Statistica Sinica **22** (2012), 1717-1736
doi:http://dx.doi.org/10.5705/ss.2010.245

A GENERALIZATION OF THE NEYMAN-SCOTT PROCESS

Chun Yip Yau and Ji Meng Loh

The Chinese University of Hong Kong and AT&T Labs-Research

Abstract: In this paper we introduce a generalization of the Neyman-Scott process Neyman and Scott (1958) that allows for regularity in the parent process. In particular, we consider the special case where the parent process is a Strauss process with offspring points dispersed about the parent points. Such a generalization allows for point realizations that show a mix of regularity and clustering in the points. We work out a closed form approximation of the K function for this model and use this to fit the model to data. The approach is illustrated by applications to the locations of a species of trees in a rainforest dataset.

Key words and phrases: Gibbs process, Neyman-Scott process, K-function, regular point process.

1. Introduction

In this paper we introduce a model for a spatial point process that contains a mix of regularity and clustering. It is a generalization of the model introduced in Neyman and Scott (1958) to fit to galaxy data. In the Neyman-Scott model, the unobserved parent points follow a Poisson process with some intensity λ_p. Each parent point produces a possibly random number of offspring points, randomly dispersed about the parent point. A realization of a Neyman-Scott process consists of the set of all the offspring points.

Our generalization involves replacing the Poisson parent process with a regular process. This allows for point patterns consisting of regularly spaced clusters of points. Gibbs processes define point models via interactions between pairs, triplets, and so on, of points. They are often studied in statistical physics and are a convenient way to specify spatial point models with regular as opposed to clustered patterns. In our generalization we use a Strauss process (Strauss (1975); Ripley and Kelly (1977)), the simplest of the Gibbs processes, as the parent point process.

The Strauss process includes the Poisson process within its parameter set. Thus our generalized model has the Neyman-Scott model as a special case. The advantage of the new Generalized Neyman-Scott (GNS) process is that it incorporates additional flexibility, allowing for regularly spaced clusters of points. For example, the Neyman-Scott process was used to model the positions of trees in a rainforest in Waagepetersen (2007). However, it is conceivable that while offspring are clustered around the parents, the parent trees that have survived are regularly spaced due to competition for nutrients and sunlight. Our model can capture this phenomenon.

There are many comprehensive overviews of spatial point processes, e.g., Ripley (1988); Stoyan and Stoyan (1994); Diggle (2003). All these works contain discussions of Gibbs (also called Markov) point processes. Taylor, Dryden and Farnoosh (2001) defined a point model, called a nearly regular point process, where the parent points lie on a fixed regular grid. Each parent point has exactly one offspring dispersed about the parent point according to a bivariate Gaussian density. They derived an expression for the K function of this process. Our model differs from the nearly regular point process in that the parent points, although also regular, are random. Also, each parent point has a random number of offspring points. Thus our model exhibits clustering at small distances that is not present in the nearly regular point process.

Møller and Torrisi (2005) introduced Generalized Shot Noise Cox Processes (GSNCP) that can be considered as Cox cluster processes with a general parent process, dispersion density, and offspring intensity that they denote by $\Phi_{\text{cent}}, k_{b_j}$, and γ_j, respectively. The GSNCP contains a very wide class of point process models since Φ_{cent} can be any point process and (b_j, γ_j) can be fixed or random with some distribution. Our GNS model is contained in this class (see Example 4 of Section 2.3 of their paper). While they derive very general results for summary statistics of GSNCPs, these formulas involve high-dimensional integrals. Møller and Torrisi (2005) provide simplifications for a few specific cases (e.g., the Neyman-Scott models). The case of Markov point processes as the parent process appears to be the most difficult for which to obtain any specific results.

We take a slightly different approach. Specifically, rather than deriving general results, we consider only the simplest of the Markov point processes. This allows us to derive an approximation to the K function of the resulting model, and to explore its use in fitting to data.

In Section 2 we describe the GNS model in more detail and work out an approximation to the K function for the model in terms of its parameters. This approximation is based on an approximation given in Isham (1984) for the second-order intensity of the Strauss process in terms of its first-order intensity. The final form of the K function depends on the dispersion function of the offspring points. We consider the cases of a uniform and a bivariate Gaussian density for the dispersion function. These models correspond to the Matérn cluster and modified Thomas processes if the parent process was Poisson. We can fit a specific Generalized Neyman-Scott model to data using least squares minimization between the theoretical and empirical K function.

For the rest of the paper, we focus on the case of a uniform dispersion function, that we call the Matérn GNS. In Section 3, we discuss and compare the form of the Matérn GNS K function with that of the Matérn cluster process. We fit the model to a rainforest dataset in Section 4, and obtain standard errors for the parameter estimates using the bootstrap. In Section 5 we present results of a simulation study in which the Matérn GNS is fit to point data simulated from Neyman-Scott and Generalized Neyman-Scott processes with a variety of parameter values. These results help us to understand better its use for modeling spatial point data. In this section we also describe results of our study into the accuracy of the Isham (1984) second-order intensity approximation. Section 6 contains a brief summary as well as a description of on-going work.

2. The Generalized Neyman-Scott Model and its K Function

We first specify the Generalized Neyman-Scott (GNS) model, and then in Section 2.1 derive an approximation to the Ripley's K-function (Ripley (1988)). The Neyman-Scott process consists of the set of clusters of offspring points, centered on an unobserved set of parent points. In the Neyman-Scott process, the parent points follow a Poisson process. In this paper, we use a Strauss process for the parent points.

The Strauss process is a special case of pairwise interaction processes. On $\mathcal{R}^2$, the Strauss process is specified by a non-negative parameter β_p and two other parameters γ_p, r_p through an interaction function ϕ where $\phi(r) = \gamma_p$ for $r \leq r_p$ and 0 otherwise. The parameter β_p controls the intensity of the process, with intensity increasing with β_p. The quantity r_p specifies the distance within which the parent points experience repulsion from other parent points. The strength of repulsion and hence the regularity of the process is determined by γ_p. Smaller values of γ_p corresponds to stronger repulsion and with $\gamma_p = 1$ (or equivalently $r_p = 0$), there is no interaction and the process is Poisson with intensity β_p. For such a process, the Papangelou conditional intensity $\lambda^*(X, \xi)$ of ξ given a point configuration X that does not contain ξ is defined to be $\prod_{x \in X} \beta_p \phi(|\xi - x|)$. We denote the intensity of this infinite Strauss process by $\lambda_p = \mathrm{E}[\lambda^*(X, 0)]$. See van Lieshout and Baddeley (1996) and Møller and Torrisi (2005) in this connection.

Each parent point produces an expected number μ_o of offspring points that are dispersed around the parent location according to a dispersion density function $k(\cdot)$ with a parameter σ_o that controls the dispersion of the offspring points. These parameters have an o subscript to signify that they are offspring parameters. Thus, if a parent point is located at $s \in \mathcal{R}^2$, its offspring points are scattered around s with intensity $Z_s(\xi) = \mu_o k(\xi - s)$. Dispersion functions include the uniform density on a disc and the bivariate Gaussian density. If the parent process were Poisson, these dispersion functions would correspond to the Matérn cluster and modified Thomas processes, respectively.

We note that Gibbs processes are stationary only when defined on an unbounded region, such as $\mathcal{R}^2$. Strauss processes can be defined on a bounded region $D \subset \mathcal{R}^2$, but the resulting process is not stationary. In applications, however, the observation region will necessarily be bounded. We take such data as being the observed portion of the infinite process. Alternatively, if D is rectangular, we can consider the region as a torus and define a stationary Strauss process on it. In our derivation, we take the parent process to be stationary,

The set of parameters for the GNS model is $\theta = (\lambda_p, \gamma_p, r_p, \mu_o, \sigma_o)$.

2.1. The GNS K function

For a stationary, istropic process, the reduced second moment function $K(h)$, for a distance h, is defined as the expected number of additional points, $N(x, h)$, within distance h of an arbitrarily chosen point x, divided by the intensity of the process, i.e.,

$$K(h) = \frac{\mathrm{E}[N(x,h)]}{\lambda}.$$

This K function is useful as a description of the clumpiness of a point process at various scales h, see e.g., Stoyan and Stoyan (1994) or Møller and Waagepetersen (2003).

In this section, we derive a closed form approximation to the K function for the GNS process. For $\xi \in \mathcal{R}^2$, let $Z(\xi)$ denote the random intensity at location ξ. Then

$$Z(\xi) = \sum_{s \in S} \mu_o k(\xi - s),$$

where S is the set of parent points, $k(\cdot)$ a probability density function representing the dispersion of the offspring points about the parent points, and μ_o the expected number of offspring points per parent. In the Matérn GNS, $k_1(s) = 1\{|s| \leq \sigma_o\}/\pi\sigma_o^2$.

Given Z, the GNS process is an inhomogeneous Poisson process. From Møller and Waagepetersen (2003) we have the pair correlation function between ξ and η as

$$g(\xi, \eta) = \mathrm{E}[Z(\xi)Z(\eta)]/\lambda(\xi)\lambda(\eta),$$

where $\lambda(\xi) = \mathrm{E}[Z(\xi)]$. In our case of stationarity and isotropy, λ does not depend on ξ and $g(\xi, \eta)$ depends only on $|\xi - \eta|$. Without loss of generality, fix ξ and let $\eta \in B(\xi, u)$, be the ball of radius u centered at ξ. Write g_ξ as the pair correlation function at ξ. Note that with stationarity, $g_\xi \equiv g$. Then $K(h) = \int_0^h 2\pi u g_\xi(u)\, du$, where

$$
\begin{aligned}
&\lambda^2 2\pi u g_\xi(u)\\
&\quad = \mathrm{E}\left[\int_{\eta\in B(\xi,u)}\left(\sum_{s\in S}\mu_o k(\xi-s)\times\sum_{s'\in S}\mu_o k(\eta-s')\right)d\eta\right]\\
&\quad = \lambda_p\mu_o^2\int_{\mathcal{R}^2}\int_{B(\xi,u)}k(\xi-s)k(\eta-s)\,d\eta\,ds\\
&\qquad + \mu_o^2\int_{\mathcal{R}^2}\int_{\mathcal{R}^2}\int_{B(\xi,u)}k(\xi-s)k(\eta-s')\lambda_{p,2}(|s-s'|)\,d\eta\,ds\,ds',\\
&\quad \approx \lambda_p\mu_o^2\int_{\mathcal{R}^2}\int_{B(\xi,u)}k(\xi-s)k(\eta-s)\,d\eta\,ds\\
&\qquad + \lambda_p^2\mu_o^2\int_{\mathcal{R}^2}\int_{\mathcal{R}^2}\int_{B(\xi,u)}k(\xi-s)k(\eta-s')\,d\eta\,ds\,ds'\\
&\qquad - \lambda_p^2\mu_o^2(1-\gamma_p)\int_{\mathcal{R}^2}\int_{\mathcal{R}^2}\int_{B(\xi,u)}k(\xi-s)k(\eta-s')1\{|s-s'|\le r_p\}\,d\eta\,ds\,ds'\\
&\quad \equiv I_1(u)+I_2(u)+I_3(u),
\end{aligned}
$$

where we have used the approximation in Isham (1984) for the stationary second-order intensity of the parent process

$$\lambda_{p,2}(|s-s'|)\approx\lambda_p^2-\lambda_p^2(1-\gamma_p)I(|s-s'|<r_p)\,. \tag{2.1}$$

Simplifying further, we have $I_1(u)=\lambda^2 f_k(u)/\lambda_p$, where $\lambda=\lambda_p\mu_o$ is the intensity of the GNS process, and $f_k(u)$ is the density function of the distance between two offspring points from the same parent when the dispersion function is given by $k(\cdot)$. Also,

$$
\begin{aligned}
I_2(u) &= \lambda^2\int_{B(\xi,u)}\int_{\mathcal{R}^2}k(\eta-s')\,ds'\,d\eta\int_{\mathcal{R}^2}k(\xi-s)\,ds\\
&= 2\pi u\lambda^2\,,\\
I_3(u) &= -\lambda^2(1-\gamma_p)\int_{\mathcal{R}^2}\left(\int_{\mathcal{R}^2}k(\eta-s')1\{\eta\in B(\xi,u)\}\,d\eta\right)\\
&\qquad\times\left(\int_{\mathcal{R}^2}k(\xi-s)1\{|s-s'|\le r_p\}\,ds\right)ds'\\
&= -\lambda^2(1-\gamma_p)\int_{\mathcal{R}^2}\frac{d}{du}\left[\int_{\mathcal{R}^2}k(\eta-s')1\{|\eta-\xi|\le u\}\,d\eta\right]\\
&\qquad\times\left(\int_{\mathcal{R}^2}k(\xi-s)1\{|s-s'|\le r_p\}\,ds\right)ds'\\
&= -\lambda^2(1-\gamma_p)\frac{d}{du}\int_{\mathcal{R}^2}\mathcal{V}_k[s',\mathcal{D}(\xi,u)]\mathcal{V}_k[\xi,\mathcal{D}(s',r_p)]\,ds',
\end{aligned}
$$

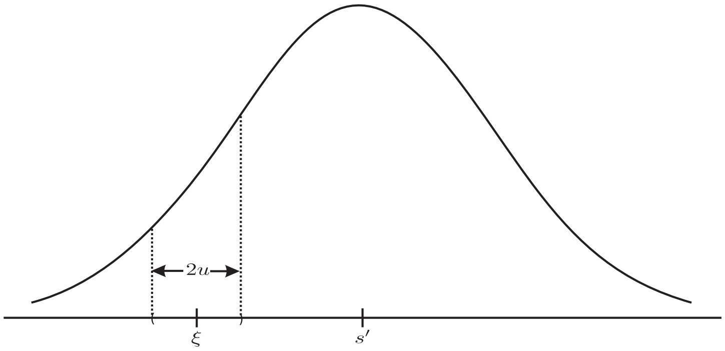

Figure 1. Diagram showing $\mathcal{V}_k[s', \mathcal{D}(\xi, u)]$ in the one-dimensional case. The Gaussian kernel is centered at s' and $\mathcal{V}_k[s', \mathcal{D}(\xi, u)]$ is the area under the kernel within the interval $(\xi - u, \xi + u)$.

where $\mathcal{V}_k(s', \mathcal{D}(\xi, u))$ is the volume under the kernel k, centered at s' within the disc $\mathcal{D}(\xi, u)$ of radius u, centered at ξ. A diagram in the one-dimensional case is shown in Figure 1.

So we have

$$K(h) \approx F_k(h)/\lambda_p + \pi h^2 - (1 - \gamma_p) \int_{\mathcal{R}^2} \mathcal{V}_k[s', \mathcal{D}(\xi, h)] \mathcal{V}_k[\xi, \mathcal{D}(s', r_p)]\, ds',$$

where $F_k(h)$ is the distribution function corresponding to the density function $f_k(h)$.

Getting an expression for $K(h)$ depends on obtaining expressions for $F_k(h)$ and $\mathcal{V}_k$. For example, for the Matérn GNS with the uniform dispersion function $k_1(s) = 1\{|s| \leq \sigma_o\}/\pi\sigma_o^2$, Stoyan, Kendall and Mecke (1995) has

$$F_{k_1}(h) = \begin{cases} 2 + [(8z^2 - 4)\cos^{-1} z - 2\sin^{-1} z & \\ \qquad + 4z\sqrt{(1 - z^2)^3} - 6z\sqrt{1 - z^2}]\frac{1}{\pi} & h \leq 2\sigma_o \\ 1 & h > 2\sigma_o, \end{cases}$$

where $z = h/2\sigma_o$.

The quantity $\mathcal{V}_{k_1}$ depends on the area of two overlapping discs. Specifically, $\mathcal{V}_{k_1}[s', \mathcal{D}(\xi, h)]$ is related to the area common to the disc of radius h centered at ξ and the disc of radius σ_o centered at s'. In particular, using elementary geometry,

$$\mathcal{V}_{k_1}[s', \mathcal{D}(\xi, h)] = \begin{cases} 1 & \text{if } |s'| \leq h - \sigma_o\,, \\ \frac{2\pi-\omega+\sin(\omega)}{2\pi} + \frac{h^2}{2\pi\sigma_o^2}(\phi - \sin(\phi)) & \text{if } h-\sigma_o < |s'| \leq \sqrt{h^2-\sigma_o^2}\,, \\ \frac{\theta-\sin(\theta)}{2\pi} + \frac{h^2}{2\pi\sigma_o^2}(\phi - \sin(\phi)) & \text{if } \sqrt{h^2-\sigma_o^2} < |s'| \leq h+\sigma_o\,, \\ 0 & \text{if } |s'| > h + \sigma_o\,, \end{cases}$$

where

$$\begin{aligned} \phi &= 2\cos^{-1}\left(\frac{h^2 + |s'|^2 - \sigma_o^2}{2h|s'|}\right)\,, \\ \theta &= 2\cos^{-1}\left(\frac{\sigma_o^2 + |s'|^2 - h^2}{2\sigma_o|s'|}\right)\,, \\ \omega &= 2\sin^{-1}\left(\frac{h}{\sigma_o}\sin\left(\frac{\phi}{2}\right)\right)\,. \end{aligned}$$

Using polar coordinates, a further simplification is

$$K(h) \approx \frac{F_k(h)}{\lambda_p} + \pi h^2 - 2\pi(1-\gamma_p)\int_0^{\sigma_o+\min(h,r_p)} s\mathcal{V}_k[s, \mathcal{D}(\xi, h)]\mathcal{V}_k[\xi, \mathcal{D}(s, r_p)]ds. \tag{2.2}$$

This formula involves a one-dimensional integral and hence is easy to compute.

For the bivariate Gaussian density, $k_2(s) = \exp(-|s|^2/2\sigma_o^2)/\sqrt{2\pi}\sigma_o$, we have, again from Stoyan, Kendall and Mecke (1995),

$$F_{k_2}(h) = -\exp\Big(-\frac{h^2}{4\sigma_o^2}\Big).$$

The quantity $\mathcal{V}_{k_2}[s', \mathcal{D}(\xi, h)]$ is somewhat more complicated: with $R = |s' - \xi|$,

$$\mathcal{V}_{k_2}[s', \mathcal{D}(\xi, h)] = \int_{R-h}^{R+h} \frac{2r}{\sqrt{2\pi}\sigma_o}\cos^{-1}\left(\frac{r^2 + R^2 - u^2}{2rR}\right)\exp\Big(-\frac{r^2}{2\sigma_o^2}\Big)dr.$$

3. Features of the GNS K Function

In this section we look at how the K function of the Matérn GNS depends on its parameters and compare its form with that of the Matérn cluster Neyman-Scott model.

First we look at the relationship between the K function in (2.2) and the parameters of the new point process, λ_p, γ_p, r_p, μ_o, and σ_o. To highlight the dependence of the K function on the parameters, we denote the K function by $K(h, \theta)$ in this section. Figure 2(A) shows a typical plot of a $K(h, \theta)$ against h for the new point process. In this example the parameters for the process are taken to be $(\lambda_p, \gamma_p, r_p, \mu_o, \sigma_o) = (3, 0.25, 0.3, 2, 0.05)$. Note that $K(h, \theta)$ is non-decreasing by definition. The function increases rapidly from 0 to $2\sigma_o = 0.1$, corresponding to the offspring points in the cluster from the same parent. The curve becomes flat in the interval $(2\sigma_o, r_p) = (0.1, 0.3)$, accounting for the fact that there are relatively few observations between clusters. Then it rises again for $h > r_p = 0.3$, as the observations in other clusters are now taken into account. From (2.2), $K(h, \theta)$ depends on λ_p only through the term $F_k(h)/\lambda_p$. Thus $K(h, \theta)$ is inversely proportion to λ_p, in line with Figures 2(C) and 2(D). From Figures 2(E) and 2(F), we observe that the degree of "flatness" in the interval $(2\sigma_o, r_p)$ is governed by γ_p, the inhibition parameter of the Strauss process. Figure 2(E) is a K function for a hard core process ($\gamma_p = 0$), where the interval $(2\sigma_o, r_p)$ is completely flat. Finally, Figures 2(G) to 2(J) indicate the relationship between the locations of the two increasing intervals and the parameters r_p, σ_o. As each parameter controls a unique feature of the K function, the parameters are identifiable from the K function. Note that in the above discussions it is assumed that $r_p > 2\sigma_o$, so that the interaction range of the parent points is larger than the dispersion radius of the offspring points. With this assumption, the K function has a fairly flat portion in the middle. In general, if $r_p < 2\sigma_o$, the dispersion of the offspring points can mask the interaction between the parent points, so that the K function cannot be differentiated from that of a regular Neyman-Scott process, and the parameters of the Generalized Neyman-Scott process become unidentifiable. This assumption is reasonable, since the model would be used only when the data exhibit a cluster pattern with regularity.

Next we compare the K function of the Neyman-Scott process and the new process. The K function of a Neyman-Scott process is

$$K(h, \theta) = \pi h^2 + \frac{F_k(h)}{\lambda_p}\,, \tag{3.1}$$

where λ_p is the intensity of the Poisson random process for the parent and $F_k(h)$ is the distribution function of the distance between two events in the same cluster (e.g., Cressie (1993)). The K function of the GNS model has an additional term corresponding to the third term on the right-hand side of (2.2) accounting for the regularity among clusters. Figure 2(B) shows the K function for a Neyman-Scott process. Comparing with Figure 2(A), the K function of Neyman-Scott process starts to behave like the parabola πh^2 for $h > 2\sigma_o$, without the relatively flat interval observed in the K function of the new process that reflects the regularity of the cluster. Therefore, by adjusting the inhibition radius r_p and degree γ_p, the new process gives more flexibility in modeling a cluster point pattern that exhibits regularity behavior.

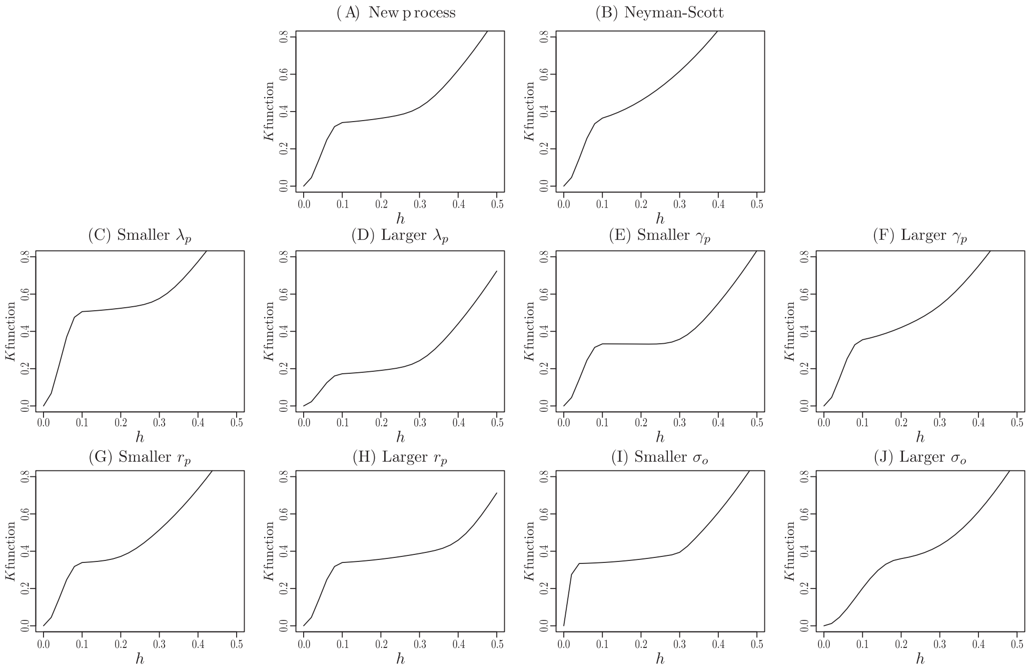

Figure 2. K functions for different models. (A): A typical K function for the new process. (B): A typical K function for a Neyman-Scott process. (C): K function for the new process with a smaller λ_p. (D): K function for the new process with a larger λ_p. (E): K function for the new process with a smaller γ_p. (F): K function for the new process with a larger γ_p. (G): K function for the new process with a smaller r_p. (H): K function for the new process with a larger r_p. (I): K function for the new process with a smaller σ_o. (J): K function for the new process with a larger σ_o.

4. Fitting to Rainforest Data

In this section we fit the new point process model to a rainforest data set consisting of the locations of Acacia Melanoceras trees. According to Seigler and Ebinger (1995), the Acacia Melanoceras has the most restricted range of all Ant-Acacias trees. Its short range could be due to the softness of its seeds that allow it to be digested instead of being dispersed by birds. The Acacia Melanoceras is sensitive to disturbances in its habitat that are “any more catastrophic than infrequent logging”, and rarely are more than two individuals found per acre in forest communities (Janzen (1974)).

The combination of restricted range of offspring and the sensitivity of the trees to disturbances suggest the possibility of the locations of these trees displaying a mix of regularity and clustering. This is borne out in Figure 3(A) which shows the locations of Acacia Melanoceras trees in a region in Barro Colarado

Island in the Panama. Note the presence of clusters of trees, regularly spaced over the region. This data set is part of a larger data set consisting of the locations of many tree species, collected when the forest there was surveyed. Several of the tree species in this data set were analyzed in Waagepetersen (2007) and Waagepetersen and Guan (2009).

We fit the Matérn Generalized Neyman-Scott process to the Acacia Melanoceras data set. To carry out model fitting, we adopted the least squares or minimum contrast method (Diggle (2003)) that matches the theoretical and empirical K functions. First, the empirical K function is estimated by

$$\hat{K}(h) = \frac{1}{\hat{\lambda} n} \sum_{i=1}^{n} \sum_{j \neq i}^{n} w^{-1}_{\mathbf{x_i},\mathbf{x_j}} 1(h_{\mathbf{x_i},\mathbf{x_j}} < h) ,$$

where $\hat{\lambda} = n/|A|$, $|A|$ is the area of the study region, $h_{\mathbf{x_i},\mathbf{x_j}}$ is the distance between the i-th and j-th points, and $w_{\mathbf{x_i},\mathbf{x_j}}$ is the corresponding weight for the correction of edge effect. Various weight functions $w_{\mathbf{x_i},\mathbf{x_j}}$ have been proposed in the literature (e.g., Diggle (2003)). In this work we choose the translation correction introduced by Ohser (1983). Then the least square estimator $\hat{\theta}$ is obtained by minimizing

$$D(\theta) = \int_{o}^{h_o} w(h) \left(\{\hat{K}(h)\}^c - \{K(h,\theta)\}^c \right)^2 dh , \tag{4.1}$$

where c and h_o are tuning constants and $w(h)$ a weighting function Diggle (2003). Although $K(h,\theta)$ is independent of the offspring intensity μ_o, we can compute the estimate

$$\hat{\mu}_o = \frac{n}{|A|\hat{\lambda}_p}$$

using the relation $\lambda = \lambda_p \mu_o$ and $\hat{\lambda} = n/|A|$. Diggle (2003) suggests that $w(h) = 1$ together with $c = 0.25$ or 0.5, and used $w(h) = 1, c = 0.5$, and $h_o = 80$m.

Figure 3(B) shows the empirical K function (solid line), together with the estimated K functions obtained by minimizing (4.1) using the theoretical K functions of the Matérn cluster process (dotted line), and of the Matérn GNS (dashed line). Notice that the empirical K function is relatively flat at around $h = 20$, indicating the possibility of regularity among clusters. As discussed in the last section, the K function of Neyman-Scott process starts to behave like the parabola πh^2 for $h > 2\sigma_o$, thus the best fitting theoretical K function is not able to capture the flat interval of the empirical K function. On the other hand, it can be seen that the Matérn Generalized Neyman-Scott model provides a better fit, with the fitted K function following more closely the flat portion of the empirical K function.

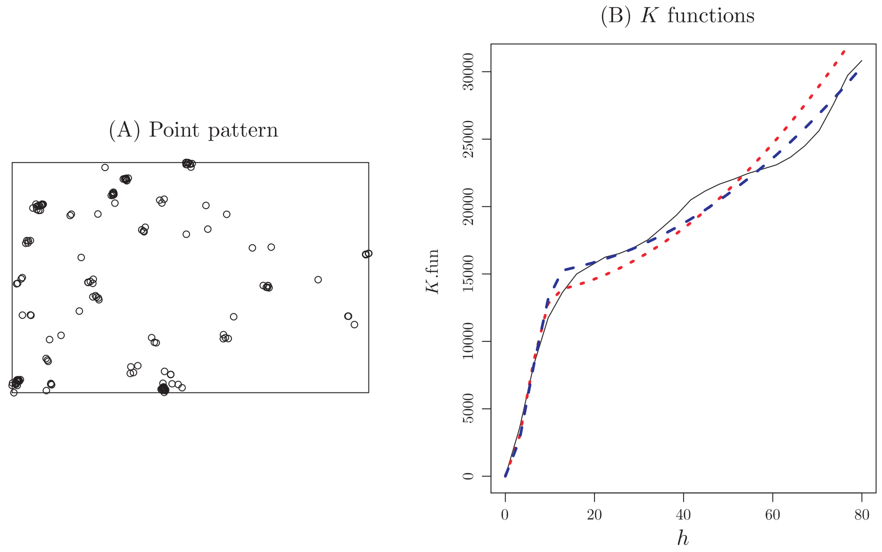

Figure 3. (A): Point pattern for the Acacia Melanoceras data set. (B): The empirical (solid line) and estimated K functions, using the Neyman-Scott model (dotted line) and the Generalized Neyman-Scott model (dashed line).

To assess the variability of the parameter estimates, we used a bootstrap approach. There are various bootstrap methods available for spatial data, including subsampling, the block bootstrap and the marked point bootstrap (Künsch (1989); Politis, Romano, and Wolf (1999); Loh and Stein (2004)). We chose to use the marked bootstrap method and expect the other bootstrap methods to yield similar results. The method works as follows. First the study region is divided into N subregions, with each subregion containing n_i points, $i = 1, \ldots, N$. Let $\mathbf{x}$ be the j-th ($j = 1, \ldots, n_i$) point of the subregion i, we assign to it the mark

$$m_{i,j}(h) = \sum_{\mathbf{y}\neq\mathbf{x}} 1(|\mathbf{x}-\mathbf{y}| < h) w(\mathbf{x},\mathbf{y}) \, .$$

Note that

$$\hat{K}(h) = \frac{|A|}{(\sum_{i=1}^{N} n_i)^2} \sum_{i=1}^{N} \sum_{j=1}^{n_i} m_{i,j}(h) \, .$$

Now the subregion i ($i = 1, \ldots, N$) is associated with the sum of the marks $M_i(h) = \sum_{j=1}^{n_i} m_{i,j}(h)$. We can sample the quantities $(M_i, n_i), i = 1, \ldots, N$, with replacement to obtain a bootstrap sample with marks given by $(\tilde{M}_j, \tilde{n}_j), j =$

Table 1. Fit to rainforest data. Least square estimates $\hat{\theta}$ of θ , together with the mean and standard deviations of the bootstrap estimates, and the lower and upper confidence limits of the 95% bootstrap confidence interval of θ.

	λ_p	γ_p	r_p	μ_o	σ_o
$\hat{\theta}$	6.685e-05	0.7576	75.64	8.379	6.764
Mean	6.927e-05	0.7544	78.44	8.730	6.679
Std	2.147e-05	0.05405	13.35	2.238	0.4339
Lower limit	1.767e-05	0.6606	56.10	2.766	6.046
Upper limit	1.033e-04	0.8751	101.00	12.030	7.742

$1, \ldots, N$. The bootstrap estimates of the K function is given by

$$\tilde{K}(h) = \frac{|A|}{(\sum_{j=1}^{N} \tilde{n}_j)^2} \sum_{j=1}^{N} \tilde{M}_j(h) \, .$$

The above procedure can be repeated B times so that we have B bootstrap estimates of K, $\tilde{K}^b(h), b = 1, \ldots, B$.

In the Acacia Melanoceras example, we take $N = 6$ and $B = 500$, i.e. we divde the observation region into two rows of three blocks for the resampling. We chose this so as to provide a balance between the size of the blocks, to retain the underlying dependence, and the number of blocks, so as not to underestimate the variability.

For each $\tilde{K}^b(h)$, a bootstrap estimate $\tilde{\theta}^b$ is obtained by least squares estimation. The bootstrap estimates of the parameters of the model, $(\tilde{\theta}^1, \tilde{\theta}^2, \ldots, \tilde{\theta}^B)$ can then be used to obtain a confidence interval for θ. A $100(1-\alpha)\%$ confidence interval for $K(r)$, called the basic bootstrap interval by Davison and Hinkley (1997), is

$$\left[2\hat{\theta} - \tilde{\theta}_{(B+1)(1-\alpha/2)}, 2\hat{\theta} - \tilde{\theta}_{(B+1)(\alpha/2)}\right] \, ,$$

where $\tilde{\theta}_{(B+1)(1-\alpha/2)}$ and $\tilde{\theta}_{(B+1)(\alpha/2)}$ are the $(B+1)(1-\alpha/2)$th and the $(B+1)(\alpha/2)$th ordered values of $(\tilde{\theta}^1, \ldots, \tilde{\theta}^B)$. Figure 4 shows histograms of the resulting bootstrap parameter estimates.

Table 1 shows the least square estimates, means, and standard deviations of the bootstrap estimates, and the lower and upper confidence limits of the 95% bootstrap confidence interval of θ. The values may be used to test for regularity in the parent process. Recall that the smaller the value of γ_p, the higher the regularity of the parent process, and the parent process reduces to the Poisson process when $\gamma_p = 1$. Thus a test for regularity can be performed by testing

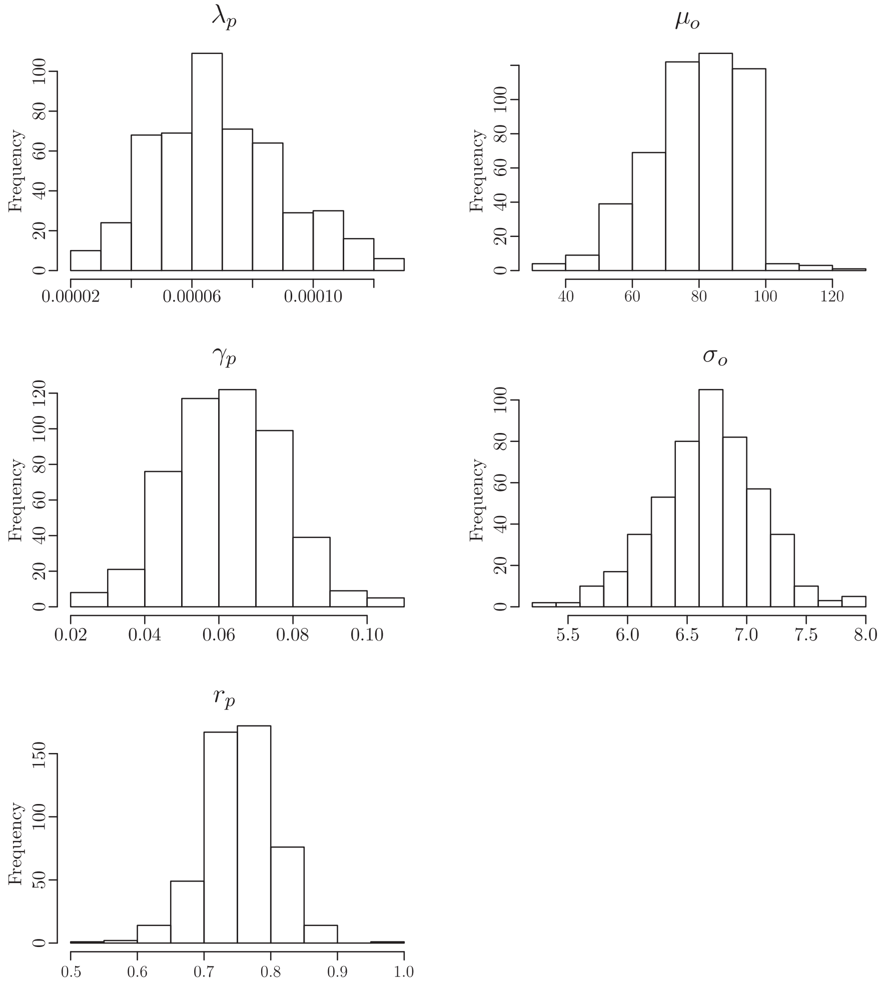

Figure 4. Histograms of the bootstrap parameter estimates from fitting the Matérn GNS to the Acacia Melanoceras data set.

whether $\hat{\gamma}_p$ lies significantly below one. In this example, the 95% confidence interval for γ_p does not cover 1, indicating a significant regularity in the parent process.

5. Simulation Study

We present the results of a simulation study to examine the performance of our procedure for fitting the Generalized Neyman-Scott model of Section 2 to data. We also present some results on our study into the accuracy of the Isham second-order intensity approximation used in our derivation of the closed form expression for the K function of the Generalized Neyman-Scott model.

5.1. Fitting the generalized Neyman-Scott model

Specifically, we used the Matérn cluster model and the Matérn Generalized Neyman-Scott model to produce simulated data, and fit a Generalized Neyman-Scott model to each set of data. For each model, we used several sets of parameter values and for each set we simulated 500 realizations on a 10×10 square. Thus for each specific model and set of parameter values, we have 500 estimates of $(\lambda_p, \gamma_p, r_p, \mu_o, \sigma_o)$.

We used the *rStrauss* function in the *spatstat* R package, Baddeley and Turner (2005), to generate realizations on a much larger region (specifically 50 x 50) and extract the points in the middle 10×10 region. Besides the observation window, the *rStrauss* function requires user-specified values for β_p, r_p, and γ_p, and produces realizations using perfect simulation (Baddeley and Turner (2005)). The quantity β_p is not the intensity, unless $\gamma_p = 1$. However, we can empirically obtain λ_p by simulating 10,000 realizations and noting the number of points produced in the central 10×10 region. The true values of λ_p indicated in Tables 2 to 4 refer to values obtained in this manner.

Table 2 shows the mean and standard deviations of the estimates obtained by fitting the Generalized Neyman-Scott model to data simulated from a Generalized Neyman-Scott model (left column) and from a Matérn cluster model (right column). Note that the Matérn cluster model is a special case of our Generalized Neyman-Scott model with $\gamma_p \equiv 1$ and r_p undefined. From Table 2, we find that the estimates are close to the true values in all six cases considered. In particular, the expected number of offspring μ_o and the disc radius of the offspring process σ_o have small bias and standard errors. For the case of Matérn cluster process we find that the standard errors of r_p are very large. This is not surprising since r_p is not well-defined here. More importantly, we find that the estimates of γ_p are close to 1, suggesting that the minimum contrast method with the Generalized Neyman-Scott model was able to correctly identify the Matérn cluster model.

We also fit the Matérn cluster model to the data simulated from the two models. The results are shown in Table 3. Comparing the right columns of Tables 2 and 3, we see that when the data were simulated from a Matérn cluster model, the biases of the estimates of λ_p, μ_o, and σ_o obtained from a Matérn cluster fit are slightly smaller than the ones obtained by fitting the Generalized Neyman-Scott model. It may seem surprising that the standard errors of the estimates of λ_p, μ_o, and σ_o are larger than the ones obtained by fitting the Generalized Neyman-Scott model. However, note that the Generalized Neyman-Scott model contains the Matérn cluster model and is thus not a wrong model. Fitting the Matérn cluster model to the data is like fitting the Generalized Neyman-Scott

Table 2. Mean and standard deviations of the estimates obtained from 500 replications of the Generalized Neyman-Scott model fit. For the left column, the data were generated from the Generalized Neyman-Scott model. For the right column, the data were generated from Matérn cluster model that corresponds to the Generalized Neyman-Scott model with $\gamma_p = 1$ and r_p undefined.

	GNS fit to GNS					GNS fit to Matérn				
	λ_p	γ_p	r_p	μ_o	σ_o	λ_p	γ_p	r_p	μ_o	σ_o
True	1.982	0.60	0.40	5	0.100	3.00	1.00	-	5	0.100
Mean	1.958	0.5546	0.4271	5.083	0.1003	2.894	0.9471	0.8915	5.216	0.1032
Std	0.1861	0.2775	0.1876	0.4234	0.003898	0.2871	0.1093	2.929	0.432	0.006235
True	2.360	0.80	0.40	4	0.050	2.50	1.00	-	4	0.100
Mean	2.341	0.7894	0.4468	4.038	0.05038	2.389	0.9483	0.7422	4.215	0.1040
Std	0.1727	0.1430	0.2308	0.2246	0.001595	0.2635	0.08071	2.024	0.4052	0.007020
True	1.731	0.40	0.40	5	0.050	2.50	1.00	-	6	0.050
Mean	1.706	0.4526	0.3788	5.032	0.05038	2.423	0.9445	0.7169	6.193	0.05153
Std	0.1215	0.1260	0.08234	0.2591	0.001330	0.2308	0.09058	1.634	0.4436	0.003027

model with γ_p constrained to 1. Having the additional parameters γ_p and r_p in the Generalized Neyman-Scott model allows the variability in the K function to be spread over these additional parameters as well.

On the other hand, we find that the biases and errors are significantly increased for the estimates of λ_p, μ_o, and σ_o when the (incorrect) Matérn cluster model is fit to data simulated from a Generalized Neyman-Scott model.

We also did a limited study on the behavior of the estimates from fitting the Generalized Neyman-Scott model under model misspecification. Specifically, the data were simulated from a different Generalized Neyman-Scott process, the Thomas GNS in which the parent process is the Strauss process but the offspring points are distributed around the parent points via a symmetrical Gaussian distribution with standard deviation σ. Table 4 shows the mean and the standard errors of the estimates from fitting the Matérn GNS. We observe that although the parameters γ_p and r_p tend to be overestimated, they are fairly close to their true values. The estimates of λ_p and the expected number of offspring μ_o have relatively small bias. Also, the estimate of the offspring radius σ_o is approximately two times σ, roughly in line with the rule-of-thumb that most of the observations fall within two standard deviations of the mean.

5.2. Isham's second-order intensity approximation

In the derivation of the K-function, the second-order intensity approximation of Strauss process (2.1) was used. To evaluate the accuracy of this approximation, we considered six sets of parameter values for the Strauss process and,

Table 3. Mean and standard deviations of the estimates obtained from 500 replications of the Matérn cluster fit. For the left column, the data were generated from the Generalized Neyman-Scott model. For the right column, the data were generated from the Matérn cluster model.

	Matérn fit to GNS					Matérn fit to Matérn		
	λ_p	γ_p	r_p	μ_o	σ_o	λ_p	μ_o	σ_o
True	1.982	0.60	0.40	5	0.10	3.00	5	0.10
Mean	2.615	-	-	3.853	0.08383	3.065	4.960	0.09963
Std	0.3123	-	-	0.4012	0.005467	0.3679	0.5669	0.008279
True	2.360	0.80	0.40	4	0.05	2.50	4	0.10
Mean	2.734	-	-	3.489	0.04503	2.543	4.000	0.1005
Std	0.2808	-	-	0.3021	0.002859	0.3378	0.4818	0.009042
True	1.731	0.40	0.40	5	0.05	2.50	6	0.05
Mean	2.329	-	-	3.736	0.04038	2.541	5.960	0.04995
Std	0.2499	-	-	0.2981	0.002168	0.2719	0 .5571	0.003772

Table 4. Mean and standard deviations of the estimates obtained from 500 replications of the Generalized Neyman-Scott model fit. The data were generated from the Thomas model, a Neyman-Scott model with the offspring distribution symmetric Gaussian with standard deviation σ. The number of offspring per cluster was Poisson with mean μ.

	Generalized Neyman-Scott (Thomas)				
	λ_p	γ_p	r_p	μ	σ
True	1.982	0.60	0.40	5.00	0.05
	$\hat{\lambda}_p$	$\hat{\gamma}_p$	$\hat{r}_p$	$\hat{\mu}$	$\hat{\sigma}_o$
Mean	2.063	0.6934	0.4424	4.801	0.09226
Std	0.1721	0.1755	0.2064	0.3228	0.003353
	λ_p	γ_p	r_p	μ	σ
True	2.360	0.80	0.40	4.00	0.025
	$\hat{\lambda}_p$	$\hat{\gamma}_p$	$\hat{r}_p$	$\hat{\mu}$	$\hat{\sigma}_o$
Mean	2.365	0.8273	0.4667	3.975	0.04754
Std	0.1871	0.1007	0.2666	0.2323	0.001915
	λ_p	γ_p	r_p	μ	σ
True	1.731	0.40	0.40	5.00	0.025
	$\hat{\lambda}_p$	$\hat{\gamma}_p$	$\hat{r}_p$	$\hat{\mu}$	$\hat{\sigma}_o$
Mean	1.733	0.4991	0.3952	4.959	0.04737
Std	0.1202	0.1253	0.09671	0.2529	0.001454

for each model, we compared the approximation (2.1) to the second-order intensity obtained empirically from 1,000 simulated realizations. The parameter values are given in Table 5. We used a 10×10 window for Models 1 to 5 and a $2,000 \times 2,000$ window for Model 6. The parameter values for Model 6 correspond

Table 5. Empirical and Isham's approximate theoretical first-order intensity for various models of Strauss process: λ_p is the empirical first order intensity obtained from 10,000 realizations of Strauss process specified by the parameter (β_p, γ_p, r_p), $\tilde{\lambda}_p$ is the approximate intensity given in (5.1).

Model	1	2	3	4	5	6
β_p	3	3	3	3	3	9e-5
γ_p	0.1	0.5	0.9	0.5	0.5	0.7576
r_p	0.5	0.5	0.5	0.2	0.8	75.64
λ_p	1.12	1.56	2.47	2.55	0.993	6.69e-5
$\tilde{\lambda}_p$	-3.36	-0.534	2.29	2.43	-6.05	5.47e-5

to the estimated values in Section 4. For each such model, we also compared the corresponding K function approximated by (2.2) and its empirical counterpart. In evaluating the approximation (2.1), the true first-order intensity was obtained from simulation as described in Section 5.1. The empirical second-order intensity function and the empirical K function were obtained by the commands *pcf* and *Kest*, respectively, in the *spatstat* R package (Baddeley and Turner (2005)).

Figure 5 illustrates the comparisons for the models we considered. For the second-order intensity function, the accuracy of the approximation is higher for higher γ_p and smaller r_p. It is encouraging to see that the empirical K function and the approximate K function are reasonably close to each other in most cases, except for a slight discrepancy for large h when γ_p is small and r_p is large. We note that there is a complex relationship among the three parameters λ_p, γ_p, and r_p of the Strauss process. For example, for fixed λ_p and γ_p, the value of r_p cannot be too large. Also, for some combinations of r_p and γ_p, the intensity λ_p needs to be small. Thus it is difficult to separate out the effects of the individual parameters on the accuracy of the approximation.

Besides Models 1 to 6, we also looked at several other additional sets of parameter values, and we found that Model 5 in Figure 5 has about the worst accuracy among these. For instance, with parameters $(\lambda_p, \gamma_p, r_p)$ equal to $(2e^{-4}, 0.2, 40)$ and $(1.7e^{-5}, 0.1, 120)$, the accuracy of approximation of the K function and the second order intensity is similar to that of Model 1 in Figure 5.

For completeness we also looked at Isham's first-order intensity approximation (Isham (1984)), that relates the first-order intensity λ_p to the parameter β_p:

$$\lambda_p \approx \beta_p(1 - (1 - \gamma_p)\pi r_p^2 \beta_p)\,. \tag{5.1}$$

Table 5 presents the approximation for λ_p from (5.1) for the models studied in Figure 5. It can be seen that the approximation can be very poor, especially when γ_p is small and r_p is large. Note that our approximation to the K function is expressed in terms of λ_p but not β_p. Thus it only involves the more reliable

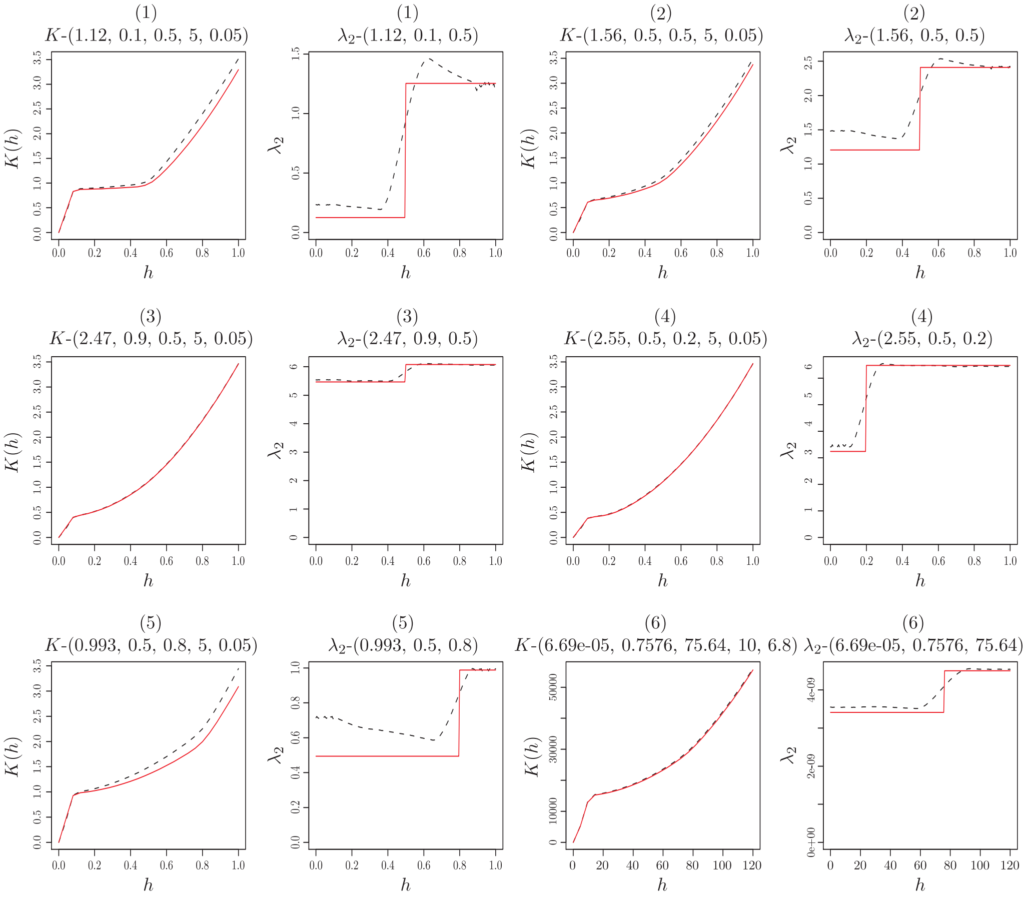

Figure 5. Empirical and Isham's approximate theoretical second-order intensity for various models of the Strauss process, and empirical and approximate theoretical K function for various models of the GNS process with the corresponding Strauss parent process. For the second-order intensity, the parameter vector $(\lambda_p, \gamma_p, r_p)$ is shown in the title. For the K function, the parameter vector $(\lambda_p, \gamma_p, r_p, \mu_o, \sigma_o)$ is shown in the title. In each graph, the solid line is the theoretical approximation and the dashed line is the empirical function.

second-order approximation (2.1), but not (5.1). Therefore (2.2) provides a good approximation to the K function of a Generalized Neyman-Scott process.

6. Summary

This paper introduced a new point process model, the Generalized Neyman-Scott model that is an extension of the Neyman-Scott process from Poisson parents to Strauss parents. This point process model allows for regular behavior among clusters of points. We found that this model to be more appropriate for the Acacia Melanoceras tree data than the usual Neyman-Scott model. Our simulation studies showed that fitting the Generalized Neyman-Scott model using a minimum contrast method based on the K function allows us to distinguish between the Matérn cluster model and the Matérn Generalized Neyman-Scott model.

The model introduced here can be easily generalized by specifying other models for the parent process. However, the resulting K function may involve higher dimensional integrals, and thus may be inexpressible in closed form. This is a focus of on-going research.

Acknowledgements

We would like to thank the associate editor and the two referees for their helpful suggestions. Research supported in part by the Direct Grant for Research, CUHK, C001-2060423.

References

Baddeley, A. and Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *J. Statist. Software* **12**, 1-42.

Cressie, N. A. C.(1993). *Statistics for Spatial Data.* Wiley, New York.

Davison, A. C. and Hinkley D. V. (1997). *Bootstrap Methods and their Applications.* Cambridge University Press, Cambridge.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns.* 2nd edition. Arnold, London.

Isham, V. (1984). Multitype Markov point process: some applications. *Proc. Roy. Soc. London Ser. A* **391**, 39-53.

Janzen, D. H. (1974). *Swollen-Thorn Acacias of Central America, Smithsonian Contributions to Botany*, Number 13. Smithsonian Institution Scholarly Press, Washington.

Künsch H. R. (1989). The Jackknife and the Bootstrap for general stationary observations. *Ann. Statist.* **17**, 1217-1241.

Loh, J.M. and Stein, M.L. (2004). Bootstrapping a spatial point process. *Statist. Sinica* **14,** 69–101.

Møller, J. and Torrisi, G.L. (2005). Generalized shot noise Cox processes. *Adv. Appl. Probab.* **37**, 48-74.

Møller, J. and Waagepetersen, R. (2003). *Statistical Inference and Simulation for Spatial Point Processes.* Chapman and Hall/CRC, Boca Raton.

Neyman, J. and Scott, E. L.(1958). Statistical approach to problems of cosmology. *J. Roy. Statist. Soc. Ser. B* **20,** 1-29.

Ohser, J. (1983). On estimators for the reduced second moment measure of point processes. *Mathematische Operationsforschung und Statistik* **14**, 63-71.

Politis, D. N., Romano, J. P. and Wolf, M (1999). *Subsampling.* Springer, Berlin.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes.* Wiley, New York.

Ripley, B. D. and Kelly, F. P. (1977). Markov point processes. *J. London Math. Soc.* **15**, 188-192.

Seigler, D. S. and Ebinger, J. E. (1995). Taxonomic Revision of the Ant-Acacias (Fabaceae, Mimosoideae, Acacia, Series Gummiferae) of the New World. *Ann. Missouri Botanical Garden* **82**, 117-138.

Stoyan, D., Kendall, W. S. and Mecke, J. (1995). *Topics in Stochastic Processes.* 2nd edition. John Wiley, New York.

Stoyan, D. and Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields.* John Wiley, New York.

Strauss, D. J. (1975). A model for clustering. *Biometrika* **63,** 467-475.

Taylor, C. C., Dryden, I. L. and Farnoosh, R. (2001). The K-function for nearly regular point processes. *Biometrics* **57,** 224-231.

van Lieshout, M. N. M. and Baddeley, A. J. (1996). A nonparametric measure of spatial interaction in point patterns. *Statist. Neerlandica* **50**, 344-361.

Waagepetersen, R. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics* **63**, 252-258.

Waagepetersen, R. and Guan, Y. (2009). Two-step estimation of inhomogeneous spatial point processes. *J. Roy. Statist. Soc. Ser. B* **71**, 685-702.

Department of Statistics, Room 110, Lady Shaw Building, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.

E-mail: cyyau@sta.cuhk.edu.hk

AT&T Labs-Research, 180 Park Ave Florham Park, NJ 07932, USA.

E-mail: loh@research.att.com

(Received October 2010; accepted September 2011)