Can a training image be a substitute for a random field model?

X. EMERY[1], C. LANTUÉJOUL[2]

[1] *University of Chile, Santiago, Chile*
[2] *MinesParisTech, Fontainebleau, France*

[1] *xemery@ing.uchile.cl*
[2] *christian.lantuejoul@mines-paristech.fr*

Introduction

Modern stochastic data assimilation algorithms may require generating ensembles of facies fields. This is typically the case in reservoir optimization where each facies field is used as input for a fluid flow exercise.

In a geostatistical context, facies fields are nothing but conditional simulations. Different approaches can be considered to produce them:

– By resorting to a spatial stochastic model such as the plurigaussian model, the Boolean model... This requires the choice of a model, the statistical inference of its parameters, the design of a conditional simulation algorithm...

– By resorting to a training image to produce multipoint simulations (MPS): no statistical inference, wide generality, conceptual simplicity...

The second approach looks miraculous. Isn't there a price to pay for it?

Outline

Compatibility between MPS's and stochastic simulations

- Principle of MPS
- Case of an infinite training image
- Case of a finite training image

Statistical considerations on template matching

- Statistical matching of a template
- Application to the estimation of the size of a training image
- Example
- A simple combinatorial remark

Compatibility between MPS's and stochastic simulations

Principle of MPS

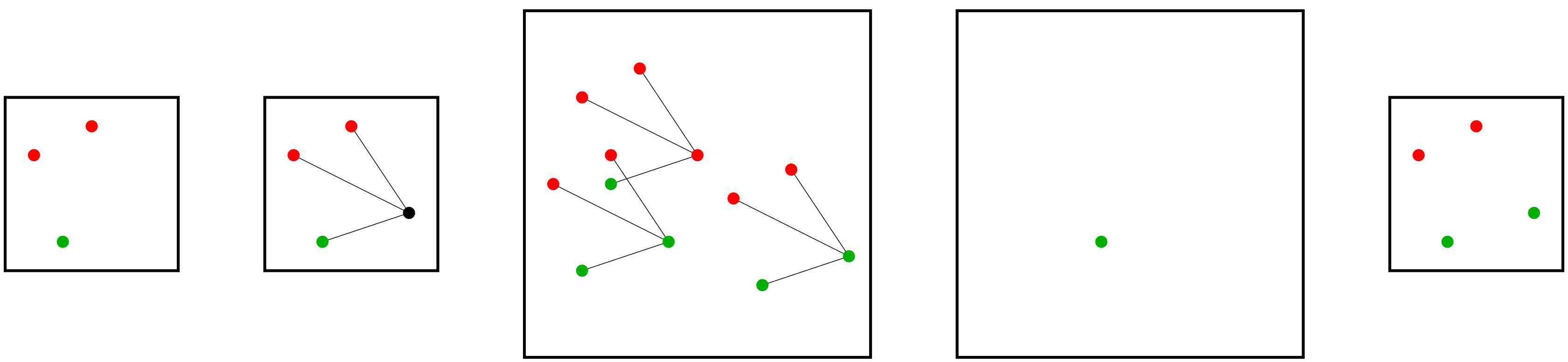
This is a sequential algorithm. Each step is as follows:

(i) a new target point is selected at random in the simulation field. It defines a template along with the already processed points;

(ii) the pixels where the template matches the training image are identified;

(iii) one pixel among those is selected at random;

(iv) its value is assigned to the target point.



(i) *(ii)* *(iii)* *(iv)*

The problem addressed

Assumption:

Suppose that the training image I is a realization, or part of a realization, of some stationary, ergodic random field (SERF) Z on $\mathbb{Z}^2$.

Z is ergodic means that its spatial distribution can be retrieved from any of its realizations:

$$P\{\cap_{i=1,n} Z(x_i) = \epsilon_i\} = \lim_{S \longrightarrow \mathbb{Z}^2} \frac{1}{\#S} \sum_{s \in S} \prod_{i=1}^{n} 1_{I(x_i+s)=\epsilon_i}$$

Question:

Does the empirical spatial distribution yielded by MPS's fit that of Z?

Case of an infinite training image

Remark:

The algorithm cannot be directly applied because the template T matches I at infinitely many points (set S_T). The target point is then assigned the value 0 or 1 with respective probabilities

$$p_0 = \lim_{S \longrightarrow \mathbb{Z}^2} \frac{1}{\#S} \sum_{s \in S \cap S_T} 1_{I(s)=0} \qquad p_1 = \lim_{S \longrightarrow \mathbb{Z}^2} \frac{1}{\#S} \sum_{s \in S \cap S_T} 1_{I(s)=1}$$

Results:

– Each MPS is a patch of the TI;

– The empirical spatial distribution fits that of Z:

If $\left(X_k, k \geq 1\right)$ is a sequence of MPS's on domain D, if $x_1, \ldots x_n \in D$ and if $\epsilon_1, \ldots, \epsilon_n \in \{0, 1\}$, then

$$= \lim_{k \longrightarrow \infty} \frac{1}{k} \sum_{\ell=1}^{k} \prod_{i=1}^{n} 1_{X_\ell(x_i)=\epsilon_i} = P\left\{\cap_{i=1,n} Z(x_i) = \epsilon_i\right\}$$

– Conditional MPS can be performed as well.

Case of a finite training image

Uncommon situation:

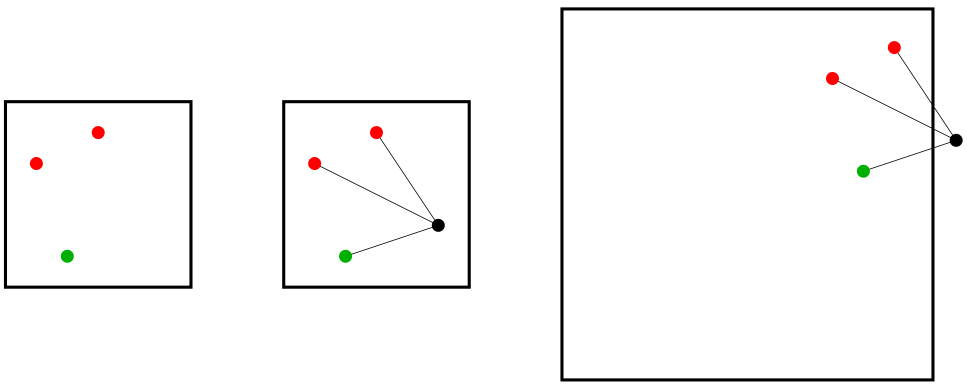
The algorithm runs till a MPS has been completed:

– Then the MPS a patch of the training image;

– Different MPS's display little variability (the training image has less variability than an entire realization, possible overlaps between MPS's).

Common situation:

The algorithm stops at one step because the training image does not match the template at any location:



How to prevent the algorithm from stopping?

Reduce the size of the template

– By discarding points of a template, spurious conditional independence relationships are introduced (Holden, 2006);

– Because of the sequential nature of the algorithm, these relationships propagate, which may lead to severe artefacts to the final outcome (Arpat, 2005).

Increase the size of the training image

– MPS algorithms works for infinitely large images

– Accordingly, it should also work provided that the training image is large enough...

Statistical considerations on template matching

Statistical matching of a template

Notation:

– Z is a binary, stationary, ergodic random field (SERF) on $\mathbb{Z}^2$;

– T is a template.

Matching:

Let $N_T(x) = 1$ if the template located at x matches Z, and 0 otherwise. N_T is also a SERF. Its mean, variance and correlation function are respectively denoted by μ_T, $\sigma_T^2 = \mu_T(1 - \mu_T)$ and ρ_T.

Matching number:

More generally, the number of times T matches Z in a finite domain V is $N_T(V) = \sum_{x \in V} N_T(x)$. We have ($\tau_h$ is the translation by vector $\vec{oh}$)

$$E\{N_T(V)\} = \mu_T \, \#V$$

$$Var\{N_T(V)\} = \sigma_T^2 \sum_{h \in \mathbb{Z}^2} \rho_T(h) \, \#(V \cap \tau_h V)$$

An asymptotic result

Heuristic approach:

$$Var\{N_T(V)\} = \sigma_T^2 \sum_{h \in \mathbb{Z}^2} \rho_T(h) \, \#(V \cap \tau_h V)$$

If the range of ρ_T is small compared to the size of V, then one heuristically has $\#(V \cap \tau_h V) \approx \#V$ whenever $\rho_T \not\approx 0$, which implies

$$Var\{N_T(V)\} \approx \sigma_T^2 \sum_{h \in \mathbb{Z}^2} \rho_T(h) \, \#V$$

Definition:

The integral $a_T = \sum_{h \in Z^2} \rho_T(h)$ of the correlation function of Z_T is called the integral range of Z_T. This is a dimensionless quantity that satisfies $0 \leq a_T \leq \infty$.

Property:

If $0 < a_T < \infty$, and if $\#V \gg a_T$, then $N_T(V)$ is approximately Gaussianly distributed with mean $\#V \mu_T$ and variance $\sigma_T^2 a_T \, \#V$

Application to the choice of V

Put $N_T(V) \approx \#V\mu_T + \sigma_T\sqrt{\#V a_T}\,Y$, where Y is a standard Gaussian variable. Accordingly, we have

$$P\{N_T(V) \geq n\} \geq 1-\alpha \quad \Longleftrightarrow \quad P\left\{Y \geq \frac{n - \#V\mu_T}{\sigma_T\sqrt{\#V a_T}}\right\} \geq 1-\alpha$$

Denoting by $y_{1-\alpha}$ the quantile of order $1-\alpha$ of Y, the latter condition will be satisfied as soon as

$$\frac{n - \#V\mu_T}{\sigma_T\sqrt{\#V a_T}} \leq y_{1-\alpha},$$

which yields

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

The right handside member is a decreasing function of μ_T and an increasing function of a_T.

Example: the discrete Boolean model

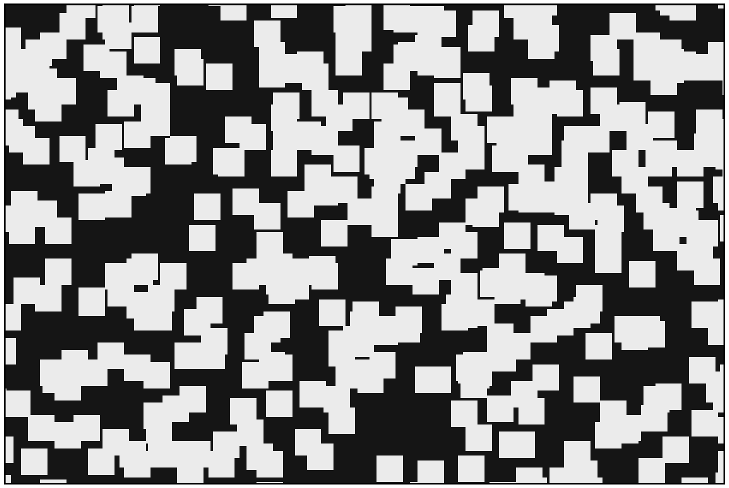
Ingredients:

– Independent Poisson variables $(N(u), u \in \mathbb{Z}^2)$ (mean value θ);

– Independent copies $\big(A_{u,n}, u \in \mathbb{Z}^2, n \leq N(u)\big)$ of a random object A.

Definition:

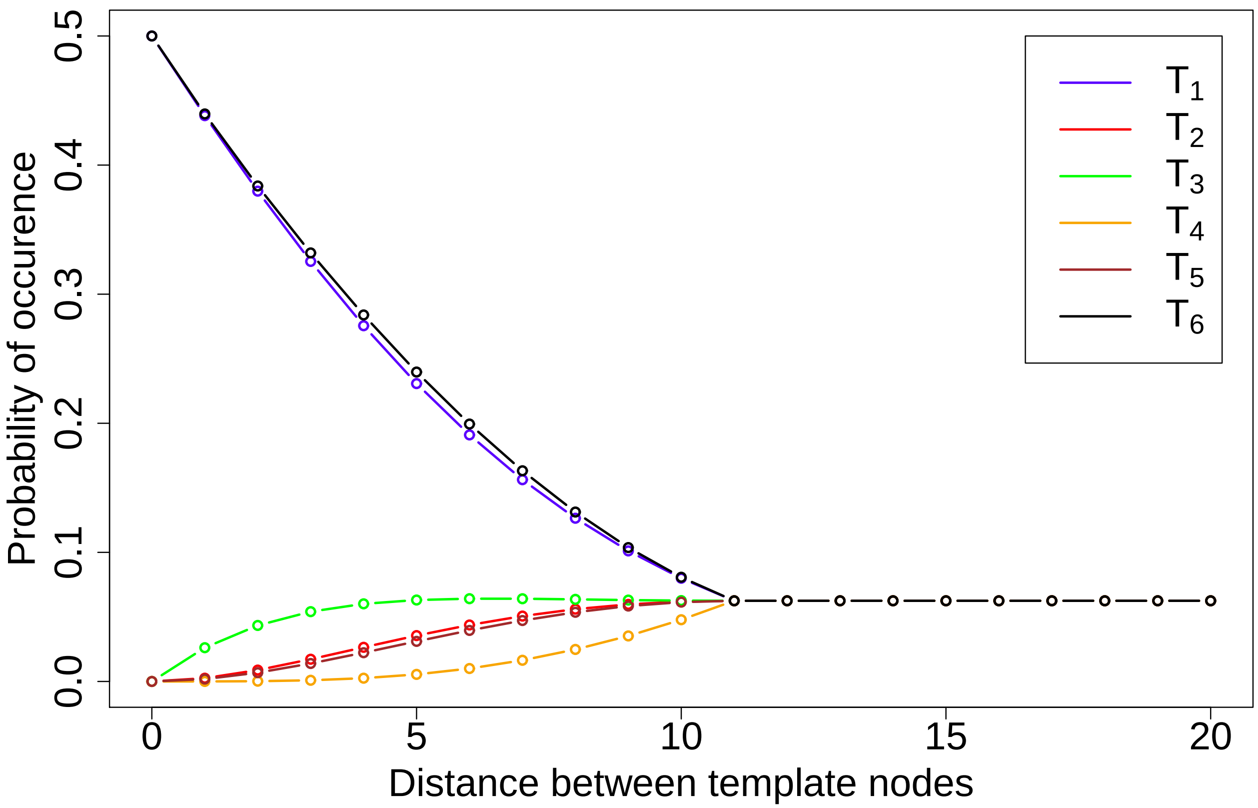
$$Z(x) = \max_{u \in \mathbb{Z}^2} 1_{x \in \tau_u A_u} \qquad A_u = \cup_{n \leq N(u)} A_{u,n}$$



Boolean model of squares of side 11. $\theta = 0.0057$ yields 50% zero proportion.

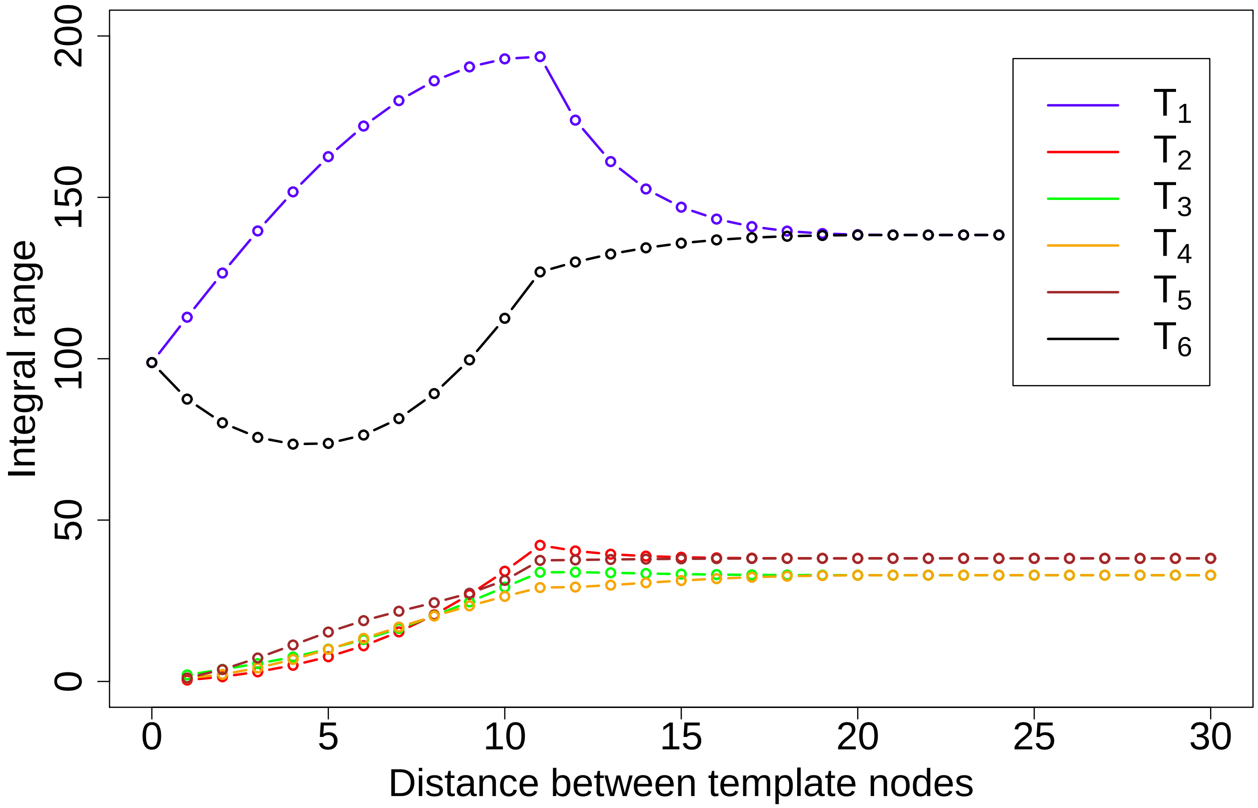
Probability of matching

$T_1 = \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix}$ $T_2 = \begin{matrix} 1 & 0 \\ 0 & 0 \end{matrix}$ $T_3 = \begin{matrix} 1 & 1 \\ 0 & 0 \end{matrix}$ $T_4 = \begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}$ $T_5 = \begin{matrix} 1 & 1 \\ 0 & 1 \end{matrix}$ $T_6 = \begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}$



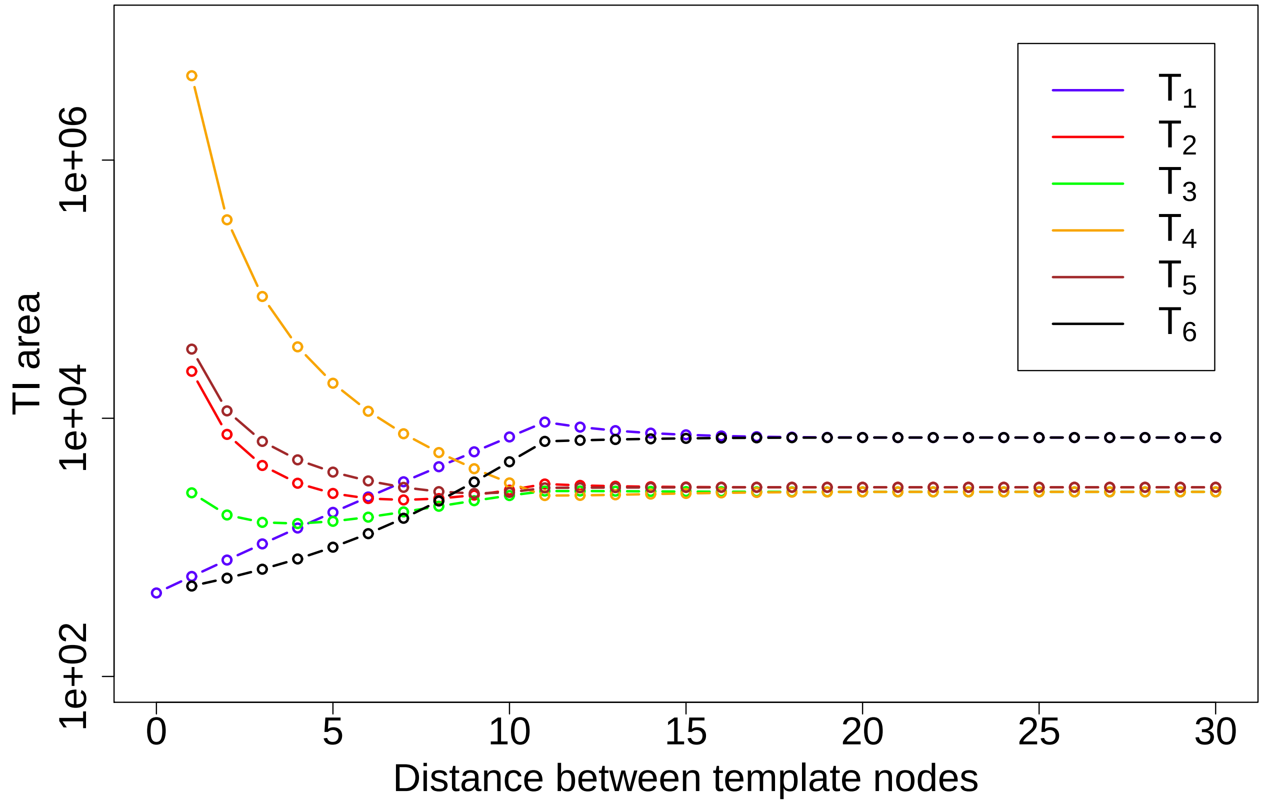
Integral range

$T_1 = \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix}$ $T_2 = \begin{matrix} 1 & 0 \\ 0 & 0 \end{matrix}$ $T_3 = \begin{matrix} 1 & 1 \\ 0 & 0 \end{matrix}$ $T_4 = \begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}$ $T_5 = \begin{matrix} 1 & 1 \\ 0 & 1 \end{matrix}$ $T_6 = \begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}$



Required area for 50 matchings in 95% cases

$T_1 = \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix}$ $T_2 = \begin{matrix} 1 & 0 \\ 0 & 0 \end{matrix}$ $T_3 = \begin{matrix} 1 & 1 \\ 0 & 0 \end{matrix}$ $T_4 = \begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}$ $T_5 = \begin{matrix} 1 & 1 \\ 0 & 1 \end{matrix}$ $T_6 = \begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}$



A simple combinatorial remark

Assumptions:

– The training image is a square of n^2 pixels;

– The population of templates considered have the same support of k pixels.

Counting:

– The total number of templates of the population is 2^k.

– The training image contains at most n^2 different templates of the population (independent of k!);

Conclusion:

– The proportion of templates present in the training image is at most $n^2/2^k$.

– To give an order of magnitude, $n = 10,000$ and $k = 100$ (square 10×10) yields an upper bound of 8×10^{-23} for the proportion, that is close to the reciprocal of the Avogadro number...