NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

# Fitting a log Gaussian Cox process with temporally varying effects — a case study

by

Janine B. Illian, Sigrunn H. Sørbye, Håvard Rue & Ditte K. Hendrichsen

PREPRINT
STATISTICS NO. 17/2010

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This preprint has URL `http://www.math.ntnu.no/preprint/statistics/2010/S17-2010.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.

# Fitting a log Gaussian Cox process with temporally varying effects – a case study

J. B. Illian[1,*], S. H. Sørbye[2], H. Rue[3] and D. K. Hendrichsen[4,5]

December 19, 2010

### Abstract

Integrated nested Laplace approximation (INLA) provides a fast and yet quite exact approach to fitting complex latent Gaussian models which comprise many statistical models in a Bayesian context, including log Gaussian Cox processes. This paper discusses how a joint log Gaussian Cox process model may be fitted to independent replicated point patterns. We illustrate the approach by fitting a model to data on the locations of muskoxen (*Ovibos moschatus*) herds in Zackenberg valley, Northeast Greenland and by detailing how this model is specified within the R-interface R-INLA. The paper strongly focusses on practical problems involved in the modelling process, including issues of spatial scale, edge effects and prior choice, and finishes with a discussion on models with varying boundary conditions.

1

## 1 Introduction

Integrated nested Laplace approximation (INLA) provides a fast and yet quite exact approach to fitting latent Gaussian models which comprise many statistical models, including models with temporal or spatial dependence structures (Rue et al., 2009). As a result, many complex models that previously required the use of time-consuming Markov chain Monte Carlo (MCMC) calculations can be fitted fast and conveniently. Log Gaussian Cox processes, a particularly flexible class of spatial point process models are a special case of latent Gaussian models. Rue et al. (2009), Illian et al. (2010) and Illian and Rue (2010) show that complex point process models, including hierarchically marked point processes may conveniently be fitted with INLA. Standard approaches to parameter estimation for complex models based on MCMC, for example, would be very cumbersome and computationally prohibitive.

[1] 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

[2] Department of Mathematics and Statistics, University of Tromsø, N-9037 Tromsø, Norway; sigrunn.sorbye@uit.no

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

[4] Department of Arctic Environment, National Environmental Research Institute, Aarhus University, DK-4000 Roskilde, Denmark

[5] Norwegian Institute for Nature Research, N-7047 Trondheim, Norway; ditte.hendrichsen@nina.no

[*] Corresponding author

Fitting spatial point process models to some spatial patterns is computationally intensive due to – amongst other things – the large number of individual points in the data set (Burslem et al., 2001; Waagepetersen, 2007; Waagepetersen and Guan, 2009; Law et al., 2009). Here, we consider a rather different situation. In some applications difficulties arise since point patterns with only a very small number of points can be collected, due to logistic limitations (e.g. for reasons of accessibility). These patterns are sometimes too small to justify the modelling of a single pattern. However, if replicates exist, a joint model of all replicates with a factor that accounts for variability among replicates caused by different conditions on different days may be more suitable. Mixed effect models for replicated point patterns have recently been considered in a frequentist approach for Gibbs processes (Illian and Hendrichsen, 2010). In that approach, parameter estimation was based on the pseudolikelihood of a Gibbs process as well as maximum quasi-likelihood optimisation. Here we discuss how data with a similar structure may be modelled with a log-Gaussian Cox process.

The data that we have available have been collected at different time points and hence we consider a complex point process model with a temporally varying effect and construct a model that may then be fitted with INLA. This approach provides us with more general modelling facilities that allow us to fit several models of varying complexity and compare their suitability for a given data set. We show in detail how this model is specified in practice, within the R-interface R-INLA (available at `www.r-inla.org`), and discuss issues involved in fitting the model to a data set derived from an ecological field study. We use a data set detailing the locations of muskoxen herds in Greenland (Illian and Hendrichsen, 2010; Schmidt et al., 2010) to illustrate the approach. In addition, we discuss issues of fitting log Gaussian Cox processes to point patterns that have been observed in observation windows where some of the edges are real edges (Illian et al., 2008) which is particularly common in animal studies.

# 2 Modelling approach

## 2.1 The data

We consider a situation where $T$ point patterns $\mathbf{x}_1, \ldots, \mathbf{x}_T$ have been observed in an observation window $S$ at different points in time, $t = 1, \ldots, T$. The objects represented by the points may be considered independent of the objects observed at other points in time conditional on common observed (spatial) covariates $z_1, \ldots, z_p$ and other unobserved covariates. The pattern may differ between the different time points as a result of observed and unobserved influences specific to a given point in time. In order to fit a single model to all replicates, ”time” may be treated as a factor in a point process model. For illustration, the approach is applied to a data set detailing the spatial locations of muskoxen (*Ovibos moschatus*) herds in Zackenberg valley, Northeast Greenland, $74°30'$N – $21°00'$W (hereafter referred to as ”muskoxen data”) at different points in time within several years (Meltofte and Berg, 2004). See Figure 1 A and B for the location of the study area within Greenland and Figure 1 C for a map of the area. The muskoxen at Zackenberg have been studied since 1996 as part of a large-scale monitoring programme in the area (Meltofte et al., 2008). The census area covers 45 square km, ranging from sea level to 600 masl. The spatial locations of muskoxen herds have been recorded weekly during the summer months, July and August, occasionally also in June. During the censuses, all herds within the census area have been mapped to the nearest 100 metres (Schmidt et al., 2010), then approached and identified to age and sex following Olesen and Thing (1989). The categories are calves, yearlings, two-year, three-year and four-years or older. There is

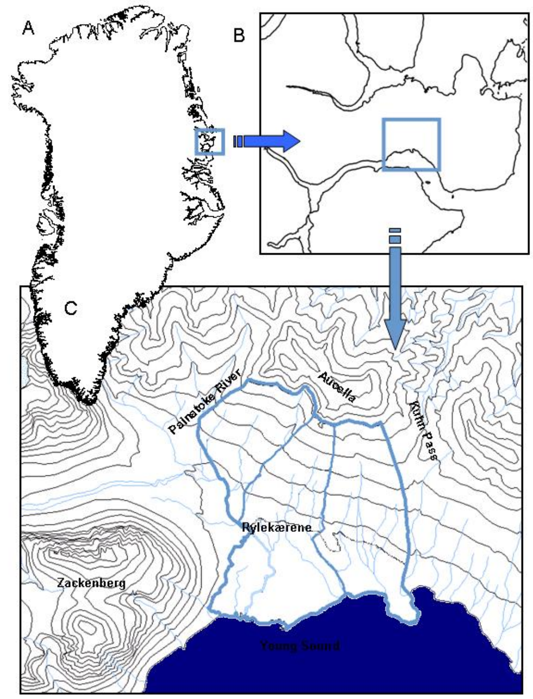


Figure 1: Map indicating the location of the Zackenberg valley within Greenland (A), within the area sourrounding the valley (B) and the structure of the Zackenberg valley and its boundaries (C).

an interest in analysing the spatial distribution of muskoxen in relation to spatial covariates such as altitude or an index of vegetation productivity, the normalized differential vegetation index (NDVI).

In this paper we analyse a subset of the muskoxen data corresponding to the years 2005–2007. The observed locations of the muskoxen herds across all time points during these years are illustrated in Figure 2, indicating that the intensity of the patterns seems to be particularly high in one specific area that forms a diagonal across the plot. Illian and Hendrichsen (2009) model a similar data set with a Gibbs process and fit the model by approximating the pseudolikelihood based on a generalised linear mixed model with Poisson outcome. They include "time" as a random effect applying the approximate Berman-Turner device (Baddeley and Turner, 2000) in a frequentist approach. Here we specify a log-Gaussian Cox process model and fit it in a Bayesian context, an approach that allows more flexibility and makes the fitting and comparison of several models of varying complexity feasible.

## 2.2 Model fitting in INLA

A log-Gaussian Cox process is a hierarchical Poisson process with random intensity $\Lambda_t(s) = \exp\{\eta_t(s)\}$, where $\eta_t = \{\eta_t(s) : s \in \mathbf{R}^d\}$ denotes a Gaussian field. This type of model is a special case of the more general class of latent Gaussian models, for which deterministic

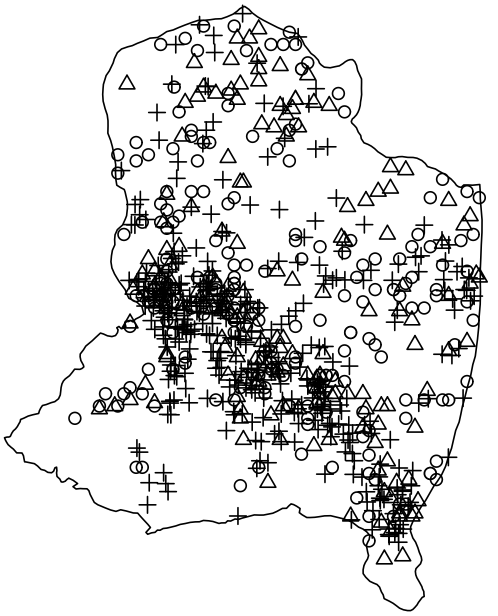


Figure 2: Plot of the locations of muskoxen herds observed in Greenland on 26 days during the summer in 2005-2007, distinguished by year; circles: 2005, triangles: 2006 and crosses: 2007.

Bayesian inference can be performed using the INLA-methodology, see Rue et al. (2009). We restrict the analysis to two-dimensional patterns and fit a log-Gaussian Cox process to a specific point pattern $\mathbf{x_t}$ discretising the observation window $S$ into $N$ grid cells $\{s_i\}_{i=1}^{N}$, where each cell has area $|s_i|$. For each time point $t = 1, \ldots, T$, let $y_{ti}$ denote the observed number of points in grid cell $s_i$. Conditional on the intensities $\Lambda_t(s_i) = \exp\{\eta_t(s_i)\}$, the joint pattern of replicates can be described as a superposition $\mathbf{x} = \bigcup_{t=1,\ldots,T} \mathbf{x}_t$ of independent Poisson processes, where

$$y_{ti}|\eta_t(s_i) \sim \mathrm{Po}(|s_i| \exp(\eta_t(s_i))). \tag{1}$$

We assume that the log-intensity of the Poisson processes can be described by a linear predictor

$$\eta_t(s_i) = \beta_0 + \beta_f + \sum_{\alpha} \beta_\alpha z_{t\alpha}(s_i) + \sum_{\gamma} f_\gamma(c_{t\gamma}(s_i)), \tag{2}$$

in which the off-set $\beta_0$ represents an intercept common to all time points while the factor $\beta_f$ accounts for variation in intensity across different time points. The linear effects of (environmental) covariates can be incorporated as fixed factors $\{z_{t\alpha}(.)\}$, which may or may not vary with time. We also include potentially smooth effects of covariates $\{c_{t\gamma}(.)\}$, in which the functions $\{f_\gamma(.)\}$ are estimated based on all replicates.

Applying the INLA-methodology, the log-intensity in (2) is estimated assigning Gaussian priors to all components of the random latent field $\zeta = \{\beta_0, \beta_f, \{\beta_\alpha\}, \{f_\gamma(.)\}\}$, as the resulting model can be viewed as a latent Gaussian model. The posterior marginals of each element $\zeta_j$ of the latent field are given by

$$\pi(\zeta_j \mid \mathbf{y}) = \int \pi(\zeta_j \mid \theta, \mathbf{y})\pi(\theta \mid \mathbf{y})d\theta, \tag{3}$$

where the vector $\theta$ denotes the hyperparameters of the model. Here, $\theta$ includes the parameters used in defining prior distributions for the precision (inverse variance) of the Gaussian priors, in which

$$\pi(\theta_j|\mathbf{y}) = \int \pi(\theta|\mathbf{y})d\theta_{-j}. \tag{4}$$

The INLA-methodology applies a nested formulation to estimate (3)–(4), combining analytical approximations with numerical integration, see Rue et al. (2009) for a thorough description. The computations are based on approximating the full conditional of the latent field by a Gaussian distribution $\tilde{\pi}_G(\zeta|\theta, \mathbf{y})$. Further, a Laplace approximation is used to estimate the posterior of the hyperparameters, given by

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

where for each $\theta$, $\zeta^*(\theta)$ is the mode of the Gaussian appproximation. Estimates of the marginals $\pi(\zeta_j|\theta, \mathbf{y})$ in (3) can be found either using a Laplace or a simplified Laplace approximation. Alternatively, the marginals can be estimated using a Gaussian approximation derived from $\tilde{\pi}_G(\zeta|\theta, \mathbf{y})$. Although the Gaussian approximation might provide some inaccuracies in estimating the marginals (Rue and Martino, 2007), this approach is used here to speed up calculations. Using numerical integration with respect to $\theta$, the resulting approximation to (3) is given by

$$\tilde{\pi}(\zeta_j \mid \mathbf{y}) = \sum_k \tilde{\pi}_G(\zeta_j \mid \theta_k, \mathbf{y})\tilde{\pi}(\theta_k \mid \mathbf{y})\Delta_k, \tag{5}$$

where $\tilde{\pi}(\theta_k \mid \mathbf{y})$ is found by numerical integration in (4) and $\Delta_k$ denotes the area weight corresponding to $\theta_k$.

### 2.3 Specifications for the muskoxen data set

In fitting (2) to the given subset of the muskoxen data, we have considered two observed environmental covariates, the altitude and the normalized differential vegetation index (NDVI). The vegetation index is not available for all time points, including the years 2006 and 2007. In various submodels for individual years this covariate has turned out to be non-significant and is hence not included in the final model.

To account for small scale spatial variation not caused by environmental covariates but by inter-individual (or here: inter-herd) interactions, we apply the ideas in Illian et al. (2010) and introduce a constructed covariate in (2), that relates each midpoint of cell $s_i \in S$ to pattern points in the neighbourhood. Specifically, for grid cell $s_i$ we define the constructed covariate

$c_t(s_i)$ to be the Euclidean distance from the midpoint $m_i$ of cell $s_i$ to the nearest point $x_{tj}$ in the pattern outside $s_i$ at time point $t$, i.e.

$$c_t(s_i) = \min_{x_{tj} \in \mathbf{x}_t} (|m_i - x_{tj}|), \qquad (6)$$

where $|.|$ denotes the Euclidean distance. However, for each time point the constructed covariate for cell $s_i$ is defined to be missing if the distance $c_t(s_i)$ is larger than the minimum Euclidean distance between $m_i$ and the border, see section 2.5 for remarks on potential edge effects. The constructed covariate reflects small-scale spatial interaction (attraction or repulsion), see Section 2.4 for further discussions on the choice of the constructed covariate. It is incorporated in (2) using a smooth function $f_{cc}(.)$ as the functional relationship between the intensity and the constructed covariate may not be linear. Specifically, the function $f_{cc}(.)$ is modelled as a first-order random walk process, using a gamma prior for the precision parameter. Similarly, large-scale spatial variation is included in (2) as a smooth function in space $f_{spat}(.)$ for all the grid cells and across all time points. Here, the spatial function is specified as a second-order random walk on a lattice, as this spatial model is supported by R-INLA. A gamma prior is used also for the precision parameter of the spatial model, see Section 2.4 and 3.1 for a discussion on appropriate choices for the prior parameters.

The resulting model for the muskoxen data set can be summarised as

$$\eta_t(s_i) = \beta_0 + \beta_{year} + \beta_1 z_1(s_i) + f_{day}(t) + f_{cc}(c_t(s_i)) + f_{spat}(s_i), \quad t = 1, \ldots T, \ i = 1, \ldots, N, \ (7)$$

where $z_1(.)$ denotes the altitude for the different grid cells. We have modelled the temporally varying effect by a factor $\beta_{year}$ accounting for the yearly increase in mean intensity observed for the years 2005-2007. In addition, we include a smooth function $f_{day}(t)$ assuming independent Gaussian observations, which accounts for random variation in intensity on different days. To ensure identifiability of the intercept, all of the smooth functions are constrained to sum to zero.

## 2.4 Issues of spatial scale – prior choices

In an analysis of a spatial pattern, it is crucial to bear in mind the spatial scales that are relevant for a specific spatial data set. Here we assume that social behaviour among the herds operates at a local spatial scale and that the association with environmental covariates operates on the scale of the variation in these covariates and hence often on a larger spatial scale.

The large-scale spatial effect in (7) is included as a spatially structured error term to account for any spatial autocorrelation unexplained by covariates in the model. The choice of the gamma prior for the precision of the spatially structured effect determines the smoothness of the spatial effect and, through this, the spatial scale at which it operates. To avoid overfitting, and in order to obtain a model describing a generally interpretable trend we choose the prior so that the spatial effect operates at a similar spatial scale as the covariate. This ensures that the spatially stuctured effect does not operate on a smaller scale than the covariate as it would otherwise be likely to explain the data better than the covariates, rendering the model rather pointless. We approach this by repeatedly fitting a simple model (see Section 3.1), comparing the estimated spatial effect to a plot of the covariate.

If small scale inter-individual spatial behaviour is of specific interest in an application it may be modelled by the constructed covariate to account for local spatial behaviour. The muskoxen herds are likely to interact mainly with the herd that is closest to them (Dice, 1952; Clark and Evans, 1955) and so the distance to the nearest neighbour has been chosen here as a constructed covariate accounting for local behaviour. This is of interest since the model may be used to reveal the relevant distances at which the herds interact. Again there is a danger of overfitting, especially since the constructed covariate is estimated directly from the point pattern. Illian et al. (2010) discuss practicalities of using a spatial constructed covariate and point out the importance of the choice of the covariate and its relevance to a specific application. The choice of grid size is also linked to issues of spatial scale. Here we apply the same grid and grid resolution as used in the data collection of locations of the muskoxen herds.

### 2.5 Edge effects and boundary conditions

Spatial point pattern data are usually observed in a finite observation window only. Since there typically is no data about the behaviour of the pattern outside the window, edge effects occur and often have to be accounted for (Illian et al., 2008), in particular in the context of stationary point processes. Edge effects have only a minor impact on large point pattern data sets such as the rainforest data discussed in Rue et al. (2009).

However, in smaller data sets, including the data considered here, edge effects are certainly relevant. Here, we consider a data set observed within a polygonal window assuming that $\Lambda_t(.)$ is defined in $\mathbf{R}^2$, i.e. that if the pattern had been observed beyond any of the edges it would have the same properties as inside the observation area. For simplicity, we currently assume that all edges are of the same quality, i.e. that they have been arbitrarily chosen and have no impact on the spatial pattern. In applications this is not necessarily the case as the boundary may reflect a true barrier beyond which the conditions are very different. This might imply that the pattern does not continue in exactly the same way and that, more importantly, the barrier has a direct influence on the pattern near the edge. We consider these issues in the discussion indicating potential approaches to solving these.

It is necessary to account for potential edge effects in the constructed covariate leading to an overestimation for cells close to the boundary as there might be points outside the observation window that are closer to a specific cell than its nearest neighbour within the window. In the calculation of the constructed covariate we hence exclude all cells for which the distance of its midpoint to the nearest point in the pattern is larger than its minimum distance to the boundary.

## 3 Results

To illustrate the modelling approach we consider a subset of the muskoxen dataset, consisting of the patterns collected on $T = 26$ different days. The observation period included weekly observations for a total of 10, 7 and 9 days during the summers of 2005-2007, respectively. Across the years, the subset consists of a total of 754 observed muskoxen herds. These are likely to include repeated observations of the same muskoxen herds at different times points. Since information on herd identity is not available we cannot account for this in the model.

We divide the observation window $S$ into a total of $N = 4533$ grid cells, each having an area equal to 0.01 km$^2$. Technically, the joint model for all locations and replicates is run in R-INLA stacking all observations in one vector $\mathbf{y} = \{y_{ti}, t = 1, \ldots, T,\ i = 1, \ldots, N\}$, having corresponding vectors for the other terms in the model. Consequently, for the given subset of the data, the length of these vectors is 117858. We fit the model in (7) to the data and apply the deviance information criterion (DIC) (Spiegelhalter et al., 2002) to assess the importance of the different terms in the model.

### 3.1 Prior choices

We initially fit a very simple model to the data where we explain the location of the herds purely based on a common spatial effect for all time points. This is done to assess the resolution of the spatial effect and simplify prior choice. We fit the model repeatedly using different values for the shape parameter of the gamma prior and compare the resulting estimated spatial effects. We deliberately choose the prior for the spatial effect manually as assessing the model based on DIC would encourage overfitting. Clearly, prior choices for the parameters could be done using the full model in equation (7) as well, but this would be more time-consuming.

We apply the simple model

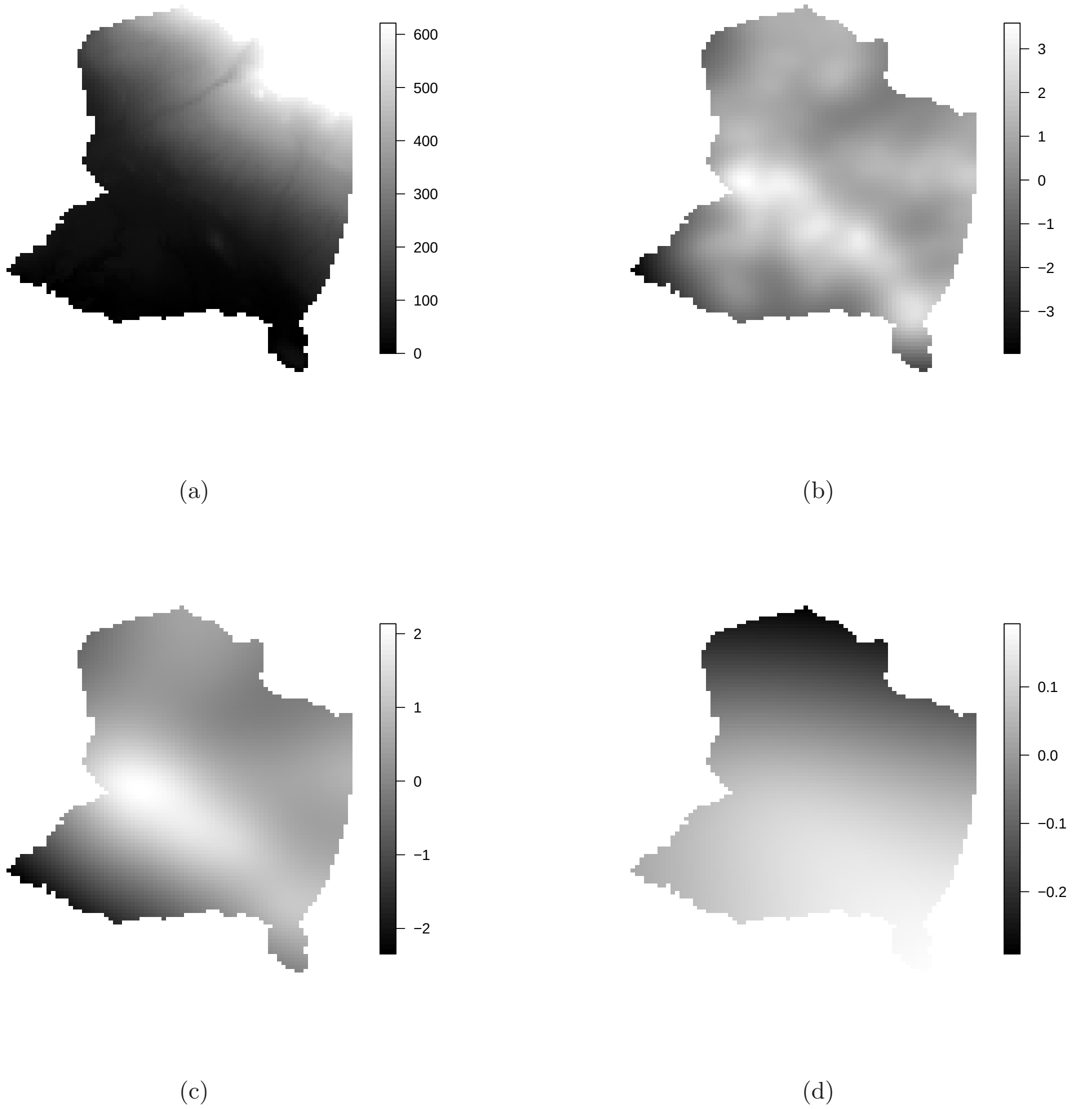
$$\eta_t(s_i) = \beta_0 + f_{spat}(s_i), \tag{8}$$

which takes about 15 seconds to run for each of the different parameter choices. The call in R-INLA to fit this simple model is

```
> formula = y ~ 1 + f(i.spat, model = "rw2d", nrow = nrow, ncol = ncol,
                 param = prior.spat)
> result = inla(formula, data=data,family = "poisson",verbose =TRUE,
                 control.inla = list(strategy = "gaussian"),
                 control.compute=list(dic=TRUE))
```

The formula accounts for a common intercept, while `f(i.spat)` represents the smooth function of the spatial effect $f_{spat}(s_i)$. The spatial model is specified as a second-order random walk on a lattice (`rw2d`) having dimension `ncol`×`nrow`, chosen to cover the observation window. The shape and scale parameter of the gamma distribution used as a prior for the precision of the spatial model, is given as the parameter vector `prior.spat`. Clearly, the function call to INLA bears strong resemblance to that for other models in `R`, such as `lm` or `glm`, and similar to the latter, the model family has to be given as `family = "poisson"`. Due to the size of the final model we apply the Gaussian approximation to estimate the marginals of each component of the latent field (`strategy=gaussian`). Also, we specify that the DIC value should be calculated (`dic=TRUE`).

Figure 3 illustrates how the spatial effect varies with different prior choices for the precision parameter in the spatial model of $f_{spat}(.)$. We compare the smoothness of the estimated surface to that of the observed altitude (Figure 3 (a)). Applying prior parameters (1, 0.001) clearly results in a low degree of smoothing. The spatial effect reflects very local behaviour (clustering at a smaller scale than the empirical covariate) and hence potentially yields a model that is overfitted to the given observations, see Figure 3 (b). In Figure 3 (c), a moderate degree of smoothing is obtained using the parameters (40, 0.001), in which the spatial effect clearly indicates the higher intensity of points in the central area that is also obvious from the plot of the pattern. A similar degree of smoothness results for a quite wide range of prior parameters and hence this prior has been chosen for the analysis of the full model. Figure 3 (d) shows the estimated spatial effect using the parameters (100, 0.001), leading to a very smooth spatial effect which seems too strong for the current application.



(a) (b)

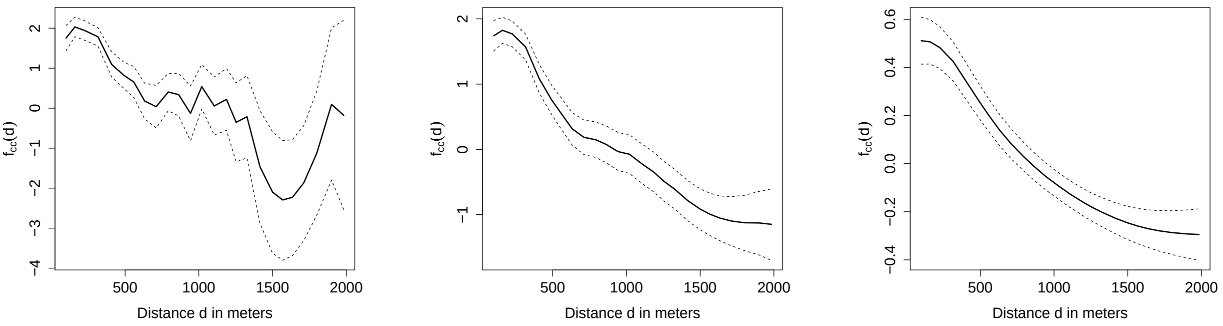
(c) (d)

Figure 3: The observed altitude (a). The estimated spatially structured effect using the parameters $(1, 0.001)$ (b), $(40, 0.001)$ (c) and $(100, 0.001)$ (d) for the gamma prior in model (8).

We take a similar approach to determine the prior for the constructed covariate and again fit a simple model to the data and observe how the smoothness of the estimated function varies with the choice of prior parameters. The model we fit here contains only the constructed covariate and a common intercept, such that

$$\eta_t(s_i) = \beta_0 + f_{cc}(c_t(s_i)). \tag{9}$$

Figure 4 (a) indicates that using a gamma prior with parameters (2, 0.01), results in a very wiggly curve. This is likely to be due to uninformative noise and stronger smoothing of the curve is required. The choice of prior parameters (60, 0.01), yields a much smoother curve, see Figure 4 (b). The estimated curve looks very similar for quite a wide range of prior parameters and this set of priors has been chosen for the final analysis. Finally, the effect of choosing an extreme prior becomes clear in Figure 4 (c) with parameters (2000, 0.01) which results in a very smooth and much flatter curve.



(a) (b) (c)

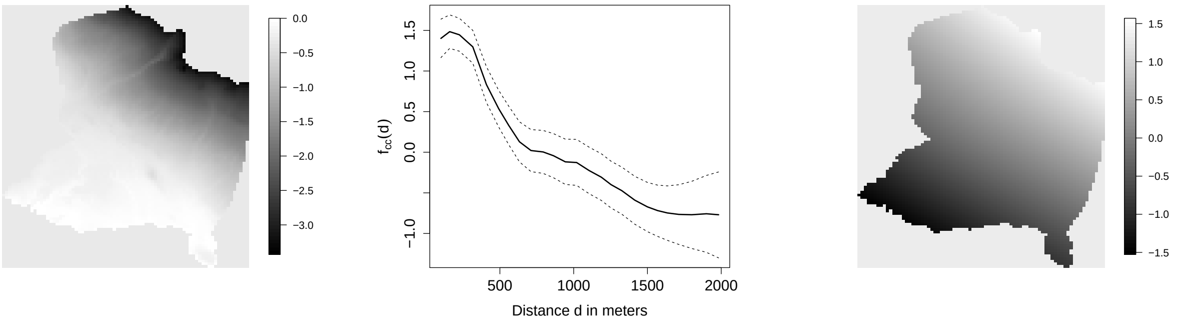
Figure 4: The estimated effect of the constructed covariate including 95% pointwise credible intervals, using the parameters $(2, 0.01)$ (a), $(60, 0.01)$ (b) and $(2000, 0.01)$ (c) for the gamma prior in model (9).

### 3.2 Model selection

Informed by the fit of the simple model in equation (8) we choose the prior parameters for the spatial effect as `prior.mod`= $(40, 0.001)$ and fit the full model in (7) to the muskoxen dataset. The formula specification to run this model in R-INLA has the following format,

```
> formula = y ~ year.factor + cov.alt + f(day, model = "iid") +
        f(inla.group(cov.cc), model = "rw1", param = prior.cc) +
        f(i.spat, model = "rw2d", nrow = nrow, ncol = ncol, param = prior.spat)
```

The model took 213 seconds to run on a 12 core 2.33 GHz machine and the DIC for this model is 8515. By specifying the year factor with three levels, we estimate a common intercept $\hat{\beta}_0 = -5.108$. The additionally estimated coefficients for the years 2006 and 2007 are $\hat{\beta}_{year} = (0.108, 0.505)$, accounting for the increase in observed muskoxen herds during these years. Having accounted for the annual effect, the estimated temporal effect as a function of days was seen to be negligible. The altitude (specified by `cov.alt`), has a significant negative effect on the intensity of the points with an estimated mean equal to $-0.006$ with a 95% credible interval given by $(-0.007, -0.004)$, see Figure 5 (a) for the resulting estimated effect of altitude. The plot of the constructed covariate clearly indicates local clustering (Figure 5 (b)). The estimated spatial effect (Figure 5 (c)) shows a smooth and non-constant surface. This surface might suggest other factors that have not been considered in the model (or in the overall study) but may influence the pattern. Most markedly, in the south west corner fewer herds appear to have been observed than the covariate altitude can explain.



(a) (b) (c)

Figure 5: The estimated effect of altitude (a), the estimated function of the constructed covariate including 95% pointwise credible intervals (b) and the estimated spatially structured effect (c) for the full model in (7).

In order to assess the effect of the various terms in (7), we repeatedly fit submodels of the full model leaving out one of the terms at a time and comparing the model fit based on the DIC. The results are summarized in Table 1.

In concrete applications several covariates are often available and the DIC may be used to identify the best model in terms of the covariates considered. We notice that DIC for a model without the year and day effects is clearly higher than the full model supporting the inclusion of the time-varying effect in the model. Likewise, the model without the altitude covariate leads to a DIC of 8558, clearly indicating that the model may be improved by including this covariate. Table 1 also illustrates that both the inclusion of a constructed covariate and a large-scale spatial effect improves the model as the spatial autocorrelation cannot be explained fully by the covariate.

# 4 Discussion

In many areas of science, rapidly improving technology facilitates data collection these days. Ecologists in particular, have become aware of the importance and relevance of spatial information for an understanding of population dynamics. For these reasons, spatially explicit data sets have become increasingly available in ecology as well as many other areas of science, including geosciences, molecular genetics, evolution and game theory 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 and accessible to applied scientists by facilitating the fitting of suitable models to help provide answers to concrete scientific questions. In order to address this issue, this paper illustrates in detail the process of fitting a spatial point process model to a realistically complex data set using modern model fitting methodology. This provides non-specialists with an illustration of a modelling approach that makes the fitting of complex spatial point process models accessible.

In the specific case considered here, a simple analysis of each of the pattern that does take the spatially explicit information into account would not be sufficient. The small size of the individual patterns would render this analysis rather uninformative such that a complex joint model of several subpatterns has to be considered in order to gain an understanding of the system. In particular, to be able to fit this complex model we apply the deterministic Bayesian INLA-methodology, to fit a model to a replicated point pattern allowing for variation among the replicates by considering time-varying effects. The INLA approach speeds up parameter estimation substantially such that we can fit several models within feasible time and use model comparison methods to identify the most suitable model.

## 4.1 Spatial scale

In general, the issue of spatial scale has to be taken seriously when fitting spatial models, irrespective of whether these are spatial point process models or other spatial models. Clearly, any model can only explain mechanisms that operate on the spatial scale that the data have been collected. In particular, in the example considered here, we find that altitude is a better predictor for the spatial distribution of muskoxen than vegetation greenness (measured as NDVI). That NDVI does not adequately explain muskoxen spatial distribution is likely to be related to spatial scale, where NDVI does not adequately capture the local variation in plant communities, growth and quality within a single valley. NDVI has become increasingly popular as a measure of primary production. It is highly correlated with herbivore distribution at larger scales (Pettorelli et al., 2006), and has become a useful tool in explaining herbivore distribution and movement in areas where direct measurements of vegetation variables such as biomass and quality is not available. However, it is less suitable in explaining habitat choice at finer scales where habitat choice is determined by several factors, including nutrient contents of foraging plants, the availability of salt licks, accessibility as well as behavioural interactions between animals (Bro-Jorgensen et al., 2008). Muskoxen have a high preference for Salix, and tend to select snow bed and moist meadows such as fens during summer time (Thing et al., 1987). Whereas moist meadows have a high biomass and a high primary production, snow beds are characterised by sparse vegetation of primarily willow, and a late green-up. NDVI is therefore unlikely to pick up the small scale variation in vegetation distribution. Altitude, by contrast, has a strong influence on vegetation communities, and may more accurately capture the variations in vegetation distribution, phenology and green-up (Mysterud et al., 2001), factors which are likely to be of more importance to muskoxen than the more crude measure of primary productivity.

## 4.2 Model comparison and assessment

The modelling approach provides methods of model comparison through the DIC so that it was possible to identify the most suitable model for the specific data set and to avoid over-parameterisation. In addition, the spatially structured effect may be used for model assessment as it reveals spatial structures in the data that have not been explained by the model and potentially points to covariates that may be included in an improved model.

In the example considered here, the model was unable to explain the low number of muskoxen herds in the south south-western section of the census area. In Zackenberg, this area together with the south-eastern peninsula and the slopes above 500 masl are characterised by a mosaic of vegetation types, but with a proportionally higher amount of low productive fell-fields and barren lands (Fredskild and Mogensen, 1997; Bay, 1998), and these areas are likely to be less preferred by muskoxen. By contrast, the diagonal band with high muskoxen densities correspond with low-lying moist fens and grassland, vegetation types which are highly preferred by muskoxen (Thing et al., 1987; Forchhammer et al., 2005). The covariates included in the current model clearly do not reflect this much detail on the type of the local vegetation.

### 4.3 Boundary conditions

The current model assumes that all edges of the observation window were artificial edges determined by the choice of the observation window and hence that the pattern was a sample from a stationary process that continues in the same way beyond the boundaries. This assumption is sensible in many data examples, where the observation window is a plot within a larger area with similar characteristics. For instance, the examples discussed in (Rue et al., 2009) and Illian et al. (2010) consider point patterns in rectangular observation window whose edges have been arbitrarily chosen. However, strictly speaking, this assumption does not hold for the concrete example data set. For instance, the southern boundary is formed by the sea and muskoxen herds would never be observed beyond this border. In some data sets a border like this might constitute a reflective boundary causing a higher intensity of animals close to it. In other cases, the animals might actively avoid the vicinity of this boundary.

Muskoxen are highly motile and in principle capable of utilising the entire census area, including the area immediately bordering the shoreline. However, their incitement to approach the boundary may depend on a number of factors, of which the distribution and quality of the vegetation is most obvious. The distribution of plant communities is strongly affected by factors such as soil quality, water and nutrient availability and exposure, all of which may be influenced by distance to the sea. It is therefore relevant to consider covariates when interpreting the effect of the boundary. In addition to the absolute nature of the shoreline as a boundary, other sections of the boundary may be characterised by reduced connectivity across the border, without being an absolute obstacle. For example, large sections of the left hand (western) and right hand (eastern) boundaries are demarcated by rivers which are passable to muskoxen, but where the degree of connectivity may vary with season (influencing water discharge and break-up of ice) and with the age or condition of the animals attempting to cross.

Situations where spatial point patterns have been observed within an observation window with a true edge have been discussed in the context of finite point processes (Illian et al., 2008). However, the given data set cannot be treated as a finite point process since it is not a truly finite process. The animals can move in and out of the area at some parts of the boundary and, more importantly, these boundaries have been arbitrarily chosen. Hence it is useful to assume that the pattern continues in the same way beyond some part of the boundary and that the edge has no effect on the pattern. In other words, a more realistic model of the data than the model that we currently consider would consider the varying nature of the boundary. This issue is likely to be relevant in many studies where spatial data on a large spatial scale have been observed. For practical reasons (e.g. accessibility or ease of data collection) observation areas are often chosen to line up with existing natural borders which might directly impact on the spatial behaviour within the resulting observation window. In some applications, the influence of the edge on the pattern might even be of primary interest to a study. Approaches that allow the fitting of complex models to data sets with complicated boundary structures are highly relevant not only to the specific data set discussed here but also to many other data sets detailing the locations of animals, such as marine mammals (Sveegaard et al., 2010).

Currently, it is not possible to account for different types of edges in INLA. However, the approaches introduced in Lindgren et al. (2010) can be used to incorporate these in a model in a straight forward way based on stochastic partial differential equations where different boundary conditions may be specified. These new developments will soon be incorporated as models in INLA and may then be applied to data sets with varying types of boundaries.

## 5 Acknowledgements

We would like to thank Mads C. Forchhammer, Niels Martin Schmidt and Mikkel Tamstorf from Greenland Ecosystem Monitoring, National Environmental Research Institute, Aarhus University for making the data available to us. We would also like to thank Finn Lindgren and Daniel Simpson, University of Trondheim for many fruitful discussions. Some of the computations for this paper were run on the RINH/BioSS Beowulf cluster at the University of Aberdeen, Rowett Institute of Nutrition and Health (http://bioinformatics.rri.518 sari.ac.uk). We thank Tony Travis for his generosity and assistance.

## References

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

Bay, C. (1998). *Vegetation mapping of Zackenberg valley, Northeast Greenland.* Danish Polar Center & Botanical Museum, University of Copenhagen.

Bro-Jorgensen, J., M. E. Brown, and N. Pettorelli (2008). Using the satellite-derived normalized difference vegetation index (NDVI) to explain ranging patterns in a lek-breeding antelope: the importance of scale. *Oecologia 158*, 1777–182.

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

Clark, P. J. and F. C. Evans (1955). Distance to nearest neighbor as a measure of spatial relationship in populations. *Ecology 35*, 445–453.

Dice, L. R. (1952). Measure of the spacing between individuals within a population. *Contrib. Lab. Vert. Biol. Univ. Mich. 55*, 1–23.

Forchhammer, M. C., E. Post, T. B. B. Berg, T. T. Høye, and N. M. Schmidt (2005). Local-scale and short-term herbivore-plant spatial dynamics reflect influences of large-scale climate. *Ecology 86*, 2644–2651.

Fredskild, B. and G. Mogensen (1997). *ZERO line. Final Report 1997. A description of the plant communities along the ZERO line from Young Sund to the top of Aucellabjerg and the common plant communities in the Zackenberg valley, Northeast Greenland.* Greenland Botanical Survey & Botanical Museum, University of Copenhagen (36 pp.).

Illian, J. B. and D. K. Hendrichsen (2010). Gibbs point processes with mixed effects. *Environmentrics 21*, 341–353.

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

Illian, J. B. and H. Rue (2010). A toolbox for fitting complex spatial point process models using integrated laplace transformation (inla). *technical report, Norges Teknisk-Naturvitenskapelige Universitet Trondheim, preprint statistics, No. 06/2010*.

Illian, J. B., S. H. Sørbye, and H. Rue (2010). A toolbox for fitting complex spatial point process models using integrated nested laplace transformation (inla). *submitted*.

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.

Lindgren, F., H. Rue, and J. Lindström (2010). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. *technical report, Norges Teknisk-Naturvitenskapelige Universitet Trondheim, preprint statistics, No. 05/2010*.

Meltofte, H. and T. B. B. Berg (2004). *Zackenberg Ecological Research Operations. BioBasis. Conceptual design and sampling procedures of the biological programme of Zackenberg Basic.* (7th ed.). National Environmental Research Institute, Department of Arctic Environment.

Meltofte, H., T. R. Christensen, B. Elberling, M. C. Forchhammer, and M. Rasch (2008). *High-Arctic Ecosystem Dynamics in a Changing Climate. Ten Years of Monitoring and Research at Zackenberg Research Station, Northeast Greenland.* Advances in Ecological Research, 40, Elsevier.

Mysterud, A., R. Langvatn, N. G. Yoccoz, and N. C. Stensetn (2001). Plant phenology, migration and geographical variation in body weight of a large herbivore: the effect of a variable topography. *Journal of Animal Ecology 70*, 915–923.

Olesen, C. R. and H. Thing (1989). Guide to field classification by sex and age of the muskox. *Canadian Journal of Zoology 67*, 1116–1119.

Pettorelli, N., J. M. Gaillard, A. Mysterud, P. Duncan, N. C. Stenseth, D. Delorme, G. V. Laree, C. Togo, and F. Klein (2006). Using a proxy of plant productivity (ndvi) to find key periods for animal performance: the case of roe deer. *Oikos 112*, 565–572.

Rue, H. and S. Martino (2007). Approximate bayesian inference for hierarchical gaussian markov random fields models. *Journal of Statistical Planning and Inference 137.*

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B 71*, 319–392.

Schmidt, N. M., T. B. G. Berg, and H. Meltofte (2010). Biobasis, conceptual design and sampling procedures of the biological monitoring programme within zackenberg basic. *The National Environmental Research Institute, Department of Arctic Environment, Aarhus University.*, 109 pp.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. V. der Linde A (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B 64*, 583–616.

Sveegaard, S., J. Teilmann, J. Tougard, R. Dietz, K. Mouritzen, G. Desportes, and U. Siebert (2010). High-density areas for harbor porpoises (*Phocoena phocoena*) identified by satellite tracking. *Marine Mammal Science DOI: 10.1111/j.1748-7692.2010.00379.x.*

Thing, H., D. R. Klein, K. Jingfors, and S. Holt (1987). Ecology of muskoxen in jameson land and northeast greenland. *Holarctic Ecology 10*, 95–13.

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.

| Model | Formula for $\eta_t$ | DIC |
|---|---|---|
| Full model | $\beta_0 + \beta_{year} + \beta_1 z_1(.) + f_{day}(.) + f_{cc}(.) + f_{spat}(.)$ | 8515 |
| Without temporal effect | $\beta_0 + \beta_1 z_1(.) + f_{cc}(.) + f_{spat}(.)$ | 8548 |
| Without altitude | $\beta_0 + \beta_{year} + f_{day}(.) + f_{cc}(.) + f_{spat}(.)$ | 8558 |
| Without constructed covariate | $\beta_0 + \beta_{year} + \beta_1 z_1(.) + f_{day}(.) + f_{spat}(.)$ | 8795 |
| Without spatial effect | $\beta_0 + \beta_{year} + \beta_1 z_1(.) + f_{day}(.) + f_{cc}(.)$ | 8561 |

Table 1: DIC values for different fitted models for the muskoxen data.