Substitution Random Fields with Gaussian and Gamma Distributions: Theory and Application to a Pollution Data Set

Xavier Emery

Abstract This paper presents random field models with Gaussian or gamma univariate distributions and isofactorial bivariate distributions, constructed by composing two independent random fields: a directing function with stationary Gaussian increments and a stationary coding process with bivariate Gaussian or gamma distributions. Two variations are proposed, by considering a multivariate directing function and a coding process with a separable covariance, or by including drift components in the directing function. Iterative algorithms based on the Gibbs sampler allow one to condition the realizations of the substitution random fields to a set of data, while the inference of the model parameters relies on simple tools such as indicator variograms and variograms of different orders. A case study in polluted soil management is presented, for which a gamma model is used to quantify the risk that pollutant concentrations over remediation units exceed a given toxicity level. Unlike the multivariate Gaussian model, the proposed gamma model accounts for an asymmetry in the spatial correlation of the indicator functions around the median and for a spatial clustering of high pollutant concentrations.

Keywords Conditional simulation · Isofactorial bivariate distribution · Bivariate Gaussian distribution · Bivariate gamma distribution · Gibbs sampler

1 Introduction

An important aspect in the analysis of regionalized variables is the modeling of local uncertainty and its incorporation in decision-making processes. In polluted site management, the planner is interested in mapping the probability that the concentration of a pollutant exceeds a regulatory threshold, given the information available at data locations. This problem can be solved by using nonlinear kriging techniques

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

like indicator, disjunctive, or multi-Gaussian kriging (Emery 2006a; Journel 1984; Matheron 1976a; Oliver et al. 1996).

In general, the decision-making process involves a more complex transfer function of the pollutant concentration. In particular, this concentration must be upscaled from the data support to that of remediation units. One possibility is to define a change-of-support model and to combine it with one of the previous nonlinear kriging techniques (Emery and Soto-Torres 2005; Matheron 1976b, 1984). Another possibility is to use conditional simulation, which provides alternative realizations of the pollutant concentration that can be averaged to the support of the remediation units. This approach is flexible as it can be used when the upscaling differs from an arithmetic averaging. For instance, the decision-maker may be concerned with the impact of short-term exposure on human health and be interested in knowing whether the maximal point-support concentration (not the average) within a remediation unit exceeds a given threshold or not.

This article deals with a family of random field models that can be used to simulate pollutant concentrations and, more generally, regionalized variables measured on a continuous quantitative scale. In the following three sections we present the models and investigate their properties. Then we describe tools to infer and validate the model parameters from a set of data. In the last section, the concepts are illustrated through a case study in polluted soil management.

2 Substitution Random Fields

2.1 Definition and Properties

A random field on $\mathbb{R}^d$, $Y = \{Y(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d\}$, is called a substitution random field if it can be written in the following fashion (Lantuéjoul 1991, 1993, 2002)

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad Y(\mathbf{x}) = X\big[T(\mathbf{x})\big], \tag{1}$$

where $T = \{T(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d\}$ is a random field on $\mathbb{R}^d$ called a directing function, and $X = \{X(t), t \in \mathbb{R}\}$ is a random field on $\mathbb{R}$ called a coding process. The univariate and bivariate distributions of Y can be determined under the assumptions that the coding process and directing function are independent; the former is stationary and the latter has stationary increments. If these conditions are satisfied, the following results hold:

(1) X and Y have the same univariate distributions.

(2) If X has a covariance function C_X, then Y has a covariance function C_Y such that (Lantuéjoul 1991, p. 402)

$$\forall \mathbf{h} \in \mathbb{R}^d, \quad \mathrm{C}_Y(\mathbf{h}) = \mathrm{E}\big\{\mathrm{C}_X\big[T(\mathbf{h}) - T(\mathbf{0})\big]\big\}, \tag{2}$$

where $\mathbf{h}$ is the lag separation vector.

(3) Isofactorial permanence: if X has isofactorial bivariate distributions, then so does Y. Moreover, if $\chi_p(X)$ is the pth factor of the bivariate distributions of

X, then $\chi_p(Y)$ is the pth factor of the bivariate distributions of Y. The covariance functions of these factors are linked by the following relationship (Matheron 1989b, p. 313)

$$\forall \mathbf{h} \in \mathbb{R}^d, \quad \mathrm{C}_Y^{(p)}(\mathbf{h}) = \mathrm{E}\big\{\mathrm{C}_X^{(p)}\big[T(\mathbf{h}) - T(\mathbf{0})\big]\big\}. \tag{3}$$

Examples and properties of isofactorial bivariate distributions can be found in the literature (Chilès and Delfiner 1999; Johnson and Kotz 1972; Lancaster 1958). In this article, we focus on two particular examples of such distributions, for which the marginals are Gaussian and gamma, and the factors are Hermite and Laguerre polynomials, respectively. Following Chilès and Delfiner (1999, p. 406) and Wackernagel (2003, p. 254–256), these isofactorial bivariate distributions will be called "Hermitian" and "Laguerre" distributions.

2.2 Conditional Simulation

Consider the problem of simulating a substitution random field and conditioning the realizations to a set of data $\{Y(\mathbf{x}_\alpha) = y_\alpha, \alpha = 1, \ldots, n\}$. This problem can be solved in four steps:

(a) Simulate $\{T(\mathbf{x}_\alpha), \alpha = 1, \ldots, n\}$ conditionally to $\{Y(\mathbf{x}_\alpha) = y_\alpha, \alpha = 1, \ldots, n\}$. Obtain a set of values $\{t_\alpha, \alpha = 1, \ldots, n\}$.
(b) Simulate T conditionally to $\{T(\mathbf{x}_\alpha) = t_\alpha, \alpha = 1, \ldots, n\}$.
(c) Simulate X conditionally to $\{X(t_\alpha) = y_\alpha, \alpha = 1, \ldots, n\}$.
(d) Obtain the simulated random field Y as per (1).

Steps (b) and (c) do not present any difficulty if conditional simulation algorithms are available for both the directing function and the coding process. As for step (a), an iterative algorithm based on the Gibbs sampler can be used. The idea is to start from a non-conditional simulation of the random variables $\{T(\mathbf{x}_\alpha), \alpha = 1, \ldots, n\}$ and to progressively modify the simulation to obtain random variables whose distribution converges to that of $\{T(\mathbf{x}_\alpha), \alpha = 1, \ldots, n\}$ conditional to $\{Y(\mathbf{x}_\alpha) = y_\alpha, \alpha = 1, \ldots, n\}$. Because the convergence is asymptotical, in practice the algorithm requires defining a maximum number of iterations. Theoretical and empirical results on the rate of convergence of the Gibbs sampler and guidelines for choosing the number of iterations have been presented by many authors (Emery 2007; Galli and Gao 2001; Lantuéjoul 2002; Tierney 1994).

Specifically, the following algorithm is proposed (Lantuéjoul 2002, p. 233; Emery 2007):

(a1) Simulate $\{T(\mathbf{x}_\alpha), \alpha = 1, \ldots, n\}$ without conditioning constraints. Obtain a set of values $\{t_\alpha, \alpha = 1, \ldots, n\}$.
(a2) Select an index α_0 at random (uniformly) in $\{1, \ldots, n\}$.
(a3) Simulate $T(\mathbf{x}_{\alpha_0})$ conditionally to $\{T(\mathbf{x}_\alpha) = t_\alpha, \alpha \neq \alpha_0\}$. Obtain a new value t'_{α_0}.
(a4) Calculate the probabilities p_{α_0} and p'_{α_0} of the events $\{X(t_{\alpha_0}) = y_{\alpha_0}\}$ and $\{X(t'_{\alpha_0}) = y_{\alpha_0}\}$ given $\{X(t_\alpha) = y_\alpha, \alpha \neq \alpha_0\}$.
(a5) Generate a uniform value u in [0, 1].
(a6) If $u p_{\alpha_0} \leq p'_{\alpha_0}$, replace t_{α_0} by t'_{α_0} (Metropolis acceptance criterion).
(a7) Go back to (a2) until the maximum number of iterations has been reached.

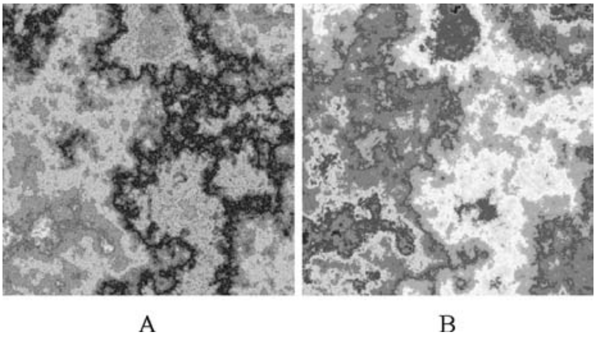

Fig. 1 Partial realizations of substitution random fields with bivariate Hermitian distributions. In every case, the directing function has a linear variogram and stationary Gaussian increments

Step (a4) assumes that the coding process has a discrete distribution. This step can be extended to the case of a continuous distribution by substituting a probability density function for a probability mass function (Tierney 1994). For practical reasons, the coding process is often assumed Markovian, which results in important simplifications in steps (a4) and (c). However, this assumption allows little flexibility in the choice of the covariance function of the substitution random field (2) as the covariance function of a Markovian process is an exponential function. The objectives of the next two sections are to remove the Markovian restriction for the coding process and to broaden the class of substitution random fields that can be simulated.

3 Substitution Random Fields with Bivariate Hermitian Distributions

3.1 Gaussian Coding Process

Let X be a stationary Gaussian random field on $\mathbb{R}$ with covariance function ρ. Its bivariate distributions have an isofactorial representation with the Hermite polynomials as the factors (Lancaster 1957). Due to isofactorial permanence, the substitution random field Y has bivariate Hermitian distributions characterized by the following factor covariance functions (3)

$$\forall \mathbf{h} \in \mathbb{R}^d,\ \forall p \in \mathbb{N}^*,\quad \mathrm{C}_Y^{(p)}(\mathbf{h}) = \mathrm{E}\big\{\rho^p\big[T(\mathbf{h}) - T(\mathbf{0})\big]\big\}. \tag{4}$$

The Markovian restriction is not necessary: the conditional distributions needed in step (a4) are Gaussian, with mean and variance equal to simple kriging predictions and simple kriging variances, respectively. Accordingly, the coding process can have any covariance model (positive semi-definite function), not just an exponential covariance.

3.2 Examples

Henceforth, we assume that the directing function T has stationary Gaussian increments (i.e. increments with multivariate Gaussian distributions) with variogram γ_T. We consider two types of covariance functions for the coding process: (1) an exponential covariance (Fig. 1A) where $\rho(\Delta t) = \exp(-a|\Delta t|)$ with $a > 0$; and (2) a

Table 1 Covariance functions for the factors of the bivariate distributions of the substitution random field. The notation $a \sim b$ means that a/b tends to 1. $G(.)$ represents the standard Gaussian cumulative distribution function, γ_T the variogram of the directing function, and c a positive real number

Covariance function of the coding process	Covariance for the pth factor of the bivariate distributions of the substitution random field	Asymptotic behavior for fixed p
Exponential	$\mathrm{C}_Y^{(p)} = 2\exp(\frac{p^2a^2\gamma_T}{2})G(-pa\sqrt{\gamma_T})$	$\mathrm{C}_Y^{(p)}(\mathbf{h}) \underset{\lvert\mathbf{h}\rvert\to+\infty}{\sim} \frac{c}{\sqrt{\gamma_T(\mathbf{h})}}$
Gaussian	$\mathrm{C}_Y^{(p)} = \frac{1}{\sqrt{1+2pa\gamma_T}}$	$\mathrm{C}_Y^{(p)}(\mathbf{h}) \underset{\lvert\mathbf{h}\rvert\to+\infty}{\sim} \frac{c}{\sqrt{\gamma_T(\mathbf{h})}}$

Gaussian covariance (Fig. 1B) where $\rho(\Delta t) = \exp(-a|\Delta t|^2)$ with $a > 0$. Table 1 gives the covariance functions for the factors of the bivariate distributions of Y; these covariance functions tend to zero if γ_T is unbounded. Even so, they are not integrable on $\mathbb{R}^d$. Their integral ranges are infinite, since the growth rate of γ_T is necessarily less than quadratic.

Having infinite integral ranges restricts the class of available covariance models and affects the properties of the substitution random field. The realizations (Fig. 1) exhibit large-range structures and appear as non-homogeneous at any scale of observation. The notion of integral range is also related to the property of ergodicity (Lantuéjoul 1991), which allows one to infer parameters such as the mean, the variance, or the variogram from a single realization of the random field. When the integral range is infinite, the realization must be known over a very large domain to accurately estimate these parameters. In the following subsections, two variations are proposed to construct models with finite integral ranges.

3.3 First Variation: Multivariate Directing Function

(1) Consider a multivariate directing function $\boldsymbol{T}$ with N mutually independent components, each of them with stationary Gaussian increments and variogram γ_T

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad \boldsymbol{T}(\mathbf{x}) = \big(T_1(\mathbf{x}), \ldots, T_N(\mathbf{x})\big) \tag{5}$$

and a stationary Gaussian coding process in $\mathbb{R}^N$, $X = \{X(\boldsymbol{t}), \boldsymbol{t} \in \mathbb{R}^N\}$, with a separable covariance

$$\mathrm{C}_X(\Delta\boldsymbol{t}) = \mathrm{C}_X(\Delta t_1, \ldots, \Delta t_N) = \prod_{i=1}^{N} \rho(\Delta t_i). \tag{6}$$

Since the coding process has bivariate Gaussian distributions, the substitution random field Y defined by putting $Y(\mathbf{x}) = X[\boldsymbol{T}(\mathbf{x})]$ for every $\mathbf{x}$ in $\mathbb{R}^d$ has bivariate Hermitian distributions (isofactorial permanence). The covariance function of the pth factor is (3)

$$\forall \mathbf{h} \in \mathbb{R}^d, \quad \mathrm{C}_Y^{(p)}(\mathbf{h}) = \mathrm{E}\left\{\prod_{i=1}^{N} \rho^p\big[T_i(\mathbf{h}) - T_i(\mathbf{0})\big]\right\} = \prod_{i=1}^{N} \mathrm{E}\big\{\rho^p\big[T_i(\mathbf{h}) - T_i(\mathbf{0})\big]\big\}$$

$$= \mathrm{E}\big\{\rho^p\big[T_1(\mathbf{h}) - T_1(\mathbf{0})\big]\big\}^N. \tag{7}$$

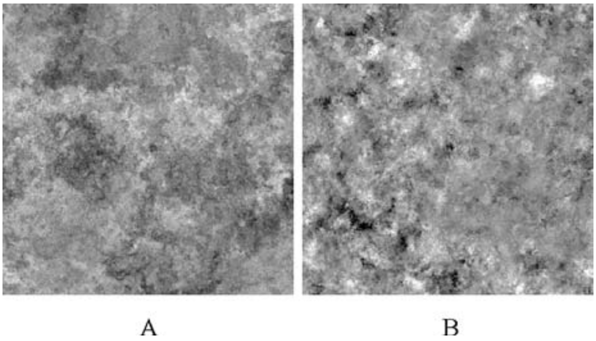

Fig. 2 Partial realizations of substitution random fields with bivariate Hermitian distributions, obtained by using a directing function with **A**, multiple components, and **B**, drift components

The second equality in (7) is justified by the independence of the components of the directing function. The factor covariance functions in (7) are simply the Nth powers of those in (4), so their integral ranges may be finite if N is large enough. Here, there are three free parameters that facilitate the modeling: (1) the variogram γ_T of the directing function; (2) the covariance ρ that characterizes the coding process; and (3) the number N of components of the directing function. Figure 2A shows a realization of a substitution random field constructed by considering a quadrivariate directing function with a linear variogram and a coding process with a separable exponential covariance.

3.4 Second Variation: Directing Function with Drift Components

First, consider a multivariate directing function $\boldsymbol{T}$ with $N + d$ components such that the first N components are mutually independent random fields with stationary Gaussian increments and variogram γ_T; and the last d components correspond to drift terms of the form

$$\forall i \in \{1, \ldots, d\},\ \forall \mathbf{x} \in \mathbb{R}^d, \quad T_{N+i}(\mathbf{x}) = \varepsilon_i b x_i \tag{8}$$

with $\varepsilon_i = 1$ or -1 with equal probability 0.5, b is a nonnegative real number, and x_i is the ith coordinate of vector $\mathbf{x}$.

Second, consider a stationary Gaussian coding process $X = \{X(\boldsymbol{t}), \boldsymbol{t} \in \mathbb{R}^{N+d}\}$ whose covariance is the product of a separable covariance on $\mathbb{R}^N$ and an isotropic covariance on $\mathbb{R}^d$

$$\mathrm{C}_X(\Delta \boldsymbol{t}) = \mathrm{C}_X(\Delta t_1, \ldots, \Delta t_{N+d}) = \prod_{i=1}^{N} \rho(\Delta t_i) \times \rho'\Big(\sqrt{\Delta t_{N+1}^2 + \cdots + \Delta t_{N+d}^2}\Big). \tag{9}$$

The resulting substitution random field has bivariate Hermitian distributions, characterized by the following factor covariances

$$\forall \mathbf{h} \in \mathbb{R}^d,\ \forall p \in \mathbb{N}^*, \quad \mathrm{C}_Y^{(p)}(\mathbf{h}) = \mathrm{E}\big\{\rho^p\big[T_1(\mathbf{h}) - T_1(\mathbf{0})\big]\big\}^N \times \rho'^p\big(b\|\mathbf{h}\|\big). \tag{10}$$

To obtain factor covariances with finite integral ranges, it is sufficient that ρ' has a finite range and that b is not zero. A realization of the proposed random field model is shown in Fig. 2B, where the covariance functions ρ and ρ' are exponential and spherical.

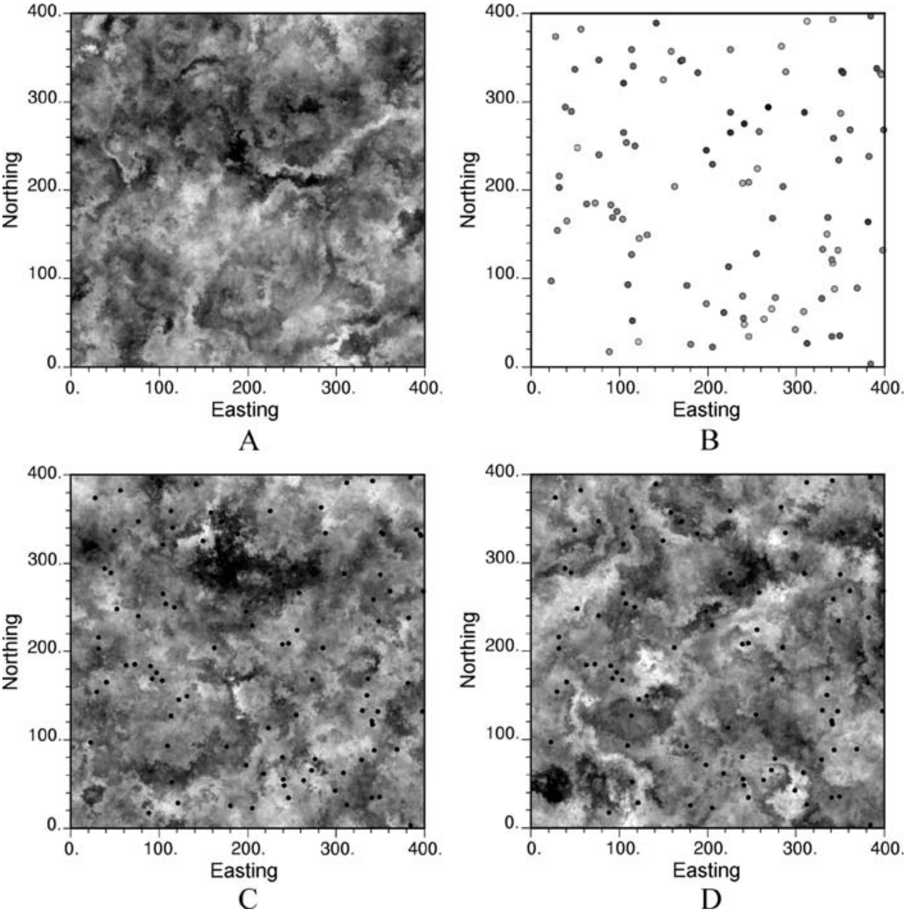

Fig. 3 **A**, non-conditional realization of a substitution random field with bivariate Hermitian distributions, **B**, conditioning data points, **C** and **D**, two conditioned realizations (the *conditioning points* are superimposed)

3.5 Conditional Simulation

The algorithm presented in the previous section (steps (a) to (d)) can be used to condition the realizations to a set of data. Concerning step (a), one has to work with vectorial components $\{\boldsymbol{t}_\alpha, \alpha = 1, \dots, n\}$ instead of scalar components. As an illustration, a non-conditional realization is generated on a 400 × 400 grid (Fig. 3A), with the following parameters

$$\begin{cases} N = 1, & \\ b = 0.01, & \\ \forall \mathbf{h} \in \mathbb{R}^d, & \gamma_T(\mathbf{h}) = \|\mathbf{h}\|, \\ \forall \Delta t \geq 0, & \rho(\Delta t) = \exp\bigl(-0.1 \Delta t^2\bigr), \\ \forall \Delta t \geq 0, & \rho'(\Delta t) = 1 - 1.5 \min(\Delta t, 1) + 0.5 \min\bigl(\Delta t^3, 1\bigr). \end{cases} \tag{11}$$

One hundred locations are then selected at random (uniformly) from the 160,000 grid nodes (Fig. 3B). The values at these locations are used as conditioning data for two new realizations (Fig. 3C and D). Note that the realizations of the substitution random field show well-structured patterns, in particular a spatial clustering of the extreme (high or low) values. These patterns are very different from those observed on realizations of stationary Gaussian random fields, for which the extreme values tend to be scattered in space (Goovaerts 1997, p. 278).

4 Substitution Random Fields with Bivariate Laguerre Distributions

4.1 Gamma Coding Process

First, consider a multivariate directing function $\boldsymbol{T} = \{\boldsymbol{T}(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d\}$ with $N + d$ components such that the first N components are mutually independent random fields with stationary Gaussian increments and variogram γ_T; the last d components are drift terms defined as per (8). Second, consider a vectorial standard Gaussian random field $\boldsymbol{U} = \{\boldsymbol{U}(\boldsymbol{t}), \boldsymbol{t} \in \mathbb{R}^{N+d}\}$ with m stationary mutually independent components, each with a covariance of the form

$$\mathrm{C}_U(\Delta\boldsymbol{t}) = \prod_{i=1}^{N} r(\Delta t_i) \times r'\Bigl(\sqrt{\Delta t_{N+1}^2 + \cdots + \Delta t_{N+d}^2}\Bigr), \tag{12}$$

where r is a covariance function on $\mathbb{R}$ and r' an isotropic covariance function on $\mathbb{R}^d$. From this vectorial Gaussian field, one can define a coding process X with a standard gamma univariate distribution with shape parameter $m/2$, by putting

$$\forall \boldsymbol{t} \in \mathbb{R}^{N+d}, \quad X(\boldsymbol{t}) = \frac{1}{2}\bigl\|\boldsymbol{U}(\boldsymbol{t})\bigr\|^2 = \frac{1}{2}\sum_{j=1}^{m} U_j^2(\boldsymbol{t}). \tag{13}$$

The correlation function of the gamma coding process is given by (9), with $\rho = r^2$ and $\rho' = r'^2$. It can be shown (Emery 2005b) that the bivariate distributions of this process have an isofactorial representation with Laguerre polynomials as the factors.

The substitution random field Y is defined as follows

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad Y(\mathbf{x}) = \frac{1}{2}\bigl\|\boldsymbol{U}\bigl[\boldsymbol{T}(\mathbf{x})\bigr]\bigr\|^2. \tag{14}$$

Because of isofactorial permanence, Y has bivariate Laguerre distributions. The covariance functions of the factors are still given by (10).

4.2 Conditional Simulation

Let $\{\mathbf{x}_\alpha, \alpha = 1, \ldots, n\}$ be the locations of the conditioning data and $\{y_\alpha, \alpha = 1, \ldots, n\}$ the values of the substitution random field observed at these locations. We suggest the following algorithm for simulating Y:

(a) Simulate $\{\boldsymbol{T}(\mathbf{x}_\alpha), \alpha = 1, \ldots, n\}$ and $\{\boldsymbol{U}[\boldsymbol{T}(\mathbf{x}_\alpha)], \alpha = 1, \ldots, n\}$ conditionally to $\{Y(\mathbf{x}_\alpha) = y_\alpha, \alpha = 1, \ldots, n\}$. Obtain two sets of vectors $\{\boldsymbol{t}_\alpha, \alpha = 1, \ldots, n\}$ and $\{\boldsymbol{u}_\alpha, \alpha = 1, \ldots, n\}$.
(b) Simulate $\boldsymbol{T}$ conditionally to $\{\boldsymbol{T}(\mathbf{x}_\alpha) = \boldsymbol{t}_\alpha, \alpha = 1, \ldots, n\}$.
(c) Simulate $\boldsymbol{U}$ conditionally to $\{\boldsymbol{U}(\boldsymbol{t}_\alpha) = \boldsymbol{u}_\alpha, \alpha = 1, \ldots, n\}$.
(d) Obtain the substitution random field as per (14).

Concerning step (a), the following iterative algorithm is proposed:

(a1) Simulate $\{\boldsymbol{T}(\mathbf{x}_\alpha), \alpha = 1, \ldots, n\}$ without conditioning constraints. Obtain a set of values $\{\boldsymbol{t}_\alpha, \alpha = 1, \ldots, n\}$.
(a2) Initialize $\{\boldsymbol{U}[\boldsymbol{T}(\mathbf{x}_\alpha)], \alpha = 1, \ldots, n\}$, by putting

$$\forall \alpha \in \{1, \ldots, n\}, \quad \boldsymbol{u}_\alpha = \sqrt{\frac{2y_\alpha}{m}}\boldsymbol{\varepsilon}_\alpha, \tag{15}$$

where $\{\boldsymbol{\varepsilon}_\alpha, \alpha = 1, \ldots, n\}$ are mutually independent random vectors, each with m independent components taking 1 or -1 with equal probability.
(a3) Select an index α_0 at random (uniformly) in $\{1, \ldots, n\}$.
(a4) Generate a uniform value v in [0, 1].
(a5) If $v > 0.5$, simulate $\boldsymbol{T}(\mathbf{x}_{\alpha_0})$ conditionally to $\{\boldsymbol{T}(\mathbf{x}_\alpha) = \boldsymbol{t}_\alpha, \alpha \neq \alpha_0\}$, obtain a new value $\boldsymbol{t}'_{\alpha_0}$ and put $\boldsymbol{u}'_{\alpha_0} = \boldsymbol{u}_{\alpha_0}$. If $v \leq 0.5$, put $\boldsymbol{t}'_{\alpha_0} = \boldsymbol{t}_{\alpha_0}$ and generate a vector $\boldsymbol{u}'_{\alpha_0}$ uniform on the hypersphere of $\mathbb{R}^m$ centered on $\mathbf{0}$ with squared radius $2y_\alpha$. The threshold 0.5 has been chosen so that there is a 50% chance of updating a $\boldsymbol{T}$-component and 50% chance of updating a $\boldsymbol{U}$-component.
(a6) Calculate the probability densities of $\boldsymbol{U}(\boldsymbol{t}_{\alpha_0})$ and $\boldsymbol{U}(\boldsymbol{t}'_{\alpha_0})$ conditionally to $\{\boldsymbol{U}(\boldsymbol{t}_\alpha) = \boldsymbol{u}_\alpha, \alpha \neq \alpha_0\}$

$$\begin{aligned} p_{\alpha_0} &= \frac{1}{(2\pi)^{m/2}\sigma_{\mathrm{SK}}^m} \exp\left\{-\frac{1}{2\sigma_{\mathrm{SK}}^2}\|\boldsymbol{u}_{\alpha_0} - \boldsymbol{u}_{\mathrm{SK}}\|^2\right\}, \\ p'_{\alpha_0} &= \frac{1}{(2\pi)^{m/2}\sigma_{\mathrm{SK}}^{\prime m}} \exp\left\{-\frac{1}{2\sigma_{\mathrm{SK}}^{\prime 2}}\|\boldsymbol{u}'_{\alpha_0} - \boldsymbol{u}'_{\mathrm{SK}}\|^2\right\}, \end{aligned} \tag{16}$$

where $\boldsymbol{u}_{\mathrm{SK}}$ and $\boldsymbol{u}'_{\mathrm{SK}}$ are the simple kriging predictions of $\boldsymbol{U}(\boldsymbol{t}_{\alpha_0})$ and $\boldsymbol{U}(\boldsymbol{t}'_{\alpha_0})$ from $\{\boldsymbol{U}(\boldsymbol{t}_\alpha) = \boldsymbol{u}_\alpha, \alpha \neq \alpha_0\}$, and σ_{SK} and σ'_{SK} are the corresponding kriging standard deviations.
(a7) Generate a uniform value u in [0, 1].
(a8) If $up_{\alpha_0} \leq p'_{\alpha_0}$, replace the former vectors $(\boldsymbol{t}_{\alpha_0}, \boldsymbol{u}_{\alpha_0})$ by the new vectors $(\boldsymbol{t}'_{\alpha_0}, \boldsymbol{u}'_{\alpha_0})$.
(a9) Go back to (a3) until the maximal number of iterations has been reached.

Note that, at each iteration, one has $\|\boldsymbol{u}_\alpha\|^2 = 2y_\alpha$ for any α in $\{1, \ldots, n\}$. Even if the Gibbs sampler is stopped before convergence, (14) will therefore be satisfied at the data locations $\{\mathbf{x}_\alpha, \alpha = 1, \ldots, n\}$.

5 Parameter Inference and Validation

In general, the available data do not have a Gaussian or gamma univariate distribution, so that a transformation is required to turn these data into normal or gamma scores (Chilès and Delfiner 1999, p. 406). The parameters of the random field model are then chosen in order to fit the empirical bivariate distributions of the transformed data.

5.1 Inference

The general substitution model (with multivariate directing function and drift components) is characterized by the following parameters: the covariance functions ρ and ρ' that define the coding process, the scalar parameters b and N, and the variogram γ_T.

The following trial-and-error strategy is proposed to infer the model parameters:

(1) Turn the original data into standard Gaussian or gamma data. A declustering technique may be needed if the data locations are not regularly spaced in $\mathbb{R}^d$. The gamma transformation requires choosing a half-integer shape parameter $m/2$, which gives the number of components of the Gaussian field $\boldsymbol{U}$ (13).

(2) Choose a set of parameters (b, N, γ_T, ρ, and ρ') so as to fit the covariance of the transformed data or, equivalently, the covariance of the first-order factor [(10) with $p = 1$]. To ease the simulation of the directing function, it is convenient to use a linear or a power variogram model for γ_T, for which many simulation algorithms are available (Chilès and Delfiner 1999; Emery and Lantuéjoul 2006). The slope of γ_T may depend on the direction under consideration if the spatial correlation of the transformed data is not isotropic.

(3) Check whether the bivariate distributions are properly fitted or not (see next subsection). If so, accept the model. Otherwise, go back to step (2) or change the target univariate distribution for data transformation (step (1)).

5.2 Validation of the Bivariate Distribution Model

Common methods to validate a bivariate distribution model rely on the analysis of indicator variograms and variograms of different orders (Emery 2005a; Goovaerts 1997).

5.2.1 Indicator Variograms

An indicator is a binary variable defined in relation to a threshold value. Its variogram can be determined by an expansion into the factor covariances (10) and computed numerically by truncating this expansion at a high order (Emery 2006b). The validation procedure consists in comparing the theoretically expected indicator variograms for several thresholds with the corresponding sample variograms.

5.2.2 Variograms of Order ω

The variogram of order ω (with $\omega > 0$) is defined as follows (Matheron 1989a, p. 30)

$$\forall \mathbf{h} \in \mathbb{R}^d, \quad \gamma_\omega^{(Y)}(\mathbf{h}) = \frac{1}{2}\mathrm{E}\{|Y(\mathbf{x}+\mathbf{h}) - Y(\mathbf{x})|^\omega\}. \tag{17}$$

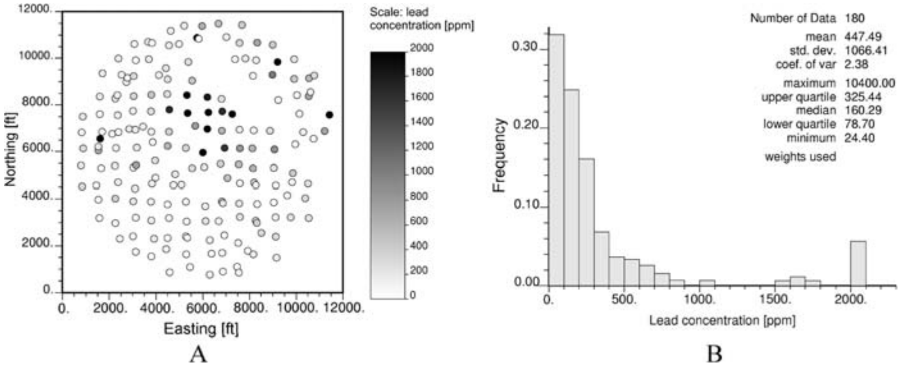

Fig. 4 **A**, data locations, and **B**, declustered lead concentration histogram. The *last bar* of the histogram represents concentrations greater than 2000 ppm

For a random field with bivariate Hermitian or Laguerre distributions characterized by the factor covariances $\{\mathrm{C}_Y^{(p)}, p \in \mathbb{N}^*\}$, one has (Emery 2005a, 2005c)

$$\gamma_\omega^{(Y)}(\mathbf{h}) \propto \left\{1 + \sum_{p\geq 1} \frac{\Gamma(p-\omega/2)}{p!\Gamma(-\omega/2)} \mathrm{C}_Y^{(p)}(\mathbf{h})\right\}. \qquad (18)$$

In practice, this formula can be approximated by truncating the expansion at a high order.

6 Application to a Pollution Data Set

In this section, the substitution model is used to characterize the concentration of a heavy metal in the topsoil of a smelter site and to determine which sub-areas of this site should be remediated.

6.1 Presentation of the Data Set

The case study deals with a soil sampling campaign performed at a smelter site in Dallas, Texas, under guidance of the U.S. Environmental Protection Agency. The lead concentration has been measured on a set of 180 soil samples located at a quasi-regular spacing of 750 feet (Fig. 4A). The area of interest is a one-mile radius circle centered on the smelter location. The lead concentration histogram shows a highly positively skewed distribution, with more than 70% of the data below 300 ppm and less than 6% above 2000 ppm (Fig. 4B). The maximum measured concentration is 10,400 ppm and is located in a junkyard on the eastern part of the area under study. This database has been documented by Isaaks (1984).

6.2 Choice of the Model and Definition of its Parameters

To decide which type of isofactorial model is more appropriate for the data, let us consider the indicator variograms for the lower and upper quartiles of the univariate

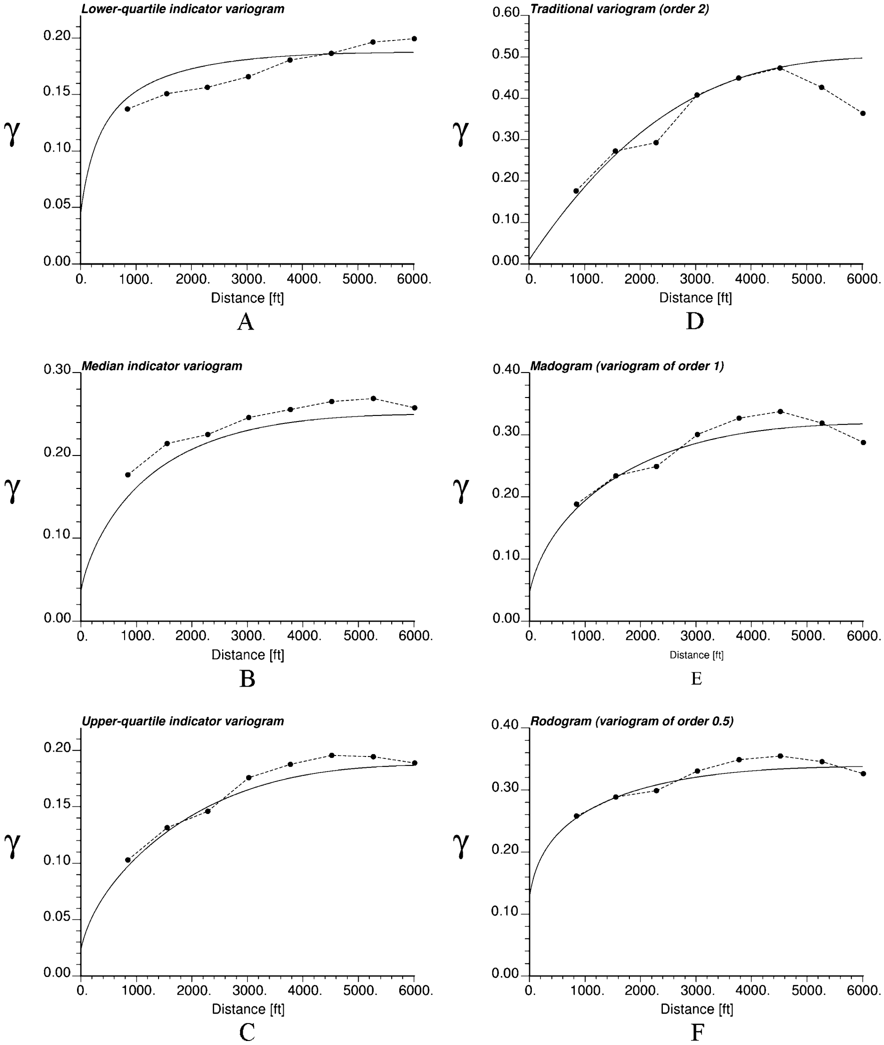

Fig. 5 **A**, **B**, **C**, quartile indicator variograms, and **D**, **E**, **F**, variograms of orders 2, 1, and 0.5 for the gamma scores data. The sample variograms are represented with *dots* and *dashed lines* and the fitted models with *solid lines*

distribution, i.e., 78.7 ppm and 325.44 ppm. The omni-directional sample variograms (Fig. 5A, C) show that the upper quartile indicator has a greater spatial correlation and lower nugget effect than the lower quartile indicator. A Gaussian model, either bivariate Gaussian or Hermitian, would therefore be ill suited as it provides the same theoretical variogram for both indicators.

Instead, a gamma model is preferred. In this model, the spatial correlation of the indicators is not symmetrical around the median threshold. The covariance of the indicator for a quantile p greater than 0.5 (above the median) has a lower slope at the origin than that of the indicator for quantile $1 - p$ (Emery 2005b, p. 428). This asymmetry is more pronounced when the shape parameter of the gamma distribution is small. Because the proposed substitution models only allow a half-integer shape parameter (13), we set this parameter to 0.5 in order to make the asymmetry in the indicator correlation as strong as possible. Accordingly, the original lead concentrations are transformed into a set of values with standard gamma univariate distribution, following the methodology proposed by Emery (2006b).

The remaining parameters of the substitution model are determined through a visual fit of the quartile indicator variograms and the variograms of order 2, 1, and 0.5 of the gamma scores data (Fig. 5). Only omni-directional variograms are considered, since the limited number of data does not allow one to detect a clear anisotropy in the spatial distribution of lead concentrations. A trial-and-error procedure leads to the following parameters

$$\begin{cases} N = 1, & \\ b = 8000, & \\ \forall \mathbf{h} \in \mathbb{R}^d, & \gamma_T(\mathbf{h}) = \|\mathbf{h}\|, \\ \forall \Delta t > 0, & \rho(\Delta t) = 0.98 \exp\left(-\dfrac{\Delta t^2}{30,000}\right), \\ \forall \Delta t \geq 0, & \rho'(\Delta t) = \left(1 - 1.5 \min(\Delta t, 1) + 0.5 \min\left(\Delta t^3, 1\right)\right)^2. \end{cases} \tag{19}$$

The model reproduces the greatest spatial correlation of the upper quartile indicator in comparison with the lower quartile and median indicators (Fig. 5A, B, and C). In the realizations, the high-value areas are expected to be spatially more continuous than the low-value and medium-value areas.

6.3 Conditional Simulation of Lead Concentrations

The lead concentrations can now be simulated conditionally to the available data. As an illustration, two realizations are displayed in Fig. 6, together with two realizations obtained by assuming that the normal scores of the lead concentrations have a multivariate Gaussian distribution (multi-Gaussian model).

By construction, all the realizations in Fig. 6 reproduce the conditioning data. However, the multi-Gaussian model does not account for the asymmetry in the spatial correlation of the indicators around the median quantile and for the spatial clustering of the high values (Chilès and Delfiner 1999, p. 101; Goovaerts 1997, p. 278). In practice, these defects are attenuated because the conditioning data tend to impose their own spatial structure over the theoretical model. But their incidence in risk assessment and decision-making in the planning of remediation measures may not be negligible, as will be illustrated in the next subsection.

6.4 Application to Site Management

Suppose that the entire area is divided into remediation units with size 60 ft × 60 ft and that the local planner decides to clean the units in which the average lead concentration is greater than 500 ppm. One is interested in knowing the distribution of the average lead concentrations over blocks with a support of 3600 ft^2, not the distribution of the point-support concentrations. In practice, the block-support concentrations

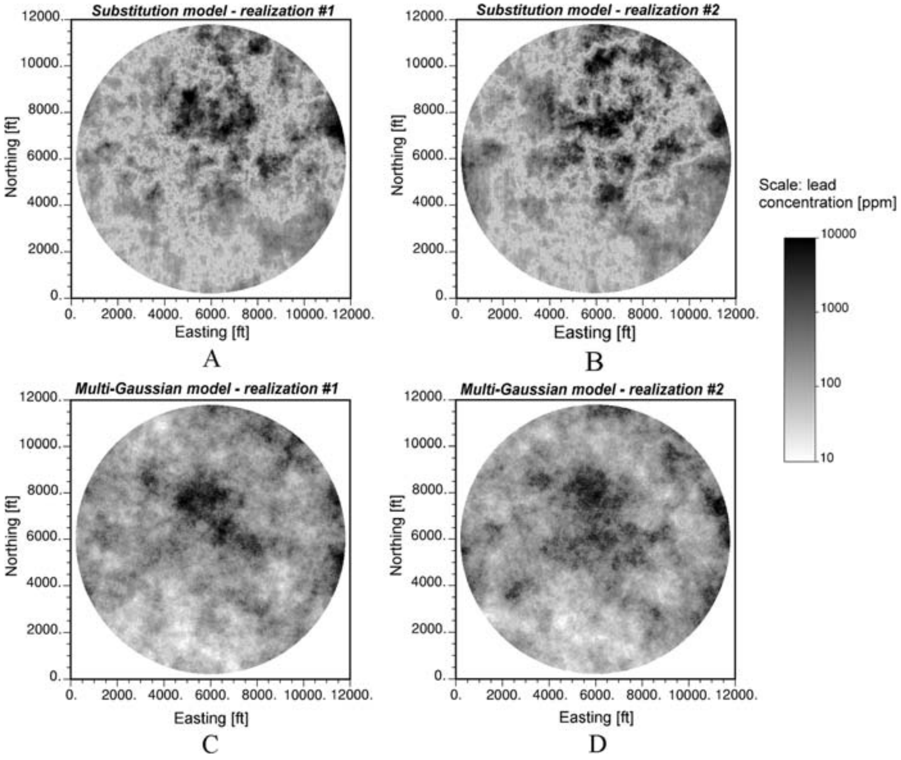
Substitution model - realization #1
Substitution model - realization #2
Multi-Gaussian model - realization #1
Multi-Gaussian model - realization #2
Northing [ft]
Easting [ft]
0.
2000.
4000.
6000.
8000.
10000.
12000.
A
B
C
D
Scale: lead concentration [ppm]
10000
1000
100
10

Fig. 6 **A**, **B**, two realizations of the substitution random field, and **C**, **D**, two realizations of a transformed Gaussian random field (*gray shades* in logarithmic scale)

are calculated by averaging the point-support concentrations simulated within each block. To minimize the effect of discretizing the block into a finite number of points, the simulation of lead concentrations should be performed on a fine grid (Chilès and Delfiner 1999, p. 572); in the present case, a 10 ft × 10 ft grid mesh was used.

As shown in Fig. 7A and B, the distribution of block-support concentrations depends on the model used for constructing the realizations. If the remediation decision is based on the multi-Gaussian model, more units have to be cleaned (17.4% of the total, while the substitution model gives a proportion of 15.3% of remediation units with concentrations above 500 ppm). If the criterion for remediation is the maximum point-support concentration over a unit instead of the average concentration, again the multi-Gaussian model overstates the proportion of units to be cleaned with respect to the substitution model (28.9% versus 23.0%) (Fig. 7C and D). These differences can be explained by the "destructuring" of the extreme values that takes place in the multi-Gaussian model; the extreme high lead concentrations tend to be scattered over the entire area, so that more units are likely to have an unacceptable average (or maximum) lead concentration (Fig. 8). The numerical results given in this example reflect the importance of the choice of the random field model representing the pollutant concentration when a change of support has to be considered for remediation decisions.

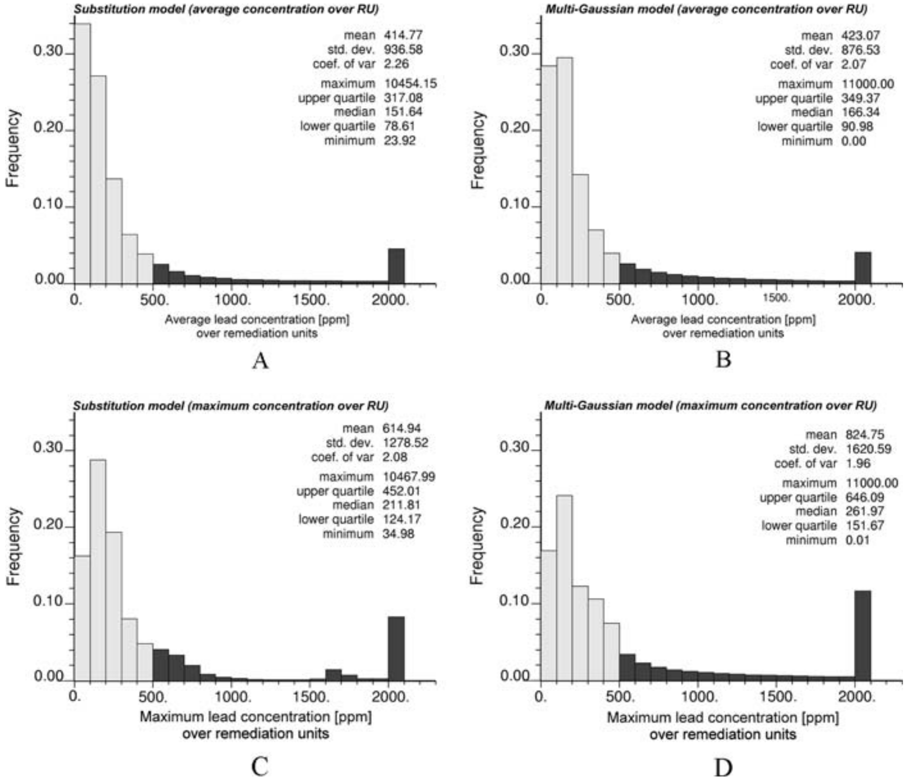

Fig. 7 Distributions of the average (**A**, **B**) and maximum (**C**, **D**) lead concentrations over remediation units with size 60 ft × 60 ft, calculated from 50 realizations. The *darker bars* correspond to upscaled concentrations above 500 ppm. The *last bar* of each histogram represents concentrations greater than 2000 ppm

7 Conclusions

Substitution random fields with Gaussian or gamma univariate distributions and isofactorial bivariate distributions can be constructed by composing two random fields: a directing function with Gaussian increments and a coding process with bivariate Gaussian or gamma distributions. Despite their limited number of parameters, these substitution models are flexible and allow one to represent regionalized variables with diverse features, such as a spatial clustering of the high values or an asymmetry in the correlation of the quantile indicators around the median threshold.

Indicator variograms and variograms of different orders can help decide which model is suitable for the available data and to infer and validate its parameters. The conditional simulation of substitution random fields requires using iterative algorithms based on the Gibbs sampler. Even if the number of iterations is limited and the Gibbs sampler is stopped before convergence, the final realizations always reproduce the conditioning data. The presented case study may encourage practitioners to use substitution random fields in application domains where modeling the spatial

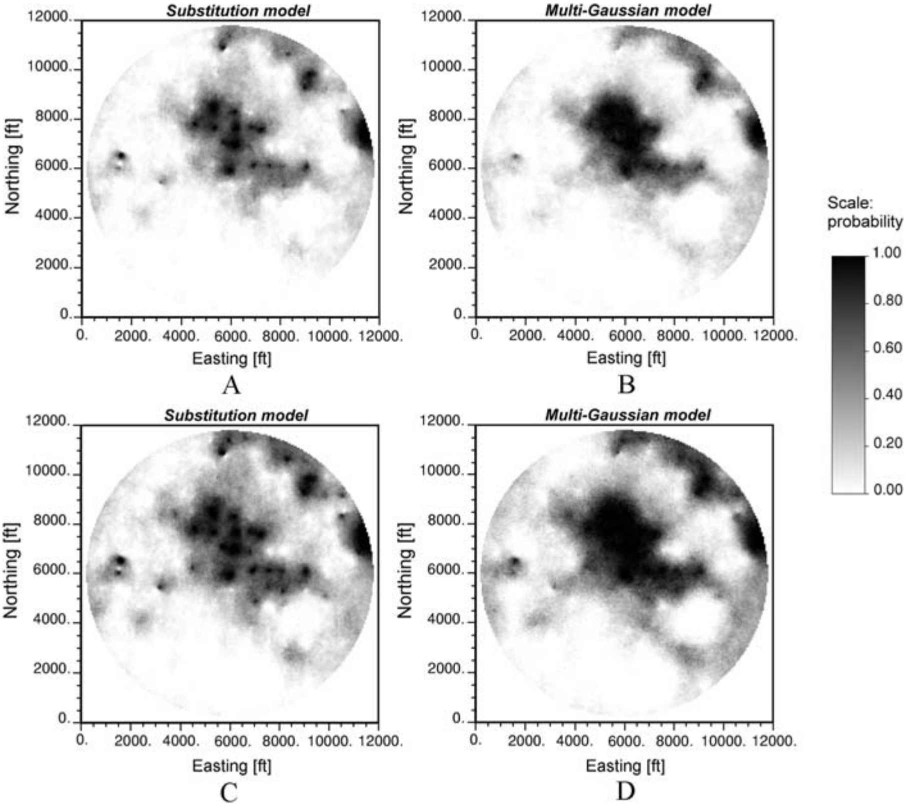

Fig. 8 Maps of the probability that the average (**A**, **B**) or maximum (**C**, **D**) lead concentrations over the remediation units exceed a toxic level of 500 ppm, calculated from 50 realizations

clustering of high values is critical to assess the risk of exceeding given levels, such as soil and environmental sciences.

Acknowledgements This article benefited from the comments of two anonymous reviewers. The author would also like to acknowledge the sponsoring by Codelco-Chile for supporting this research.

References

Chilès JP, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. Wiley, New York, 695 p

Emery X (2005a) Variograms of order ω: a tool to validate a bivariate distribution model. Math Geol 37(2):163–181

Emery X (2005b) Conditional simulation of random fields with bivariate gamma isofactorial distributions. Math Geol 37(4):419–445

Emery X (2005c) Geostatistical simulation of random fields with bivariate isofactorial distributions by adding mosaic models. Stoch Environ Res Risk Assess 19(5):348–360

Emery X (2006a) Multigaussian kriging for point-support estimation: incorporating constraints on the sum of the kriging weights. Stoch Environ Res Risk Assess 20(1–2):53–65

Emery X (2006b) A disjunctive kriging program for assessing point-support conditional distributions. Comput Geosci 32(7):965–983

Emery X (2007) Using the Gibbs sampler for conditional simulation of Gaussian-based random fields. Comput Geosci 33(4):522–537

Emery X, Lantuéjoul C (2006) TBSIM: A computer program for conditional simulation of three-dimensional Gaussian random fields via the turning bands method. Comput Geosci 32(10):1615–1628

Emery X, Soto-Torres JF (2005) Models for support and information effects: a comparative study. Math Geol 37(1):49–68

Galli A, Gao H (2001) Rate of convergence of the Gibbs sampler in the Gaussian case. Math Geol 33(6):653–677

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York, 480 p

Isaaks EH (1984) Risk qualified mappings for hazardous waste sites: a case study in distribution free geostatistics. Unpubl master's thesis, Department of Applied Earth Sciences, Stanford University, Stanford, 85 p

Johnson NL, Kotz S (1972) Distributions in statistics: continuous multivariate distributions. Wiley, New York, 333 p

Journel AG (1984) The place of non-parametric geostatistics. In: Verly G, David M, Journel AG, Maréchal A (eds) Geostatistics for natural resources characterization, vol 1. Reidel, Dordrecht, pp 307–335

Lancaster HO (1957) Some properties of the bivariate normal distribution considered in the form of a contingency table. Biometrika 44(1–2):289–292

Lancaster HO (1958) The structure of bivariate distributions. Ann Math Stat 29(3):719–736

Lantuéjoul C (1991) Ergodicity and integral range. J Microsc 161(3):387–403

Lantuéjoul C (1993) Substitution random functions. In: Soares A (ed) Geostatistics Tróia'92, vol 1. Kluwer Academic, Dordrecht, pp 37–48

Lantuéjoul C (2002) Geostatistical simulation, models and algorithms. Springer, Berlin, 256 p

Matheron G (1976a) A simple substitute for conditional expectation: the disjunctive kriging. In: Guarascio M, David M, Huijbregts CJ (eds) Advanced geostatistics in the mining industry. Reidel, Dordrecht, pp 221–236

Matheron G (1976b) Forecasting block grade distribution: the transfer functions. In: Guarascio M, David M, Huijbregts CJ (eds) Advanced geostatistics in the mining industry. Reidel, Dordrecht, pp 237–251

Matheron G (1984) Isofactorial models and change of support. In: Verly G, David M, Journel AG, Maréchal A (eds) Geostatistics for natural resources characterization, vol 1. Reidel, Dordrecht, pp 449–467

Matheron G (1989a) The internal consistency of models in geostatistics. In: Armstrong M (ed) Geostatistics, vol 1. Kluwer Academic, Dordrecht, pp 21–38

Matheron G (1989b) Two types of isofactorial models. In: Armstrong M (ed) Geostatistics, vol 1. Kluwer Academic, Dordrecht, pp 309–322

Oliver MA, Webster R, McGrath SP (1996) Disjunctive kriging for environmental management. Environmetrics 7(3):333–357

Tierney L (1994) Markov chains for exploring posterior distributions. Ann Stat 22(4):1701–1762

Wackernagel H (2003) Multivariate geostatistics—an introduction with applications. Springer, Berlin, 387 p