A computer package for modeling and simulating regionalized count variables⋆

Xavier Emery [a,b*], Jaime Hernández [c]

[a] *Department of Mining Engineering, University of Chile, Santiago, Chile*

[b] *Mining Technology Center, University of Chile, Santiago, Chile*

[c] *Department of Forest Resources Management, University of Chile, Santiago, Chile*

Abstract

Regionalized variables with discrete distributions are commonly associated with counts of individuals (precious stones in ore deposits, wild animals in ecosystems, trees in forests...) that can be represented by a spatial point process. In this paper, we propose to model the point distribution by a Cox process, i.e., a Poisson point process with a random regionalized intensity. The model is parsimonious and versatile, as it allows fitting the histogram of the count variable, its variogram and madogram. Simulation conditional to data is performed by recourse to iterative algorithms based on the Gibbs sampler.

⋆ Code available from server at http://www.iamg.org/CGEditor/index.htm

* Corresponding author. Tel.: + 56 2 978 4498.

E-mail address: xemery@ing.uchile.cl

Computer programs are provided for parameter inference and simulation, and an application to a forestry dataset is presented.

Keywords: spatial count data; spatial point process; Cox process; conditional simulation; Gibbs sampler.

1. Introduction

Discrete observations (counts) associated with spatial point processes are encountered in various fields of applications, e.g., evaluation of mineral deposits (counts of diamonds in kimberlite pipes, or of gold grains in alluvial placer deposits), forestry (counts of trees of a given species), ecology (sightings of wild animals), epidemiology (disease mapping based on reported infection cases), pest management (counts of infected plants), environmental sciences (radioactivity counts).

Several approaches are available to model the distribution of count variables and to predict the outcomes at unobserved locations. One option is the application of kriging methods. Apart from traditional linear kriging, which can be used for both continuous and discrete variables, specific approaches have been developed for dealing with count data, among which one can mention transitive kriging (Rivoirard *et al*., 2000) and disjunctive kriging (Matheron, 1984; Armstrong and Matheron, 1986; Emery, 2006).

Another option is the recourse to hierarchical models, in which the count variable is driven by a latent (hidden) random field that accounts for the spatial variations in the counts. The construction is hierarchical as, once the latent random field is fixed, the counts at different locations are assumed to be mutually independent. These models have been widely used for developing variants of kriging, such as binomial and Poisson kriging (McNeill, 1991; Oliver *et al*., 1993; Monestiez *et al*., 2006), as well as generalized linear predictors within a Bayesian framework (Diggle *et al*., 1998; Hrafnkelsson and Cressie, 2003; Banerjee *et al*., 2003). Most often, the latent random field is chosen as a monotonic transform of a Gaussian random field, e.g., a lognormal random field; a few works also propose the use of gamma random fields (Wolpert and Ickstadt, 1998).

In this work, we will consider a specific model (the Cox point process) for describing the spatial distribution of discrete observations or counts. The model is hierarchical and driven by a random field (known as *potential*) that controls the variations of the number of points in space. The goals of our contribution are the following:

- To propose a methodology for inferring the model parameters based on the fitting of the histogram and spatial continuity measures of the count data. The Bayesian framework will not be considered.

- To present a parametric family of random fields for modeling the potential, including Gaussian and gamma random fields as particular cases.

- To present algorithms for conditional simulation, in order to assess the uncertainty in the outcomes of the discrete (count) variable.

- To provide a set of computer programs and to illustrate them on a case study.

The aforementioned kriging approaches (transitive, disjunctive, binomial, Poisson kriging) are helpful for spatial prediction and for quantifying the uncertainty in the outcome of the variable at a location without observation. In comparison, conditional simulation allows a deeper analysis, since it provides measures of the joint uncertainty at multiple locations (Goovaerts, 2001). For instance, in natural resources exploitation, simulation is of great interest for quantifying the selectivity of the exploitation and the amounts of recoverable resources, for optimizing the sequence of extraction, and for assessing the technical and financial risks associated with the quantity and quality of extracted resources (e.g., Godoy and Dimitrakopoulos, 2004; Dowd and Dare-Bryan, 2005; Leite and Dimitrakopoulos, 2007; Nicholas *et al*., 2008).

2. The Cox process

2.1. Model description

Consider a Poisson point process in R^d with a regionalized intensity $\theta = \{\theta(\mathbf{x}), \mathbf{x} \in \mathsf{R}^d\}$. This process is characterized by the following properties (Lantuéjoul, 2002):

1) The number $N(v)$ of points contained in a finite domain $v \subset \mathbb{R}^d$ is a Poisson random variable with parameter

$$\theta(v) = \int_v \theta(\mathbf{x})\, d\mathbf{x} \tag{1}$$

2) If $\{v_i, i = 1 \dots m\}$ is a set of pairwise disjoint domains of $\mathbb{R}^d$, then the random variables $\{N(v_i), i = 1 \dots m\}$ are mutually independent.

Because of the last property, there is no stochastic dependence between the numbers of points observed in non-overlapping domains of space. Hence, the spatial structure of the point process is entirely controlled by the (deterministic) regionalized intensity function θ, which must be explicitly modeled by the user.

In practice, the intensity of the point process is often uncertain in areas without data, so that it is convenient and more parsimonious to use a stochastic modeling of this intensity. This leads to the so-called *Cox process* or *doubly stochastic Poisson point process* (Cox, 1955), in which the regionalized intensity is replaced by a random field $\Theta = \{\Theta(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d\}$ called *potential*. The spatial distribution of points is characterized by the following probabilities:

$$\text{Prob}\{\bigcap_{i=1}^{m} N(v_i) = n_i\} = E\left\{\prod_{i=1}^{m} \exp\{-\Theta(v_i)\} \frac{\Theta(v_i)^{n_i}}{n_i!}\right\}, \tag{2}$$

for all finite set of pairwise disjoint domains $\{v_i, i = 1\ldots m\}$ and nonnegative integers $\{n_i, i = 1\ldots m\}$.

In general, there exists a stochastic dependence between the numbers of points observed in non-overlapping domains, due to the spatial dependence structure of the potential field Θ. This random field measures the propensity of any domain to contain points of the process because of geographical, geological or environmental factors.

The Cox process has been used in the geosciences for modeling natural hazards (Jaquet and Carniel, 2001) and natural resources, e.g., the distribution of trees in forests (Matérn, 1986; Stoyan and Penttinen, 2000) or of precious stones in ore deposits (Kleingeld and Lantuéjoul, 1993; Kleingeld *et al*., 1997).

2.2. Modeling the univariate distribution of the potential field

Assume that the numbers of points have been counted in a set of non-overlapping domains $\{v_i, i = 1\ldots m\}$, all with the same support (shape and orientation) as a reference domain v. According to Equation (2), the univariate distribution of the counts is given by:

$$\forall n \in \mathsf{N}, P(n) = \text{Prob}\{N(v) = n\} = E\left\{ \exp\{-\Theta(v)\} \frac{\Theta(v)^n}{n!} \right\}. \qquad (3)$$

In practice, knowing the distribution of the counts, i.e., the distribution of $N(v)$, one wants to determine that of $\Theta(v)$. A flexible model is obtained by considering that $N(v)$ follows a generalized Sichel distribution:

$$\forall n \in \mathbb{N}, P(n) = \frac{(\frac{a}{b})^{\alpha/2} K_{\alpha+n}(2\sqrt{(a+1)b})}{n!\,(\frac{a+1}{b})^{(\alpha+n)/2} K_{\alpha}(2\sqrt{ab})}, \quad (4)$$

with $a > 0$, $b > 0$, $\alpha \in \mathbb{R}$, and K_α the modified Bessel function of the second kind of index α. In such a case, $\Theta(v)$ follows a generalized inverse Gaussian distribution with density (Jørgensen, 1982; Lantuéjoul, 2002):

$$\forall \theta \in \mathbb{R}_+, f(\theta) = \frac{(\frac{a}{b})^{\alpha/2}}{2K_{\alpha}(2\sqrt{ab})} \theta^{\alpha-1} \exp\{-a\theta - \frac{b}{\theta}\}. \quad (5)$$

The motivation for choosing a generalized Sichel distribution is the capability to model a very large family of univariate distributions for the count data, in particular highly skewed distributions such as stone number frequencies in diamondiferous deposits (Sichel, 1973). The inference of the parameters (a, b, α) can be done by fitting the count data histogram, based on the moments of this histogram and/or on the frequencies of specific classes (for instance, $P(0)$, $P(1)$ and $P(2)$). One can also use maximum likelihood techniques (Stein *et al*., 1987) or iterative algorithms in order to minimize a given goodness-of-fit criterion (Press *et al*., 2007). Particular cases include:

- $\alpha = -1/2$: this yields an inverse Gaussian distribution for $\Theta(v)$ and a standard Sichel distribution (Sichel, 1974) for $N(v)$;

- $b \to 0$ and $\alpha > 0$: this yields a gamma distribution for $\Theta(v)$ and a negative binomial distribution for $N(v)$:

$$\forall n \in \mathbb{N}, P(n) = \frac{\Gamma(\alpha + n)}{\Gamma(\alpha) n!} \left(\frac{a}{a+1} \right)^{\alpha} \left(\frac{1}{a+1} \right)^{n}. \quad (6)$$

Other choices of univariate distributions are possible. For instance, one can assume that the potential field has a lognormal distribution, which yields a point process known as log-Gaussian Cox process (Møller *et al*., 1998; Brix and Diggle, 2001). Nevertheless, because the lognormal distribution only depends on two parameters, this is a less flexible choice than the generalized inverse Gaussian distribution and will not be considered in this work.

2.3. Modeling the spatial structure of the potential field

There exists a simple relationship between the covariance function of the potential field and that of the counts (Chilès and Delfiner, 1999; Lantuéjoul, 2002):

$$\forall \mathbf{h} \in \mathbb{R}^d, \operatorname{cov}\{\Theta(v), \Theta(v_{\mathbf{h}})\} = \operatorname{cov}\{N(v), N(v_{\mathbf{h}})\} - E\{N(v \cap v_{\mathbf{h}})\}, \quad (7)$$

where $v_{\mathbf{h}}$ represents domain v shifted by vector $\mathbf{h}$. In particular, if v and $v_{\mathbf{h}}$ do not overlap, the last term in Equation (7) vanishes, and the covariance of the potential field is the same as that of the counts. At $\mathbf{h} = \mathbf{0}$, one gets:

$$\mathrm{var}\{\Theta(v)\} = \mathrm{var}\{N(v)\} - E\{N(v)\}. \tag{8}$$

This equation proves that $N(v)$ cannot be Poisson distributed (its variance is greater than its mean), unless the variance of $\Theta(v)$ is zero.

The knowledge of the covariance function of the potential random field still leaves a large indetermination on this field, and many models can be designed. In this work, we consider the case when the potential $\Theta(v)$ is of the following form:

$$\Theta(v) = \phi((\delta + Y(v))^2) = \psi(Y(v)) \tag{9}$$

with

- ϕ a monotonic function (either increasing or decreasing)
- $\delta \in \mathbb{R}_+$
- $\{Y(v), v \in \mathbb{R}^d\}$ a stationary standard Gaussian random field with correlation function $\rho(\mathbf{h})$
- $\forall y \in \mathbb{R}, \psi(y) = \phi((\delta + y)^2)$.

The motivation for the model proposed in (9) is its simplicity, insofar as it depends on few parameters, and its versatility. In particular, one has the following extreme cases:

- $\delta = +\infty$: the potential is the transform of a standard Gaussian random field (Kleingeld *et al*., 1997);

- $\delta = 0$: the potential is the transform of a gamma random field with shape parameter 0.5 (Møller and Waagepetersen, 2003).

In the general case, $(\delta + Y(v))^2$ has a non-central chi square distribution with one degree of freedom and shift parameter δ. The recourse to an underlying Gaussian random field $Y(v)$ will also facilitate the design of conditional simulation algorithms, as it will be seen in Section 3.

2.4. Steps for parameter inference

The proposed model depends on the parameters (a,b,α) of the count distribution, the shift parameter δ, the correlation function $\rho(\mathbf{h})$, and on whether the transformation function ϕ is increasing or decreasing. Given a finite set of count data $\{N(v_i), i = 1\ldots m\}$, the inference of these parameters can be performed by trials and errors, according to the following steps:

(1) Find the parameters (a,b,α) of the generalized Sichel distribution (Eq. 4) that best fit the count data histogram. One can use empirical approaches (e.g., a visual comparison

of the data histogram and model distribution, or a fitting of the first-order moments or of specific frequencies of the data histogram), maximum likelihood approaches, or goodness-of-fit optimization approaches. According to Equation (5), the univariate distribution of the potential field will be a generalized inverse Gaussian distribution with parameters (a,b,α).

(2) Choose a value for the shift parameter δ and a behavior for the transformation function ϕ (increasing or decreasing).

(3) Construct a conversion table between the potential field and its chi square transform:

(a) Simulate a large number of independent random variables with generalized inverse Gaussian distribution.

(b) Simulate a large number of chi-squared random variables with 1 degree of freedom and shift parameter δ.

(c) The conversion table, which models the transformation function ϕ, is obtained by constructing a quantile-quantile plot between the sets of simulated values obtained in the previous steps (a) and (b).

(4) Calculate the sample variogram of the available count data for various lag vectors (or lag classes if distance and/or angle tolerances are taken into account). Calculations can

be done by the method-of-moments approach or by recourse to robust variogram estimators (Chilès and Delfiner, 1999).

(5) Propose a model for the correlation function ρ of the Gaussian random field $Y(v)$.

(6) For each lag vector **h** used in the calculation of the sample variogram (step (4)):

(a) Simulate a large number of bi-Gaussian pairs $\{Y(v),Y(v_{\mathbf{h}})\}$ with correlation $\rho(\mathbf{h})$

(b) Shift and square the simulated values, then back-transform by using the conversion table obtained at step (3). Obtain realizations of $\{\Theta(v),\Theta(v_{\mathbf{h}})\}$.

(c) Simulate independent Poisson random variables with parameters $\Theta(v)$ and $\Theta(v_{\mathbf{h}})$. Obtain realizations of $\{N(v),N(v_{\mathbf{h}})\}$ and calculate their variogram.

(7) If the variogram calculated at step (6) does not satisfactorily fit the sample variogram obtained in (4), then go back to step (5) and propose another correlation function ρ, or go back to step (2) and change the shift parameter δ or the assumed behavior of ϕ. The quality of the variogram fitting can be assessed graphically or statistically, for instance by calculating sums of squared errors or weighted sums of squared errors of the fitted models.

Unlike Bayesian approaches, here the model parameters (a,b,α,δ,ρ) are assumed perfectly known. To better determine the spatial structure of the potential field, the above procedure (steps 4 to 7) can also be applied to structural tools other than the variogram, in particular indicator variograms associated with specific thresholds, or variograms of different orders, e.g., the madogram (first-order variogram) (Emery, 2005).

3. Conditional simulation

We now address the problem of simulating the count variable $N(v)$ in a bounded domain of $\mathbb{R}^d$, conditionally to a set of data $\{N(v_i) = n_i,\ i = 1\dots m\}$. In the following, to shorten notations, we will write, for any subset $J \subseteq I = \{1,\dots m\}$, N_J, n_J, Y_J and y_J instead of $\{N(v_j),\ j \in J\}$, $\{n_j, j \in J\}$, $\{Y(v_j), j \in J\}$ and $\{y_j, j \in J\}$. We also assume that the supports $\{v_i,\ i = 1\dots m\}$ are the same up to a translation and do not overlap.

The conditional simulation consists of the following steps.

(1) Simulate Y_I conditionally to the data $N_I = n_I$.

(2) Simulate the Gaussian random field $Y(v)$ at the target locations in $\mathbb{R}^d$ conditionally to the vector Y_I obtained at step (1). This can be done by using any multivariate Gaussian simulation algorithm (sequential Gaussian, Choleski decomposition of the covariance

matrix, continuous or discrete spectral method, turning bands…) (Chilès and Delfiner, 1999).

(3) At each target location v, derive the potential $\Theta(v)$ as per Equation (9) and simulate an independent Poisson random variable $N(v)$ with parameter $\Theta(v)$.

Steps (2) and (3) are straightforward and the only difficulty concerns step (1). According to Bayes' theorem and Equations (2) and (9), the conditional density probability function of Y_I is

$$g(y_I \mid n_I) \propto g(y_I) \prod_{i=1}^{m} \exp\{-\psi(y_i)\} \frac{\psi(y_i)^{n_i}}{n_i!}, \tag{10}$$

with $g(y_I)$ the prior standard multivariate Gaussian density of Y_I. A random vector with the conditional distribution (10) can be obtained by using the *Gibbs sampler*, an iterative algorithm originally designed by Geman and Geman (1984). Given the vector $Y_I = y_I$ in the current state of the sampler, one iteration consists of the following steps:

(a) Define a random permutation $\{i_1,\ldots i_m\}$ of I.

(b) Set $k = 1$.

(c) Set $i = i_k$ and denote $J = I - \{i\}$.

(d) Simulate Y_i conditional on $Y_J = y_J$. Since $Y(v)$ is a stationary Gaussian random field, the conditional distribution of Y_i is Gaussian, with mean equal to the simple kriging prediction of Y_i from Y_J and variance equal to the simple kriging variance (Chilès and Delfiner, 1999). Let y_i' denote the new simulated value of Y_i.

(e) Calculate the conditional probabilities associated with the former and new states:

$$\begin{aligned} p_i &= \text{Prob}\{N_I = n_I \mid Y_i = y_i, Y_J = y_J\} \\ &= \exp\{-\psi(y_i)\}\frac{\psi(y_i)^{n_i}}{n_i!}\prod_{j\in J}\exp\{-\psi(y_j)\}\frac{\psi(y_j)^{n_j}}{n_j!} \end{aligned} \tag{11}$$

$$\begin{aligned} p_i' &= \text{Prob}\{N_I = n_I \mid Y_i = y_i', Y_J = y_J\} \\ &= \exp\{-\psi(y_i')\}\frac{\psi(y_i')^{n_i}}{n_i!}\prod_{j\in J}\exp\{-\psi(y_j)\}\frac{\psi(y_j)^{n_j}}{n_j!} \end{aligned} \tag{12}$$

(f) Simulate a uniform random variable U on [0,1].

(g) If $p_i U < p_i'$, substitute y_i' for y_i, in accordance with Metropolis acceptance criterion (Metropolis *et al*., 1953). According to Equations (11) and (12), the substitution takes place if:

$$U\exp\{-\psi(y_i)\}\psi(y_i)^{n_i} < \exp\{-\psi(y_i')\}\psi(y_i')^{n_i}. \tag{13}$$

(h) Set $k = k+1$.

(i) If $k \leq m$, go back to step (c).

The iterative algorithm (steps (a)-(i)) must be run until a maximum number of iterations has been reached, ensuring the convergence of the simulated vector Y_I to a vector following the conditional distribution (10). The whole procedure must then be repeated as many times as necessary (using different seeds for random number generation) to obtain the required number of realizations of Y_I.

To initialize the Gibbs sampler, it is convenient to simulate each component Y_i according to its distribution conditional on $N_i = n_i$ only. According to Bayes' theorem and Equations (3) and (9), the density of this conditional distribution is

$$g(y_i \mid n_i) = C\, g(y_i) \exp\{-\psi(y_i)\} \frac{\psi(y_i)^{n_i}}{n_i!}, \tag{14}$$

where C is a normalization constant, while $g(y_i)$ is the prior standard Gaussian density of Y_i. The conditional density (14) is bounded by:

$$g(y_i \mid n_i) \leq C\, g(y_i) \exp\{-n_i\} \frac{n_i^{n_i}}{n_i!}. \tag{15}$$

This inequality allows simulating $Y_i \mid N_i = n_i$ by rejection sampling (Von Neumann, 1951; Freulon 1994):

(i) Simulate Y_i as a standard Gaussian random variable.

(ii) Simulate a uniform random variable U on [0,1].

(iii) If $U \exp\{-n_i\} n_i^{n_i} \leq \exp\{-\psi(Y_i)\} \psi(Y_i)^{n_i}$, deliver Y_i. Otherwise, go back to step (i).

Rejection sampling can also be used in the iterative steps of the Gibbs sampler, which leads to the following acceptance criterion at step (g) (corrected from Kleingeld *et al*., 1997):

$$U \exp\{-n_i\} n_i^{n_i} < \exp\{-\psi(y_i')\} \psi(y_i')^{n_i} . \tag{16}$$

However, criterion (16) turns out to be more restrictive than the Metropolis criterion (13) (the chance of accepting the new simulated value y_i' is lower), which implies a slower convergence (hence, the need for a larger number of iterations) if it were used in the Gibbs sampler.

To illustrate the proposed algorithm, a non-conditional realization is generated on a 400 × 400 grid (Fig. 1A), with the following parameters:

$$\begin{cases} a = b = 0.5 \\ \alpha = -0.5 \\ \delta = 0 \\ \phi \text{ increasing} \\ \forall \mathbf{h} \in \mathsf{R}^2, \rho(\mathbf{h}) = \text{isotropic cubic model with range } 80 \end{cases} \tag{17}$$

One hundred locations are selected at random (uniformly) among the 160,000 grid nodes (Fig. 1B). The values simulated at these locations are then used as conditioning data for two new realizations (Fig. 1C and 1D). In each case, the Gibbs sampler uses one hundred iterations and the Gaussian random field $Y(v)$ is simulated by the turning bands algorithm with one thousand regularly-distributed lines. One simulation takes about five minutes of CPU time on a Pentium 2.0 Ghz, consisting of 3.9 minutes for the Gibbs sampler and 1.1 minute for the turning bands simulation over the 160,000 grid nodes.

The algorithm can become time-consuming if the conditioning data set is very large (say, over a few thousand data), insofar as it requires solving a kriging system in the Gibbs sampler (step 1d) and in the conditional simulation of $Y(v)$ (step 2). To speed up the algorithm, one can perform kriging in a moving neighborhood; the search for nearest neighbors can be made faster by using a super-block strategy or multidimensional tree structures (Friedman *et al*., 1975; Deutsch and Journel, 1992).

4. Program description

4.1. Inference program

The algorithm for parameter inference proposed in Section 2.4 is implemented in a Matlab program called **COXINFER** that can be used in workspaces of up to three dimensions. The input arguments are the parameters (a,b,α) of the univariate count distribution (Eq. 4), the shift parameter δ (set to a positive value if the transformation function ϕ is assumed to be increasing, and to a negative value otherwise), the variogram model of the underlying Gaussian random field $Y(v)$ (the sum of a nugget effect and one or more nested structures), and the parameters for variogram calculation (lag, number of lags, azimuth and dip). The outputs are the values of the variogram and madogram of the count variable $N(v)$ for the specified direction and lag distances, and a conversion table between the potential random field and its chi square transform (Eq. 9).

The reader is referred to the header of the program file for details on the input and output parameters. Alternatively, program COXINFER can be used with an external parameter file (by default, COXINFER.PAR), in which case there is no need to enter the parameters in the Matlab workspace (Table 1).

The simulation of generalized inverse Gaussian random variables (step 3a of Section 2.4) is done by subroutine GIGRND, which uses an acceptance and rejection method proposed by Devroye (1986), except in the following cases:

- $b = 0$ and $\alpha \geq 1$: the Matlab subroutine GAMRND is used;

- $b = 0$ and $0 < \alpha < 1$: an acceptance-rejection method (Ahrens and Dieter, 1974) is used to simulate gamma random variables with shape parameter less than 1 (subroutine GAMRND2);

- $\alpha = -1/2$: random variables with an inverse Gaussian distribution are simulated via the following algorithm (Michael *et al*., 1976; Matheron, 1985):

  a) Simulate a standard Gaussian variable X and set $S = \dfrac{2X^2}{\sqrt{ab}}$

  b) Calculate $Z = \dfrac{1}{2}(S + 2 - \sqrt{S^2 + 4S})$

  c) The simulated variable is $Z\sqrt{\frac{b}{a}}$ with probability $\frac{1}{1+Z}$ or $\frac{1}{Z}\sqrt{\frac{b}{a}}$ with probability $\frac{Z}{1+Z}$.

4.2. Conditional simulation program

The conditional simulation algorithm presented in Section 3 has been implemented in a program called **COXSIMU**. The second step of the algorithm (conditional simulation of

the Gaussian random field $Y(v)$) uses the turning bands algorithm and relies on a former program by Emery and Lantuéjoul (2006) that offers the following functionalities:

1) *Versatility*: there is no restriction on the number of nested structures contained in the correlation model for $Y(v)$ (the most commonly used basic models are available), nor on the spatial configuration and on the number of locations targeted for simulation. In particular, the program can easily handle several millions of locations that may not be regularly spaced, a feature that cannot be addressed by simulation methods such as the LU decomposition of the covariance matrix, discrete Fourier and circulant-embedding algorithms. Simulation is restricted to three-dimensional workspaces, or to sub-spaces by setting one or two of the coordinates to a constant value.

2) *Efficiency*: the turning bands algorithm is very fast because it simplifies the simulation of $Y(v)$ to that of a series of one-dimensional random fields. Moreover, at each target location, all the realizations of $Y(v)$ are conditioned by solving a single kriging system. Kriging can be performed in a moving neighborhood defined by an ellipsoid divisible into octants, or in a unique neighborhood (provided that the conditioning data set is not too large, say, less than a few thousands data). With the use of a moving neighborhood, a super-block search is implemented and there is practically no limit to the number of conditioning data, other than memory capacity.

3) *Accuracy*: the 1D random fields are simulated continuously (no discretization) along the lines, which allows reproducing the correlation model of $Y(v)$ without bias, even if the simulation is performed at unevenly spaced locations.

The input parameters (see header of the program file for details) consist of:

- information on the locations targeted for simulation: coordinates, block discretization if a change of support (regularization) has to be considered;

- information on the conditioning data: coordinates, count values, trimming limits, parameters (a,b,α) of the univariate distribution model;

- spatial structure of the potential field: shift parameter δ, conversion table, correlation model for $Y(v)$;

- simulation parameters: numbers of lines to use for turning bands simulation, number of realizations to generate, seed for random number generation, number of iterations to use for the Gibbs sampler;

- information on kriging neighborhood: radii, angles, octant division and number of data per octant;

- information on the output: name of output file, presence of a header in this file.

The output of program COXSIMU is an external ASCII file with the simulated values of the count variable (one column per realization). For simulations at regular grid nodes, the nodes are ordered point by point to the east, then row by row to the north, and finally level by level upward.

As for the inference program, COXSIMU can also be used with an external parameter file (by default, COXSIMU.PAR), without the need to specify the input arguments in the Matlab workspace (Table 2). In such a case, the conditioning data and the conversion table must be stored in ASCII files without header.

5. A case study in forest resources evaluation

In this section, the proposed model and programs are applied to a forestry case study. The dataset consists of 108 measurements from a ground survey in a forest domain located in southern Chile (Arauco Province) and owned by Bosques Arauco S.A. Each measurement corresponds to a rectangular plot of about 500m^2 and indicates the number of radiata pines (coniferous trees) observed in this plot. Additional variables not considered in this work are the mean tree height and tree basal area.

The data are located on a quasi-regular grid with a mesh of about 150m, and the histogram of counts is slightly skewed with most of the values between 15 and 50 (Fig. 2).

To assess the wood resources that can be recovered, it is of interest to predict the number of trees, both globally over the entire domain and locally, and to measure the uncertainty in this number. To fulfill these goals, the spatial distribution of trees will be modeled by a Cox process, as described in Section 2. Even if this model may be ill-suited to represent the processes generated by local competition between trees at a short scale, this should no longer be the case at larger scales (plot scale and stand scale), where the tree distribution is the expression of aggregation processes under micro-site conditions.

The histogram of tree counts (Fig. 2) has a variance greater than the mean, which complies with the requirement for using a Cox point process (Eq. 8). In the following, it will be modeled by a negative binomial distribution. The mean and variance of this distribution can be determined by that of the count histogram, or by examining the variogram of the count data (Fig. 3):

- the mean should be no more than the nugget effect, i.e., approximately 25 (Eq. 7)
- the variance should coincide with the variogram sill, i.e., approximately 120.

Based on these statements, the parameters of the negative binomial distribution (Eq. 6) are found to be $a = 0.263$ and $\alpha = 6.58$.

The remaining parameters of the model (correlation function $\rho(\mathbf{h})$, shift parameter δ and conversion table between count data and chi square random field) are determined by trials and errors, by recourse to program COXINFER. We attempted to find the best fitting of the sample variogram and sample madogram of the count data, calculated along the main

directions of anisotropy (N40ºW and N50ºE) (Fig. 3). To the authors' opinion, fitting both the variogram and madogram is desirable in order to provide an adequate modeling of the bivariate (and, hopefully, multivariate) distributions of the count variable, which might not be the case when fitting the sole variogram (Chilès and Delfiner, 1999; Emery, 2005). The following parameters are finally chosen:

- $\delta = 5$;
- ϕ = increasing function;
- $\rho(\mathbf{h})$ = spherical model with range 1200m (N40ºW) and 650m (N50ºE).

Note that the model parameters have been determined only on the basis of the univariate and bivariate distributions of the count data, and higher-order distributions (multiple-point statistics) have not been examined. Therefore, as long as the sample histogram, variogram and madogram are deemed reliable, there is no real danger of over-fitting, i.e., the selected parameters are expected not to exceed the content of information contained in the data.

To put the fitted model to the test, leave-one-out cross-validation is performed. At each data location, the number of trees is simulated 1000 thousand times conditionally to the remaining data, and a prediction is obtained by averaging the simulated values. Table 3 shows that the prediction errors have an average close to zero and that the regression of the true upon the predicted numbers of trees has a slope close to one, which indicates that predictions do not suffer from global or conditional bias. From the set of simulated values, it is also possible to construct probability intervals and to verify that the proportions of

data belonging to the intervals match the interval probabilities, e.g., by checking that the *goodness statistics* (Goovaerts, 2001) is close to one (Table 3).

Having determined and validated the model parameters, we can now simulate the numbers of trees over the study domain by using program COXSIMU. The domain is covered by a regular grid containing 5663 cells with size 22.4m × 22.4m, so that each cell corresponds to the same areal support as the data (i.e., $500m^2$). Program COXSIMU is run by using 100 iterations for the Gibbs sampler, 1000 lines for the turning bands simulation, and a unique neighborhood for the conditioning data selection. The CPU time for constructing 1000 realizations is 4.8 hours (i.e., an average of 17 seconds per realization) on a Pentium 2.0 Ghz. As an illustration, Figure 4 presents the maps of two realizations and of the average of 1000 realizations, while Table 4 gives statistics of the number of trees over the realizations.

For comparison, ordinary kriging of the count data provides an estimated total of 176,230 trees over the entire domain, which is very close to the expected number (175,545) found via the conditional simulation approach. Also, the kriging prediction is almost the same as the average of realizations (correlation coefficient equal to 0.997). However, in contrast to kriging, simulation provides a set of realizations that reproduce the spatial variability and allows determining confidence intervals on the number of trees (either globally or locally) and quantifying the risk of not meeting planned production targets.

For instance, let us assume that the area will be harvested by skidding at a rhythm of 7.45 hectares per day, according to the sequence indicated in Fig. 5A. The entire area, covering 283.1 hectares, will be harvested in 38 days, with an average recovery of 4620 trees per day. To schedule the wood production, it is of interest to evaluate the number of days for which the harvest is significantly less than the average, say, less than 4000 trees per day. The daily harvest can be calculated on each conditional realization, yielding a set of 1000 recovery curves (Fig. 5B). It is found that the number of days with a harvest smaller than 4000 trees is comprised between 0 (best case) and 15 (worst case), with an average of 9.7. These results cannot be obtained with kriging approaches, insofar as they depend on the multivariate distributions of the numbers of trees in space; they can be used by the landowner to modify the harvest sequence or to schedule the harvesting of other sites, in order to reduce the variability in the daily stream sent to the pulp or saw mills.

6. Conclusions

One motivation of this work was to present a stochastic model for representing discrete regionalized variables associated with count observations. The proposed Cox process turns out to be versatile and parsimonious, as it depends upon few parameters that allow fitting the main characteristics of the discrete variable: univariate distribution (histogram) and spatial continuity (variogram, madogram, or other continuity measures).

Because the process is driven by a potential obtained by transforming a Gaussian random field, the conditional simulation is easily performed by recourse to iterative algorithms. Of

course, many other random field models could be designed for representing the potential, in particular models based on operations and/or combinations of Gaussian random fields, for which the Gibbs sampler is well-suited (Emery, 2007, 2008).

The proposed approach can also be enriched by including covariates spatially correlated with the potential field. For instance, in the presented forestry case study, one could think of physiographic variables such as terrain elevation, slope and profile curvature to explain the spatial variations of the number of trees. Adaptations of the simulation algorithm to the multivariate framework are straightforward (the mean and variance of the conditional distributions of $Y(v)$ needed in the Gibbs sampler stage are obtained by using co-kriging instead of kriging).

A difficulty of the algorithm (not tackled in this work) is that all the samples must have congruent supports and the total domain must be divisible into the sample support. If these requirements are not fulfilled, compensation techniques must be introduced to reshape and resize the sample supports to proper dimensions (Ferreira and Lantuéjoul, 2007).

7. References

Ahrens, J.H., Dieter, U., 1974. Computer methods for sampling from gamma, beta, Poisson and binomial distributions. Computing 12 (3), 223-246.

Armstrong, M., Matheron, G., 1986. Disjunctive kriging revisited, Part I. Mathematical Geology 18 (8), 711-728.

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2003. Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall/CRC, Boca Raton, 472 pp.

Brix, A., Diggle, P.J., 2001. Spatiotemporal prediction for log-Gaussian Cox processes. Journal of the Royal Statistical Society B63 (4), 823-841.

Chilès, J.P., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty. Wiley, New York, 695 pp.

Cox, D.R., 1955. Some statistical models connected with series of events. Journal of the Royal Statistical Society B17, 129-164.

Devroye, L., 1986. Non-Uniform Random Variate Generation. Springer-Verlag, New York, 843 pp.

Deutsch, C.V., Journel, A.G., 1992. GSLIB: Geostatistical Software Library and User's Guide. Oxford University Press, New York, 340 p.

Diggle, P.J., Tawn, J.A., Moyeed, R.A., 1998. Model-based geostatistics (with discussion). Applied Statistics 47 (3), 299-350.

Dowd, P.A., Dare-Bryan, P.C., 2005. Planning, designing and optimising production using geostatistical simulation. In: Dimitrakopoulos, R., Ramazan, S., (Eds.), The International Symposium on Orebody Modelling and Strategic Mine Planning: Uncertainty and Risk Management. Hyatt Regency, Perth, pp. 321-337.

Emery, X., 2005. Variograms of order ω: a tool to validate a bivariate distribution model. Mathematical Geology 37 (2), 163-181.

Emery, X., 2006. A disjunctive kriging program for assessing point-support conditional distributions. Computers & Geosciences 32 (7), 965-983.

Emery, X., 2007. Using the Gibbs sampler for conditional simulation of Gaussian-based random fields. Computers & Geosciences 33 (4), 522-537.

Emery, X., 2008. Substitution random fields with Gaussian and gamma distributions: theory and application to a pollution data set. Mathematical Geosciences 40 (1), 83-99.

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

Ferreira, J., Lantuéjoul, C., 2007. Compensation for sample mass irregularities in core sampling for diamonds and its impact on grade and variography. In: Costa, J.F., Koppe, J., (Eds.), Third World Conference on Sampling and Blending. Fundação Luiz Englert, Publication Series Nº1/2007, Porto Alegre, pp. 3-15.

Freulon, X., 1994. Conditional simulation of a Gaussian vector with non linear and/or noisy observations. In: Armstrong, M., Dowd, P.A., (Eds.), Geostatistical Simulations. Kluwer Academic, Dordrecht, pp. 57-71.

Friedman, J.H., Baskett, F., Shustek, L.J., 1975. An algorithm for finding nearest neighbors. IEEE Transactions on Computers 24 (10), 1000-1006.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence 6 (6), 721-741.

Godoy, M., Dimitrakopoulos, R., 2004. Managing risk and waste mining in long-term production scheduling of open-pit mines. SME Transactions 316, 43-50.

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103 (1-2), 3-26.

Hrafnkelsson, B., Cressie, N., 2003. Hierarchical modeling of count data with application to nuclear fall-out. Environmental and Ecological Statistics 10 (2), 179-200.

Jaquet, O., Carniel, R., 2001. Stochastic modelling at Stromboli: a volcano with remarkable memory. Journal of Volcanology and Geothermal Research 105 (3), 249-262.

Jørgensen, B., 1982. Statistical properties of the generalized inverse Gaussian distribution. Lecture Notes in Statistics, Springer-Verlag, New York, 188 pp.

Kleingeld, W.J., Lantuéjoul, C., 1993. Sampling of orebodies with a highly dispersed mineralization. In: Soares, A., (Ed.), Geostatistics Tróia' 92. Kluwer Academic, Dordrecht, pp. 953-964.

Kleingeld, W.J, Thurston, M.L, Prins, C.F., Lantuéjoul, C., 1997. The conditional simulation of a Cox process with application to deposits with discrete particles. In: Baafi, E.Y, Schofield, N.A., (Eds.), Geostatistics Wollongong' 96. Kluwer Academic, Dordrecht, pp. 683-694.

Lantuéjoul, C., 2002. Geostatistical Simulation: Models and Algorithms. Springer, Berlin, 256 pp.

Leite, A., Dimitrakopoulos, R., 2007. Stochastic optimisation model for open pit mine planning: application and risk analysis at copper deposit. IMM Transactions (Mining Technology) 116 (3), 109-118.

Matérn, B., 1986. Spatial Variation, Second edition. Springer-Verlag, Berlin, 151 pp.

Matheron, G., 1984. Isofactorial models and change of support. In: Verly, G., David, M., Journel, A.G., Maréchal, A., (Eds.), Geostatistics for Natural Resources Characterization. Reidel, Dordrecht, pp. 449-467.

Matheron, G., 1985. Comparaison de quelques distributions du point de vue de la sélectivité (Study of the selectivity index properties in the case of different statistical distributions). Sciences de la Terre 24, 1-21.

McNeill, L., 1991. Interpolation and smoothing of binomial data for the Southern African Bird Atlas Project. South African Statistical Journal 25 (2), 129-136.

Metropolis, N., Rosenbluth, A.W., Teller, A.H., Teller, E., 1953. Equation of state calculation by fast computing machines. Journal of Chemical Physics 21 (6), 1087-1092.

Michael, J.R., Schucany, W.R., Haas, R.W., 1976. Generating random variates using transformations with multiple roots. The American Statistician 30 (2), 88-90.

Møller, J., Syversveen, A.R., Waagepetersen, R.P., 1998. Log-Gaussian Cox processes. Scandinavian Journal of Statistics 25 (3), 451-482.

Møller, J., Waagepetersen, R.P., 2003. Statistical Inference and Simulation for Spatial Point Processes. Chapman and Hall/CRC, Boca Raton, 300 pp.

Monestiez, P., Dubroca, L., Bonnin, E., Durbec, J.P., Guinet, C., 2006. Geostatistical modelling of spatial distribution of Balaenoptera physalus in the Northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. Ecological Modelling 193 (3-4), 615-628.

Nicholas, G., Coward, S., Ferreira, J., 2008. Financial risk assessment using conditional simulations in an integrated evaluation model. In: Ortiz, J.M., Emery, X., (Eds.),

Proceedings of the Eighth International Geostatistics Congress Geostats 2008. Quebecor World Chile, Santiago, pp. 759-768.

Oliver, M.A., Lajaunie, C., Webster, R., Muir, K.R., Mann, J.R., 1993. Estimating the risk of childhood cancer. In: Soares, A., (Ed.), Geostatistics Tróia' 92. Kluwer Academic, Dordrecht, pp. 899-910.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 2007. Numerical Recipes: the Art of Scientific Computing. Cambridge University Press, Cambridge, 1256 pp.

Rivoirard, J., Simmonds, J., Footes, K.G., Fernandes, P., Bez, N., 2000 Geostatistics for Estimating Fish Abundance. Wiley, New York, 216 pp.

Sichel, H.S., 1973. Statistical valuation of diamondiferous deposits. Journal of the South African Institute of Mining and Metallurgy 73 (7), 235-243.

Sichel, H.S., 1974. On a distribution representing sentence-length in written prose. Journal of the Royal Statistical Society A137, 25-34.

Stein, G.Z., Zucchini, W., Juritz, J.M., 1987. Parameter estimation for the Sichel distribution and its multivariate extension. Journal of the American Statistical Association 82 (399), 938-944.

Stoyan, D., Penttinen, A., 2000. Recent applications of point process methods in forestry statistics. Statistical Science 15 (1), 61-78.

Von Neumann, J., 1951. Various techniques used in connection with random digits. U.S. National Bureau of Standards, Applied Mathematics Series 12, 36-38.

Wolpert, R.L, Ickstadt, K., 1998. Poisson / gamma random field models for spatial statistics. Biometrika 85 (2), 251-267.

FIGURE CAPTIONS

Figure 1. A, non-conditional realization, B, conditioning data, C and D, two conditional realizations

Figure 2. Location map and histogram of count data

Figure 3. Sample (circles and dashed lines) and modeled (solid lines) variograms and madograms of count data along main anisotropy directions

Figure 4. Two realizations and average of 1000 realizations of number of trees

Figure 5. A, Harvesting sequence and B, recovery curves (number of harvested trees per day) for 1000 realizations

TABLES

```
                Parameters for COXINFER
                ***********************

START OF PARAMETERS:
0.5  0.5  -0.5                  % parameters (a,b,alpha) of distribution modeling the count data
0.8                             % shift parameter (delta) for chi square random field
2 0.15                          % number of nested structures, nugget effect
1 0.45 170 120 100 30 0 0 1     %        1st structure: it cc a1 a2 a3 ang1 ang2 ang3 bb
2 0.40 100 100  50  0 0 0 1     %        2nd structure: it cc a1 a2 a3 ang1 ang2 ang3 bb
30  0                           % azimuth, dip for variogram calculations
20.0                            % lag distance
10                              % number of lags
countvariogram.out              % name of output file with variogram / madogram
table.trn                       % name of output file with transformation table

Available covariance model types:
        1: spherical
        2: exponential
        3: gamma (parameter bb > 0)
        4: stable (parameter bb < 2)
        5: cubic
        6: Gaussian
        7: cardinal sine
        8: J-Bessel (parameter bb > 0.5)
        9: K-Bessel (parameter bb > 0)
       10: generalized Cauchy (parameter bb > 0)
       11: exponential sine
```

Table 1. Default parameter file for program COXINFER

```
                 Parameters for COXSIMU
                 **********************

START OF PARAMETERS:
0                                 % type of simulation: 0=gridded locations; 1=scattered locations
locations.prn                     % if =1: file with coordinates of locations for simulation
1 2 3                             %        columns for location coordinates
0.0 0.0 0.0                       % if =0: x0, y0, z0
100 100 10                        %        nx, ny, nz
1.0 1.0 10.0                      %        dx, dy, dz
1 1 1                             % block discretization (1 1 1 for point-support simulation)
counts.dat                        % file with conditioning data
1 2 3                             %        columns for coordinates
4                                 %        column for count data
-1  100                           %        trimming limits for count data
0.5  0.5  -0.5                    % parameters (a,b,alpha) of distribution modeling the count data
0.8                               % shift parameter (delta) for chi square random field
table.trn                         % file with conversion table (raw-chi square) for the potential field
2 0.15                            % number of nested structures, nugget effect
1 0.45 170 120 100 30 0 0 1 1000 %        1st structure: it cc a1 a2 a3 ang1 ang2 ang3 bb nlines
2 0.40 100 100  50  0 0 0 1 1000 %        2nd structure: it cc a1 a2 a3 ang1 ang2 ang3 bb nlines
30                                % number of realizations
9784498                           % seed for random number generation
500                               % maximum number of locations to simulate simultaneously
100                               % maximum number of iterations for the Gibbs sampler
200 200 100                       % maximum search radii in the rotated system
30 0 0                            % angles for search ellipsoid
1                                 % divide ellipsoid into octants? 1=yes, 0=no
4                                 % optimal number of data per octant (if octant=1) or in total (if 0)
coxsimu.out                       % name of output file
1                                 % create a header in output file? 1=yes, 0=no

Available covariance model types:
        1: spherical
        2: exponential
        3: gamma (parameter bb > 0)
        4: stable (parameter bb < 2)
        5: cubic
        6: Gaussian
        7: cardinal sine
        8: J-Bessel (parameter bb > 0.5)
        9: K-Bessel (parameter bb > 0)
       10: generalized Cauchy (parameter bb > 0)
       11: exponential sine
```

Table 2. Default parameter file for program COXSIMU

Statistics	*Actual value*	*Ideal value*
mean error	0.089	0
mean absolute error	5.143	smallest possible
mean squared error	48.37	smallest possible
regression slope	0.956	1
goodness statistics	0.927	1

Table 3. Cross-validation statistics on 108 data, in order to check for accuracy of predictions and of local uncertainty measures.

	minimum	maximum	mean	quantile 5%	quantile 95%
Total number of trees	163,347	185,592	175,545	169,528	181,405

Table 4. Statistics on simulated numbers of trees within entire domain (5662 nodes, corresponding to a surface of 283.1 ha).