Artifacts near Conditioning Data in MPS Gibbs Sampling

Steven Lyster and Clayton V. Deutsch

Centre for Computational Geostatistics
Department of Civil and Environmental Engineering
University Alberta

Introduction

Multiple-point statistics (MPS) methods are used to model geologic phenomena in more realistic and robust ways than traditional geostatistics. Rather than considering linear estimates based on individual conditioning points, MPS use several points simultaneously to determine conditional probabilities. These higher-order relations more accurately describe complex features than the standard Gaussian model of spatial structure. MPS methods have primarily focused on simulation of facies and geologic structures rather than continuous properties; this is probably due to the relative ease with which training images may be created for geologic structures, as opposed to the dense sampling necessary to infer high-order moments for continuous data.
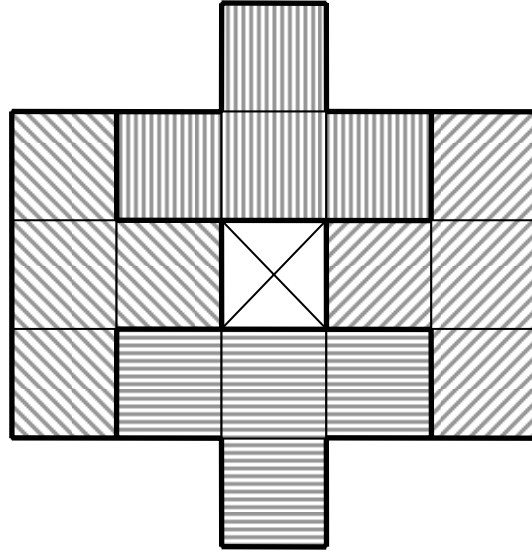
Several methods for reproducing MPS have been explored previously, some of which include the single normal equation (SNE) approach first proposed by Guardiano and Srivastava (1993) and further expanded by Strebelle and Journel (2000); simulated annealing (Deutsch, 1992); Gibbs sampling (Srivastava, 1992; Lyster et al, 2006); and a neural network iterative scheme (Caers and Journel, 1998). With the exception of the SNE methodology, all of these proposed algorithms are iterative systems. Also, these methods draw the necessary statistics from a training image (TI), deemed fully representative of the area in question.

The Gibbs sampler is a statistical algorithm which repeatedly resamples variables in an existing state to obtain an approximate sample from the joint distribution. This is advantageous in that it requires only the conditional distributions to fully define the joint distribution (Casella and George, 1992; Robert and Casella, 2004). When modeling geologic structures the joint distribution is the arrangement of the facies in the area of interest; each sample from the joint distribution is a possible realization used for characterizing the uncertainty. Using this idea it is possible to produce simulated realizations of resource deposits without having to characterize the full multivariate distribution; when the multivariate distribution is not Gaussian the full distribution is very difficult to characterize mathematically.

All iterative algorithms have drawbacks, some of which are time requirements, effects of starting images, and artifact or discontinuities appearing at conditioning data. The latter will be discussed in this paper, and how these artifacts develop in a Gibbs sampler algorithm.

Gibbs Sampler Algorithm Using MPS

The algorithm discussed in this paper is described in Lyster et al, 2006. In a Gibbs sampler framework, an initial image (or state) is iteratively perturbed, drawing from the conditional distributions. As the full conditionals are not explicitly known they must be estimated; this is accomplished through the use of multiple-point events (MPEs). Figure 1 shows an example of four four-point MPEs. Dividing the MPS template into several pieces in this way is meant to capture the structure of the TI while avoiding problems caused by small-scale randomness present in the initial images used for simulation.

Figure 1: Example of four multiple-point events with four points each.

To estimate the conditional probabilities of each facies at a given location, the following equation is used:

$$P^*(k) = \sum_{i=1}^{M} \lambda_i^k \cdot [I(E_i) - P(E_i)] + P(k) \tag{1}$$

Where:

$P^*(k)$ is the estimated probability of facies k at the current point;

λ_i^k is the weight given to MPE i for the probability of k;

E_i is one of the MPEs near the current point;

$I(E_i)$ is the indicator of the MPE; it is 1 for all events considered;

$P(E_i)$ is the global probability of E_i, also denoted as f_i;

$P(k)$ is the global probability of k, also denoted as P_k

Minimizing the variance of this estimate leads to the set of equations

$$\sum_{j=1}^{M} \lambda_j^k \cdot Cov\{E_i, E_j\} = Cov\{E_i, k\} \quad i = 1, \ldots, M \tag{2}$$

with the multiple-point covariance defined as

$$Cov\{E_i, E_j\} = E\{[I(E_i) - P(E_i)] \cdot [I(E_j) - P(E_j)]\} = P(E_i \cap E_j) - P(E_i) \cdot P(E_j) \tag{3}$$

Using the above equations, a conditional probability for each facies is calculated and then a new value for the location is drawn.

There are a few key modifications that have been made to the algorithm for this paper. The first is the addition of a "noise reduction factor" (NRF) to help control the randomness in realizations; because the global P_k values are included in Equation 1, facies with high global proportions may be drawn even in areas where they should be unlikely to appear. The NRF is a user-specified parameter; if the equation
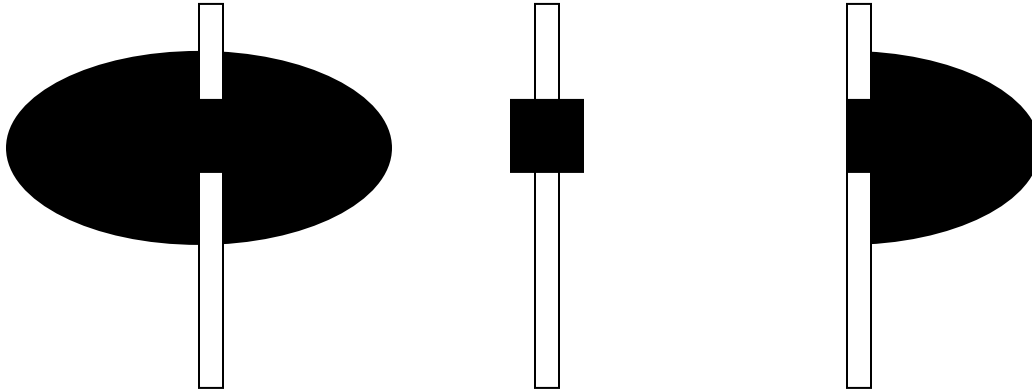
$$\frac{P^*(k)}{P(k)} < NRF \qquad (4)$$

is satisfied, then the probability of facies k is set to zero. This prevents facies with very small conditional probabilities from being drawn, speeding up convergence and reducing the randomness of the image. The effect of this on the convergence of the Gibbs sampler will not be discussed here.

Another modification to the algorithm is the random path followed; in the original algorithm the path started near the conditioning data and spiraled away. This type of path will be used, but in addition a fully random path will be examined as well.

Artifacts at Conditioning Data

Iterative simulation algorithms often have problems relating to discontinuities at hard conditioning data (Deutsch, 1992). These problems occur because the data are frozen in place, but the surrounding points have no such constraints. As the simulation proceeds, objects dilate and contract; these changes are not governed completely by the data and may in fact be caused more by the simulated nodes themselves. Figure 2 shows three examples of potential artifacts that may appear at strings of conditioning data.

Figure 2: Three artifact scenarios that may arise at hard conditioning data.

In Figure 2, the left image shows a case where the geo-object produced by the simulation (an ellipse) began to form around the black facies in the string of data; as the object formed, it grew to the size suggested by the statistics (and found in the TI) despite the fact that this did not honour the conditioning data properly. This kind of discontinuity will be referred to as Type I. The middle image in Figure 2 is nearly the opposite case: the conditioning data are only honoured by the points immediately adjacent, while the rest of the expected geo-object has contracted to almost nothing. This class of discontinuity will be referred to as Type II; note that this will also be used in the following discussions for cases where the geo-object has totally disappeared. The right-hand image is a mixture of the two cases. From the left side the geo-object has contracted to nothing, while on the right the black conditioning data formed the "seed" for a geo-object which has grown larger than suggested by the data. This hybrid discontinuity will be designated Type III for this paper. All three of these kinds of discontinuities are undesirable.

To explore the nature of artifact formation, a simple example study was undertaken. A simple TI, consisting only of ellipses and background facies, was created. Another image was created with the same structure and used as the true data, from which samples were taken. The TI and true data are shown in Figure 3. The images are both 400 x 200 pixels in size and have approximately 35% black (ellipse) pixels. Two vertical pseudo-well strings of data were taken from the true image at X coordinates 100 and 300, or one-quarter from the left and right edges of the image.

Because the Gibbs sampler algorithm is an iterative process, an initial image is needed for the simulation to proceed. Two cases were examined: an initial image created using the two-point statistics of the TI, and one created using only the univariate proportions with facies randomly selected at every point. These initial images are shown in Figure 4. To eliminate edge effects, the realizations created were 408 x 208 pixels with a four-pixel "buffer zone" being trimmed after the MPS simulation was finished. The initial images honour the pseudo-well data taken from the true data.

As mentioned above, both spiral and random paths were used to explore the impact of path selection on artifacts. The main drive for using a spiral path is to increase the range of influence of the conditioning data; however, this process may actually speed up the propagation of artifacts. Therefore, four total cases will be examined:

1. Two-point initial image with spiral path

2. Random initial image with spiral path

3. Two-point initial image with random path

4. Random initial image with random path

The results of these cases are shown in Figures 5 through 8. One hundred iterations (or loops) were performed for each case of the example. The states for the cases are shown in the figures after 10, 50, and 100 loops. All of the MPS simulations used six four-point events, with a noise reduction factor of 0.3.

Looking at Figures 5 through 8, it appears as though the random initial image simulations reproduce the features of the training image (and true data) better than those using two-point initial images; the images which had more long-range continuity initially ended up with too much continuity and curvilinearity. This relates to the artifacts in that cases 1 and 3 did not produce significantly different results; the choice of a spiral or random path appears to have little impact on the outcome of the simulation when the initial image already has substantial reproduction of some features and lower-order statistics. In those two cases, type I and III artifacts appear quite common after 50 and 100 loops. This is likely caused by the pre-existing geo-objects near the data being either dilated to the point of artifacting, or else contracted to nothing on the "smaller" side of the data (ie, the side with less of the object). It is quite notable that after 10 loops the artifacts do not appear to be prevalent. This supports the idea that the artifacts develop over time.

There are significant differences between cases 2 and 4. With a random initial image and a spiral path, the majority of geo-objects form clustered around the strings of data; when using a random path the geo-objects are formed gradually from the noise in the image and therefore do not show any preference to be near the data. Both case 2 and case 4 show very significant artifacting; however, the types of artifacts are different. Case 2, using a spiral path, has type I artifacts as the geo-objects formed almost symmetrically around the stings of data and overgrew what is suggested by the samples. Case 4 uses a random path and has a mixture of all three artifact types, with type III being most prevalent. This last case appears to best honour the spatial structure of the TI with the ellipses being randomly distributed about the field; however, where these ellipses formed did not seem to be controlled by the data.

Both spiral and random paths show artifacting, but from the examples shown here it seems as though a spiral path imparts no noticeable advantage. The one common characteristic of the artifact formation for all four cases is that the artifacts became progressively worse as the simulations proceeded. This may be the most important result, as proper stopping criteria in the algorithm could fix many artifacts before they begin.

Possible Solutions

As mentioned above, implementing good stopping criteria in the algorithm could prevent artifacting to some extent. However, this may not be possible in all cases; the simulation may not converge soon enough to allow stopping before artifacts form, or some discontinuities may appear very early and thus not be avoided by even the best stopping criteria.

One other possible solution is special weighting of the data points; this amounts to locally varying proportions near the data, which would hopefully avert discontinuities. However, too much weighting of the data could come close to being deterministic setting of the simulated values in the local neighbourhood

near samples. A slightly different approach would be to use two-point statistics to simulate the nearest unsampled points to the data; then, "freeze" these locations for the MPS Gibbs sampler portion of the simulation. This would result in the data being honoured explicitly and with no artifacts or discontinuities, with randomness between realizations as to the exact structure near the data. However, this could also cause artifacts at the edge of the frozen locations instead of at the data.
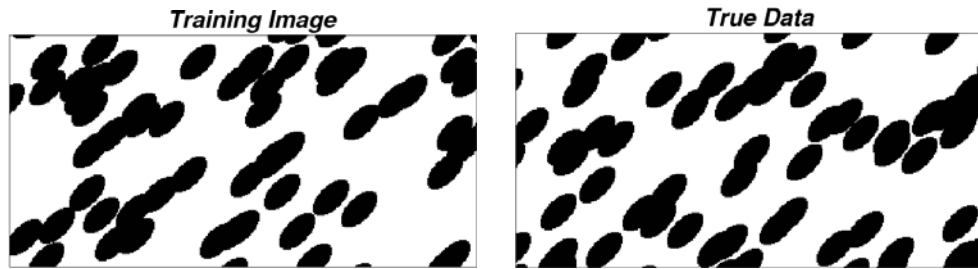
Another solution proposed is to metropolize the Gibbs sampler algorithm (as in Liu, 1995) so as to reject some undesirable changes and reduce the randomness of the simulation. Statistically, this amounts to a slower exploration of the solution space of the problem. This approach would hopefully prevent the edges of existing geo-objects from expanding or contracting due to random fluctuations once they have already formed in a consistent way with the data. The accept/reject rate could also be used as stopping criteria for the algorithm.
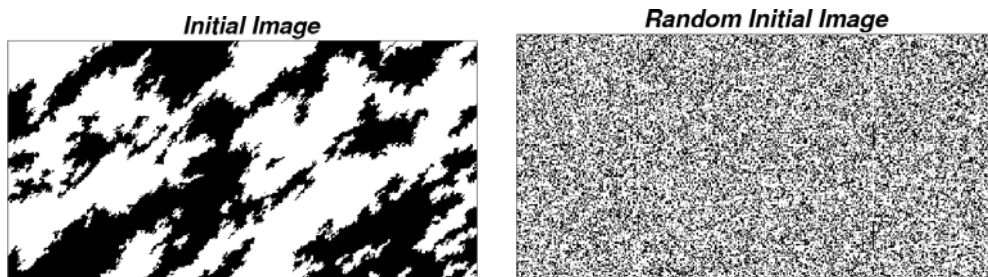
Future Work

The most likely fix to the artifact problems currently seen is a combination of the solutions suggested above, as well as other ideas as they arise. Finding and enforcing optimal stopping criteria will have a great impact on the progression of artifacts. Locally varying facies proportions and freezing of simulated nodes near data could also help. A hybrid approach utilizing all of these methods is probably the best approach, and will be explored.
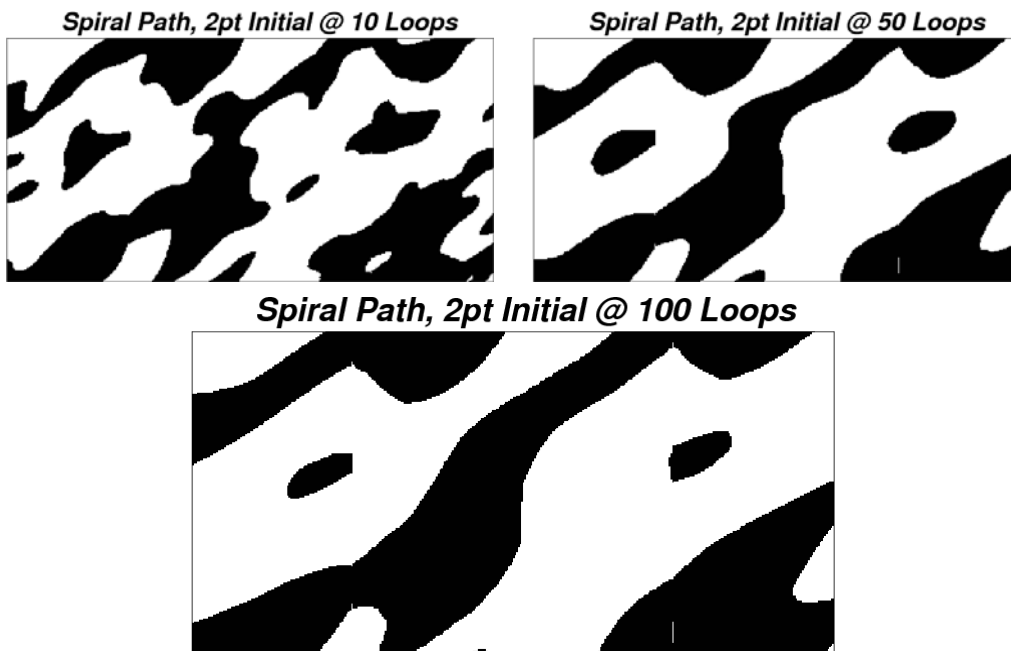
References

Boisvert, J.B. (2007) *Mineral Deposit Modeling with Pseudo-Genetically Constructed Training Images*. M.Sc. Thesis, University of Alberta, 72 p.

Caers, J. and Journel, A.G. (1998) Stochastic Reservoir Simulation Using Neural Networks Trained on Outcrop Data. *SPE Annual Technical Conference and Exhibition*, New Orleans, Oct. 1998, pp 321-336. SPE #49026.

Casella, G. and George, E.I. (1992) Explaining the Gibbs Sampler. *The American Statistician*, Vol. 46, No. 3, Aug. 1992, pp 167-174.

Deutsch, C.V. (1992) *Annealing Techniques Applied to Reservoir Modeling and the Integration of Geological and Engineering (Well Test) Data*. Ph.D. Thesis, Stanford University, 304 p.

Guardiano, F.B. and Srivastava, R.M. (1993) Multivariate Geostatistics: Beyond Bivariate Moments. Soares, A., Editor, *Geostatistics Troia '92*, Vol. 1, pp 133-144.

Liu, J. (1995) Metropolized Gibbs Sampler: An Improvement. Technical Report, Dept. of Statistics, Stanford University, 12 p.

Lyster, S. and Deutsch, C.V. (2006) TISIS: A Program to Perform Full Indicator Cosimulation Using a Training Image. *Centre for Computational Geostatistics*, No. 8, 16 p.

Lyster, S., Deutsch, C.V., and Dose, T. (2006) A New MPS Simulation Algorithm Based on Gibbs Sampling. *Centre for Computational Geostatistics*, No. 8, 26 p.

Robert, C.P. and Casella, G. (2004) *Monte Carlo Statistical Methods, 2nd Ed*. Springer Science and Business Media LLC, New York, 645 p.

Srivastava, M. (1992) Iterative Methods for Spatial Simulation. *Stanford Center for Reservoir Forecasting*, No. 5, 24 p.

Strebelle, S.B. and Journel, A.G. (2000) Sequential Simulation Drawing Structures From Training Images. Kleingeld, W.J. and Krige, D.G., Editors, *6th International Geostatistics Congress*, 12 p.
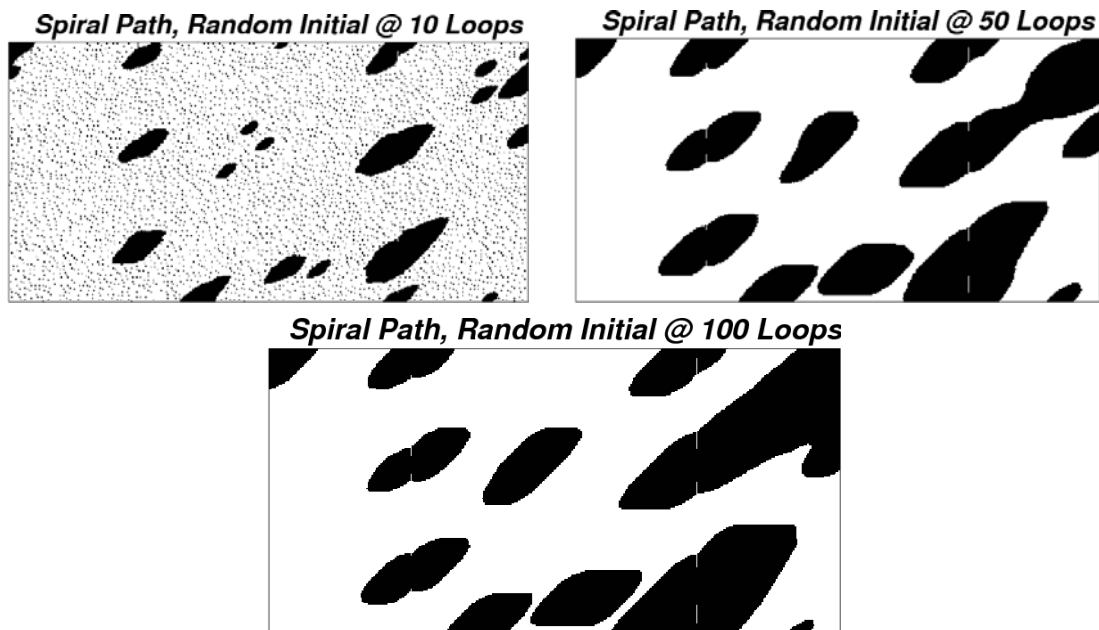
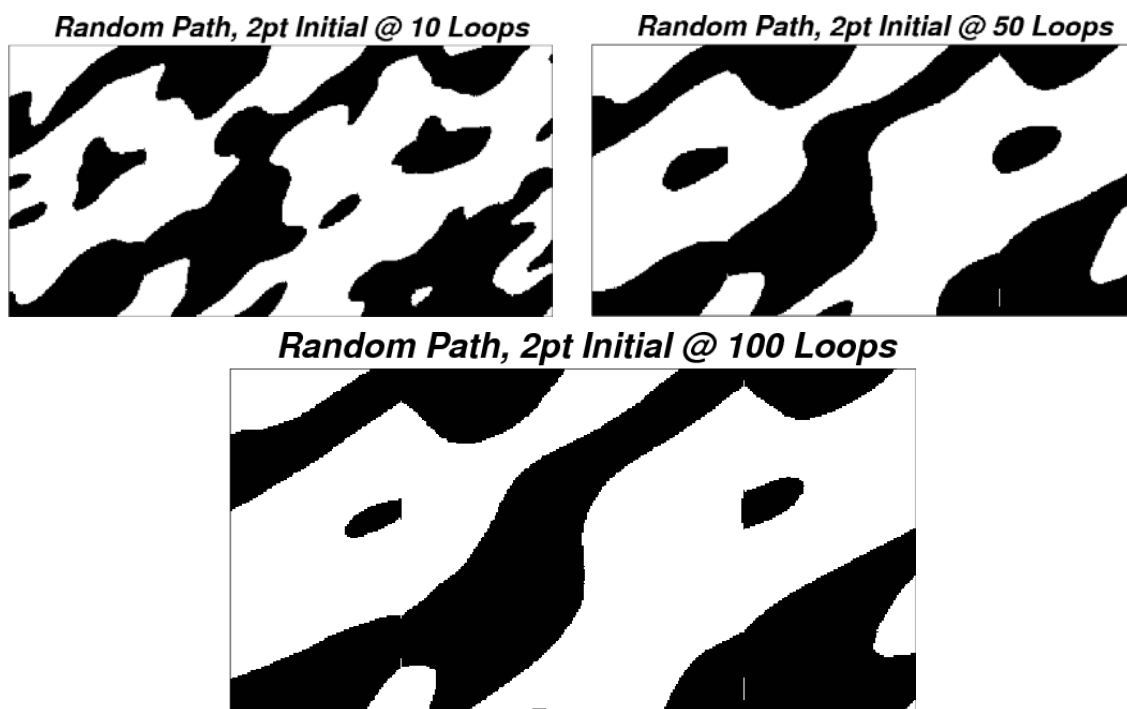**Figure 3:** Training image (top) and true data (bottom) used in the examples.

Figure 4: Initial image created using two-point statistics (top) and univariate proportions only (bottom).

Figure 5: Progression of the example using a spiral path and an initial image created with two-point statistics.
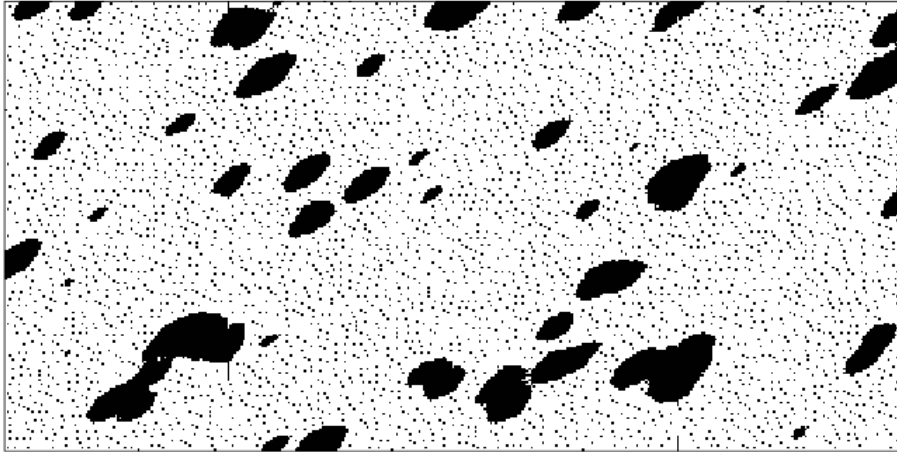
Figure 6: Progression of the example using a spiral path and an initial image created randomly with univariate statistics.

Figure 7: Progression of the example using a random path and an initial image created with two-point statistics.

Random Path, Random Initial @ 10 Loops

Random Path, Random Initial @ 50 Loops

Random Path, Random Initial @ 100 Loops

Figure 8: Progression of the example using a random path and an initial image created randomly with univariate statistics.