LOG GAUSSIAN COX PROCESSES: A STATISTICAL MODEL FOR ANALYZING STAND STRUCTURAL HETEROGENEITY IN FORESTRY

Jesper Møller^{1}, Anne Randi Syversveen^{2} and Rasmus Plenge Waagepetersen^{3}

[1]*Aalborg University, Department of Mathematics, Fredrik Bajers Vej 7E, DK-9220 Aalborg Ø, Denmark (jm@math.auc.dk).* [2]*The Norwegian University of Science and Technology, Department of Mathematical Sciences, N-7034 Trondheim, Norway (annerand@math.ntnu.no).* [3]*University of Aarhus, Departments of Mathematical Sciences, DK-8000 Aarhus C, Denmark (rasmus@mi.aau.dk).*

Abstract: The appealing properties of planar Cox processes directed by a log Gaussian intensity process have recently been investigated in a longer joint paper by the authors. The purpose of this contribution is to illustrate this by an example of application in forestry.

Keywords: clustering of trees; environmental heterogeneity; estimation and model checking; point process; prediction; summary statistics.

1 Introduction and data

Log Gaussian Cox processes provide a rich class of models for clustered point patterns and they are appealing from a theoretical as well as an applied point of view. This contribution surveys some of the ideas and results in Møller *et al.* (1996) on log Gaussian Cox processes by considering an example of application in foresty.

The data consists of 126 Scots pine saplings in a square plot of $10 \times 10m^2$. This is shown in Figure 1 a) where the square plot is normalized to the unit square. The pine forest has grown naturally in Eastern Finland and the data have previously been analyzed by Penttinen *et al.* (1992) and Stoyan and Stoyan (1994), who both fitted a Matérn cluster process to the data. The fit in both papers seems quite good, but one may object to that the same summary statistic has been used for estimation and model check.

In the following we show how a nonparametric statistical analysis of such a point pattern is typically performed and how this leads to proposing and analyzing a parametric model for a log Gaussian Cox process for the data. It is concluded that this model gives a better fit than using a Mátern cluster process. Finally, it is shown how the unobserved environmental heterogeneity in the soil may be investigated through an empirical Bayesian analysis.

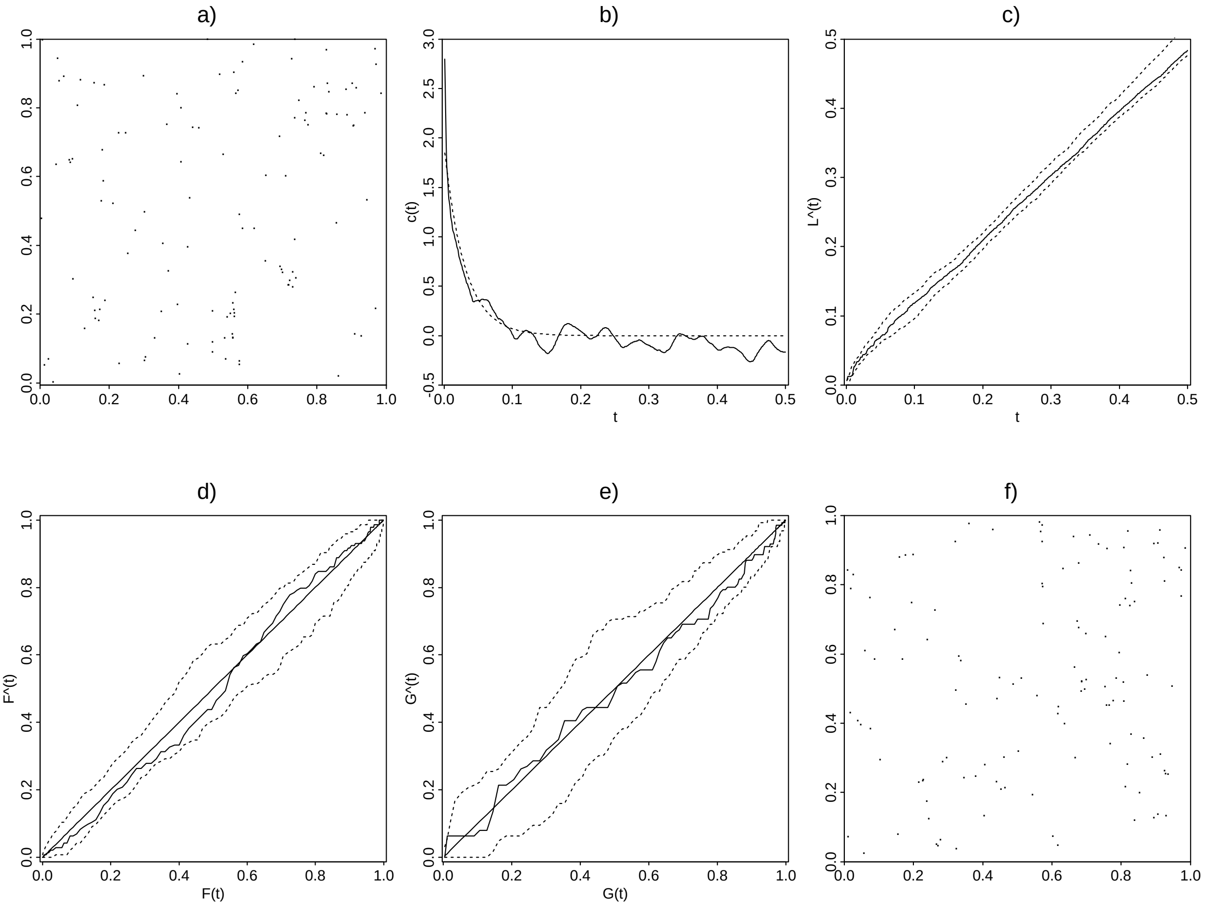

Figure 1: a) Pine data. b)-e) Different summary statistics (see the text for explanations). f) Simulated pattern under the estimated log Gaussian Cox process.

2 Nonparametric analysis of the Scots pine saplings

We consider the point pattern in Figure 1 a) as a realization of a planar point process X observed within the square plot (normalized to a unit square in Figure 1). Briefly, this means that X is a random countable set of points in the plane. The distribution of X is assumed to be invariant under translations and rotations in the plane (no derivation from this assumption seems present in the data possibly because of the small size of the observation window).

The *intensity* ρ of X, i.e. the mean number of trees per unit area (here $10\times 10m^2$), is naturally estimated by $\hat{\rho} = 126$. Another informative summary statistic is the *pair correlation function* g; heuristically, $\rho^2 g(||s_1 - s_2||)$ is the probability of having a tree in each of two infinitesimally small regions of areas ds_1 and ds_2 located at the points s_1 and s_2, respectively. The logarithm of a nonparametric estimate $\hat{g}$ of the pair correlation function is plotted in Figure 1 b) (solid line). In the case of complete spatial randomness of the trees, $\log g = 0$; the curve in Fig. 1 b) clearly shows a derivation from this and it in fact indicates a clustering of the trees.

3 A parametric log Gaussian Cox process model for the Scots pine saplings

3.1 Definition of log Gaussian Cox processes

Cox processes provide flexible and statistically tractable models for aggregated point patterns such as in Figure 1 a) where the clustering is due to a *stochastic environmental heterogeneity* in the soil. In a *log Gaussian Cox process* the heterogeneity is modelled by $\Lambda = \exp(Y)$ where Y is a planar Gaussian process, i.e. the joint distribution of any finite vector $(Y(s_1), \dots, Y(s_n))$ is Gaussian. Further, conditional on Λ, the distribution of X is a Poisson process with intensity function Λ; this means, losely speaking, that apart from what may be explained by the environmental heterogeneity there is no interaction between the trees, and if Λ is known then $\Lambda(s)ds$ is the probability of having a tree in the infinitesimally small region ds. Finally, by assumption of stationarity and isotropy, the distribution of Y and hence X is specified by the mean $\mu = EY(s)$ and the covariance function $c(r) = Cov(Y(s_1), Y(s_2))$ where $r = ||s_1 - s_2||$ is the distance between the points s_1 and s_2.

As shown in Møller *et al.* (1996) there is a one-to-one correspondence between the first and second order properties of X respective Y, since

$$\rho = \exp(\mu + \sigma^2)\,, \quad g(r) = \exp(c(r)), \tag{1}$$

where $\sigma^2 = c(0) = Var(Y(s))$ is the variance of the Gaussian process. Consequently, *the intensity and the pair correlation function completely determine the distribution* of both trees and the environmental heterogeneity. Further theoretical properties are investigated in Møller *et al.* (1996) and this enables us to construct summary statistics and estimation procedures as exemplified in Subsections 3.2 and 3.3.

3.2 Estimation

Since the distribution of a log Gaussian Cox process is completely determined by the intensity and the pair correlation function, we suggest to base the statistical estimation on (1) using the nonparametric estimates of the intensity and the pair correlation function.

The shape of the nonparametric estimate $\hat{c}(r) = \log(\hat{g}(r))$ of the covariance shown in Figure 1 b) (solid line) suggests to use the exponential covariance function $c(r) = \sigma^2 \exp(-r/\beta)$ as a parametric model for the covariance. We then need to estimate the parameters $\sigma^2 > 0$ and $\beta > 0$, and for this we use a *minimum contrast estimation procedure* where $\int_\epsilon^{0.1} (\hat{c}(r)^{1/2} - c(r)^{1/2})^2 dr$ is minimized with respect to (σ^2, β); here ϵ denotes the smallest distance between the observed trees. As explained in Møller *et al.* (1996) this procedure can be fairly fast performed. We found $\hat{\sigma}^2 = 1.91$ and $\hat{\beta} = 1/33$; the dotted line in Figure 1 b) shows $\hat{\sigma}^2 \exp(-r/\hat{\beta})$, the estimated exponential covariance function.

3.3 Model checking

A second order characteristic closely related to the pair correlation function is the L-function defined by $L = \sqrt{K/\pi}$ where $K(t) = 2\pi \int_0^t rg(r)dr,\ t > 0$. The plot in Figure 1 c) shows a nonparametric estimate $\hat{L}$ for the data (solid line); apart from very small distances t, $\hat{L}$ is between the upper and lower envelopes (dotted lines) for the L-function based on 19 simulations of the fitted model. Our model shows a better fit with respect to the L-function than the Matérn cluster model used by Stoyan and Stoyan (1994).

Yet other summary statistics have been considered in order to check the appropiateness of the estimated log Gaussian Cox process: the distribution function F for the distance between an arbitrary but fixed location to the nearest tree; the distribution function G for the distance from a 'typical' tree to its nearest tree; and a certain third-order summary statistic (see Møller *et al.*, 1996). The plots in Figure 1 d) and e) show nonparametric estimates $\hat{F}$ and $\hat{G}$ based on the data against the mean of these estimates obtained from 99 simulations under the estimated log Gaussian Cox process. The plots show a reasonable good fit to the chosen model and $\hat{F}$ and $\hat{G}$ fall within the upper and lower envelopes based on the 99 simulations. For the Matérn cluster model fitted by Stoyan and Stoyan (1994) we have also created plots similar to d) and e) which

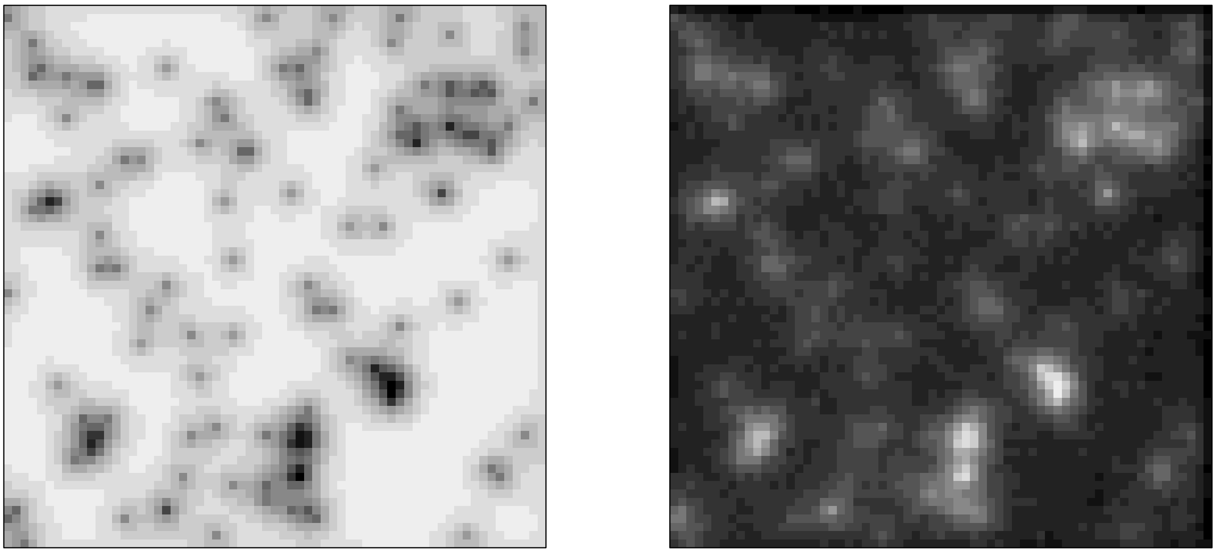

Figure 2: Posterior mean (left) and variance (right) of the log intensity surface.

indicate that our model fits the data better. Also the third-order summary statistic (omitted here to save space) gives no reason to doubt our model assumptions, while it raises serious doubt about the appropiateness of the Matérn cluster process as a model for the data. Finally, Figure 1 f) shows a simulated point pattern under the estimated log Gaussian Cox process, which apart from randomness looks much the same as the data in a).

4 Predicting the enviromental heterogeneity

Using the estimated log Gaussian Cox process we are able to predict the environmental heterogeneity by simulating from the conditional distribution of the log intensity surface Y given the observed pattern of trees; this may be considered as an a posteriori distribution obtained by an *empirical Bayesian approach.* The posterior mean $E(Y|\mathsf{data})$ as shown in Figure 2 is a good indication of the environmental heterogeneity. The posterior mean is ranging from 3.24 (white) to 7.59 (black). Figure 2 also shows the posterior variance $Var(Y|\mathsf{data})$ ranging from .69 (white) to 1.76 (black). By comparing the two plots in Figure 2 we see that the posterior variance is smallest where the posterior mean is largest and vice versa. In other words, given the observed trees the posterior variability in the intensity surface is largest where fewest Scots pine saplings are found.

Incidentally, our methods may also be used for predicting the pattern of trees outside the observation window. For further details the reader is referred to Møller *et al.* (1996).

References.

Møller, J., Syversveen, A.R. and Waagepetersen, R.P. (1996). Log Gaussian Cox processes. R-96-2036, Department of Mathematics, Aalborg University. 40 pages. To appear in *Scandinavian Journal of Statistics.*

Penttinen, A., Stoyan, D. and Henttonen, M.H. (1992). Marked point processes in forest statistics. *Forest Science* **38** (4), 806-824.

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