Math Geosci (2014) 46:133–147
DOI 10.1007/s11004-013-9492-z

SPECIAL ISSUE

Can a Training Image Be a Substitute for a Random Field Model?

Xavier Emery · Christian Lantuéjoul

Received: 17 December 2012 / Accepted: 8 September 2013 / Published online: 27 September 2013

Abstract In most multiple-point simulation algorithms, all statistical features are provided by one or several training images (TI) that serve as a substitute for a random field model. However, because in practice the TI is always of finite size, the stochastic nature of multiple-point simulation is questionable. This issue is addressed by considering the case of a sequential simulation algorithm applied to a binary TI that is a genuine realization of an underlying random field. At each step, the algorithm uses templates containing the current target point as well as all previously simulated points. The simulation is validated by checking that all statistical features of the random field (supported by the simulation domain) are retrieved as an average over a large number of outcomes. The results are as follows. It is demonstrated that multiple-point simulation performs well whenever the TI is a complete (infinitely large) realization of a stationary, ergodic random field. As soon as the TI is restricted to a limited domain, the statistical features cannot be obtained exactly, but integral range techniques make it possible to predict how much the TI should be extended to approximate them up to a prespecified precision. Moreover, one can take advantage of extending the TI to reduce the number of disruptions in the execution of the algorithm, which arise when no conditioning template can be found in the TI.

Keywords Multiple-point simulation · Stationarity · Ergodicity · Integral range

X. Emery (✉)
Department of Mining Engineering, University of Chile, Avenida Tupper 2069, Santiago, Chile
e-mail: xemery@ing.uchile.cl

X. Emery
Advanced Mining Technology Center, University of Chile, Avenida Beauchef 850, Santiago, Chile

C. Lantuéjoul
Mines ParisTech, 35 rue Saint-Honoré, 77305 Fontainebleau, France
e-mail: christian.lantuejoul@mines-paristech.fr

1 Introduction

Initiated by Guardiano and Srivastava (1993), multiple-point statistics (MPS) have known a surge of interest once that Strebelle (2002) discovered efficient ways to implement them for geostatistical simulation. Nowadays, MPS are the object of extensive developments, especially in hydrology and reservoir engineering, where the challenges are to deal with continuous variables (Mariethoz et al. 2010), to account for nonstationarities (Strebelle and Zhang 2005) and to accommodate various types of constraints (Hu and Chugunova 2008; Renard et al. 2011). The reasons that explain the success of MPS techniques include their generality, their capability to reproduce complicated shapes as well as their conceptual simplicity. Owing to the experience gained over the last years, the status of the training image (TI) that is generally used to compute MPS is worthwhile being revisited. As mentioned by Ortiz (2008), its link with random fields has not been properly addressed. This is what is investigated in this paper. To this end, a TI that is a genuine realization of a stationary, ergodic random field (SERF) is considered. The objective is to check whether or not the multiple-point techniques can produce outcomes that are statistically acceptable as realizations of the SERF. For the sake of simplicity, the SERF is assumed binary (taking values in $\{0, 1\}$). Discrete SERFs (taking a finite set of values) could be treated similarly. Two different cases will be successively considered, depending on whether the TI is defined in the whole space or in a limited domain. A note on terminology: throughout the paper, the word "outcomes" refers to the simulations produced by multiple-point algorithms, whereas the word "realizations" refers to those constructed from random field models.

2 Basic Concepts and Notation

2.1 Translation, Dilation, and Erosion Operators

Throughout this paper, the workspace is $\mathbb{Z}^2$ and its origin is denoted by $\mathbf{o}$. Let $\mathbf{h}$ be a point of $\mathbb{Z}^2$ and X be a subset of $\mathbb{Z}^2$. Then the translation of X with respect to $\overrightarrow{\mathbf{oh}}$ is denoted by $\tau_{\mathbf{h}} X$. Let K be another subset of $\mathbb{Z}^2$. The dilation and the erosion of X by K are respectively defined as

$$\delta_K X = \left\{\mathbf{h} \in \mathbb{Z}^2 : \tau_{\mathbf{h}} K \cap X \neq \emptyset\right\} \qquad \varepsilon_K X = \left\{\mathbf{h} \in \mathbb{Z}^2 : \tau_{\mathbf{h}} K \subset X\right\}.$$

2.2 Level Set, Template, and Hit-or-Miss Transform

A binary image I with support $D(I) \subset \mathbb{Z}^2$ can be characterized by its level sets

$$S_0(I) = \left\{\mathbf{x} \in D(I) : I(\mathbf{x}) = 0\right\} \qquad S_1(I) = \left\{\mathbf{x} \in D(I) : I(\mathbf{x}) = 1\right\}.$$

Let $T = (K_0, K_1)$ be an ordered pair of subsets of $\mathbb{Z}^2$, which will be called a template. The hit-or-miss transform of I by T is the set of points where the pair (K_0, K_1) fits the values of I (Serra 1982) (Fig. 1)

$$\eta_T^I = \left\{\mathbf{x} \in \mathbb{Z}^2 : \tau_{\mathbf{x}} K_0 \subset S_0(I),\ \tau_{\mathbf{x}} K_1 \subset S_1(I)\right\}.$$

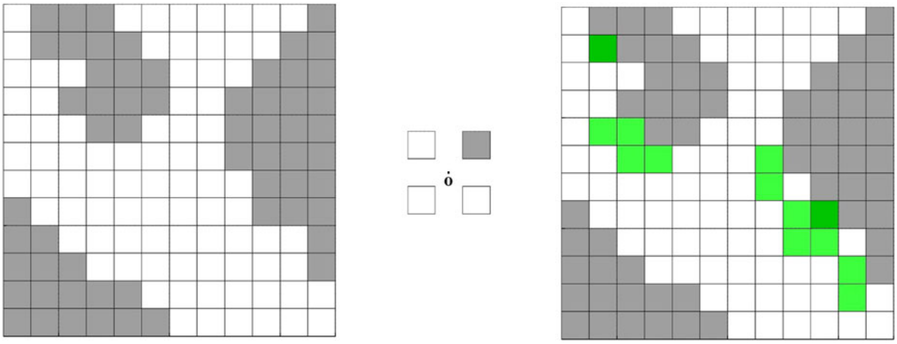

Fig. 1 Example of a hit-or-miss transform: the original image I with levels sets painted in *white and gray* (*left*), a four-point template T centered at the origin (*middle*), and the result η_T^I (set of points painted in *green*) (*right*)

2.3 Random Fields, Stationarity, and Ergodicity

Consider now a binary random field Z defined on $\mathbb{Z}^2$. Its statistical properties are characterized by the spatial distribution of its level sets, that is the set of probabilities of the form $\text{Prob}\{K_0 \subset S_0(Z), K_1 \subset S_1(Z)\}$ when K_0 and K_1 are finite subsets of $\mathbb{Z}^2$. Z is stationary if its spatial distribution is invariant under translation

$$\forall \mathbf{h} \in \mathbb{Z}^2 \quad \text{Prob}\{\tau_{\mathbf{h}} K_0 \subset S_0(Z), \tau_{\mathbf{h}} K_1 \subset S_1(Z)\} = \text{Prob}\{K_0 \subset S_0(Z), K_1 \subset S_1(Z)\}.$$

When Z is stationary, Z is also ergodic if its spatial distribution can be retrieved from any of its realizations, say z (Matheron 1989; Chilès and Delfiner 2012)

$$\text{Prob}\{K_0 \subset S_0(Z), K_1 \subset S_1(Z)\} = \lim_{\lambda \to \infty} \frac{\#[\eta_T^z \cap Sq(\lambda)]}{\#Sq(\lambda)}, \tag{1}$$

where $T = (K_0, K_1)$, $Sq(\lambda)$ is the square of side length $2\lambda + 1$ centered at $\mathbf{o}$, and # indicates cardinality. For example, a random field with no spatial structure, made of independent and identically distributed Bernoulli random variables, is stationary and ergodic, by virtue of the strong law of large numbers. In contrast, a random field made of the same Bernoulli random variable at every point of $\mathbb{Z}^2$ is not ergodic: Indeed, the values observed in one realization are all the same, providing a marginal distribution with zero variance, whereas the true underlying distribution is a Bernoulli distribution.

3 Multiple-Point Simulation with an Infinite TI

3.1 Nonconditional Algorithm

Let I be a TI that is an entire realization over $\mathbb{Z}^2$ of a binary SERF Z. The following algorithm sequentially simulates the level sets S_0 and S_1 in a finite domain D. In this algorithm, $\mathscr{U}(A)$ and $\mathscr{U}$ denote the uniform distribution over A and over $]0, 1[$, respectively.

Algorithm 1

(i) set $S_0 = S_1 = \emptyset$;
(ii) generate $\mathbf{x} \sim \mathscr{U}(D \backslash (S_0 \cup S_1))$, and put $T_0 = (S_0 \cup \{\mathbf{x}\}, S_1)$, $T_1 = (S_0, S_1 \cup \{\mathbf{x}\})$;
(iii) compute

$$p = \lim_{\lambda \to \infty} \frac{\#[\eta^I_{T_1} \cap Sq(\lambda)]}{\#[\eta^I_{T_0} \cap Sq(\lambda)] + \#[\eta^I_{T_1} \cap Sq(\lambda)]};$$

(iv) generate $u \sim \mathscr{U}$, and put $S_1 = S_1 \cup \{\mathbf{x}\}$ if $u < p$ and $S_0 = S_0 \cup \{\mathbf{x}\}$ otherwise;
(v) if $S_0 \cup S_1 \neq D$, then go to (ii);
(vi) stop.

Algorithm 1 is nothing but a standard sequential algorithm, except that p is computed from the TI instead of being assigned the value $\text{Prob}\{Z(\mathbf{x}) = 1 \mid Z(S_0) = 0, Z(S_1) = 1\}$ calculated from the random field model. One must ensure that p takes the correct value and is not of the form $0/0$ at any time. The proof is deferred to Appendix A.

3.2 Conditional Algorithm

Let now C_0 and C_1 be conditioning data points for both level sets. A natural idea is to run the nonconditional algorithm by assuming that the conditioning data points correspond to already simulated values, which boils down to replacing $S_0 = S_1 = \emptyset$ by $S_0 = C_0$ and $S_1 = C_1$ at step (i) of Algorithm 1. This effectively works, provided that $\text{Prob}\{Z(C_0) = 0, Z(C_1) = 1\} > 0$. In the opposite case, the conditioning information is not compatible with the random field model from which the TI is derived. The algorithm stops because no data template can be found in the TI.

4 Multiple-Point Simulation with a Finite TI

In this section, the TI considered is a realization of a SERF that is limited to a finite domain $D(I)$.

4.1 Nonconditional Algorithm

It does not look significantly different from that of the infinite case. The only difference is that the computation of the Bernoulli parameter p is simplified.

Algorithm 2

(i) set $S_0 = S_1 = \emptyset$;
(ii) generate $\mathbf{x} \sim \mathscr{U}(D \backslash (S_0 \cup S_1))$, and put $T_0 = (S_0 \cup \{\mathbf{x}\}, S_1)$, $T_1 = (S_0, S_1 \cup \{\mathbf{x}\})$;
(iii) compute

$$p = \frac{\#\eta^I_{T_1}}{\#\eta^I_{T_0} + \#\eta^I_{T_1}};$$

```
(iv) generate $u \sim \mathscr{U}$, and put $S_1 = S_1 \cup \{\mathbf{x}\}$ if $u < p$ and $S_0 = S_0 \cup \{\mathbf{x}\}$ otherwise;
(v) if $S_0 \cup S_1 \neq D$, then go to (ii);
(vi) stop.
```

This basically corresponds to the SNESIM algorithm (Strebelle 2002), except that the neighboring data search is not restricted. At each stage, all the previously simulated values are used as conditioning data, not only those that are close to the target point $\mathbf{x}$. As a matter of fact, the computation of parameter p at step (iii) is not even required. Because $\eta^I_{T_0} \cap \eta^I_{T_1} = \emptyset$, one has $\text{Prob}\{I(\tau_{\mathbf{y}}\{\mathbf{x}\}) = 1\} = p$ if $\mathbf{y}$ is a uniform point of $\eta^I_{T_0} \cup \eta^I_{T_1}$. Accordingly, borrowing an idea from Mariethoz et al. (2010) (direct sampling), Algorithm 2 can be rewritten as follows.

Algorithm 3

```
(i) set $S_0 = S_1 = \emptyset$;
(ii) generate $\mathbf{x} \sim \mathscr{U}(D \backslash (S_0 \cup S_1))$, and put $T_0 = (S_0 \cup \{\mathbf{x}\}, S_1)$, $T_1 = (S_0, S_1 \cup \{\mathbf{x}\})$;
(iii) generate $\mathbf{y} \sim \mathscr{U}(\eta^I_{T_0} \cup \eta^I_{T_1})$;
(iv) put $S_{I(\tau_{\mathbf{y}}\{\mathbf{x}\})} = S_{I(\tau_{\mathbf{y}}\{\mathbf{x}\})} \cup \{\mathbf{x}\}$;
(v) if $S_0 \cup S_1 \neq D$, then go to (ii);
(vi) stop.
```

Algorithm 3 shows that only the templates contained in the TI can be reproduced. Consequently, the final outcome is nothing but a piece of the TI. An immediate implication is that a TI can produce but a limited number of outcomes. This may be a source of problems when many outcomes are required to derive confidence limits for regional features. Another limitation of this algorithm is that $\eta^I_{T_0} \cup \eta^I_{T_1}$ may be empty, in which case no uniform point $\mathbf{y}$ can be selected at step (iii), and no random value can be assigned to the current point $\mathbf{x}$. In such a case, the algorithm stops. Assuming that the support of the TI is convex, this may happen when $\mathbf{x}$ does not belong to the convex hull of $S_0 \cup S_1$.

4.2 Conditional Algorithm

Exactly as for an infinite TI, Algorithms 2 and 3 can be made conditional by prescribing the conditions $T_c = (C_0, C_1)$ to be satisfied at the initial step. However, the conditioning data may be incompatible with the TI, i.e $\eta^I_{T_c} = \emptyset$, even if they are compatible with the SERF model (i.e., $\text{Prob}\{Z(C_0) = 0, Z(C_1) = 1\} > 0$). Moreover, the difficulties encountered in the nonconditional case remain. In particular, the number of possible outcomes may be extremely limited, as they should correspond to pieces of the TI that fit the conditioning data.

4.3 Reducing Templates

As mentioned above, step (iii) of Algorithms 2 and 3 fail if none of the templates T_0 and T_1 is found in the TI. To bypass this problem, one solution is to reduce these templates, by discarding the points of S_0 and S_1 that lie outside a neighborhood of the current point $\mathbf{x}$ or that contain less information (Strebelle 2002; Liu 2005;

Eskandaridalvand and Srinivasan 2010). This amounts to replacing the conditional distribution $\text{Prob}\{Z(\mathbf{x}) = 1 | Z(S_0) = 0, Z(S_1) = 1\}$ by a distribution conditioned to reduced sets $\text{Prob}\{Z(\mathbf{x}) = 1|Z(S'_0) = 0, Z(S'_1) = 1\}$, where $S'_0 \subset S_0$ and $S'_1 \subset S_1$. Putting $S''_0 = S_0 \backslash S'_0$ and $S''_1 = S_1 \backslash S'_1$, this implicitly entails the introduction of a conditional independence relationship between $Z(\mathbf{x})$ and $Z(S''_0 \cup S''_1)$ given $Z(S'_0 \cup S'_1)$. One consequence is that the multivariate distribution of the outcome, $Z(D)$, is likely to differ from the distribution of the underlying SERF. A simple but illustrative example is given by Holden (2006), who considers a one-dimensional training image I such that the width of the level sets $S_0(I)$ and $S_1(I)$ is always strictly greater than one unit. By using a restricted neighborhood in the simulation algorithm, he points out a situation in which the outcomes exhibit level sets with a unit width, thus in disagreement with the statistical features of the TI. Another example is provided by Arpat (2005), who shows that the template reduction is likely to produce large-scale artifacts in the final outcomes. The artifacts may be not negligible because each simulated data point is subsequently used as conditioning data for simulating the points of D that have not been processed yet, so that the error made in the conditional distribution at each point propagates to the next ones. Other solutions have been proposed to circumvent these problems, such as accepting the simulation of templates that are similar to those of the training image (according to a given similarity measure) or resimulating points that provoke inconsistencies with the training image (Arpat 2005; Strebelle and Remy 2005; Mariethoz et al. 2010). In the next sections, we will not dwell on these ideas and rather turn to the following question: What is the area required for the TI to avoid, with a given confidence level, blockings in Algorithms 2 and 3?

5 TI Representativeness

To summarize the previous two sections, the sequential algorithm gives satisfactory outcomes whenever applied to a TI that is an entire realization of a SERF and may fail otherwise. This conclusion is a bit schematic because one can surmise that the larger the TI, the more satisfactory the outcomes. The problem of course is to understand what does one mean by “large.”

5.1 Integral Range

Let Z be a binary SERF on $\mathbb{Z}^2$. For each template $T = (K_0, K_1)$, an indicator random field Z_T can be defined as

$$Z_T(\mathbf{x}) = 1_{\mathbf{x} \in \eta_T^Z}.$$

It is not difficult to show that Z_T is also a binary SERF on $\mathbb{Z}^2$. Its mean, variance, and correlation function are respectively denoted by μ_T, σ_T^2 and ρ_T. Let also V be a finite subset of $\mathbb{Z}^2$, and $Z_T(V)$ the average of $Z_T(\mathbf{x})$ when $\mathbf{x}$ scours V

$$Z_T(V) = \frac{1}{\#V} \sum_{\mathbf{x} \in V} Z_T(\mathbf{x}).$$

Clearly, the expected value of $Z_T(V)$ is μ_T, which is the probability of occurrence of template T. Its variance can be written as (Matheron 1971; Chilès and Delfiner 2012)

$$\mathrm{Var}\big\{Z_T(V)\big\} = \frac{\sigma_T^2}{(\#V)^2} \sum_{\mathbf{h}\in\mathbb{Z}^2} \rho_T(\mathbf{h})\,\kappa_V(\mathbf{h}),$$

where $\kappa_V(\mathbf{h}) = \#(V \cap \tau_{\mathbf{h}} V)$ is the geometric covariogram of V. Now, if the range of ρ_T is small compared to the size of V, then one heuristically has $\kappa_V(\mathbf{h}) \approx \kappa_V(\mathbf{o}) = \#V$ wherever $\rho_T(\mathbf{h}) \not\approx 0$, which leads to

$$\mathrm{Var}\big\{Z_T(V)\big\} \approx \frac{\sigma_T^2}{\#V} \sum_{\mathbf{h}\in\mathbb{Z}^2} \rho_T(\mathbf{h})$$

when $\#V$ is large. The quantity $a_T = \sum_{\mathbf{h}\in\mathbb{Z}^2} \rho_T(\mathbf{h})$ is called the integral range of ρ_T (Matheron 1989). It is nonnegative (but possibly infinite) and depends on both the template T and the spatial distribution of Z. For instance, if $T = (\emptyset, \mathbf{o})$, then $Z_T(\mathbf{x}) = Z(\mathbf{x})$ and a_T exclusively depends on the correlation function of Z: it decreases when its nugget effect increases or when its range decreases, that is, when Z has a poor spatial structure. If $0 < a_T < \infty$, then it can be shown (Lantuéjoul 1991) that

$$\mathrm{Var}\big\{Z_T(V)\big\} \approx \frac{\sigma_T^2 a_T}{\#V} \quad \#V \gg a_T.$$

As $\#V$ is large, the central limit theorem is applicable and states that the distribution of $Z_T(V)$ is approximatively normal with mean μ_T and variance $\sigma_T^2 a_T/\#V$. From this, one can assess, up to a prespecified precision, how large should V be to contain at least n copies of template T. Starting from $Z_T(V) \approx \mu_T + \sigma_T\sqrt{a_T/\#V}\,Y$ where Y is a standard normal variable, one has

$$\mathrm{Prob}\big\{\#V\, Z_T(V) \geq n\big\} \geq 1-\alpha \quad \text{iff} \quad \mathrm{Prob}\left\{Y \geq \frac{n - \#V\mu_T}{\sigma_T\sqrt{\#V a_T}}\right\} \geq 1-\alpha.$$

Denoting by y_α the quantile of order α of Y, the latter condition will be satisfied as soon as

$$\frac{n - \#V\mu_T}{\sigma_T\sqrt{\#V a_T}} \leq y_\alpha,$$

that is

$$\sqrt{\#V} \geq \frac{-\sigma_T\sqrt{a_T}\,y_\alpha + \sqrt{\sigma_T^2 a_T y_\alpha^2 + 4\mu_T n}}{2\mu_T}.$$

Because $\sigma_T^2 = \mu_T(1-\mu_T)$ and $y_\alpha < 0$ in the standard case when α is less than 0.5, this formula simplifies into

$$\sqrt{\#V} \geq \frac{\sqrt{(1-\mu_T)a_T y_\alpha^2} + \sqrt{(1-\mu_T)a_T y_\alpha^2 + 4n}}{2\sqrt{\mu_T}}. \tag{2}$$

This equation shows that the area required for the TI to ensure the occurrence of a given template at least n times increases when the template becomes scarce (μ_T small) or when the integral range associated with the template becomes extended (a_T large).

5.2 Example: The Stationary Boolean Model

This example has been chosen because of its conceptual simplicity and its mathematical tractability. Intuitively speaking, the Boolean model is an aggregate of independent and possibly overlapping objects. This model was extensively studied by Matheron (1975) in the continuous case. In the present paper, a discrete version is proposed, which is more appropriate to our objectives. The construction of a discrete stationary Boolean model rests on the following ingredients. Independent Poisson random variables with parameter θ, $(N(\mathbf{u}), \mathbf{u} \in \mathbb{Z}^2)$, are attached to the points of $\mathbb{Z}^2$, along with independent copies $(A_{\mathbf{u},n},\ \mathbf{u} \in \mathbb{Z}^2,\ n \leq N(\mathbf{u}))$ of some random finite reference set A of $\mathbb{Z}^2$. The indicator function of the Boolean model is defined as

$$Z(\mathbf{x}) = \max_{\mathbf{u} \in \mathbb{Z}^2} 1_{\mathbf{x} \in \tau_{\mathbf{u}} A_{\mathbf{u}}},$$

where $A_{\mathbf{u}} = \bigcup_{n \leq N(\mathbf{u})} A_{\mathbf{u},n}$ if $N(\mathbf{u}) > 0$ and $\emptyset$ otherwise. The statistical properties of the Boolean model are completely specified by the Poisson parameter θ and the distribution of the reference set A. The standard statistics of the Boolean model are its mean $\mu = 1 - e^{-\theta \kappa_A(\mathbf{o})}$, its variance $\sigma^2 = e^{-\theta \kappa_A(\mathbf{o})}[1 - e^{-\theta \kappa_A(\mathbf{o})}]$ and its correlation function $\rho(\mathbf{h}) = \frac{e^{\theta \kappa_A(\mathbf{h})} - 1}{e^{\theta \kappa_A(\mathbf{o})} - 1}$. Besides the Poisson parameter, all three involve nothing but the geometric covariogram $\kappa_A(\mathbf{h})$ of the reference set A. Now formulae become more complicated when the template cardinality exceeds two points. For instance, the probability that a subset K_0 avoids all objects of the Boolean model is

$$\mathrm{Prob}\bigl\{Z(K_0) = 0\bigr\} = \exp\bigl(-\theta\, E\{\#\delta_{K_0} A\}\bigr), \tag{3}$$

where $\delta_{K_0} A$ is the dilation of A by K_0. More generally, for $T = (K_0, K_1)$, the inclusion-exclusion formula gives

$$\mu_T = \mathrm{Prob}\bigl\{Z(K_0) = 0, Z(K_1) = 1\bigr\} = \sum_{L \subset K_1} (-1)^{\#L} \exp\bigl(-\theta\, E\{\#\delta_{K_0 \cup L} A\}\bigr). \tag{4}$$

Equations (3) and (4) can be established by analogy with the continuous case (Lantuéjoul 2002); a formal proof is given in Appendix B. To fix ideas, a Boolean model of squares with fixed side length $a = 11$ is considered next (Fig. 2). The Poisson parameter is chosen to yield a 50 % proportion of zeros (explicitly $\theta = 0.0057$). In a first instance, templates supported by the four vertices of a square of side d ($\mathbf{x}_1 = (0,0)$, $\mathbf{x}_2 = (d,0)$, $\mathbf{x}_3 = (d,d)$ and $\mathbf{x}_4 = (0,d)$) are examined. Among the $2^4 = 16$ possible templates, only 6 are retained using symmetry arguments, namely (Fig. 3)

$$\begin{array}{lll} T_1 = \bigl((\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4), \emptyset\bigr) & T_2 = \bigl((\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3), \mathbf{x}_4\bigr) & T_3 = \bigl((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_3, \mathbf{x}_4)\bigr) \\ T_4 = \bigl((\mathbf{x}_1, \mathbf{x}_3), (\mathbf{x}_2, \mathbf{x}_4)\bigr) & T_5 = \bigl(\mathbf{x}_1, (\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)\bigr) & T_6 = \bigl(\emptyset, (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)\bigr) \end{array} \tag{5}$$

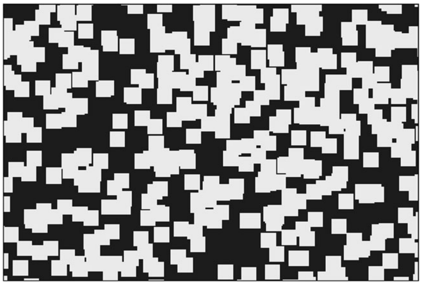

Fig. 2 An example of Boolean realization with a square reference set

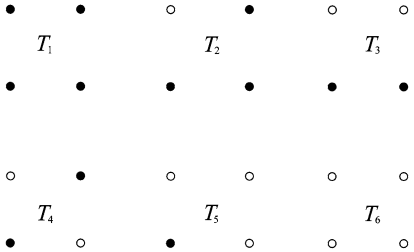

Fig. 3 Four-point templates T_1 to T_6. *Black circles* indicate zero values, *white circles* one values

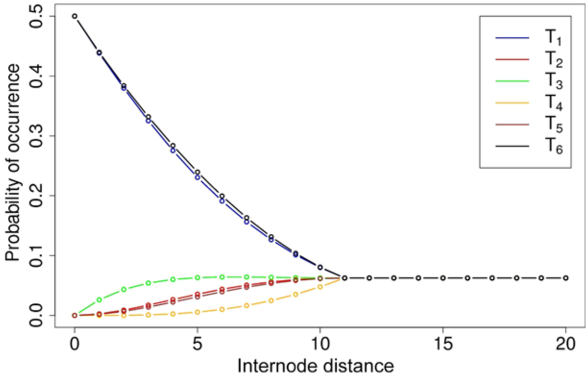

Fig. 4 Probabilities of occurrence of the four-point templates T_1 to T_6

Figure 4 shows their probabilities of occurrence $\{\mu_{T_i}, i = 1, \ldots, 6\}$ when the internode distance d ranges from 0 to 20. For $d = 0$, the only possible templates are T_1 and T_6 that contain only zeros or ones, and $\mu_{T_1} = \mu_{T_6} = 0.5$. When d increases, the chance that the four points belong to the same level set decreases. As all probabilities are related by the formula $\mu_{T_1} + 4\mu_{T_2} + 4\mu_{T_3} + 2\mu_{T_4} + 4\mu_{T_5} + \mu_{T_6} = 1$, the other probabilities increase. Note, however, that the growths are rather differentiated. For instance, μ_{T_4} increases more slowly than μ_{T_3}. Although both T_3 and T_4 have two points with a value of 1, those of T_3 may belong to the same object whereas those of T_4 must belong to different objects, which is not so likely owing to the Boolean parameters. When d exceeds the side length of the objects (i.e., $d \geq 11$), the points of all templates take independent values. As a result, the six templates have the same probability of occurrence ($0.5^4 = 0.0625$).

Now, Fig. 4 says little about the spatial arrangements of a template $T = (K_0, K_1)$ in a Boolean realization. This issue is investigated by considering the SERF Z_T as

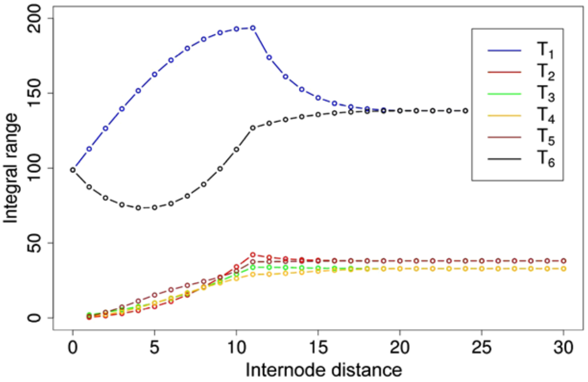

Fig. 5 Integral ranges associated with the four-point templates T_1 to T_6

in Sect. 5.1. Its mean μ_T is given by Eq. (4) and its variance σ_T^2 is equal to $\mu_T(1-\mu_T)$ since Z_T is a binary random field. Its correlation function satisfies $\sigma_T^2\rho_T(\mathbf{h}) = \mu_{T(\mathbf{h})} - \mu_T^2$ where $T(\mathbf{h})$ is the template $(K_0 \cup \tau_{\mathbf{h}} K_0, K_1 \cup \tau_{\mathbf{h}} K_1)$. Because $\rho_T(\mathbf{h})$ vanishes when $\mathbf{h} \notin Sq(a+d)$, the integral range of Z_T can be written

$$a_T = \sum_{\mathbf{h} \in Sq(a+d)} \rho_T(\mathbf{h}).$$

Figure 5 shows the integral ranges of the six templates $T_1, \ldots, T_6$ as the internode distance d varies from 0 to 20. It appears that the curves associated with T_1 and T_6 substantially differ, although their probability curves are similar. To explain what happens, put $K = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$. It is not difficult to establish that $S_1(Z_{T_6}) = \varepsilon_K S_1(Z)$. If the internode distance is small compared to the object size, then each connected component of $S_1(Z)$ (that is a union of objects) remains nonempty by erosion and produces a cluster of T_6 templates. On the other hand, one can also establish $S_1(Z_{T_1}) = \varepsilon_K S_0(Z)$. Yet, there is no minimal size for the connected components of $S_0(Z)$. They may vanish by erosion even at small distances. Accordingly, the T_1-templates are not arranged in clusters. They are more scattered than the T_6-templates, which results in a larger integral range. At large distances, all eroded connected components may be empty and the clustering effect does not hold any longer: Z_{T_1} and Z_{T_6} have the same correlation function $\rho^4(\mathbf{h})$, hence the same integral range.

One can derive the area of the TI required to contain at least $n = 50$ templates with 95 % confidence, using Eq. (2) with $y_\alpha = -1.64$. The results are displayed on Fig. 6. It can be seen that several templates, such as T_4 with a unit internode distance, require a very large TI (area greater than 4, 500, 000, corresponding to a TI with a side length close to 200 times the range of the correlation function ρ along the main axes) in order to be found numerous enough. Even if one decreases the number n of template occurrences, the TI area is likely to remain large, insofar as it mainly depends on the integral range associated with the template, as expressed by the numerator of Eq. (2). One may argue that, in the model, template T_4 with a unit internode distance corresponds to a quite rare event (Fig. 4), as it happens only when two objects are touching at one vertex. This event may, however, be relevant as it can affect the connectivity of the simulated structure.

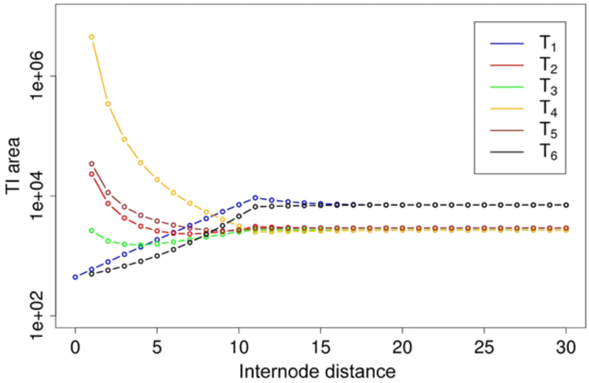

Fig. 6 TI areas required to get at least 50 copies of the four-point templates T_1 to T_6 with 95 % confidence

To illustrate the effect of the template cardinality, let us now consider the template $T = (S_0, \emptyset)$, where S_0 is a square grid with mesh d and with $k \times k$ points ($k \geq 1$). Figure 7 shows the probability of occurrence of T, the integral range of ρ_T and the TI area required to contain at least 50 copies of T with 95 % confidence (Eq. (2)), as a function of the grid mesh d, for $k = 1, 2, \ldots, 12$. In comparison with Fig. 6, the TI area dramatically increases with the template cardinality and with the grid mesh. This is mainly explained because the integral range (in the numerator of Eq. (2)) takes relatively high values (between 30 and 200), while the probability of occurrence of the template (in the denominator of Eq. (2)) strongly decreases as the template cardinality and the grid mesh increase.

As an example, for a 12×12 template with $d = 2$, the required TI area is $227, 600$, corresponding to a TI with a side length of about 43 times the range of correlation of the underlying SERF along the main axes, although the probability of occurrence of the template is not negligible (0.002). These figures show that, even when working with rather common templates, a quite large TI may be necessary to contain "sufficiently enough" copies of these templates and to avoid the sequential simulation algorithm to be trapped. For rare templates, the required TI area is out of practical reach. For instance, the 12×12 template with $d = 3$ leads to an area of $18, 416, 000$, corresponding to a side length of almost 400 times the range of correlation along the main axes; although quite low (0.000015), the probability of occurrence of this template is much larger than the average probability of a 12×12 template (10^{-43}).

6 Conclusions

The motivation of this paper was to determine whether or not a training image (TI) could replace a random field model for geostatistical simulation. To this end, the TI has been considered as a realization of an underlying random field, assumed stationary and ergodic, so that the statistics calculated over a large image should coincide with the model statistics. It has been shown that sequential simulation based on multiple-point statistics successfully reproduces the model statistics when the TI is infinitely large, but the algorithm may fail when the TI is bounded, the data templates being less and less likely to be found in the TI as their cardinality increases.

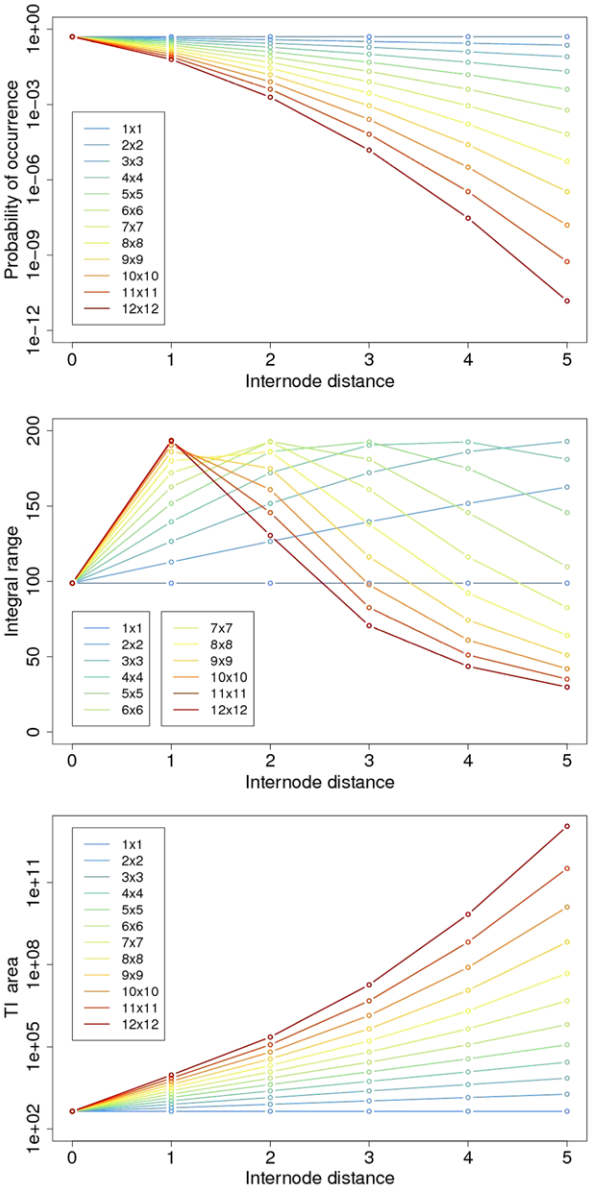

Fig. 7 Probabilities of occurrence, integral ranges and TI areas required to get at least 50 copies with 95 % confidence, for templates supported by $k \times k$ points

To overcome this problem, two approaches can be considered. The first one consists in reducing the templates, but the price is a loss of accuracy in the outcomes. The second one consists in enlarging the TI, in order to get closer to the ideal case of an infinite TI. In turn, this approach raises the question of the representativeness of the TI: how large should it be to ensure, with a given confidence level, that a given template T is found sufficiently enough times? The answer to that question is given by Eq. (2) and depends on the template probability of occurrence (μ_T) and on the spatial distribution of the random field, via the integral range (a_T) associated with the template. Numerical experiments conducted on specific templates with a Boolean model as the underlying random field suggest that the required TI area considerably increases with the template cardinality and with its rareness.

Our results go against the statement of Journel and Zhang (2006), who claim that a large TI can be seen as a statistically explicit prior random field model. Indeed, unless an extremely large TI is considered or the templates are restricted to a few points, the statistical properties of a random field have actually very little chance to be retrieved starting from a bounded TI. In such a situation, practitioners should be aware that multiple-point simulation algorithms are not really geostatistical simulation algorithms, but rather stochastic computer-aided design algorithms: Although these algorithms are successful in reproducing complex heterogeneous structures in geosciences applications, they cannot be associated with a well-defined random field model and their outcomes are dependent on proper implementation parameters.

Acknowledgements The authors are grateful to the editors and the three anonymous reviewers for their constructive comments on a former version of this paper.

Appendix A: Proof of Algorithm 1

The proof is established by induction. Let S_0 and S_1 be the current level sets of the outcome, and suppose that $\mathrm{Prob}\{Z(S_0) = 0, Z(S_1) = 1\} = \beta > 0$. Put $p = \lim_{\lambda\to\infty} \frac{N_\lambda}{D_\lambda}$ with

$$N_\lambda = \frac{\#[\eta^I_{T_1} \cap Sq(\lambda)]}{\#Sq(\lambda)} \qquad D_\lambda = \frac{\#[\eta^I_{T_0} \cap Sq(\lambda)] + \#[\eta^I_{T_1} \cap Sq(\lambda)]}{\#Sq(\lambda)}.$$

By the ergodic property (Eq. (1)), one has

$$\begin{aligned}\lim_{\lambda\to\infty} N_\lambda &= \mathrm{Prob}\big\{Z(S_0) = 0, Z\big(S_1 \cup \{\mathbf{x}\}\big) = 1\big\}\\ \lim_{\lambda\to\infty} D_\lambda &= \mathrm{Prob}\big\{Z\big(S_0 \cup \{\mathbf{x}\}\big) = 0, Z(S_1) = 1\big\} + \mathrm{Prob}\big\{Z(S_0) = 0, Z\big(S_1 \cup \{\mathbf{x}\}\big) = 1\big\}\\ &= \mathrm{Prob}\big\{Z(S_0) = 0, Z(S_1) = 1\big\}.\end{aligned}$$

As $\lim_{\lambda\to\infty} D_\lambda > 0$, it follows

$$p = \frac{\lim_{\lambda\to\infty} N_\lambda}{\lim_{\lambda\to\infty} D_\lambda} = \mathrm{Prob}\big\{Z(\mathbf{x}) = 1 \mid Z(S_0) = 0, Z(S_1) = 1\big\}.$$

Let S'_0 and S'_1 be the next level sets obtained once $\mathbf{x}$ has been allocated. Note that p can take all values on $[0, 1]$. If $p < 1$, step (iv) of Algorithm 1 shows that $\mathbf{x}$ can be assigned the value 0, in which case $S'_0 = S_0 \cup \{\mathbf{x}\}$ and $S'_1 = S_1$, and one has $\mathrm{Prob}\{Z(S'_0) = 0, Z(S'_1) = 1\} = (1 - p)\beta > 0$. Similarly, if $p > 0$, then $\mathbf{x}$ can be assigned the value 1, in which case $S'_0 = S_0$ and $S'_1 = S_1 \cup \{\mathbf{x}\}$, and one has $\mathrm{Prob}\{Z(S'_0) = 0, Z(S'_1) = 1\} = p\beta > 0$. Consequently, one has $\mathrm{Prob}\{Z(S'_0) = 0, Z(S'_1) = 1\} > 0$ whatever the allocation of $\mathbf{x}$. The induction hypothesis is thus preserved, which proves the correctness of the sequential algorithm for infinite TI's.

Appendix B: Proof of Eqs. (3) and (4)

To calculate $\text{Prob}\{Z(K_0)=0\}$, the starting point is to express that none of the Boolean objects hits K_0

$$\text{Prob}\big\{Z(K_0)=0\big\}=\text{Prob}\big\{\forall\mathbf{u}\in\mathbb{Z}^2,\forall n\le N(\mathbf{u}),\tau_{\mathbf{u}}A_{\mathbf{u},n}\cap K_0=\varnothing\big\}.$$

The right-hand side is now expanded using the fact that the Boolean model is made of independent objects in independent Poisson numbers

$$\begin{aligned}\text{Prob}\big\{Z(K_0)=0\big\}&=\prod_{\mathbf{u}\in\mathbb{Z}^2}\text{Prob}\big\{\forall n\le N(\mathbf{u}),\tau_{\mathbf{u}}A_{\mathbf{u},n}\cap K_0=\varnothing\big\}\\&=\prod_{\mathbf{u}\in\mathbb{Z}^2}\sum_{n=0}^{\infty}\exp(-\theta)\frac{\theta^n}{n!}\big(\text{Prob}\{\tau_{\mathbf{u}}A\cap K_0=\varnothing\}\big)^n\\&=\prod_{\mathbf{u}\in\mathbb{Z}^2}\exp\big(-\theta+\theta\,\text{Prob}\{\tau_{\mathbf{u}}A\cap K_0=\varnothing\}\big)\\&=\prod_{\mathbf{u}\in\mathbb{Z}^2}\exp\big(-\theta\,\text{Prob}\{\tau_{\mathbf{u}}A\cap K_0\neq\varnothing\big).\end{aligned}$$

Moreover, one has $\tau_{\mathbf{u}}A\cap K_0\neq\varnothing$ if and only if $A\cap\tau_{-\mathbf{u}}K_0\neq\varnothing$, that is, $-\mathbf{u}\in\delta_{K_0}A$. Accordingly,

$$\begin{aligned}\text{Prob}\big\{Z(K_0)=0\big\}&=\exp\Big(-\theta\sum_{\mathbf{u}\in\mathbb{Z}^2}\text{Prob}\{-\mathbf{u}\in\delta_{K_0}A\}\Big)\\&=\exp\Big(-\theta\sum_{\mathbf{u}\in\mathbb{Z}^2}E\{1_{-\mathbf{u}\in\delta_{K_0}A}\}\Big)\\&=\exp\Big(-\theta E\Big\{\sum_{\mathbf{u}\in\mathbb{Z}^2}1_{-\mathbf{u}\in\delta_{K_0}A}\Big\}\Big)\\&=\exp\big(-\theta E\{\#\delta_{K_0}A\}\big)\end{aligned}$$

as announced in Eq. (3).

To prove Eq. (4), rewrite the probability as the expectation of an indicator function

$$\begin{aligned}\text{Prob}\big\{Z(K_0)=0,Z(K_1)=1\big\}&=E\Big\{1_{Z(K_0)=0}\prod_{\mathbf{x}_1\in K_1}1_{Z(\mathbf{x}_1)=1}\Big\}\\&=E\Big\{1_{Z(K_0)=0}\prod_{\mathbf{x}_1\in K_1}(1-1_{Z(\mathbf{x}_1)=0})\Big\}.\end{aligned}$$

By expanding, one obtains

$$\text{Prob}\big\{Z(K_0)=0,Z(K_1)=1\big\}=E\Big\{1_{Z(K_0)=0}\sum_{L\subset K_1}(-1)^{\#L}1_{Z(L)=0}\Big\}$$

$$= \sum_{L\subset K_1} (-1)^{\#L} E\{1_{Z(K_0\cup L)=0}\}$$

$$= \sum_{L\subset K_1} (-1)^{\#L} \operatorname{Prob}\big\{Z(K_0\cup L)=0\big\}.$$

Equation (4) is derived by replacing $\operatorname{Prob}\{Z(K_0\cup L)=0\}$ by its expression given in Eq. (3).

References

Arpat G (2005) Sequential simulation with patterns. PhD dissertation, Stanford University

Chilès J, Delfiner P (2012) Geostatistics. Modeling spatial uncertainty. Wiley, New York

Eskandaridalvand K, Srinivasan S (2010) Reservoir modelling of complex geological systems—a multiple-point perspective. J Can Pet Technol 49(8):59–69

Guardiano F, Srivastava R (1993) Multivariate geostatistics: beyond bivariate moments. In: Soares A (ed) Geostatistics Tróia'92. Kluwer, Dordrecht, pp 133–144

Holden L (2006) Markov random fields and multipoint statistics. In: Proceedings of the 10th European conference on the mathematics of oil recovery. European Association of Geoscientists & Engineers, Amsterdam

Hu L, Chugunova T (2008) Multiple-point geostatistics for modeling subsurface heterogeneity. Water Resour Res 44:W11413

Journel A, Zhang T (2006) The necessity of a multiple-point prior model. Math Geol 38(5):591–610

Lantuéjoul C (1991) Ergodicity and integral range. J Microsc 161(3):387–404

Lantuéjoul C (2002) Geostatistical simulation: models and algorithms. Springer, Berlin

Liu Y (2005) An information content measure using multiple-point statistics. In: Leuangthong O, Deutsch C (eds) Geostatistics Banff 2004, vol 2. Springer, Dordrecht, pp 1047–1054

Mariethoz G, Renard P, Straubhaar J (2010) Direct sampling method to perform multiple-point geostatistical simulation. Water Resour Res 46(1):1–22

Matheron G (1971) The theory of regionalized variables and its applications. Ecole des Mines, Paris

Matheron G (1975) Random sets and integral geometry. Wiley, New York

Matheron G (1989) Estimating and choosing. Springer, Berlin

Ortiz J (2008) An overview of the challenges of multiple-point geostatistics. In: Ortiz J, Emery X (eds) Proceedings of the eighth international geostatistics congress, vol 1. Gecamin, Santiago, pp 11–20

Renard P, Straubhaar J, Caers J, Mariethoz G (2011) Conditioning facies simulations with connectivity data. Math Geosci 43(8):879–903

Serra J (1982) Image analysis and mathematical morphology. Academic Press, London

Strebelle S (2002) Conditional simulation of complex geological structures using multiple-point statistics. Math Geol 34(1):1–22

Strebelle S, Remy N (2005) Post-processing of multiple-point geostatistical models to improve reproduction of training patterns. In: Leuangthong O, Deutsch C (eds) Geostatistics Banff 2004, vol 2. Springer, Dordrecht, pp 979–988

Strebelle S, Zhang T (2005) Non-stationary multiple-point geostatistical models. In: Leuangthong O, Deutsch C (eds) Geostatistics Banff 2004, vol 1. Springer, Dordrecht, pp 235–244