CLUSTER DETECTION IN SPATIAL DATA USING MARKED POINT PROCESSES

BY

Radu S. Stoica
Emilie Gay

Research Report No. 11
July 2005

Unité de Biométrie
Institut National de la Recherche Agronomique
Avignon, France
http://www.avignon.inra.fr/biometrie

Cluster detection in spatial data using marked point processes

R.S. Stoica [1] and E. Gay[2]

INRA Unité Biomètrie
Domaine St. Paul, site Agroparc
84914 Avignon Cedex 9, France

ABSTRACT
This paper proposes a marked point process approach for cluster detection in spatial data. The cluster pattern is supposed made of random interacting disks. The proposed model has two components. The first component is related to the location of the disks in the data field, and it is defined as an inhomogeneous Poisson process. The second one is related to the interaction between disks and it is constructed by the superposition of an area-interaction and a pairwise interaction processes. The model is tested on spatial data coming from animal epidemiology. Statistical descriptors of the cluster are given. These descriptors are the sufficient statistics of the proposed model.

2000 Mathematics Subject Classification: 60G55, 60J22, 62M30, 62M40
Keywords and Phrases: cluster detection, spatial data analysis, marked point processes, area and pair-wise interaction models, Markov chain Monte Carlo simulation, statistical inference, animal epidemiology, subclinical mastitis.

1 Introduction

Pattern detection in digital images using the marked point processes approach is based on two key ideas [2, 10, 18, 27, 31, 32, 33]. First, a pattern is a finite random set $\mathbf{y} = \{y_1, y_2, \dots y_n\}$ with its elements beeing simple interacting objects. Second, the objects forming the pattern are driven by a marked point process.

Let $p(\mathbf{y}|\theta)$ be the probability density of such a process, with θ the model parameters. Under these assumptions, the pattern to be detected $\widehat{\mathbf{y}}$ is esti-

[1] radu.stoica@avignon.inra.fr
[2] emilie.gay@avignon.inra.fr

mated by the configuration of objects maximizing this probability density :

$$\widehat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\Omega}\{p(\mathbf{y}|\theta)\}$$

The probability density can be written as follows

$$p(\mathbf{y}|\theta) \propto \exp\left[-U(\mathbf{y},\theta]\right) = \exp\left[-\left(U_{\mathbf{d}}(\mathbf{y},\theta) + U_i(\mathbf{y},\theta)\right)\right]. \qquad (1)$$

$U(\mathbf{y},\theta)$ is the Gibbs energy of the system. The term $U_{\mathbf{d}}(\mathbf{y},\theta)$ is called the data energy and is related to the position of the objects in the image domain, whereas $U_i(\mathbf{y},\theta)$ represents the interaction energy, and is related to the interaction between the objects in the configuration $\mathbf{y}$.

The strong point of this approach is that it considers the image as a collection of objects instead of numerical values. This leads to robust methods in terms of noise and detection. In remote sensed image analysis, patterns with a complex geometry such as road networks or aligned structures of buildings, were detected using marked point processes manipulating interacting random segments or rectangles [5, 14, 26, 35].

Nevertheless when using such methods, there is an obvious dependence between the model parameters and the obtained result. This fact leads to some natural questions. Are the objects detected, because they realy exist or because we are "insisting" in finding them ? Is their presence detected because of some random effects exhibited by the data ?

The aim of this paper is to apply this approach to cluster detection in spatial data and to bring particular attention to these questions. Similarly with the problems in image analysis, the cluster pattern is considered made of simple interacting objects. These objects are disks with random center and random radius. These disks are driven by a marked point process with probability density having the form (1). The proposed approach is tested on spatial data coming from animal epidemiology.

Spatial data is the generic name for data sets with elements having two components, one related to the location of the element, the second one designing their characteristics. Clearly, digital images are spatial data. In several situations, the location of the elements contained in the data is not always distributed on a regular grid : for instance, the positions of earthquakes in a geographical region, together with their corresponding magnitudes.

Cluster detection is of great importance when analysing spatial data, since it may reveal un-normal behaviour of the monitored phenomenon. Generally, a cluster is considered as a geographically bounded group of occurences of sufficient size and concentration to be unlikely to have occured by chance [6].

The cluster detection problem received a lot of attention when the data consist of point - or event - locations only [1, 4, 13, 17, 20, 24]. The case of spatial data as defined previously was tackled via the construction of spatial interpolators, so to obtain a smooth "map" of the studied phenomenon. Peakedness in this "map", it is often considered as an indicator of cluster presence. Still, the difference between clustering effects and extra variances in the observed process, cannot be easily stated [16, 17, 39].

The structure of the paper is as follows. First, some modelling tools based on marked point processes theory are given. The available data is then presented, together with some exploratory analysis. This data comes from animal epidemiology, hence the presence of clusters may reveal an unnormally high concentration of the disease in a special region. The next section is dedicated to the construction of a marked point process model able to detect cluster patterns in the presented data. The results obtained are shown and interpreted in the fifth section. Finally, conclusions and perspectives are depicted.

2 Modelling framework based on marked point processes

2.1 Definitions. General facts.

Let K be a compact subset of strictly positive measure $0 < \nu(K) < \infty$ in the Lebesgue measure space $(\mathbb{R}^2, \mathcal{B}, \nu)$. Different characteristics or marks may be attached to points in K. Let $(M, \mathcal{M}, \nu_M)$ be the probability measure space of these marks.

A marked point process with locations in K and marks in M is a measurable mapping from some probability space into $(\Omega, \mathcal{F})$. A marked point process is usually called an object point process if the marks represent the geometrical parameters of an object. Here $\Omega = \cup_{n=0}^{\infty} \Xi_n$ is the configuration space, with Ξ_n the set of all unordered configurations $\mathbf{y} = \{(k_i, m_i)\}_{i=1}^{n}$ consisting of not necessarily distinct marked points $y_i = (k_i, m_i) \in K \times M$. Ξ_0 is the empty configuration. $\mathcal{F}$ is the $\sigma-$algebra generated by the mappings that count the number of marked points in Borel sets $A \subseteq K \times M$.

At our knowledge the simplest marked point process is the Poisson marked point process of probability measure

$$\begin{array}{rcl} \mu(F) & = & \sum_{n=0}^{\infty} \frac{e^{-\nu(K)}}{n!} \int_{K \times M} \cdots \int_{K \times M} \mathbf{1}_F\{(k_1, m_1) \ldots, (k_n, m_n)\} \\ & & \times d\nu(k_1) \cdots d\nu(k_n) d\nu_M(m_1) \ldots d\nu_M(m_n) \end{array} \quad (2)$$

for all $F \in \mathcal{F}$. According to a Poisson law of intensity $\nu(K)$, this process distributes points uniformly in K. The point marks are chosen independently according to ν_M.

The Poisson marked point process does not take into account interaction between the marked points. Indeed, more complicated models may be constructed by specifying a Radon-Nikodým derivative $p(\mathbf{y})$ with respect to μ.

Stability conditions need to be fulfiled, for any probability density of a marked point process. The Ruelle's stability condition [34] requires

$$\frac{p(\mathbf{y})}{p(\emptyset)} \leq \Lambda^{n(\mathbf{y})} \tag{3}$$

to hold for a finite constant $\Lambda > 0$ and any $\mathbf{y} \in \Omega$. $n(\mathbf{y})$ is the cardinality of $\mathbf{y}$. The marked point process fulfiling (3) is called stable.

The local stability is a stronger condition than (3) and it is defined as follows

$$\lambda(\xi; \mathbf{y}) = \frac{p(\mathbf{y} \cup \{\xi\})}{p(\mathbf{y})} \leq \Lambda \tag{4}$$

with $\Lambda > 0$ finite, for all $\mathbf{y} \in \Omega$ and $\xi \in K \times M$. $\lambda(\xi; \mathbf{y})$ is called the Papangelou conditional intensity. In this case the marked point process is said to be locally stable.

The stability of a point process (3) ensures the integrability of its probability density with respect to the Poisson reference measure. The local stability (4) often guarantees the necessary convergence properties of the Monte Carlo dynamics simulating a point process.

For a rigorous and detailed presentation of the marked point processes theory, we recommend the monographs [19, 25, 29].

2.2 Tools for modelling

As stated in the beginning the paper, the key idea throughout this work is to consider the cluster pattern as a collection of random disks driven by a marked point process. In order to detect and to form cluster patterns, such a marked point process has to detect the regions in the data where the objects are situated, to agregate these objects to form the clusters and simultaneously to spread them throughout the entire location space in the data.

Clearly, more complicated processes than the reference measure (2) are needed. In the following we will present three marked point processes. These processes are modelling tools in the construction of the solution for our problem.

Tool 1. Inhomogeneous Poisson process *With respect to the reference measure μ, the probability density of an inhomogeneous point proces is defined by*

$$p(\mathbf{y}) \propto \prod_{(k,m)\in\mathbf{y}} \beta(k,m) \tag{5}$$

with the intensity $\beta : K \times M \to \mathbb{R}_+$ beeing a bounded function.

Note that, the Papangelou conditional intensity is given by

$$\lambda(\xi;\mathbf{y}) = \beta(\xi) = \beta(k_\xi, m_\xi).$$

Tool 2. Area interaction process *Consider the mark space $M = [r_{\min}, r_{\max}]$. In the following, $y_i = (k_i, m_i)$ represents the parameters of a disk, i.e. its center and radius, respectively.*

The set $Z(\mathbf{y}) = \cup_{i=1}^{n(\mathbf{y})} b(k_i, m_i)$ is the union of all the disks $b(k_i, m_i) = \{a \in \mathbb{R}^2 : d(a, k_i) \leq m_i\}$.

The area interaction process has the following probability density

$$p(\mathbf{y}) \propto \beta^{n(\mathbf{y})} \gamma_a^{-\nu[Z(\mathbf{y})]} \tag{6}$$

with respect to the reference process (2). The model parameters are $\beta, \gamma_a > 0$.

The local stability ratio is given by

$$\lambda(\xi;\mathbf{y}) = \beta\gamma_a^{-\nu[b(k_\xi, m_\xi)\setminus Z(\mathbf{y})]}.$$

Tool 3. A pairwise interaction - Strauss like - process *As in the previous example, let $\mathbf{y}$ represent a random configuration of disks parameters.*

Let us suppose that overlapping between these objects has to be taken into account. Here, we consider that two disks of parameters $y_i = (k_i, m_i)$ and $y_j = (k_j, m_j)$ overlap and we write $y_i \sim_o y_j$, if the following relation is verified

$$d(k_i, k_j) \leq \max\{m_i, m_j\}$$

The object point process that takes into account the defined interaction has the probability density

$$p(\mathbf{y}) \propto \beta^{n(\mathbf{y})} \gamma_o^{n_o(\mathbf{y})} \tag{7}$$

with respect to the reference measure (2). The model parameters are $\beta > 0$ and $\gamma_o \in (0, 1)$. The sufficient statistics $n(\mathbf{y})$ and $n_o(\mathbf{y})$ represent the total *number of disks in* $\mathbf{y}$ *and the number of pairs of different disks in* $\mathbf{y}$ *that overlap.*

Here, the Papangelou ratio is

$$\lambda(\xi;\mathbf{y}) = \beta\gamma_o^{\tilde{n}_o(\xi,\mathbf{y})}$$

with $\tilde{n}_o(\xi,\mathbf{y})$ *beeing the number of disks in* $\mathbf{y}$ *that overlap with* ξ*.*

The marked point processes presented in these examples are all locally stable. Furthermore these processes belong to the class of (Ripley-Kelly) Markov point processes [30].

The area interaction process was proposed to model patterns of objects that cluster [3]. The model given by (6) forms configurations of disks that tend to be "regular" if $\gamma_a < 1$ or "clustered" whenever $\gamma_a > 1$. A similar model to (7) - in fact the original Strauss process - was used to form clustered patterns using a parameter $\gamma_o > 1$. It turned out that, in this case, the corresponding probability density is not integrable [11, 38]. Some more recent models propose alternative ways of clustering objects defining the connectivity of an object [21, 36].

The Gibbs energy of these processes is $U(\mathbf{y}) = -\log p(\mathbf{y})$. The definition of a model in energetical terms is sometimes prefered in modelling because of the more intuitive description of the considered phenomenon. In this case, the Papangelou ratio represents the energetical contribution of a particle to the considered system.

2.3 Simulation methods

Several Monte Carlo techniques are available for simulating marked point processes : spatial birth-and-death processes, reversible jump dynamics or more recently exact simulation techniques [7, 8, 9, 12, 19, 22, 28].

Exact simulation methods have the advantage of indicating by themselves when convergence is reached. Still, methods such as coupling from the past or clan of ancestors are efficient in practice only within a limited range of parameters [22]. Since, the spatial birth-and-death processes are the core of the mentioned exact simulation methods, the same drawbacks may discard this choice, too.

The Metropolis-Hastings paradigm - a particular case of the reversible jump framework - is generaly prefered because the models can be easily simulated for the whole range of parameters. Furthermore, this technique allows the use of transition kernels tailored to the model to simulate. The simulations throughout this paper are done using a Metropolis-Hastings dynamics as in [7, 8].

2.4 Inference

Our problem consists of infering the cluster pattern - *i.e.* the positions and the characteristics of the cluster pattern $\mathbf{y}$ - from the spatial data $\mathbf{d}$. The estimator obtained maximizing (1) clearly depends on the model parameters.

The ideal solution would be to simultaneously perform pattern detection and parameter estimation. The image analysis community using Markov random fields showed a great interest for this kind of approaches [15, 40, 41]. Intuitively, a theoretical extrapolation of these methods for marked point processes can lead to a possible answer to this question. Still, the image segmentation method presented by [15] performs also parameter estimation but no convergence guarantees are given. Ideas used for parameter estimation in image analysis [40, 41] can be also found in the point processes litterature [8]. The parameter estimation for marked point processes can be formulated under the complete or the missing data framework. This formulation requires a total - or a partial - observation of the sufficient statistics of the pattern to be detected. Clearly, for the problem on hand such an observation is not available. The only thing we observe is the data field "hidding" the pattern we are looking for.

Under these circumstances, a possible answer is to model the parameters too, using a prior law $p(\theta)$. The uniform law is commonly used, when no particular knowledge about the parameters is available. Hence, the pattern estimator can be written as follows

$$\widehat{\mathbf{y}} = \arg\max_{\Omega\times\Theta} p(\mathbf{y},\theta) = \arg\max_{\Omega\times\Theta} p(\mathbf{y}|\theta)p(\theta) \tag{8}$$

where Θ represents the parameters space.

The estimated object configuration given by (8) can be computed using a simulated annealing algorithm [18, 36].The obtained solution is a random object configuration in the conifiguration sub-space maximizing $p(\mathbf{y},\theta)$. The obtained solution is not unique.

Therefore, we are interested in how often a spatial region $\mathcal{R}$ in the data is considered to be part of the pattern. Let us define the following quantity

$$N_{\mathcal{R}} = \mathbb{E}[\mathbf{1}\{\mathcal{R} \subseteq Z(Y)\}] = \int_{\Omega\times\Theta} \mathbf{1}\{\mathcal{R} \subseteq Z(\mathbf{y})\} p(\mathbf{y},\theta)\mu(d\mathbf{y})d\theta \tag{9}$$

as the visit number of the spatial region $\mathcal{R}$ by the random pattern Y. Since, the integral in (9) is not available analytically, the Monte Carlo approximation

$$\widehat{N_{\mathcal{R}}} = \frac{1}{J}\sum_{j=1}^{J} \mathbf{1}\{\mathcal{R} \subseteq Z(Y_j)\} \tag{10}$$

may be used. $\{Y_1, \ldots, Y_J\}$ are samples of $p(\mathbf{y}, \theta)$.

There is an analogy of the equation (9) with formulas coming from stochastic geometry. In [37], the authors derive analytical expressions for the Boolean model. One direct application of such formulas is to test the hypothesis of completely randomness - *Poissonianity* - for a pattern of objects. In the following, we will use the presented methods in order to make inference not related to the pattern of objects, but rather to the data "covered" by the pattern.

3 Presentation of the data

In this paper, the data to analyse are made of elements having two components. The first component represents the location of a farm on the territory of France, given by the center of the commune to which the farm belongs. Hence, the data exhibits multiple points at the same location. These farms are dairy herds breeding Holstein cows only. To each farm a continous variable is attached. This variable represents the annual somatic cell score, an indicator for subclinical mastitis. The range of the variable value is approximately between 1 and 5. In contrast with the high values, the low ones are interpreted as "healthy". Each data set represents the registration of more than $30,000$ farms with a general cellular score computed for a whole year. Five data sets are available, representing the years 1996 to 2000, included.

This disease is endemic. Its random spread all over the territory is considered as usual by the epidemiologists. Therefore, the detection of spatial clusters is important, since it may indicate a locally un-normal situation.

The cluster definition adopted in the beginning of the paper forces us to empiricaly decide when the cellular score is high. In Figure 1 the histograms for each data set are plotted. Through all the cases, a significant Gaussian shape can be noticed. For each data set i, a threshold value is computed

$$d_{0i} = \hat{\mu_i} + \hat{\sigma_i}$$

where $\hat{\mu_i}$ and $\hat{\sigma_i}$ are the corresponding estimated mean and standard deviation, respectively. These threshold values are shown in Figure 2 and they are to be used in the data energy term of the proposed model.

4 Model construction for cluster detection

In the following, the modelling tools presented previously are integrated in a model able to detect cluster patterns in the presented data. This is done

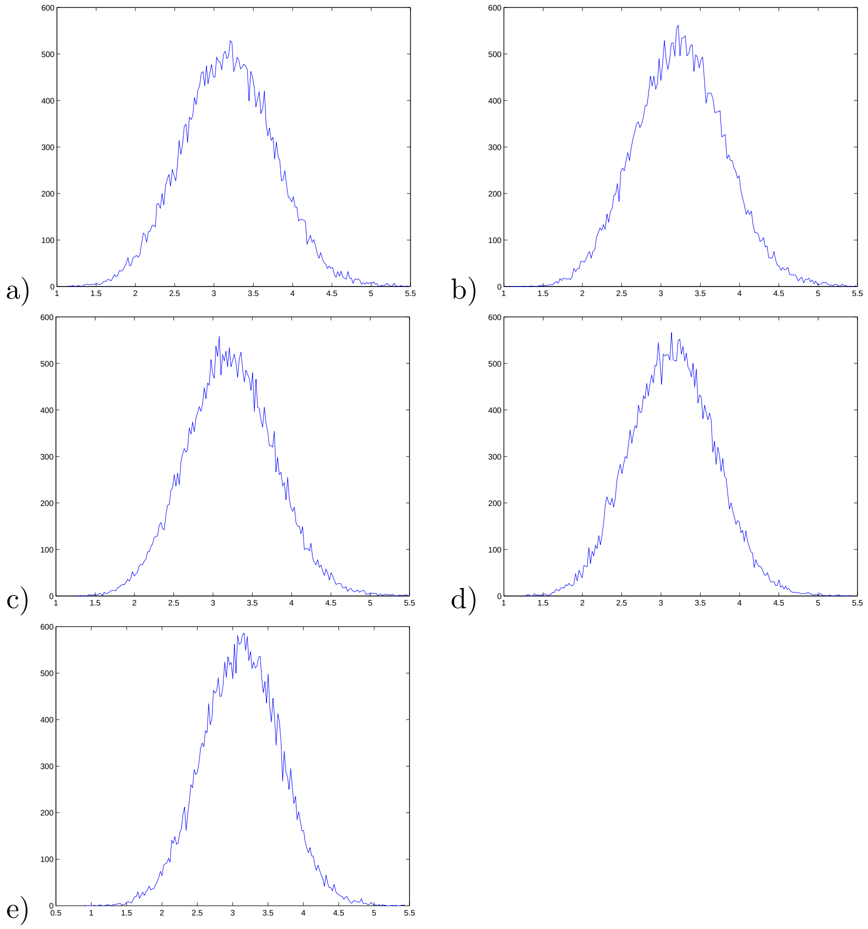

Figure 1: Histogram of the cellular score values of the year : a) 1996, b) 1997, c) 1998, d) 1999 and e) 2000.

Year	1996	1997	1998	1999	2000
d_{0i}	3.7673	3.8183	3.7812	3.6846	3.6795

Figure 2: Threshold values of for the data sets.

by specifying the Gibbs energy functions $U_{\mathbf{d}}(\mathbf{y},\theta)$ and $U_i(\mathbf{y},\theta)$ defining (1).

4.1 The data energy

The term $U_{\mathbf{d}}(\mathbf{y},\theta)$ verifies whether a random disk belongs to the cluster pattern or not. A random disk y is considered to be a part of the pattern if the number of covered farms $n_{\mathbf{d}}(y)$ is higher than a fixed value n_0. In the same time, we want to avoid the detection of clusters of "healthy" farms. Hence, $\bar{\mathbf{d}}(y)$ the estimated mean of the cellular score values covered by a disk is tested against the threshold d_0 computed previously. Under the Gaussian assumption for the covered values, a rejection region $W_{\mathbf{d}}(y,d_0)$ is computed by the Student test. The significance level of the Student test has a fixed value.

All these considerations lead us to the following expression for the energy contribution of a disk, :

$$v(y) = \mathbf{1}\{n_{\mathbf{d}}(y) > n_0\}\mathbf{1}\{\bar{\mathbf{d}}(y) \notin W_{\mathbf{d}}(y,d_0)\}[\bar{\mathbf{d}}(y) - d_0 + v_{\max}] - v_{\max}$$

where $v_{\max}$ is a positive fixed value. Its role is to penalize those disks in the object configuration that do not fulfil the enumerated conditions.

The data energy of a cluster pattern is the sum of the contributions of all the disks in the configuration:

$$U_{\mathbf{d}}(\mathbf{y},\theta) = -\sum_{i=1}^{n(\mathbf{y})} v(y).$$

It is easy to check that

$$U_{\mathbf{d}}(\mathbf{y}) - U_{\mathbf{d}}(\mathbf{y} \cup \{\zeta\}) \leq \max\{\bar{\mathbf{d}}(\zeta) - d_0, -v_{\max}\}$$

is a bounded quantity, since it is data conditioned. Therefore, the inhomogeneous Poisson process defined by $\exp[-U_{\mathbf{d}}(\mathbf{y})]$ is locally stable.

4.2 The interaction energy

Random disks tend to form clusters if they are driven by an area interaction process (6) with a parameter $\gamma_a > 1$. In the same time, the area-interaction process helps the model to better fit the data. By this, we understand that the area covered by a disk does not over exceed the region underlined by the farms positions. An example is given in Figure 3. The region induced by the same configuration of points is better fitted by the small circle, than by the big one.

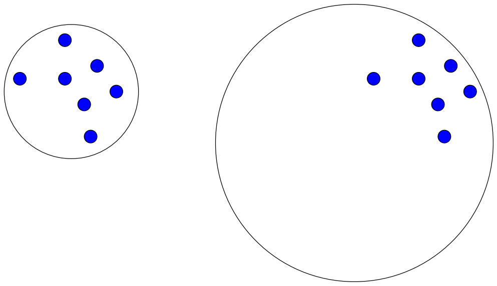

Figure 3: Two different disks around the same point configuration.

The cluster regions formed only using the area interaction process are made of disks that overlap, minimizing the occupied area. Nevertheless, the disks have to "search" for clusters through all the data locations and to not group together in an isolated region. Somehow, the disks need to be "encouraged" to cluster and to spread simultaneously. To lower the effect of the area interaction process, the pairwise interaction process given by (7) is superposed to it. This process introduces a penalty between overlapping disks.

Under these considerations, the interaction energy of a disks configuration can be written as follows:

$$U_i(\mathbf{y}, \theta) = \nu[Z(\mathbf{y})] \log \gamma_a - n_o(\mathbf{y}) \log \gamma_o. \tag{11}$$

with $\log \gamma_a$ and $\log \gamma_o$ parameters of the model. The local stability of the marked point process induced by (11) is easy to prove, since this process is the superposition of two locally stable marked point processes.

5 Experiments and results

In this section, we first present the choices for the model parameters together with some details related to the simulation dynamics. Cluster detection results on the presented data together with some statistical descriptors of the clusters are shown and interpreted.

The location space is given by the rectangle $K = [0, 317] \times [0, 318]$. The radii of the disks are continuously uniformly distributed on $M = [1, 10]$. One unit length corresponnds to 3 km distance in the real world.

The parameters for the data term are considered fixed. In this paper, their corresponding values were set as follows : $n_0 = 4$,$v_{\max} = 20$ and

5% for the significance level of the Student test. For each data set, the corresponding threshold value in Figure 2 is asigned to d_0.

The parameters vector θ contains only those parameters related to the interaction energy only. Hence, we have $\theta = (\log \gamma_a, \log \gamma_o)$ defined on the parameter space $\Theta = [0, 0.125] \times [-0.1, 0]$. The prior law $p(\theta)$ is the uniform distribution over Θ.

Sampling from $p(\mathbf{y}, \theta)$ is done in two steps. First, a parameter value is chosen with respect $p(\theta)$. Then, conditionally on θ, a new cluster pattern is sampled from $p(\mathbf{y}|\theta)$. The conditional law is simulated using a Metropolis-Hastings algorithm [7, 8]. Three types of moves are used for the construction of the transition kernel : add, delete and modify a disk to/from the current configuration. An iteration consists of two steps : first a sample from $p(\theta)$ is chosen and second, 3000 Metropolis-Hastings moves for $p(\mathbf{y}|\theta)$ are performed.

The simulated annealing algorithm samples from $p(\mathbf{y}, \theta)^{\frac{1}{T}}$, while T goes slowly to zero. Its implementation is based on the method previously described. The authors in [36] prove the convergence of the simulated annealing for simulating marked point processes, when a logarithmic cooling schedule is used. Therefore, here the temperature is lowered as follows

$$T_n = \frac{T_0}{\log(n) + 1}$$

with $T_0 = 10.0$.

5.1 Cluster detection

For each data set, 50000 iterations of the simulated annealing algorithm were carried out. The spatial domain was divided into square cells by a regular grid of size 317 × 318, hence the cells have an approximate area of 9km^2. For each cell, its corresponding visit number was calculated using (10), while running the simulated annealing. Samples were picked up every 10 iterations. The disks configurations together with its corresponding visit number map that are obtained for the data set of year 1996 are shown in Figure 4.

In both representations, we observe a massive cluster structure in the center of the spatial domain. A little bit lower around the point $(150, 150)$ in Figure 4a, a small disk can be observed. The same region in the visit number map looks like rather an important cluster region. In the same time, to the small cluster detected up the massive cluster, around the point $(175, 210)$ in Figure 4a, it corresponds in the visit number map to a less significative region. So, the visit number map can be used as a visual indicator of the “quality” of the detected clusters. The results obtained over the data sets from of years 1997 to 2000 are represented in Figure 5.

A lot of care has to be taken when using such a visit number map. Taking into account how this map was calculated, we can assign the computed visit number to each separate cell only. When computing the visit number for a greater region, we have to introduce each time the region of interest in the formula (9).

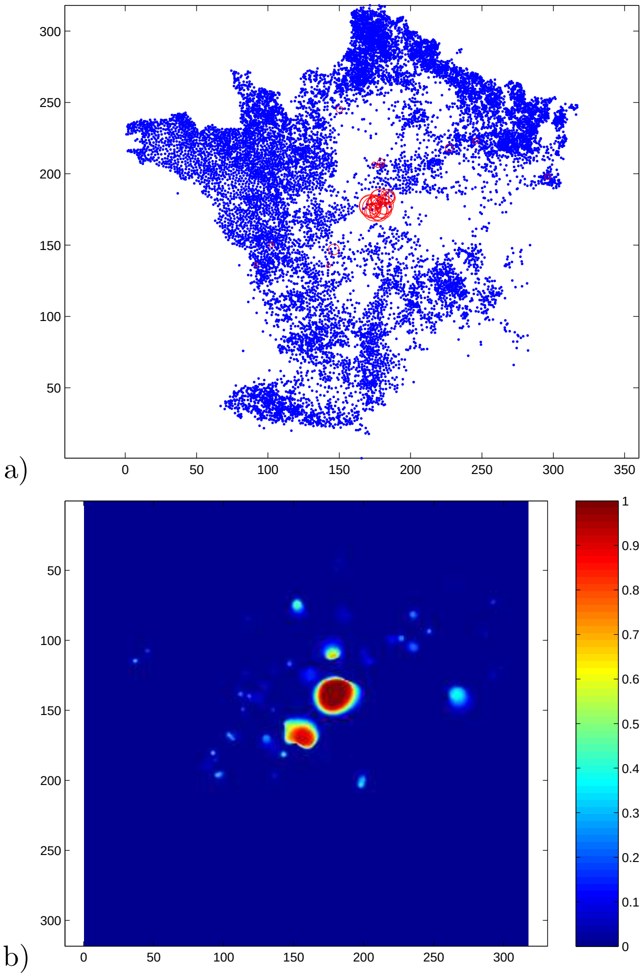

Figure 4: Cluster detection for the data set of the year 1996 : a) disks configuration obtained with the simulated annealing algorithm ; b) visit number map.

5.2 Statistical description of the cluster pattern

For the data set of year 1996 the cellular scores were permuted. 100 such fields were generated. For each field a cluster detection was performed using a simulated annealing algorithm running at fixed temperature $T = 1$. The same algorithm was run for the data set from of year 1996. 10000 iterations of the algorithm were carried out for each data set.

The visit number map obtained for the not permuted data is shown in Figure 6a. With respect to the permuted data, for each cell of the grid, the maximum of the visit number over all the 100 fields was computed. The result is plotted in Figure 6b. Significant differences can be noticed in terms of visit number value and areas of the connected components. The cluster regions observed in the not permuted data have a higher visit number values and a greater area.

During this experiment, the sufficient statistics of the model were observed every 10 iterations. For a pattern $\mathbf{y}$, these statistics are $n(\mathbf{y})$ the total number of disks, $\nu(Z(\mathbf{y}))$ the area of the pattern and $n_o(\mathbf{y})$ the number of pairs of overlapping disks. In Figure 7, the cumulative means of the sufficient statistics obtained for the year 1996 are shown. For the 100 fields of permuted data, the mean of the suffcient statistics was computed for each of them. These values are plotted in Figure 8. Over the 100 permuted data fields, the maximum value of the mean of each statistic is also indicated.

These experiments indicate good discriminant properties of the sufficient statistics of the proposed model. When clusters are detected in a data set, this is clearly indicated by the sufficient statistics of the model.

6 Conclusions and perspectives

In this paper we have proposed for the problem of cluster detection in spatial data a methodology based on marked point processes theory. Using this theoretical tool and making hypothesis on the observed phenomenon lead us to the construction of a marked point process model for the cluster pattern. Simulating this model allows statistical inference for the cluster pattern.

The proposed model is constructed by the superposition of three point processes : an inhomogeneous point process - the data term, an area interaction and a Strauss like process - the interaction term. Each of these processes plays its own role. The inhomogeneous point process detects the regions in the data where the somatic cell score is high. Using such a term only, has two drawbacks. First, an extra-detection of the clusters can occur as explained by the Figure 3. Second, the minimization of the energy func-

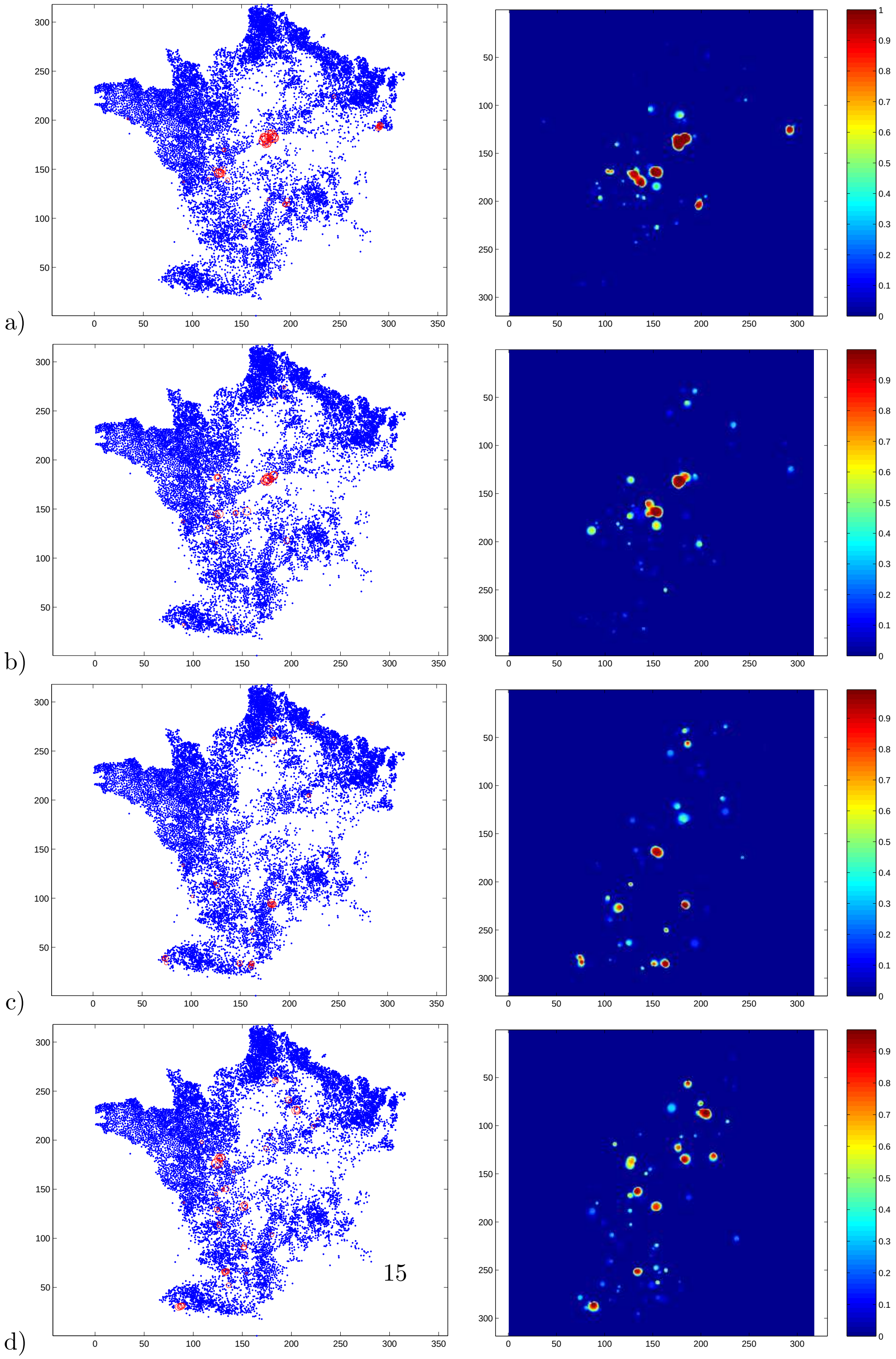

Figure 5: Cluster detection for the data sets of years : a) 1997, b) 1998, c)1999, d)2000. Left column : disks configuration obtained with the simulated annealing algorithm. Right column : visit number map.

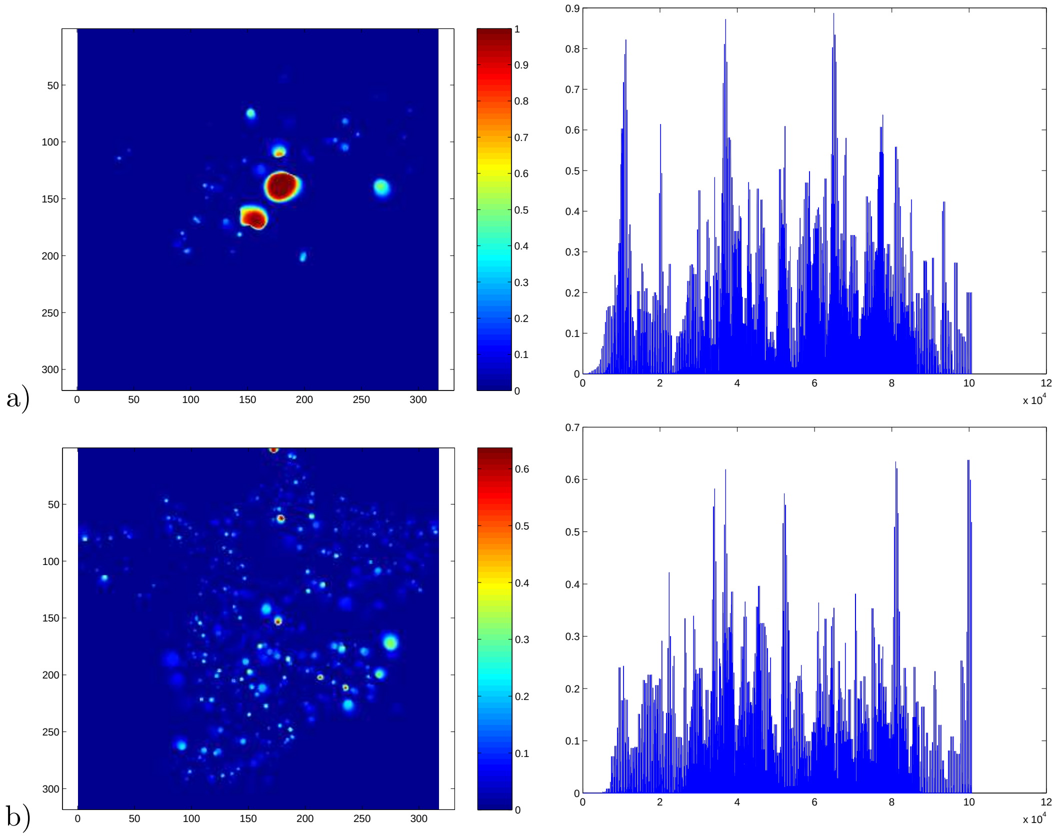

Figure 6: Visit number map comparison : a) visit map number for the year 1996 ; b) maximum visit number for the permuted data. Left : visual (matrix) presentation. Right : profile (vectorial) presentation.

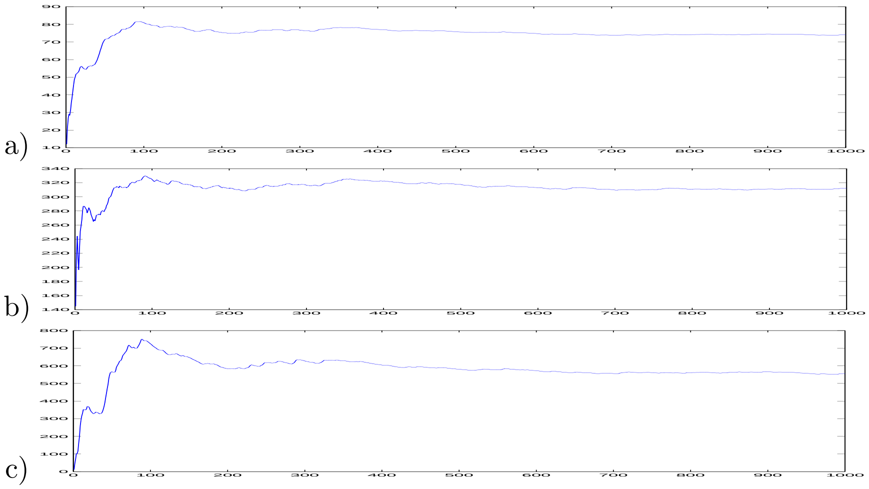

Figure 7: Sufficient statistics obtained for the year 1996 : a) $\bar{n}(\mathbf{y}) = 74.10$ cumulative mean for the total number of disks; b) $\bar{\nu}(Z(\mathbf{y})) = 312.46$ cumulative mean for the area of the pattern (mesured in cells) ; c) $\bar{n}_o = 555.08$ cumulative mean for the number of pairs of overlapping disks.

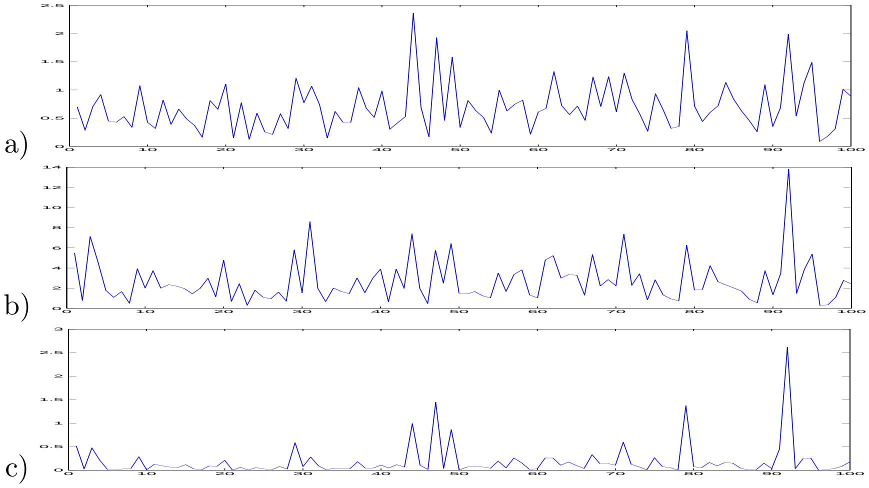

Figure 8: Mean of the sufficient statistics obtained for the 100 field of the permuted data of the year 1996 : a) mean of the total number of disks : maximum value $\bar{n}(\mathbf{y}) = 2.36$; b) mean of the area of the pattern (mesured in cells) : maximum value $\bar{\nu}(Z(\mathbf{y})) = 13.83$; c) mean of the number of pairs of overlapping disks : maximum value $\bar{n}_o = 2.62$.

tion may attract all the disks in a single region, since the disks exhibit no interaction. The area interaction component remediates the first drawback. The Strauss-like componet deals with the second one : penalizing the disks that overlap too much, forces the model to look for cluster through all the location space.

The parameters for the data term are all pre-fixed and there is no special indication for choosing $p(\theta)$. All these values were chosen studying configurations of objects that are to be favoured or not. Sampling the joint law of the pattern and the parameters helps in obtaining a natural weightening of the contribution of each component for the interaction term. This is compromise solution, since sampling from $p(\theta|\mathbf{y})$ in this context, it is far from beeing trivial [23].

The visit number for a region enabled us to build visit number maps. These are good indicators of the cluster presence and spread in the teritory. These maps are robust with respect the model parameters. They are somehow shape smoother of the cluster pattern obtained by the detection algorithm.

The sufficient statistics of the model are another indicator of the cluster pattern presence and consistency. The total number of objects and the total area occupied by the cluster pattern give indications about the size of the cluster pattern. The number of pairs of overlapping objects indicates the "strength" of these clusters.

Some perspectives may be outlined. The proposed model allows the computation of the area and the perimeter of a connected component in a cluster pattern. An interesting question is whether it is possible to compute average quantities of these characteristics. Adapting the parameter estimation method [23] to the present approach can be a way to eliminate compromise solutions for the pattern detection. Studying patterns made of objects having different shapes or introducing a time dimension for the models, are open and challenging problems.

From a more applied point of view, we intent to apply this approach to data coming from other epidemiological domains.

Aknowledgements

The authors are grateful to D. Allard, A. Kretzschmar, E. Parent, R. Senoussi and G. Thebaud for helpful discusssions.

References

[1] D. Allard and C. Fraley. Nonparametric maximum likelihood estimation of features in spatial point processes using Voronoi tessellation. *Journal of the American Statistical Association* , 92:1485–1493, 1997.

[2] A.J. Baddeley and M.N.M. van Lieshout. Stochastic geometry models in high-level vision. In K.V. Mardia and G.K. Kanji (eds.), *Statistics and Images Volume 1, Advances in Applied Statistics, a supplement to Journal of Applied Statistics*, 20:231–256, Abingdon, 1993, Carfax.

[3] A.J. Baddeley and M.N.M. van Lieshout. Area-interaction point processes. *Annals of the Institute of Statistical Mathematics*, 47:601–619, 1995.

[4] S.D. Byers and A.E. Raftery. Bayesian estimation in segmentation of spatial point processes using Voronoi tilings, in: A.B. Lawson and D.G.T. Denison (eds.), *Spatial cluster modelling*, CRC Press/Chapman and Hall, Boca Raton, 2002.

[5] X. Descombes, R.S. Stoica, L. Garcin and J. Zerubia. A RJMCMC algorithm for object processes in image processing. *Monte Carlo Methods and Applications*, 7 : 149–156, 2001.

[6] P. Elliott and J. Wakefield. Disease clusters : should they be investigated, and, if so, when and how? *Journal of Royal Statistical Society, Series A*, 164:3–12, 2001.

[7] C.J. Geyer and J. Møller. Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, 21:359–373, 1994.

[8] C.J. Geyer. Likelihood inference for spatial point processes, in: O.E. Barndorff-Nielsen, W.S. Kendall and M.N.M. van Lieshout (eds.), *Stochastic geometry, likelihood and computation*, CRC Press/Chapman and Hall, Boca Raton, 1999.

[9] P.J. Green. Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82:711-732, 1995.

[10] M.B. Hansen, J. Møller and F.Aa. Tøgersen. Bayesian contour detection in a time series of ultrasound images through dynamic deformable template models. *Biostatistics*, 3:213–228,2002.

[11] F.P. Kelly and B.D. Ripley. A not on Strauss's model for clustering. *Biometrika*, 63(2):357–360, 1976.

[12] W. S. Kendall and J. Møller. Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Advances in Applied Probability (SGSA)*, 32:844–865, 2000.

[13] M. Kulldorff. A spatial scan statistic. *Communication in Statistics : Theory and Methods*, 26(6):1481–1496, 1997.

[14] C. Lacoste, X. Descombes and J. Zerubia. A comparative study of point processes for line network extraction in remote sensing. *Research report No. 4516*, INRIA Sophia-Antipolis, 2002.

[15] S. Lakshmanan and H. Derin. Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 11 : 799–813.

[16] A.B. Lawson and M. Kulldorff. A review of cluster detection methods, in: A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J-F. Viel and R. Bertollini (eds.), *Disease mapping and risk assessment for public health*, John Wiley and Sons, 1999.

[17] A.B. Lawson and D.G.T. Denison. Spatial cluster modelling: an overview, in: A.B. Lawson and D.G.T. Denison (eds.), *Spatial cluster modelling*, Chapman and Hall/CRC, 2002.

[18] M.N.M. van Lieshout. Stochastic annealing for nearest-neighbour point processes with application to object recognition. *Advances in Applied Probability*, 26 : 281 –300, 1994.

[19] M.N.M. van Lieshout. *Markov point processes and their applications.* London/Singapore: Imperial College Press/World Scientific Publishing, 2000.

[20] M.N.M van Lieshout and A.J. Baddeley. Extrapolating and interpolating spatial patterns, in: A.B. Lawson and D.G.T. Denison (eds.), *Spatial cluster modelling*, Chapman and Hall/CRC, 2002.

[21] M.N.M. van Lieshout and R.S. Stoica. The Candy model revisited: properties and inference. *Statistica Neerlandica*, 57:1–30, 2003.

[22] M.N.M. van Lieshout and R.S. Stoica. Perfect simulation for marked point processes. *CWI Research Report* PNA-0306, 2003.

[23] J. Møller, A.N. Pettitt, K.K. Berthelsen and R.W. Reeves. An efficient Markov chain Monte Carlo method for distributions with intractable normalizing constants. *Research report R-2004-02*, Department of Mathematical Sciences, Aalborg University, 2004.

[24] J. Møller and R.P. Waagepetersen. Statistical inference for Cox processes, in: A.B. Lawson and D.G.T. Denison (eds.), *Spatial cluster modelling*, Chapman and Hall/CRC, 2002.

[25] J. Møller and R.P. Waagepetersen. *Statistical inference for spatial point processes*, Chapman and Hall/CRC, 2003.

[26] M. Ortner, X. Descombes and J. Zerubia. Building extraction from digital elevation model. *Research report No. 4517*, INRIA Sophia Antipolis, 2002.

[27] A. Pievatolo and P. J. Green. Boundary detection through dynamic polygons. *Journal of the Royal Statistical Society, Series B*, 60 : 609–626,1998.

[28] C.J. Preston. Spatial birth-and-death processes. *Bulletin of the International Statistical Institute*, 46:371–391, 1977.

[29] R.-D. Reiss. *A course on point processes.* Springer-Verlag, New-York, 1993.

[30] B.D. Ripley and F.P. Kelly. Markov point processes. *Journal of the London Mathematical Society*, 15:188–192, 1977.

[31] H. Rue and O.K. Husby. Identification of Partly Destroyed Objects using Deformable Templates. *Statistics and Computing*, 8 : 221–228, 1998.

[32] H. Rue and A.R. Syversveen. Bayesian object recognition with Baddeley's Delta loss. *Advances in Applied Probability (SGSA)*, 30 : 64–84, 1998.

[33] H. Rue and M. Hurn. Bayesian object identification. *Biometrika*, 3 : 649–660, 1999.

[34] D. Ruelle. *Statistical mechanics.* Wiley, New York, 1969.

[35] R.S. Stoica, X. Descombes and J. Zerubia. A Gibbs point process for road extraction in remotely sensed images. *International Journal of Computer Vision*, 57(2):121–136, 2004.

[36] R.S. Stoica, P. Gregori and J. Mateu. Simulated annealing and object point processes : tools for analysis of spatial patterns. To appear in *Stochastic Processes and their Applications.*

[37] D. Stoyan, W.S. Kendall and J. Mecke. *Stochastic geometry and its applications*, John Wiley and Sons, 1995.

[38] D.J. Strauss. A model for clustering. *Biometrika*, 62(2):467–475,1975.

[39] T. Tango. Comparison of general tests for spatial clustering, in: A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J-F. Viel and R. Bertollini, *Disease mapping and risk assessment for public health*, John Wiley and Sons, 1999.

[40] L. Younes Estimation and annealing for Gibbsian fields. *Annales de l'Instiut Henri Poincaré*, 24(2):269–294, 1988.

[41] L. Younes Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields*, 82:625–645, 1989.