Characterization of High Order Correlation for Enhanced Indicator Simulation

Julián Ortiz C. (jmo1@ualberta.ca)
Department of Civil & Environmental Engineering, University of Alberta

Abstract

Geostatistical simulations aim to mimic the real variations of the underlying phenomena. Traditional simulation approaches only use two point statistics such as the variogram. In presence of non linear features or large range connectivity, such traditional methods do not provide good reproduction of those features. Connectivity of high and low values critical for most studies, e.g., permeability flow paths in a petroleum reservoir, grades in a mineral deposit, and concentrations of a pollutant in a environmental study.

The proposed methodology builds on the theory of runs used in statistics to characterize high order correlation in sequences of data. These "runs" are equivalent to high order indicator covariances and can be used in the context of extended normal equations or projection theory for estimation or simulation. Implementation of the extended normal equations using the information provided by runs is proposed for more realistic simulations. The methodology is discussed with exploratory examples.

The examples present runs above and below different quantiles to characterize high order correlation in pseudo-random series and in correlated series; maps of frequency of length of runs above quantiles and maps of differences in frequencies with respect to the random case were constructed to show the influence of correlation.

Introduction

Geostatistical techniques are used to simulate realizations of regionalized variables to help in decision making; however, the realizations must correctly reproduce important aspects of the true distribution. When the heterogeneities are not well reproduced in the resulting model or the uncertainty is inaccurate, incorrect predictions and wrong decisions may be made.

One feature of classical geostatistical simulation techniques, such as Gaussian and indicator approaches, is that they account only for two- point statistics through a covariance or variogram function. This limitation is mainly due to the difficult inference of higher order statistics [1, 3, 12, 20]. Simulation techniques would be improved with multiple-point statistics, since the resulting realizations would share more quantitative information than two-point statistics. More realistic numerical models likely lead to better decisions.

Several researchers have tried to incorporate multiple-point statistics in geostatistics, by using simulated annealing or training images [1, 8, 10, 20]. Most of them have failed in their applicability, because they are extremely CPU time consuming or because they require too many parameters to be set. Training images have been used to extract multiple-point statistics from outcrops, which supposedly represent the domain under study. The training

image may consist of a conceptual geological model of the site being characterized. We will never know how representative the training image is of the domain of interest.

The indicator approach is based on the use of the rank order of the data. Conditional probabilities are identified by the conditional expectations of the indicator transforms at particular thresholds. Kriging or the normal equations are used to estimate the conditional expectations as a function of the indicator values of the data and previously simulated locations at the same threshold being estimated. Cokriging can be used to consider the cross-correlations between indicators at different thresholds. The products of indicators would require inference of high order correlation functions. This would lead to the best *non*-linear unbiased estimator. As mentioned, high correlation is generally ignored, which reduces the information space and degrades the resulting estimation or simulation.

Some interesting results in number theory motivate us to study application to the indicator approach. This research aims to quantify high order correlation using the frequency of runs above and below different thresholds. Once high order correlation is known, the extended kriging equations may be applied in simulation with the consequent improvement on the result.

The applicability of this method is seen mainly (but not restricted to) using well data in petroleum applications and drillhole data in mining applications where the number of data is enough to calculate the expected lengths of runs. **Figure 1** presents a well with 22 composites (samples of equal length). The actual values are shown as a solid line, while the sample values are shown as black dots (they represent the average of the values in the sample, assuming no sampling and sample preparation errors). Given 5 thresholds, $z_i, i = 1, ..., 5$, a run of length l can be seen as the event of having l consecutive samples with grade higher than the threshold. In the example here presented runs are represented by thicker solid lines. For z_1, there is one run of length 16; for z_2, there is one run of length 13; for z_3, there are two runs, one of length 4 and the other of length 5; for z_4 there are three runs of lengths 2, 2, and 1 respectively; finally, for z_5 there are no runs in this example. The probability of having a run of length l is equivalent to the expected value of the product of l indicators corresponding to samples in a row or column. This high order moment is a multiple point statistic that would better characterize the true multivariate distribution or "spatial law".

Literature Review

This literature review focuses on the theory of runs, indicator formalism and extended normal equations.

Theory of Runs

The results presented here are based on a paper by A. M. Mood published in 1940 [16]. This work summarizes most of the work done previously by other authors, and can be considered as the basis for the majority of the subsequent statistical studies on runs (see for example [5, 6, 9, 17, 18, 21]). The author derives the "*distribution of runs of given length both from random arrangements of fixed numbers of elements of two or more kinds, and from binomial and multinomial populations*". He also gives the limiting form of these

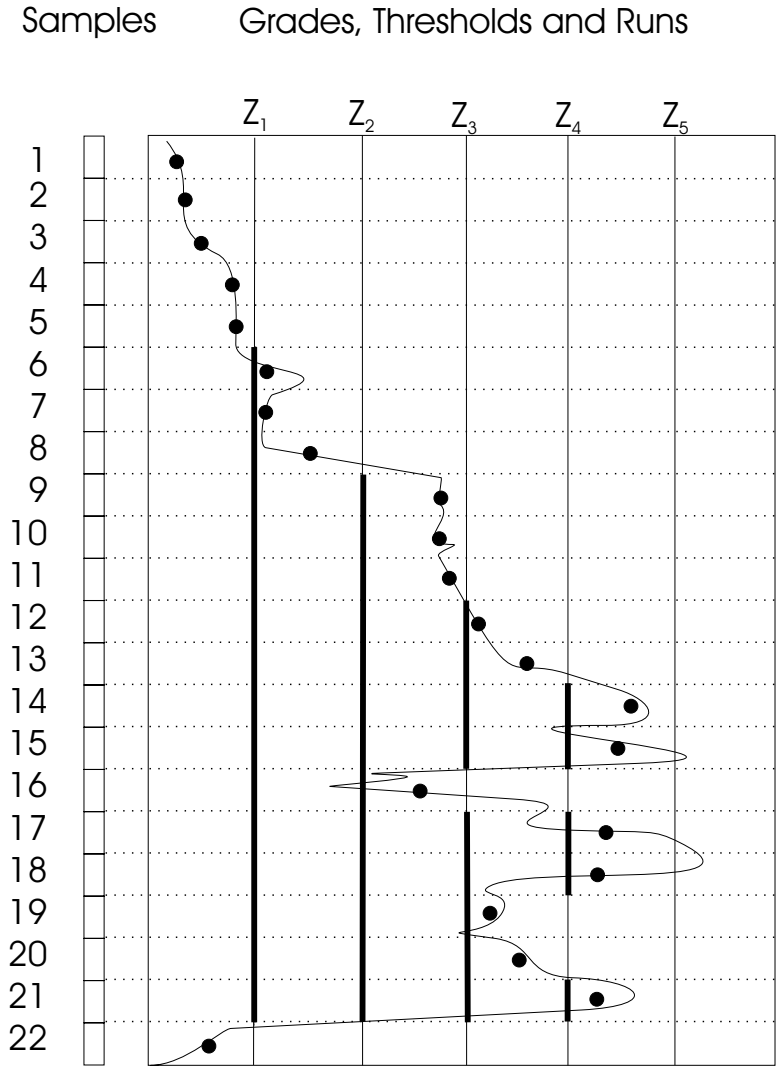

Figure 1: Example of calculation of runs in a drillhole with 22 samples. The solid line represents the actual grade, the black dots are the sample values. The runs are presented as thick solid lines under each threshold $z_i, i = 1, ..., 5$.

distributions as the sample size increases. Those limits distributions are all normal. The results are based on combinatorial analysis results, so independence between the elements is assumed at all times. Since those results are to be applied in the indicator framework, only the results for elements of two kinds are of interest for this research.

Lets first consider a sequence of uniform random numbers between α and β. We can set a threshold t and then rename each number with an a (above), if it is greater than t, or a b (below), if it is less or equal to t. Since the numbers are uniformly distributed, we can consider that a's are drawn from a Bernoulli distribution with probability $p = \frac{\beta - t}{\beta - \alpha}$. b's are drawn with probability $q = 1 - p$. Now that we have a sequence of a's and b's, we can count the length of strings of a's and b's. This is what we call runs.

For uniform random sequences, the distribution of runs of given lengths is known, so this property can be used to test pseudo-random number generators [2, 13, 14, 16, 19]. The special cases $(\alpha, \beta) = (0, \frac{1}{2})$ or $(\frac{1}{2}, 1)$ originated the so called tests of runs above and below the mean (or the median).

The following example shows how to calculate the runs for a sequence of uniform random numbers between 0 and 1. Consider the median as a threshold with the following sequence:

$$0.35,\ 0.56,\ 0.12,\ 0.11,\ 0.84,\ 0.76,\ 0.77,\ 0.45,\ 0.61,\ 0.51,\ ...$$

This sequence generates the following sequence of a's and b's:

$$b,\ a,\ b,\ b,\ a,\ a,\ a,\ b,\ a,\ a,\ ...$$

The sequence of lengths of runs above/below the median would be:

$$1,\ 1,\ 2,\ 3,\ 1,\ 2,...$$

The same procedure can be applied for thresholds other than the median. If we consider that there are n_1 a's and n_2 b's (and we define $n = n_1 + n_2$) then the proportions $p = \frac{n_1}{n}$ and $q = \frac{n_2}{n}$ of values above and below the threshold can be calculated. The total number of runs above and below t should follow a normal distribution with the following mean and variance [16]:

$$E[r] = 2 \cdot n \cdot p \cdot q$$

$$\sigma_r^2 = 4 \cdot n \cdot p \cdot q \cdot (1 - 3 \cdot p \cdot q)$$

When the threshold is the median (or the mean) of a uniform distribution then, the parameters are simply:

$$E[r] = \frac{n}{2} \qquad \sigma_r^2 = \frac{n}{4}$$

The number of runs above t of length i can be calculated as:

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

where the factorial $x^{(a)}$ corresponds to $x^{(a)} = x \cdot (x-1) \cdot (x-2) \cdot ... \cdot (x-a+1)$. The number of runs above t of length greater or equal to k, s_{1i}, can also be calculated:

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

Most of the moments of the distribution of runs for a random uniform case can be predicted. Analytical or approximate expressions for correlated sequences have to be found, whi (of particular interest is the multigaussian distribution.

Indicator Formalism

The non-parametric formalism of indicators was introduced in 1983 by A. G. Journel [11]. Many authors have presented this approach in great detail (e.g. see [4, 8]). This method avoids the need of a multigaussian assumption at the bivariate level and therefore the problem of maximum entropy implicit in that assumption. Nonetheless, notice that higher statistics (trivariate and higher level) are subject to the Central Limit Theorem, and therefore Gaussianity is implicit. Indicators characterize the random variable differently for high and low values. The Gaussian formalism assumes the behavior is symmetric with respect to the median.

The basic idea is to use a probability coding which takes into account the rank ordering of the data. For each threshold z_k, the data are coded using the following indicator function:

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

where $z(\mathbf{u}_\alpha)$ is the value at the data location $\mathbf{u}_\alpha$. There are now $n \cdot K$ indicator values with n data. This can be interpreted as a probability:

$$i(\mathbf{u}_\alpha; z_k) = Prob\{z(\mathbf{u}_\alpha) \leq z_k\} = F_{\mathbf{u}_\alpha}(z_k)$$

Notice that the constraint intervals and soft data can also be coded as indicators [4, 8]. We are mainly interested in the coding of hard data at this time.

The distribution of uncertainty of the regionalized variable can be inferred by kriging the indicator function at every threshold. Each one of the K sets of n indicator data can be used to estimate the value of the indicator at an unsampled location, i.e. the probability of having $z(\mathbf{u}) \leq z_k$. The indicators at different thresholds could be used as secondary variables for cokriging [4, 7, 8], but that is not commonly done because of the additional CPU and inference requirements.

Indicator simulation uses the conditional cumulative distribution function obtained through kriging for Monte Carlo simulation. It is important to emphasize that the conditional information considers both actual data and previously simulated values. In this way, the model covariance between all locations is reproduced.

Generalized Indicator Algorithm

As mentioned before, the kriging algorithm gives the best linear unbiased estimator; however there are non-linear components that are dropped, which degrades the result due to the reduction of the information space. A generalization of the indicator algorithm is presented to clarify where the present implementation of indicator kriging comes from and where it could be improved.

Consider N dependent events $A_j, j = 1, ..., N$. They can be sequentially simulated using the following expression, that comes from a repeated application of Bayes postulate:

$$P(A_j, j = 1, ..., N) = \begin{array}{l} P(A_N|A_j, j = 1, ..., N-1) \cdot P(A_{N-1}|A_j, j = 1, ..., N-2) \cdot \\ P(A_{N-2}|A_j, j = 1, ..., N-3) \cdot ... \cdot P(A_2|A_1) \cdot P(A_1) \end{array}$$

This relation is general and exact [11]. Two problems arise to implement this technique:

- Inference of the $(N-1)$ conditional probabilities $P(A_i|A_j, j = 1, ..., i-1), i = 2, ..., N$, and
- The size of the conditioning information increases from n to $n+N-1$, i.e. the kriging system or normal equations to be solved becomes very large.

Due to the difficult inference of those conditional probabilities, some approximations are often made to facilitate implementation of sequential indicator simulation.

The conditional probability $F(\mathbf{u}; z_k|(n))$ is the conditional expectation of an indicator random variable $I(\mathbf{u}; z_k)$ given the (n) data:

$$F(\mathbf{u}; z_k|(n)) = P(Z(\mathbf{u}) \leq z|(n)) = E\{I(\mathbf{u}; z_k)|(n)\}$$

$$\begin{array}{l} P\{Z(\mathbf{u}) \leq z_{k_0}|Z(\mathbf{u}_\alpha) = z_\alpha, \alpha \in (n)\} = \\ \quad E\{I(\mathbf{u}; z_{k_0})|I(\mathbf{u}_\alpha; z_k) = i(\mathbf{u}_\alpha; z_k), k = 1, ..., K; \alpha \in (n)\} \end{array}$$

where z_{k_0} is one of the k thresholds considered. The conditional probability $F(\mathbf{u}; z_k|(n))$ should be obtained by cokriging the unknown indicator using the $n \cdot K$ indicators; however, it is often assumed that the indicators for the same threshold z_{k_0} are more correlated with the indicator that is being estimated than the indicators for other thresholds, which is true in most cases since the calculation of cross-correlations is too demanding. Thus the probability is approximated by:

$$\begin{array}{l} P\{Z(\mathbf{u}) \leq z_{k_0}|Z(\mathbf{u}_\alpha) = z_\alpha, \alpha \in (n)\} = \\ \quad E\{I(\mathbf{u}; z_{k_0})|I(\mathbf{u}_\alpha; z_{k_0}) = i(\mathbf{u}_\alpha; z_{k_0}), \alpha \in (n)\} \end{array}$$

All cross-correlation between indicators at different thresholds and multiple point indicators are then ignored. This first approximation is done not because the inference of the cross-correlations is difficult, but because, in general, the improvement in the resulting simulation does not justify the increased work.

The conditional expectation can be written as a function of the conditioning information in the following manner:

$$\begin{array}{l} E\{I(\mathbf{u}; z_{k_0})|I(\mathbf{u}_\alpha; z_{k_0}) = i(\mathbf{u}_\alpha; z_{k_0}), \ \ \alpha \in (n)\} = \phi\{i(\mathbf{u}_\alpha; z_{k_0}), \alpha \in (n)\} \\ \quad = a_0 + \displaystyle\sum_{\alpha \in (n)} a_1(\alpha) \cdot i(\mathbf{u}_\alpha; z_{k_0}) \\ \quad + \displaystyle\sum_{\alpha \in (n)} \sum_{\alpha' \in (n), \alpha \neq \alpha'} a_2(\alpha, \alpha') \cdot i(\mathbf{u}_\alpha; z_{k_0}) \cdot i(\mathbf{u}_{\alpha'}; z_{k_0}) + ... \\ \quad + a_n \cdot \displaystyle\prod_{\alpha \in (n)} i(\mathbf{u}_\alpha; z_{k_0}) \end{array}$$

The classical application of indicators considers only the use of univariate and bivariate statistics; the reliable inference and positive definite modeling of higher order covariances is difficult. The first $(n+1)$ terms of the previous expansion are retained. The use of higher order statistics is possible through the use of extended normal equations [10]. This approach was originally based on multiple-point statistics inferred from training images. The use of runs to estimate high order covariances avoids the need for training images by inferring the multiple-point covariances directly.

Extended Normal Equations

The 2^n coefficients a_0, $a_1(\alpha)$, $a_2(\alpha, \alpha')$, ..., a_n in the last expression correspond to the full indicator kriging weights and can be determined by an extended system of 2^n normal equations, which imposes the orthogonality of the error vector to the 2^n combination of data events defined by the indicator functions [10, 15]. The applications reviewed are based on multiple-point statistics extracted from training images. This ensures the positive definiteness of the multiple-point covariances. One of the problems that will have to be solved in this research is the positive definiteness of the high order statistics based on runs.

Proposed Methodology

A number of steps are required. Exploratory examples will help understand the statistical basis of runs, as well as the details of the indicator simulation methods currently used. A multiple-point tool equivalent to the variogram or covariance for two-point statistics must be proposed that is understandable, easy to calculate and intuitive. This tool must possess some properties: positive definiteness must be ensured if we want to apply it in a kriging-type framework. With that multiple-point measure of correlation, the implementation of the algorithm to simulate regionalized variables including high order correlation must be developed. Several approaches are proposed. Practical applications should be presented to compare the results of the enhanced algorithm to current application. The cases when it is worth to apply the proposed methodology should be explored.

Exploratory examples

Some exploratory examples must be developed to have a deeper understanding of each topic discussed in the Literature Review.

- Current application of indicator techniques must be reviewed to explain and understand their limitations and possible areas of improvement.
- Examples of runs and a better understanding of their behavior with correlated sequences are required to judge their possible application as well as their limitations. Calculations of the frequency of runs of different lengths for different thresholds using sequences of random and correlated values are presented.

Parametrization or analytical expression of the distribution of runs

An approximate or analytical expression for the frequency of runs of a given length should be developed. A measure of departure from Gaussianity could be found that may be useful to quantify the improvement with respect to a multigaussian simulation technique.

A definition of a "cumulative run" may be required, since the sole use of runs when the number of data is limited may not be robust enough for inference. A cumulative run would consider that a run of length l is *also* 2 runs of length $l-1$, 3 runs of length $l-2$, and so on. In general, a run of length l corresponds also to c cumulative runs of length $l-c+1$, with $c = 1, ..., l$, as shown in **Figure 2**.

The cumulative runs should be robust and a parametrization (or in some cases an analytical expression) could be used as data reduction, so that with a few statistics (parameters of the distributions) we can characterize the complete high order correlation behavior of the variable under study.

Implementation of an enhanced indicator simulation algorithm

An enhanced indicator simulation algorithm that accounts for multiple- point statistics generated using the runs is required for practical applications. As mentioned before, well data is an obvious source of data on runs since wells provide strings of equidistant samples. The extension to three dimensions should be straightforward once the one dimensional high order correlation is well understood. Several issues must be addressed here:

- Positive definiteness of the multiple-point statistic used is required if a kriging-like procedure is applied.
- The simulated values should be drawn via Monte Carlo simulation from a conditional probability. This probability will have two different components: firstly, a usual two-point statistic (variogram or covariance) has to be used to impose the correlation between the sample data and the location being simulated; secondly, the correlation between multiple-point events and a single point (the location being simulated) has to be considered to inject the right connectivity or non linear characteristic to the simulated model.
- Order relations could be avoided by simulating the thresholds hierarchically. The lowest threshold would be simulated first, starting with a grid filled with zeros, that is, all values are initially set to be greater than the lowest threshold. Then, the nodes less than the highest threshold are simulated. The second highest threshold will have a restricted field; only the values that are less than the highest threshold will be visited during the simulation. It will continue in this fashion until the last threshold is simulated.

Exploratory Examples

In this section, the number of runs above and below thresholds are calculated for different correlated series. The results are compared with the theoretical distribution for uncorrelated sequences. Then, the frequencies of lengths of runs above thresholds are plotted in

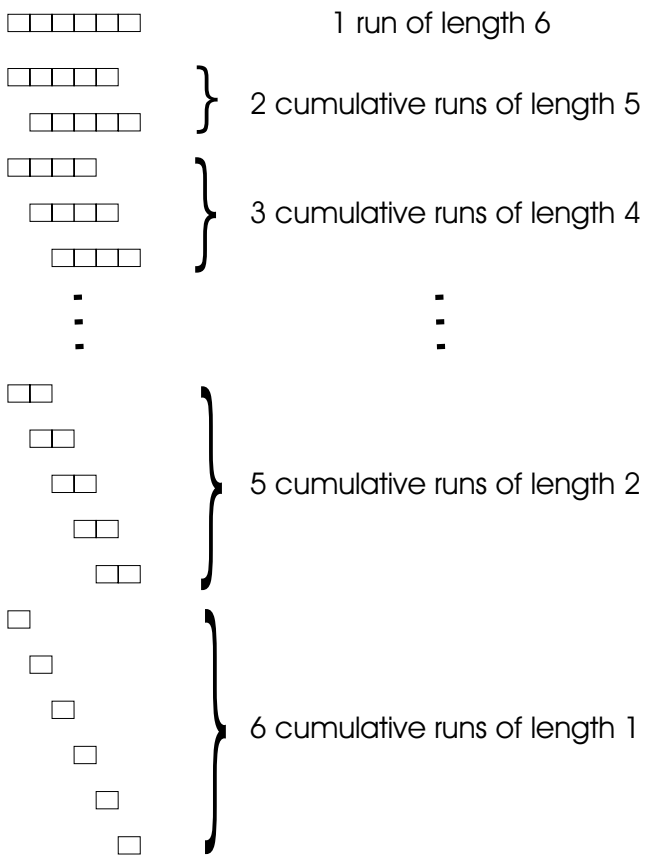

Figure 2: Example to illustrate the concept of "cumulative runs". One single run of length 6 corresponds to 2 cumulative runs of length 5, 3 cumulative runs of length 4,..., and 6 cumulative runs of length 1.

1,000 Sequences of 10,000 Random Numbers - `mcorn`				
Threshold	Theoretical Mean	Theoretical Std. Dev.	Observed Mean	Observed Std. Dev.
0.2	3200	57.69	3199	59.20
0.4	4800	51.85	4800	51.65
0.6	4800	51.85	4799	51.97
0.8	3200	57.69	3200	58.32

Table 1: Theoretical and Observed results - `mcorn`

a map, along with the curve of average length for a given threshold. Again, different two-point variogram functions are used in order to see the differences with the random case. Thresholds have been chosen as regularly spaced quantiles.

Finally, maps of differences between the observed frequencies of lengths of runs in correlated sequences and the expected frequencies for the random case were plotted, showing again different responses given different two-point variogram functions.

Distribution of Total Number of Runs Above and Below Thresholds

Using the pseudo-random number generator `mcorn`, 1,000 sequences of 10,000 uniform pseudo-random numbers were generated. The number of runs above and below 4 thresholds were counted and compared with the theoretical limit distribution.

Histograms showing the distribution of total number of runs above and below the corresponding thresholds are shown in **Figure 3**. The theoretical parameters of the distribution are summarized in **Table 1** and compared with the observed ones. Both the mean and the standard deviation are close to their theoretical values.

Comparison of Different Variogram Functions

Recall that the distribution of runs of elements above and below a threshold (i.e. assumed independently drawn from a Bernoulli distribution with probabilities p and $q = 1 - p$, respectively) are asymptotically normally distributed with the following parameters:

$$\mu = 2 \cdot n \cdot p \cdot q$$

$$\sigma^2 = 4 \cdot n \cdot p \cdot q \cdot (1 - 3 \cdot p \cdot q)$$

Two different random number generators were compared with the expected number of runs (for uncorrelated values). `mcorn` and `acorni` showed a very good reproduction of the theoretical mean, as presented in **Figure 4**. The standard deviation is not as smooth as the mean, but notice the good reproduction at extremes; this is common to all cases presented. The mean and standard deviation of the total number of runs above and below each threshold was calculated as an average over 100 sequences of 1000 values each.

Series of correlated data were generated using moving average simulation and simulated annealing. The first example considers a triangular variogram function (this variogram model is valid in one dimension only):

$$\gamma(\mathbf{h}) = \begin{cases} h, & \text{if } h \leq a \\ a, & \text{if } h > a \end{cases}$$

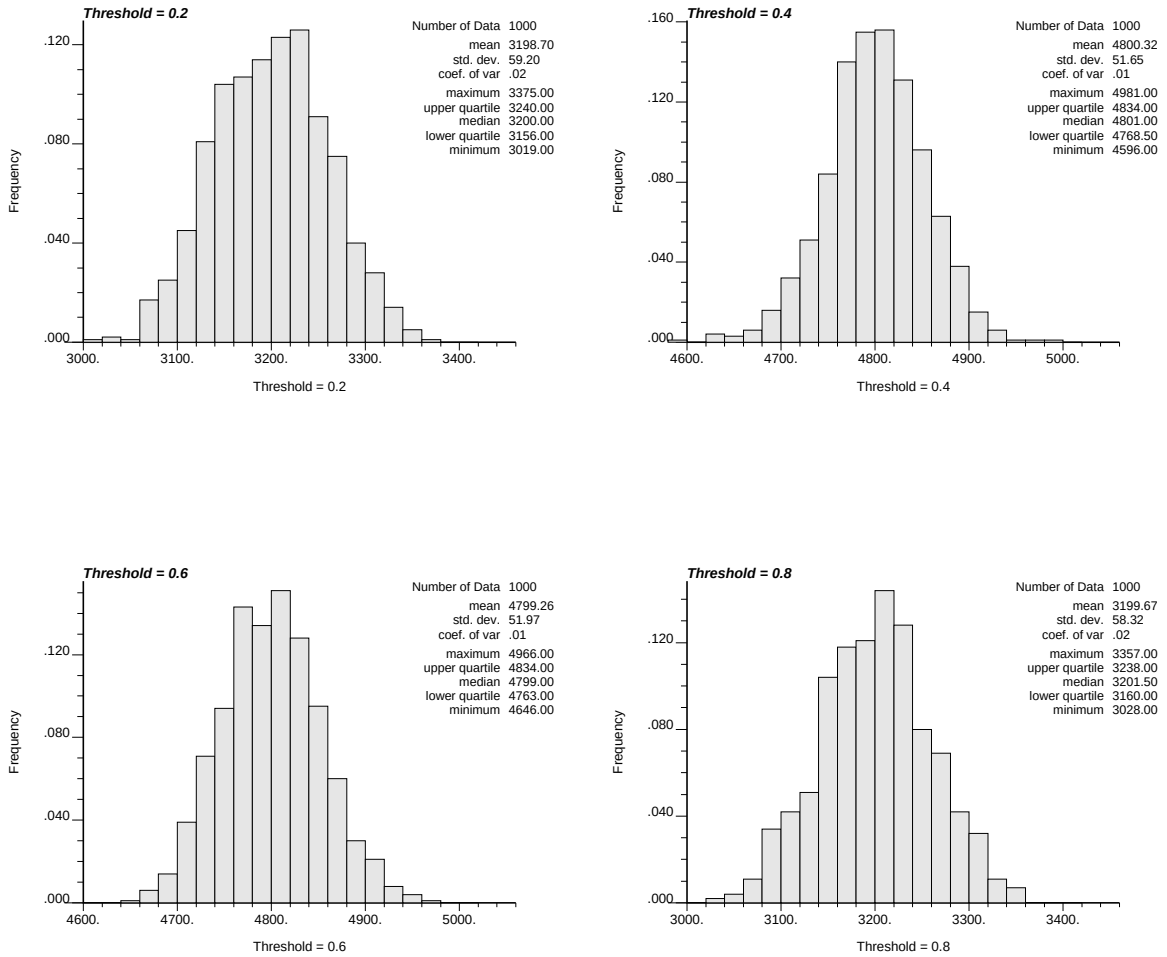

Figure 3: Histograms of total number of runs for different thresholds - 1,000 sequences generated with `mcorn`.

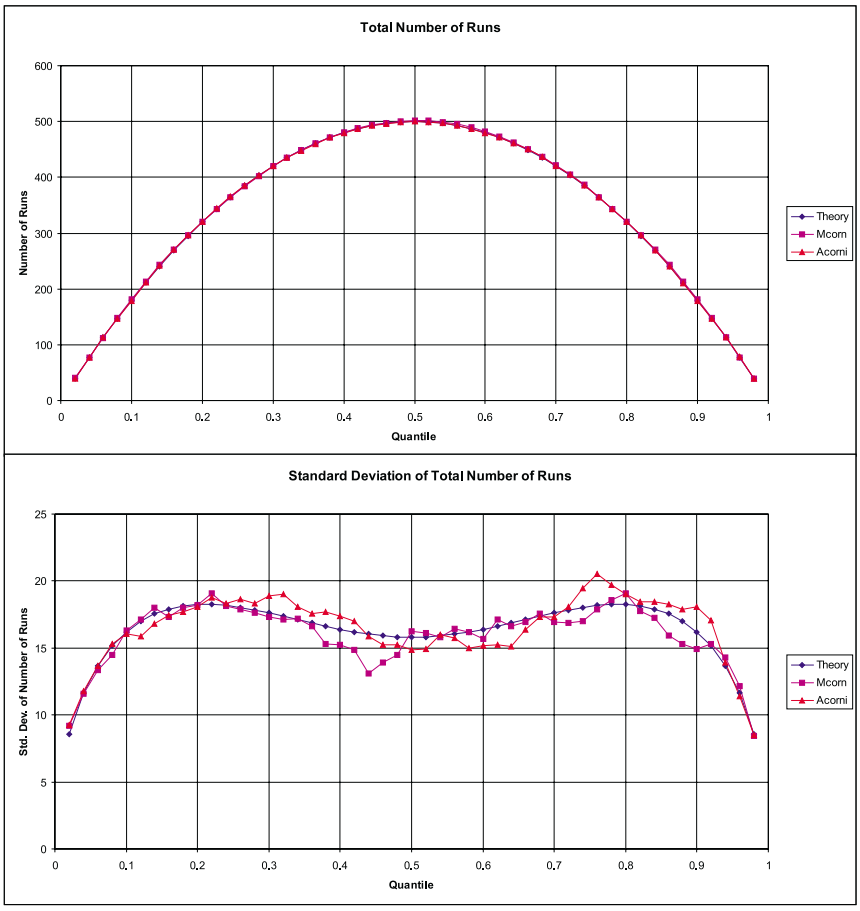

Figure 4: Mean and standard deviation of total number of runs above and below thresholds (quantiles) for `mcorn` and `acorni`, compared with the theoretical expected values.

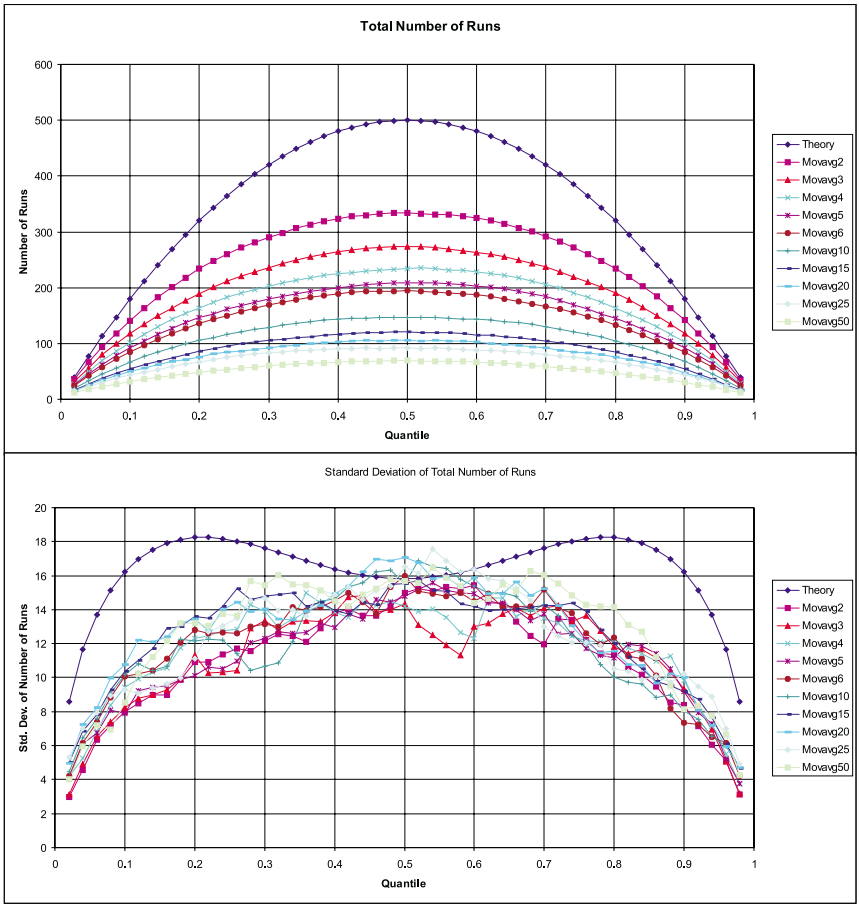

Figure 5: Mean and standard deviation of total number of runs above and below thresholds (quantiles) for sequences with a triangular variogram function generated using moving average, compared with the theoretical expected values for uncorrelated series.

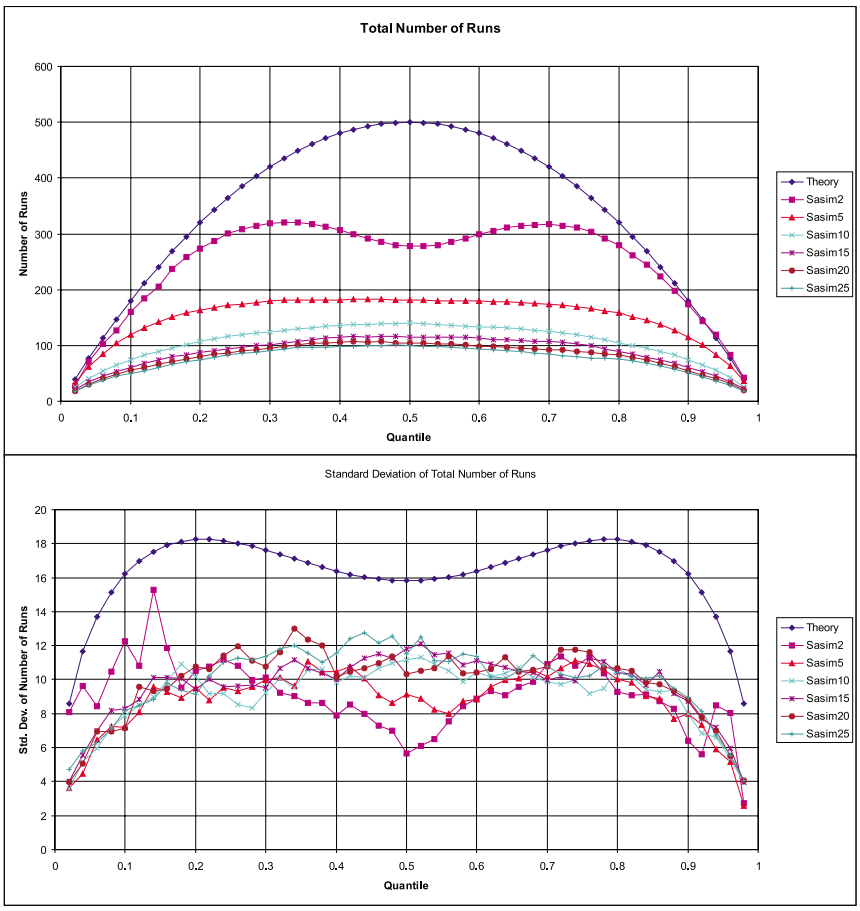

Figure 6: Mean and standard deviation of total number of runs above and below thresholds (quantiles) for sequences with a triangular variogram function generated using simulated annealing, compared with the theoretical expected values for uncorrelated series.

A variety of ranges were evaluated using sequences generated by moving average (**Figure 5**). The curves of mean and standard deviation of the total number of runs depart predictably from the uncorrelated case. When correlation increases, runs tend to be longer, so there are less than in the random case. For some ranges (5, 10, 15, 20, and 25 units) simulated annealing was used to generate correlated series. **Figure 6** gives the result for a triangular variogram function. Some different variogram models were explored with similar results: in all the cases, the mean number of runs decreases when the sequence has a greater correlation range.

Three different seed numbers were used with a fixed range (equal to 5 units). The results showed that there is no significant differences between the sequences generated with different seed numbers. Notice that for every sequence in those examples a different seed number was used.

In general, the curve of mean total number of runs is quite smooth and well behaved, however, the standard deviation is not as stable as the mean. Differences between moving average and simulated annealing is likely due to the random function implicit in each method. In the first case, gaussianity is derived from the averages and the Central Limit Theorem. In the case of simulated annealing, the random function is unknown. In order to visualize the differences between different variogram models and between the methods used to generate the sequences, **Figure 7** is presented comparing the result for a correlation range of 5 units. The theoretical result for uncorrelated sequences is plotted as a reference. The same comparison was done for other ranges. An interesting and consistent difference between the results given by moving average and simulated annealing is demonstrated here. In all the cases the moving average method (Gaussian) generates standard deviations closer to the random case than the simulated annealing technique. This situation can be explained by the maximum entropy property of the Gaussian model. Differences between the triangular, spherical and exponential variogram are due to the different correlation for a given distance, as presented in **Figure 8**.

Maps of Frequencies of Lengths Above each Threshold

In order to obtain a plot easily understandable and that clearly reflects changes in the high order behavior of the variable, a map of frequencies of lengths of runs above each threshold along with a curve showing the average length of runs above each threshold has been implemented.

Using sequences with ranges of 2, 5, 10, 15, 20, and 25 units, the number of runs above each threshold were calculated. The decision of using only the runs above (instead of runs above and below) was taken to avoid hiding differences in the continuity of high and low values.

Figure 9 shows the maps for random sequences generated with `acorni` and `mcorn`. **Figure 10** shows the maps for sequences generated using moving average with a triangular variogram model. In **Figure 11** sequences with the same variogram model were generated using simulated annealing.

In all the cases, when the range increases, the cloud of non zero frequencies grows to the right and up, because when the range of correlation is greater, long runs are more likely to be found.

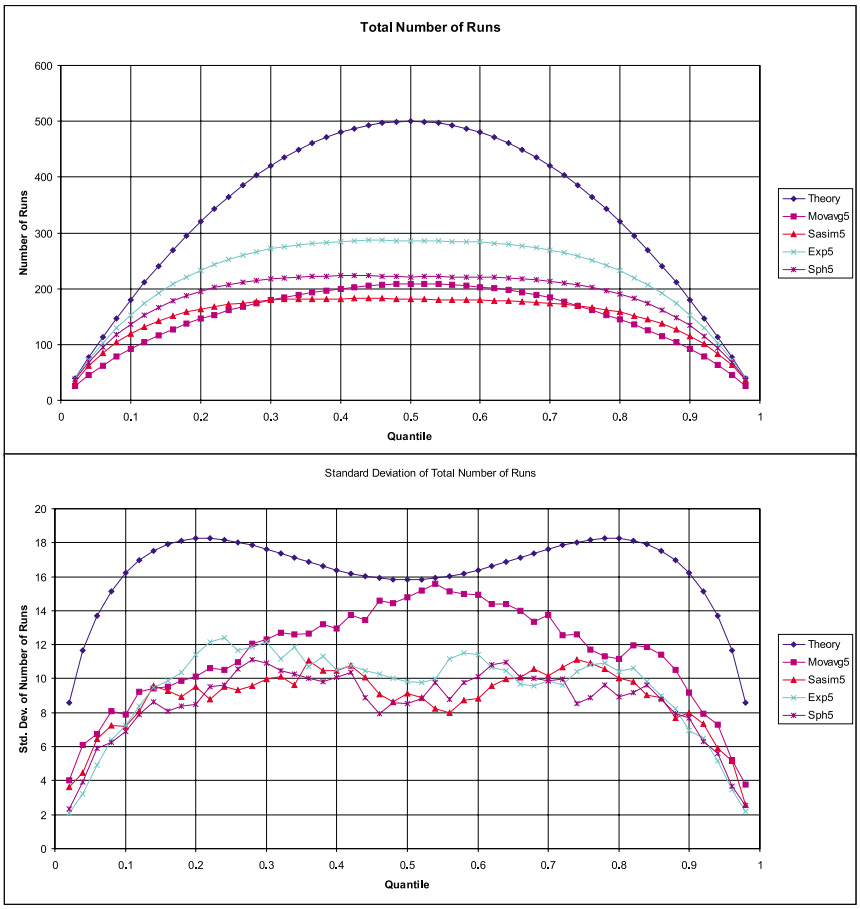

Figure 7: Mean and standard deviation of total number of runs above and below thresholds (quantiles) for sequences with a range of 5 and different variogram function (sasim and movavg have a triangular variogram).

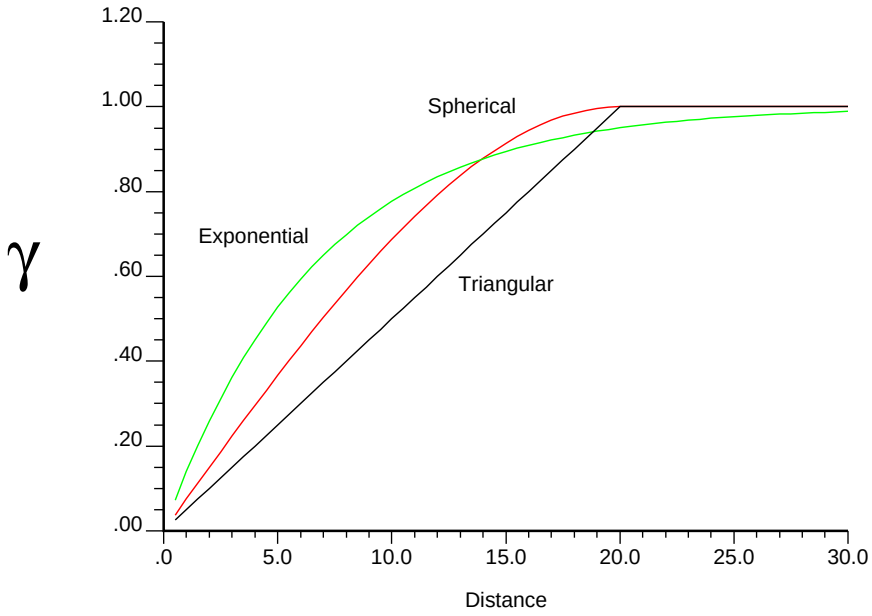

Figure 8: Variogram models used in the examples (relative shape for effective range of 20).

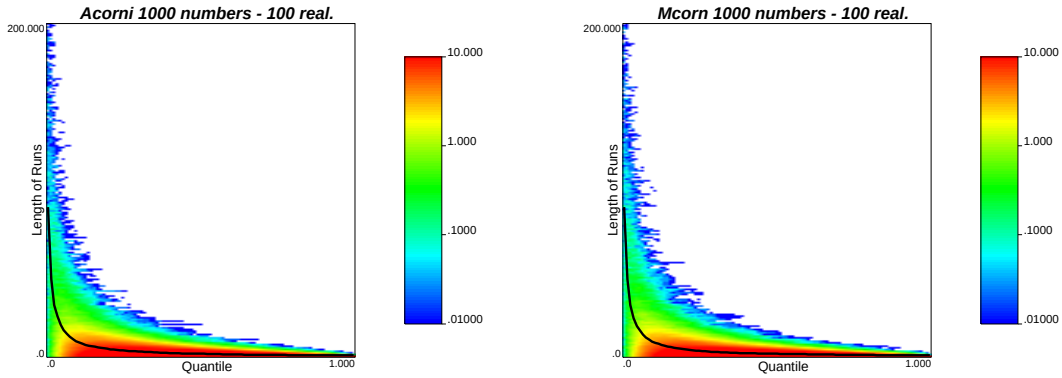

Figure 9: Map of frequency of lengths of runs above quantiles for sequences generated with `acorni` and `mcorn`. The solid line shows the average length as a function of the quantile.

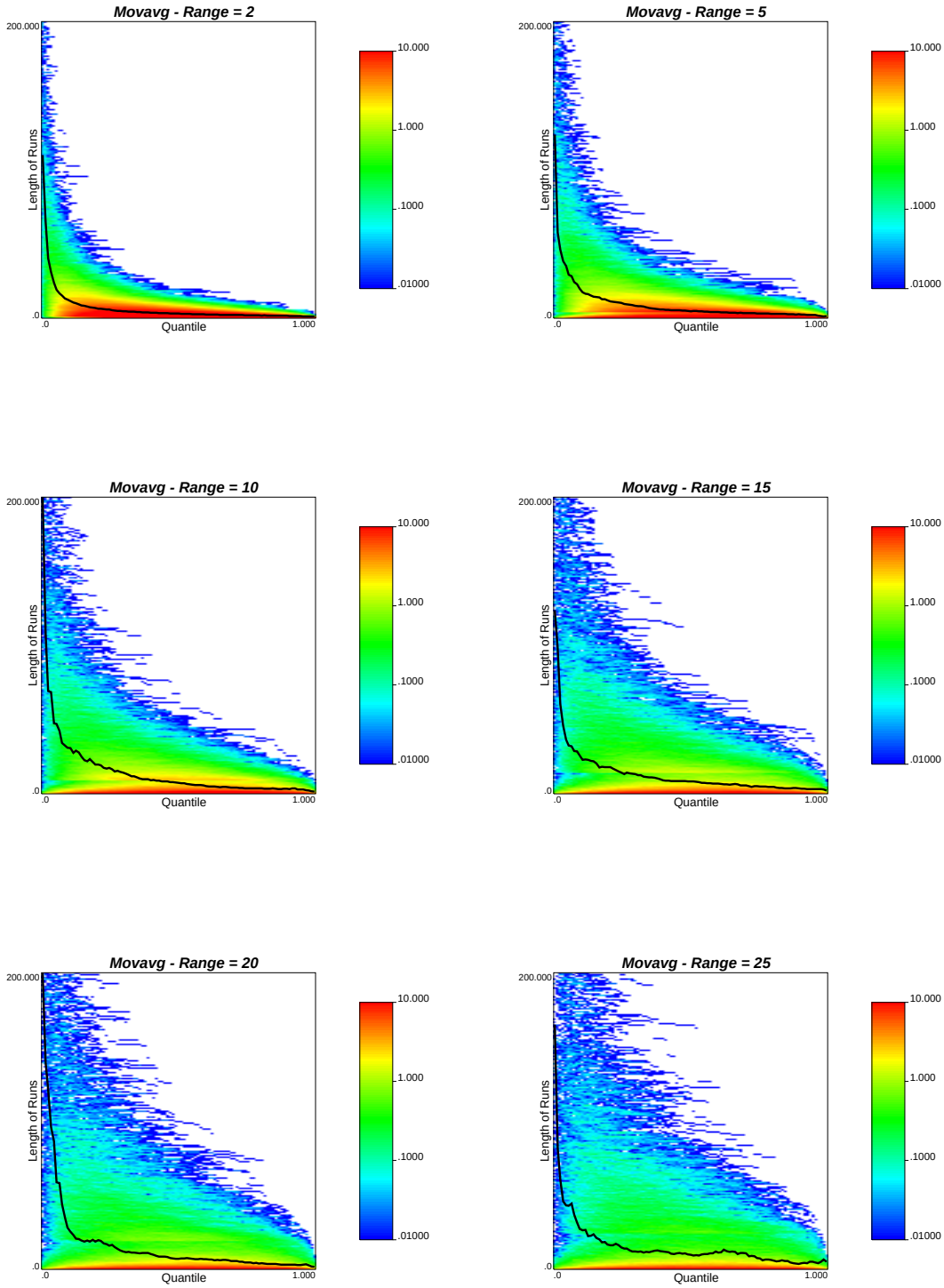

Figure 10: Map of frequency of lengths of runs above quantiles for sequences generated by moving average (triangular variogram model). The solid line shows the average length as a function of the quantile.

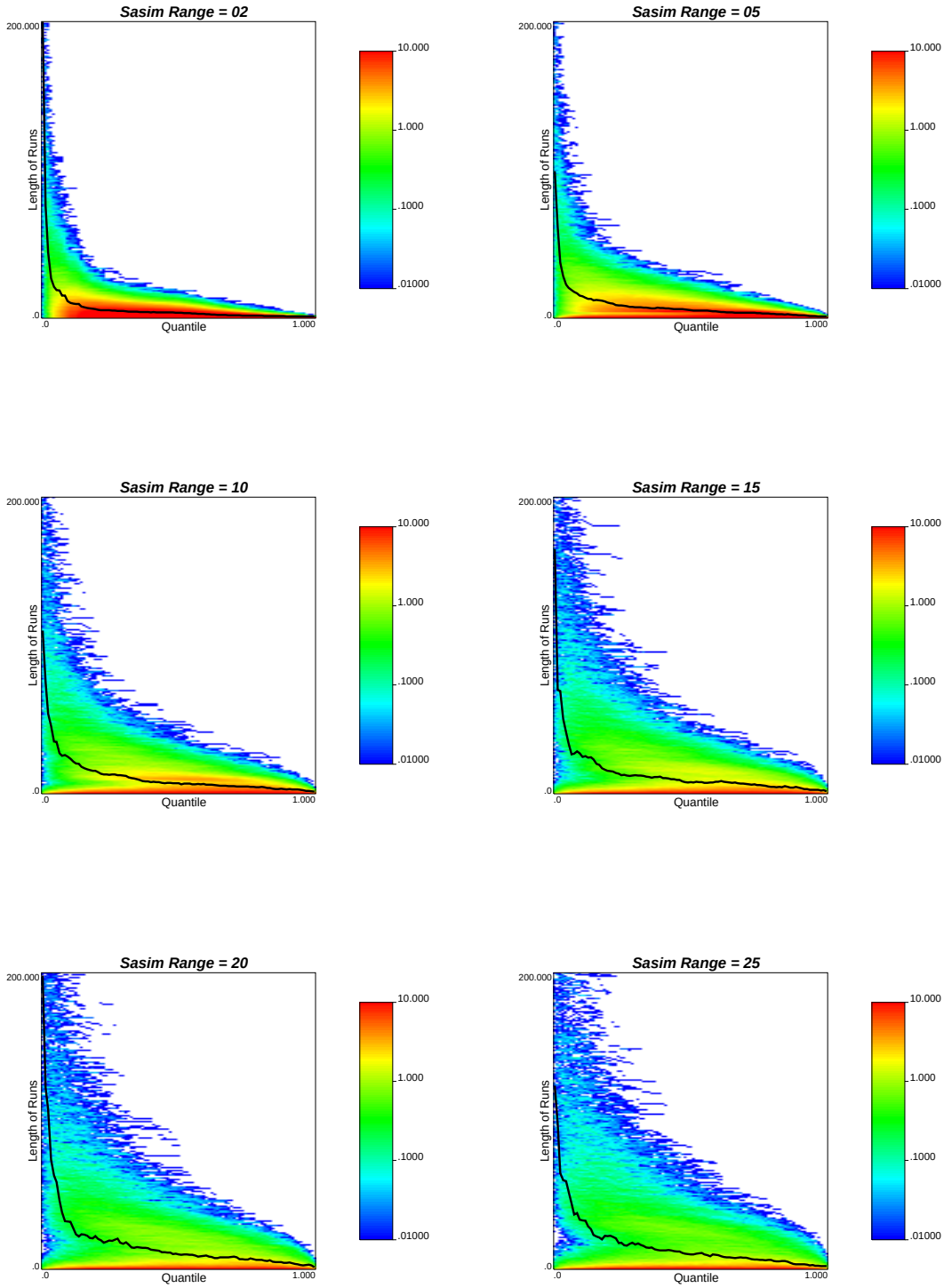

Figure 11: Map of frequency of lengths of runs above quantiles for sequences generated by simulated annealing (triangular variogram model). The solid line shows the average length as a function of the quantile.

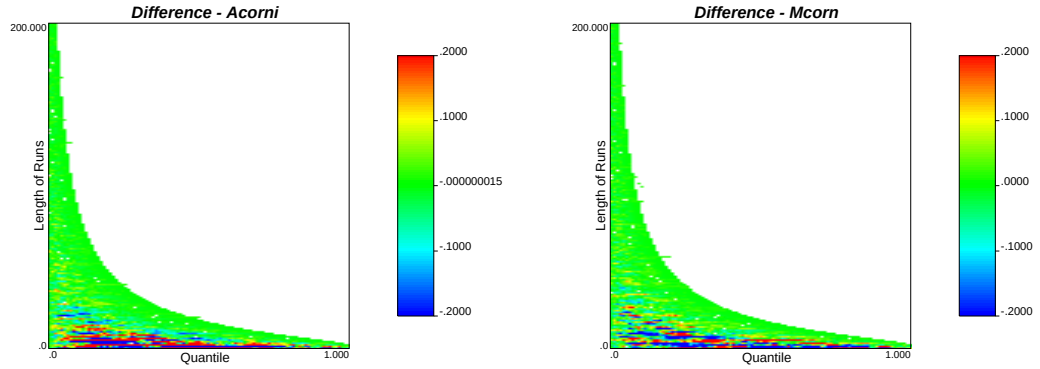

Figure 12: Map of differences of frequencies of lengths of runs above quantiles for sequences generated with `acorni` and `mcorn`.

When different models of correlation are used, slight differences in the cloud of frequencies can be seen. The curve of average lengths also changes when different variogram models are used.

The next section presents another way to look at high order correlation. Subtracting the expected frequencies of lengths of runs for the random case to the observed frequencies, maps of differences were generated.

Maps of Differences in Frequencies of Lengths of Runs

The expected number of runs above a threshold of a given length i for a random sequence, r_{1i}, can be expressed as [16]:

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

where $n_1 = n \cdot p_1$, $n_2 = n \cdot p_2$, and $x^{(a)} = x \cdot (x-1) \cdot \ldots \cdot (x-a+1)$

The difference between frequencies observed from correlated sequences and the expected for the random case were calculated. **Figure 12** shows the maps of differences using the pseudo-random sequences generated by `acorni` and `mcorn`. This just shows that the pseudo-random number generators do not depart significantly from the theoretical values. **Figure 13** shows the maps of differences for sequences generated using moving average with a triangular variogram model. The differences for the same variogram model, but for sequences generated with simulated annealing are presented in **Figure 14**. Again, different correlation functions were used with similar results and are not shown in this paper. A characteristic zone where the observed frequencies are lower than the expected ones is repeatedly seen for all ranges and variogram models; there are fewer short runs. Then, there is a zone where the observed frequencies are higher than the expected for the random case: there are more longer runs. In other words, when correlated sequences are used, there are less short runs and more long ones than in the random case.

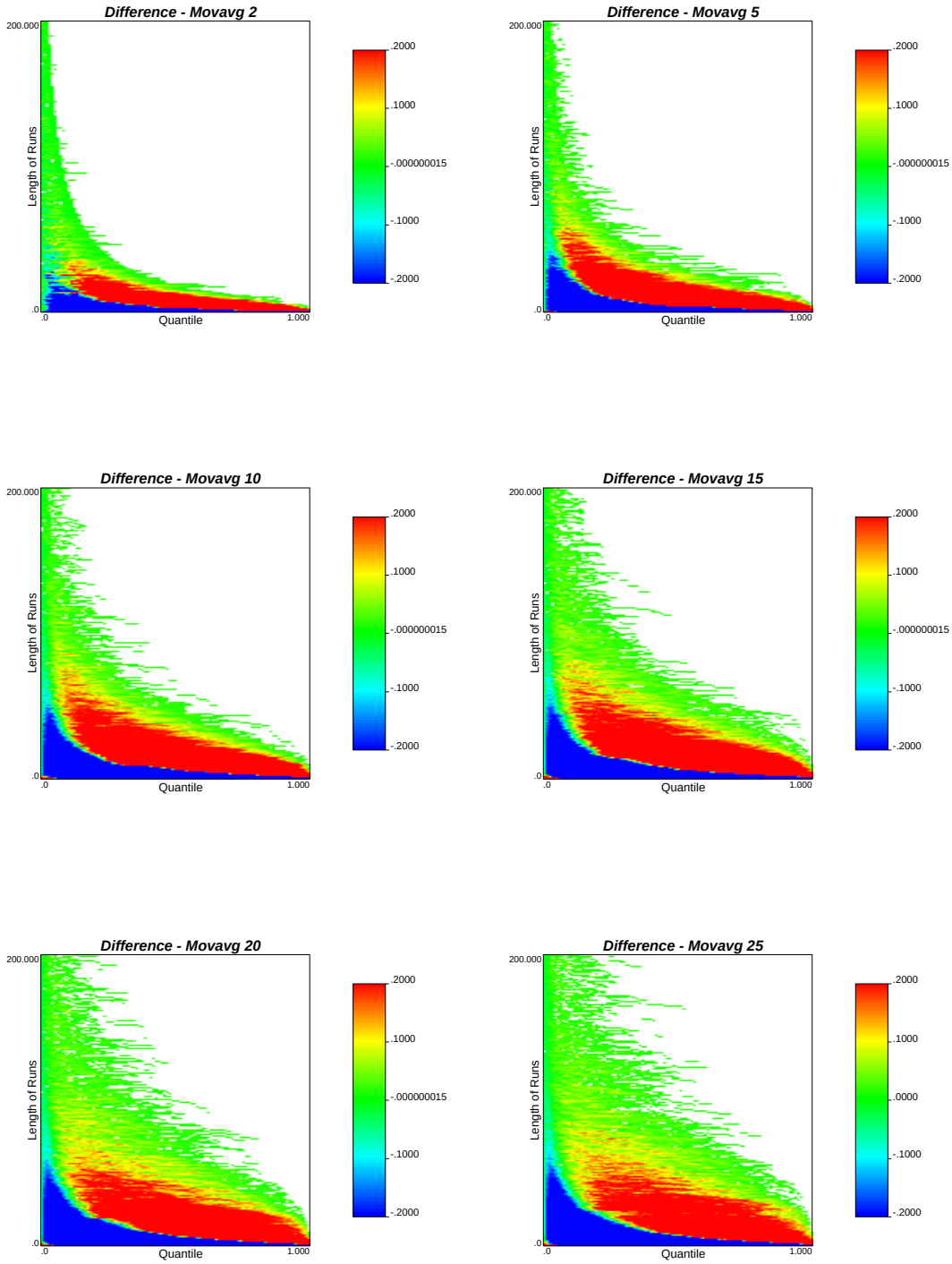

Figure 13: Map of differences of frequencies of lengths of runs above quantiles for sequences generated by moving average (triangular variogram model).

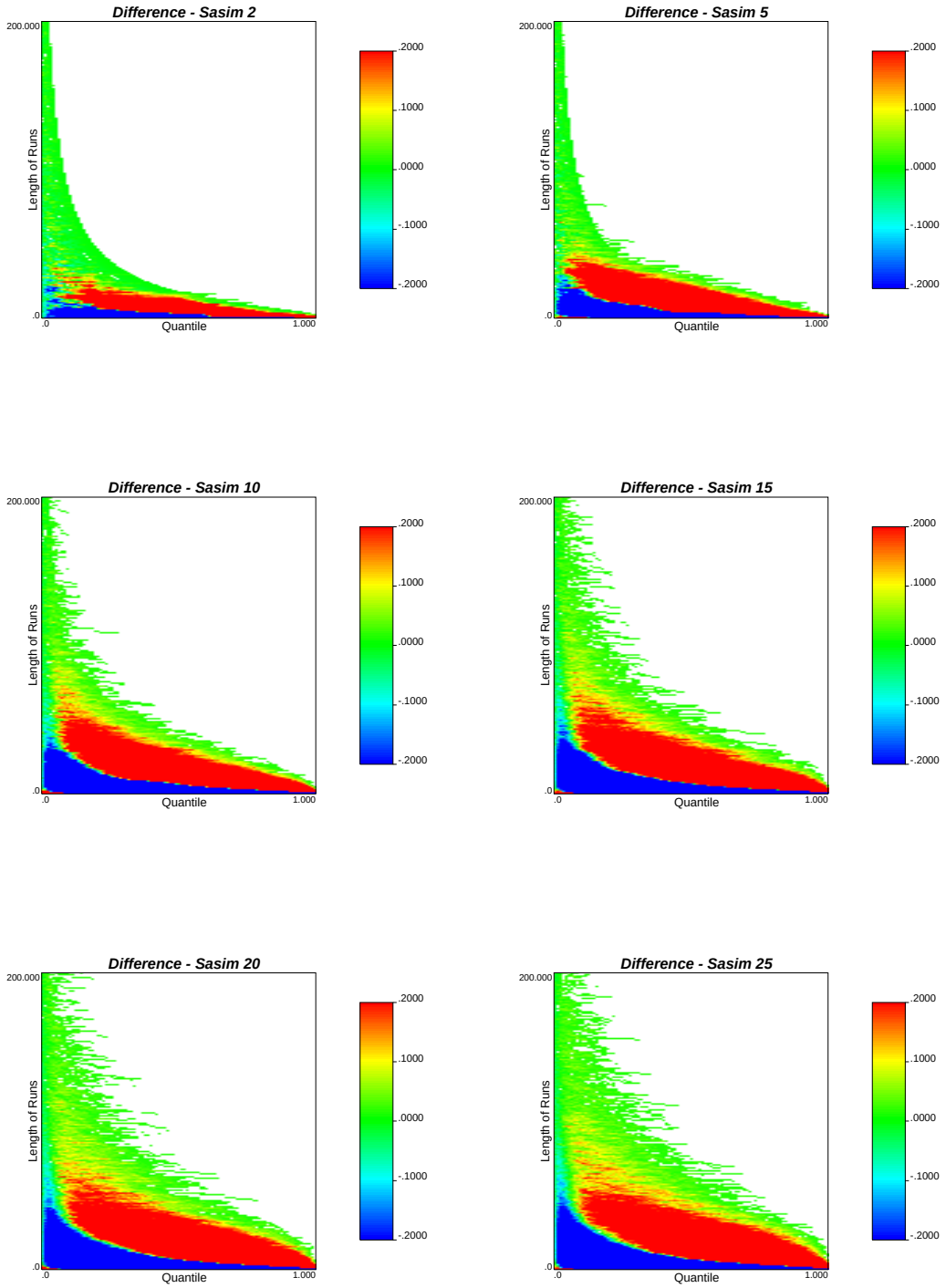

Figure 14: Map of differences of frequencies of lengths of runs above quantiles for sequences generated by simulated annealing (triangular variogram model).

Further Work

Additional bibliographic research will be done related to methods to quantify spatial variability, the indicator approach, and multiple-point statistics. The extended kriging equations and the deduction of the best *non* linear unbiased estimator have to be presented. An analytical expression for the expected distribution when the data are correlated must be deduced if possible. A parametrization of the distributions of lengths of runs could be performed.

A hierarchical sequential indicator simulation program is under development that has the promise to integrate *runs* data in addition to classical variogram information. Early results will be presented at the CCG meeting in March 2001.

References

[1] 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.

[2] C. V. Deutsch. A comparative study of pseudo-random number generators. In *Report 5*, Stanford, CA, March 1992. Stanford Center for Reservoir Forecasting.

[3] C. V. Deutsch and E. Gringarten. Accounting for multiple-point continuity in geostatistics modeling. In *6th International Geostatistics Congress*, Cape Town, South Africa, April 2000. Geostatistical Association of Southern Africa.

[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] J. C. Fu and M. V. Koutras. Distribution theory of runs: a markov chain approach. *Journal of the American Statistical Association*, 89:1050–1058, September 1994.

[6] L. A. Goodman. Simplified runs tests and likelihood ratio tests for markoff chains. *Biometrika*, 45:181–197, June 1958.

[7] P. Goovaerts. Comparative performance of indicator algorithms for modeling conditional probability distribution functions. *Math Geology*, 26(3):385–410, 1994.

[8] P. Goovaerts. *Geostatistics for Natural Resources Evaluation.* Oxford University Press, New York, 1997.

[9] B. F. Green, J. E. Keith Smith, and L. Klem. Empirical tests of an additive random number generator. *Journal of the ACM (Association for Computing Machinery)*, 6:527–537, October 1959.

[10] 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.

[11] A. G. Journel. Nonparametric estimation of spatial distribution. *Mathematical Geology*, 15:445–468, 1983.

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

[13] M. G. Kendall and B. B. Smith. Randomness and random sampling numbers. *Journal of the Royal Statistical Society*, 101:147–166, 1938.

[14] D. E. Knuth. *The Art of Computer Programming*, volume 2, Seminumerical Algorithms. Addison-Wesley, 1969.

[15] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, New York, 1969.

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

[17] F. Mosteller. Note on an application of runs to quality control charts. *Annals of Mathematical Statistics*, 12:228–232, June 1941.

[18] P. W. Shaughnessy. Multiple runs distributions: Recurrences and critical values. *Journal of the American Statistical Association*, 76:732–736, September 1981.

[19] R. M. Srivastava. Testing of sequential indicator simulation. Personal communication, November 1991.

[20] 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.

[21] B. V. Sukhatme. On certain probability distributions arising from points on a line. *Journal of the Royal Statistical Society, Series B (Methodological)*, 13:219–232, 1951.