INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Optimization Techniques for Energy Minimization Problem in a Marked Point Process Application to Forestry

Guillaume Perrin — Xavier Descombes — Josiane Zerubia

N° 5704

September 2005

Thème COG

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--5704--FR+ENG

Optimization Techniques for Energy Minimization Problem in a Marked Point Process Application to Forestry

Guillaume Perrin , Xavier Descombes , Josiane Zerubia

Thème COG — Systèmes cognitifs
Projet Ariana

Rapport de recherche n° 5704 — September 2005 — 31 pages

Abstract: We use marked point processes to detect an unknown number of trees from high resolution aerial images. This approach turns to be an energy minimization problem, where the energy contains a prior term which takes into account the geometrical properties of the objects, and a data term to match these objects onto the image. This stochastic process is simulated via a Reversible Jump Markov Chain Monte Carlo procedure, which embeds a Simulated Annealing scheme to extract the best configuration of objects.
We compare in this paper different cooling schedules of the Simulated Annealing algorithm which could provide some good minimization in a short time. We also study some adaptive proposition kernels.

Key-words: Marked point process, RJMCMC, Simulated Annealing, cooling schedule, Dobrushin's contraction coefficient, application to forestry.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Recuit simulé adaptatif pour la minimisation d'énergie dans le cadre de processus ponctuels marqués appliqués à la foresterie

Résumé : Dans ce rapport de recherche, nous utilisons les processus ponctuels marqués afin d'extraire un nombre inconnu d'objets dans des images aériennes. Ces processus sont définis par une énergie, qui contient un terme a priori formalisant les interactions entre objets ainsi qu'un terme d'attache aux données. Nous cherchons à minimiser cette énergie, afin d'obtenir la meilleure configuration d'objets, à l'aide d'un recuit simulé qui s'inscrit dans l'algorithme d'échantillonnage MCMC à sauts réversibles.
Nous comparons ici différents schémas de décroissance de température, et présentons certaines méthodes qui permettent d'améliorer la convergence de l'algorithme en un temps fini.

Mots-clés : Processus ponctuels marqués, RJMCMC, recuit simulé, schéma de décroissance de la température, coefficient de contraction de Dobrushin, application à la foresterie.

Contents

1 Introduction

We aim at extracting tree crowns from remotely sensed images in order to assess some useful parameters such as the number of trees, their diameter, and the density of the stem. This problem has been widely tackled in the literature over the past years. In the case of color infrared images, some methods use a pixel based approach and give the delineation of the tree crowns [Gou98], other ones use an object based approach by modelling a synthetic tree crown template to find the tree top positions [Lar99].

In [PDZ05], we proposed to use a marked point process approach which can embed most of the geometric properties describing the distribution of the trees, especially in plantations where we obtained good results. Indeed, marked point processes enable to model complex geometrical objects in a scene and have been exploited for different applications in image processing [DKL+04]. The context is stochastic, and our goal is to minimize an energy on the state space of all possible configurations of objects, using some Markov Chain Monte Carlo (MCMC) algorithms and Simulated Annealing (SA). In this report, we will focus on the optimization problem.

The first section is dedicated to recall some definitions about marked point processes. Then, we present our model adapted to tree crown extraction, and the SA algorithm. Some convergence theorems are detailed. In the last section, we perform a range of tests in order to study acceleration techniques that can be used to get good results in a faster way.

2 Definitions and Notations

For more details about marked point processes we refer to [vL00], and for their applications to image processing to [DKL+04].

2.1 Marked point process

Let S be a set of interest, called the state space, typically a subset of $\mathbb{R}^n$. A configuration of objects in S is an unordered list of objects :

$$\mathbf{x} = \{x_1, \ldots, x_n\}, x_i \in S, i = 1, \ldots, n \tag{1}$$

A point process X in S is a measurable mapping from a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ to configurations of points of S, in other words a random variable whose realizations are random configurations of points. These configurations $\mathbf{x}$ belong to

$$\Psi = \bigcup_n \Psi_n \tag{2}$$

where Ψ_n contains all configurations of a finite number n of points of S.

A marked point process living in $S = \mathcal{P} \times \mathcal{K}$ is a point process where some marks in $\mathcal{K}$ are added to the positions of the points in $\mathcal{P}$. A configuration of objects $\mathbf{x} = \{(p_1, k_1), \ldots, (p_n, k_n)\}$ is also a finite set of marked points. The marks are some parameters that fully describe the object. For example, ellipses are described by the position of their center, their major and minor axis, and their orientation.

The most obvious example of point processes is the homogeneous Poisson process of intensity measure $\nu(.)$, proportional to the Lebesgue measure on S. It induces a complete spatial randomness, given the fact that the positions are uniformly and independently distributed.

2.2 Application to object extraction

The marked point process framework has been successfully applied in different image analysis problems [DKL$^+$04], the main issue being that we do not know a priori the number of objects to be extracted. Since some pioneering work for cell segmentation [BVL93, RH99], it has been used for road network detection [LDZB04, SDZ00], building extraction [Ort04, ODZ04], and tree crown extraction on LIDAR [And03] or on near infrared images [PDZ05].

The approach consists of modelling an observed image $\mathcal{I}$ (see Fig. (3) lefthandside) as a realization of a marked point process of geometrical objects, such as discs, ellipses, or rectangles for instance. The position state space $\mathcal{P}$ will be given by the image support, and the mark state space $\mathcal{K}$ will be some compact set of $\mathbb{R}^d$.

2.3 Energy minimization problem

We consider the probability distribution $\mu(.)$ of an homogeneous Poisson process living in S with intensity measure $\nu(.)$, which gives us a probability measure on Ψ (see [vL00]). We have $\forall B \in \mathcal{B}(\Psi)$:

$$\mu(B) = e^{-\nu(S)} \left(\mathbf{1}_{[\emptyset \in B]} + \sum_{n=1}^{\infty} \frac{\mu_n(B)}{n!} \right) \tag{3}$$

where

$$\mu_n(B) = \int \ldots \int \mathbf{1}_{[\{x_1,\ldots,x_n\}\in B]}\nu(dx_1)\ldots\nu(dx_n).$$

Then, if the probability distribution of a marked point process X , written $\mathcal{P}_X(.)$, is uniformly continuous with respect to $\mu(.)$, the Radon Nikodym theorem (for more details see [Hal50] for example) defines its unnormalized density $f(.)$ with respect to this dominating reference measure as :

$$\mathcal{P}_X(d\mathbf{x}) = \frac{1}{Z}f(\mathbf{x})\mu(d\mathbf{x}) = \frac{1}{Z}\exp(-U(\mathbf{x}))\mu(d\mathbf{x}) \tag{4}$$

where Z is a normalizing constant, and $U(\mathbf{x})$ the energy of the configuration $\mathbf{x}$.

Within the Bayesian framework, given the data $\mathcal{I}$, $\mathbf{x}$ is assumed to have been generated thanks to a posterior distribution $f(\mathbf{x}|\mathcal{I})$, which can be written as :

$$f(\mathbf{x}|\mathcal{I}) \propto f_p(\mathbf{x})\mathcal{L}(\mathcal{I}|\mathbf{x}) \tag{5}$$

From now on we will write $f(\mathbf{x}) = f(\mathbf{x}|\mathcal{I})$. We aim at finding the Maximum A Posteriori estimator $\mathbf{x}_{MAP}$ of this density, which is also the minimum of the Gibbs energy $U(\mathbf{x})$. As for many energy minimization problems, the prior term $f_p(\mathbf{x})$ can be seen as a regularization or a penalization term, while the likelihood $\mathcal{L}(\mathcal{I}|\mathbf{x})$ can be seen as a data term. We note Ψ_{min} the subset of Ψ where the energy is minimized, and $U_{min} > -\infty$ the value of this energy. By analogy, $U_{max} \leq +\infty$ is the maximum of $U(\mathbf{x})$.

The landscape of $U(\mathbf{x})$ in the problem of tree crown extraction is very elaborate as we will see in the sequel. $U(\mathbf{x})$ contains a lot of local minima [PDZ05]. That is why the classical SA scheme has to be adapted in order to give a good estimation of $\mathbf{x}_{MAP}$ in a reasonable time.

2.4 Simulation of Marked Point Processes

A marked point process X is fully defined by its unnormalized density $f(\mathbf{x})$ under a reference measure, which is often in practice the homogeneous Poisson measure. Sampling a realization of this process is not obvious, this requires some MCMC algorithms, with $\mathcal{P}_X(d\mathbf{x})$ as equilibrium distribution.

In particular, the Reversible Jump MCMC (RJMCMC) algorithm [Gre95] allows us to build a Markov Chain (X_n) which jumps between the different dimensions of Ψ. At each step, the transition of this chain is managed by a set of proposition kernels $\{Q_m(\mathbf{x},.)\}_{m\in M}$ which propose the transformation of the current configuration $\mathbf{x}$ into a new configuration $\mathbf{y}$. This move is accepted with a probability $\alpha = \min\{1, R(\mathbf{x},\mathbf{y})\}$, where :

$$R(\mathbf{x},\mathbf{y}) = \frac{\mathcal{P}_X(d\mathbf{y})Q_m(\mathbf{y},d\mathbf{x})}{\mathcal{P}_X(d\mathbf{x})Q_m(\mathbf{x},d\mathbf{y})} \tag{6}$$

is called the Green ratio.

The Markov chain (X_n) converges ergodically to the distribution $\mathcal{P}_X$ under some stability condition on the Papangelou conditional intensity [vL00] which must be bounded :

$$\lambda(\mathbf{x},u) = \frac{f(\mathbf{x}\cup\{u\})}{f(\mathbf{x})} \leq M \tag{7}$$

with $M > 0$ finite, for all $\mathbf{x} \in \Psi$ and $u \in S$. When (7) holds, the ergodic convergence of (X_n) is ensured (see [GM98], with a kernel containing uniform birth and death).

Model optimization is achieved by this RJMCMC algorithm embedded in a SA scheme. This consists of sampling $f_n(.) = f^{\frac{1}{T_n}}(.)$ instead of $f(.)$, where T_n is a temperature parameter which tends to zero as $n \to \infty$. The Markov chain (X_n) is now nonhomogeneous, and the convergence properties detailed in [GM98] do not apply anymore. This will be detailed in the next sections.

3 Our model for tree crown extraction

We aim at extracting tree crowns from remotely sensed images of forests. Our data contain infrared information, which enhances the chlorophyllean matter of the trees. To perform this object extraction, we use a marked point process of ellipses. The associated state space S is therefore a bounded set of $\mathbb{R}^5$:

$$S = \mathcal{P} \times \mathcal{K} = [0, X_M] \times [0, Y_M] \times [a_m, a_M] \times [b_m, b_M] \times [0, \pi[\tag{8}$$

where X_M and Y_M are respectively the width and the length of the image $\mathcal{I}$, (a_m, a_M) and (b_m, b_M) respectively the minimum and the maximum of the major and the minor axes, and $\theta \in [0, \pi[$ the orientation of our objects. In practice we set $a_m = b_m$ and $a_M = b_M$. In order not to have two points of S representing the same object, we set $b \leq a$.

As explained in Section (2.3), we work in the Bayesian framework : the density $f(\mathbf{x})$ of one configuration is split into a prior term and a likelihood term.

3.1 Prior energy $U_p(\mathbf{x})$

The prior term gives a general aspect of the solution we desire, by adding some constraints to the configurations that should fit the data. As we are working on plantations of poplars, we model the periodic pattern of the alignments. Thus, we consider three components in the prior energy (see [PDZ05] for more details) :

- a repulsive term between two overlapping objects $x_i \sim_r x_j$ in order to avoid over-detection. An overlapping coefficient $\mathcal{A}(x_1, x_2) \in [0, 1]$ penalizes more or less $x_1 \sim_r x_2$ depending on the overlapping area :

$$U_r(\mathbf{x}) = \gamma_r \sum_{x_i \sim_r x_j} \mathcal{A}(x_i, x_j), \ \ \gamma_r \in \mathbb{R}^+ \tag{9}$$

- an attractive term that favours regular alignments in the configuration. A quality function $\mathcal{Q}(x_1, x_2) \in [0, 1]$ quantifies the quality of the alignment of two objects $x_1 \sim_a x_2$, with respect to two predefined vectors of alignments :

$$U_a(\mathbf{x}) = \gamma_a \sum_{x_i \sim_a x_j} \mathcal{Q}(x_i, x_j), \ \ \gamma_a \in \mathbb{R}^- \tag{10}$$

- for stability reasons and because of the attractive term, we have to avoid extreme closeness of objects. This can be done by adding a hard core constraint in our prior process :

$$U_h(\mathbf{x}) = \begin{cases} +\infty & \text{if } \exists (x_i, x_j) \in \mathbf{x} \mid d(x_i, x_j) < 1 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

The first two neighbourhood relationships are shown in Fig. (1). Then, the prior energy is :

$$U_p(\mathbf{x}) = U_r(\mathbf{x}) + U_a(\mathbf{x}) + U_h(\mathbf{x}) \tag{12}$$

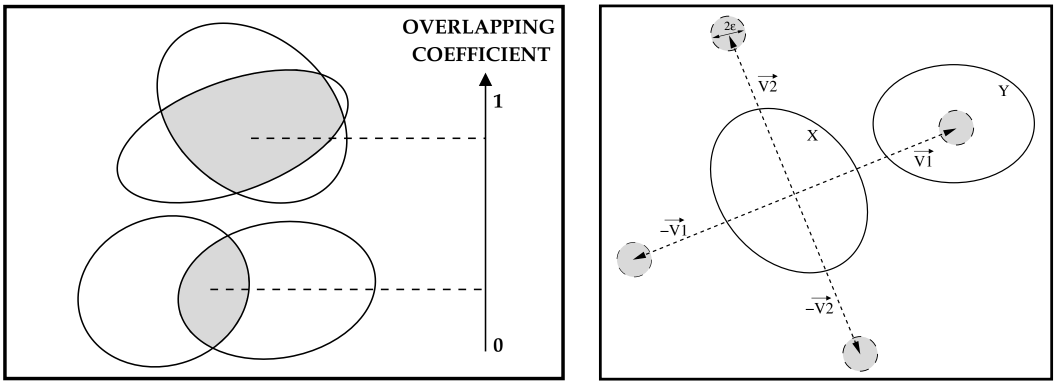

Figure 1: Lefthandside : two overlapping objects and the quality of this interaction. Righthandside : The four regions around one object where some alignments are favoured.

3.2 Likelihood $\mathcal{L}(\mathcal{I}|\mathbf{x})$

The likelihood of the data $\mathcal{I}$ given a configuration $\mathbf{x}$ is a statistical model of the image. We consider that the data can be represented by some Gaussian mixture of two classes (the trees with some high grey value and the background with low grey value), where each pixel is associated to one of these two classes :

- $\mathcal{C}_i = \mathcal{N}(m_i, \sigma_i)$ for the pixels inside at least one of the objects of the configuration,
- $\mathcal{C}_o = \mathcal{N}(m_o, \sigma_o)$ for the pixels outside.

Other models to define the likelihood are studied in [PDZ05].

3.3 RJMCMC algorithm

As explained in Section (2.4), we use a RJMCMC dynamics in order to sample our marked point process. The global proposition kernel $Q(\mathbf{x}, .)$ contains uniform birth and death (Q_{BD}), translation (Q_T), dilation (Q_D), rotation (Q_R), split and merge (Q_{SM}), and birth and death in a neighbourhood (Q_{BDN}). More details about this kernel can be found in [PDZ05]. Fig. (2) gives an overview of some of these moves.

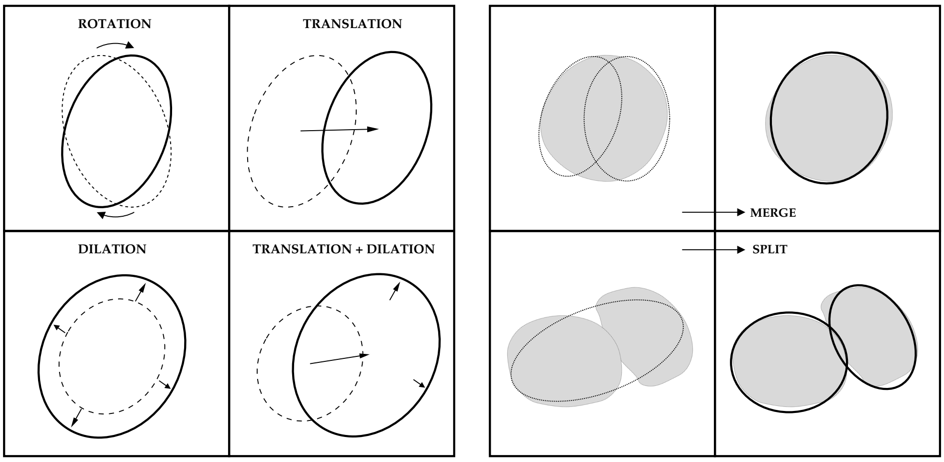

Figure 2: Lefthandside : non jumping kernels, i.e. moves that do not affect the number of objects of the configuration. Righthandside : split and merge kernel, the grey shapes can be seen as the trees we want to extract.

3.4 Data and results

Here we present one extraction result obtained on an aerial image of forests provided by the French Forest Inventory (IFN). The parameters of the model can be found in [PDZ05]. Our goal is to compare this result to those obtained in the next sections, that is why a very slow plateau cooling schedule was chosen in order to have a good estimation of U_{min}. Tab. (1) presents some statistics related to the simulation, and Fig. (3) shows the image and the extraction result.

Number of iterations	$N = 30$ millions
Length of plateau	$k = 5000$
Cooling parameter	$\alpha = 0.9995$
Final energy	$U_N = 134662$
Number of objects	$n(X_N) = 292$

Table 1: Statistics about the extraction after a slow decrease of the temperature.

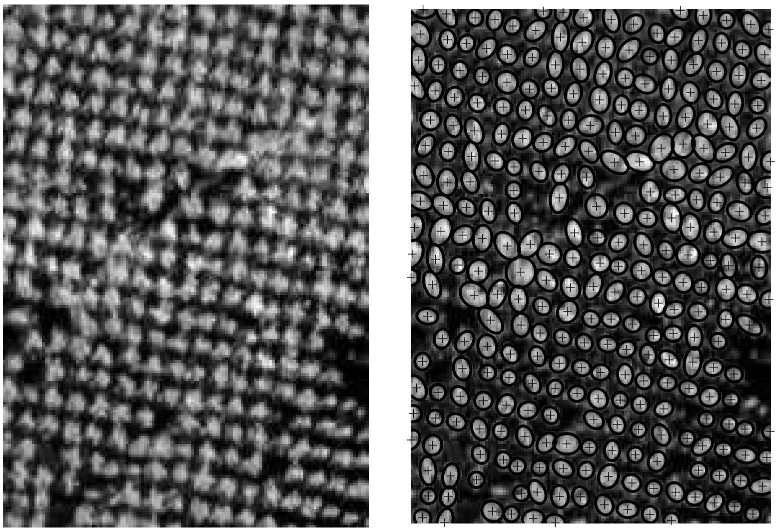

Figure 3: Lefthandside : some data provided by IFN (resolution 50cm/pixel). Righthandside : the poplar extraction result, after 30 millions iterations (45 minutes on a Red Hat Linux 3GHz machine).

In the following, as our algorithm is stochastic, each scenario will be simulated 10 times, and the statistics will represent the mean values of the statistics observed during these simulations.

4 Simulated Annealing

The SA algorithm is a stochastic global optimization procedure which exploits an analogy between the search of the minima $\mathbf{x} \in \Psi_{min}$ and a statistical thermodynamic problem. As explained in [vLA87], the thermal mobility of molecules is lost when the temperature is slowly decreased, going from a liquid state to a pure crystal which is a stable state of minimum energy. Cerny [Cer85] and Kirpatrick [KGV83] simultaneously applied this analogy and proposed an algorithm using a Metropolis Monte Carlo method and some temperature parameters, in order to find the global minima of an energy function $U(x)$. Ever since it has been widely used in many optimization problems including applications in image analysis [Win03]. In this section, we present this algorithm, its convergence properties and some ideas to accelerate it.

4.1 The algorithm

At each iteration of the SA algorithm, a candidate point y is generated from the current position x using the distribution $Q(x, .)$, and accepted or refused via an acceptance ratio $\alpha(x, y)$. This acceptance ratio is controlled by a parameter T called the temperature [Cer85, KGV83], which slowly decreases to 0 during the algorithm :

$$\alpha(x, y) = \min\left\{1, \exp\left(-\frac{U(y) - U(x)}{T}\right)\right\} \tag{13}$$

The principle of SA is also to accept some random ascent moves since the temperature parameter is high, in order to escape local minima of the energy. When $T \to \infty$, the algorithm tends to sample a uniform distribution on S, while for $T \to 0$ it tends to a Dirac distribution on Ψ_{min}. The evolution of the temperature parameter T_n during the optimization process is called the cooling schedule.

SA also samples a nonhomogeneous Markov chain (X_n) which should find the states of minimum energy Ψ_{min} with an appropriate cooling schedule. A schedule is said to be asymptotically good [Aze92] if it satisfies :

$$\lim_{n \to \infty} P(X_n \in \Psi_{min}) = 1 \tag{14}$$

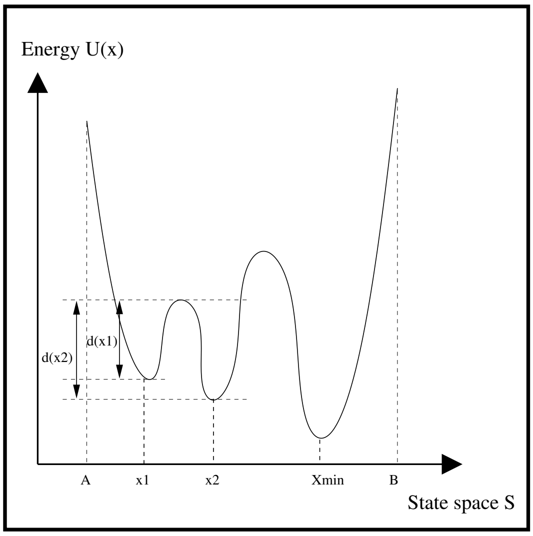

Figure 4: Starting from point B, both SA algorithm and determinist algorithms like steepest descent find the global minimum X_{min}. Starting from A, SA jumps the local minima, but the determinist algorithm is stuck in $x1$. $d(x1)$ and $d(x2)$ are respectively the depths of the local minima $x1$ and $x2$.

4.2 Convergence results

Originally, SA algorithm was introduced for combinatorial problems (especially the Traveling Salesman Problem and Graph Partitioning) on some finite solution set S, but where $\sharp(S)$[1] was very large. Thus, most of the convergence proofs have been obtained in the case of $\sharp(S) < \infty$.

Geman and Geman [GG84] showed that some logarithmic decrease of the temperature

$$\lim_{n\to\infty} T_n \log(n) \geq K > 0 \tag{15}$$

with K large enough, depending on $\Delta = U_{max} - U_{min}$, was sufficient to guarantee the convergence (see Eq. (14)). Then, Hajek [Haj85] proved that K could be the maximum of the depths d_{x_m} of local minima x_m. He also linked the constant K with the energy landscape. We will see how crucial it is to adapt SA to the energy of the problem at hand. Details about this theorem, and speed of convergence can be found in [vLA87].

More recently, some results have been established in the case of general state spaces for continuous global optimization. Most of the time, the case of compact sets was studied [Loc00], while in [ABD01, HS91] more general state spaces were taken. The logarithmic schedule $T_n \geq \frac{K}{\log(n+2)}$ was proven to be asymptotically good for some constant K depending on the energy landscape. This cooling schedule can be accelerated with some restrictions on the proposal distribution $Q(x,.)$.

Finally, point process theory and SA have been linked in [vL93], where some convergence proofs of inhomogeneous Markov chain were established in the case of a birth and death process, and in [SGM04] for general RJMCMC dynamics.

4.3 Convergence theorems in the RJMCMC framework

The proofs of convergence of the non-stationary Markov chain use the Dobrushin's contraction coefficient of the proposition kernel, they can be found in [vL93, SGM04]. We just recall the definitions and the principal theorems of convergence in the sequel.

[1] $\sharp(S)$ is the cardinal of the set S

Definition

Total Variation

Let $\mu_1(.)$ and $\mu_2(.)$ be some probability measures on a measurable space $(\Psi, \mathcal{B}(\Psi))$. Their total variation distance is defined as the maximal difference on measurable subsets $A \in \mathcal{B}(\Psi)$:

$$\|\mu_1 - \mu_2\|_{TV} = \sup_{A \in \mathcal{B}(\Psi)} |\mu(A) - \nu(A)| \tag{16}$$

In the continuous case, if both $\mu_1(.)$ and $\mu_2(.)$ are absotulely continuous with respect to some measure $\mu(.)$ with Radon-Nikodym derivatives f_{μ_1} and f_{μ_2},

$$\|\mu_1 - \mu_2\|_{TV} = \frac{1}{2} \int_\Psi |f_{\mu_1}(\mathbf{x}) - f_{\mu_2}(\mathbf{x})| \mu(d\mathbf{x}) \tag{17}$$

Definition

Dobrushin's contraction coefficient

For a transition probability kernel $Q(.,.)$ on $(\Psi, \mathcal{B}(\Psi))$, Dobrushin's contraction coefficient $c(Q)$ is defined by :

$$c(Q) = \sup_{\mathbf{x},\mathbf{y} \in \Psi} |Q(\mathbf{x}, .) - Q(\mathbf{y}, .)|_{TV} \tag{18}$$

Theorem

Limit theorem

Let (X_n) *be a non-stationary Markov process with a transition kernel* $Q_n(.,.)$ *defined by :*

$$Q_n(\mathbf{x}, A) = Q^{\delta}(\mathbf{x}, A)$$

where $Q^{\delta}(.,.)$ *is the application* δ*-times of the transition kernel* $Q(.,.)$ *defined by the Metropolis Hastings updates of the RJMCMC algorithm. Assume also that conditions*

$$\sum_{n=1}^{\infty} \|\pi_n - \pi_{n+1}\|_{TV} < \infty \qquad (19)$$

and

$$\lim_{n\to\infty} c(Q_{n_0,n}) = 0 \qquad (20)$$

hold $\forall n_0 \geq 0$ *and transition probabilities* $Q_{n_0,n}(\mathbf{x}, A) = \mathbb{P}(X_n \in A | X_{n_0} = \mathbf{x})$*. Then, the limit* $\lim_{n\to\infty} \pi_n = \pi_\infty$ *exists and also* $\lim_{n\to\infty} \pi_0 Q_{0,n} = \pi_\infty$ *in total variation, for any initial distribution* π_0*.*

We need now to prove conditions (19) and (20), which is done in the following.

Theorem **Summability condition**

Let Ψ^* be the set of configurations minimizing the energy $U(\mathbf{x})$ of the model previously defined, and assume that $\mu(\Psi^*) > 0$. Let T_n be a sequence of temperatures such that $\lim_{n\to\infty} T_n = 0$ and consider the probability densities $f^n(.)$ given by

$$f_n(\mathbf{x}) = f(\mathbf{x})^{\frac{1}{T_n}}$$

with respect to the reference measure $\mu(.)$ on Ψ. Then the sequence $\pi_n = \int f_n d\mu$ converges in total variation to the uniform distribution on Ψ^*.

Moreover, the sequence of probability densities satisfies Eq. (19) :

$$\sum_{n=1}^{\infty} \|\pi_n - \pi_{n+1}\|_{TV} < \infty$$

Lemma **Dobrushin condition**

Let $Q_n(\mathbf{x}, A) = Q^{\delta}(\mathbf{x}, A)$ be the transition kernel introduced previously. If the step δ is a finite integer, such that $\delta \geq n(\mathbf{x})$ for all the configurations of objects $\mathbf{x} \in \Psi$, then the Dobrushin's condition (20) holds.

As a conclusion, if such a $\delta \geq n(\mathbf{x})$ can be found (this is our case thanks to the hard core energy which bounds the total number of objects), the nonhomogeneous Markov chain converges and a logarithmic cooling schedule in the SA algorithm (see [SGM04]) is sufficient.

4.4 Acceleration methods

In the previous subsection, we noticed that a logarithmic decrease was needed to ensure the convergence of SA. However, these schedules require too much computation time and are not achievable in practice. Even if we are bound to lose the convergence conditions, we prefer implementing some faster cooling schedules. Many methods have been proposed in that prospect (see [Aze92, Fac00, SSF02, Var96] for instance), and we will test some of them in our application.

The first family of methods consists in adapting the cooling schedule itself [Fac00, vLA87, Var96]. Most of the cooling schedules in the literature rely on exponential, also called geometrical, cooling schedules of the form :

$$T_n = T_0 * a^n \tag{21}$$

with $a < 1$ and very close to 1 (see [vLA87] for more details). A slight adaptation of this schedule brings us to the fixed length plateau cooling schedule, where the temperature parameter is decreased every $k * n$ iterations of the algorithm (k fixed, $n \in \mathbb{N}$). This enables the Markov chain to have more time, at a given temperature, to reach its equilibrium distribution $\mathcal{P}_X^{T_n}(dx) = f(x)^{\frac{1}{T_n}}\mu(dx)$. We will see in the next section the influence of the parameters a and k.

Another family of methods consists in adapting the candidate distribution at each step [GM93, HS91, Ing96]. In the Fast SA [SH87] or in the Adaptive SA (ASA) [Ing96] for instance, the convergence relies on the asumption that we globally explore at each step the feasible region, i.e. that each state is reachable from the current position X_k, though the probability of sampling points far from X_k decreases to 0 when $n \to \infty$. An additional temperature parameter t_n is added in the proposition density of the next candidate, which can decrease to 0 much faster than logarithmically (exponentially for the ASA), while the cooling schedule of the temperature parameter $T_n \to 0$ without any speed constraint.

In practice, we will run some finite schedules, starting from a temperature T_0 and ending with a temperature T_N. Some papers tackle this problem of optimal schedules in some finite time in combinatorial optimization problems [BK94, SK91]. The goal is to find the optimal temperature schedule for a given number of iterations N, i.e. the schedule that gives the lowest expected value of energy. This leads to the "Best So Far" for instance, which consists in taking the minimum value of the energy $U(X_n)$ encountered during the simulation, instead of the "Where You Are" (last) value $U(X_N)$.

However, finite schedules lead all to what is called broken ergodicity [SSF02], especially in applications with a continuous state space of varying dimension. The Markov chain is no longer ergodic because the state space is too big to be explored and would require a huge number of iterations at each temperature, even if there is a nonzero probability of reaching any state from any other state. This phenomenon have been studied in [Ort04], and a new adaptive annealing schedule which takes into account the convergence delay has been proposed. This one was inspired by [Fac00, TH95], where the authors proposed to decrease the temperature in a plateau cooling schedule only if the mean of the energy during the current plateau was bigger than the former one :

$$T_{n+1} = \begin{cases} T_n \text{ if } E[U(X)]_n \leq E[U(X)]_{n-1} \\ a * T_n \text{ otherwise} \end{cases} \tag{22}$$

where $E[U(X)]_n = \frac{1}{k}\sum_{i=kn}^{i=(k+1)n} U(X_i)$. In practice this schedule is very long, that is why Ortner [Ort04] proposed to accept the decrease with less constraints. Moreover, he remarked that the selection of the minimum was done during a specific period of the cooling schedule, called the critical temperature. He also constructed an heuristic that enables the cooling schedule to decrease faster when the region is not interesting, and to go slower, or even to warm the temperature, when the temperature is critical. This adaptive schedule is very interesting because it fits the energy landscape of the problem.

To conclude, in complex optimization problems in a finite time like the problem we have to solve, the adaptability of the algorithm is of high interest, in order to avoid too much computation. We compare in the next section some few results obtained with some of these schedules.

5 Comparative results

Different experiments were carried out in order to compare some cooling schedules of the SA in our energy minimization problem. As explained above, we simulated 10 times each experiment to avoid too much imprecision.

5.1 Initial and final temperature

We still have not discussed the choice of the two bounds T_0 and T_N. In the literature, it is often suggested to link T_0 with the standard deviation of the energy $U(x)$ of random objects $x \in S$, typically twice as large (see [Whi84]). We can, for example, estimate this value by sampling at infinite temperature. For the stopping temperature T_N, it is more difficult to estimate a good value in continuous problems, while in discrete problems [Whi84] suggests taking it of the order of the smallest energy scale. Generally speaking, the cooling schedule should always take into account the energy to be optimized, its scale, its landscape, the number and the size of local minima (see [SSF02]).

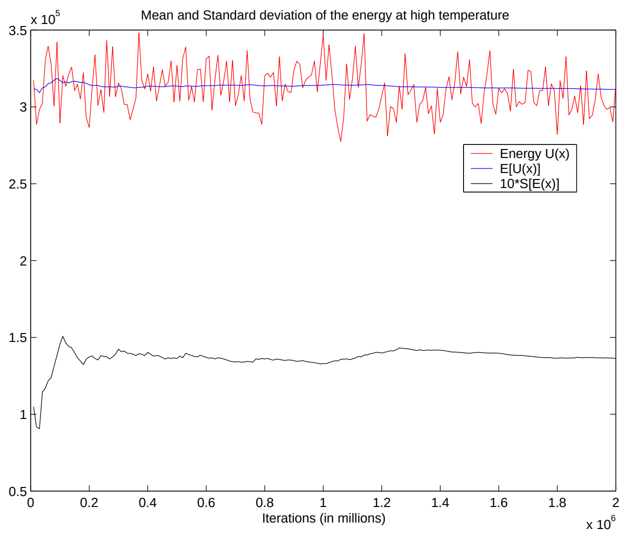

Figure 5: Estimation of the standard deviation of the energy during an infinite temperature simulation.

First, we would like to assess the influence of the initial temperature T_0 in our problem. To that prospect, we use Fig. (5) and take twice the standard deviation as a first initialization of the temperature : $T_0 \simeq 25000$. Then, we compare this value with a bigger value $T_0 = 1000000$ and some smaller ones $T_0 = 100$ and $T_0 = 1$, in a plateau cooling schedule of fixed length $k = 5000$. Only a is varying in order to end the cooling schedule at the same final temperature : $T_N = 10^{-9}$. Some statistics are presented in Tab. (2), and compare the mean value of the last energy $U(X_N)$ (Where-You-Are), the standard deviation of this energy, and the mean value of the total number of extracted objects.

What can be deduced from these results is that the estimation of the initial temperature with the standard deviation is not the best one, considering the final value of the energy. With $T_0 = 100$, we obtain better results for the expectation of the

	$T_0 = 1000000$	$T_0 = 25000$	$T_0 = 100$	$T_0 = 1$
Iterations ($*10^6$)	2	2	2	2
Plateau length k	500	500	500	500
Cooling parameter a	0.9914	0.9923	0.9937	0.9948
Mean $E[U(X_N)]$	136317	136200	**136002**	136713
Standard deviation $\sigma[U(X_N)]$	216.12	222.05	139.07	314.65
Mean $E[n(X_N)]$	282.0	283.6	284.3	285.3

Table 2: Different starting temperatures T_0, with the same number of iterations ($N = 2000000$) and the same final temperature $T_N = 10^{-9}$.

final energy and for its variance, which is the lowest. One can suppose that at higher temperature, we lose some time and that the selection of the minimum begins around $T = 100$ (critical temperature). When the cooling starts with $T_0 = 1 < 100$, it is 'too late'.

Then, we could try to change the final temperature, starting from $T_0 = 100$. The results are presented in Tab. (3). It is interesting to note that it is less important to end with a very low temperature than it is to spend more time around the critical temperature. It seems that $T_N = 10^{-4}$ is a good final temperature, but the varaiance of the final energy is high. We also deduce that the interesting part of the schedule is $100 > T > 10^{-4}$. The best schedule would be perhaps the one which decreases quickly to this critical temperature, then waits, and then goes fast to a very low temperature for a quasi-deterministic schedule.

5.2 Parameters a and k of the plateau cooling schedule

The problem in this cooling schedule is also to determine the two parameters a and k. During how many iterations should the Markov chain X_n stay at a given temperature ? Some tests have been performed in order to see the influence of these parameters. The results are shown in Tab. (4). It appears that the parameters a and k (keeping the same number of iterations) do not have a big impact on the results. This makes sense, because whatever they are, the Markov chain will spend the same time in the critical zone of the temperature.

	$T_N = 10^{-3}$	$T_N = 10^{-4}$	$T_N = 10^{-5}$	$T_N = 10^{-7}$	$T_N = 10^{-9}$	$T_N = 10^{-13}$
Iter. ($*10^6$)	2	2	2	2	2	2
k	500	500	500	500	500	500
a	0.9971	0.9966	0.996	0.9948	0.9937	0.9914
$E[U(X_N)]$	135914	**135843**	135848	135961	136002	136713
$\sigma[U(X_N)]$	113.43	239.11	140.88	187.77	139.07	96.70
$E[n(X_N)]$	284.1	284.7	285.0	284.9	284.3	284.6

Table 3: Different ending temperatures T_N, with the same number of iterations ($N = 2000000$) and the same starting temperature $T_0 = 100$.

	$k = 10$	$k = 100$	$k = 500$	$k = 1000$
Iter. ($*10^6$)	2	2	2	2
a	0.99987	0.99874	0.9937	0.9874
$E[U(X_N)]$	135944	136054	136002	135876
$\sigma[U(X_N)]$	151.24	266.65	139.07	226.07
$E[n(X_N)]$	286.1	285.1	284.3	286.4

Table 4: Different fixed length plateau schedules, with the same number of iterations ($N = 2000000$), initial temperature $T_0 = 100$ and stopping temperature $T_N = 10^{-9}$

5.3 Adaptive cooling schedule

In the previous results, we noticed that some better values for T_0 and T_N could be found. Unfortunately, we cannot afford doing so many simulations for every image for which we want to use our model. That is why the cooling schedule proposed in [Ort04, ODZ04] is interesting, because it adapts the cooling speed to the energy landscape.

We still set $k = 500$, and $k' = 50$ the length of the 10 sub-plateaus. We note

$$E[U(X)]_n^i = \frac{1}{k'} \sum_{j=kn+ik'}^{j=kn+(i+1)k'} U(X_j) \tag{23}$$

the mean of the energy on the i^{th} sub-plateau. We have $E[U(X)]_n = \frac{1}{10}\sum_{i=0}^{i=9} E[U(X)]_n^i$. A decrease of the temperature would be accepted if at least one sub-plateau has a mean energy $E[U(X)]_n^i$ lower than the global former energy $E[U(X)]_{n-1}$. Moreover, the cooling parameter a_n now depends on n, and we can accelerate or even warm the temperature (if the ergodicity is broken) according to :

$$T_{n+1} = \begin{cases} \frac{1}{a_n} * T_n \text{ if } \sharp\left\{E[U(X)]_n^i \leq E[U(X)]_{n-1}\right\} = 0 \text{ and } a_n = a_n^{\frac{1}{r}} \\ a_n * T_n \text{ if } \sharp\left\{E[U(X)]_n^i \leq E[U(X)]_{n-1}\right\} \in [1,4] \\ a_n * T_n \text{ if } \sharp\left\{E[U(X)]_n^i \leq E[U(X)]_{n-1}\right\} \geq 5 \text{ and } a_n = a_n^r \end{cases} \tag{24}$$

We use $r = 0.9$, and threshold the parameter a_n in order that $0.96 < a_n < 0.996$. Obviously, starting from a temperature T_0 and ending at a temperature T_N, we cannot predict the number of iterations it will take.

In Tab. (5), we compare the results of different adaptive schedules with $T_0 = 25000$ and $T_N = 10^{-9}$. S1 is the simple plateau schedule with the constant speed studied before ($a = 0.9923$). Then, S2 and S3 are some adaptive plateau schedules for which the speed changes during the simulation in accordance with the previous observations (see Fig. (6), lefthandside). S2 and S3 spend respectively $p = 80\%$ and $p = 90\%$ of the time in the critical zone $100 > T > 10^{-4}$. S4 is the limit adaptive geometric schedule ($p = 100\%$ studied before) where $T_0 = 100$ and $T_N = 10^{-4}$. Finally, S5 is the adaptive schedule presented above, which accelerates or warm the temperature.

	S1	S2 : 80%	S3 : 90%	S4 : 100%	S5
Iter. ($*10^6$)	2	2	2	2	$E[N] = 2.03$
Plateau length k	500	500	500	500	500
Cooling parameters	$a = 0.9923$	$a_1 = 0.9789$ $a_2 = 0.9957$	$a_1 = 0.9583$ $a_2 = 0.9962$	$a = 0.9966$	$0.96 < a < 0.996$
$E[U(X_N)]$	136200	136048	135952	135843	**135805**
$\sigma[U(X_N)]$	222.05	218.65	193.58	239.11	132.84
Mean $E[n(X_N)]$	283.6	283.5	284.5	284.7	285

Table 5: Comparison of different schedules with $T_0 = 25000$ and $T_N = 10^{-9}$.

As expected, the adaptive schedule makes the most of the N iterations and spend much more time in the interesting part of the temperature (see Fig. (6)). We see

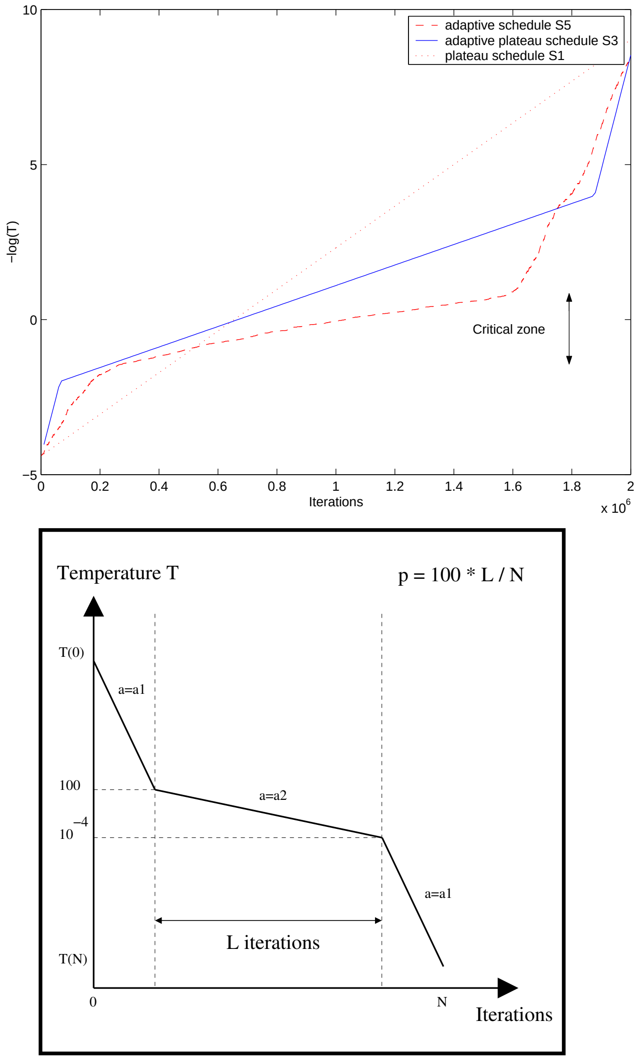

Figure 6: Top : Comparison of the adaptive plateau schedule and the adaptive schedule. Bottom : Cooling schedule of an adaptive plateau schedule, p being the percentage of iterations spent in the critical zone. For $T_0 > T > 100$ and $10^{-4} > T > T_N$, $a_n = a_1$. For $100 > T > 10^{-4}$, $a_n = a_2$. In order to spend more time in the critical zone, $1 > a_2 > a_1$.

that our estimation (around $T = 100$) of the beginning of the critical zone is quite good.

5.4 Adapting the proposition kernel to the temperature

A last optimization can be performed on the proposition kernel $Q(x, y)$. Indeed, this influences the global optimum of the SA, considering that at low temperature, the system is looking preferably for small perturbation. We also decrease along with the temperature our parameters in the translation, the rotation, and the dilation moves in order to propose smaller perturbations. Results are shown in Tab. (6). We can see that the results are much better with an adaptive proposition kernel. This can be understood because whatever the temperature is, the energy of the configuration should vary at the same order on the scale of one transition [SSF02]. At a very low temperature for example, many costly energy evaluations will be required to reach any improvement of the objective function if we propose big moves to some of the objects of the configuration. It seems that keeping a good acceptance ratio even for low temperatures is crucial.

	classical kernel	adaptive kernel
Iterations ($*10^6$)	2	2
Plateau length k	500	500
Cooling parameter a	0.9923	0.9923
$E[U(X_N)]$	136200	**135669**
$\sigma[U(X_N)]$	222.05	78.8
$E[n(X_N)]$	283.6	284.1

Table 6: Comparison between a classical kernel and an adaptive proposition kernel.

6 Conclusion

In this report, we have performed some optimization tests on a marked point process problem algorithm applied to tree crown extraction. It is in fact an energy minimization, which can be studied using a SA scheme embedded in RJMCMC procedure.

First, it appears that the adaptive cooling schedule proposed in [Ort04] gives better results that any geometric schedule, knowing the starting and the ending temperatures of the schedule. It fits well the energy landscape, and accelerates or warms the temperature in order to minimize the broken ergodicity phenomenon, which always occurs in finite schedules. Then, adapting the proposition kernel itself is also interesting. It increases the acceptance ratio of the proposition kernel for low temperatures, and also finds better local minima of the energy.

Future work could involve the implementation of another optimization algorithm, such as genetic algorithms. Moreover, other techniques which could improve the Markov chain speed will be studied. Among all, Data Driven Markov Chain Monte Carlo [TZ02] achieves high acceleration by using the data (clustering or edge detection in the image segmentation problem for instance) to compute importance proposal probabilities which drive the Markov chain.

7 Acknowledgement

The work of the first author has been partly supported by a grant from the MAS Laboratory of Ecole Centrale Paris. The authors would like to thank the French Forest Inventory (IFN) for providing the data and for interesting discussions. Part of this work has been conducted within the INRIA ARC "Mode de Vie" joint research program.

References

[ABD01] C. Andrieu, L.A. Breyer, and A. Doucet. Convergence of Simulated Annealing using Foster-Lyapunov Criteria. *Journal of Applied Probability*, 38(4):975–994, 2001.

[And03] H.E. Andersen. *Estimation of Critical Forest Structure Metrics through the Spatial Analysis of Airborne Laser Scanner Data*. PhD thesis, College of Forestry Resources, University of Washington, Seattle, USA, 2003.

[Aze92] R. Azencott, editor. *Simulated Annealing. Parallelization Techniques.* John Wiley and Sons, 1992.

[BK94] K.D. Boese and A.B. Kahng. Best-so-far vs Where-you-are : Implications for Optimal Finite Time Annealing. *System and Control Letters*, 22:71–78, 1994.

[BVL93] A. Baddeley and M.N.M. Van Lieshout. Stochastic Geometry Models in High-level Vision. In K.V. Mardia, editor, *Statistics and Images*, volume 1, pages 231–256. 1993.

[Cer85] V. Cerny. Thermodynamical Approach to the Traveling Salesman Problem: an Efficient Simulation Algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, 1985.

[DKL+04] X. Descombes, F. Kruggel, C. Lacoste, M. Ortner, G. Perrin, and J. Zerubia. Marked Point Process in Image Analysis : from Context to Geometry. In *SPPA Conference*, Castellon, Spain, April 2004. invited paper.

[Fac00] A. Fachat. *A Comparison of Random Walks with Different Types of Acceptance Probabilities*. PhD thesis, University of Chemnitz, Germany, May 2000.

[GG84] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions and Bayesian Restoration of Images. *IEEE PAMI*, 6:721–741, 1984.

[GM93] S.B. Gelfand and S.K. Mitter. Metropolis-type Annealing Algorithms for Global Optimization in R^d. *SIAM Journal of Control and Optimization*, 31(1):111–131, 1993.

[GM98] C.J. Geyer and J. Moller. Likelihood Inference for Spatial Point Processes. In O.E. Barndoff Nielsen, W.S. Kendall, and M.N.M. van Lieshout, editors, *Stochastic Geometry, Likelihood and Computation*. Chapmann and Hall, London, 1998.

[Gou98] F.A. Gougeon. Automatic Individual Tree Crown Delineation using a Valley-following Algorithm and Rule-based System. In D.A. Hill and D.G. Leckie, editors, *Proc. of the International Forum on Automated Interpretation of High Spatial Resolution Digital Imagery for Forestry*, pages 11–23, Victoria, British Columbia, Canada, February 1998.

[Gre95] P.J. Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika 82*, pages 711–7320, 1995.

[Haj85] B. Hajek. Cooling Schedules for Optimal Annealing. *Mathematics of Operations Research*, 13(2):311–329, 1985.

[Hal50] P.R. Halmos. *Measure Theory*. Springer-Verlag, 1950.

[HS91] H. Haario and E. Saksman. Simulated Annealing Process in General State Space. *Advances Applied Probability*, 23:866–893, 1991.

[Ing96] L. Ingber. Adaptive Simulated Annealing : Lessons Learned. *Control and Cybernetics*, 25(1):33–54, 1996.

[KGV83] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science, Number 4598*, 220:671–680, May 1983.

[Lar99] M. Larsen. Individual Tree Top Position Estimation by Template Voting. In *Proc. of the Fourth International Airborne Remote Sensing Conference and Exhibition / 21^{st} Canadian Symposium on Remote Sensing*, volume 2, pages 83–90, Ottawa, Ontario, June 1999.

[LDZB04] C. Lacoste, X. Descombes, J. Zerubia, and N. Baghdadi. A Bayesian Geometric Model for Line Network Extraction from Satellite Images. In *ICASSP Conference*, Montreal, Quebec, Canada, May 2004.

[Loc00] M. Locatelli. Simulated Annealing Algorithms for Continuous Global Optimization : Convergence Conditions. *Journal of Optimization Theory and Applications*, 104:121–133, 2000.

[ODZ04] M. Ortner, X. Descombes, and J. Zerubia. A Reversible Jump MCMC Sampler for Object Detection in Image Processing. In *MC2QMC Conference*, Antibes Juan Les Pins, France, 2004. to be published in LNS-Springer Verlag.

[Ort04] M. Ortner. *Processus Ponctuels Marqués pour l'Extraction Automatique de Caricatures de Bâtiments à partir de Modèles Numériques d'Élévation.* PhD thesis, University of Nice-Sophia Antipolis, France, October 2004. (in French).

[PDZ05] G. Perrin, X. Descombes, and J. Zerubia. Point Processes in Forestry : an Application to Tree Crown Detection. Research Report 5544, INRIA, April 2005.

[RH99] H. Rue and M. Hurn. Bayesian object identification. *Biometrika*, 86:649–660, May 1999.

[SDZ00] R. Stoica, X. Descombes, and J. Zerubia. Road Extraction in Remotely Sensed Images using a Stochastic Geometry Framework. In *Proc. of the International Workshop Bayesian Inference and Maximum Entropy Methods*, Gif-sur-Yvette, France, 2000.

[SGM04] R. Stoica, P. Gregori, and J. Mateu. Simulated Annealing and Object Point Processes : Tools for Analysis of Spatial Patterns. Technical Report 69, University Jaume I, Castellon, Spain, 2004.

[SH87] H. Szu and R. Hartley. Fast Simulated Annealing. *Physics Letters A*, 122(3-4):157–162, 1987.

[SK91] P.N. Strenski and S. Kirkpatrick. Analysis of Finite Length Annealing Schedules. *Algorithmica*, 6:346–366, 1991.

[SSF02] P. Salamon, P. Sibani, and R. Frost. *Facts, Conjectures, and Improvements for Simulated Annealing.* SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2002.

[TH95] R. Tafelmayer and K.H. Hoffmann. Scaling Features in Complex Optimization Problems. *Computer Physics Communications*, 86:81–90, 1995.

[TZ02] Z.W. Tu and S.C. Zhu. Image Segmentation by Data-Driven Markov Chain Monte Carlo. *IEEE PAMI*, 24(5), 2002.

[Var96] J.M. Varanelli. *On the Acceleration of Simulated Annealing*. PhD thesis, University of Virginia, Charlottesville, USA, May 1996.

[vL93] M.N.M. van Lieshout. Stochastic Annealing for Nearest Neighbour Point Process with Application to Object Recognition. Technical Report BS-R9306, Centruum vor Wiskunde en Informatica, Amsterdam, 1993.

[vL00] M.N.M. van Lieshout. *Markov Point Processes and their Applications.* Imperial College Press, London, 2000.

[vLA87] P.J.M. van Laarhoven and E.H.L. Aarts. *Simulated Annealing : Theory and Applications*. D. Reidel, Boston, 1987.

[Whi84] S.R. White. Concepts of Scale in Simulated Annealing. In *IEEE Proc. of the 1984 International Conference on Computer Design*, pages 646–651, 1984.

[Win03] G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer-Verlag, 2^d edition, 2003.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399