A CASE STUDY ON POINT PROCESS MODELLING IN DISEASE MAPPING

Viktor Beneš[12], Karel Bodlák[1], Jesper Møller[3] and Rasmus Waagepetersen[3]

[1]Charles University in Prague, Faculty of Mathematics and Physics, Sokolovská 83, 18675 Prague 8, Czech Republic, benesv@karlin.mff.cuni.cz
[2]Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 18208 Prague 8, Czech Republic
[3]Aalborg University, Department of Mathematical Sciences, Fredrik Bajersvej 7G, DK-9220 Aalborg, Denmark, jm@math.aau.dk, rw@math.aau.dk

ABSTRACT

We consider a data set of locations where people in Central Bohemia have been infected by tick-borne encephalitis (TBE), and where population census data and covariates concerning vegetation and altitude are available. The aims are to estimate the risk map of the disease and to study the dependence of the risk on the covariates. Instead of using the common area level approaches we base the analysis on a Bayesian approach for a log Gaussian Cox point process with covariates. Posterior characteristics for a discretized version of the log Gaussian Cox process are computed using Markov chain Monte Carlo methods. A particular problem which is thoroughly discussed is to determine a model for the background population density. The risk map shows a clear dependency with the population intensity models and the basic model which is adopted for the population intensity determines what covariates influence the risk of TBE. Model validation is based on the posterior predictive distribution of various summary statistics.

Keywords: background intensity, Bayesian estimation, L-function, log Gaussian Cox spatial point process.

INTRODUCTION

The aims of statistical disease mapping are to characterize the spatial variation of cases of a disease and to investigate connections with possible covariates. In particular it is of interest to identify areas with an elevated disease risk. The data may be a point pattern showing e.g. home residences of diseased people or locations where people have acquired an infection. Often, the data are aggregated so that only counts of diseased people within subregions of the study region are available. Indeed, most statistical analyses reported in the literature are based on a so-called area level approach, where a model for aggregated data is used

after an initial aggregation. However, if a spatial point pattern is available, it is more natural to use a spatial point process model. Recent surveys of both the area level approach and point process modelling in disease mapping are given by Diggle (2000), Richardson (2003), and the accompanying discussion by Knorr-Held (2003) and Møller (2003). For a discussion of statistical analysis of disease mapping in general, see Lawson *et al.* (2001), Lawson (2001), and the references therein.

In this paper, we consider a point process approach to the analysis of a data set of positions of locations where people in Central Bohemia have been infected by tick-borne encephalitis (TBE). Specifically we consider a log Gaussian Cox point process (LGCP), where covariates concerning occurrence of different forest types, altitude, and the population density are used in the modelling of the spatially varying intensity of TBE infections. LGCPs were independently introduced in astronomy by Coles and Jones (1991) and in statistics by Møller *et al.* (1998); see also Møller and Waagepetersen (2002). To the best of our knowledge we here for the first time consider a Bayesian approach to inference for LGCPs with non-aggregated data.

A particular problem is the determination of the 'background intensity' of humans being at risk, cf. Diggle (2000) and Lawson (2001). Raw geographical population data connects population numbers to home locations, but typically people get infected during visits to more or less distant surroundings of their homes. This is an additional complication compared with spatial analysis of chronic diseases like cancer, where the objective may be to study association between disease incidence and risk factors at the home locations. We consider various approaches to smoothing of population data, where the smoothing to some extent is connected to the movement of people around their homes.

The data and previous analyses in Zeman (1997) and other papers are described in more detail in Section Data and background. Section Bayesian analysis using LGCP considers estimation of the background intensity, modelling of the risk function in terms of a LGCP depending on covariates, and our Bayesian approach to inference using Markov chain Monte Carlo methods. The results for different models of the background intensity are discussed in the last Section.

DATA AND BACKGROUND

DESCRIPTION OF DATA AND PROBLEM SETTING

TBE is an infectious debilitating illness which is transmitted by parasitic ticks and which occasionally afflicts humans. Epidemiologists and medical practitioners making decisions on prophylactic measures deal with the problem of estimating the risk that a human gets infected by TBE at a specific location, cf. Zeman (1997). Since field studies of potential animal hosts are expensive, usually the

data for statistical analysis consist of case locations and a population map. Moreover, explanatory variables of geographical nature which may influence the risk of infection are often available from geographical information systems.

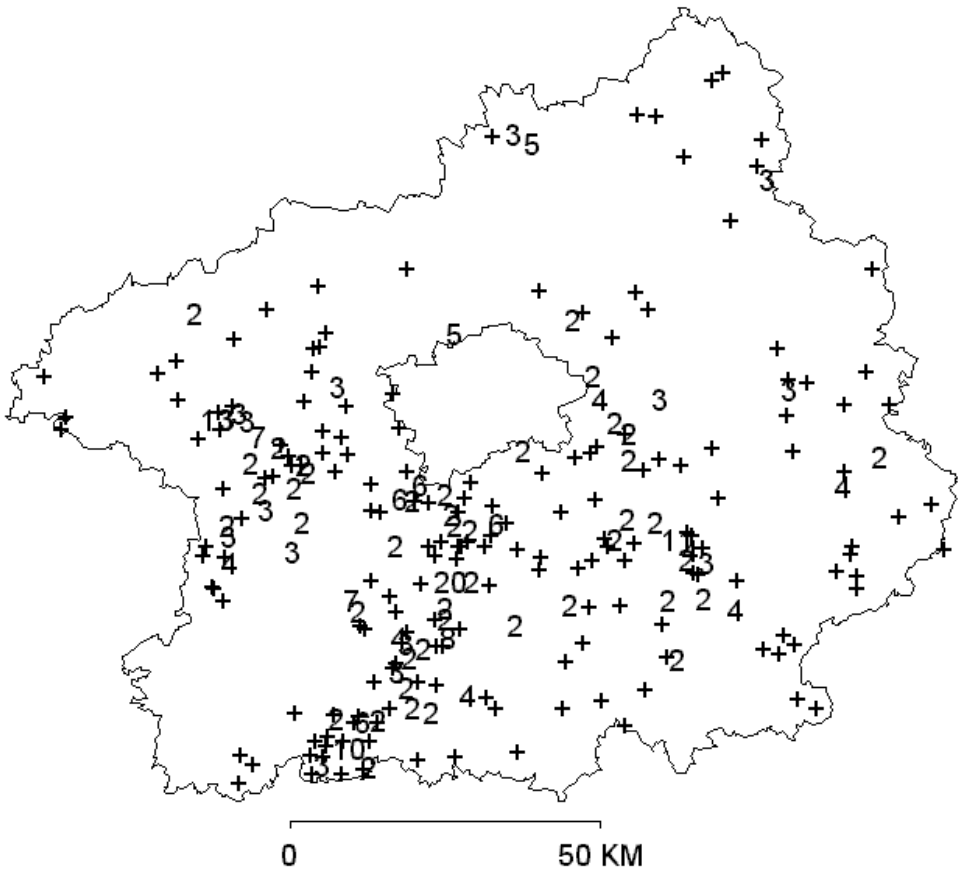

Figure 1: Locations of infection of 446 cases of tick-borne encephalitis in Central Bohemia. For each distinct location the number of cases associated with the location is shown (a plus corresponds to one case).

Fig. 1 shows the point pattern of locations of 446 reported cases of TBE in Central Bohemia during 1971-93. The empty space in the middle of the figure corresponds to the capital Prague, and the total area of Central Bohemia is about 11860 km^2. This data set was first studied in Zeman (1997). Only 255 distinct points are visible due to ties in the data caused by positional error where several cases in an area have been associated with a common representative point. The distinct points in Fig. 1 are marked with the number of cases associated with each point. Information concerning the magnitude of the positional error is not available.

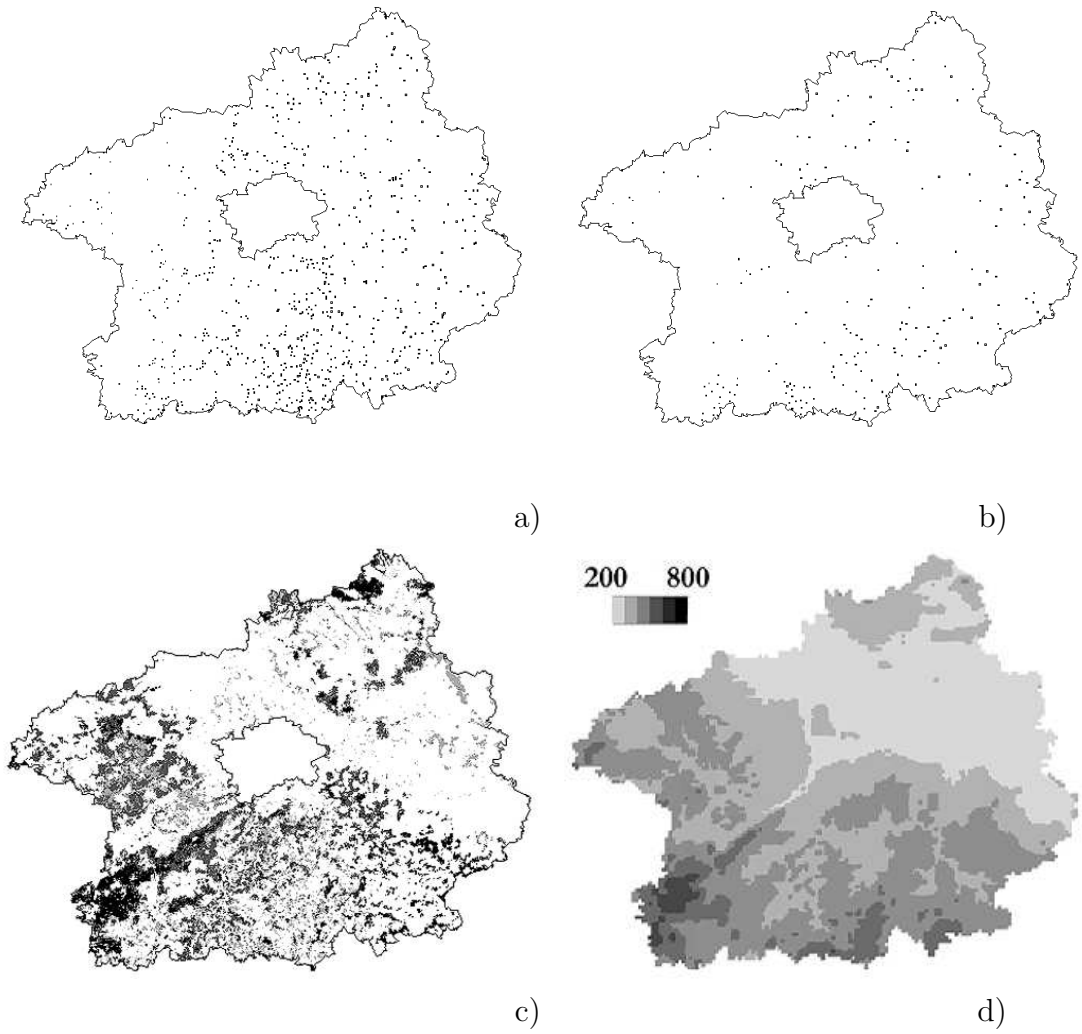

Figure 2: a) Locations of small forests (10-50 ha) (independent of the forest type). b) Locations of medium forests (50-150 ha). c) Three types of forest. Conifer: black; mixed: dark grey; foliate: light grey. d) Map of altitudes (in metres).

Different covariates are shown in Fig. 2a–d. Fig. 2a and Fig. 2b shows the locations of forests of areas between 10-50 and 50-150 ha, respectively. This covariate information is possibly relevant since ticks can be transmitted by deers and other animals living in small forest areas. Fig. 2c shows the subareas of three different forest types (conifer, foliate, and mixed forest). Fig. 2b and Fig. 2c are obtained from satellite images of LANDSAT-5 MSS with resolution power of 80×80 m^2. Finally, Fig. 2d shows a map of altitudes obtained from the Institute of Military Topography, Dobruska. Some other candidates, e.g. a covariate indicating the vicinity of a river, were not included in the analysis.

Finally, population data from the National Census Bureau, Prague, are avail-

able. For the Central Bohemia they consist of the number of inhabitants in 3582 administrative units. In Fig. 3 each unit is represented by a disc with center at a census point and radius given by $0.02\sqrt{\#\text{inhabitants in the unit}}$ km (chosen to get good resolution outside cities). Clusters of discs correspond to larger towns and cities. The total number of inhabitants is 1,112,717 and the largest city has about 74,000 inhabitants.

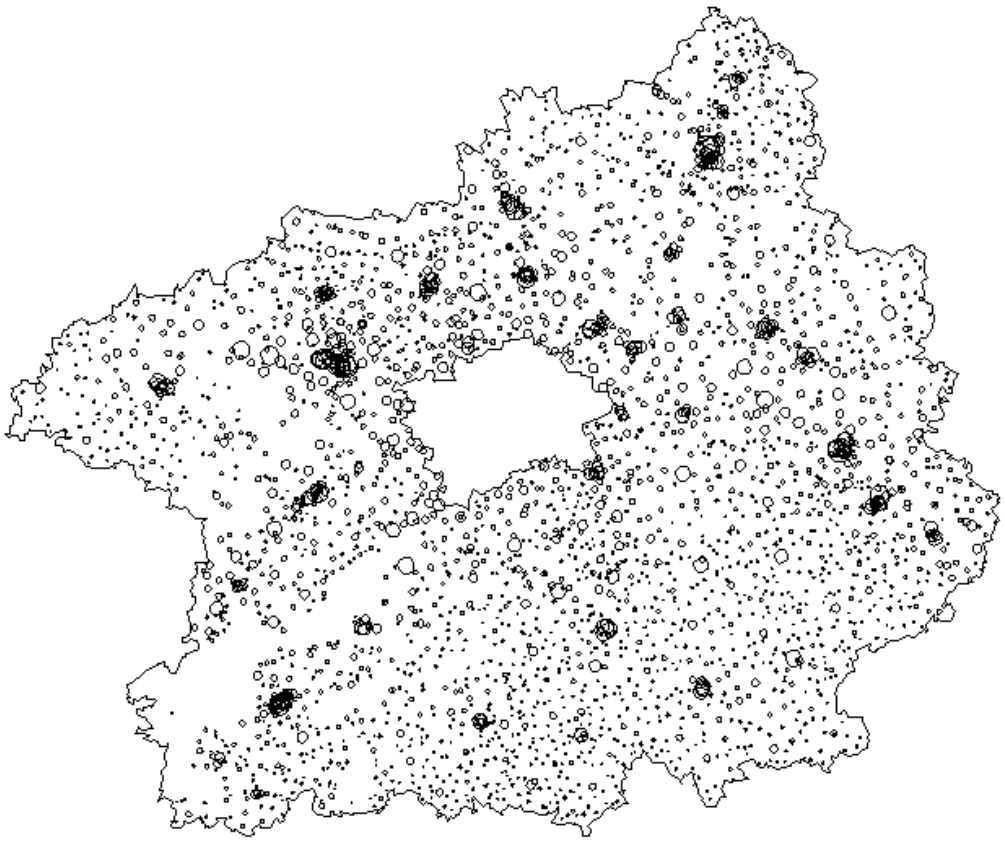

Figure 3: Population at 3582 administrative units in Central Bohemia represented by discs (see the text in the subsection Description of data and problem setting).

PREVIOUS DATA ANALYSIS

Zeman (1997) considers both the point pattern of TBE cases and another point pattern of cases for a related disease, Lyme borreliosis (LB). The LB data consist of paired data: 866 reported locations of infection during 1987-91 in Central Bohemia and the home location of each infected person. Zeman (1997) uses the distances between cases of infection and home location to obtain a kernel for smoothing the population map. Apart from this smoothed population map no

other covariates are included in Zeman's analysis where the intensity functions of TBE and LB cases (each considered as a realization of a point process) are estimated by kernel methods. For each disease, Zeman (1997) obtains a risk map by the ratio of the estimated intensity function and the smoothed population map (Bithell, 1990).

A similar ratio estimator of the risk map is suggested by Krejčíř (2000) where again both the TBE and the LB data are analyzed, using only the population data as an explanatory variable. He assumes that each point pattern of cases is a realization of an inhomogeneous Poisson point process with an intensity function constructed from beta splines, where the coefficients are estimated by a maximum likelihood method.

Incorporating the other explanatory variables in the model has so far only been studied in connection to two area level approaches for the TBE data. Mašata (1999) divides Central Bohemia into 41 irregular subregions, and he includes three covariates (the area in percentage of conifer, mixed, and foliate forests in each subregion) in a Bayesian Gaussian-Gaussian model (Stern and Cressie, 1999). Jiruše *et al.* (2004) use a subdivison of 141 squares of 10×10 km^2, and include the same types of covariates as Mašata (1999) together with the mean altitude and the proportion covered by small forests in each square. They use first a generalized linear model and the Akaike Information Criterion to optimize the set of parameters, and second an empirical Bayesian approach to estimate the risk. Jiruše *et al.* (2004) compare the credibility intervals for risk estimators obtained by their method with that of Mašata (1999), and conclude that rather similar results are obtained in subregions with a large risk for infection, although it is only the model in Mašata (1999) which incorporates spatial dependence.

The results of the above-mentioned papers are further discussed in final Section.

BAYESIAN ANALYSIS USING LOG GAUSSIAN COX PROCESSES

Let S denote the region of Central Bohemia and x the locations of observed tick cases. Below we describe a hierarchical model. At the first level, x is assumed to be a realization of a Poisson process X with an intensity function which is a product of a background intensity and a risk function as described in the first subsection of this section. Estimation of the background intensity is discussed in the second subsection. At the next level a log linear model for the risk function is proposed in the third subsection, incorporating the covariate information and a Gaussian process so that the uncertainty of the estimated background intensity is taken into account. At the final stage hyper priors on the unknown parameters for the covariates and the Gaussian field are imposed, whereby a posterior is

obtained in the last subsection of this section. For computational reasons certain approximations of the posterior are required.

Strictly speaking, the multiple points in x can not occur under the proposed model. However, our approximate approach only utilizes counts of locations within certain small cells and this makes the results less sensitive to the presence of ties in x.

A SIMPLIFIED MODEL

Our modelling of the TBE data is motivated by the following simplifying considerations, which are similar to one of the steps in the construction of a Neyman-Scott process (Neyman and Scott, 1958; Diggle, 1983).

In the observation period 1971-93 a number $m = 1,112,717$ of persons are living at home locations $h_1, \ldots, h_m \in S$, and the ith person makes a number N_i of visits to the surroundings of h_i, $i = 1, 2, \ldots, m$. The N_i's are assumed to be independent and Poisson distributed with mean $\lambda > 0$ independent of i. Given the N_i, the location of each visit of the ith person is distributed according to some density g_{h_i}, and the locations of visits of all persons are assumed to be independent. For a visit to a location $s \in S$, there is associated a probability $\pi(s)$ for getting an infection during the visit. The random set of locations where persons have been infected is then a Poisson process with intensity function of the form

$$\Lambda(s) = \rho(s)\pi(s), \; s \in S \tag{1}$$

where $\rho(s) = \lambda \sum_{i=1}^{m} g_{h_i}(s)$ is the *background intensity* of humans visiting s.

ESTIMATION OF BACKGROUND INTENSITY

The background intensity $\rho(s)$ is a crucial component of the modelling. As it is unknown, we discuss below how it may be estimated.

For the LB data both locations of infection and home are available. Under the crude assumption that the densities g_{h_i} are of the form $g_{h_i}(s) = g(\|s - h_i\|)$ one may as in Zeman (1997) try to estimate g from the LB data. Recall that if f denotes the density of $\|Z\|$ for a two-dimensional random variable Z with isotropic density $g_h(z) = g(\|z - h\|)$, then g and f are related by

$$g(\|z\|) = f(\|z\|)/(2\pi\|z\|), \quad z \in \mathbb{R}^2. \tag{2}$$

Zeman (1997) fits a power regression to a histogram for the log distances between home and place of infection. He then obtains an expression $\tilde{f}(h) = ah^b$ for the density of the distances and uses this as a kernel for smoothing of the population data. Strictly speaking $\tilde{f}$ is not a proper density on $\mathbb{R}_+$, and apparently Zeman (1997) is not using the correct transformation Eq.2 to obtain a density on $\mathbb{R}^2$.

We try another approach where we fit a non-parametric kernel density estimate $\hat{f}$ to the distances between home and place of infection in the LB data. The density $\hat{f}$ is subsequently transformed by Eq.2 into a density $\hat{g}$. The kernel estimate of the background intensity is finally

$$\hat{\rho}(s) = \lambda \sum_{j \in U} K_j \hat{g}(\|s - u_j\|) \tag{3}$$

where U is the index set of the administrative units, K_j is the number of persons associated with the jth unit, and u_j is the census point of the unit, cf. Fig. 3. Here λ is for the moment left unspecified as it is absorbed into another parameter introduced in the next subsection. Note that we are ignoring the fact that people in the jth unit live in the vicinity of u_j and not exactly at u_j.

The kernel estimate and alternative models for the background intensity are further discussed in the last subsection of this Section.

PRIOR DISTRIBUTIONS AND LIKELIHOOD USING A LGCP

We model π in Eq.1 by a log linear model,

$$\pi(s) = \exp\left(\beta^{\mathsf{T}} d(s) + Y(s)\right) \tag{4}$$

where $Y(s)$ is a zero-mean Gaussian process, $\beta = (\beta_0, \ldots, \beta_6)^{\mathsf{T}}$ is a regression parameter, and $d(s) = (1, d_1(s), \ldots, d_6(s))^{\mathsf{T}}$. Here β_0 is an intercept, and $d_1(s), \ldots, d_6(s)$ are the 6 covariates associated with the location $s \in S$, where the index corresponds to the following:

$1 \sim$ forest 10-50 ha, $2 \sim$ forest 50-150 ha, $3 \sim$ conifer,
$4 \sim$ mixed forest, $5 \sim$ foliate, $6 \sim$ altitude.

That means $d_1(s), \ldots, d_5(s)$ are zero-one variables corresponding to absence-presence of the attribute at location s. The role of $\exp(Y(s))$ is partly to model deviations of $\rho(s)/\hat{\rho}(s)$ from one. Therefore we do not constrain Eq.4 to be less than one. Actually, in the previous section, λ is absorbed in $\exp(\beta_0)$ and we replace the unknown ρ by the estimate Eq.3 with $\lambda = 1$. Then $\pi(s)$ is more precisely a *relative risk function*, since for $s_1, s_2 \in S$, $\pi(s_1)/\pi(s_2)$ is the ratio of risk functions at the locations s_1 and s_2.

We assume that Y is second-order stationary and isotropic with exponential covariance function, i.e.

$$\mathrm{Cov}(Y(s_1), Y(s_2)) = c(\|s_1 - s_2\|; \sigma^2, \alpha) = \sigma^2 \exp(-\|s_1 - s_2\|/\alpha) \tag{5}$$

where $\sigma^2 > 0$ is the variance and $\alpha > 0$ is the correlation parameter. A log Gaussian Cox process (LGCP) is then obtained by assuming that conditionally

on $Y = (Y(s))_{s\in S}$ and $\theta = (\beta, \sigma, \alpha)$, the TBE cases form a Poisson process X with intensity function $\hat{\rho}(s)\pi(s)$.

We view the Gaussian distribution for Y as a prior and the conditional distribution of X given (Y, θ) as the likelihood. Furthermore, a hyper prior density $p(\theta)$ for θ is imposed; specific hyper priors are considered in the next Section. Notice that the likelihood depends on θ only through β, and it has density

$$p(x|Y,\beta) = \exp\Big(|S| - \int_S \hat{\rho}(s)\exp\big(\beta^{\mathsf{T}}d(s) + Y(s)\big)\mathrm{d}s\Big)\times \\ \times \prod_{\xi\in x}\hat{\rho}(\xi)\exp\big(\beta^{\mathsf{T}}d(\xi) + Y(\xi)\big) \tag{6}$$

with respect to the unit rate Poisson process on S where $|\cdot|$ denotes area.

POSTERIOR AND DISCRETIZATION

The posterior, that is, the conditional distribution of (Y, θ) given $X = x$, can be specified as follows.

Suppose that $p(\theta)$ is proper and let E_θ denote expectation conditionally on θ. For $n \geq 1$ and pairwise distinct $s_1, \ldots, s_n \in S$, let $f_{(s_1,\ldots,s_n)}(\cdot|\theta)$ denote the conditional density of $(Y(s_1), \ldots, Y(s_n))$ given θ. The posterior density of $(Y(s_1), \ldots, Y(s_n), \theta)$ given $X = x$ is defined by

$$f_{(s_1,\ldots,s_n)}(y_1,\ldots,y_n,\theta|x) \propto E_\theta[p(x|Y,\beta)|Y(s_1) = y_1,\ldots,Y(s_n) = y_n]\times \\ \times f_{(s_1,\ldots,s_n)}(y_1,\ldots,y_n|\theta)p(\theta). \tag{7}$$

The posterior of the process (Y, θ) given $X = x$ is then given by the consistent set of finite-dimensional posterior distributions with densities of the form Eq.7 for $n \geq 1$ and pairwise distinct $s_1, \ldots, s_n \in S$. If $p(\theta)$ is improper we define the posterior similarly provided it is well-defined; i.e. provided

$$\int E_\theta[p(x|Y,\beta)]p(\theta)\mathrm{d}\theta < \infty.$$

The integral in Eq.6 depends on the continuous random field Y which cannot be represented on a computer. In practice the integral is approximated by a Riemann sum. The region S is appropriately translated and embedded in a rectangular region, say a square $[0, b]^2$ of side length $b > 0$. For an integer M the square is divided into a lattice of M^2 subsquares C_η^M, $\eta \in I^M = \{1, \ldots, M^2\}$, and in each subsquare indexed by η the covariate value is constant, represented by an average value $\tilde{d}^M(\eta) = \int_{C_\eta^M} d(s)\mathrm{d}s/|C_\eta^M|$. In Waagepetersen (2003) the approximate posterior based on the discretization is described and it is proved that under certain conditions, expectations computed with respect to the approximate posterior converges to the corresponding expectations with respect to Eq.7 when M tends to infinity.

RESULTS

In this section we discuss the results for the TBE data obtained by the Bayesian approach described above. Details concerning estimation of population intensity are given in the first subsection together with specification of priors, and posterior results are discussed in the second subsection. Model selection is adressed in the third subsection. The estimated relative risk map is discussed in the fourth subsection. Concluding remarks are given in last subsection of this section.

SPECIFICATION OF POPULATION INTENSITY MODEL AND PRIOR DISTRIBUTIONS

For the discretized LGCP, S is rescaled and embedded in a unit square which is divided into a grid of square cells C_η^M, $\eta \in I^M$ where $M = 65$. Thereby Central Bohemia is covered by 2166 cells, and S occupies about 51% of the unit square. Other values $M = 17$ and $M = 33$ considered yielded too coarse discretizations while $M = 129$ was computationally too demanding.

The following models of the population intensity are considered, setting $\lambda = 1$.

- Model W (constant ρ): $\hat{\rho}_{\text{const}}(s) = \sum_{j \in U} K_j / |S|$ is constant.
- Model P (kernel based on paired data): $\hat{\rho}_{\text{pair}}(s) = \sum_{j \in U} K_j \hat{g}(\|s - u_j\|)$ is estimated as in Eq.3 using the paired LB data.
- Models B,D, and E (Gaussian kernel): $\hat{\rho}_\tau(s) = \sum_{j \in U} K_j g(\|s - u_j\|; \tau)$ where $g(;\tau) : \mathbb{R}_+ \to \mathbb{R}_+$ is given by $g(h;\tau) = \phi(s;\tau)$ for $s \in \mathbb{R}^2$ with $\|s\| = h$ and where ϕ is the density for a two-dimensional radially symmetric Gaussian distribution with zero mean and standard deviation τ. Model B: $\tau = 0.7\,\text{km}$. D: $\tau = 2.5\,\text{km}$. E: $\tau = 5\,\text{km}$.

For the corresponding LGCPs, model W is equivalent to the limiting case $\tau \to \infty$ for a Gaussian kernel. Figure 4 shows a selection of the different kernels $g(\cdot)$. Note that the tail for P falls between the tails of D and E.

For all the population intensity models we use independent hyper priors for β, σ, and α given by

$$p_1(\beta) \propto 1, \quad \beta \in \mathbb{R}^7,$$
$$p_2(\sigma) \propto \exp(-10^{-6}/\sigma)/\sigma, \quad \sigma > 0,$$
$$p_3(\alpha) \propto 1/\alpha, \quad -6.91 < \log \alpha < -1.10.$$

The improper prior p_1 is completely flat and the improper p_2 yields an essentially flat prior for $\log \sigma$. The limits for the log uniform prior p_3 are chosen subjectively in order to accomodate a reasonable range of strengths of correlation. By similar arguments as in the proof of Proposition 1 in Christensen *et al.* (2001), a proper posterior is obtained for the discretized LGCP.

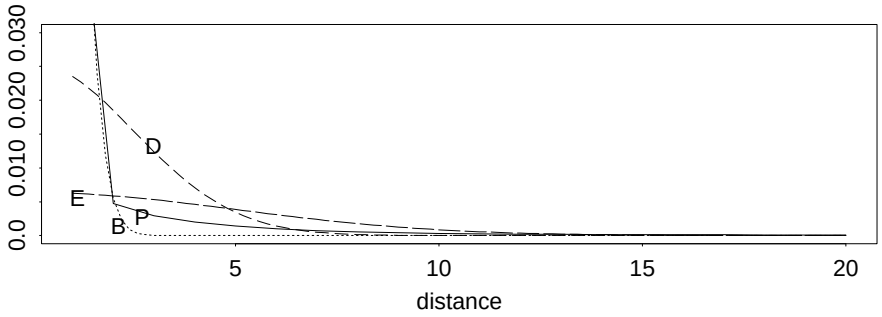

Figure 4: Kernels for models P, B, D, and E.

POSTERIOR RESULTS AND COMPARISON OF MODELS

The posterior means and posterior probabilities reported in this section are computed using a Markov chain Monte Carlo (MCMC) algorithm discussed in Waagepetersen (2003). Posterior means for β_i and for the different models are shown in Table 1. The numbers in parentheses are the probabilities $p_i = P(\beta_i > 0|x)$. Under model B, p_4 and p_5 indicate that the presence of mixed forest (β_4) or foliate forest (β_5) increases the risk of infection; in Jiruše *et al.* (2004) the presence of mixed forest is concluded to be a significant covariate. For all the population intensity models there is evidence that the presence of coniferous forest decreases the risk of infection and, except for B, that a high altitude increases the risk of infection. The posterior means are rather sensitive to the choice of population intensity. The qualitative results based on the posterior probabilities p_i are on the other hand rather similar for all population intensity models except model B.

	β_0	β_1	β_2	β_3	β_4	β_5	β_6
W	-10.1 (.0)	-1.2 (.3)	-.5 (.4)	-4.5 (.0)	.1 (.6)	.2 (.6)	1.5 (1.0)
P	-9.7 (.0)	-1.0 (.4)	-.1 (.5)	-3.3 (.0)	.2 (.6)	.4 (.7)	1.0 (1.0)
B	-9.8 (.0)	-1.6 (.3)	-.0 (.5)	-1.8 (.1)	2.2 (1.0)	2.2 (1.0)	-0.1 (.5)
D	-9.6 (.0)	-2.1 (.2)	-.2 (.5)	-3.4 (.0)	-.2 (.4)	.3 (.6)	1.0 (1.0)
E	-9.8 (.0)	-1.6 (.3)	-.4 (.4)	-4.0 (.0)	-.3 (.3)	.0 (.5)	1.3 (1.0)

Table 1: Posterior means for β_i and $p_i = P(\beta_i > 0 \,|\, x)$ (in parentheses), $i = 0, \ldots, 6$, under models W, P, B, D, and E.

The posterior means and standard deviations for σ and $\log \alpha$ are rather comparable for the different models. The posterior means for σ ranges between 2.0 (model W) and 2.4 (B), and the standard deviation between 0.1 (W) and 0.2 (model D). The posterior means for $\log \alpha$ are between -3.8 (B) and -3.3 (D), while the standard deviations take the value 0.2. The posterior mean of the

empirical mean of Y is close to zero for all models, and the posterior mean of its empirical standard deviation is a bit larger than 2 and close to the posterior mean of σ for all models. Notice that Y is playing an important role in the model since the posterior for σ is concentrated on an interval far from zero. For the exponential correlation function with $\log \alpha = -3.8$, the correlation is bigger than 0.01 for distances less than 15 km on the physical scale. Finally, let

$$\Lambda^M(s) = \hat{\rho}(\eta) \exp\left(\beta^{\mathsf{T}} \tilde{d}^M(\eta) + Y(\eta)\right),\ s \in C_\eta^M, \tag{8}$$

denote the intensity function of the discretized LGCP. The posterior mean

$$E[\int_S \Lambda^M(s)\mathrm{d}s \,|\, x]$$

of the intensity function integrated over Central Bohemia is between 445.5 and 446.1 (close to the number of observed cases) for the different models.

MODEL SELECTION

As the posterior results depend much on the choice of population map, one may naturally ask from which model conclusions should be drawn. In the Bayesian framework there exist several tools for model selection including Bayes factors, posterior predictive distributions, and, of course, an extended Bayesian analysis where prior probabilities are also assigned to the different models in question.

We restrict attention to the consideration of posterior predictive distributions, basically because this is supported by our present software. Consider a summary statistic $U(x)$ computed from the data x. The idea is to assess the fit of a posterior model by comparing $U(x)$ with the posterior predictive distribution; i.e. in our case the distribution of $U(X)$ where X is a Cox process with random intensity surface distributed as $[\Lambda^M|x]$, see Eq.8. Below we consider two types of summary statistics: the counts $n^M(\eta)$, $\eta \in I^M$, (number of points of X in subsquare C_η^M) and a variant of the K-function.

For the counts $n^M(\eta)$, $\eta \in I^M$, we just compare $n^M(\eta)$ with the posterior predictive mean $\hat{\lambda}_\eta = |C_\eta^M| E[\Lambda^M(\eta)|x]$ and compute discrepancy statistics

$$\chi_1^2 = \sum_{\eta \in I^M \cap S} (n^M(\eta) - \hat{\lambda}_\eta)^2 \quad \text{and} \quad \chi_2^2 = \sum_{\eta \in I^M \cap S} (n^M(\eta) - \hat{\lambda}_\eta)^2 / \hat{\lambda}_\eta.$$

The values of these χ^2-statistics can be used to rank the different models according to their predictive performance. Note that χ_2^2 is more tolerant towards deviations between $n^M(\eta)$ and the posterior predictive mean $\hat{\lambda}_\eta$ when $\hat{\lambda}_\eta$ is large. The values of χ_1^2 under the different models are B: 120, W: 171, E: 179, P: 184, D: 192. However, the picture is different for χ_2^2 where we have W: 1054, P: 1096, E: 1112, D: 1133, B: 1344. This is consistent with the degree of smoothness of the

population maps employed so that the smallest values of χ^2_2 are obtained for the models with the smoothest population maps. Another approach for using the cell counts $n^M(\eta)$ would be to consider so-called cross-validation predictive densities (Gelfand, 1996), but this is computationally quite demanding in our setting. In the following, we restrict attention to models B and W.

Our LGCP X can be extended to a so-called second-order intensity-reweighted stationary point process on $\mathbb{R}^2$ for which an extension of the K-function can be defined; for details see Baddeley *et al.* (2000). If $\lambda(\cdot)$ denotes the intensity function for X, the inhomogeneous K-function denoted K_{inhom} is given by

$$K_{\text{inhom}}(t) = \frac{1}{|A|} E\Big[\sum_{\xi \in X \cap A} \sum_{\substack{\eta \in X: \\ \xi \neq \eta}} \frac{1(\|\xi - \eta\| < t)}{\lambda(\xi)\lambda(\eta)} \Big] \tag{9}$$

for $t > 0$ and an arbitrary $A \subset \mathbb{R}^2$ with $0 < |A| < \infty$, $1(.)$ is the indicator function of the event in brackets. It is common practice to transform K_{inhom} into $L_{\text{inhom}}(t) = \sqrt{K_{\text{inhom}}/\pi}$ which is equal to t for a Poisson process. From Eq.9 we obtain an estimate of K_{inhom} by omitting the expectation, letting $A = S$, and replacing X with the observed data x; here we ignore the edge effects caused by unobserved tick infections outside S. Furthermore, the unknown $\lambda(\cdot)$ is replaced by the maximum likelihood estimate under the Poisson model corresponding to model W without Y (or equivalently with $\sigma^2 = 0$).

Fig. 5 shows the estimated $L_{\text{inhom}}(t) - t$. Notice that the estimate is bigger than zero which indicates clustering — this is in accordance with the results which showed that Y was not a negligible part of the model. Theoretically, $L_{\text{inhom}}(0) = 0$, while the behaviour of the estimate for small values of t is an artifact due to the multiple points in the data. The dashed curves in Fig. 5 are envelopes, i.e. pointwise minima and maxima for estimates of $L_{\text{inhom}}(t) - t$ computed from 39 point patterns simulated under the posterior predictive distributions corresponding to models B and W, respectively. If the observed data were generated by one of the posterior predictive distributions, then for each $t > 0$, there is 5% probability that the estimate of $L_{\text{inhom}}(t) - t$ from the data falls outside the envelopes. If we disregard the small t-values ($t < 1$ for model B, $t < 2$ for model W), then neither of the posterior predictive distributions seem to be in conflict with the observed data.

RELATIVE RISK MAP

The posterior means of the relative risk function under models B and W are shown in Figure 6 on a log scale. More precisely, in order to compare the results for models W and B we plot

$$\log E\big[\exp(\beta^T \tilde{d}^M(\eta) + Y(\eta) - \max_{\xi \in I^M \cap S} \{\beta^T \tilde{d}^M(\xi) + Y(\xi)\}) | x \big], \quad \eta \in I^M \cap S,$$

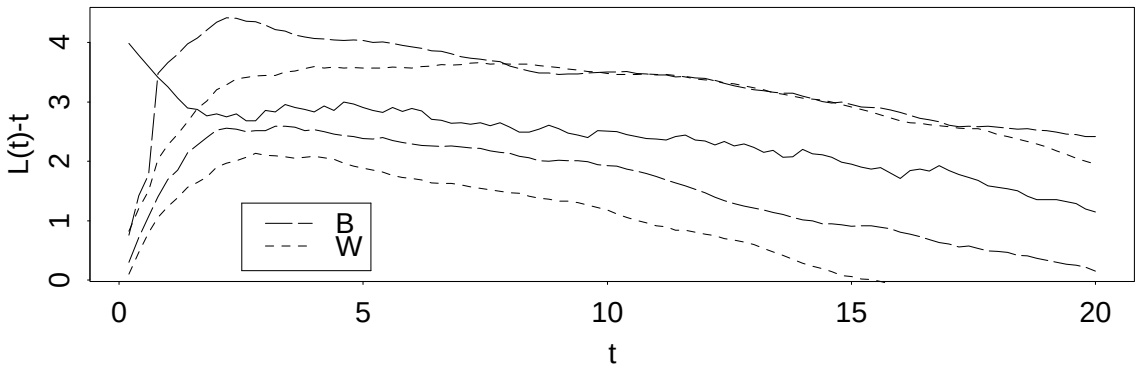

Figure 5: Estimated $L_{\text{inhom}}(t) - t$ (solid line) versus t (measured in km) under models B and model W.

for each model. The relative risk function is less varying under model W than under model B; for model B the smallest and the largest value are $\exp(-11.06)$, $\exp(-0.13)$, respectively and for the model W the values $\exp(-9.18)$ and $\exp(-0.05)$.

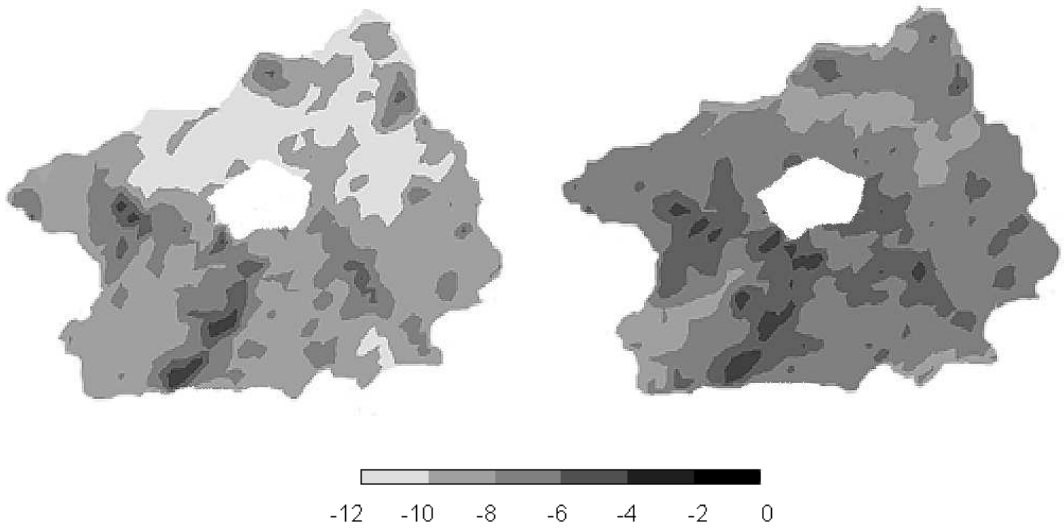

Figure 6: Maps of the logarithm of the posterior mean of the relative risk function divided by its maximal value. Left: model B. Right: model W.

Jiruše *et al.* (2004) compare their results with those of Mašata (1999) in a plot showing the credibility intervals of the relative risk function evaluated separately for each cell (ordered with increasing risk) in the irregular division of 41 cells used in Mašata (1999). Fig. 7 shows 2.5% and 97.5% posterior quantiles for the log relative risk function. The uncertainty is large; for model B and the cell with the largest mean posterior relative risk the 2.5% and 97.5% quantiles are -1.32 and

0, respectively; for model W the corresponding numbers are -0.48 and -0.01, respectively. The oscillations of the quantiles are smaller for model W than model B due to the constant population intensity for model W.

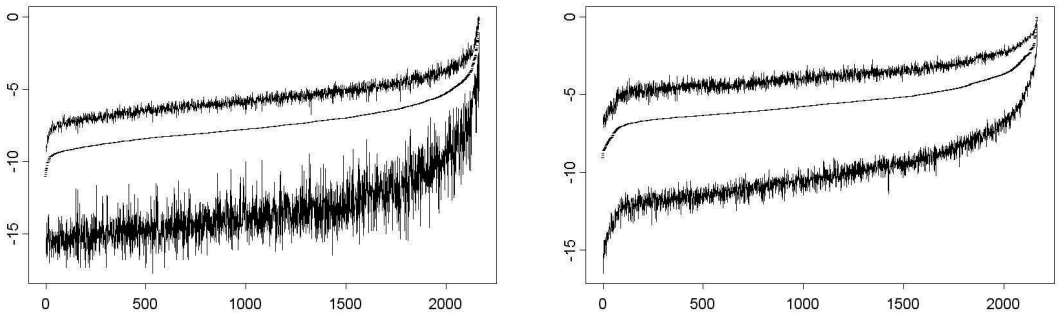

Figure 7: The log mean posterior relative risk function divided by its maximal value for all 2166 cells in ascending order of the mean posterior relative risk. The lower and upper curves are 2.5% and 97.5% posterior quantiles. Left: model B. Right: model W.

DISCUSSION

Comparing Fig. 6 with the maps in Zeman (1997), Mašata (1999), and Krejčíř (2000), the overall features are rather similar (no such map is shown in Jiruše *et al.*, 2004). However, our analysis appears to be more detailed than those in Zeman (1997) and Krejčíř (2000), since they do not include the covariate information and because of larger flexibility in our approach. Jiruše *et al.* (2004) and Mašata (1999) deal with covariates but using an area level approach. Thanks to the point process setting used in the present paper, we have provided a more detailed modelling and analysis of the spatial dependence (recall that Mašata (1999) uses only 41 irregular cells and Jiruše *et al.* (2004) only 141 cells of size 10×10 km^2, and it is only in Mašata (1999) that spatial dependence is incorporated).

The results concerning which covariates are important for predicting tick infections depend much on which population map is used. The best fit is according to the statistic χ^2_2 obtained with the population intensity model W. A uniform population intensity is on the other hand not realistic. Model P is obtained empirically from paired data, but under very crude assumptions. With the data available it seems hard to make a definite choice between the different population intensity models considered. One should therefore consider either of the two following possibilities: 1) collect more data from which a satisfactory population intensity map could be constructed, or 2) include uncertainty concerning the population map in the analysis e.g. by introducing a prior for the various maps, or

perhaps just on τ if one restricts attention to the Gaussian smoothing kernels.

In many examples of disease mapping one fixed population map is regarded as the truth, but our results suggest that such an approach can easily be misleading. In fact, the estimation of background rates in disease mapping has been an issue since 1990 when first Diggle (1990) used a plug-in estimator and subsequently Lawson and Williams (1994) demonstrated that use of different estimators for the background intensity can critically affect the relative risk surface estimation.

ACKNOWLEDGEMENTS
The authors wish to thank both referees for helpful comments and Dr. Petr Zeman, Regional Center of Hygiene, Dittrichova 17, Prague 2, Czech Republic, for encouraging their interest in the problem and for providing the data sets. VB and KB were supported by grants GAČR 201/03/0946 and MSM 0021620839. JM and RW were supported by MaPhySto – Centre for Mathematical Physics and Stochastics, funded by a grant from The Danish National Research Foundation, and by the European Union's Network "Statistical and Computational Methods for the Analysis of Spatial Data. ERB-FMRX-CT96-0096".

References

Baddeley AJ, Møller J, Waagepetersen R (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. Statistica Neerlandica 54:329–50.

Bithell JF (1990). An application of density estimation to geographical epidemiology. Statistics in Medicine 9:691–701.

Christensen OF, Møller J, Waagepetersen R (2001). Geometric ergodicity of Metropolis-Hastings algorithms for conditional simulation in generalised linear mixed models. Methodology and Computing in Applied Probability 3:309–27.

Coles P, Jones B (1991). A lognormal model for the cosmological mass distribution. Monthly Notices of the Royal Astronomical Society 248:1–13.

Diggle P (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. Journal of Royal Statistical Society Series A 153:349–62.

Diggle P (2000). Overview of statistical methods for disease mapping and its relationship to cluster detection. In: Elliott P, Wakefield JC, Best NG, Briggs DJ, eds., Spatial Epidemiology: Methods and Applications. Oxford: Oxford University Press, 87–103.

Diggle PJ (1983). Statistical Analysis of Spatial Point Patterns. Academic Press, London.

Gelfand AE (1996). Model determination using sampling-based methods. In: Gilks WR, Richardson S, Spiegelhalter DJ, eds., Markov Chain Monte Carlo in Practice. London: Chapman & Hall, 145–62.

Jiruše M, Machek J, Beneš V, Zeman P (2004). A Bayesian estimate of the risk of tick-borne diseases. Applications of Mathematics 49:389–404.

Knorr-Held L (2003). Some remarks on Gaussian Markov random field models for disease mapping. In: Green PJ, Hjort NL, Richardson S, eds., Highly Structured Stochastic Systems. Oxford: Oxford University Press, 260–4.

Krejčíř P (2000). A maximum likelihood estimator of an inhomogeneous Poisson point process intensity using beta-splines. Kybernetika 36:455–64.

Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel JF, Bertollini R, eds. (2001). Disease Mapping and Risk Assessment for Public Health. Chichester: Wiley.

Lawson AB (2001). Statistical Methods in Spatial Epidemiology. Chichester: Wiley.

Lawson AB, Williams FLR (1994). Armadale: a case-study in environmental epidemiology. Journal of Royal Statististical Society Series A :285–98.

Mašata M (1999). Assessment of risk of infection by means of a Bayesian method. In: Beneš V, Janáček J, Saxl I, eds., Proceedings of the International Conference on Stereology, Spatial Statistics, and Stochastic Geometry. Praha: Union of Czech Mathematicians and Physisists, 197–202.

Møller J (2003). A comparison of spatial point processes in epidemiological applications. In: Green PJ, Hjort NL, Richardson S, eds., Highly Structured Stochastic Systems. Oxford: Oxford University Press, 264–8.

Møller J, Syversveen A, Waagepetersen R (1998). Log Gaussian Cox processes. Scandinavian Journal of Statistics 25:451–82.

Møller J, Waagepetersen RP (2002). Statistical inference for Cox processes. In: Lawson AB, Denison D, eds., Spatial Cluster Modelling. Chapman and Hall/CRC, Boca Raton, 37–60.

Neyman J, Scott EL (1958). Statistical approach to problems of cosmology. Journal of the Royal Statistical Society B 20:1–43.

Richardson S (2003). Spatial models in epidemiological applications. In: Green PJ, Hjort NL, Richardson S, eds., Highly Structured Stochastic Systems. Oxford: Oxford University Press, 237–59.

Stern HS, Cressie N (1999). Inference for extremes in disease mapping. In: Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel JF, Bertollini R, eds., Disease Mapping and Risk Assessment for Public Health. New York: Wiley, 63–84.

Waagepetersen R (2003). Convergence of posteriors for discretized log Gaussian Cox processes. Statistics and Probability Letters 66:229–35.

Zeman P (1997). Objective assessment of risk maps of tick-borne encephalitis and lyme borreliosis based on spatial patterns of located cases. International Journal of Epidemiology 26:1121–30.