A Spatially Discrete Approximation to Log-Gaussian Cox Processes for Modelling Aggregated Disease Count Data

Olatunji Olugoke Johnson, Peter Diggle and
Emanuele Giorgi

CHICAS, Lancaster Medical School,
Lancaster University, Lancaster, UK

January 29, 2019

Abstract

In this paper, we develop a computationally efficient discrete approximation to log-Gaussian Cox process (LGCP) models for the analysis of spatially aggregated disease count data. Our approach overcomes an inherent limitation of spatial models based on Markov structures, namely that each such model is tied to a specific partition of the study area, and allows for spatially continuous prediction. We compare the predictive performance of our modelling approach with LGCP through a simulation study and an application to primary biliary cirrhosis incidence data in Newcastle-Upon-Tyne, UK. Our results suggest that when disease risk is assumed to be a spatially continuous process, the proposed approximation to LGCP provides reliable estimates of disease risk both on spatially continuous and aggregated scales. The proposed methodology is implemented in the open-source R package `SDALGCP`.

Keywords: disease mapping; geostatistics; log-Gaussian Cox process; Monte Carlo maximum likelihood.

1 Introduction

In this paper our concern is to make inference on a spatially continuous disease risk surface using aggregated counts of reported disease cases, say y_i, over regions $\mathcal{R}_i$ forming a partition of a geographical area of interest A. In this context, information on risk factors and on the population at risk may also be available, possibly at different spatial scales. We shall denote these by $d(x)$ and $m(x)$, respectively, when available on a spatially continuous scale, and by d_i and m_i when they are spatially aggregated.

1

Existing methods from small area estimation (SAE) only allow spatial prediction at the aggregated level of the regions $\mathcal{R}_i$ and are usually based on a Gaussian Markov random field (GMRF) structure (Besag, 1974; Rue and Held, 2005). Typically, non-zero elements of the precision matrix of a GMRF are restricted to contiguous pairs of the $\mathcal{R}_i$. Hence, the formulation and interpretation of a GMRF is tied to the specific partition of A, which will usually have been drawn up for administrative, historical, or other reasons unrelated to the disease aetiology. The use of such models also becomes impractical when the spatial units $\mathcal{R}_i$ change over time. Wall (2004) points out that the use of GMRFs is especially problematic when dealing with irregular geometries, which can induce counter-intuitive forms for the correlation structure between variables associated with the $\mathcal{R}_i$.

The geostatistical paradigm, unlike SAE, treats disease risk as a spatially continuous phenomenon irrespective of the data-format. Diggle et al. (2013) argue that the analysis of spatially aggregated counts can be regarded as a special case of the class of geostatistical problems and propose to model the y_i as an aggregated realisation of a Log-Gaussian Cox process (LGCP). Unlike GMRFs, LGCPs allow for prediction of disease risk at any spatial scale, while avoiding the ecological fallacy (Wakefield and Shaddick, 2006). However, fitting of LGCP models using the aggregated counts y_i is computationally demanding due to the iterative imputation of the unobserved locations for each reported case within a region $\mathcal{R}_i$ (Li et al., 2012).

In this paper, our objective is to develop a computationally efficient approximation to LGCPs in order to predict disease risk at any desired spatial scale. We argue that this provides a more realistic alternative to GMRF models when LGCPs are not computationally feasible, and can also be used as an exploratory tool in order to inform more complex modelling apporaches based on LGCPs.

In Section 2 of the paper, we review existing methods for modelling spatially aggregated disease counts. In Section 3, we develop a computationally efficient spatially discrete approximation to LGCP models. In Section 4 we carry out a simulation study to investigate the predictive performance of the proposed approximation and compare this with an exact fitting algorithm for LGCP models. In Section 5 we show an application of the method to a data-set on primary biliary cirrhosis (PBC) incidence in Newcastle, UK. Section 6 is a concluding discussion on the advantages and limitations of the proposed approach.

The method has been implemented in the open-source R package SDALGCP (Johnson et al., 2018), available from the Comprehensive R Network Archive. The R code for reproducing the results of Section 5 is available as supplementary material.

2 Existing methods for modelling spatially aggregated disease counts data

2.1 Gaussian Markov random field models

Let Y_i denote the reported disease count in region $\mathcal{R}_i$. Conditionally on a zero-mean Gaussian process $S = (S_1, \ldots, S_n)$, assume that the Y_i are mutually independent Poisson random variables with expectations

$$\lambda_i = m_i \exp\{d_i^\top \beta + S_i\}, i = 1, \ldots, n \tag{1}$$

where β is a vector of regression coefficients. Spatially discrete models are then developed by specifying the precision matrix for the Gaussian process S. Here, we focus on the two most commonly used formulations for S, namely the conditional autorgressive (CAR) and intrinsic conditional autoregressive (ICAR) models.

Let $i \sim j$ be a shorthand notation for "$\mathcal{R}_i$ and $\mathcal{R}_j$ are neighbours". A CAR model then assumes that

$$S_i | S_{-i} \sim N \left(\rho_c \sum_{j \sim i} c_{ij} S_j, \tau_i^2 \right), \tag{2}$$

where $S_{-i} = \{S_j : j \neq i\}$, ρ_c is the spatial dependence parameter and c_{ij} are known quantities such that $c_{ij} \neq 0$ if and only if $j \sim i$ and $j \neq i$. It follows from Brook's Lemma (Brood, 1964) and the Hammersley-Clifford Theorem (Besag, 1974) that the joint distribution of S is a multivariate zero-mean Gaussian distribution with covariance matrix

$$(I - \rho_c C)^{-1} \tilde{D}, \tag{3}$$

where $\tilde{D} = \{\tau_1^2, \ldots, \tau_n^2\}$, while the specification of C is generally tied to the specific arrangement of the partition of the region of interest. The most common approach is to set $c_{ij} = 1$ if $j \sim i$ and 0 otherwise. The matrix in (3) is then a valid covariance matrix if $\xi_{max}^{-1} < \rho_c < \xi_{min}^{-1}$ (Cressie, 1993, pg. 472), where ξ_{min} and ξ_{max} are the minimum and maximum eigenvalues of C, respectively. Scaling of the matrix C so as to obtain a weighted average of the S_j in (2) also implies that $-1 < \rho_c < 1$.

The ICAR model is a special case of the CAR model when $\rho_c = 1$ in (2). Although this leads to an improper distribution for S because of the singularity of its covariance matrix, the associated conditional distribution of S given Y is proper.

2.2 Log-Gaussian Cox process models

A spatial point process is a stochastic mechanism that generates a countable set of events $x_i \in \mathbb{R}^2$. The class of inhomogeneous Poisson processes with intensity $\lambda(x)$ is defined by the following postulates.

3

1. The number of events, $N(\mathcal{A})$, in any planar region $\mathcal{A} \subset \mathbb{R}^2$ follows a Poisson distribution with mean $\int_{\mathcal{A}} \lambda(x) dx$.

2. Conditionally on $N(\mathcal{A})$, each event in $\mathcal{A}$ is an independent random sample from a distribution on $\mathcal{A}$ with probability density function proportional to $\lambda(x)$.

A Cox process (Cox, 1955) is defined by a non-negtive-valued stochastic process $\Lambda(x)$ such that, conditional on a realisation of $\Lambda(x)$, the process is an inhomogenous Poisson process with intensity $\Lambda(x)$. If we assume that $\log\{\Lambda(x)\} = S(x)$ is a Gaussian process, we obtain the log-Gaussian Cox process (LGCP); for more details on the theoretical properties of LGCPs, see Møller et al. (1998).

Diggle et al. (2013) develop a modelling framework for aggregated disease count data using LGCPs. They assume that, conditionally on $S(x)$, the Y_i are mutually independent Poisson variables with means

$$\int_{\mathcal{R}_i} m(x) \exp\{d(x)^{\top}\beta + S(x)\} \, dx, \tag{4}$$

where $d(x)$ is a vector of covariates at location x with associated regression coefficients β. Unlike the spatially discrete models described in the previous section, an LGCP is not tied to any particualr partition of the area of interest and therefore provides a route to a solution to the problem of combining information at multiple spatial scales. However, this is offset by a substantial increase in the computational burden arising from the need to impute the unobserved locations for each of the reported cases within each of the $\mathcal{R}_i$, $i = 1, \ldots, n$ (Li et al., 2012). In the next section, we circumvent this issue by proposing a spatially discrete approximation to $S(x)$ which allows to model the counts y_i as the realisation a Poisson log-linear mixed model.

3 A spatially discrete approximation to Log-Gaussian Cox processes

Let $w_i(x)$ be a positive function with domain $\mathcal{R}_i$, such that $\int_{\mathcal{R}_i} w_i(x) \, dx = 1$. Using the same notation as in Section 2.2, we approximate the conditional log-intensity of an LGCP as piecewise constant by taking its weighted average over $\mathcal{R}_i$ to give

$$
\begin{aligned}
\log\{\Lambda(x)\} &\approx \int_{\mathcal{R}_i} w_i(x) \left[d(x)^{\top}\beta^* + S^*(x) \right] \, dx \\
&= \int_{\mathcal{R}_i} w_i(x) \, d(x)^{\top}\beta^* \, dx + \int_{\mathcal{R}_i} w_i(x) \, S^*(x) \, dx \\
&= d_i^{\top}\beta^* + S_i^*, \ x \in \mathcal{R}_i, \tag{5}
\end{aligned}
$$

where β^* is a vector of regression coefficients for the aggregate explanatory variables d_i and $S^*(x)$ is a Gaussian process. The rationale for using the weighting

function $w_i(x)$ is to account for the potential non-homogeneous distribution of disease cases within a region $\mathcal{R}_i$. For example, a larger number of cases may concentrate in more densely populated areas, thus a natural choice for $w_i(x)$ would be to set this equal to $m(x)/m_i$ with $m_i = \int_{\mathcal{R}_i} m(x)dx$, if $m(x)$ is available. If $m(x)$ is instead unavailable, a pragmatic approach would be to set $w_i(x) = 1/|\mathcal{R}_i|$.

Following from (5), we obtain the following approximation for the conditional mean of the counts Y_i

$$
\begin{aligned}
\lambda_i = \int_{\mathcal{R}_i} m(x)\Lambda(x)\, dx &\approx \int_{\mathcal{R}_i} m(x)\exp\left\{d_i^\top \beta^* + S_i^*\right\}\, dx \\
&= m_i \exp\{d_i^\top \beta^* + S_i^*\} \\
&= m_i \exp\{\eta_i\} \\
&= \mu_i.
\end{aligned}
\tag{6}
$$

The joint distribution of $S^* = (S_1^*, \ldots, S_n^*)$ is multivariate Gaussian with zero mean and covariance function

$$
\text{Cov}\{S_i^*, S_j^*\} = \sigma^2 \int_{\mathcal{R}_i} \int_{\mathcal{R}_j} w_i(x)w_j(x')\, \rho(\|x - x'\|; \phi)\, dx\, dx',
\tag{7}
$$

where $\|\cdot\|$ is the Euclidean distance and $\rho(\cdot; \phi)$ is the isotropic and stationary covariance function of $S^*(x)$ indexed by the parameter ϕ. Hence, the resulting model (6) falls under the class of generalized linear mixed models. Also, note that the variance of S_i^* depends on the size and shape of $\mathcal{R}_i$, with larger regions leading to smaller variances.

We now provide further details on the computation of the covariance function in (7). Among the class of isotropic and stationary covariance functions for $S^*(x)$ in (6), one of the most commonly used is the Matérn covariance function (Stein, 2012), which has expression

$$
\text{Cov}\{S^*(x), S^*(x')\} = \frac{\sigma^2}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{u}{\phi}\right)^\kappa \mathcal{K}_\kappa \left(\frac{u}{\phi}\right),
\tag{8}
$$

where $u = \|x - x'\|$ is the Euclidean distance between any two locations x and x', σ^2 is the variance, ϕ is a scale parameter that regulates the rate at which the spatial correlation decays for increasing distance u, κ is the shape parameter that determines the differentiability of the process S and $\mathcal{K}_\kappa(\cdot)$ is the modified Bessel function of the second kind of order $\kappa > 0$. Estimating κ reliably requires a large amount of densely sampled data, which in this context is not available. As a pragmatic apporach, we then set $\kappa = 0.5$ which reduces (8) to

$$
\text{Cov}\{S^*(x), S^*(x')\} = \sigma^2 \exp\{-u/\phi\}
$$

corresponding to a mean-square continuous process.

We approximate (7) as a discrete sum over L_i and L_j randomly chosen points in $\mathcal{R}_i$ and $\mathcal{R}_j$ to give

$$\int_{\mathcal{R}_i} \int_{\mathcal{R}_j} w_i(x) w_j(x') \, \rho(\|x - x'\|; \phi) \, dx \, dx' \approx$$
$$\frac{\sum_{k=1}^{L_i} \sum_{k'=1}^{L_j} w_i(x_k) w_j(x_{k'}) \, \rho(\|x_k - x_{k'}\|; \phi)}{\sum_{k=1}^{L_i} \sum_{k'=1}^{L_j} w_i(x_k) w_j(x_{k'})}, \tag{9}$$

To attain a good spatial coverage of $\mathcal{R}_i$ and $\mathcal{R}_j$, we propose to draw each of the x_k and $x_{k'}$ in the above equation using a class of inhibition processes (Diggle, 2013, pp. 110-116) which combines simple sequential inhibition with rejection sampling. More specifically, we proceed through the following steps.

1. Compute $w_{max} = \max_{x \in \mathcal{R}_i} w_i(x)$.

2. Generate x_{prop} over $\mathcal{R}_i$ from a homogeneous Poisson process with intensity w_{max}.

3. Compute $p(x_{prop}) = w_i(x_{prop})/w_{max}$.

4. Generate a sample u from the uniform distribution on $(0,1)$.

5. If $k = 1$, set $x_1 = x_{prop}$ if $u \le p(x_{prop})$; for $k > 1$ and given $\{x_j : j = 1, \ldots, k-1\}$, set $x_k = x_{prop}$ if $u \le p(x_{prop})$ and x_{prop} falls at the intersection of $\mathcal{R}_i$ and $\{x \in \mathcal{R}_i : \|x - x_j\| > \delta(1 - w(x_j)/w_{max})\}$. Otherwise, reject x_{prop}.

6. Repeat 2 to 5, until $k = L_i$.

To identify a suitable value for L_i (the total number of generated points within $\mathcal{R}_i$), a possible solution is to use the packing density for a sequential inhibitory point process given by

$$\gamma = \frac{L_i \pi \delta^2}{4|\mathcal{R}_i|}, \tag{10}$$

where δ is the minimum permissible distance between points. The maximum possible value for γ is obtained by close-packed discs whose centres form an equilateral triangular lattice with sides of length $\delta = \pi/\sqrt{12}$. Through a simulation study, Tanemura (1979) suggested to set $\gamma = 0.55$ in order to achieve good spatial coverage in a relatively small number of iterations. Once γ and δ are fixed, we can then obtain L_i through equation (10).

An alternative solution is to leave choose γ as a function of ϕ using the following adaptive algorithm.

1. For a given ϕ, initialize a batch size k and a relative tolerance ϵ;

2. Locate k quadrature points with packing intensity $\gamma(k) = k\pi\delta^2/4|\mathcal{R}_i|$, evaluate the integral in (9) and denote its value as I_{old};

3. Add k points using a packing intensity $\gamma(k)/2$, re-evaluate the integral and denote its value as I_{new};

4. If $I_{new} = I_{old}$, stop the algorithm. Otherwise, set $I_{new} = I_{old}$, add k points with $\gamma(k)/3$ and repeat until $|I_{old} - I_{new}| < \epsilon|I_{new}|$.

Since ϕ is almost always unknown, the additional computational burden from the adaptive algorithm can be remarkable. When fitting the model in (6) (see next section for more details), our recommendation is to use the non-adaptive algorithm first, in order to locate the likely value of ϕ, and then to run a final estimation using the adaptive algorithm.

3.1 Monte Carlo maximum likelihood

We carry out parameter estimation for the model in (6) using the Monte Carlo maximum likelihood (MCML) method (Christensen, 2012).

Let $f(\cdot)$ be a shorthand notation for "the density function of $\cdot$". Let $y^\top = (y_1, \ldots, y_n)$ and linear predictor $\eta^\top = (\eta_1, \ldots, \eta_n)$; it then follows that conditionally on $S^* = (S_1^*, \ldots, S_n^*)^\top$, the joint distribution of Y is

$$f(y|\eta) = \prod_{i=1}^{n} f(y_i|\eta_i),$$

where

$$f(y_i|\eta_i) \propto \exp\{y_i \log \mu_i - \mu_i\}.$$

Let $\psi = (\beta, \sigma^2, \phi)$ denote the vector of the model parameters, then the likelihood function for ψ is obtained by integrating out S^*, i.e.

$$L(\psi) \quad = \quad \int_{\mathbb{R}^n} f(y|\eta) \; f(\eta; \psi) \, d\eta. \tag{11}$$

In (11) $f(\eta; \psi)$ is a multivariate Gaussian distribution function with mean $D\beta$, where D denote a matrix of explanatory variables, and covariance matrix Σ, whose (i, j)-th entry is given by (7). To reduce the computational burden accrued from the numerical approximation (9), we restrict the maximization of (11) to a finite set of predefined values for ϕ and, for each of these, pre-compute the covariance matrix Σ together with its inverse, determinant and Cholesky decomposition.

Since the high-dimensional integral in (11) cannot be solved analytically, we use Monte Carlo methods for the approximation of the likelihood. Let ψ_0 denote our best guess of ψ. We re-write (11) as

$$L(\psi) \quad \propto \quad E_{\eta|y}\left[\frac{f(\eta; \psi)}{f(\eta; \psi_0)}\right], \tag{12}$$

where the expectation E is taken with respect to the conditional distribution of η given y with parameters vector ψ_0. We provide the proof of this in Appendix A of the supplementary material.

To generate N samples, say $\eta_{(j)}$, from the conditional distribution of η given y, we use a Monte Carlo Markov chain (MCMC) algorithm implemented in the `Laplace.sampling.MCML` function in the PrevMap package (Giorgi and Diggle, 2017). This function uses a Metropolis-adjusted Langevin MCMC algorithm to update the standardised vector of random effects, $\tilde{\eta} = \hat{\Sigma}^{-\frac{1}{2}}(\eta - \hat{\eta})$, where $\hat{\eta}$ and $\hat{\Sigma}$ are the mode and the inverse of the negative Hessian of $f(\eta; \psi_0)$ at $\hat{\eta}$. We can then approximate the likelihood function in (12) as

$$L(\psi) \approx L_N(\psi) \quad = \quad \frac{1}{N} \sum_{j=1}^{N} \frac{f(\eta_{(j)}; \psi)}{f(\eta_{(j)}; \psi_0)}. \tag{13}$$

As $N \to \infty$, in the above equation, $L_N(\psi)$ converges to $L(\psi)$ (Geyer and Thompson, 1992; Geyer, 1994, 1996).

Finally, we maximize (13) using a constrained quasi-Newton optimization algorithm, implemented in the `nlminb` function in the R software environment, by providing analytical expressions for the first and second derivatives of (13) with respect to ψ. If $\hat{\psi}_N$ denote the resulting MCML estimate, we then set $\psi_0 = \hat{\psi}_N$ and repeat the previous steps until convergence.

3.2 Continuous spatial prediction

We now consider the problem of carrying out spatial prediction of $S^*(x)$ at a pre-defined location x within the study area A. Using the same notation as in the previous section, we first note that

$$
\begin{aligned}
f(S^*(x)|y) \quad &= \quad \int_{\mathbb{R}^n} f(\eta, S^*(x)|y) \, d\eta \\
&= \quad \int_{\mathbb{R}^n} f(\eta|y) f(S^*(x)|\eta, y) \, d\eta \\
&= \quad \int_{\mathbb{R}^n} f(\eta|y) f(S^*(x)|\eta) \, d\eta. \tag{14}
\end{aligned}
$$

Hence, we sample from $f(S^*(x)|y)$ using the following two-step procedure: (1) draw samples $\eta_{(j)}$, for $j = 1, \ldots, N$ from $f(\eta|y)$ using the MCMC algorithm described in the previous section; (2) for each $\eta_{(j)}$, for $j = 1, \ldots, N$ simulate from $f(S^*(x)|\eta_{(j)})$, a Gaussian distribution with mean $\mu^*(x) = c(x)^\top \Sigma^{-1}(\eta_{(j)} - D\beta)$ and variance $v^2(x) = \sigma^2 - c(x)^\top \Sigma^{-1} c(x)$, where $c(x)^\top = (c_1(x), \ldots, c_n(x))$, $c_i(x) = \sigma^2 \int_{\mathcal{R}_i} w(x)\rho(\|x - x'\|) \, dx'$, and we use (9) to approximate the integral. The resulting samples from $f(\eta|y)$ can then be used to compute non-linear properties of $S^*(x)$ and to summarise these using, for example, predictive means and standard errors.

4 Simulation Study

We now conduct a simulation study to assess the predictive performance of the proposed approximation in (3) when the underlying process is an LGCP model.

We simulate $B = 1,000$ data-set of counts using the administrative boundaries of the lower layer super output areas (LSOAs) in Newcastle-Upon-Tyne, UK, as in the application of Section 5. We specify the offsets $m(x)$ using population density estimates from the OpenPopGrid database (Murdock et al., 2015) and simulate the locations of the events using an inhomogeneous Poisson process with intensity $m(x) \exp\{S(x)\}$. We define three scenarios by setting the standard deviation of the Gaussian random field $S(x)$ to $\sigma = 0.706$ and let ϕ (whose unit of measure is kilometres) vary over the set $\{100, 800, 1500\}$. The value of the standard deviation corresponds to the posterior mean obtained from the fitted LGCP in the application to primary biliary cirrhosis data described in the next section. Finally, for each of the $1,000$ simulated data-sets of aggregated counts at LSOA level, we fit the following models.

- *LGCP.* We use a Bayesian data augmentation technique implemented in the `lgcp` package (Taylor et al., 2015). We overlay a computational grid at a spacing of of 300×300 metres onto the area of interest and fit the model in (4). We run 3,100,000 iterations of the MCMC algorithm with a burn-in of 100,000 samples and then retain every 300-th sample.

- *Spatially discrete approximation (SDA) to LGCP.* We fit the approximation in (3) using a population weighted average (SDA I, with $w_i(x) = m(x)/m_i$) and simple average (SDA II, with $w_i(x) = 1/|\mathcal{R}_i|$) of the log-intensity. For both, we use the MCML method described in Section 3.1 and run 110,000 iterations of the MCMC algorithm with a burn-in of 10,000 samples and then retain every 10-th sample.

We summarise the results from the simulation study through the bias, root-mean-square-error (RMSE) and the 95% coverage probability (CP) for the incidence at LSOA level, λ_i, and for the spatially continuous relative risk, $\exp\{S(x)\}$. Let $\lambda_i^{(j)}$ denote the true simulated incidence for $\mathcal{R}_i$ at the j-th simulation; hence

$$BIAS = \frac{1}{nB} \sum_{i=1}^{n} \sum_{j=1}^{B} (\hat{\lambda}_i^{(j)} - \lambda_i^{(j)}),$$

$$RMSE = \sqrt{\frac{1}{nB} \sum_{i=1}^{n} \sum_{j=1}^{B} (\hat{\lambda}_i^{(j)} - \lambda_i^{(j)})^2},$$

$$CP = \frac{1}{nB} \sum_{i=1}^{n} \sum_{j=1}^{B} I(\lambda_i^{(j)} \in CI_{0.95}^{(j)}),$$

where $\hat{\lambda}_i^{(j)}$ is the mean of the predictive distribution for $\lambda_i^{(j)}$ and $I(\lambda_i^{(j)} \in CI_{0.95}^{(j)})$ is an indicator function that takes value 1 if $\lambda_i^{(j)}$ falls inside the 95% prediction interval and 0 otherwise. Similarly, we compute the three indices for the relative risk $\exp\{S(x)\}$ by averaging each of these over the regular grid at a spacing of 300 metres covering the whole of Newcastle-Upon-Tyne, UK.

Table 1 reports the results for the prediction of λ_i, the incidence at LSOA level. We observe that SDA I and II have a slightly lower bias and RMSE than LGCP in all three scenarios, with SDA I having the best performance. The coverage probability is close to the 95% nominal level for all three models.

The results for the spatially continuous relative risk, $\exp\{S(x)\}$, are shown in Table 2. LGCP has the lowest bias and RMSE followed by SDA I in all three scenarios, with larger differences for $\phi = 800$ and $\phi = 1500$. Both SDA I and II are more conservative than LGCP and provide prediction intervals with a larger coverage than the nominal level, as the result of a large RMSE. We also observe that the use of the population weighted average in SDA I leads to a tangible reduction in RMSE and bias with respect to SDA II.

5 Application: mapping of primary biliary cirrhosis risk

We analyse incidence data on primary biliary cirrhosis (PBC) in Newcastle-Upon-Tyne, UK, obtained from the original study carried out by Prince et al. (2001); the data-set is freely available from the `lgcp` R package. The data consist of geo-referenced cases of definite or probable PBC between 1987 and 1994. The objective of this analysis is to quantify the difference in the predictive inferences between the gold-standard LGCP model and the proposed spatially discrete approximation (or SDA), on PBC incidence at LSOA level and the spatially continuous relative risk surface. In the case of SDA, we fit the population weighted (SDA I) and simple average (SDA II) versions described in the previous section. For the sake of completeness, we also fit the Besag et al. (1991) model, one of most commonly used approaches in small area estimation, with linear predictor

$$\log \lambda_i = d_i^\top \beta^* + S_i + Z_i$$

where S_i is a zero-mean intrinsically autoregressive process with variance σ^2 and Z_i is Gaussian noise with variance τ^2. In all four models, we use the following set of covariates, which are available as piecewise constant at LSOA level: (1) proportion of males, (2) proportion of the population experiencing income deprivation, (3) proportion of the population experiencing employment deprivation, (4) deprivation due to poor education, (5) barriers to housing and services index, (6) deprivation due to crime, (7) living environment deprivation. Each of these variables is publicly available from the UK Government online archives (webarchive.nationalarchives.gov.uk). The regression coefficients are denoted by β_i in the LGCP model and by β_i^* in the BYM and SDA models, for $i = 0, 1, \ldots, 7$, with $i = 0$ corresponding to the intercept.

For the SDA models, we run 110,000 iterations of the MCMC algorithm with a burn-in of 10,000 samples, and then retain every 10-th sample. We discretise ϕ using 100 equally spaced values between 50 and 2000 meters. The resulting 10,000 samples show a good mixing of the chain and small autocorrelation at lags smaller than 4 (Appendix B of the supplementary material).

For the LGCP approach, we specify independent priors as follows: $\log \sigma \sim N(\log 1, 0.15)$, $\log \phi \sim N(\log 500, 2)$ and $(\beta_0, \dots, \beta_7) \sim MVN(0, 10^6 I)$. We run 3,100,000 iterations of the MCMC algorithm with a burn-in of 100,000 samples and retain every 300-th sample so as to obtain a set of 10,000 weakly dependent samples. Trace plots and correlograms indicate convergence of the MCMC algorithm for both the random effects and the model parameters (Appendix C of the supplementary material).

Fitting of the BYM model is carried out by iterating the MCMC algorithm 1,100,000 times with a burn-in of 100,000 samples and retaining every 100-th sample. The mixing diagnostic plots for the MCMC chain are shown in Appendix D of the supplementary material.

Tables 3 to 5 report the point and intervals estimates for the parameters of each of the fitted models. The differences amongst the point estimates of the regression coefficients from the four models are small except for deprivation due to education, although this is not statistically significant at the conventional 5% level in any of the four models.

Figure 1 shows a map of the estimates of PBC incidence at LSOA level from the four models. Each of these gives a qualitatively similar spatial pattern as also indicated by the scatter plots of Figure 2. The same holds for the standard errors of PBC incidence (Figure 3).

Figure 4 shows the map of the estimated continuous relative risk surface $\exp\{S(x)\}$ over a 300×300 meters regular grid covering the whole of the study area. Figures 5 and 6 indicate that the point estimates from the LGCP and SDA models are strongly similar but the standard errors from the latter are larger. This is consistent with our results from the simulation study of the previous section.

6   Discussion

In this article we have developed a spatially discrete approximation (SDA) to log-Gaussian Cox process (LGCP) models in order to carry out spatial prediction of disease risk at any desired spatial scale using spatially aggregated disease count data.

As variation in disease risk occurs in a spatial continuum irrespective of the format in which the data are available, we consider the LGCP framework to be a natural statistical paradigm for modelling aggregated disease count data. However, when computational constraints make the fitting of an LGCP infeasible, we argue that SDA provides a computationally efficient solution while respecting the spatially continuous nature of disease risk. SDA also overcomes some of the limitations inherent to other spatially discrete models, such as CAR models. In addition to providing spatially continuous predictions, SDAs can also deal with the issue of changing administrative boundaries over time and allow incorporation of covariates available at any spatial scale.

Kelsall and Wakefield (2002) developed a similar approach to the proposed SDA for modelling count data available at areal level. Specifically, by assuming

an intercept-only model, they approximate (4) using a multivariate log-Gaussian distribution with mean

$$E[\lambda_i] = \exp\{\beta_0 + \sigma^2/2\}$$

and covariance

$$\mathrm{Cov}\{\lambda_i, \lambda_j\} = \exp\{\beta_0 + \sigma^2/2\} \times \left[\int_{\mathcal{R}_i} \int_{\mathcal{R}_j} w_i(x) w_j(x') \exp\{\sigma^2 \rho(\|x - x'\|; \phi)\} \, dx \, dx' - 1\right].$$

Kelsall and Wakefield (2002) then advocate the use of the log-Gaussian approximation as a Bayesian prior for spatial smoothing but no reference is made to the LGCP framework. In this paper, instead, our objective was to develop a computationally efficient approximation to the LGCP model which, in Bayesian terms, is our chosen prior for modelling disease risk.

In fitting SDA models, most of the computational burden is due to the approximation of the integral in (7), which defines the area-level correlation between the spatial random effects. In our example, the SDA model is about 5 to 15 times faster to fit than the LGCP model, depending on the number of values used to discretise the scale of the spatial correlation ϕ.

We conclude that SDA is a reliable approximation to LGCP for carrying out predictions at areal-level, both in terms of point predictions and in the quantification of uncertainty. It also provides spatially continuous predictions in disease risk that are comparable to those from LGCP, but with larger standard errors and more conservative predictions intervals.

Finally, extension to the spatio-temporal case of the method discussed in this paper is possible and is work in progress. For example, let us consider counts y_{it} for the region $\mathcal{R}_i$ over the time interval $(t, t+1)$. Let $S(x, t)$ be a spatio-temporal Gaussian process with covariance function

$$\mathrm{cov}\{S(x, t), S(x', t')\} = \sigma^2 \exp\{-|t - t'|/\psi\} \exp\{-\|x - x'\|/\phi\}.$$

By modelling the y_{it} as realisations of a spatio-temporal log-Gaussian Cox process with conditional intensity $\Lambda(x, t) = \exp\{\alpha + S(x, t)\}$, we can then approximate this with a spatio-temporally discrete Gaussian process $S_t^* = (S_{1t}^*, \ldots, S_{nt}^*)$, such that

$$S_t^* = \varphi S_{t-1}^* + W_t, 0 < \varphi < 1,$$

where the temporal innovation W_t is modelled as a multivariate Gaussian distribution with covariance matrix given by (7). Preliminary results suggest that the reduction in computing time with respect to a spatio-temporal LGCP model is substantially larger than that observed for the purely spatial scenario presented in this paper.
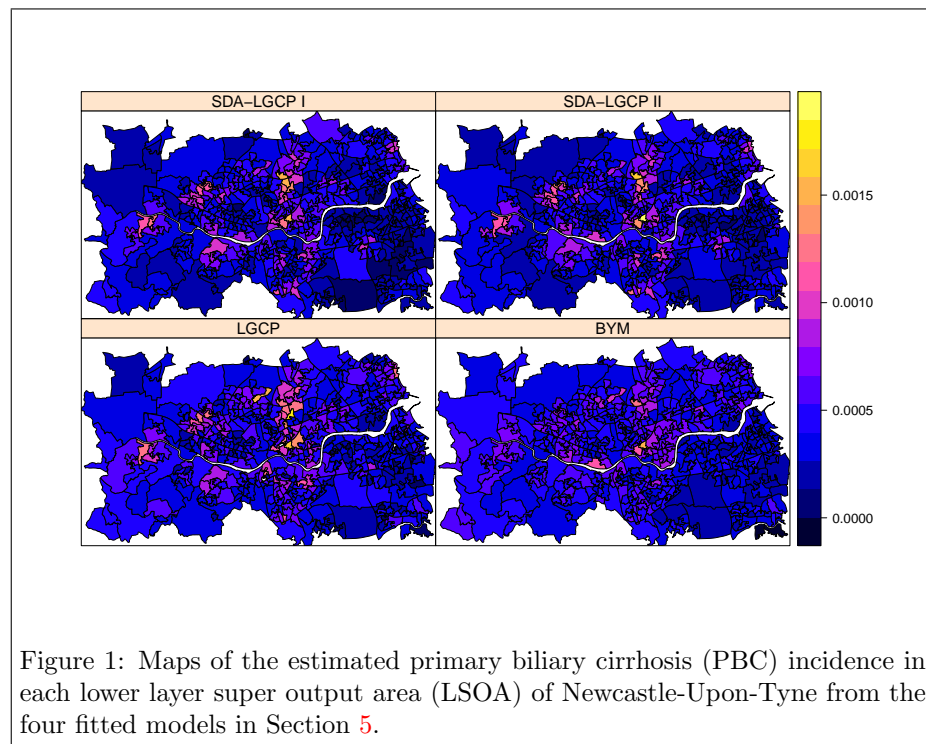
Acknowledgements

Funding

Figure 1: Maps of the estimated primary biliary cirrhosis (PBC) incidence in each lower layer super output area (LSOA) of Newcastle-Upon-Tyne from the four fitted models in Section 5.

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.
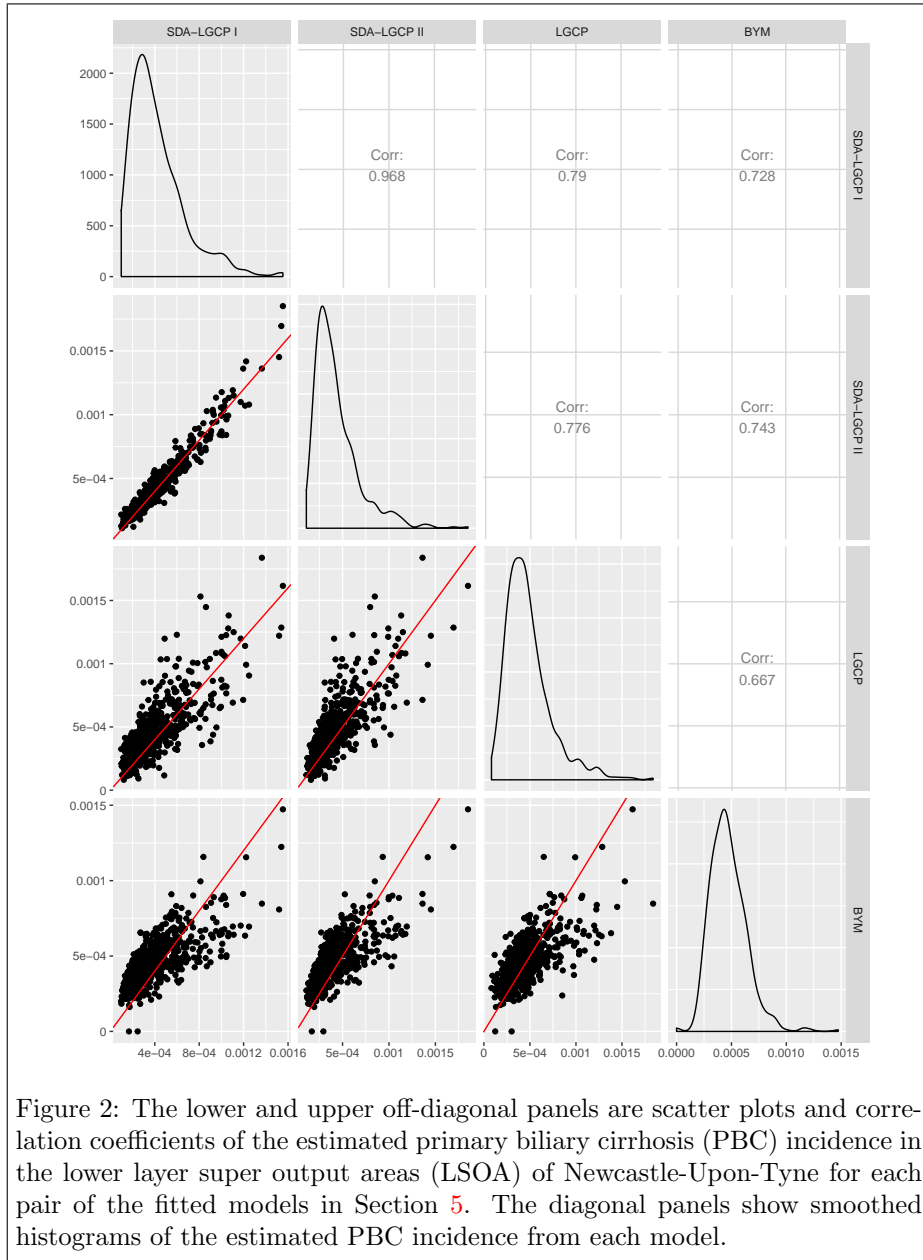
Figure 2: The lower and upper off-diagonal panels are scatter plots and correlation coefficients of the estimated primary biliary cirrhosis (PBC) incidence in the lower layer super output areas (LSOA) of Newcastle-Upon-Tyne for each pair of the fitted models in Section 5. The diagonal panels show smoothed histograms of the estimated PBC incidence from each model.
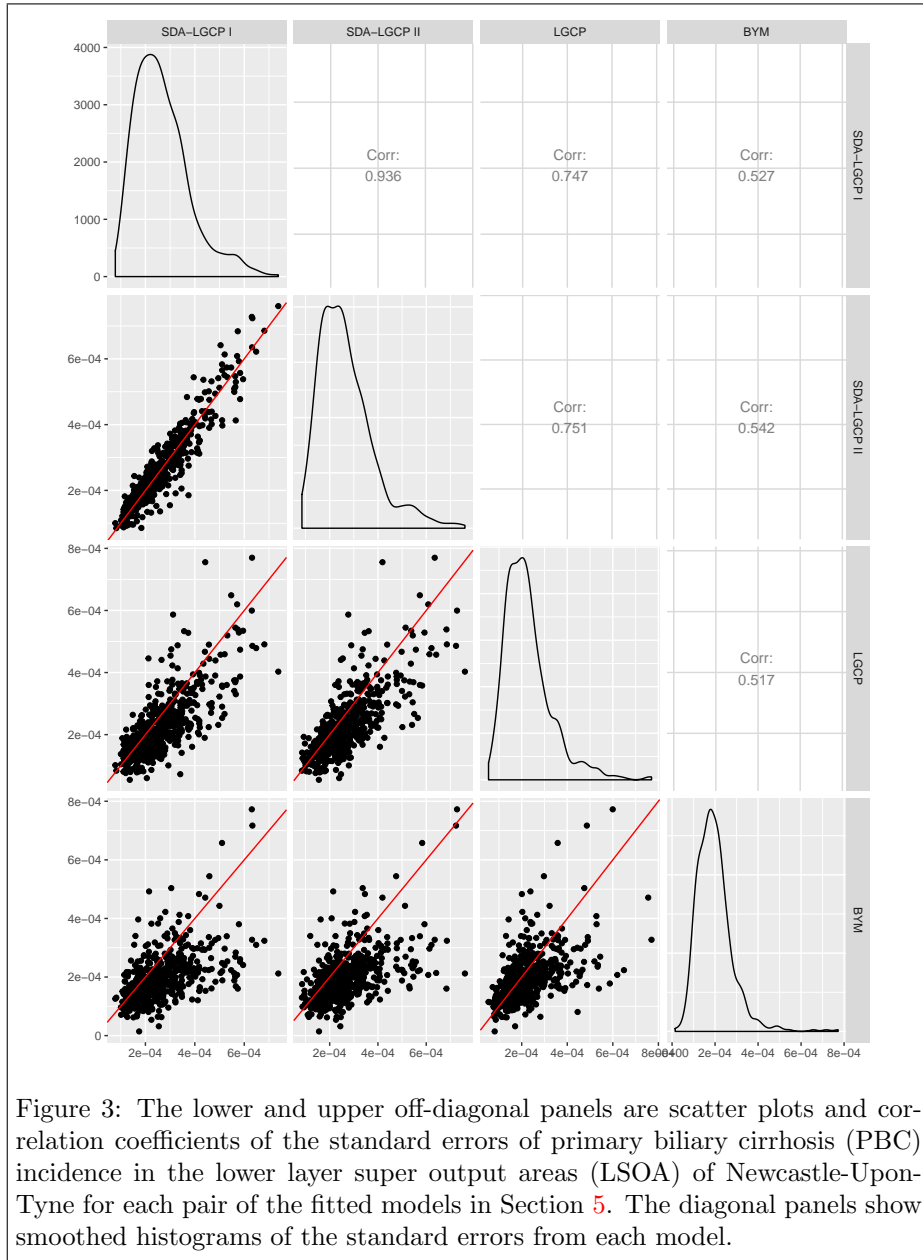
Figure 3: The lower and upper off-diagonal panels are scatter plots and correlation coefficients of the standard errors of primary biliary cirrhosis (PBC) incidence in the lower layer super output areas (LSOA) of Newcastle-Upon-Tyne for each pair of the fitted models in Section 5. The diagonal panels show smoothed histograms of the standard errors from each model.
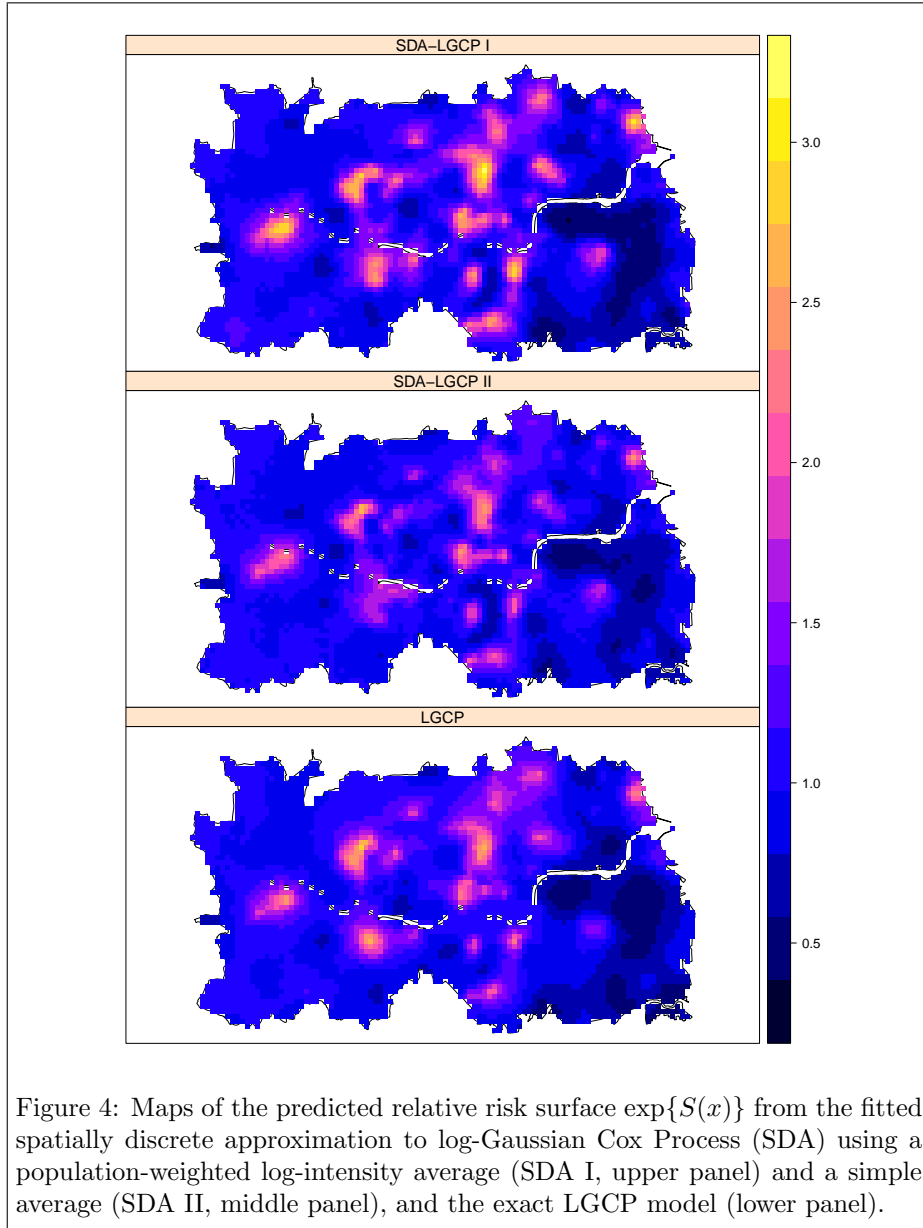
Figure 4: Maps of the predicted relative risk surface $\exp\{S(x)\}$ from the fitted spatially discrete approximation to log-Gaussian Cox Process (SDA) using a population-weighted log-intensity average (SDA I, upper panel) and a simple average (SDA II, middle panel), and the exact LGCP model (lower panel).
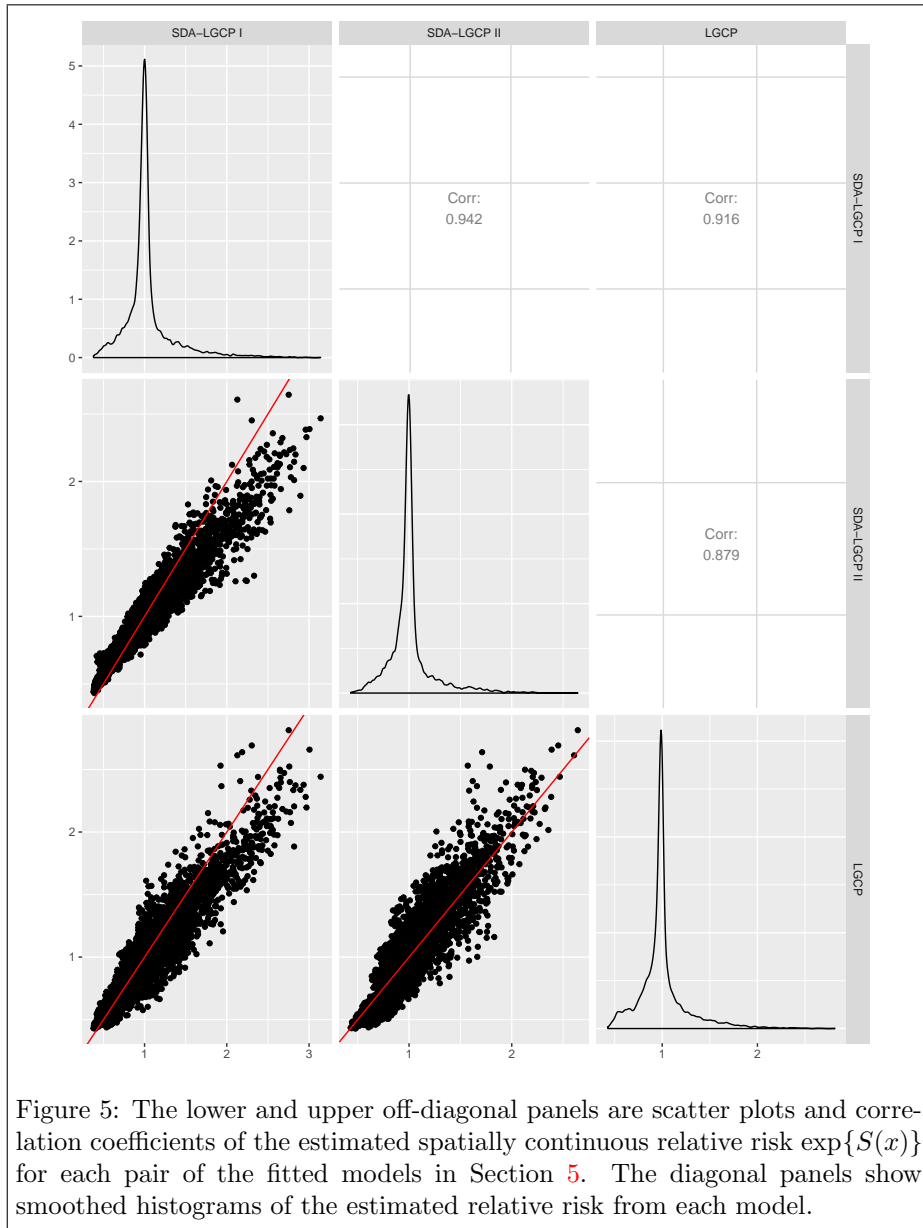
16

Figure 5: The lower and upper off-diagonal panels are scatter plots and correlation coefficients of the estimated spatially continuous relative risk $\exp\{S(x)\}$ for each pair of the fitted models in Section 5. The diagonal panels show smoothed histograms of the estimated relative risk from each model.
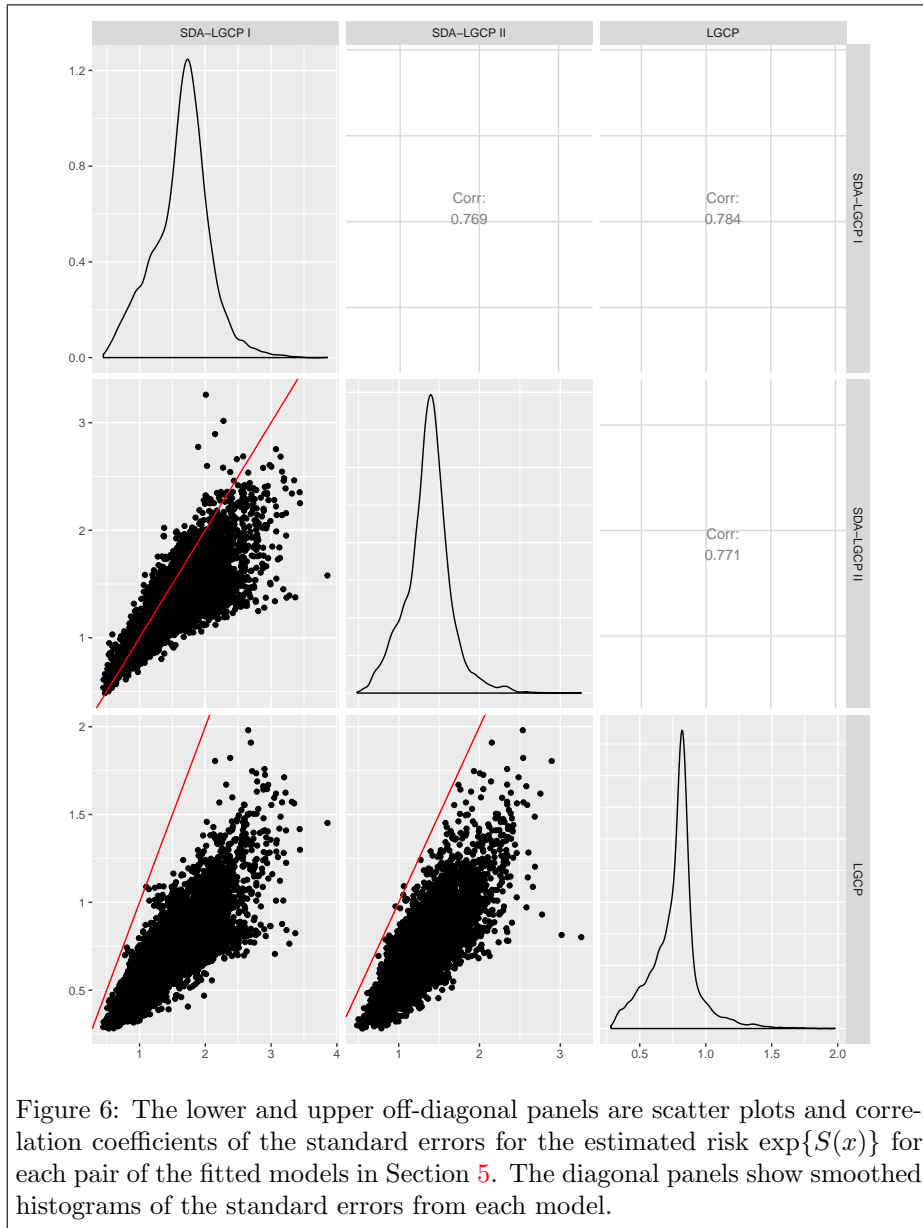
Figure 6: The lower and upper off-diagonal panels are scatter plots and correlation coefficients of the standard errors for the estimated risk $\exp\{S(x)\}$ for each pair of the fitted models in Section 5. The diagonal panels show smoothed histograms of the standard errors from each model.

Brood, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neihbor systems. *Biometrika*, 51:481–483.

Christensen, O. F. (2012). Monte carlo maximum likelihood in model-based geostatistics. *Journal of computational and graphical statistics.*

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 129–164.

Cressie, N. A. (1993). Statistics for spatial data: Wiley series in probability and mathematical statistics. *Applied probability and statistics, rev. edn, New York: Wiley.*

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns.* CRC Press.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, pages 542–563.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B*, 56:261–274.

Geyer, C. J. (1996). Estimation and optimization of functions. In Gilks, W., Richardson, S., and Spiegelhalter, D., editors, *Markov Chain Monte Carlo in Practice*, pages 241–258. London: Chapman and Hall.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B*, 54:657–699.

Giorgi, E. and Diggle, P. J. (2017). PrevMap: An R package for prevalence mapping. *Journal of Statistical Software*, 78(8):1–29.

Johnson, O., Giorgi, E., and Diggle, P. (2018). *SDALGCP: Spatially Discrete Approximation to Log-Gaussian Cox Processes for Aggregated Disease Count Data.* R package version 0.1.0.

Kelsall, J. and Wakefield, J. (2002). Modeling spatial variation in disease risk: a geostatistical approach. *Journal of the American Statistical Association*, 97(459):692–701.

Li, Y., Brown, P., Gesink, D. C., and Rue, H. (2012). Log gaussian cox processes and spatially aggregated disease incidence data. *Statistical methods in medical research*, 21(5):479–507.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482.

Murdock, A., Harfoot, A., Martin, D., Cockings, S., and Hill, C. (2015). Open-popgrid: an open gridded population dataset for england and wales. open-popgrid.geodata.soton.ac.uk. GeoData, University of Southampton.

Prince, M. I., Chetwynd, A., Diggle, P., Jarner, M., Metcalf, J. V., and James, O. F. (2001). The geographical distribution of primary biliary cirrhosis in a well-defined cohort. *Hepatology*, 34(6):1083–1088.

Rue, H. and Held, L. (2005). *Gaussian random Markov fields*. Chapman & Hall/CRC.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Tanemura, M. (1979). On random complete packing by discs. *Annals of the Institute of Statistical Mathematics*, 31(1):351–365.

Taylor, B., Davies, T., Rowlingson, B., and Diggle, P. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-gaussian cox processes in r. *Journal of Statistical Software*, 63:1–48.

Wakefield, J. and Shaddick, G. (2006). Health-exposure modeling and the ecological fallacy. *Biostatistics*, 7(3):438–455.

Wall, M. M. (2004). A close look at the spatial structure implied by the car and sar models. *Journal of statistical planning and inference*, 121(2):311–324.

Table 1: Average bias, root-mean-square-error (RMSE) and the 95% coverage probability (CP) for the LSOA incidence, λ_i, from the simulation study of Section 4.

| | $\phi = 100$ | | | $\phi = 800$ | | | $\phi = 1500$ | | |
	SDA I	SDA II	LGCP	SDA I	SDA II	LGCP	SDA I	SDA II	LGCP
Bias	-0.006	-0.007	-0.009	-0.002	-0.003	-0.004	-0.008	-0.008	-0.011
RMSE	0.020	0.021	0.022	0.003	0.004	0.006	0.027	0.029	0.030
95%CP	0.960	0.962	0.948	0.962	0.963	0.952	0.963	0.967	0.945

Table 2: Average bias, root-mean-square-error (RMSE) and the 95% coverage probability (CP) for the spatially continuous relative risk, $\exp\{S(x)\}$, from the simulation study of Section 4.

	$\phi = 100$			$\phi = 825$			$\phi = 1500$		
	SDA I	SDA II	LGCP	SDA I	SDA II	LGCP	SDA I	SDA II	LGCP
Bias	-0.575	-0.582	-0.566	0.842	0.965	-0.108	0.299	0.316	0.227
RMSE	2.590	2.800	0.045	0.439	0.531	0.005	2.070	2.260	0.002
95%CP	0.988	0.990	0.940	0.979	0.983	0.948	0.975	0.982	0.942

Table 3: Point estimates and 95% confidence intervals (CI) for the model parameters of the spatially discrete approximation to log-Gaussian Cox processes based weighted population average (SDA I) and simple average (SDA II) of the log-intensity.

	SDA I		SDA II	
Parameter	Estimate	95% CI	Estimate	95% CI
σ	1.054	(0.995, 1.109)	1.061	(1.003, 1.116)
ϕ	878.947	-	626.316	-
β_0^*	-6.267	(-7.196, -5.337)	-5.171	(-6.152, -4.190)
β_1^*	-3.474	(-5.329, -1.619)	-5.887	(-7.888, -3.886)
β_2^*	-2.377	(-3.284, -1.469)	-2.377	(-3.284, -1.469)
β_3^*	4.634	(3.355, 5.913)	6.510	(5.129, 7.891)
β_4^*	0.003	(-0.001, 0.007)	-0.002	(-0.007, 0.002)
β_5^*	-0.024	(-0.031, -0.017)	-0.016	(-0.022, -0.009)
β_6^*	-0.040	(-0.123, 0.042)	0.026	(-0.063, 0.114)
β_7^*	0.006	(0.001, 0.010)	0.001	(-0.004, 0.007)

Table 4: Point estimates and 95% credible intervals (CRI) for the model parameters of the Besag-York-Mollié model.

Parameter	Estimate	95% CRI
τ^2	0.094	(0.017, 0.366)
ν^2	0.017	(0.003, 0.123)
β_0^*	-4.858	(-7.461, -2.203)
β_1^*	-6.153	(-11.806, -0.679)
β_2^*	-2.951	(-5.701, -0.254)
β_3^*	6.830	(3.106, 10.624)
β_4^*	0.001	(-0.011, 0.012)
β_5^*	-0.021	(-0.038, -0.003)
β_6^*	-0.068	(-0.283, 0.154)
β_7^*	0.004	(-0.010, 0.017)

Table 5: Posterior median and 95% credible intervals (CRI) for the model parameters of the log-Gaussian Cox process model.

Parameter	Posterior median	95% CRI
σ	0.706	(0.484, 0.963)
ϕ	986.668	(359.716, 2329.131)
β_0	-14.499	(-18.624, -10.279)
β_1	-9.573	(-18.587, -0.676)
β_2	-1.956	(-6.445, 2.451)
β_3	5.710	(-0.422, 11.849)
β_4	-0.001	(-0.020, 0.018)
β_5	-0.031	(-0.058, -0.006)
β_6	-0.118	(-0.434, 0.206)
β_7	0.013	(-0.009, 0.034)