POST-PROCESSING OF MULTIPLE-POINT GEOSTATISTICAL MODELS TO IMPROVE REPRODUCTION OF TRAINING PATTERNS

SEBASTIEN STREBELLE[1] and NICOLAS REMY[2]
[1] ChevronTexaco Energy Technology Company,
6001 Bollinger Canyon Road, San Ramon, CA 94583, USA
[2] Department of Geological and Environmental Sciences
Stanford University, Stanford, CA 94305, USA

Abstract. In most petroleum and groundwater studies, flow performance is highly dependent on the spatial distributions of porosity and permeability. Because both porosity and permeability distributions primarily derive from facies deposition, facies should be the first property to be modeled when characterizing a reservoir. Yet, traditional geostatistical techniques, based on variogram reproduction, typically fail to model geologically-realistic depositional facies. Indeed, variograms only measure facies continuity between any two points in space; they cannot account for curvilinear and/or large-scale continuous structures, such as sinuous channels, that would require inferring facies joint-correlation at many more than two locations at a time.

Multiple-point geostatistics is a new emerging approach wherein multiple-point facies joint-correlation is inferred from three-dimensional training images. The simulation is pixel-based, and proceeds sequentially: each node of the simulation grid is visited only once along a random path, and simulated values become conditioning data for nodes visited later in the sequence. At each unsampled node, the probability of occurrence of each facies is estimated using the multiple-point statistics extracted from the training image. This process allows reproducing patterns of the training image, while honoring all conditioning sample data.

However, because of the limited size of the training image, only a very limited amount of multiple-point statistics can be actually inferred from the training image. Therefore, in practice, only a very few conditioning data close to the node to be simulated are used whereas farther away data carrying important large-scale information are generally ignored. Such approximation leads to inaccurate facies probability estimates, which may create "anomalies", for example channel disconnections, in the simulated realizations. In this paper, a method is proposed to use more data for conditioning, especially data located farther away from the node to be simulated. A measure of consistency between simulated realizations and training image is then defined, based on the number of times each simulated value, although initially identified as a conditioning datum to simulate a nearby node, had to be ignored eventually to be able to infer from the training image the conditional probability distribution at that node. Re-simulating the most inconsistent node values according to that measure enables improvement in the reproduction of training patterns without any significant increase of computation time. As an application, that post-processing process is used to remove channel disconnections from a fluvial reservoir simulated model.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 979–988.

1 Introduction

Most geological environments are characterized by successive depositions of elements, or rock bodies, through time. These elements are traditionally grouped into classes, commonly named "depositional facies", that correspond to particular combinations of lithology, physical and biological structures. For example, the typical depositional facies encountered in fluvial environments are high permeability sand channels, and sometimes, levies and splays with variable ranges of permeability.

Reservoir heterogeneity, hence flow performance, is primarily controlled by the spatial distribution of those depositional facies. Thus, a best practice would consist of modeling first depositional facies, and then populating each simulated facies with its corresponding specific porosity and permeability distributions. Yet traditional facies modeling techniques show severe limitations:

1. Variogram-based techniques, for example sequential indicator simulation (Deutsch and Journel, 1998), do not allow modeling geologically-realistic depositional elements because identification of two-point statistics, as modeled by the variogram, is not sufficient to characterize curvilinear or long-range continuous facies such as sand channels (Strebelle, 2000).
2. Object-based modeling techniques (Viseur, 1997) allow modeling quite realistic elements, but their conditioning is still commonly limited to a few wells.

An alternative technique proposed by Guardiano and Srivastava (1993) consists of going beyond the two-point statistics variogram by extracting multiple-point statistics from a training image. The training image can be defined as a three-dimensional conceptual geological model that depicts the geometry of each depositional facies expected to be present in the subsurface, as well as the complex spatial relationships existing among the different facies. Training images are typically obtained by interpreting available field data (cores, well logs, seismic), but also by using information from nearby field analogues and outcrop data. Figure 1a displays an example of a training image for a 2D horizontal section of a fluvial reservoir. That particular image was hand-drawn by a geologist, then numerically digitized. In 3D applications, three-dimensional training images are preferentially created using unconditional object-based modeling techniques.

Multiple-point statistics (MPS) simulation consists of reproducing patterns displayed in the training image, and anchoring them to the data actually sampled from the reservoir under study. In more detail, let S be the categorical variable (depositional facies) to be simulated, and s_k, k=1…K, the K different states (facies types) that the variable S can take. MPS simulation is a pixel-based technique that proceeds sequentially: all simulation grid nodes are visited only once along a random path and simulated node values become conditioning data for cells visited later in the sequence. At each unsampled node $\mathbf{u}$, let d_n be the data event consisting of the n closest conditioning data $S(\mathbf{u_1})=s(\mathbf{u_1})\ldots S(\mathbf{u_n})=s(\mathbf{u_n})$, which may be original sample data or previously simulated node values. The probability that the node $\mathbf{u}$ be in state s_k given d_n is estimated using Bayes' relation:

$$\text{Prob}\{S(\mathbf{u})=s_k \mid d_n\}=\frac{\text{Prob}\{S(\mathbf{u})=s_k \text{ and } d_n\}}{\text{Prob}\{d_n\}}$$

Prob$\{S(\mathbf{u})=s_k$ and $d_n\}$ and Prob$\{d_n\}$ are multiple-point statistics moments that can be inferred from the training image:

1. Prob$\{d_n\}$=c(d_n)/N_{TI}, where N_{TI} is the size of the training image, and c(d_n) is the number of replicates of the conditioning data event d_n that can be found in the training image. By replicates, we mean training data events that have the same geometrical configuration and the same data values as d_n.
2. Prob$\{S(\mathbf{u})=s_k$ and $d_n\}$=c$_k$(d_n)/N_{TI}, where c$_k$(d_n) is the number of training replicates, among the c(d_n) previous ones, associated to a central value $S(\mathbf{u})$ in state s_k.

The conditional probability of occurrence of state s_k at location **u** is then identified as the proportion of state s_k obtained from the central values of the training d_n -replicates:

Prob$\{S(\mathbf{u})=s_k \mid d_n\}$=c$_k$($d_n$)/c($d_n$)   (1)

The original MPS simulation implementation proposed by Guardiano and Srivastava was extremely cpu-time demanding since, at each node **u** to be simulated, the whole training image had to be scanned anew to search for training replicates of the local conditioning data event. Strebelle (2000) proposed decreasing the cpu-time by storing ahead of time all conditional probability distributions that can be actually inferred from the training image in a dynamic data structure called a search tree. More precisely, given a conditioning data search window *W*, which may be a search ellipsoid defined using GSLIB conventions (Deutsch and Journel, 1998), τ_N denotes the data template (geometric configuration) consisting of the *N* vectors $\{\mathbf{h}_\alpha, \alpha=1\ldots N\}$ corresponding to the *N* relative grid node locations included within *W*. Prior to the simulation, the training image is scanned with τ_N, and the numbers of occurrences of all the training data events associated with τ_N are stored in the search tree. During the simulation, at each unsampled node **u**, τ_N is used to identify the conditioning data located in the search neighborhood *W* centered on **u**. d_n denoting the data event consisting of the *n* conditioning data found in *W* (original sample data or previously simulated values, $n \le N$), the local probability distribution conditioned to d_n is retrieved directly from the above search tree; the training image need not be scanned anew. Furthermore, to decrease the memory used to build the search tree and the cpu-time needed to retrieve conditional probabilities from it, a multiple-grid approach was implemented that consists of simulating a series of nested and increasingly-finer grids, and rescaling the data template τ_N proportionally to the node spacing within the grid being simulated (Tran, 1994; Strebelle, 2000). That multiple-grid approach enables the reproduction of the large-scale structures of the training image while keeping the size of the data template τ_N reasonably small ($N \le 100$).

The MPS simulation program **snesim** (Strebelle, 2000) is applied to the modeling of a 2D horizontal section of a fluvial reservoir. The training image depicts the prior conceptual geometry of the sinuous sand channels expected to be present in the subsurface (Figure 1a). The size of that image is 250*250=62,500 pixels, and the channel proportion is 27.7%. An isotropic 40-data template was used to build the search trees for each of the four nested grids considered in the multiple-grid simulation approach. The unconditional MPS model generated by **snesim** reproduces reasonably well the patterns displayed by the training image, although some channel disconnections can be observed (Figure 1b).

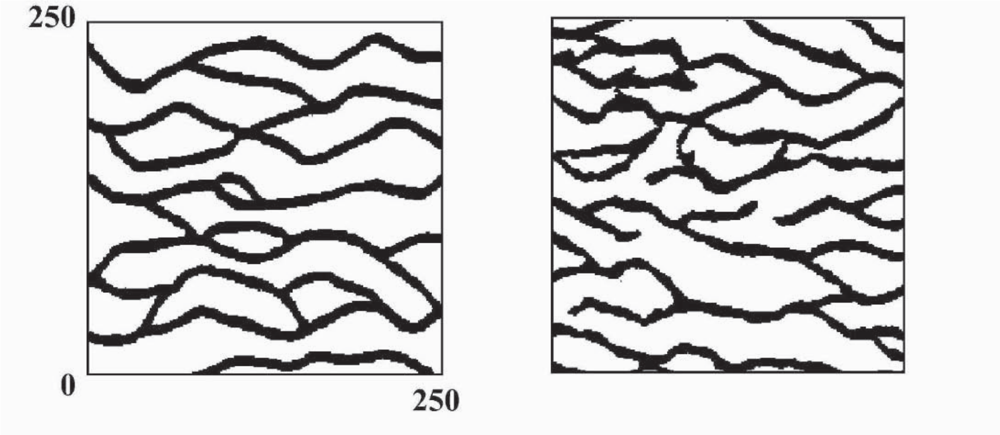

Figure 1. a) Training image used for the simulation of a 2D horizontal section of a fluvial reservoir (left); b) Corresponding MPS model generated by **snesim** (right)

If channel disconnections were believed to occur in the reservoir being modeled, due for example to some faults, the training image should display such disconnections. In the case above, the channel disconnections are not consistent with the information carried by the training image, thus they should be treated as "anomalies" that need to be corrected. In this paper we analyze why "anomalies" appear in MPS models. Then we propose modifying the **snesim** algorithm to decrease the number of anomalies, and we introduce a post-processing technique to remove those remaining.

2 Multiple-point statistics inference limitation

Inferring from the training image the probability conditional to a data event d_n requires finding at least one occurrence of d_n in that training image. However the likelihood that not a single exact replicate of the data event d_n can be found increases dramatically with the size n of that data event. Indeed, for an attribute S taking K possible states, the total number of possible data events associated with a given n-data template τ_n is K^n (for n=50 and K=2, $K^n \approx 10^{15}$!), while the total number of data events associated with τ_n scanned in the training image is always necessarily smaller than the size N_{TI} of that training image (typically less than a few millions nodes).

When no occurrence of a conditioning data event d_n is present in a training image, the solution proposed by Guardiano and Srivastava (1993), and implemented in the original **snesim** program, consists of dropping the farthest away conditioning data until at least one training replicate of the resulting smaller conditioning data event can be found.

However, *n'* being the number of conditioning data actually used to estimate the conditional probability distribution at the unsampled node **u**, critical information, in particular large-scale information, may be ignored when dropping the *(n-n')* farthest away conditioning data. Such approximation may lead to the inaccurate estimation of some conditional facies probability distributions, and the subsequent simulation of facies values that may not be consistent with the information carried by the dropped conditioning data.

To illustrate the above explanation for the presence of anomalies in MPS models, we plotted in Figure 2b the locations of the nodes that were simulated using less than 10 conditioning data in the MPS model created in the previous section (Figure 1b, repeated in Figure 2a). As expected, a close correspondence can be observed between the locations of those poorly-conditioned nodes and the channel disconnection occurrences.

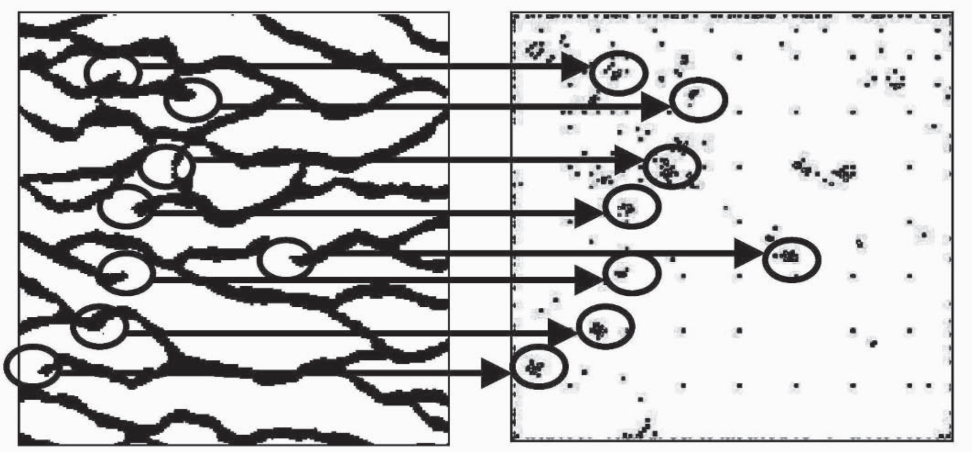

Figure 2. a) MPS model of a 2D horizontal section of a fluvial reservoir generated using **snesim** (left); b) Locations of the nodes simulated using less than 10 conditioning data (right). The arrows show the correspondence between the main clusters of poorly-conditioned nodes and the channel disconnections observed in the MPS model.

In the next section we propose modifying the original **snesim** implementation to decrease the number of dropped conditioning data.

3 Enhanced method to infer conditional probabilities

In the original **snesim** program, the facies probability distribution conditional to a data event $d_n = \{S(\mathbf{u_1})=s(\mathbf{u_1})\ldots\ S(\mathbf{u_n})=s(\mathbf{u_n})\}$ (or $d_n = \{s(\mathbf{u_1})\ldots s(\mathbf{u_n})\}$ to simplify the notations), is estimated using the following process:

- Retrieve from the search tree the number c(d_n) of d_n-replicates that can be found in the training image.
- If c(d_n) is greater or equal to 1, identify the conditional facies probabilities as the facies proportions of type (1). Otherwise drop the farthest away conditioning datum, reducing the number of conditioning data to (n-1).

Retrieve again from the search tree the number of training replicates of that lesser data event $d_{n-1}=\{s(\mathbf{u_1})\ldots s(\mathbf{u_{n-1}})\}$, and so on… until at least one replicate of the sub-data event $d_n=\{s(\mathbf{u_1})\ldots s(\mathbf{u_{n'}})\}$ ($n'\leq n$) can be found in the training image.

- If the number of conditioning data drops to 1, and still no training replicate of d_1 can be found, then the conditional facies probabilities are identified as the target marginal facies proportions of the simulation.

Instead of starting from the full data event d_n, and dropping conditioning data, one can obtain the exact same result using a reverse process, starting from the smallest possible sub-data event d_1, and adding conditioning data until the corresponding conditional probability distribution (cpdf) cannot be inferred anymore from the training image. In more detail, that inverse process consists of the following steps:

- Retrieve from the search tree the number of replicates of the sub-data event $d_1=\{s(\mathbf{u_1})\}$, consisting of a single conditioning datum, that closest to the node **u** to be simulated.
- If no training replicate of d_1 can be found, the local conditional facies probabilities are identified as the target marginal facies proportions of the simulation. Otherwise retrieve again from the search tree the number of replicates of the larger sub-data event $d_2=\{s(\mathbf{u_1}),s(\mathbf{u_2})\}$, consisting of the two conditioning data closest to **u**.
- If no training replicate of d_2 can be found, the probability distribution conditional to d_1 is used to simulate **u**. Otherwise retrieve from the search tree the number of replicates of the larger sub-data event d_3 consisting of the three conditioning data closest to **u**, and so on… until at least one replicate of the sub-data event $d_{n'}$ ($n'\leq n$) can be found in the training image, but no replicate of $d_{n'+1}$.

Because the dropped conditioning data may carry critical information, especially information about large-scale training structures, we propose extending the previous reverse process to retain additional conditioning data beyond $s(\mathbf{u_{n'}})$:

- Given that no replicate of $d_{n'+1}=\{s(\mathbf{u_1})\ldots s(\mathbf{u_{n'+1}})\}$ can be found in the training image, drop $s(\mathbf{u_{n'+1}})$, but add the next conditioning datum $s(\mathbf{u_{n'+2}})$. Retrieve from the search tree the number of replicates of the resulting sub-data event $\{s(\mathbf{u_1})\ldots s(\mathbf{u_{n'}}),s(\mathbf{u_{n'+2}})\}$ (or $d_{n'+2}-\{s(\mathbf{u_{n'+1}})\}$)
- If no training replicate of the previous sub-data event can be found, drop $s(\mathbf{u_{n'+2}})$, otherwise keep that conditioning datum. In both cases, consider the resulting data event, and add the next conditioning datum $s(\mathbf{u_{n'+3}})$, and so on… until the last conditioning datum $s(\mathbf{u_n})$ is reached.

This new conditional probability distribution function (cpdf) estimation method enables the retention of more conditioning data, as confirmed by its application to the previous fluvial reservoir section: only 127 nodes were simulated using less than 10 conditioning data in the new MPS model (Figure 3d) versus 285 in the original one (Figure 2b, repeated in 3c). In particular, the additional conditioning data used are located farther away from the node to be simulated. Therefore large-scale information was better integrated in the new MPS model (Figure 3b) than in the original one (Figure 1b, repeated in Figure 3a), and a significant number of channel disconnections were removed. A post-processing technique is proposed in the next section to remove the remaining anomalies.

Note also that the new cpdf estimation method requires only a minor amount of additional cpu-time: generating a simulated realization using the new cpdf estimation method took 16.3 seconds versus 16.0 seconds for the original simulated realization on a 660MHz SGI Octane II.

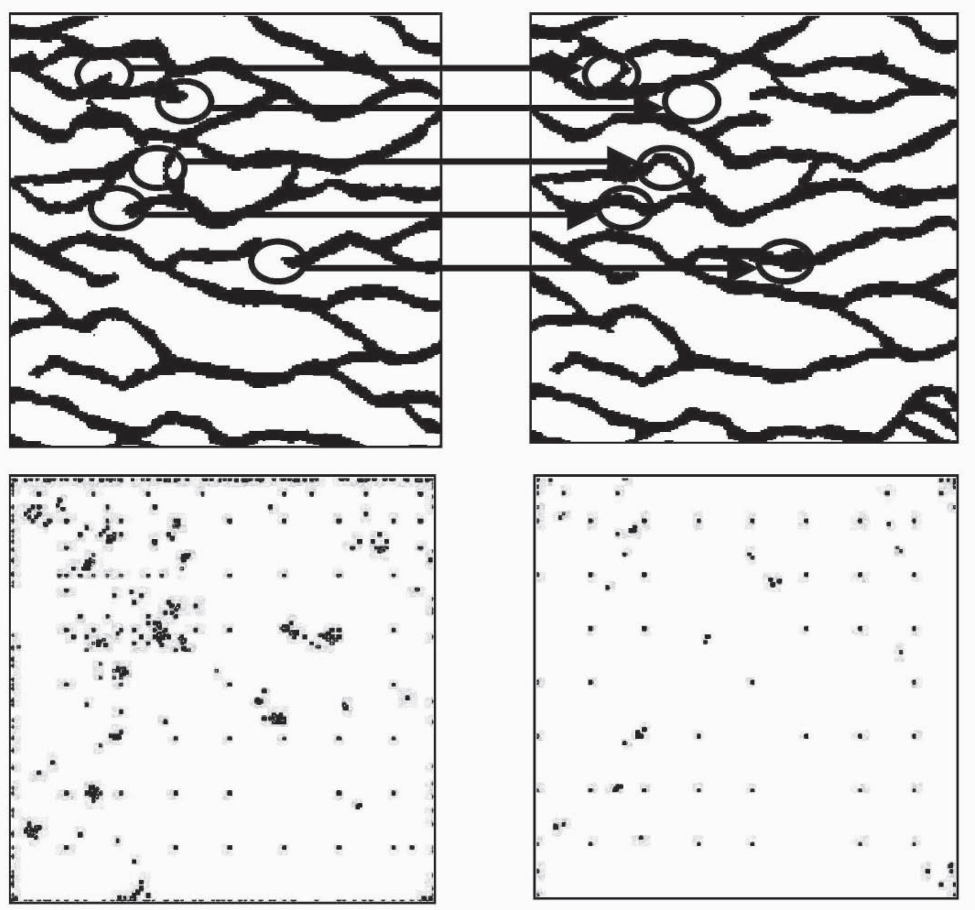

Figure 3. MPS models generated using a) the original cpdf estimation method (top left); b) the new cpdf estimation method (top right). Locations of the nodes simulated using less than 10 conditioning data c) in the original MPS model (bottom left); d) in the new MPS model (bottom right). The arrows show the channel disconnections that were removed from the original model.

4 A new post-processing algorithm

A post-processing algorithm was proposed by Remy (2001) to remove anomalies from MPS models using a two-step process:

- First, given a data template τ_N, identify in the simulated realization all data events associated with τ_N that do not occur in the training image
- Then re-simulate the grid nodes of those data events.

However, because of the limited size of the template τ_N, only small-scale anomalies could be corrected. Furthermore, a better identification of the nodes to be re-simulated is proposed in this section.

When estimating local cpdf's, conditioning data are dropped until at least one replicate of the resulting conditioning data event can be found in the training image. Considering a larger training image could provide additional possible patterns, thus decreasing the number of dropped conditioning data. But in many cases, a datum actually must be dropped because the information it carries is not consistent with the information carried by the other nearby conditioning data. Dropped conditioning data may be then a good indicator of the local presence of anomalies. This is confirmed by the good spatial correlation observed between the channel disconnections of the previous fluvial reservoir MPS model (Figure 3b, repeated in Figure 4a) and the clusters of nodes where conditioning data were dropped (Figure 4b).

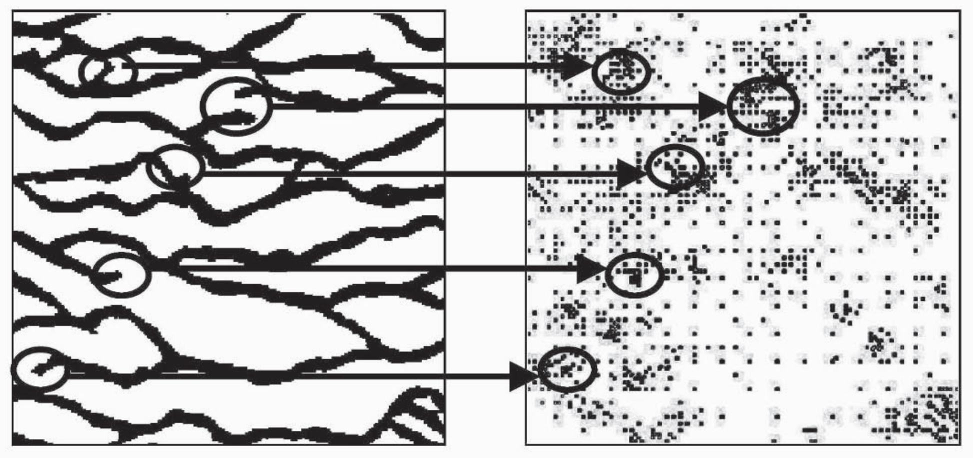

Figure 4. a) MPS model obtained using the new cpdf estimation method proposed in the previous section (left); b) locations of the conditioning data dropped (right). The arrows show the correspondence between channel disconnections and clusters of dropped conditioning data.

We propose marking the locations of the conditioning data dropped as the simulation progresses from one grid node to the other and revisiting these locations later. The new **snesim** implementation using that post-processing method proceeds in the following steps:

1. Define a random path visiting once all unsampled nodes.

2. At each unsampled node **u**, retrieve the local conditional probability distribution using the new cpdf estimation method previously described, and mark the nodes corresponding to the conditioning data dropped. Draw a simulated value for node **u**.
3. Move to the next node along the random path and repeat step 2.
4. Once all grid nodes have been visited and simulated, remove values from the nodes that have been marked, provided that they do not correspond to original sample data.
5. Repeat steps 1 to 4 several times until the generated image is deemed satisfactory according to some convergence criterion, for example until the number of nodes to be re-simulated stops decreasing.

To correct anomalies at all scales, this post-processing technique needs to be applied to every nested (fine or coarse) grid used in the multiple-grid simulation approach implemented in **snesim**. Figure 5 shows two post-processed MPS models of the previous fluvial reservoir. The number of channel discontinuities in both models is much lower than in the original model of Figure 4a.

Six post-processing iterations were performed on average per nested grid, and 43 nodes were re-simulated on average per iteration. The additional cpu-time required for the post-processing is relatively small: generating one realization using post-processing took 19.2 seconds on average, versus 16.3 seconds without post-processing.

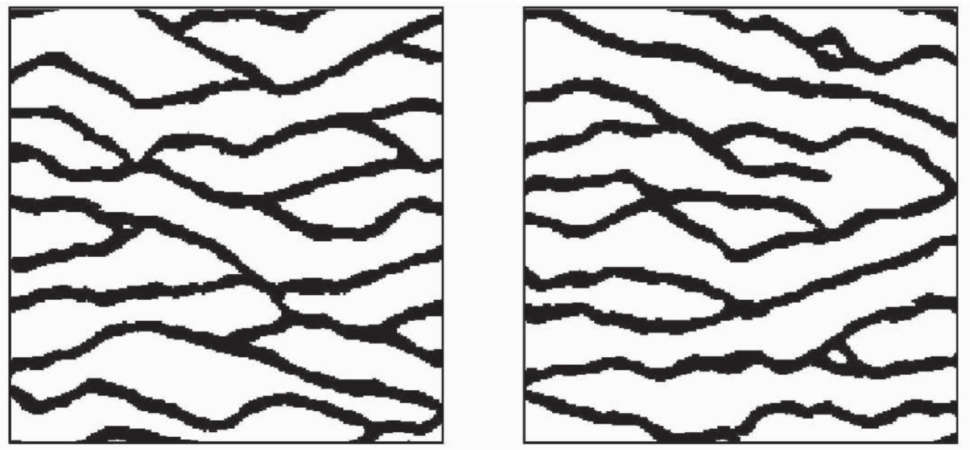

Figure 5. Two MPS models generated using the post-processing.

In 3D applications, the number of nodes to be re-simulated may be much higher than in 2D. Thus, instead of simply marking the nodes where conditioning data were dropped, we propose measuring the local consistency of the MPS model with regard to the training image by counting the number of times each simulated value, although being initially identified as a conditioning datum to simulate a nearby node, had to be ignored eventually to be able to infer from the training image the conditional probability distribution at that node. The post-processing consists then of re-simulating only those nodes where the previous consistency measure is greater than a given threshold.

5 Conclusion

A new version of the multiple-point statistics simulation program **snesim** with integrated post-processing is presented in this paper. In this new program, the method used to infer local conditional facies probability distributions is modified to increase the number of conditioning data actually used in that inference process. This new estimation method removes from multiple-point geostatistical models a great number of anomalies, i.e. simulated patterns that were not present in the training image. Then a post-processing technique is proposed to reduce the number of remaining anomalies. The application of that post-processing to a 2D horizontal section of a fluvial reservoir shows that the cpu-time needed to run the new modified **snesim** is comparable with that of the original **snesim**, while the number of anomalies decreases dramatically.

References

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

Guardiano, F. and Srivastava, R.M., Multivariate Geostatistics: Beyond Bivariate Moments, in Soares, A., editor, *Geostatistics-Troia*, vol. 1, p. 133-144. Kluwer Academic Publications, 1993.

Remy, N.. Multiple-point statistics for image post-processing, in *Report 14, Stanford Center for Reservoir Forecasting*, 2001.

Strebelle, S., *Sequential Simulation Drawing Structures from Training Images*, Ph.D. Thesis, Department of Geological and Environmental Sciences, Stanford University, 2000.

Tran, T., Improving Variogram Reproduction on Dense Simulation Grids, *Computers and Geosciences***,** vol. 20, no. 7, 1994, p. 1161-1168.

Viseur, S., Stochastic Boolean Simulation of Fluvial Deposits: a New Approach Combining Accuracy and Efficiency, paper SPE 56688 presented at the 1999 SPE Annual Technical Conference and Exhibition, Houston, Oct. 3-6, 1999.