Perfect sampling for posterior landmark distributions with an application to the detection of disease clusters

Marc A. Loizeaux* and Ian W. McKeague†
Department of Statistics
Florida State University
Tallahassee, FL 32306-4330

May 28, 2001

Abstract

We study perfect sampling for the posterior distribution in a class of spatial point process models introduced by Baddeley and van Lieshout (1993). For Neyman–Scott cluster models, perfect sampling from the posterior is shown to be computationally feasible via a coupling-from-the-past type algorithm of Kendall and Møller. An application to data on leukemia incidence in upstate New York is presented.

1 Introduction

Bayesian cluster models based on spatial point processes were originally introduced by Baddeley and van Lieshout (1993), primarily for applications in computer vision. Disease clustering applications have also played a prominent role in the development of these models, as surveyed by Lawson and Clarke (1999). An important special case is the Neyman–Scott process in which the observations arise from a superposition of inhomogeneous Poisson processes associated with underlying landmarks (Neyman and Scott, 1972); van Lieshout (1995) focused on this case.

Markov chain Monte Carlo (MCMC) techniques are indispensable for the application of point process models in statistics, see, e.g., the survey of Møller (1999). Following

*Partially supported by NSA Grant MDA904-99-1-0070.
†Partially supported by NSF Grant 9971784.
[1]Equipment support was provided under ARO Grant DAAG55-98-1-0102 and NSF Grant 9871196.
[2]*AMS 1991 subject classifications.* Primary: 62M30; secondary: 62H30.
[3]*Key words and phrases: disease mapping, landmark estimation, Markov chain Monte Carlo, spatial point processes.*
[4]Running head: Perfect sampling in spatial statistics

the seminal work of Propp and Wilson (1996), Kendall and Møller (2000) developed a version of perfect simulation for locally stable point processes; see the article of Møller (2001) in this volume. This raises the possibility of constructing perfect samplers for the posterior distribution in Bayesian cluster models. Perfect samplers deliver an exact draw from the target distribution; this is a distinct advantage over traditional MCMC schemes which are often plagued by convergence problems. For some recent applications of perfect simulation in statistics, see Green and Murdoch (1999), Møller and Nicholls (1999) and Casella et al. (1999).

In this article we show that the posterior in the Baddeley–van Lieshout class of Bayesian cluster models is locally stable, provided the prior is locally stable and the likelihood satisfies some mild conditions. This has two important consequences: the posterior density is *proper* (has unit total mass), and the Kendall–Møller algorithm is potentially applicable. However, the Kendall–Møller algorithm is known to be computationally feasible only under a monotonicity condition: the Papangelou conditional intensity needs to be attractive (favoring clustered patterns), repulsive (discouraging clustered patterns), or a product of such terms. We show that perfect sampling is feasible for the Neyman–Scott process when the prior satisfies this monotonicity condition.

We present an application to data on leukemia incidence in an eight county area of uptstate New York during the years 1978–82. The study area includes 11 inactive hazardous waste sites. We assess the possibility of an increased leukemia incidence rate in the proximity of these sites. There is an extensive literature on the analysis of these data, recent contributions being Ghosh et al. (1999), who applied a hierarchical Bayes generalized linear model approach, and Ahrens et al. (1999), who adjusted for covariate effects using a log-linear model. Our results suggest that there is an elevated leukemia incidence rate in the neighborhood of one of the sites.

The paper is organized as follows. In Section 2 we develop the main result of the paper showing that the posterior is locally stable, and examine the Neyman–Scott model in detail. Section 3 contains the application to disease clustering. Concluding remarks are given in Section 4.

2 Bayesian cluster models

2.1 Preliminaries

The basic framework comes from Carter and Prenter (1972), see also Møller (1999). Let W be a compact subset of the plane representing the study region. A realization of a point process in W is a finite set of points $\mathbf{x} = \{x_1, x_2, \ldots, x_{n(\mathbf{x})}\} \subset W$, where $n(\mathbf{x})$ is the number of points in $\mathbf{x}$. If $n(\mathbf{x}) = 0$, write $\mathbf{x} = \emptyset$ for the empty configuration. Let Ω denote the exponential space of all such finite point configurations in W, and furnish it with the σ-field $\mathcal{F}$ generated by sets of the form $\{\mathbf{x} : n(\mathbf{x} \cap B) = k\}$, where $B \in \mathcal{B}$, the Borel σ-field on W, and $k = 0, 1, 2, \ldots$.

A standard way of constructing an Ω-valued point process X is by specifying an unnormalized density f with respect to the distribution π of the unit rate Poisson process on W. The unnormalized density f (or corresponding process X) is said to be locally stable if there is a constant $K > 0$ such that $f(\mathbf{x} \cup \{\xi\}) \leq K f(\mathbf{x})$ for all $\mathbf{x} \in \Omega$, $\xi \in W \backslash \mathbf{x}$. Local stability implies that the Papangelou conditional intensity

$$q(\mathbf{x}, \xi) = \frac{f(\mathbf{x} \cup \{\xi\})}{f(\mathbf{x})}, \ \mathbf{x} \in \Omega, \ \xi \in W \backslash \mathbf{x},$$

(with $0/0 = 0$) is bounded.

Most point processes that have been suggested for modeling spatial point patterns are locally stable, including the Strauss (1975) process and the area-interaction process of Baddeley and van Lieshout (1995). The Strauss process, used later in this article, has unnormalized density $f(\mathbf{x}) = \beta^{n(\mathbf{x})} \gamma^{t(\mathbf{x})}$, where $\beta > 0$, $0 < \gamma \leq 1$ and $t(\mathbf{x})$ is the number of unordered pairs of points in $\mathbf{x}$ which are within a specified distance r of each other. The Strauss process only models repulsive pairwise interaction.

2.2 Posterior distribution

The observed point configuration which arises from the landmarks $\mathbf{x}$ will be denoted $\mathbf{y} = \{y_1, y_2, \ldots, y_{n(\mathbf{y})}\} \subset W$, and assumed to be non-empty. The prior and observation models are specified by point processes on W. The prior distribution of landmarks corresponds to a point process X having density $p_X(\mathbf{x})$ with respect to π.

The likelihood is defined in terms of an unnormalized density $f(\cdot|\mathbf{x})$. Thus, for a given set of landmarks $\mathbf{x}$, the density of the observed point process Y with respect to π is

$$p_{Y|X=\mathbf{x}}(\mathbf{y}) = \alpha_Y(\mathbf{x}) f(\mathbf{y}|\mathbf{x}),$$

where

$$\alpha_Y(\mathbf{x}) = \left(\int_\Omega f(\mathbf{v}|\mathbf{x}) \, \pi(d\mathbf{v}) \right)^{-1}$$

is the normalizing constant. We assume that $f(\mathbf{y}|\mathbf{x})$ is jointly measurable in $\mathbf{x}$ and $\mathbf{y}$.

From Bayes formula, the posterior density of X with respect to π is

$$p_{X|Y=\mathbf{y}}(\mathbf{x}) \propto \alpha_Y(\mathbf{x}) f(\mathbf{y}|\mathbf{x}) p_X(\mathbf{x}). \tag{2.1}$$

The following theorem provides sufficient conditions for the posterior to be locally stable. We assume local stability of the prior $p_X(\cdot)$ and of the likelihood $f(\mathbf{y}|\cdot)$ (for each fixed $\mathbf{y}$). In addition, $f(\mathbf{y}|\cdot)$ is assumed to satisfy the following *local growth condition*: there exists a constant $L > 0$ such that

$$f(\mathbf{y}|\mathbf{x} \cup \{\xi\}) \geq L f(\mathbf{y}|\mathbf{x}) \tag{2.2}$$

for all $\mathbf{x}$, $\mathbf{y} \in \Omega$, $\xi \in W \backslash \mathbf{x}$. The term 'local growth condition' is used here because we view it as being dual to local stability.

Theorem 2.1. *Suppose $p_X(\cdot)$ and $f(\mathbf{y}|\cdot)$ (for each $\mathbf{y}$) are locally stable, and $f(\mathbf{y}|\cdot)$ satisfies the local growth condition (2.2). Then the posterior (2.1) is locally stable.*

Proof. It suffices to show that $\alpha_Y(\cdot)$ is locally stable, because $p_X(\cdot)$ and $f(\mathbf{y}|\cdot)$ are assumed to be locally stable, and locally stability is preserved under products. Given $\xi \in W\backslash\mathbf{x}$,

$$\begin{aligned}\alpha_Y(\mathbf{x}\cup\xi)^{-1} &= \int_\Omega f(\mathbf{v}|\mathbf{x}\cup\xi)\,\pi(d\mathbf{v}) \\ &\geq L\int_\Omega f(\mathbf{v}|\mathbf{x})\,\pi(d\mathbf{v}) \\ &= L\alpha_Y(\mathbf{x})^{-1},\end{aligned}$$

completing the proof.

The Kendall–Møller algorithm uses the method of dominated coupling-from-the-past to obtain perfect samples from a locally stable point process as the equilibrium distribution of a spatial birth-and-death process. The algorithm is computationally feasible if the Papangelou conditional intensity $q(\mathbf{x},\xi)$ is either attractive or repulsive, or a product of such terms. In the *attractive* case, $q(\mathbf{x},\xi) \leq q(\mathbf{x}',\xi)$ whenever $\xi \notin \mathbf{x}'$ and $\mathbf{x} \subset \mathbf{x}'$; in the *repulsive* case $q(\mathbf{x},\xi) \geq q(\mathbf{x}',\xi)$ whenever $\xi \notin \mathbf{x}'$ and $\mathbf{x} \subset \mathbf{x}'$.

2.3 Neyman–Scott model

In this section we focus on the Neyman–Scott model in which the observation process Y is the superposition of $n(\mathbf{x})$ independent inhomogeneous Poisson processes Z_{x_i} and a background Poisson noise process of intensity $\epsilon > 0$. The intensity $h(\cdot|x_i)$ of Z_{x_i} is specified parametrically, and the prior $p_X(\mathbf{x})$ is assumed to be locally stable. Here and in the application in the next section we assume the Thomas intensity model

$$h(t|x) = \frac{\kappa}{2\pi\sigma^2}e^{-||t-x||^2/2\sigma^2}, \tag{2.3}$$

where $\kappa, \sigma > 0$. For convenience let $\kappa^* = \frac{\kappa}{2\pi\sigma^2}$. In this case $f(\mathbf{y}|\mathbf{x}) = \prod_{j=1}^{n(\mathbf{y})}\mu(y_j|\mathbf{x})$, where

$$\mu(t|\mathbf{x}) = \epsilon + \sum_{i=1}^{n(\mathbf{x})} h(t|x_i)$$

is the conditional intensity at t of Y given $\mathbf{x}$. We now check the relevant conditions of Theorem 2.1.

To show that $f(\mathbf{y}|\cdot)$ is locally stable, note that for $\xi \in W\backslash\mathbf{x}$

$$f(\mathbf{y}|\mathbf{x}\cup\{\xi\}) = \prod_{j=1}^{n(\mathbf{y})}\left(\epsilon + \sum_{i=1}^{n(\mathbf{x})} h(y_j|x_i) + h(y_j|\xi)\right)$$

$$
\begin{aligned}
&\le \prod_{j=1}^{n(\mathbf{y})} \left(\epsilon \left(1 + \frac{h(y_j|\xi)}{\epsilon} \right) + \sum_{i=1}^{n(\mathbf{x})} h(y_j|x_i) \right) \\
&\le K_{\text{lik}}(\mathbf{y}) f(\mathbf{y}|\mathbf{x}),
\end{aligned}
$$

where

$$K_{\text{lik}}(\mathbf{y}) = \sup_{\xi \in W} \prod_{j=1}^{n(\mathbf{y})} \left(1 + \frac{h(y_j|\xi)}{\epsilon} \right) < \infty.$$

To check the local growth condition, note that for $\xi \in W \backslash \mathbf{x}$

$$f(\mathbf{y}|\mathbf{x} \cup \{\xi\}) = \prod_{j=1}^{n(\mathbf{y})} (\mu(y_j|\mathbf{x}) + h(y_j|\xi)) \ge f(\mathbf{y}|\mathbf{x}),$$

uniformly in $\mathbf{y}$, and we can use $L = 1$. Thus the conditions of Theorem 2.1 are satisfied, so the posterior is locally stable.

The Papangelou conditional intensity corresponding to $f(\mathbf{y}|\cdot)$ is

$$\frac{f(\mathbf{y}|\mathbf{x} \cup \{\xi\})}{f(\mathbf{y}|\mathbf{x})} = \prod_{j=1}^{n(\mathbf{y})} \left(1 + \frac{h(y_j|\xi)}{\epsilon + \sum_{i=1}^{n(\mathbf{x})} h(y_j|x_i)} \right),$$

for $\xi \in W \backslash \mathbf{x}$, which is clearly decreasing in $\mathbf{x}$, thus repulsive.

Noting that

$$\alpha_Y(\mathbf{x}) = \exp \left\{ \int_W (1 - \mu(t|\mathbf{x}))\, dt \right\},$$

we find that the Papangelou conditional intensity corresponding to $\alpha_Y(\cdot)$ is

$$\frac{\alpha_Y(\mathbf{x} \cup \{\xi\})}{\alpha_Y(\mathbf{x})} = \exp \left\{ - \int_W h(t|\xi)\, dt \right\},$$

for $\xi \in W \backslash \mathbf{x}$, which does not depend on $\mathbf{x}$.

We conclude that the posterior is of a feasible form for implementing the Kendall–Møller algorithm if the Papangelou conditional intensity corresponding to the prior is a product of repulsive or attractive components.

3 Application

In this Section we present an application to data on leukemia incidence in an eight county area of uptstate New York. There is an extensive literature on the analysis of these data, see, e.g., Waller et al. (1992, 1994), Ghosh et al. (1999) and Ahrens et al. (1999).

The study area is comprised of 790 census tracts and leukemia incidence was recorded by the New York Department of Health for each census tract during the years 1978–82,

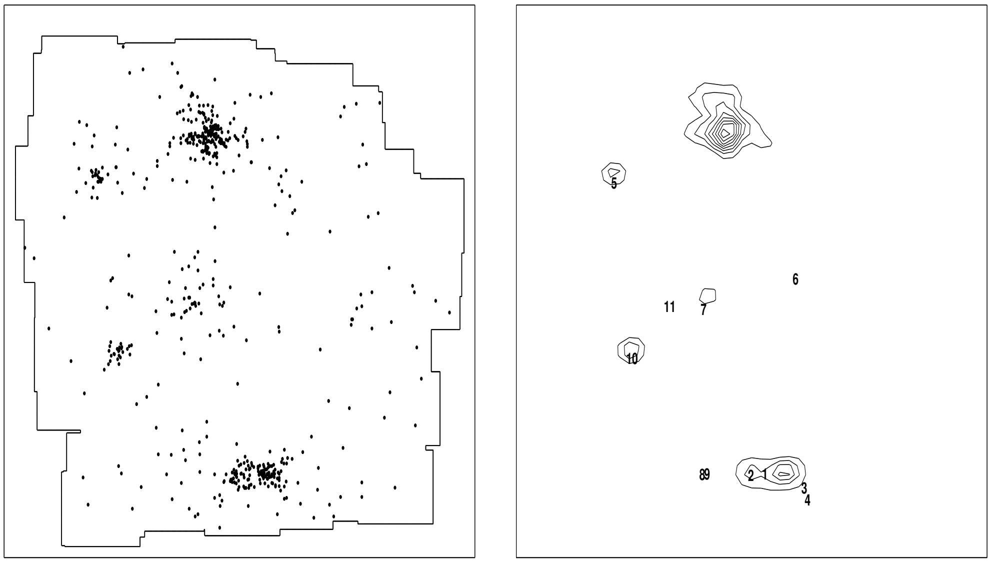

Figure 1: Left: locations of 552 leukemia cases in upstate New York, along with an approximate outline of the eight county study region. The rectangular region is 1×1.2 square units. Right: contour plot of the population density $\lambda(t)$, and the locations of the 11 hazardous waste sites.

see Waller et al. (1992). The study area includes 11 inactive hazardous waste sites. The goal is to assess the possibility of an increased leukemia incidence rate in the proximity of these sites.

The locations of the centroids of the census tracts are available, but precise locations of the leukemia cases are not. Our methods require the precise locations, so we randomly dispersed the cases throughout their corresponding census tracts; if there was exactly one case in a tract, we placed it at the centroid, see the left panel of Figure ??. (Sensitivity analysis indicates that this approximation makes no difference to our conclusions.) In some instances a case could not be associated with a unique census tract, resulting in fractional counts. Our approach does not accomodate this type of data, so we follow Ghosh et al. (1999) and group the 790 census tracts into 281 blocks in order to identify most of the cases with a specific block. Less than 10% of all cases could not be identified with a specific block, and such cases are excluded from our analysis.

Our analysis is based on the Neyman–Scott model with the leukemia intensity rate specified by

$$\mu(t|\mathbf{x}) = \lambda(t)\Big(\epsilon + \sum_{i=1}^{n(\mathbf{x})} h(t|x_i)\Big),$$

where $\lambda(t)$ adjusts for population density, $\epsilon > 0$ and $h(t|x)$ is the Thomas intensity (2.3). Our earlier treatment of the Neyman–Scott model extends without change to this form of

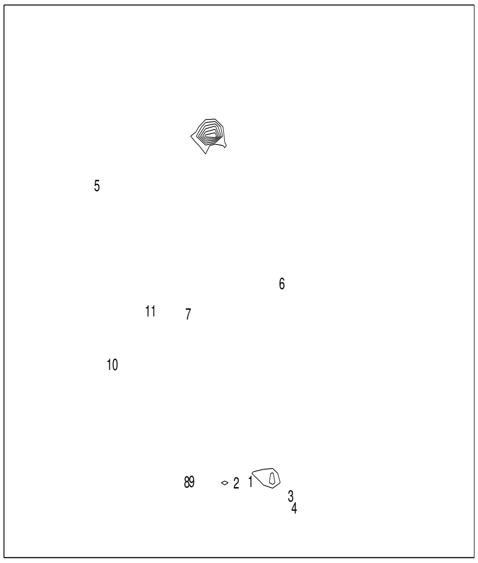

Figure 2: Posterior intensity map for the leukemia data based on the Neyman–Scott model with $\epsilon = 5.2 \times 10^{-4}$, $\sigma = 0.01$, $\kappa^* = 0.23\epsilon$, and a Strauss prior with interaction radius $r = 0.1$, $\beta_X = 0.5$ and $\gamma_X = 0.1$. Locations of the 11 hazardous waste sites are included.

the model because $\lambda(t)$ does not depend on $\mathbf{x}$. We use a Strauss prior for the landmarks $\mathbf{x}$. For $\lambda(t)$ we used a smoothed version of the population density based on the 1980 U.S. census, see the right panel of Figure ??; this plot also gives the locations of the 11 inactive hazardous waste sites suspected of causing elevated leukemia incidence rates.

Figure ?? gives the posterior intensity for the landmarks based on the data shown in the left panel of Figure ??; we used 1000 samples drawn using the Kendall–Møller algorithm. Note that one of the waste sites (site 1) is located close to an area of high posterior intensity.

To assess the significance of an elevated leukemia rate in the neighborhood of a given site, we compare the 'observed' with the 'expected' posterior landmark distribution. The relevant null hypothesis here is that the leukemia cases form an inhomogeneous Poisson process with intensity $\rho\lambda(t)$, where ρ is the average leukemia rate throughout the study region. To sample from the null distribution, we generated an artificial data set using independent Poisson counts for each census tract, then analyzed the artificial data the same way as the original data.

In Figure ?? we compare the observed and expected posterior probabilities of at least one landmark within a given distance (0–7.2 kms) of each waste site. The 20 dotted lines correspond to samples from the null distribution, and the 5 solid lines correspond to the data (with the leukemia cases randomly dispersed throughout their corresponding census tracts). The plots provide evidence of elevated leukemia rates in the neighborhood of site 1.

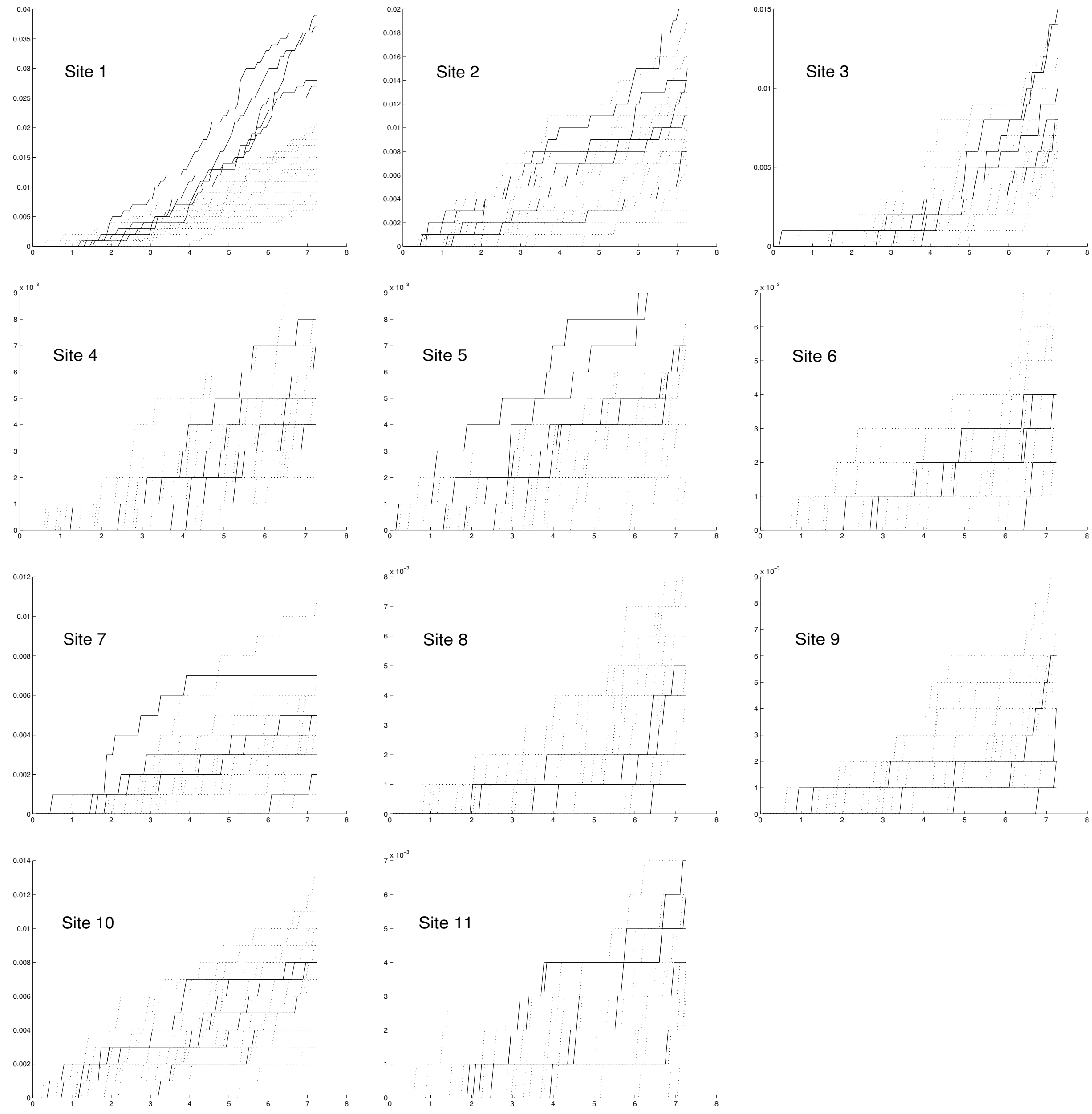

Figure 3: Posterior observed (solid lines) and expected (dotted lines) probabilities of at least one landmark within a given distance (in kms) of each waste site.

4 Conclusion

In this article we have developed perfect sampling for the posterior distribution in Bayesian cluster models for spatial point processes. We have isolated conditions under which perfect sampling using the Kendall–Møller algorithm is applicable. The algorithm is shown to be feasible under mild conditions on the prior and the likelihood, and, in particular, for the useful special case of the Neyman–Scott model when the prior is repulsive.

We are currently working on a more detailed study of this topic in which we examine an extended formulation of the Baddeley–van Lieshout cluster model and provide other examples in which perfect sampling from the posterior is feasible.

Acknowledgements. We thank George Casella, John Staudenmayer and Lance Waller for help with the leukemia incidence data. We also thank Jesper Møller for several useful comments.

References

Ahrens, C., Altman, N., Casella, G., Eaton, M., Hwang, J. T. G., Staudenmayer, J. and Stefanscu, C. (1999). Leukemia clusters and TCE waste sites in upstate New York: How adding covariates changes the story. Preprint.

Baddeley, A. J. and van Lieshout, M. N. M. (1993). Stochastic geometry models in high-level vision. In K. V. Mardia and G. K. Kanji, editors, *Advances in Applied Statistics*, Statistics and Images: 1, 231–256, Carfax Publishing.

Baddeley, A. J. and van Lieshout, M. N. M. (1995). Area-interaction point processes. *Ann. Inst. Statist. Math.* 47, 601–619.

Casella, G., Mengersen, K. L., Robert, C. P. and Titterington, D. M. (1999). Perfect slice samplers for mixtures of distributions.
`ftp://ftp.ensae.fr/pub/labo_stat/CPRobert/perfect.ps.gz`.

Geyer, C. J. (1999). Likelihood inference for spatial point processes. In O. E. Barndorff–Nielsen, W. S. Kendall, and M. N. M. van Lieshout, editors, *Stochastic Geometry: Likelihood and Computation*, pp. 79–140. Chapman and Hall.

Geyer, C. J. and Møller, J. (1994). Simulation and likelihood inference for spatial point processes. *Scand. J. Statist.* 21, 359–373.

Ghosh, M., Natarajan, K., Waller, L. and Kim, D. (1999). Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping. *J. Statist. Planning Inference* 75, 305–318.

Green, P. J. and Murdoch, D. J. (1999). Exact sampling for Bayesian inference: towards general purpose algorithms. In J. M. Bernardo et al., editors, *Bayesian Statistics 6*, Oxford University Press. Presented as an invited paper at the 6th Valencia International Meeting on Bayesian Statistics, Alcossebre, Spain, June 1998.

Häggström, O., van Lieshout, M. N. M. and Møller, J. (1999). Characterization results and Markov chain Monte Carlo algorithms including exact simulation for some spatial point processes. *Bernoulli* 5, 641–659.

Kendall, W. S. and Møller, J. (2000). Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Adv. Appl. Probab.* 32, 844–865.

Lawson, A. B. and Clark, A. (1999). Markov chain Monte Carlo for putative sources of hazard and general clustering. In A. B. Lawson, D. Böhning, A. Biggeri, E. Lesaffre, J.-F. Viel, editors, *Disease Mapping and Risk Assessment for Public Health.* Wiley.

Møller, J. (1999). Markov chain Monte Carlo and spatial point processes. In O. E. Barndorff–Nielsen, W. S. Kendall, and M. N. M. van Lieshout, editors, *Stochastic Geometry: Likelihood and Computation*, pp. 141–172. Chapman and Hall.

Møller, J. (2001). A review on perfect simulation in stochastic geometry. In this volume.

Møller, J. and Nicholls, G. K. (1999). Perfect simulation for sample-based inference. `http://www.math.auckland.ac.nz/~nicholls /linkfiles/papers/perfect.sim.temp1.ps.gz`.

Neyman, J. and Scott, E. L. (1972). Processes of clustering and applications. In P. A. W. Lewis, editor, *Stochastic Point Processes*, Wiley, New York, pp. 641–681.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* 9, 223–252.

Strauss, D.J. (1975). A model for clustering. *Biometrika* 62, 467–475.

Van Lieshout, M. N. M. (1995). *Stochastic Geometry Models in Image Analysis and Spatial Statistics.* CWI Tract 108. Stichting Mathematisch Centrum, Amsterdam.

Waller, L., Turnbull, B., Clark, L. and Nasca, P. (1992). Chronic disease surveillance and testing of clustering of disease and exposure: application to leukemia incidence and TCE-contaminated dump sites in upstate New York. *Environmetrics* 3, 281–300.

Waller, L., Turnbull, B., Clark, L. and Nasca, P. (1994). Spatial pattern analysis to detect rare disease clusters. In N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, J. Greenhouse, editors, pp. 3–23, *Case Studies in Biometry.* Wiley.