MARKOV CHAIN MONTE CARLO METHODS FOR CLUSTERING OF IMAGE FEATURES

M N M van Lieshout

University of Warwick
United Kingdom

A J Baddeley

University of WA
Australia

1 Introduction

The identification of centres of clustering is of interest in many areas of applications, for instance edge detector output has to be grouped into meaningful curves. In this paper we argue that stochastic geometry models are helpful both in providing models for clustering and as a prior distribution to combat overestimation of the number of clusters and to improve robustness.

The general idea in connection with object recognition was proposed by Baddeley and Van Lieshout [1]. See also Van Lieshout [13]. Independently, in an epidemiological context, a different Gibbs sampler technique for detection of cluster centres in a Cox process was developed by Lawson [11]. Earlier attempts include [3, 5, 14, 15]. For extensions and more examples see Lawson et al. [12].

2 Cluster processes

The data consist of a set of 'events' $\mathbf{y} = \{y_1, \ldots, y_m\} \subseteq T$, where $T \subseteq \mathbb{R}^2$ (say) is the window of observation and it is required to determine the locations of an unspecified number of cluster centres $\mathbf{x} = \{x_1, \ldots, x_n\} \subseteq U$, $n \geq 0$. In the simplest case, $U = T$ or a somewhat larger set to account for edge effects, but note that U may be a different space as in Section 5 below. In probability theory, a set of points such as $\mathbf{x}$ or $\mathbf{y}$ are realisations of a random *point process*. An arbitrary dummy centre x_0 is introduced both for technical reasons, and to allow for events not belonging to any cluster.

The mathematical model we adopt is an *independent cluster model* (see eg. [19]) where, conditional on $\mathbf{x} = \{x_0, \ldots, x_n\}$, the observations result from the superposition of independent, finite point processes

$$\mathbf{y} = \bigcup_{i=0}^{n} \mathbf{Z}_{x_i}.$$

For simplicity, we confine ourselves here to the case where each set of 'daughters' $\mathbf{Z}_u$, $u \in U$, is a realisation of an *inhomogenous Poisson process* on T with *intensity function* $h(\cdot \mid u) : T \to [0, \infty)$. By the superposition property, the combined offspring form a Poisson process with intensity function

$$\lambda(\cdot \mid \mathbf{x}) = \sum_{i=0}^{n} h(\cdot - x_i).$$

In other words, given a fixed, finite reference measure μ on T, representing deterministic but spatially varying factors, the number of events $n(\mathbf{y})$ in $\mathbf{y}$ is Poisson distributed with mean $\int_T \lambda(t \mid \mathbf{x})\, d\mu(t)$ and conditionally on $n(\mathbf{y}) = m$ the events have joint density

$$f_m(\mathbf{y} \mid \mathbf{x}) = \frac{\prod_{j=1}^{m} \lambda(y_j \mid \mathbf{x})}{\left(\int_T \lambda(t \mid \mathbf{x})\, d\mu(t)\right)^m}$$

with respect to μ^m.

3 Detection of cluster centres

We interpret the identification of centres of clustering as a statistical estimation problem. The unknown configuration of centres is regarded as a model parameter, and upon observation of $\mathbf{y}$, has likelihood $l(\mathbf{x} \mid \mathbf{y}) = f(\mathbf{y} \mid \mathbf{x})$. A maximum likelihood estimator solves

$$\hat{\mathbf{x}} = \operatorname{argmax}_{\mathbf{x}} f(\mathbf{y} \mid \mathbf{x}),$$

but note that there is no guarantee these equations have a solution, nor that a solution is unique. Indeed, a maximum likelihood estimator of $\mathbf{x}$ may run into difficulties similar to those encountered in the context of object recognition [1]. If h is smooth and almost flat near its maximum, and the data pattern is 'dense', the maximum likelihood estimate tends to contain multiple responses to each true cluster.

To overcome these problems we propose a prior distribution to penalise scenes that contain too many 'similar' centres. A suitable choice is a Markov spatial process [2, 17]. For brevity, we consider only pairwise interactions models, defined by their density

$$p(\mathbf{x}) = \alpha \beta^{n(\mathbf{x})} \prod_{x_i \sim x_j} g(x_i, x_j) \qquad (1)$$

Image Processing And Its Applications, 4–6 July 1995
Conference Publication No. 410, © IEE 1995.

(with respect to a unit rate Poisson model). Here, α is the normalising constant, $\beta > 0$ a model parameter. The product ranges over all pairs of similar objects $x_i \sim x_j$ and $g(\cdot,\cdot) : U \times U \to [0,\infty)$ is the interaction function. The case $g \equiv 1$ is a Poisson process with rate β; for $g(\cdot,\cdot) < 1$, configurations with many similar centres are unlikely.

An important property of (1), and of Markov models in general, is that the likelihood ratio

$$\frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} = \beta \prod_{x_i \sim u} g(x_i, u)$$

depends only on those $x_i \in \mathbf{x}$ that are similar to u, signifying that all interaction is 'local'. In the statistical physics interpretation, $-\log p(\mathbf{x} \cup \{u\}) + \log p(\mathbf{x})$ is the energy required to add a new point u to an existing configuration $\mathbf{x}$; in probabilistic terms $\lambda(u;\mathbf{x}) = p(\mathbf{x} \cup \{u\})/p(\mathbf{x})$ is the Papangelou conditional intensity at u given the rest of the pattern $\mathbf{x}$ on $U \setminus \{u\}$, see [6]. Roughly speaking, $\lambda(u;\mathbf{x})du$ is the conditional probability of a point in the infinitesimal region du centred at u given the configuration agrees with $\mathbf{x}$ outside this region.

After collection of the data, the posterior density of $\mathbf{x}$ is

$$p(\mathbf{x} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x})$$

by an application of Bayes' formula. However, due to the normalising constant, this distribution tends to be rather intractable and we have to resort to Markov chain Monte Carlo methods [4] for optimisation and sampling. The basic idea is to build a Markov chain (or in continuous time, a Markov process) with the target distribution $p(\mathbf{x} \mid \mathbf{y})$ as equilibrium and transitions that are 'easy' to perform.

The classical approach in a point process context is via spatial birth-and-death processes [2, 16, 18] but other methods such as Metropolis-Hastings [9], simulated tempering or annealing can be used as well. Furthermore, functionals of the posterior distribution such as the distribution of the number of clusters, the probability that there is no cluster in a particular region, and the first-order intensity of cluster locations can be estimated.

All these techniques iteratively update the cluster centre configuration by addition and deletion, with transition criteria based on the posterior likelihood ratios. For instance, a spatial birth-and-death process is a continuous time Markov process for which the only transitions are the birth of a new object (instantaneous transition from $\mathbf{x}$ to $\mathbf{x} \cup \{u\}$) or the death of an existing one (transition from $\mathbf{x}$ to $\mathbf{x} \setminus \{x_i\}$). Transitions are governed by death rate $D(\cdot,\cdot)$ and birth rate $B(\cdot,\cdot)$ as follows.

- the probability of a death $\mathbf{x} \to \mathbf{x} \setminus \{x_i\}$ during a time interval $(t, t+h)$, $h \to 0$, is $D(\mathbf{x} \setminus \{x_i\}, x_i)h + o(h)$;
- the probability of a birth $\mathbf{x} \to \mathbf{x} \cup \{u\}$ during time $(t, t+h)$, where u lies in a given subset $F \subseteq U$, is $B(\mathbf{x}, F)h + o(h)$;
- the probability of more than one transition during $(t, t+h)$ is $o(h)$.

We will assume that $B(\mathbf{x},\cdot)$ has a density $b(\mathbf{x},\cdot)$ with respect to some finite reference measure ν on U, so that intuitively $b(\mathbf{x},u)$ is the transition rate for a birth $\mathbf{x} \to \mathbf{x} \cup \{u\}$.

Good choices for the transition rates are

$$b(\mathbf{x},u) = \frac{p(\mathbf{x} \cup \{u\} \mid \mathbf{y})}{p(\mathbf{x} \mid \mathbf{y})}$$

and, writing $n(\mathbf{x})$ for the number of non-dummy centres in $\mathbf{x}$, $D(\mathbf{x} \setminus \{x_i\}, x_i) = 1/n(\mathbf{x})$, which results in detailed balance between births and deaths:

$$b(\mathbf{x},u)p(\mathbf{x} \mid \mathbf{y}) = D(\mathbf{x},u)p(\mathbf{x} \cup \{u\} \mid \mathbf{y}).$$

To avoid explosion, i.e. an infinite number of transitions occurring in finite time, the rates have to satisfy certain assumptions [16]. Typically in this context we need an inhibitory prior $g(\cdot,\cdot) \leq 1$ and an upper bound on the daughter intensity $h(\cdot)$ [13].

The Metropolis-Hastings method is a two-step discrete time chain, which proposes to change the current state $\mathbf{x}$ to a new candidate state $\mathbf{x}'$, randomly sampled from a probability density $q(\mathbf{x},\cdot)$. The proposal $\mathbf{x}'$ is accepted with probability

$$A(\mathbf{x},\mathbf{x}') = \min\left\{1, \frac{p(\mathbf{x}' \mid y)\, q(\mathbf{x}',\mathbf{x})}{p(\mathbf{x} \mid y)\, q(\mathbf{x},\mathbf{x}')}\right\}.$$

It is easily verified that transitions out of state $\mathbf{x}$ are belanced by transitions into $\mathbf{x}$:

$$p(\mathbf{x}' \mid \mathbf{y})q(\mathbf{x}',\mathbf{x})A(\mathbf{x}',\mathbf{x}) = p(\mathbf{x} \mid \mathbf{y})q(\mathbf{x},\mathbf{x}')A(\mathbf{x},\mathbf{x}').$$

The transition densities $q(\cdot,\cdot)$ are built as follows:

- with probability $q(\mathbf{x})$ generate a new point u from a density $b(\mathbf{x},u)$ with respect to ν;
- otherwise (with probability $1 - q(\mathbf{x})$) delete a point $x_i \in \mathbf{x}$ at random;

The simplest choices for $b(\cdot,\cdot)$ and $q(\cdot)$ are

$$q(\cdot) \equiv \frac{1}{2},\ b(\cdot,\cdot) \equiv \frac{1}{\nu(U)},$$

but this may tend to generate too many proposals with low acceptance probabilities. Another possibility resembling spatial birth-and-death processes is to set

$$b(\mathbf{x},u) = \frac{f(\mathbf{y} \mid \mathbf{x} \cup \{u\})\, p(\mathbf{x} \cup \{u\})}{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})} \frac{1}{B(\mathbf{x})}$$

and

$$q(\mathbf{x}) = \frac{B(\mathbf{x})}{n(\mathbf{x}) + B(\mathbf{x})},$$

where

$$B(\mathbf{x}) = \int_U \frac{p(\mathbf{x} \cup \{u\} \mid \mathbf{y})}{p(\mathbf{x} \mid \mathbf{y})}\, d\nu(u)$$

or some other density that tends to select proposals u where $p(\mathbf{x} \cup \{u\} \mid \mathbf{y})$ is large.

Under mild conditions, the resulting Markov chains converge to $p(\mathbf{x} \mid \mathbf{y})$ [9].

From a computational point of view, to (say) add $u \in U$ to $\mathbf{x}$ we need to evaluate

$$\frac{f(\mathbf{y} \mid \mathbf{x} \cup \{u\})}{f(\mathbf{y} \mid \mathbf{x})} = H(u) \prod_{j=1}^{m} \left[1 + \frac{h(y_j \mid u)}{\sum_{i=0}^{n} h(y_j \mid x_i)}\right], \tag{2}$$

where $H(u) = \exp\left\{-\int_T h(t \mid u)\, d\mu(t)\right\}$, and

$$\frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})}. \tag{3}$$

Ratios (2) are straightforward to calculate and can be compared to the Hough transform [10] in the sense that each point y_j votes with variable strength for a cluster centre at point u. Ratios of the form (3) are easy to compute if $p(\cdot)$ is a nearest-neighbour Markov point process (1). In particular, the intractable normalising constant cancels out.

Parameters in f and p can be estimated in advance or during iteration. Alternatively, a Gibbs sampler can be specified but note that this approach requires an extra set of prior distributions, one for each parameter.

4 Offspring labelling

So far, we concentrated on estimating the process $\mathbf{x}$ of cluster centres, but it might well be of interest to label the observed events by the cluster they belong to. This opens the possibility to estimate functionals such as the probability that two data points are siblings, or the distribution function of offspring displacements.

In order to be able to do inference on cluster membership, we need the full distribution of $\mathbf{w} = \{(x_0, Z_{x_0}), \ldots, (x_n), Z_{x_n})\}$, parents $\{x_1, \ldots, x_n\}$ marked by their offspring Z_{x_i}, $i = 0, \ldots, n$.

Recall that (in contrast to e.g. the analysis of finite mixture models [7]) the number of points in the pattern is not fixed in advance. This is quite crucial, as it implies the impossibility to build a Gibbs sampler, alternatingly sampling the marks (or centres) given data and centers (or marks). We propose a two-step Metropolis-Hastings algorithm [13], where the transitions consist of births and deaths of centres, followed by an adjustment of the marks, by drawing from the conditional distribution of offspring. It is important to note that, when a new configuration $\mathbf{x}'$ is generated from the current $\mathbf{x}$ and offspring $\mathbf{Z}$, $\mathbf{x}'$ does not have to conform to $\mathbf{Z}$ (eg. $\mathbf{x}'$ could have one point fewer than $\mathbf{x}$). We then update the daughter sets to $\mathbf{Z}'$ and require conformity with $\mathbf{x}'$. Specifically, we have the following algorithm. Given an initial parent configuration $\mathbf{x}$, a labelling $\mathbf{Z_x} = (Z_{x_i})_{i=0}^{n(\mathbf{x})}$ and data $\mathbf{y}$, alternate between a Metropolis-Hastings and an adjustment step as follows.

1. given $\mathbf{w} = \{(x_0, Z_{x_0}), \ldots, (x_n, Z_{x_n})\}$ and $\mathbf{y}$, perform a Metropolis-Hastings step for $\mathbf{x}$ only, *independent* of the marks $\mathbf{Z_x}$, yielding the new configuration $\mathbf{x}'$;

2. given $\mathbf{x}'$, $\mathbf{y}$ reallocate the offspring by a Gibbs step, sampling $\mathbf{Z'_{x'}}$ from the conditional distribution $P(\mathbf{Z'_{x'}} \mid \mathbf{x}', \mathbf{y})$. Replace $\mathbf{x}$ by $\mathbf{x}'$, $\mathbf{Z_x}$ by $\mathbf{Z'_{x'}}$ and return to step 1.

The detailed balance equations reduce to those for the centres only (see Section 3) and hence convergence of the Markov process to $p(\mathbf{w} \mid \mathbf{y})$ follows under similar conditions.

In step 2, an assignment of data to clusters can be interpreted as an *ordered* partition $\phi : \mathbf{y} \to \mathbf{x}$ of the data. The labels $\phi(\cdot)$ ascribing each observation to a cluster centre are independent, with probabilities

$$\frac{h(y_j \mid x_{\phi(y_j)})}{\sum_i h(y_j \mid x_i)}.$$

Hence, simulation of the mark process is easy.

The independence of labels provides a justification for the *nearest-parent classifier* [11]. Given a set of sites $\mathbf{y} = \{y_1, \ldots, y_m\}$ and a set of parents $\mathbf{x} = \{x_0, \ldots, x_n\}$, suppose the task is to assign to each y_j a label $\phi(j) \in \{0, \ldots, n\}$ corresponding to centre $x_{\phi(j)}$. By independence, a maximum likelihood classifier is

$$\hat{\phi}(j) = \text{argmax}_{i=0,\ldots,n} h(y_j \mid x_i), \tag{4}$$

which assigns each point to the closest centre if $h(y_j \mid x_i)$ is a decreasing function of $||y_j - x_i||$

To conclude this section, if one considers *unordered* partitions of the data instead of unordered ones, a Gibbs sampler [4] can be developed. Intuitively, these unordered partition elements correspond to 'sibling sets' generated by a common, if unspecified, parent. Then, by allowing parents not having any offspring at all, the number of cluster centres can be changed, but the conditional distributions involved are rather awkward [13].

5 Applications

The simplest examples arise in spatial statistics where $T \subseteq U$ are bounded regions in the plane. Typical applications include forestry, where the task is to reconstruct the ancestors of the current generation of trees in a wood or the epidemiological analysis of rare diseases.

As an illustration, Figure 1 shows the locations of 62 redwood seedlings in a square of side approximately 23 m. The data was extracted by Ripley [18] from a larger data set in Strauss [20]. A biological explanation [20] for the apparent clustering is the presence of stumps known to exist in the plot, but whose position has not been recorded. Our goal is to reconstruct these centres of clustering.

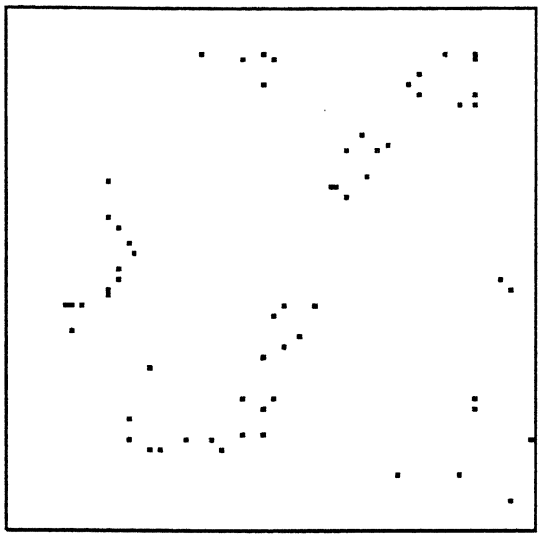

Figure 1: Positions of 62 redwood seedlings in a unit square (Ripley 1977).

Previous analyses of this data set include Strauss [20] who attempted to fit a pairwise interaction Markov model, Ripley [18], Diggle [8] and Lawson [11]. From a biological point of view, the above works tend to find an implausibly large number of stumps (clusters); 25 in [8] and 16 in [11], making a case for a Bayesian analysis with a prior distribution penalising configurations with too many stumps close together. The analysis below is taken from Van Lieshout [13].

Following [8, 11] we assume the number of daughters per parent is Poisson and seedlings follow a radially symmetric Gaussian distribution around their ancestor. In contrast to the aforementioned papers, a pairwise interaction Markov prior (1) with strict inhibition $g \equiv \gamma < 1$ is introduced. We used an interaction distance $r = .084$ [8], that is $u \sim v$ if and only if u and v are less than r apart.

Using a spatial birth-and-death sampler, a realisation from the posterior distribution of stumps is shown in Figure 2. Figure 3 displays the estimated posterior intensity, roughly speaking the posterior 'probability' that a point u belongs to $\mathbf{x}$. For reasons of clarity, here *black* corresponds to high values.

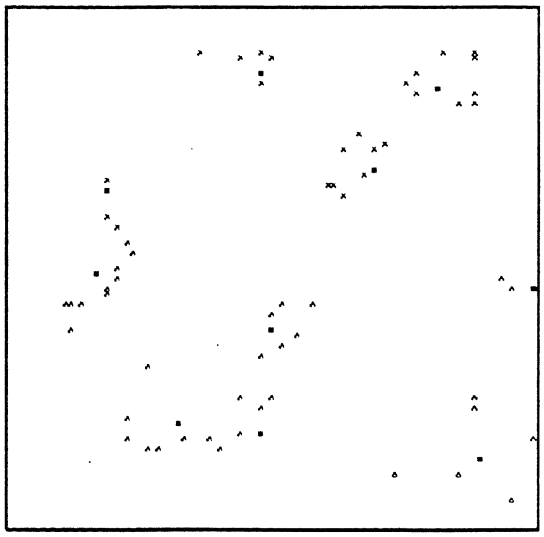

Figure 2: Realisation from the posterior distribution taken after 2 time units (black), for a gaussian model with $\mu = 6.5$, $\sigma = .05$ and a prior with $\log\beta = \log\gamma = -10$, $r = .084$. The data is displayed in grey.

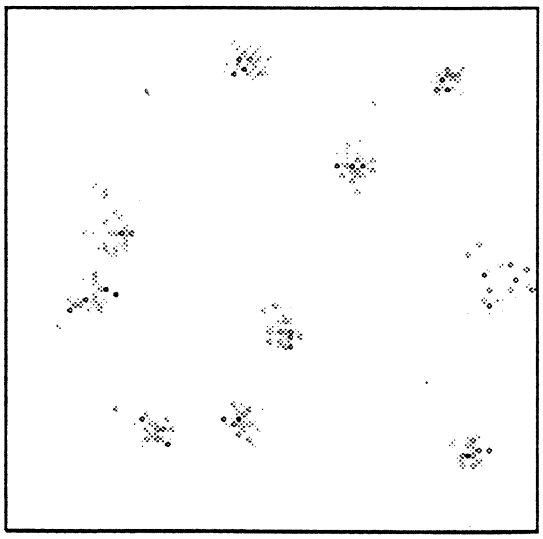

Figure 3: Posterior intensity of redwood seedlings (Ripley) estimated over 50 time units, for a gaussian model with $\mu = 6.5$, $\sigma = .05$ and a prior with $\log\beta = \log\gamma = -10$, $r = .084$.

For an analysis of the full data [20] see [13].

As another example [1], suppose that the data again consist of a point pattern in a bounded region $T \subseteq \mathbb{R}^2$, but that the points are believed to lie close to a curve and the objective is to estimate the curves. This includes the image analysis task of joining a dot pattern into a curvilinear boundary,

but also eg. the identification of ancient roads or trade routes given information about the location of archaeological finds such as pottery or coins [19, p. 139], or the analysis of earthquake occurrences in relation to geographical fault patterns.

Our final example is the problem of identifying large scale edges in a scene using the output of a low-level edge detector. The 'data' $\mathbf{y}$ consist of a pattern of line segments and the objective is to cluster them around a small number of larger line segments [14].

Let T denote the set of possible outputs of the low-level edge detector. For example these may be line segments restricted to have unit length (= 1 pixel width) and orientation which is a multiple of 45 degrees. The space U of objects we are looking for in this case also contains line segments, but of unrestricted length and orientation.

Hence, $\mathbf{y}$ is a superposition of conditionally independent line segment processes $\mathbf{Z}_{x_i}$ associated with each true line segment x_i. Typically the expected number of segments in $\mathbf{Z}_{x_i}$ will depend on the length of x_i. The benefits of a prior model for $\mathbf{x}$ include the ability to encourage long lines and continuity between lines, and to penalise lines that cross one another.

References

[1] A. J. Baddeley and M. N. M. van Lieshout. Stochastic geometry models in high-level vision. K. V. Mardia and G. K. Kanji (Eds.) Statistics and Images, Volume 1. *Journal of Applied Statistics*, 20:233–258, 1993.

[2] A. J. Baddeley and J. Møller. Nearest-neighbour Markov point processes and random sets. *International Statistical Review*, 57:89–121, 1989.

[3] M. Baudin. Note on the determination of cluster centres from a realization of a multidimensional Poisson cluster process. *Journal of Applied Probability*, 20:136–143, 1983.

[4] J. Besag, P.J. Green, D. Higdon and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 1995. To appear.

[5] J. Besag and J. Newell. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*, 154:143–155, 1991.

[6] D.J. Daley and D. Vere-Jones. *An introduction to the theory of point processes.* Springer Verlag, New York, 1988.

[7] J. Diebolt and C.P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, 56:363–375, 1994.

[8] P. J. Diggle. *Statistical analysis of spatial point patterns.* Academic Press, London, 1983.

[9] C. J. Geyer and J. Møller. Simulation procedures and likelihood inference for spatial point processes. Research report 260, Mathematical Institute, University of Aarhus, january 1993.

[10] J. Illingworth and J. Kittler. A survey of the Hough transform. *Computer Vision, Graphics and Image Processing*, 44:87–116, 1988.

[11] A. Lawson. Discussion contribution. *Journal of the Royal Statistical Society, Series B*, 55:61–62, 1993.

[12] A. B. Lawson, M. N. M. van Lieshout and A. J. Baddeley. Markov chain Monte Carlo methods for clustering of spatial point processes. In preparation, 1995.

[13] M. N. M. van Lieshout. *Stochastic geometry models in image analysis and spatial statistics*, PhD thesis Vrije Universiteit Amsterdam, 1994. To appear as monograph in CWI-tract series, 1995.

[14] P. Nacken. A metric for line segments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:1312–1318, 1993.

[15] S. Openshaw, M.G. Charlton, C. Wymer, and A.W. Craft. A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal on Geographical Information Systems*, 1:335–358, 1987.

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

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

[18] B. D. Ripley. Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:172 - 212, 1977.

[19] D. Stoyan, W. S. Kendall and J. Mecke. *Stochastic Geometry and its Applications.* John Wiley and Sons, Chichester, 1987.

[20] D.J. Strauss. A model for clustering. *Biometrika*, 62:467–475, 1975.