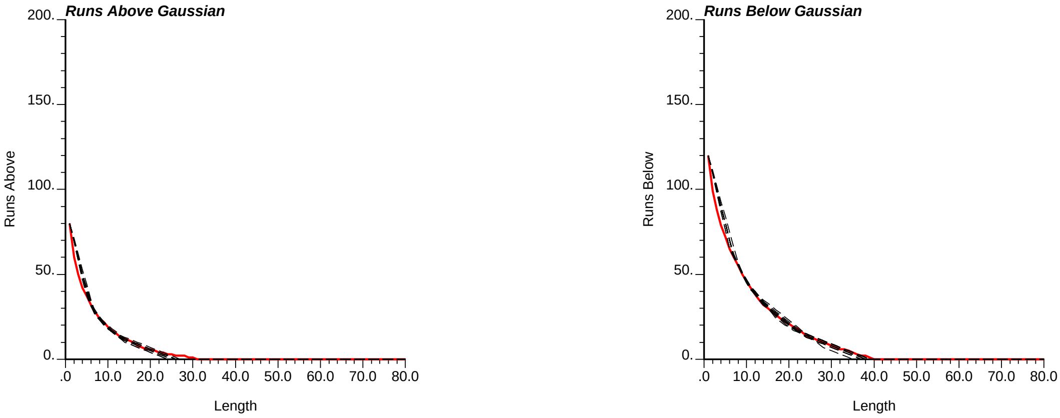


Figure 10: Reproduction of runs above and below the median for a binary array obtained by truncating a multi-Gaussian sequence.

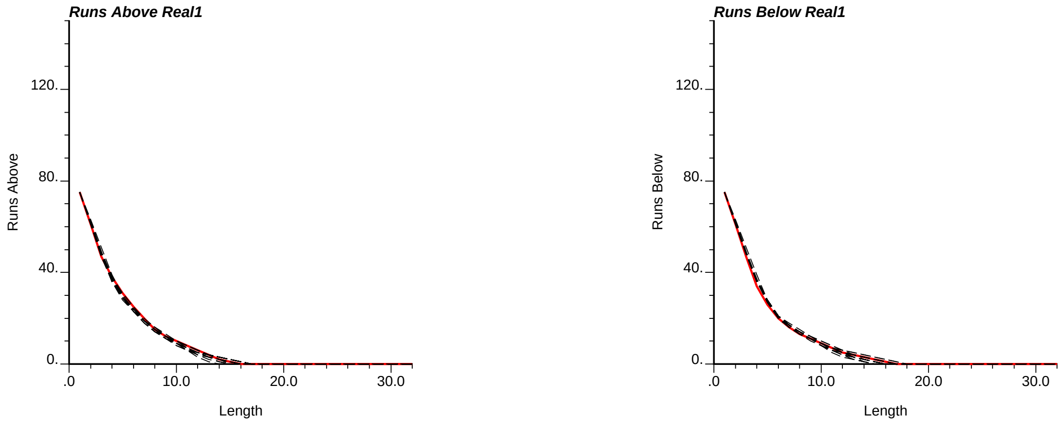


Figure 11: Reproduction of runs above and below the median for a binary array obtained from a realistic exhaustive data set.

showed very good reproduction of the reference statistics as shown on **Figure 11**.

### Real Data 2

A second realistic example was done. Again the data come from an image that shows realistic geological features. The histogram of runs was again very well reproduced (**Figure 12**).

## Discussion and Further Work

The use of runs should improve the performance of numerical models after a transfer function. The use of multiple-point statistics extracted from sampling data, rather than from a training image, permits the reproduction of complex features without doubting on their representativeness.

The algorithm presented must be generalized to handle conditioning data, scaling of the multiple point statistics and to generate three dimensional realizations. Some of the steps that should be considered next are simulating multiple thresholds hierarchically, and then generalizing to three dimensions. The use of runs in off diagonal directions has to be considered, along with the possible need to also incorporate the indicator variograms as target statistics.

The relationship between runs and the indicator variogram deserves special attention. In one dimension, given the histograms of runs above and below a threshold, the realizations based solely on these statistics will generate a range of indicator variograms, since they will only sample the combinatorial space of the total number of samples. The goal, for now is to see if that entire sample space can be explored with the method proposed. The next step will be to assess the performance of a realization that honors the runs (but not the indicator variogram) after a transfer function. We expect a better result from the realization that uses multiple-point statistics even when the two-point statistics are not honored.

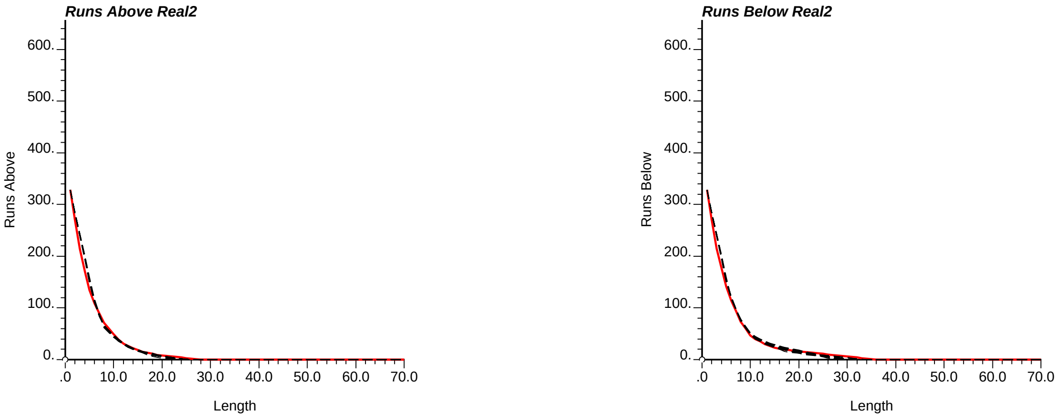


Figure 12: Reproduction of runs above and below the median for a binary array obtained from a second realistic exhaustive data set.

# References

[1] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–598, September 1985.

[2] J. Caers. Stochastic simulation using neural networks. In *Report 11, Stanford Center for Reservoir Forecasting*, Stanford, CA, May 1998.

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

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

[5] F. B. Guardiano and R. M. Srivastava. Borrowing complex geometries from training images: The extended normal equations algorithm. In *Report 5*, Stanford, CA, May 1992. Stanford Center for Reservoir Forecasting.

[6] J. W. Harris and H. Stocker. *Handbook of Mathematics and Computational Science.* Springer, 1998.

[7] A. G. Journel and F. Alabert. Non-Gaussian data expansion in the earth sciences. *Terra Nova*, 1:123–134, 1989.

[8] A. M. Mood. The distribution theory of runs. *Annals of Mathematical Statistics*, 11:367–392, December 1940.

[9] J. Ortiz C. Characterization of high order correlation for enhanced indicator simulation. In *Centre For Computational Geostatistics*, volume 3, Edmonton, AB, 2001.

[10] R. M. Srivastava. Iterative methods for spatial simulation. In *SCRF report*, Stanford, CA, May 1992.

[11] S. Strebelle and A. G. Journel. Sequential simulation drawing structures from training images. In *6th International Geostatistics Congress*, Cape Town, South Africa, April 2000. Geostatistical Association of Southern Africa.

[12] L. Wang. Modeling complex reservoir geometries with multiple-point statistics. Technical report, Stanford Center for Reservoir Forecasting, Stanford, CA, May 1995.