NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

A toolbox for fitting complex spatial point process models using integrated Laplace transformation (INLA)

by

Janine B. Illian and Håvard Rue

PREPRINT
STATISTICS NO. 6/2010

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This preprint has URL `http://www.math.ntnu.no/preprint/statistics/2010/S6-2010.pdf`

Håvard Rue has homepage: `http://www.math.ntnu.no/~hrue`
E-mail: `hrue@math.ntnu.no`
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491
Trondheim, Norway.

A toolbox for fitting complex spatial point process models using integrated Laplace transformation (INLA)

Janine B. Illian[*], Håvard Rue[†]

April 5, 2010

1 Introduction

These days a large variety of often rather complex statistical models can be fitted routinely to equally complex data sets as a result of widely accessible high-level statistical software, such as `R` (R Development Core Team, 2009) or `winbugs`, (Lunn et al., 2000). For instance, even the non-specialist user can estimate parameters in generalised linear mixed models or run a Gibbs sampler to fit a model in a Bayesian setting, and expert programming skills are no longer required. Researchers from many different disciplines are now able to analyse their data with sufficiently complex methods rather than resorting on simpler yet non-appropriate methods In addition, methods for the assessment of a model's fit as well as for the comparison of different models are widely used in practical applications.

The routine fitting of spatial point process models to (complex) data sets, however, is still in its infancy. This is despite a rapidly improving technology that facilitates data collection, and a growing awareness of the importance and relevance of small scale spatial information. Spatially explicit data sets have become increasingly available in many areas of science, including plant ecology (Burslem et al., 2001; Law et al., 2001), animal ecology (Forchhammer and Boomsma, 1995, 1998), geosciences (Naylor et al., 2009; Ogata, 1999), molecular genetics (Hardy and Vekemans, 2002), evolution (Johnson and Boerlijst, 2002) and game theory (Killingback and Doebeli, 1996) with the aim of answering a similarly broad range of scientific questions. Currently, these data sets are often analysed with methods that do not make full use of the available spatially explicit information. Hence there is a need for making the existing point process methodology available, but also so for extending it and accessible to applied scientists by facilitating the fitting of suitable models to help provide answers to concrete scientific questions.

There have been previous advances in facilitating routine model fitting for spatial Gibbs point processes, which model local spatial interaction in point patterns. Most markedly, the work by Baddeley and Turner (2000) has facilitated the routine fitting of Gibbs point processes based on an approximation of the pseudolikelihood to avoid the issue of intractable normalising constants (Lawson, 1992) as well as the approximate likelihood approach by Huang and Ogata (1999). Work by Baddeley et al. (2005) and Stoyan and Grabarnik (1991) has provided methods for model assessment for some Gibbs processes. Many of the methods have been made readily available through the library `spatstat` for `R` (Baddeley and Turner, 2005).

[*] Centre for Research into Ecological and Environmental Modelling The Observatory, University of St Andrews, St Andrews KY16 9LZ, Scotland, janine@mcs.st-and.ac.uk

[†] Department of Mathematical Sciences Norwegian University of Science and Technology, 7491 Trondheim, Norway, havard.rue@math.ntnu.no

However, most Gibbs process models considered in the literature are relatively simple. In an attempt to generalise the approach in Baddeley and Turner (2005), Illian and Henrichsen (2009) include random effects in Gibbs point processes but more complex models, such as hierarchical models or models including quantitative marks currently cannot be fitted in this framework. Similarly, there has been little discussion on explicit model comparison or assessment for complex models. The methods considered in Baddeley et al. (2005) and Stoyan and Grabarnik (1991) are restricted to relatively simple models. Furthermore, both the estimation based on the maximum likelihood and on the pseudolikelihood are approximate such that inference is not straight forward. The approximations become less reliable with increasing interaction strength (Baddeley and Turner, 2000) – in other words in realistic situations that researchers are particularly interested in, estimates are particularly unreliable.

Cox processes are another, particularly flexible class of spatial point process models (Møller and Waagepetersen, 2007). Even though many theoretical results have been discussed in the literature for these (Møller and Waagepetersen, 2004) the practical fitting of Cox point process models to point pattern data remains difficult due to intractable likelihoods. Fitting a Cox process to data is typically based on Markov chain Monte Carlo (MCMC) methods. These require expert programming skills and can also be very time-consuming to both tune and run (Møller and Waagepetersen, 2004) such that fitting complex models can easily become computationally prohibitive. Methods for model comparison or model assessment for Cox processes have not been discussed in the literature and there have been no attempts at providing routine model fitting approaches for these processes. As a result, Cox processes have been rarely used in the applied literature to answer concrete scientific questions based on realistic data.

Realistically complex data sets often consist of the exact spatial locations of the objects or events of interest as well as of further information on these objects, i.e. marks. They are often more complex than the classical data sets that have been analysed with point process methodology in the past. For example, some of these data sets consist of a large number of points representing the locations of objects in space (Burslem et al., 2001). The data often contain further information on the properties of the objects, i.e. on several potentially dependent qualitative as well as quantitative marks. In applications, the focus of the study may be to either model the patterns given the marks or in modelling the marks while taking the underlying spatial structure into account (Moore et al., 2010). Frequently, several (spatial) covariates have been collected such there is an interest in deciding which of these are relevant (John et al., 2007). Currently, Cox process models have not been fitted to data sets of such a complexity.

Applied researchers are aware that spatial behaviour tends to vary at a number of spatial scales as a result of different underlying mechanisms that drive the pattern (Wiegand et al., 2007; Latimer et al., 2009). Local spatial behaviour is often of specific interest but the spatial structure also varies on a larger spatial scale due to the influence of observed or unobserved spatial covariates. Cox processes model spatial patterns relative to observed or unobserved spatial trends and would be ideal models for these data sets. In particular, they are realistic models for many data sets, since the spatial trend is modelled as a random field. I.e. it is assumed to be stochastic and hence not fully known, which is rarely the case in applications.

However, two main issues arise – the Cox process models that have been considered so far are not complex enough to be suitable for many realistic data sets, and they typically do not consider local and larger scale spatial structures within the same model. More specifically, regarding the first issue, most applications of Cox process models have focused on the analysis of relatively small spatial patterns in terms of the locations of individual species. Very few attempts have been made of fitting models to both the pattern and the marks, in particular not to patterns with multiple dependent continuous marks (Ho and Stoyan, 2008; Myllymäki and Penttinen, 2009).

As far as the second issue is concerned, a specific strength of spatial point process models is their ability to take into account detailed information at very small spatial scales contained in spatial point pattern data, in terms of the local structure formed by an individual and its neighbours. So far, Cox processes – unlike Gibbs or Neyman-Scott processes – have often been used to relate the locations of individuals to phenomena that typically operate on larger spatial scales, i.e. to environmental trends. However, spatial point data sets are often collected with a specific interest in the local behaviour of individuals, such as spatial interaction or local clustering (Law et al., 2001; Latimer et al., 2009).

In this paper, we aim at proposing solutions to both these issues. We provide modern model fitting methodology for spatial point pattern data similar to what is common in other areas of statistics and has become a standard in many areas of application. We introduce a general modelling framework that provides a toolbox enabling the routine fitting of models to realistically complex data sets. We use integrate nested Laplace transformation (INLA) (Rue et al., 2009) to fit these models which speeds up parameter estimation substantially such that Cox processes can be fitted within feasible time. In order to makes the methods accessible to non-specialists an R package that may be used to run INLA is available and contains specific functions for fitting spatial point process models, see `http://www.r-inla.org/`.

We suggest an approach to fitting Cox process models that reflect both the local spatial structure and spatial behaviour at a larger spatial scale by using constructed covariates that account for spatial behaviour at different spatial scales. In doing this, we avoid issues with intractable normalising constants common in Gibbs process models and are able to consider local spatial structures within the flexible class of Cox processes. In addition, the class of models that we suggest here is general enough to allow very general hierarchical models to be fitted including models of both the spatial pattern and associated marks.

This paper is structured as follows. The general methodology is introduced in Section 2. In Section 3 we investigate the idea of mimicking local spatial behaviour by using constructed covariates in the context of simulated data with known spatial structures. Section 4 discusses a model for a large data set with both observed and constructed covariates and apply this to a rainforest data set. A hierarchical approach is considered in Section 5, where both (multiple) marks and the underlying pattern in a complex data set are modelled.

2 Methods

2.1 Spatial point process models

Point processes have been discussed in detail in the literature, see Stoyan et al. (1995); van Lieshout (2000); Diggle (2003); Møller and Waagepetersen (2004); Illian et al. (2008) for details. Here we aim at modelling a spatial point pattern $\mathbf{x} = (\xi_1, \ldots, \xi_n)$ regarding it as a realisation from a spatial point process $\mathbf{X}$. Note that for simplicity we consider only point processes in $\mathbb{R}^2$ but the approaches can be generalised to point patterns in higher dimensions.

We refer the reader to the literature for more details about different (classes of) spatial point process models such as the simple Poisson process, the standard null model of complete spatial randomness as well as the rich class of Gibbs (or Markov) processes (van Lieshout, 2000). Here, we generalise the class of Cox processes, in particular log Gaussian Cox processes. Cox processes lend themselves well to modelling spatial point pattern data with spatially varying environmental conditions (Møller and Waagepetersen, 2007) as they model spatial patterns based on an underlying (or latent) random field $\Lambda(\cdot)$ that describes the random intensity, assuming independence given this field. In other words, given the random field the point pattern forms a Poisson process. Log-Gaussian Cox processes as considered for example in Møller et al. (1998), Møller and Waagepetersen (2004) and Møller and Waagepetersen (2007) are a particularly flexible class where $\Lambda(s)$ has the form $\Lambda(s) = \exp\{Z(s)\}$, and $\{Z(s)\}$ is a zero mean Gaussian random field, $s \in \mathbb{R}^2$. Other examples of Cox processes include the shot noise Cox process Møller and Waagepetersen (2004).

Here, we consider a general class of spatial point process models based on log Gaussian Cox processes that admits both small and larger scale spatial behaviour. In order to do so, we employ spatially continuous constructed covariates exploiting the spatial structure contained in the spatial pattern to reflect behaviour at different spatial resolutions. This approach is discussed in detail in Section 3 where different constructed convariates are considered in the context of simulated patterns with known spatial structures.

2.2 Integrated nested Laplace approximation (INLA)

Cox processes are a special case of the very general class of *latent Gaussian models*, models of an outcome variable $\mathbf{y}$ that assume independence conditional on some underlying latent field ζ and hyperparameters θ. Rue et al. (2009) show that if ζ has a sparse precision matrix and the number of hyperparameters is small (i.e. ≤ 7), inference is fast, in particular with the integrated nested Laplace approximation (INLA) approach they have developed.

The main aim of the INLA approach is to approximate the posteriors of interest, the posterior marginals

$$\pi(\zeta_i|\mathbf{y}),$$

i.e. the marginal posteriors for the latent field and

$$\pi(\theta|\mathbf{y})$$

i.e. the marginal posterior for the hyperparameters and use these to calculate posterior means, variances etc.

These posteriors can be written as:

$$\pi(\zeta_i|\mathbf{y}) = \int \pi(\zeta_i|\theta, \mathbf{y})\pi(\theta|\mathbf{y})d\theta \tag{1}$$

$$\pi(\theta_j|\mathbf{y}) = \int \pi(\theta|\mathbf{y})d\theta_{-j} \tag{2}$$

This nested formulation is used to compute $\pi(\zeta_i|\mathbf{y})$ by approximating $\pi(\zeta_i|\theta, \mathbf{y})$ and $\pi(\theta|\mathbf{y})$ and by then using numerical integration to integrate out θ. This is feasible since the dimension of θ is small. Similarly $\pi(\theta_j|\mathbf{y})$ is calculated by approximating $\pi(\theta|\mathbf{y})$ and integrating out θ_{-j}.

The marginal posterior in equations (1) and (2) can be approximated as

$$\tilde{\pi}(\theta|\mathbf{y}) \propto \frac{\pi(\zeta, \theta, \mathbf{y})}{\tilde{\pi}_G(\zeta|\theta, \mathbf{y})} |_{\zeta=\zeta*(\theta)},$$

where $\tilde{\pi}_G(\zeta|\theta, \mathbf{y})$ is the Gaussian approximation to the full conditional of ζ and $\zeta * (\theta)$ is the mode of the full conditional for ζ for a given θ. This makes sense since the full conditional of a zero mean Gauss Markov random field can often be well approximated by a Gaussian distribution by matching the mode and the curvature at the mode (Rue and Held, 2005). Rue et al. (2009) show that this fact and the nested approach makes the approximation very accurate if applied to latent Gaussian models but substantially reduce the time required for fitting latent Gaussian models.

In the specific case of the log Gaussian Cox process the observation window is discretised into $N = n_{row} \times n_{col}$ grid cells $\{s_{ij}\}$ with area $|s_{ij}|$, $i = 1, \ldots, n_{row}, j = 1, \ldots, n_{col}$. This approach is taken in the examples in Sections 3, 4 and 5. The points in the point pattern are then denoted by $\xi_{ijk_{ij}}$ and the observed number y_{ij} of points in the grid cells, $k_{ij} = 1, \ldots, y_{ij}$. Conditional on $\eta_{ij} = Z(\xi_{ijk_{ij}})$ we consider

$$y_{ij}|\eta_{ij} \sim Po(|s_{ij}| \exp(\eta_{ij})), \qquad (3)$$

see Rue et al. (2009).

Based on this approach, existing spatial point process methodology may be generalised to enable the fitting of realistically complex models. We extend the ideas in Rue et al. (2009) and show that many types of realistically complex and flexible models can be fitted to point pattern data with complex structures with the INLA approach within reasonable computation time. This includes large point patterns with covariates operating on a large spatial scale and local clustering (Section 4) as well as point pattern data with several dependent marks which also depend on the intensity of the pattern (Section 5). Furthermore, the approach enables us to apply methods for both model comparison based on the deviance information criterion (DIC) and model assessment based on spatially structured and unstructured error fields.

3 Using constructed covariates to account for local spatial structure

The primary aim of using constructed covariates is to incorporate local spatial structure into a model while accounting for spatial variation (heterogeneity) and uncertainty at a larger spatial scale. Through this approach we can fit complex models with local spatial structures within the flexible class of Cox process models and avoid issues with intractable normalising constants common in the context of Gibbs processes. This is because the covariates operate directly on the mean (i.e. or, in other words, on the intensity of the pattern) rather than on the likelihood or the conditional intensity (Schoenberg, 2005).

In general, the constructed covariates we consider here are first order summary characterstics defined for any location in the observation window yielding an estimate of an (unobserved) random field that reflects spatial processes such as local interaction or competition in every location. In a similar way, one could consider constructed marks based on first or second order summary characteristics (Illian et al., 2008) that are only defined for the points in the pattern and include these in the model; we do not consider these here.

Clearly, in applications the covariates have to be carefully constructed depending on the questions of interest and different types of constructed covariates may be suitable in different contexts. Hence, our exposition below cannot be exhaustive; we discuss several examples that we consider to be of general relevance and apply some of these in the examples in Section 4 and 5. To illustrate the use of constructed covariates we simulate point patterns from various classical point process models. Note, however, that we do not aim at explicitly estimating the parameters of these models but at assessing whether known spatial structures may be detected and described through the use of constructed covariates as suggested here.

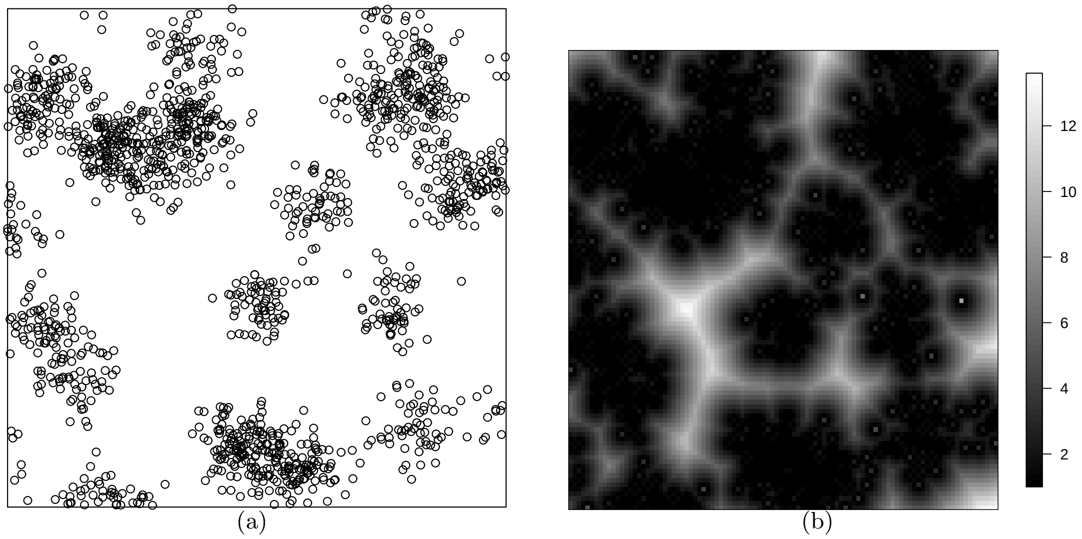
For most constructed covariates the functional relationship between the outcome variable and the covariate will not be obvious and might often not be linear. We thus estimate this relationship explicitly by a function and inspect this estimate to gain further information on the form of the spatial dependence.

3.1 Constructed covariates for local interaction

Constructed covariates are used to enable the modelling of spatial behaviour at different spatial scales. Clearly, as the behaviour at a larger spatial scale is not independent of the behaviour at a smaller scale, it is inherently difficult to disentangle the two. The covariates have to be carefully constructed and background knowledge as well as methods of model comparison might have to be used to inform the choice of the relevant spatial scale(s) for a specific pattern. With the aim of exploring the behaviour of constructed covariates here, we consider a few simple situations of simulated data sets. In the applications we have in mind, such as those discussed in the examples below, the data structure is typically much more complicated. The simulated data sets are only considered here to illustrate the use of constructed covariates and to assess their performance when the underlying spatial structure is known.

3.1.1 Nearest point distance

The first constructed covariate that we consider is the simple distance to the nearest point of the process. We assess the performance of the covariate for both clustered and regular patterns. For the former, we simulate a pattern from a spatially homogeneous Thomas process (Neyman and Scott, 1952) in the unit square, with parameters $\kappa = 20$ (the intensity of the Poisson process of cluster centres), $\sigma = 0.04$ (the standard deviation of the distance of a process point from the cluster centre) and $\mu = 60$ (the expected number of points per cluster). Figure 1 (a) shows a realisation from this process.



(a) (b)

Figure 1: Simulated Thomas process (a) and the associated constructed covariate d_{np} (b).

We consider a simple log Gaussian Cox process with

$$\Lambda(s) = \exp\{\beta \cdot c(s)\}, \tag{4}$$

i.e. $Z(.) = 0$ and where β is a parameter describing the dependance on the constructed covariate $c(s)$.

The unit square is discretised as discussed above using a 100×100 grid. For the mid-points of the grid cells we find the Euclidean distance to the nearest point in the pattern as a constructed covariate as

$$d_{np}(C(s_{ij})) = \min_{\xi_i \in \mathbf{x}}(\|C(s_{ij}) - \xi_i\|), \tag{5}$$

where $C(s_{ij})$ denotes the centre point of cell $\{s_{ij}\}$ and $\|\cdot\|$ the Euclidean distance. Refer to Figure 1 (b) for a plot of d_{np} for the pattern in Figure 1 (a). Note that this constructed covariate is related to the classical spherical contact distribution function (or empty space function) (Diggle, 2003; Illian et al., 2008). Noting that the dependence on the constructed covariate might not be linear, we fit a model to the data as indicated in equation (3) with

$$\eta_{ij} = \mu + f(d_{np}(C(s_{ij}))), \tag{6}$$

where μ is an intercept and $f(.)$ is an unknown function of the constructed covariate d_{np}.

The appropriate call to the `INLA` library in `R` is then

```
formula = Y ~ f(inla.group(dnp), model="rw1")
inla(formula, data=data.frame(Y,dnp), family="poisson", E=E)
```

where `Y` are the cell counts, `dnp` is the constructed covariate as defined in equation (5) and `E` is the area of the grid cells. Figure 2 shows the estimated functional relationship between the constructed covariate and the predicted values. The function clearly indicates that at small values of the covariate the intensity is negatively related to the constructed covariate reflecting the clustering a smaller distances. At larger distances (> 0.08) the function levels out indicating that at these distances the covariate and the intensity are unrelated reflecting the random behaviour of the pattern in locations outside the clusters.

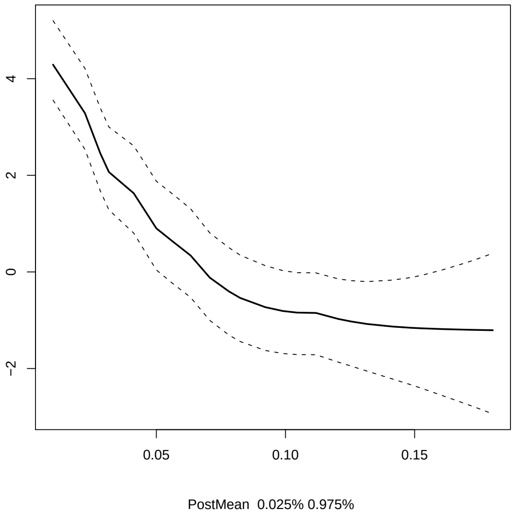
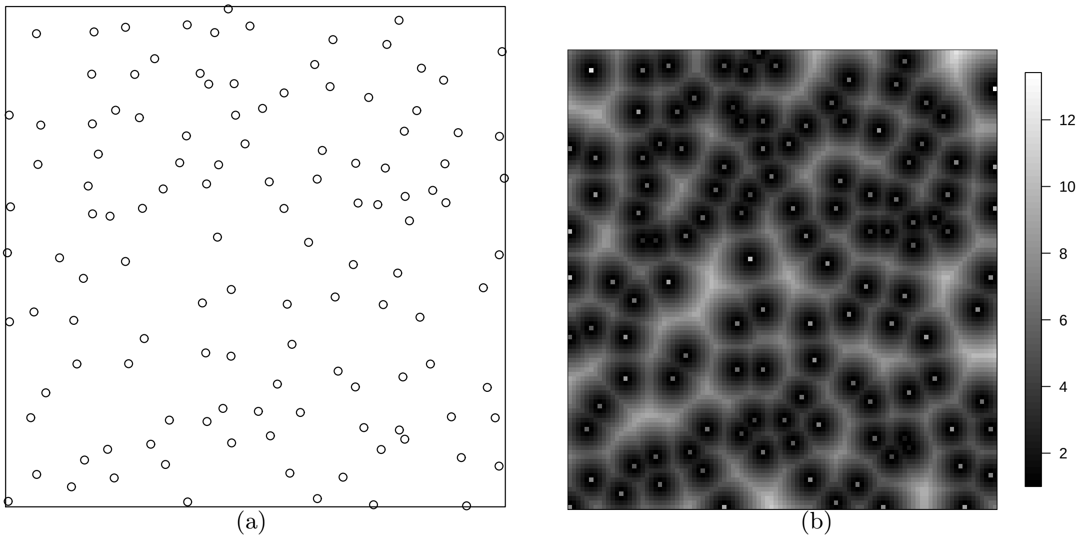


Figure 2: Functional relationship between constructed covariate d_{nn} and the predicted values for the realisation from a Thomas process shown in Figure 1.

To contrast these observation with results for regular patterns we simulate two patterns from a homogeneous Strauss process (Strauss, 1975) on the unit square, one with strong repulsion (intensity parameter $\beta = 350$, interaction parameter $\gamma = 0.01$ and interaction radius $r = 0.05$ and one with medium repulsion $\beta = 700$, $\gamma = 0.5$ and $r = 0.05$, respectively).

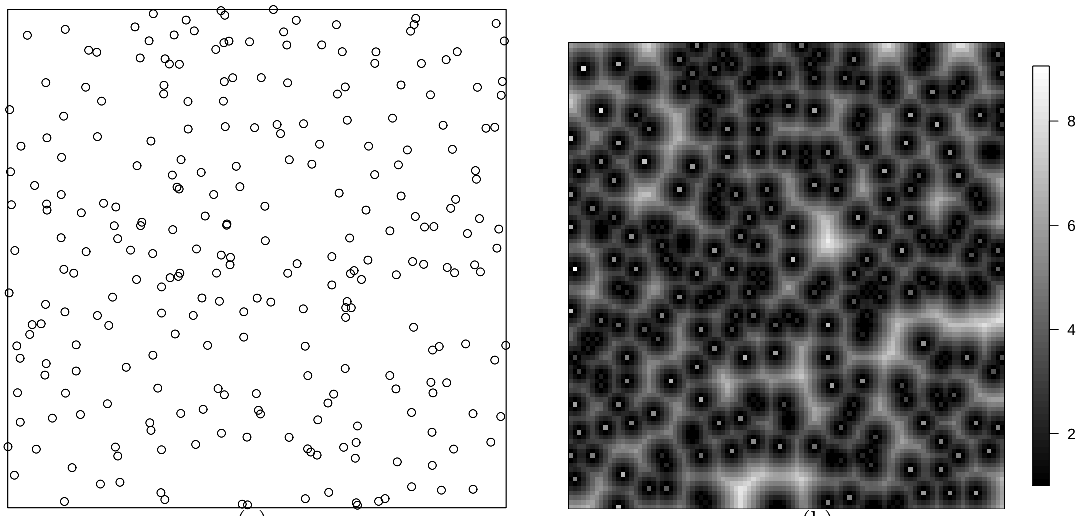
Figure 3 (a) shows a realisation from the process with strong repulsion and Figure 3 (b) shows the estimated constructed covariate d_{np} for this pattern. Figure 4 (a) shows a realisation from the process with medium repulsion and Figure 4 (b) shows the estimated constructed covariate.

We again fit the simple model in equation (6) to the data using the same constructed covariate as above. Figure 5 shows the estimated functional relationship between the constructed covariate and the outcome variable for the pattern with strong repulsion and Figure 6 for the pattern with medium



(a) (b)

Figure 3: Simulated Strauss process with strong repulsion ($r = 0.05$, $\beta = 350$, $\gamma = 0.01$) (a) and the associated constructed covariate d_{np} for this pattern (b)



(a) (b)

Figure 4: Simulated Strauss process with medium repulsion ($r = 0.05$, $\beta = 700$, $\gamma = 0.5$ (a) and the associated constructed covariate d_{np} for this pattern (b)

repulsion. Both functions indicate that at small values of the covariate the intensity is positively related to the constructed covariate clearly reflecting repulsion. At larger distances (> 0.05) the functions distinctly level out, indicating that beyond these distances the covariate and the intensity are unrelated, i.e. the spatial pattern shows random behaviour. In other words, the functional relationship not only characterises the pattern as regular but also correctly identifies the interaction distance as 0.05. In addition, the different degrees of repulsion are also reflected in the functional relationship. The results for the pattern with strong repulsion show both a larger slope in the functional relationship for distances smaller than 0.05 and a more pronounced kink at 0.05.

Note that the constructed covariate based on the nearest neighbour distance can be generalised by regarding it as a graph operating on the point pattern. Specifically, the nearest neighbour relationship can be expressed by connecting each point to its nearest neighbour in a graph. The local structure in a spatial point pattern may be reflected by other, more complicated graphs (see Rajala and Illian

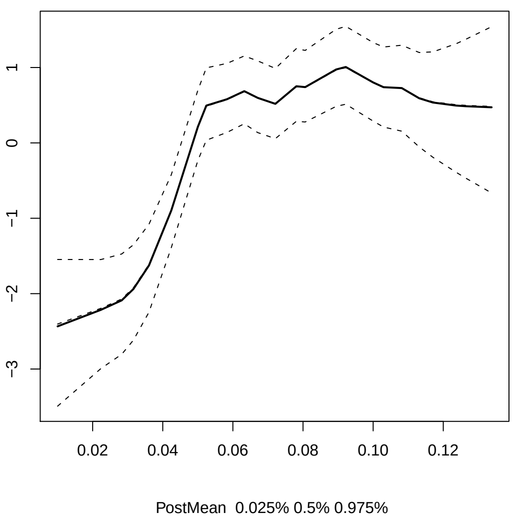


Figure 5: Functional relationship between constructed covariate d_{nn} and outcome variable for the realisation of a Strauss process with strong repulsion as shown in Figure 3 (a).

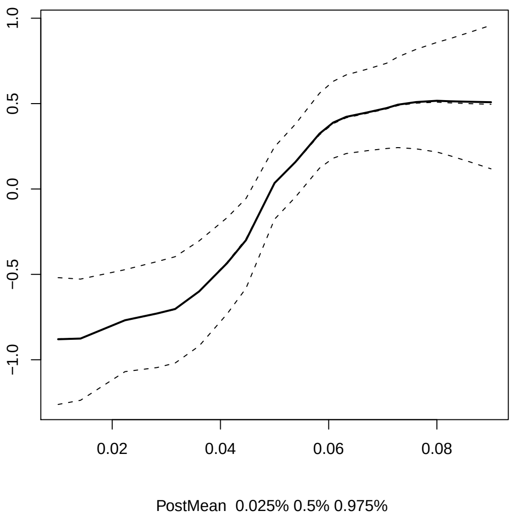


Figure 6: Functional relationship between constructed covariate d_{nn} and outcome variable for the realisation of a Strauss process with medium repulsion as shown in Figure 4 (a).

(2010)). Hence the approach can be generalised by either using different types of graphs to form a large class of potential constructed covariates.

3.1.2 Distance from nearest cluster centre

Clearly, considering only the nearest neighbours does not exploit all the available information contained in a spatial pattern and is a rather "short-sighted" measure that provides only a very local description. Another approach to constructing a covariate is to employ a clustering algorithm to identify clusters in the data and to calculate the distance of each grid cell location to the nearest cluster centre. As an example we use the k-means algorithm (Everitt et al., 2001) to identify clusters and associated cluster centres in a pattern. Clearly this approach is only suitable for clustered data.

We again generate a pattern from a homogeneous Thomas process, but with different parameters $\kappa = 10$, $\sigma = 0.05$ and $\mu = 80$. Since the k-means algorithm contains a stochastic element and hence does not produce a unique solution we run the algorithm repeatedly (100 times) with $k = 10$ and calculate d_{ncc}, the average Euclidean distance to the nearest cluster centre for each grid cell in 100×100 cell grid over 100 runs.

Figure 7 shows a realistion of the Thomas process along with the constructed covariate d_{ncc}.

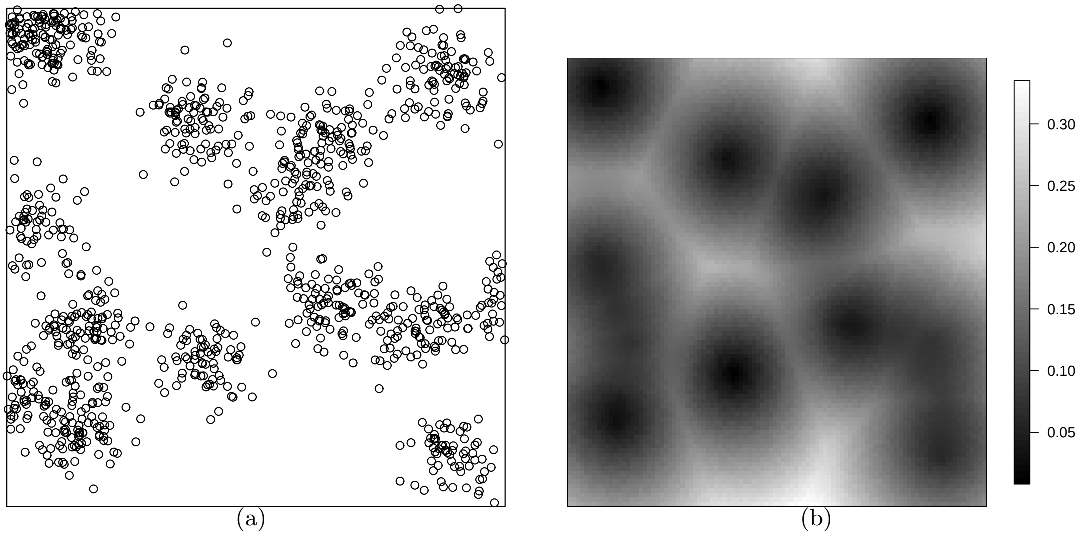


Figure 7: Simulated Thomas process (a) and the associated constructed covariate d_{ncc} (b).

We fit a simple log Gaussian Cox process model to the simulated data as in equation (8) and similar to above

$$\eta_{ij} = \mu + f(d_{ncc}(C(s_{ij}))), \tag{7}$$

where $f(.)$ is again an unknown function of the constructed covariate d_{ncc}.

The appropriate call to INLA is

```
formula = Y ~ f(inla.group(dncc), model="rw1",)
inla(formula, data=data.frame(Y,dncc), family="poisson", E=E)
```

where `Y` and `E` are as above and `dncc` is the distance to the nearest cluster centre.

Figure shows the estimated functional relationship between the intensity and the covariate d_{ncc}. The clustering behaviour of the pattern has clearly been detected by the covariate and is reflected in the negative relationship between the covariate and the intensity.

3.2 Constructed covariates for local interaction with trend

So far, we have only considered constructed covariates for homogeneous patterns. However, the main reason for using constructed covariates is that we aim at distinguishing behaviour at different spatial resolutions since in applications, including those we discuss below, patterns might be locally clustered but often are also inhomogeneous, i.e. the intensity of the pattern varies in space at a larger spatial scale. This a clearly an ill-posed problem since the behaviour at one spatial scale is not uniquely distinguishable from that at different spatial scale and background knowledge on suitable scales benefits the modelling process (Diggle, 2003). We use two further constructed covariates in this context - mainly to show that there is a wide range of potential constructed covariates that may be used to reveal local spatial behaviour.

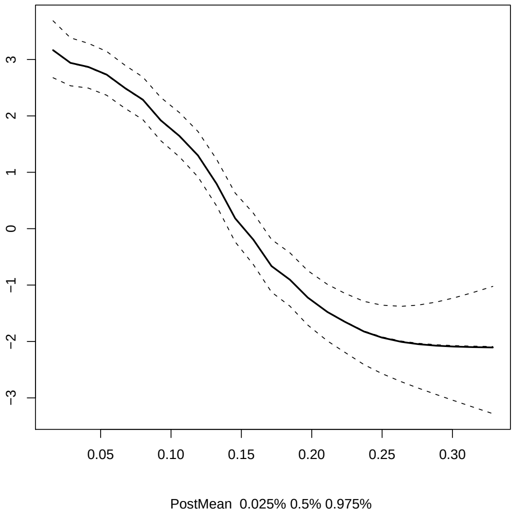


Figure 8: Functional relationship between constructed covariate d_{nc} and outcome variable for the realisation of a Thomas process as shown in Figure 7 (a); the value of k was chosen as 10.

3.2.1 Local density

We illustrate the use of constructed covariates in this context by simulating from an inhomogeneous Thomas process where the parent pattern is inhomogeneous. We generate a pattern from a Thomas process with parameters $\sigma = 0.02$ and $\mu = 20$ and a simple trend function for the intensity of parent points $\kappa(x_1, x_2) = 100x_1$. A realisation of the process is shown in Figure 9.

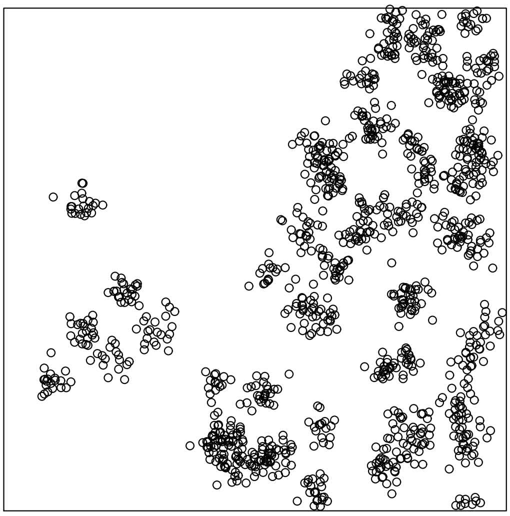


Figure 9: Realisation of an inhomogeneous Thomas process with parameters $\sigma = 0.02$ and $\mu = 20$ and trend function $\kappa(x_1, x_2) = 100x_1$

As we are interested in characterising local spatial behaviour we use a Gaussian kernel estimator with a small bandwidth (standard deviation of the Gaussian kernel = 0.02) to reflect local fluctuations in intensity as the constructed covariate. A plot of this covariate is shown in Figure 10. We consider a log Gaussian Cox process with

$$\Lambda(s) = \exp\{\beta \cdot c(s) + Z(s)\}, \tag{8}$$

where $Z(.)$ is a random field and β is a parameter describing the dependance on the constructed

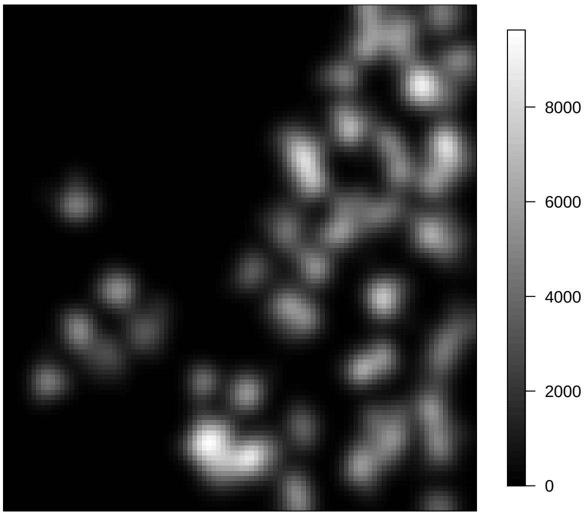


Figure 10: Local fluctuations in intensity estimated for pattern in Figure 9

covariate. After discretising the data, we fit the following model to the simulated data

$$\eta_{ij} = \mu + f(\hat{\lambda}_{0.02}(C(s_{ij}))) + f_s(s_{ij}), \tag{9}$$

where η_{ij}, $C(s_{ij})$ and $f(.)$ are as above, $\hat{\lambda}_{0.02}(C(s_{ij}))$ is the estimated density in $C(s_{ij})$ with bandwidth = 0.02 and $f_s(s_{ij})$ is a spatially structured effect, chosen to be a second order random walk on a lattice. We use vague gamma priors for the hyperparameter of the spatially structured effect.

Figure 11 shows a plot of the relationship between the constructed covariate and the outcome variable for the realisation of an inhomogeneous Thomas process shown in Figure 9. The relationship is clearly not linear and a strong local clustering effect has been picked up by the model. Figure 12 shows a plot of the estimated spatially structured effect for the same pattern. It still shows evidence of local clustering but at a slightly larger scale than the constructed covariate. This is certainly partly due to the fact that large scale inhomogeneity and local clustering are not independent and cannot be completely separated. However, a clear trend is visible in the estimated surface and has been accounted for by the model, i.e. local clustering and large scale inhomogeneity have been separated.

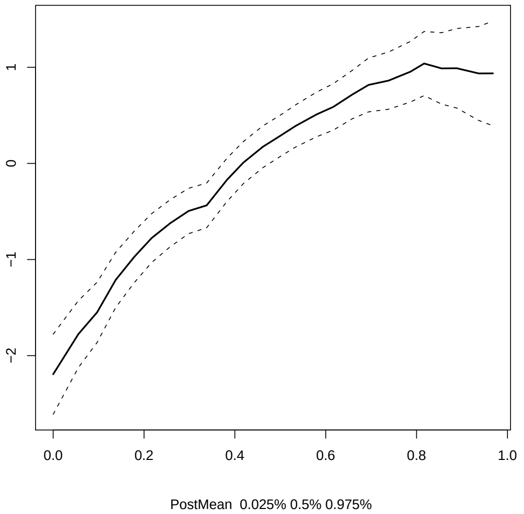


Figure 11: Functional relationship between constructed covariate and outcome variable for the realisation of an inhomogeneous Thomas process as shown in Figure 9

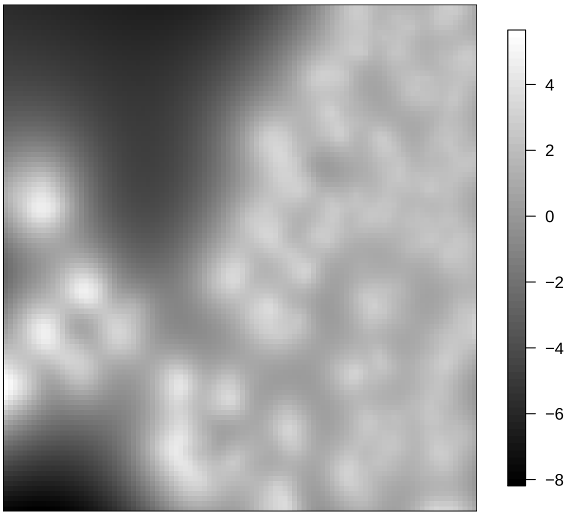


Figure 12: Estimated spatially structured effect for the realisation of an inhomogeneous Thomas process as shown in Figure 9

Note the constructed covariate considered here cannot be interpreted in the same straight forward way as the covariates discussed above as these were based on a distance whereas here we are estimating the density of points. Also note that this approach may be extended to allow for clustering at various spatial scales, potentially using several constructed covariates with a series of different bandwidths in order to pick up clustering at different spatial scales.

3.3 Discussion constructed covariates

With the aim of incorporating spatial behaviour at a local spatial scale we have considered a number of constructed covariates for simulated point patterns in various different scenarios. In all cases it was possible to identify the local spatial structure very clearly. The constructed covariates not only take account of local spatial model but may also be used to characterise the spatial behaviour. In several examples, the functional form of the dependence of the intensity on the constructed covariates clearly reflected the character of the local behaviour for instance in terms of clustering or regularity.

We have run a large number of simulations repeatedly sampling from the same point process models and always got very similar results with the same interpretation. The results from simulations from the same models with different sets of parameters were essentially the same. We also simulated patterns from a homogeneous Poisson process. As expected, when we fitted models such as those in equations 6 or 8 to these patterns, the functional relationship was always non-significant (results not shown).

Certainly, the covariates and the associated parameters such as bandwidth or number of clusters have to be chosen very carefully and background knowledge on the expected spatial scale of local clustering for instance can greatly improve the fit of a model. For instance, for the example we discuss in Section 4 some prior information on the range of local clustering is available. In cases like this, several models with varying values for the parameters related to local clustering around this prior value may be fitted and compared using the DIC.

Note if the constructed covariate d_{ncc} is used standard methods for determining the parameter k such as the test may be used. However it is not appropriate to use the compare different models based on the Deviance Information Criterion (DIC). This is due to the fact that while standard methods for determining the optimal choice of k aim at detecting the optimal number of clusters by minimising within cluster variation and maximising between-cluster variation whereas the DIC optimises the model fit. If the number of clusters is very high ($\approx$ number of points) the fit is optimised but the covariate does no longer reflect clustering at an appropriate spatial scale. If the DIC is to be used to estimate parameters for the constructed covariate, the parameter has to be directly related to the spatial resolution.

Note also that through the use of kernel estimators for the estimation of the local density our approach bears resemblance to shot-noise Cox processes) as well as to convolution based approaches (Calder and Cressie, 2007).

Admittedly, the use of constructed covariates is of a rather subjective nature and benefits from background knowledge on the data. However, when fitting a model which is purely based on empirical covariates similarly subjective decisions are usually made as these have been specifically chosen as potentially influencing the outcome variable, based on background knowledge. As mentioned above, the treatment of constructed covariates cannot be exhaustive here since different constructed covariates may be useful in different contexts. Clearly, many other approaches to defining constructed covariates may be considered. As indicated above these may be constructed based on different types of graphs (Rajala and Illian, 2010) and may also be of a second order nature. Similarly, approached based on morphological functions (Illian et al., 2008) may be used for this purpose.

4 Large point patterns with environmental covariates and local interaction

In order to illustrate our modelling approach and in particular the use of constructed covariates in practice we model a point pattern data set with a large number of points. We use a constructed covariate to account for local clustering as well as empirical covariates likely to impact on the larger scale spatial behaviour. To capture behaviour at spatial scales not accounted for by the covariates we include a Gaussian random field. An (i.i.d.) error field is used to detect any remaining spatial structures in the model and may be interpreted as a spatial residual term.

4.1 Data structure

We consider a spatial point pattern $\mathbf{x} = (\xi_1, \ldots, \xi_n)$ where the number of points n is potentially very large. The aim is to fit a model to $\mathbf{x}$. The point intensity is assumed to depend on one or several (observed or unobserved) random fields $Z_1, \ldots, Z_q$ as well as environmental covariates $z_1, \ldots, z_p$. We assume that the objects represented by the points cluster locally. In the application we have in mind (see Section 4.3) this clustering is a result of locally operating seed dispersal mechanisms.

4.2 Modelling approach

The point pattern $\mathbf{x}$ is modelled using a log Gaussian Cox process X with random intensity $\Lambda(.)$, where

$$\begin{aligned}\Lambda(s) &= \exp\{Z(s) + U(s) + \beta_0 + \beta_1 z_1(s) + \ldots + \beta_p z_p(s) \\ &+ f(c(s))\},\end{aligned} \tag{10}$$

where $\{Z(s)\}$ is a (stationary and isotropic) Gaussian random field and $U(s)$ is an error field with zero mean and variance covariance matrix I, $s \in S \subset \mathbb{R}^2$. The $z_1(.), \ldots, z_p(.)$ are (environmental) covariates for which data exist in each $s \in S$, and $c(.)$ is a constructed covariate reflecting local clustering. Since the functional relationship between the intensity Λ and the constructed covariates may not be linear we consider an unknown function of the constructed covariate, $f(.)$, and investigate its shape, in analogy to approaches taken in Section 3.

4.3 Application to example data set

4.3.1 The rainforest data

Some extraordinarily detailed multi-species maps are being collected in tropical forests as part of an international effort to gain greater understanding of these ecosystems (Condit, 1998; Hubbell et al., 1999; Burslem et al., 2001; Hubbell et al., 2005). These data comprise the locations of all trees with diameters at breast height (dbh) 1 cm or greater, a measure of the size of the trees (dbh), and the species identity of the trees. The data usually amount to several hundred thousand trees in large (25 ha or 50 ha) plots that have not been subject to any sustained disturbance such as logging. The spatial distribution of these trees is likely to be determined by both spatial varying environmental conditions and local dispersal.

Recently, spatial point process methodology has been applied to analyse some of these data sets (Law et al., 2009; Wiegand et al., 2007) using non-parametric descriptive methods. In addition, there have been a few explicit modelling attempts, including Waagepetersen (2007); Waagepetersen and Guan (2009). Rue et al. (2009) model the spatial pattern formed by a tropical rain forest tree species solely on the underlying environmental conditions and use the INLA approach to fit the model. We analyse the same data here. Since the spatial structure in a forest also reflects dispersal mechanisms we use a constructed covariate to account for local clustering. We use local density as estimated by a Gaussian kernel for this purpose and choose a standard deviation of 20, based on background knowledge about the dispersal typical distance (D. Burslem, personal communication).

Figure 13 shows the spatial pattern formed by trees of the species *Beilschmidia* on Barro Colorado Island (BCI) and Figure 14 shows a plot of the normalised covariates elevation and gradient.

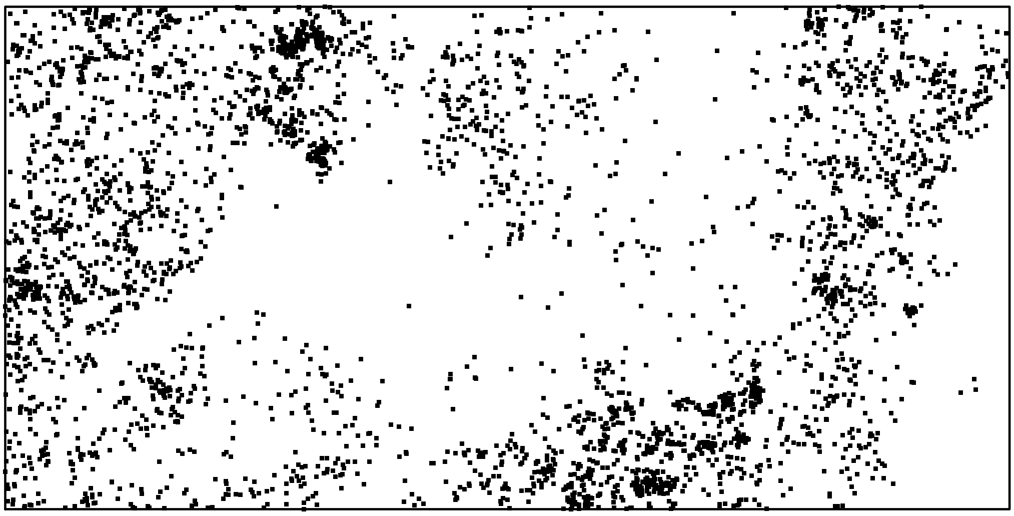


Figure 13: Spatial pattern of the species Beilschmiedia on Barro Colorado Island, Panama.

Figure 15 shows a plot of the normalised constructed covariate. We discretise the observation window as discussed above. For the locations of the trees we thus have for η_{ij} (see equation (3))

$$\eta_{ij} = f_s(s_{ij}) + u_{ij} + \beta_0 + \beta_1 \cdot z_{1ij} + \beta_2 \cdot z_{2ij} + f(c_{ij}),$$

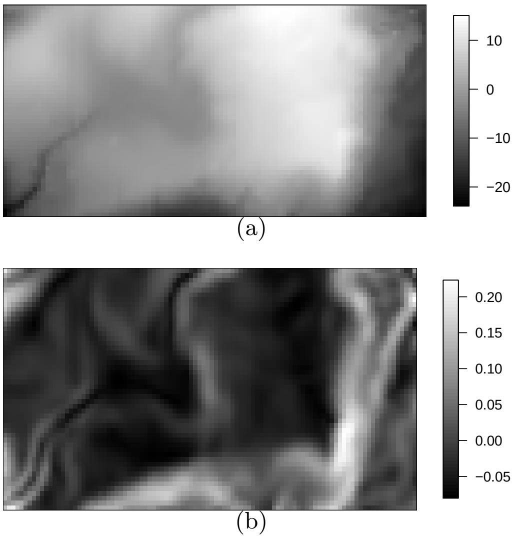


Figure 14: Normalised covariates evaluation (a) and gradient (b) in rainforest plot on Barro Colorado Island, Panama.

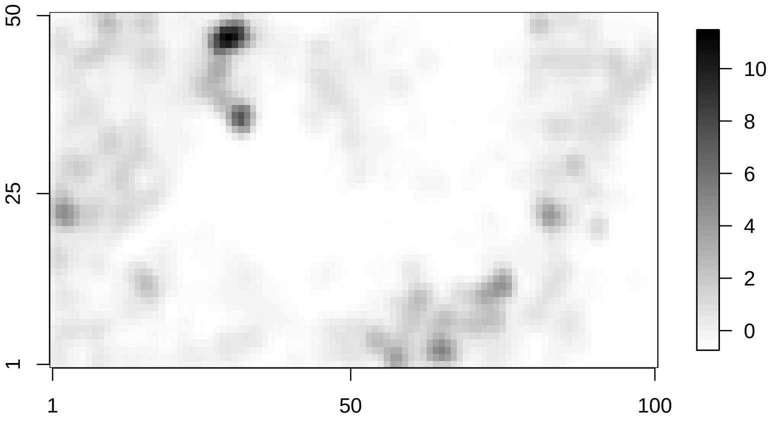


Figure 15: Normalised constructed covariate reflecting local clustering in the rainforest data set

where $f_s(s_{ij})$ is a spatially structured location effect reflecting the random field $\{Z(.)\}$, chosen to be a second order random walk on a lattice. We use vague gamma priors for the hyperparameter of the spatially structured effect. u_{ij} is an unstructured random effect reflecting the error field $\{U(.)\}$. z_{1ij} and z_{2ij} are observed covariates (elevation and gradient) for each grid cell. The c_{ij} are the values of the constructed covariate in grid cell k_{ij} reflecting local clustering.

4.3.2 Results

The parameter estimates of the fixed effects are $\beta_1 = 0.056, \beta_2 = 0.127$. These results indicate that trees are more likely to be located in areas where the two environmental covariates have high values. However, only the effect of the second covariate (gradient) is significant as the 95% posterior interval (0.051, 0.204) does not cover 0 whereas the effect of the first covariate (elevation) is not significant; posterior interval: (-0.134, 0.246). The effect of the constructed covariate that reflects small scale clustering is plotted in Figure 16. A significant local clustering effect has been identified by the model.

In Figure 17 we consider a plot of the structured (top) and the unstructured (bottom) effect. The

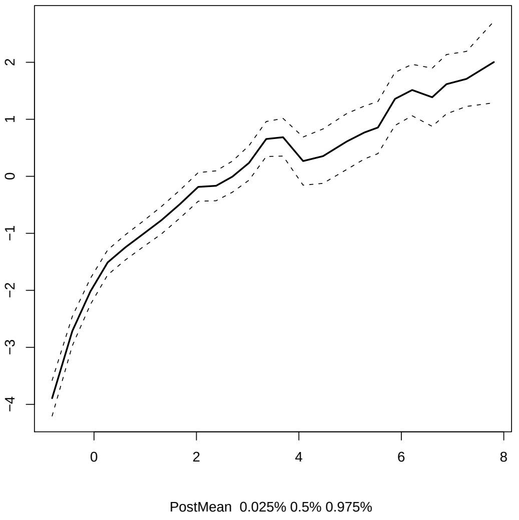
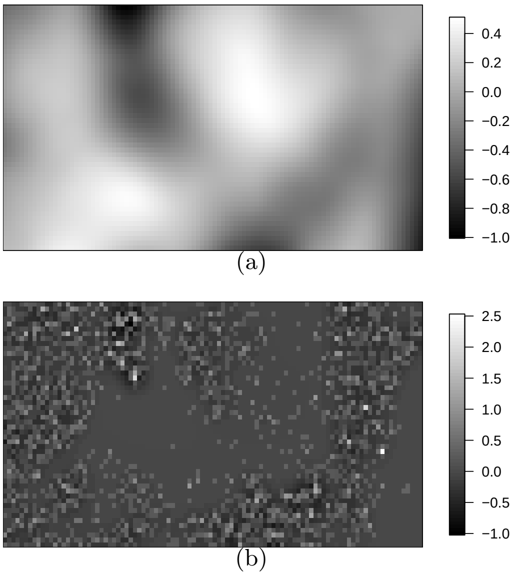


Figure 16: Effect of the constructed covariate reflecting local clustering for the rainforest data.



(a)

(b)

Figure 17: The structured (a) and the unstructured effect (b) for the rainforest data.

structured spatial effect clearly accounts for spatial behaviour at a scale larger than that reflected in the constructed covariate. In addition, a residual spatial structure is clearly visible in Figure 17 (bottom) indicating that the model does not sufficiently explain all the spatial structure contained in the data. This might be due to further covariates that operate at a different spatial scale than the structured effect and the covariates that have been currently considered. Clearly, the plot indicates that the current model appears to be unable to account for the complete absence of Beilschmidia trees in the central area (see Figure 13); current covariates cannot fully explain this absence.

4.4 Discussion rainforest data

In this section, we consider a Cox process model for a point pattern data set with a large number of points and two observed covariates. Waagepetersen (2006), Waagepetersen and Guan, 2009, Gimona et al. (2009) model the patterns formed by rainforest tree species with this data structure using a Thomas processes to include local clustering resulting from seed dispersal. This approximate approach is based on the minimum contrast method for parameter estimation. Rue et al. 2009 consider the same data in the context of Cox processes to demonstrate that log Gaussian Cox processes can conveniently be fitted to a large spatial point pattern using INLA solely on basis of the environmental covariates. We generalise this approach and fit a model based on the two empirical covariates and a constructed covariate that reflects local clustering as a result of local seed dispersal, as discussed above.

Fitting the model took about 30 minutes on a standard PC. It would hence be possible to fit the model with different standard deviations for the Gaussian kernel of the constructed covariate, reflecting different dispersal distances and use the DIC to decide on an appropriate choice of spatial scale. Similarly, if a larger number of empirical covariates is available, the DIC could be used for model choice.

An inspection of the error field indicates that some spatial structure still remains in the data which cannot be explained by the current model and that spatial correlation may operate at a different spatial scale than the structured spatial effect and the covariates considered. It is likely that further covariates that operate on intermediate spatial scales, e.g. data on soil properties may improve the model. It is not surprising to see that the pattern exhibits spatial dependence at a number of spatial scales, in particular given the size of the pattern.

We were able to fit a complex model to this large data set, indicating that local spatial behaviour in a real data set can be identified by including a constructed covariate. A benefit of using INLA for parameter estimation is that both model comparison and model assessment are relatively straight forward. These are certainly of practical value in many applications. In particular, the structured and unstructured spatial effects may be interpreted as spatial residuals and used to suggest potential further covariates to the applied user to improve the model.

The approach discussed here can be easily extended to accommodate more complex models, including models with larger number of environmental covariates or a model of both the spatial pattern and associated marks, along the lines of the model discussed in Section 5. For instance, this could include a model of both the spatial pattern and the size or the growth of the trees.

5 Modelling marks and pattern in a marked point pattern with multiple marks

Here we aim at including (several dependent) marks in the model and model these in dependence on and together with the spatial structure. Situations where marks depend on the intensity of the point pattern have been considered in Ho and Stoyan (2008) andMyllymäki and Penttinen (2009). Also note the work by Diggle et al. (2010), where a point process with intensity dependent marks is used in the context of preferential sampling in geostatistics. All of these approaches are related to the approach taken here but consider a single mark only.

5.1 Data structure

We consider a spatial point pattern $\mathbf{x} = (\xi_1, \ldots, \xi_n)$ together with several types of – nonindependent – associated marks. For simplicity we consider only two marks $\mathbf{m}_1 = (m_{11}, \ldots, m_{1n})$ and $\mathbf{m}_2 = (m_{21}, \ldots, m_{2n})$ here but the approach can be generalised in a straight forward way to more than two marks. The $\mathbf{m}_1$ are assumed to follow an exponential family distribution $F_{1\theta_1}$ with parameter vector $\theta_1 = (\theta_{11}, \ldots, \theta_{1q})$ and to depend on the intensity of the point pattern and the $\mathbf{m}_2$ are assumed to follow a different exponential family distribution $F_{2\theta_2}$ with parameter vector $\theta_2 = (\theta_{21}, \ldots, \theta_{2q})$ and to depend on the intensity of the point pattern but also on the marks $\mathbf{m}_1$. Without loss of generality the parameters θ_{11} and θ_{21} are the location parameters of the distributions F_1 and F_2, respectively.

5.2 Modelling approach

We model the pattern by using a hierarchically marked log Gaussian Cox process. Again the locations are modelled as a log Gaussian Cox process with random intensity

$$\Lambda(s) = \exp\{\beta_1 Z(s) + U(s)\},$$

where, as above, $\{Z(.)\}$ is a Gaussian random field, $s \in S \subset \mathbb{R}^2$ and $\{U(.)\}$ an error field. In the example considered below data on (environmental) covariates are not available. For the sake of the exposition we do not consider constructed covariates here either but empirical and constructed covariates can easily be included in a similar model.

The marks $\mathbf{m}_1$ are modelled as:

$$m_1(\xi_k)|\Lambda(\xi_k) \sim F_{1\theta_1}(\theta_{11} = \mu_1 + \beta_2 \cdot Z(\xi_k) + V(\xi_i), \theta_{12}, \ldots, \theta_{1q}),$$

where $\xi_k \in \mathbf{x}$, $\{Z(.)\}$ is as above and $\{V(.)\}$ is another error field. The marks $\mathbf{m}_2$ are modelled as:

$$m_2(\xi_k)|\Lambda(\xi_k) \sim F_{2\theta_2}(\theta_{21} = \mu_2 + \beta_3 \cdot Z(\xi_k) + \beta_4 \cdot m_1(\xi_k) + W(\xi_k), \theta_{22}, \ldots, \theta_{2q}),$$

where $\xi_k \in \mathbf{x}$, $\{Z(.)\}$ are as above and $\{W(.)\}$ is another error field. The $(\theta_{12}, \ldots, \theta_{1q})$ and the $(\theta_{22}, \ldots, \theta_{2q})$ are hyperparameters.

5.3 Application to example data set

5.3.1 Koala data

Koalas are arboreal marsupial herbivores native to Australia with a very low metabolic rate. They rest motionless for about 18 to 20 hours a day, sleeping most of that time. They feed selectively and live almost entirely on eucalyptus leaves. Whereas these leaves are poisonous to most other species, the koala gut has adapted to digest them. It is likely that the animals preferentially forage leaves that are high in nutrients and low in toxins (FPC) as an extreme example of evolutionary adaptation. An understanding of the koala - eucalyptus interaction is crucial for conservation efforts (Moore et al., 2010).

The complete data set consists of the locations of 915 eucalyptus trees. For each tree, information on the leave chemistry in terms of a measure of the palatability of the leaves (the "leave marks" $\mathbf{m}_L$) is available, which is assumed to be normally distributed and to depend on the intensity of the point pattern. In addition, the "frequency marks" $\mathbf{m}_F$ describe for each tree the diurnal tree use by individual koalas collected at monthly intervals between 1993 and March 2004. The $\mathbf{m}_F$ are assumed to follow a Poisson distribution and to depend on the intensity of the point pattern as well as on the leave marks. For reasons of data ownership we have selected only a subset of the data in a smaller observation window here and fit the model to these data points. Clearly, it is straight forward to fit the model to the larger data set.

There is no additional covariate data available, hence for the locations of the trees we simply have

$$\eta_{ij} = \mu_{int} + \beta_1 \cdot f_s(s_{ij}) + u_{ij},$$

where $f_s(s_{ij})$ is a spatially structured effect reflecting $\{Z(.)\}$, chosen to be a second order random walk on a lattice. We use vague gamma priors for the hyperparameter of the spatially structured effect. u_{ij} an unstructured random effect reflecting $\{U(.)\}$.

For the leave marks we have $m_L(\xi_{ijk_{ij}})|\kappa_{ijk_{ij}} \sim N(\kappa_{ijk_{ij}}, \sigma^2)$ with

$$\kappa_{ijk_{ij}} = \mu_L + \beta_2 \cdot f_s(s_{ij}) + v_{ijk_{ij}},$$

where $f_s(s_{ij})$ is a spatially structured location effect and $v_{ijk_{ij}}$ another unstructured random effect, reflecting $\{V(.)\}$

The frequency marks are assumed to follow a Poisson distribution and to depend on the leave marks. We have that $m_F(\xi_{ijk_{ij}})|\nu_{ijk_{ij}} \sim Po(\exp(\nu_{ijk_{ij}}))$ with

$$\nu_{ijk_{ij}} = \mu_F + \beta_3 \cdot f_s(s_{ij}) + \beta_4 \cdot m_L(\xi_{ijk_{ij}}) + w_{ijk_{ij}},$$

where $f_s(s_{ij})$ is the same spatially structured location effect as above and $w_{ijk_{ij}}$ another unstructured random effect, reflecting $\{W(.)\}$

5.3.2 Results

The fit of an initial model that assumes that the influence of random field $\{Z(.)\}$ is different for the spatial pattern and the two marks, i.e. that $\beta_1 \neq \beta_2 \neq \beta_3$, indicates that $\beta_1 \approx \beta_2 \neq \beta_3$ (results not shown). We hence fit a second model assuming that $\beta_1 = \beta_2 \neq \beta_3$ and compare the model fit based on the deviance information criterion (DIC). The second model has a smaller DIC; we hence only present the results for this model.

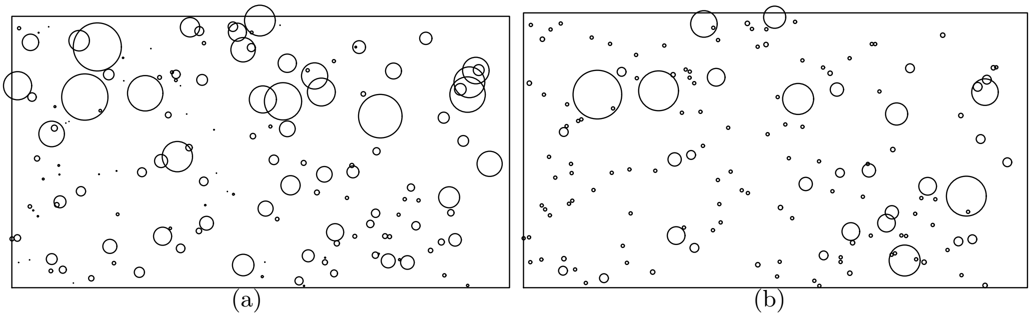


Figure 18: Spatial pattern of the koala data; the diameters of the circles reflect the value of the leave marks (a) and the frequency marks (b) respectively.

The mean of the posterior density for the parameter of interest, β_4 is 1.573; posterior interval (1.279, 1.895), indicating a significant positive influence of palatability on the frequency of koala visits to the trees.

Figure 19 (a) shows a plot of the structured spatial effect $f_s(s_{ij})$ and Figure 19 (b-d) of the three unstructured spatial effects u_{ij} , $v_{ijk_{ij}}$ and $w_{ijk_{ij}}$. The smooth spatially structured effect shows a clear trend. On inspection of the unstructured spatial effect little remaining structure is apparent in any of the three plots indicating that the covariates and the structured spatial effect have explained most of the variability in the data, both for the spatial pattern and the two marks.

5.4 Discussion koala data

The example considered in this section is a marked Cox process model, i.e. a model of both the spatial pattern and two types of dependent marks allowing us to learn about the spatial pattern at

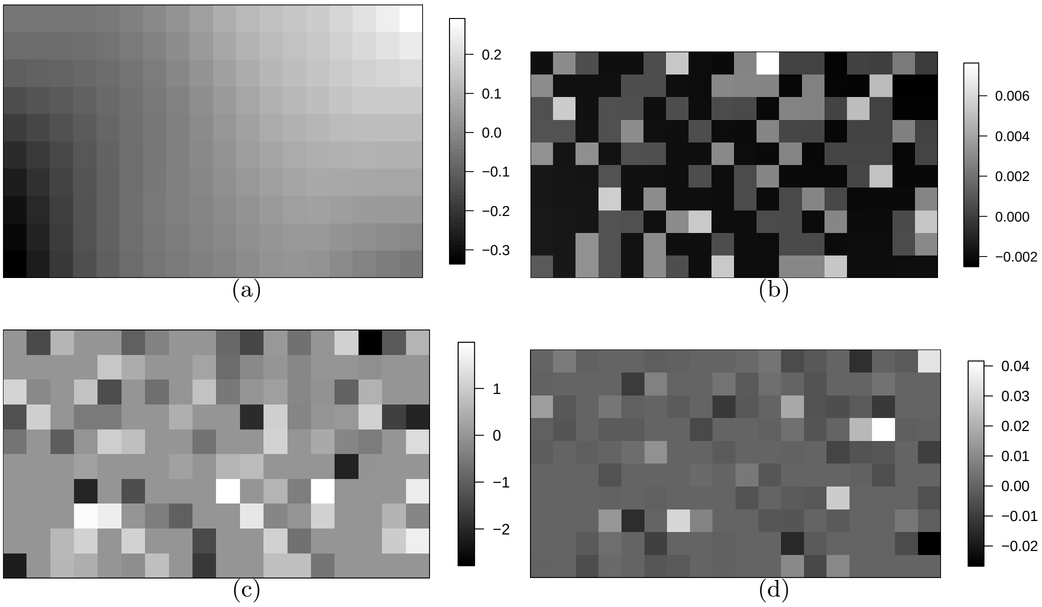


Figure 19: Plots of structured spatial effect (a) and the three unstructured effects (b - d).

the same time as about the marks. In cases where the marks are of primary scientific interest one could view this approach as a model of the marks which implicitly takes the spatial dependence into account by modelling it alongside the marks.

The frequency marks are modelled as Poisson counts, the leaf marks as Gaussian data and the spatial locations based on a structured and an unstructured spatial effect but no covariates. This may easily be extended to include constructed covariates to incorporate local spatial behaviour but was omitted here for simplicity. Similarly, empirical covariates may be included in the model but have not been available here.

Model assessment indicates little remaining spatial structure in the unstructured random field hence indicating that the model sufficiently explains the spatial dependence in the data. This does, however, not mean that that the model discussed here is the perfect model. For instance, the koala count data are overdispersed (mean = 1.014, variance = 3.834) such that it might make sense to consider a negative binomial distribution instead of the Poisson distribution for the frequency marks.

Data sets of the type considered here are relatively common within ecology. For instance metapopulation data (Hanski, 2009) typically consist of the locations of sub-populations and their properties where ecologists are usually interested in modelling the properties given the underlying spatial structure.

6 Discussion

There is an increasing interest in modelling spatial structures in many disciplines and researchers have become familiar with fitting a large range of different models to complex data sets using R. Hence methodology that allows the routine fitting of spatial point process models to real life data sets is likely to be welcome by many applied researchers. This paper provides a very flexible framework for routinely fitting models to (potentially) complex spatial point pattern data using a model class that accounts for both local and global spatial behaviour.

There is an extensive literature on descriptive and non-parametric approaches to the analysis of spatial point patterns, specifically on (functional) summary characteristics describing first and second order spatial behaviour, in particular on Ripley's K-function and the pair correlation function. In both the statistical and the applied literature these have been discussed far more frequently than modelling approaches and provide an elegant means for characterising the properties of spatial patterns (Illian et al., 2008). A thorough analysis of a spatial point pattern typically includes an extensive exploratory analysis and in many cases it may even seem unnecessary to continue the analysis and fit a spatial point process model to a pattern. The fact that an exploratory analysis based on a functional summary characteristics such as Ripley's K–function or the pair-correlation function considers spatial behaviour at a multitude of spatial scales make this approach particularly appealing. However, with an increasing complexity of the data it becomes less obvious how suitable summary characterstics should be defined for these and a point process model may be a suitable alternative. For example, it is not obvious how one would jointly analyse the two different marks together with the pattern in a purely exploratory analysis of the koala data set discussed in Section 5 but it is straight forward to do this with a hierarchical model as discussed above. In addition, most exploratory analysis tools assume the process to be first order stationary or at least second order reweighted stationary (Baddeley et al. 2000)– a situation that is both rare and difficult to assess in applications, in particular in the context of realistic and complex data sets. The approach discussed here does not make any assumptions about stationarity but explicitly includes the inhomogeneity into the model. In other words, through the use of constructed covariates our approach combines the use of (currently first order) functional summary characteristics with modelling.

Large scale spatial behaviour may be incorporated into a Gibbs process model as a parametric or non-parametric, yet deterministic, trend while it is treated as a stochastic process in itself here. Modelling the spatial trend in a Gibbs process hence often assumes that an explicit and deterministic model of the trend as a function of location (and spatial covariates) is known (Baddeley and Turner, 2005). Even in the non-parametric situation the estimated values of the underlying spatial trend are considered fixed values, which are neither subject to stochastic variation nor to measurement error. Since it is based on a latent random field the approach discussed here substantially differs from the Gibbs process approach and assumes a hierarchical, doubly stochastic structure. This very flexible class of point processes provides models of local spatial behaviour relative to an underlying large scale spatial trend. In realistic applications this spatial trend is not known. Values of the covariates that are continuous in space are typically not known everywhere and have been interpolated and it is likely that spatial trends exist in the data that cannot be accounted for by the covariates. The spatial trend is hence not regarded as deterministic but modelled as a random field consisting of spatially structured as well as spatially unstructured random effects and fixed or random covariate effects.

In summary, we combine the flexibility of the log Gaussian Cox process that results from its doubly stochastic structure with the use of constructed covariates to reflect local spatial behaviour. This enables us to fit point processes that reflect spatial structures typically modelled by Cox processes as well as those modelled by Gibbs processes but avoid untractable normalising constants.

In addition, in many scientific areas, spatial point pattern data have been collected in the field, i.e. not as part of a controlled experiment. Hence often many potential factors and dependence structures have not been controlled for but yet are likely to impact on the spatial pattern, in particular on the large scale behaviour. This typically results in the collection of a large number of covariates as well as marks and hence requires highly complex models but it is not always clear, which of these influence the spatial structure. In order to be able to use spatial point process methodology to decide which covariates are relevant. Scientists need to be able to fit models quickly and easily enough such that several potential models can be fitted and compared, and that the most suitable model can be chosen within reasonable time. This requires suitable tools for statistical inference to appropriately interpret the fitted models, including methods for model comparison as well as for model assessment.

In this paper, we develop methodology for routinely fitting suitably complex point process models with little computational effort that take into account both local spatial structure and spatial behaviour at a larger scale. The local spatial structure is described by carefully chosen constructed covariates which we discuss in detail in Section 3. We consider complex data examples and incorporate constructed covariates into a model. We also demonstrate how marks can be included in a joint model of the marks and the locations and hence how the marks can be modelled while the spatial structure is implicitly accounted for.

The two very different examples indicate that our approach can be applied in a wide range of situations and is flexible enough to facilitate the fitting of other even more complex models. For both data sets the parameter estimation procedure took only a few minutes to run. This indicates that it is hence feasible to fit several related models to realistically complex data sets if necessary and use the DIC to aid the choice of covariates. The posterior distributions of the estimated parameters were used to assess the significance of the influence of different covariates in the models. Through the use of a structured spatial effect and an unstructured spatial effect it has been possible to assess the quality of the model fit for each of the three models. Specifically, the structured spatial effect can be used to reveal spatial correlations in the data that have not been explained with the covariates and may help researchers identify suitable covariates to incorporate into the model. The unstructured spatial effect may be regarded as a spatial residual. Remaining spatial structure visible in it indicates further unexplained spatial dependence in the data most likely at a different spatial resolution than the spatial autocorrelation explained by the structured spatial effect or the spatial covariates.

In summary, methodology discussed here, together with the specific `R` library (`http://www.r-inla.org/`), makes the complex spatial point process models accessible to scientists and provides them with a toolbox for routinely fitting and assessing the fit of suitable and realistic point process models to complex spatial point pattern data.

Acknowledgements

The rainforest data set was collected with the support of the Center for Tropical Forest Science of the Smithsonian Tropical Research Institute and the primary granting agencies that have supported the BCI plot. The BCI forest dynamics research project was made possible by National Science Foundation grants to Stephen P. Hubbell: DEB-0640386, DEB-0425651,DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-9615226, DEB-9405933, DEB-9221033, DEB-9100058, DEB-8906869, DEB-8605042,DEB-8206992, DEB-7922197, support from the Center for Tropical Forest Science, the Smithsonian Tropical Research Institute, the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Celera Foundation, and numerous private individuals, and through the hard work of over 100 people from 10 countries over the past two decades. The plot project is part the Center for Tropical Forest Science, a global network of large-scale demographic tree plots.

We thank Colin Beale, University of York and Ben Moore, Macauley Land Research Institute, Aberdeen for extended discussions on the koala data.

References

Baddeley, A. and R. Turner (2000). Practical maximum pseudolikelihood for spatial point processes. *New Zealand Journal of Statistics 42*, 283–322.

Baddeley, A., R. Turner, J. Møller, and M. Hazelton (2005). Residual analysis for spatial point processes (with discussion). *Journal of Royal Statistical Society Series B 67*, 617–666.

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

Burslem, D. F. R. P., N. C. Garwood, and S. C. Thomas (2001). Tropical forest diversity – the plot thickens. *Science 291*, 606–607.

Calder, C. A. and N. A. Cressie (2007). Some topics in convolution-based spatial modeling. In *Proceedings of the 56th Session of the International Statistics Institute.*

Condit, R. (1998). *Tropical Forest Census Plots.* Springer-Verlag and R. G. Landes Company, Berlin, Germany, and Georgetown, Texas.

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

Diggle, P., R. Menezes, and T. Su (2010). Geostatistical inference under preferential sampling (with discussion). *Journal of the Royal Statistical Society Series C 59*, 191– 232.

Everitt, B., S. Landau, and M. Leese (2001). *Cluster Analysis.* Arnold, London.

Forchhammer, M. C. and J. Boomsma (1995). Foraging strategies and seasonal diet optimization of muskoxen in west greenland. *Oecologia 104*, 169–180.

Forchhammer, M. C. and J. Boomsma (1998). Optimal mating strategies in nonterritorial ungulates: a general model tested on muskoxen. *Behavioural Ecology 9*, 136–143.

Hanski, I. (2009). Metapopulations and spatial population processes. In S. A. Levin (Ed.), *The Princeton Guide to Ecology*, pp. 177 –185. Princeton University Press, Princeton.

Hardy, O. and X. Vekemans (2002). SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes 2*, 618–620.

Ho, L. P. and D. Stoyan (2008). Modelling marked point patterns by intensity-marked Cox processes. *Statistical Probability Letters 78*, 1194 – 1199.

Huang, F. and Y. Ogata (1999). Improvements of the maximum pseudo-likelihood estimators in various spatial statistical models. *Journal of Computational and Graphical Statistics 8*, 519–530.

Hubbell, S. P., R. Condit, and R. B. Foster (2005). Barro Colorado Forest Census Plot Data.

Hubbell, S. P., R. B. Foster, S. T. O'Brien, K. E. Harms, R. Condit, B. Wechsler, S. J. Wright, and S. L. de Lao (1999). Light gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science 283*, 283: 554–557.

Illian, J. B. and D. K. Henrichsen (2009). Gibbs point processes with fixed and random effects. *Environmentrics, DOI: 10.1002/env.1008*.

Illian, J. B., A. Penttinen, H. Stoyan, and D. Stoyan (2008). *Statistical Analysis and Modelling of Spatial Point Patterns.* Wiley, Chichester.

Johnson, C. R. and M. C. Boerlijst (2002). Selection at the level of the community: the importance of spatial structure. *Trends in Ecology & Evolution 17*, 83–90.

Killingback, T. and M. Doebeli (1996). Spatial evolutionary game theory: Hawks and doves revisited. *Proceedings of the Royal Society of London,. B 263*, 1135–1144.

Latimer, A. M., S. Banerjee, S. S, E. S. Mosher, and J. A. Silander Jr (2009). Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. *Ecology Letters 12*, 144154.

Law, R., J. B. Illian, D. F. R. P. Burslem, G. Gratzer, C. V. S. Gunatilleke, and I. A. U. N. Gunatilleke (2009). Ecological information from spatial patterns of plants: insights from point process theory. *Journal of Ecology 97*, 616–628.

Law, R., D. Purves, D. Murrell, and U. Dieckmann (2001). Causes and effects of small scale spatial structure in plant populations. In J. Silvertown and J. Antonovics (Eds.), *Integrating Ecology and Evolution in a spatial context*, pp. 21–44. Blackwell Science, Oxford.

Lawson, A. (1992). On fitting non-stationary Markov point process models on GLIM. In Y. Dodge and J. Whittaker (Eds.), *COMPSTAT: Proceedings 10th Symposium on Computational Statistics*, Volume 1, pp. 35–40. Physica Verlag.

Lunn, D., A. Thomas, N. Best, and D. Spiegelhalter (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing 10*, 325–337.

Møller, J., A. R. Syversveena, and R. P. Waagepetersen (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics 25*, 451–482.

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

Møller, J. and R. P. Waagepetersen (2007). Modern statistics for spatial point processes (with discussion). *Scandinavian Journal of Statistics 34*, 643–711.

Moore, B. D., I. R. Lawler, I. Wallis, C. M. Beale, and W. J. Foley (2010). Palatability mapping: a koala's eye view of spatial variation in habitat quality. *Ecology to appear.*

Myllymäki, M. and A. Penttinen (2009). Conditionally heteroscedastic intensity-dependent marking of log gaussian cox processes. *Statistica Neerlandica 63*, 450 – 473.

Naylor, M., J. Greenhough, J. McCloskey, A. Bell, and I. Main (2009). Statistical evaluation of characteristic earthquakes in the frequency-magnitude distributions of sumatra and other subduction zone regions. *Geophysical Research Letters 36, doi:10.1029/2009GL040460.*

Neyman, J. and E. L. Scott (1952). A theory of the spatial distrbution of galaxies. *Astrophysical Journal 116*, 144–163.

Ogata, Y. (1999). Seismicity analysis through point-process modeling: A review. *Pure and Applied Geophysics 155*, 471–507.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rajala, T. A. and J. B. Illian (2010). Graph-based description of mingling and segregation in multi-type spatial point patterns. *under submission.*

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields.* Chapman & Hall/CRC, Boca Raton.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B 71*, 1–35.

Schoenberg, F. (2005). Consistent parametric estimation of the intensity of a spatial-temporal point process. *Journal of Statistical Planning and Inference 128*, 79–93.

Stoyan, D. and P. Grabarnik (1991). Second-order characteristics for stochastic structures connected with gibbs point processes. *Mathematische Nachrichten 151*, 95–100.

Stoyan, D., W. Kendall, and J. Mecke (1995). *Stochastic Geometry and its Applications* (2nd ed.). John Wiley & Sons, London.

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

van Lieshout, M. (2000). *Markov point processes and their applications.* Imperial College Press, London.

Waagepetersen, R. and Y. Guan (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society, Series B 71*, to appear.

Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics 95*, 351–363.

Wiegand, T., S. Gunatilleke, N. Gunatilleke, and T. Okuda (2007). Analysing the spatial structure of a Sri Lankan tree species with multiple scales of clustering. *Ecology 88*, 3088–3012.