Modelling point patterns with linear structures

Møller, Jesper; Rasmussen, Jakob Gulddahl

Published in:
Stereology and Image Analysis, Ecs10

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

Stereology and Image Analysis. Ecs10 - Proceedings of the 10th European Congress of ISS,
(V.Capasso et al. Eds.), The MIRIAM Project Series, ESCULAPIO Pub. Co., Bologna, Italy, 2009

MODELLING POINT PATTERNS WITH LINEAR STRUCTURES

JESPER MØLLER AND JAKOB G. RASMUSSEN

Department of Mathematical Sciences, Aalborg University, Fredrik Bajersvej 7G, 9220 Aalborg, Denmark
e-mail: jm@math.aau.dk, jgr@math.aau.dk

ABSTRACT

Many observed spatial point patterns contain points placed roughly on line segments. Point patterns exhibiting such structures can be found for example in archaeology (locations of bronze age graves in Denmark) and geography (locations of mountain tops). We consider a particular class of point processes whose realizations contain such linear structures. Such a point process is constructed sequentially by placing one point at a time. The points are placed in such a way that new points are often placed close to previously placed points, and the points form roughly line shaped structures. We consider simulations of this model and compare with real data.

Keywords: Archaeology; Dirichlet Tesselation; Geology; Likelihood; Simulation; Spatial Point Processes.

1 INTRODUCTION

Many observed spatial point patterns contain points placed roughly on line segments; we will refer to these structures as linear structures. In the data section below we consider two datasets, both of which contain linear structures (see Figures 3 and 4). The first data set is the locations of barrows (bronze age burial sites) in a region of Denmark, and the other data set is the locations of mountain tops in a region of Spain.

Blackwell (2001), Blackwell & Møller (2002), and Skare *et al.* (2006) consider point process models with linear structures close to the edges of (deformed) Dirichlet (or Voronoi) tessellations. However, for the two abovementioned data sets and many others, the exact mechanism responsible for the formations of lines is unknown. Thus the development of tractable and practically useful spatial point process models capable of producing point patterns with linear structures becomes important.

In this paper we develop a particular class of such models using a sequential construction by placing one point at a time. The model is easy to simulate and its likelihood function is known on closed form. Perhaps somewhat surprising it is a flexible model for producing linear structures without incorporating any lines into the model.

The paper is organized as follows. Section 2 defines the model, Section 3 presents the data sets, Section 4 concerns simulation of the model, and finally Section 5 discusses inference, model checking, and extentions of the model.

2 MODEL

Figures 3 and 4 show two kinds of points, those roughly located along lines, and others which seem to be distributed fairly randomly across the observation region. We model this by a superposition of two point processes called the 'cluster process' and the 'background process'. Briefly, the cluster process is constructed sequentially, and each cluster point can be of two types: 'dependent' cluster points and 'independent' cluster points, where the independent cluster points (and also the background points) are independent and uniformly distributed, while each dependent cluster point is attracted by previously generated cluster points.

2.1 LIKELIHOOD

This section specifies the likelihood when we have no missing data in the following sense. The likelihood is given below by the joint distribution of the cluster process $x_c = (x_1, \ldots, x_k)$ and the background process $x_b = (x_{k+1}, \ldots, x_n)$, where the n points $x_1, \ldots, x_n$ are contained in a given bounded convex region $W \subset \mathbb{R}^2$ of area $|W| > 0$. The assumption that W is convex becomes important later. In our applications, the data $z = (z_1, \ldots, z_n)$ is a permutation of $x = (x_1, \ldots, x_n)$. This permutation as well as k and the knowledge whether each z_i is a cluster or background point are unknown, i.e., they constitute the missing data.

We let $m = n - k$ denote the number of background points, and make the following model assumptions, where $0 \leq p \leq 1$, $0 \leq q \leq 1$, and $\lambda > 0$ are model parameters:

(i) The number of points n is fixed.

(ii) The number of cluster points k is a random variable following a binomial distribution with index n and probability q.

(iii) Conditional on k, we have that x_c and x_b are independent.

(iv) Conditional on k, the m background points in x_b are independent and uniformly distributed on W (a socalled binomial point process on W).

(v) Conditional on k, if $k > 0$ then the first cluster point x_1 follows a uniform distribution on W, and if $2 \leq i \leq k$ and we also condition on $x_1, \ldots, x_{i-1}$ then the ith cluster point x_i follows a density $f(\cdot|x_1, \ldots, x_{i-1}; p, \lambda)$ with respect to Lebesgue measure on W. Further,

$$f(\cdot|x_1, \ldots, x_{i-1}; p, \lambda) = p \times h(\cdot|\{x_1, \ldots, x_{i-1}\}; \lambda) + (1-p) \times \frac{1}{|W|} \quad (1)$$

depends only on $(x_1, \ldots, x_{i-1})$ through the point pattern $\{x_1, \ldots, x_{i-1}\}$, and the density $h(\cdot|\{x_1, \ldots, x_{i-1}\}; \lambda)$ is specified below by formula (4).

In the mixture density (1), the uniform density on W is used for the distribution of an independent cluster point, and the density $h(\cdot|\{x_1, \ldots, x_{i-1}\}; \lambda)$ for the distribution of a dependent cluster point. Note that an independent cluster point x_i is statistically independent of previous cluster points $x_1, \ldots, x_{i-1}$, while it influences the distribution of later dependent cluster points. Moreover, (1) implies that the location of a new cluster point does not depend on the time-ordering of the previous cluster points.

One way of simulating our model is by first generating mutually independent and uniformly distributed points $y_1, \ldots, y_n$ in W. We independently divide these points into background points, independent cluster points, and dependent cluster points in accordance to the probabilities $(1-q)$, $q(1-p)$, and pq, respectively. If y_i is a background or independent cluster point, then $x_i = y_i$. If y_i is a dependent cluster point, it is transformed into a dependent cluster point x_i, depending on other cluster points x_j with $j < i$ as specified below, and involving some further simulation steps given by (A)-(D) also below.

Combining (i)-(v), we obtain that

$$\pi(x_c, x_b|q, p, \lambda) = \binom{n}{k} q^k \left(\frac{1-q}{|W|}\right)^m \times \prod_{i=1}^{k} f(x_i|x_1, \ldots, x_{i-1}; p, \lambda) \quad (2)$$

is the joint density of (x_c, x_b) with respect to the measure ν on $\cup_{l=0}^{n} W^l \times W^{n-l}$ given by $\nu = \sum_{l=0}^{n} \nu_k$, where ν_l is the product measure of Lebesgue measure on W^l and Lebesgue measure on W^{n-l} (with obvious modifications if $l = 0$ or $l = n$). In (2) and elsewhere, for notational convenience, we interpret $f(\cdot|x_1, \ldots, x_{i-1}; p, \lambda)$ as the uniform density on W if $i = 1$.

If we had 'no missing data' in the sense that (x_c, x_b) is observed but we do not know whether each cluster point is an independent or dependent cluster point, then (2) would specify the likelihood for $\theta = (q, p, \lambda)$. However, when considering the data in the data section, the following quantities u, ω, η are missing data. Let $u = (u_1, \ldots, u_n)$ where $u_i = 1$ if z_i is one of the cluster points, and $u_i = 0$ if z_i is one of background points. Given the value of u, define the permutation $\omega = (\omega_1, \ldots, \omega_k)$ of those i with $u_i = 1$ such that $x_c = (z_{\omega_1}, \ldots, z_{\omega_k})$, and the permutation $\eta = (\eta_1, \ldots, \eta_m)$ of those i with $u_i = 0$ such that $x_b = (z_{\eta_1}, \ldots, z_{\eta_m})$. In other words, z_{ω_i} is the ith cluster point, and z_{η_j} is the jth background point. Thus (x_c, x_b) is in a one-to-one correspondence with (z, u, ω, η), with a density which for each fixed value of (u, ω) is constant for all possible values of η, cf. (2). Consequently, conditional on the data z, we have that (u, ω) is in a one-to-one correspondence with x_c and the point pattern $\{x_{k+1}, \ldots, x_n\}$ of background points, with probability mass density

$$\pi(u, \omega|z; \theta) \propto \frac{1}{k!} q^k \left(\frac{1-q}{|W|}\right)^m \times \prod_{i=1}^{k} f(x_i|x_1, \ldots, x_{i-1}; p, \lambda). \quad (3)$$

2.2 THE CONDITIONAL DENSITY OF DEPENDENT CLUSTER POINTS

We now turn to specifying a particular form of the conditional density of dependent cluster points h, such that realizations of the model exhibit linear structures. Conditional on pairwise distinct cluster points $x_1, \ldots, x_{i-1}$ with $2 \leq i \leq k$, we define the density $h(x_i|\{x_1, \ldots, x_{i-1}\}; \lambda)$ in (1) by

$$h(x_i|\{x_1, \ldots, x_{i-1}\}; \lambda) = \frac{l_i^2 \exp(-r_i^2/\lambda)}{\lambda |W| (1 - \exp(-l_i^2/\lambda))}, \quad (4)$$

for $0 < r_i < l_i$, where the notation means the following. Let $\|\cdot\|$ denote Eucledian distance, and

$$C_j = \{\xi \in \mathbb{R}^2 : \|\xi - x_j\| \leq \|\xi - x_{j'}\|, j' = 1, \ldots, i-1\}$$

the cells of the Dirichlet (or Voronoi) tessellation of $\mathbb{R}^2$ with nuclei $x_1, \ldots, x_{i-1}$ (Okabe *et al.*, 2000), where $j =$

$1,\ldots,i-1$. Then x_i belongs almost surely to a unique Dirichlet cell, say C_j, and $C_j \cap W$ is convex (this is where the assumption that W is convex is used). Define $r_i = \|x_i - x_j\|$, and l_i as the length of the line segment through x_i and with endpoints at x_j and the boundary of $C_j \cap W$. See the example in Figure 1, where $i = 4$ and $j = 2$. Under the distribution (4), x_i appears in cell C_j with probability $p_j = |C_j \cap W|/|W|$. If we condition on that $x_i \in C_j$, and $N_2(x_j, \sigma^2 I)$ denotes the radially symmetric bivariate normal distribution with mean x_j and standard deviation $\sigma = \sqrt{\lambda/2}$, then x_i follows the restriction of $N_2(x_j, \sigma^2 I)$ to $C_j \cap W$.

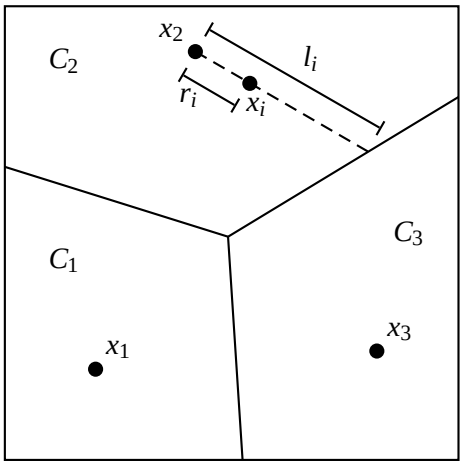

Fig. 1. *Example with $i = 4$ and three cluster points x_1, x_2, x_3, and their respective Dirichlet cells C_1, C_2, C_3. The new cluster point x_i and the distances l_i and r_i are shown.*

Neither the calculation of the distribution $p_1,\ldots,p_{i-1}$ or the construction of the entire Dirichlet tessellation is needed when evaluating the density (4) or simulating from this distribution as explained in the following.

To evaluate the density (4) we use the following steps.

(a) Find the closest point x_j to x_i with $j < i$, the half-line L_j with endpoint x_j and passing through x_i, and the intersection point v_j between L_j and the boundary of W. Calculate $l_j = \|v_j - x_j\|$.

(b) For each $j' \in \{1,\ldots,i-1\} \setminus \{j\}$, find the line $L_{j'}$ passing through $(x_j + x_{j'})/2$ and perpendicular to the line through x_j and $x_{j'}$. If $v_{j'}$ is the intersection point between L_j and the $L_{j'}$, then calculate $l_{j'} = \|v_{j'} - x_j\|$. If $L_j \cap L_{j'} = \emptyset$, then set $l_{j'} = \infty$.

(c) Return $r_i = \|x_i - x_j\|$ and $l_i = \min\{l_1,\ldots,l_{i-1}\}$.

Figure 2 shows an example, where $i = 5$, step (c) returns $l_5 = \|v_3 - x_4\|$, and the area around x_4 bounded by the lines L_1, L_2, L_3 and the boundary of W is the Dirichlet cell C_4.

We can easily make a simulation under (4) by the following steps.

(A) Generate a uniform point y_i in W, which is independent of $x_1,\ldots,x_{i-1}$.

(B) Find the (almost surely unique) closest point x_j to y_i ($1 \le j < i$), the half-line L_j through y_i and with endpoint x_j, and the distance l_i from x_j to the intersection point between L_j and the boundary of W.

(C) Generate r_i^2 from an exponential distribution with parameter $1/\lambda$ and restricted to the interval $(0, l_i^2)$.

(D) Return x_i as the point on L_j with distance r_i from x_j.

In (B), we calculate l_i in the same way as in (a)-(c) above.

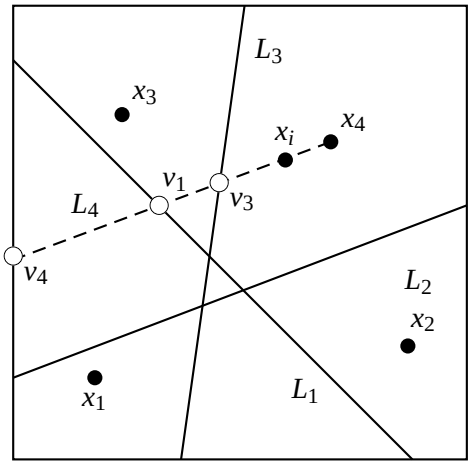

Fig. 2. *Example with $i = 5$, showing four cluster points $x_1,\ldots,x_4$ and a new cluster point x_i (filled circles), where x_i is closest to x_4. The half line L_4 (dashed line), the lines L_1, L_2, L_3 (solid lines), and the intersection points v_1, v_3 and v_4 (empty circles) are also shown.*

3 DATA

The first data set is the location of barrows in a 15×15 km region in Western Jutland, Denmark. Barrows, which are bronze age burial sites resembling small hills, are important sources of information for archaeologists. Contrary to other areas of Denmark, a relatively large proportion of the barrows are still present in the Western Jutland due to less intensive agriculture. Figure 3 shows the locations of the barrows.

The spatial distribution of barrows across Denmark shows a variety of patterns, particularly clusters of points along various lines, where some lines seem to stretch across the landscape for hundreds of kilometers. The barrow lines have traditionally been regarded as reflecting a prehistoric system of roads, cf. Müller (1904), though there are other potential explanations for this phenomenon, see e.g. Sahlquist (2001).

The model in this paper has the following interpretation in the context of this data: the barrows are placed according to a 'local decision-making rule', where we interpret y_i as the location where a person died, and

- the survivers decide if the point should be a background point, independent cluster point, or dependent cluster point
- in case y_i is a dependent cluster point, the person is buried in a barrow close to the closest previous cluster point, justifying the terminology 'cluster process' for x_c
- if instead the y_i is a background point or an independent cluster point, then he is buried where he died.

In other words, if the model produces point patterns resembling the data, this indicates that placing barrows close to previously placed barrows may be an alternative explanation to the observed linear structures in the point pattern of barrows.

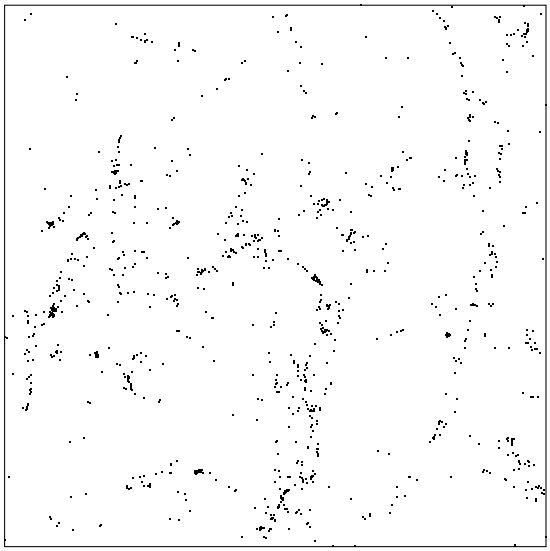

Fig. 3. *The locations of barrows in a* 15×15 *km region.*

The other data set is the location of mountain tops in a 10.5×7.5 km region in Northern Spain. The data has been taken from a hiking map, and is shown in Figure 4. Many mountain tops are located along linear structures, which of course is a consequence of the fact that many tops are located along the mountain ridges. However, visually the linear structures are somewhat obscured by the many tops located off the ridges. Note that the height of each top is known, and many of the tops not located on the ridges are lower than the ones on the ridges, but for this paper we will ignore the additional information of height and only consider the point pattern of positions. In our model, all the tops off the ridges are simply set to be background points.

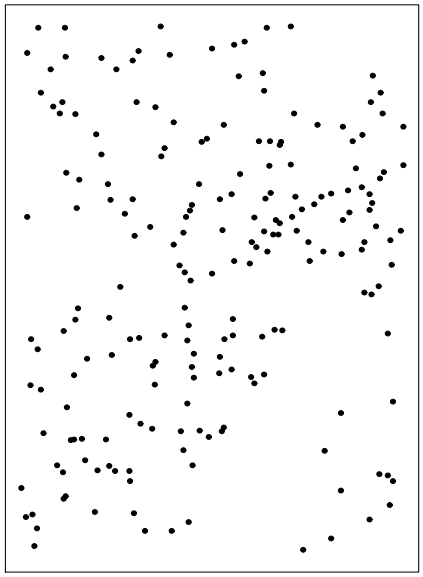

Fig. 4. *The locations of mountain tops in a* 10.5×7.5 *km region.*

4 SIMULATION

We now show some simulations of the model with various choices of parameters to see how flexible the model is and whether it can produce point patterns with some resemblance to the data.

Figure 5 shows simulated point pattern with the same number of points as in the barrows data set, and where the parameters are $(p, q, \sigma) = (0.98, 0.95, 150\,\mathrm{m})$. These parameters have been chosen by simulating the model with various parameter settings and choosing the simulation that visually resembles the data best. The first observation we make is that the model is capable of producing linear structures, although the mechanism behind this model is only a method of attracting new points to

previously placed points, and no actual line segments are incorporated into the model. Furthermore, there are many similarities between the patterns in Figures 3 and 5: long linear structures with short linear structures extending from them, and large gaps with no or few points. The model has a higher tendency to produce short linear structures extending from the long linear structures than the data in this particular simulation, but we should of course remember that the simulation has been made with a rather arbitrary choice of parameters and other parameter settings may produce better fits; we return briefly to the issue of parameter estimation in Section 5, but a full discussion of this topic is beyond the scope of this paper. One obvious feature of the data that is not found in the simulation is a few small, but densely packed, clusters of points; such clusters will never appear in the model nomatter the choice of parameters.

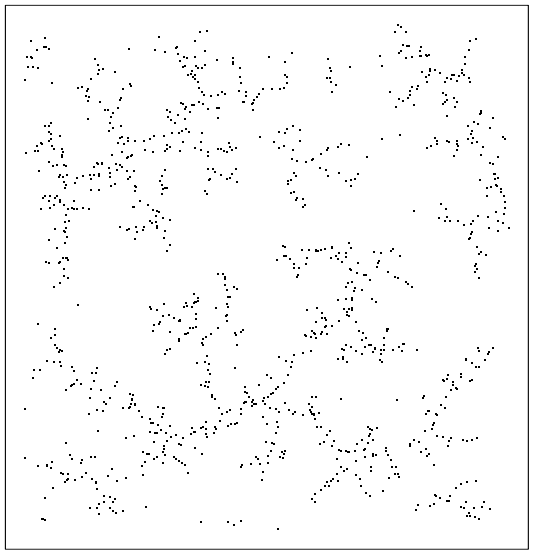

Fig. 5. *A simulation with* $n = 1147$ *and parameters* $(p,q,\sigma) = (0.98, 0.95, 150\,\mathrm{m})$

Figure 6 shows a simulation with the same number of points as in the mountain data set, and with parameters $(p,q,\sigma) = (0.98, 0.98, 400\,\mathrm{m})$; again the parameters have been chosen by trial and error. Comparing with the data in Figure 4, we see no obvious discrepancies. Both contain medium length linear structures, gaps with no or few points, and quite many solitary points. We have also made another simulation with parameters $(p,q,\sigma) = (0.95, 1.00, 200\,\mathrm{m})$ mainly to illustrate that with an adjustment to the parameters we can get linear structure which are much more visible; see Figure 7. Obviously this simulation has much clearer linear structures than the data.

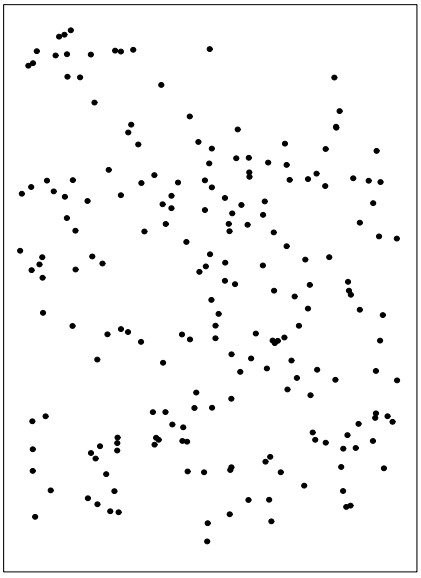

Fig. 6. *A simulation with* $n = 203$ *and parameters* $(p,q,\sigma) = (0.98, 0.98, 400\,\mathrm{m})$

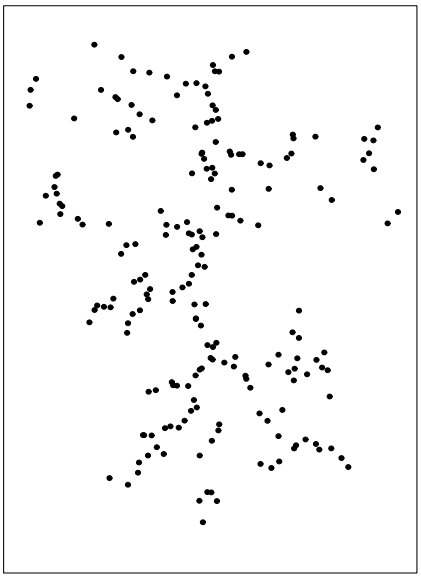

Fig. 7. *A simulation with* $n = 203$ *and parameters* $(p,q,\sigma) = (0.95, 1.00, 200\,\mathrm{m})$

5 DISCUSSION

This paper is exploring the limits of our model by comparing simulations to actual observed data sets.

Obviously, a proper statistical analysis should involve a much more thorough treatment of the data.

We intend to develop a Bayesian model with priors for the three parameters (p, q, σ). Although the model is simple to simulate due to its sequential construction, it does not seem possible to estimate parameters analytically, using e.g. the posterior mean. However, since the likelihood is known completely (except for missing data), an MCMC based approach using a Metropolis-within-Gibbs algorithm can be used for making approximate posterior simulations of the parameters and the missing data. Hastings ratios for updates both for parameters and missing data are easily found using equations (1), (3) and (4). As a by-product of this approach, we can also estimate the missing data, which means that this model can be used to estimate whether or not a particular point belongs to a linear structure.

Another issue is that of model checking. A common way to check the fit of a point process model is to estimate various summary statistics and compare with theoretical calculations for the fitted model, or compare with simulations from the model, if theoretical calculations are intractable. Many standard summary statistics are available (see e.g. Møller & Waagepetersen (2004)), but few of them are useful for checking the linear structures which are the focus of this model. Developing summary statistics specifically aimed at the shape or size of the linear structures are important in checking the fit of the model.

Finally, there are many theoretically interesting or practically useful extensions of the model which can be explored. The model in this paper is homogeneous in the sense that all points are initially placed according to a homogeneous binomial process (dependent cluster points are then moved). Incorporating covariate information into the model to obtain an inhomogeneous model can be practically useful. For example, information about the terrain types throughout the landscape could potentially improve the model for the barrows. Another generalisation is to extend the model to $\mathbb{R}^2$. Only minor modifications are required to obtain a stationary point process, e.g. using Poisson processes rather than binomial processes, and changing the order of the cluster points such that all independent cluster appear first to form a Dirichlet tesselation of $\mathbb{R}^2$. Although the model construction is easy, many aspects of this model can prove difficult, such as perfect simulation on a finite subset (i.e., simulation without edge effects). The infinite model is also of practical relevance since it gets rid of the artificial boundaries of the finite model (this is the reason why Figure 7 has almost no points close to the boundary).

ACKNOWLEDGEMENTS

This research was supported by the Danish Natural Science Research Council (grant 272-06-0442, “Point Process Modelling and Statistical Inference”). The authors would like to thank Geoff Nicholls, Kasper Lampert Johansen, Steffen Terp Laursen, Mads Kähler Holst, Øjvind Skare, and Dietrich Stoyan for useful discussions.

REFERENCES

Blackwell, P. G. (2001). Bayesian inference for a random tessellation process. *Biometrics*, 57: 502–507.

Blackwell, P. G. and Møller, J. (2002). Bayesian analysis of deformed tessellation models. *Advances in Applied Probability*, 35: 4–26.

Müller, S. (1904). Vei og bygd i sten- og bronzealderen. *Aarbøger for Nordisk Oldkyndighed og Historie*: 1–64.

Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall, Boca Raton, Florida.

Okabe, A., Boots, B., Sugihara, K. and Chiu, S. N. (2000). *Spatial Tessellations. Concepts and Applications of Voronoi Diagrams*. 2nd edition. Wiley, Chichester.

Sahlquist, L. (2001). Territorial behaviour and communication in a ritual landscape. *Geografiska Annaler*, 83 B, 2: 79–102.

Skare, Ø., Møller, J., and Jensen, E. B. V. (2006). Bayesian analysis of spatial point processes in the neighbourhood of Voronoi networks. *Statistics and Computing*, 17: 369–379.