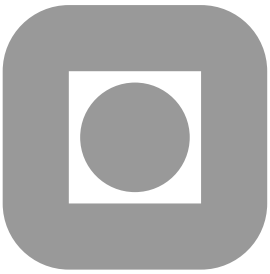
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Modelling spatial variation in disease risk using Gaussian Markov random field proxies for Gaussian random fields

by

Turid Follestad and Håvard Rue

PREPRINT
STATISTICS NO. 3/2003



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
http://www.math.ntnu.no/preprint/statistics/2003/S3-2003.ps
Turid Follestad has homepage: http://www.math.ntnu.no/~follesta
E-mail: follesta@stat.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology,
N-7034 Trondheim, Norway.

Modelling spatial variation in disease risk using Gaussian Markov random field proxies for Gaussian random fields

Turid Follestad and Håvard Rue

Department of Mathematical Sciences
Norwegian University of Science and Technology

Abstract

Analyses of spatial variation in disease risk based on area-level summaries of disease counts are most often based on the assumption that the relative risk is uniform across each region. Such approaches introduce an artificial piecewise-constant relative risk-surface with discontinuities at regional boundaries. A more natural approach is to assume that the spatial variation in risk can be represented by an underlying smooth relative risk-surface over the area of interest. This approach was taken by Kelsall and Wakefield (2002), who used an underlying Gaussian random field (GRF) to derive a multivariate log-Normal distribution of the risk at the regional level. The derivation rely on the approximation $\sum_i \exp(x_i) \approx \exp(\sum_i x_i)$, which is frequently used in similar contexts in the geostatistics literature, but the different sizes and shapes of the regions typically found in disease mapping applications indicate that the validity of the approximation is questionable.

We propose an approach to the modelling of a smoothly varying risk surface based on aggregated data avoiding this approximation. We also derive computationally efficient block MCMC-algorithms using a re-formulation of the geostatistical GRF model using Gaussian Markov random fields (GMRFs). We make extensive use of recent developments for GMRFs, including a method for fitting GMRFs to Gaussian random fields and computationally efficient algorithms for GMRFs based on numerical methods for sparse matrices. We demonstrate our approach on simulated data as well as a set of German oral cavity cancer mortality data from the period 1986–90, which have been previously analysed in the literature.

1 Introduction

Disease maps displaying the geographical variability of disease incidence or mortality rates across a region of interest, are valuable tools in spatial epidemiology. By studying a disease map, regions with particularly high or low rates can be identified, and this information can be used as input to ethological studies as a guideline in defining and validating hypotheses about a disease. For an overview of the history of disease mapping, see e.g. Walter (2000). Disease incidence or mortality data can be available as point data for which the exact location of each case is known, or more commonly as aggregated or areal summary data, often due

to confidentiality reasons. For rare and non-infectious diseases, the aggregated incidence or mortality counts $y_i;\ i = 1, \ldots, m$ in a set of m regions are commonly assumed to be conditionally independent given the stratum-weighted relative risks R_i of the regions, and to follow Poisson distributions with mean given by $E_i R_i$. The value E_i represents the expected number of cases in region i, typically given as a population-weighted sum of stratum-specific probabilities of disease, computed from the data assuming uniform risk across the study area. The maximum likelihood estimate of the relative risk in region i is the standardised mortality (or incidence) ratio SMR y_i/E_i. Figure 1 shows the observed aggregated counts and SMR for a set of data on mortality from oral cavity cancer in Germany, that will be analysed in Section 6 (Knorr-Held and Raßer, 2000). From Figure 1 we observe that there is a tendency toward high risk in the north-east and in the south-west and low risk in the east. However, for small populations at risk and for rare diseases, the SMR as an estimator

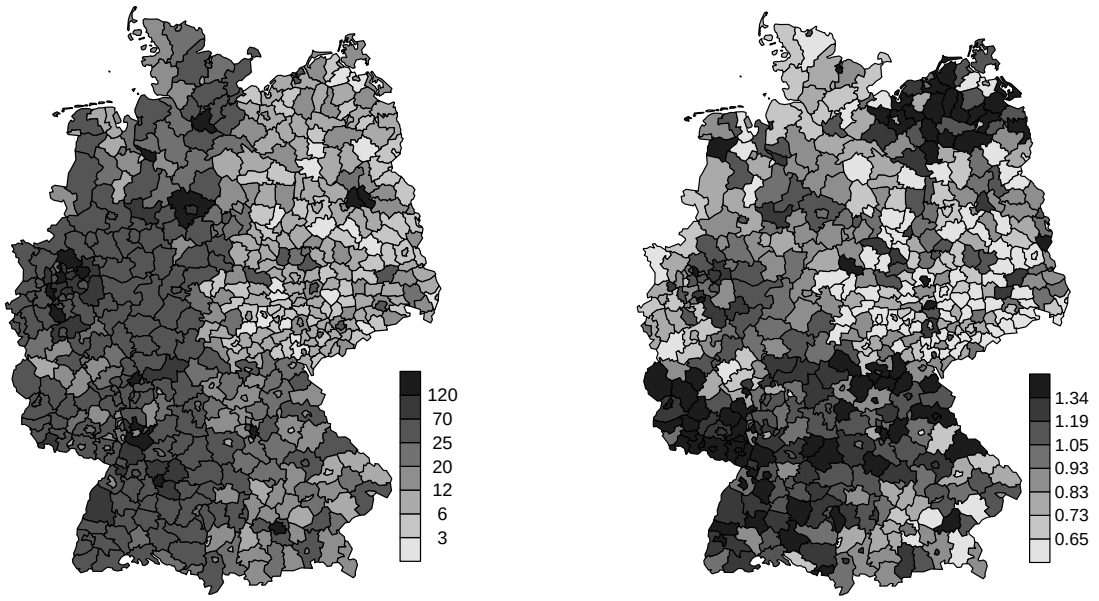


Figure 1: The observed counts (left) and the standardised mortality ratio (SMR) (right) for the German oral cavity cancer data.

of the relative risk can be highly variable. It can give rise to spurious estimates of high risk in regions with low populations, masking the true spatial pattern of the risk over the area of interest. Therefore, conclusions drawn from maps of the SMR can be misleading. To overcome this problem, a number of authors have developed statistical approaches to improve on raw estimates of disease risk. Reviews of statistical methods for mapping disease risk are provided by e.g. Diggle (2000), Wakefield, Best and Waller (2000) and Mollié (1996), the latter two focusing on Bayesian approaches.

Taking a Bayesian approach, the risk estimates of sparsely populated or low frequency regions are smoothed toward an overall prior mean. Since it is often the case that the relative risk tend to be similar in neighbouring regions, disease maps can also be improved on by adding spatial correlation to the prior model. The estimates of the risk in each region can then "borrow strength" from neighbouring regions. This can be accommodated by includ-

ing a spatially structured component within a random effects model for the disease risk, an approach first taken by Clayton and Kaldor (1987). A commonly used approach, proposed by Besag, York and Mollié (1991), is to model the log relative risk as

$$\log(R_i) = \boldsymbol{\beta}^T \boldsymbol{z}_i + u_i + v_i, \tag{1}$$

where $\boldsymbol{z}_i$ is a vector of covariates, including an intercept term, and u_i and v_i are spatially structured and unstructured random effects, respectively. The spatially structured random effect is assigned a Gaussian Markov random field prior, such that $f(u_i|\boldsymbol{u}_{-i}) = f(u_i \mid \boldsymbol{u}_{\delta i})$, where $\boldsymbol{u}_{-i}$ denotes all elements of the vector $\boldsymbol{u}$ except element i, and $\delta(i)$ is the set of neighbouring regions of region i. To specify the Markov random field prior, we need to define which regions are neighbours. The level of aggregation of areal summary data is often defined by administratively specified regions, and therefore alternative definitions to the square neighbourhoods often used in the case of lattice data are needed. An approach taken by many authors, e.g. Clayton and Kaldor (1987), Bernardinelli, Pascutto, Best and Gilks (1997), Knorr-Held and Besag (1998) and Waller, Carlin, Xia and Gelfand (1997), is to define two regions as neighbours if they share a common boundary. This will work well if the regions do not differ much in size and shape, but this is often not the case. An alternative to the adjacency approach is to specify the joint distribution of the heterogeneity effects u_i, defining spatial structure of the covariance matrix as a function of differences between the region centres (e.g. Wakefield and Morris, 2001; Wakefield and Morris, 1999). However, similar objections apply to this method as to the adjacency based methods, as the size and shape of the regions are still not taken into account, and in both cases the inference will depend on the level of aggregation of the data.

The method of Besag et al. (1991) does not naturally allow for discontinuities in the spatial structure of the risk. In a recent paper, Fernández and Green (2002) present an alternative approach, developing a spatially structured mixture model where GMRF priors are specified for the weights in the mixture. Using a mixture of Poisson distributions, the method is applied in a disease mapping context, and it is illustrated how the approach represents an improvement over the method of Besag et al. (1991) in cases where the spatial pattern has step-like discontinuities. The approach is related to that of Knorr-Held and Raßer (2000) identifying clusters of constant risk.

In general, spatial heterogeneity of the disease risk will be a confounder for unmeasured spatially structured factors influencing the disease risk. In most cases, there is no reason that these risk factors are region specific and discontinuous at region boundaries. Thus, the relative risk is not expected to be constant within regions and disjoint across regions. On the contrary, it seems reasonable to believe that the underlying risk surface is varying continuously over the region of study. In cases where the observations can be regarded as point data, a smoothly varying risk surface and the corresponding hyper-parameters can be estimated using extensions of classical geostatistical or point process approaches. Using data for which the exact locations are known, Diggle, Tawn and Moyeed (1998) propose a model-based geostatistical approach embedding the classical linear geostatistical methods for Gaussian data within a framework analogous to the generalised linear models (McCullagh and Nelder, 1989) for mutually independent data. Consequently, they allow for data for which the stochastic variation is assumed to be non-Gaussian. Another approach is taken by Best, Ickstadt and Wolpert (2000), who specify a Poisson-Gamma random field model for the disease risk. The approach is based on the methodological framework presented in Wolpert

and Ickstadt (1998) and extended in Ickstadt and Wolpert (1999) to include location-specific covariates measured at different levels of spatial aggregation and individual attributes like age and gender. The point locations of individual cases and the corresponding attributes are regarded as a marked point process, and the spatial structure of the residual risk surface is represented by a kernel smoothed Gamma random field. The risk surface can be estimated at any level of spatial aggregation.

When the disease incidence or mortality data are only available as aggregated counts, the approaches to risk-surface estimation described above are not directly applicable. Kelsall and Wakefield (2002) propose a geostatistical approach to modelling the joint distribution of the area-level relative risks in such situations. They specify a model for an underlying continuously varying risk surface $R(s);\ s \in \mathcal{A}$, assuming the log risk surface $S(s) = \log(R(s))$ to be a realisation of a Gaussian random field (GRF). Based on this GRF model, area-level relative risks R_i in a set of regions $\mathcal{A}_i;\ i = 1, \ldots, m$, forming a partition of the study region $\mathcal{A}$, are defined by

$$R_i = \int_{\mathcal{A}_i} R(s) f_i(s) ds, \tag{2}$$

where $f_i(s);\ i = 1, \ldots, m$ are weight functions depending on the stratum-specific population density distribution in region $\mathcal{A}_i$. Conditionally on these relative risks, the data are assumed to be independent realisations from a Poisson distribution with mean $R_i E_i$. To allow for computational feasibility, they approximate the joint distribution of the region-level risks $R_i;\ i = 1, \ldots, m$ by a multivariate log-Normal distribution with moments that are derived from the moments of the Normal distribution of $S(s)$, using numerical methods to evaluate the integrals involved. The approximation is essentially equivalent to approximating the distribution of $S_i = \log(R_i)$ by the distribution of $\int_{\mathcal{A}_i} \log R(s) f_i(s) ds$. As pointed out by the authors, the approximation is best when the size of the regions are relatively small and the regions are of about the same shape, and the log-Gaussian assumption is exact only in the limit when the regions are of the same shape and size, and the size tends to zero. The parameters of the model, including the log-risk surface at a set of locations s_k, are estimated by Markov chain Monte Carlo methods (Gilks, Richardson and Spiegelhalter, 1996), using Gibbs sampling in combination with adaptive rejection sampling.

We propose an alternative approach to the estimation of a smooth risk surface based on aggregated count data, representing the Gaussian random field defining the prior for the log-risk surface by a Gaussian Markov random field defined on a lattice. The basis of the model formulation is as in Kelsall and Wakefield (2002), but while they use the geostatistical model to derive an approximation to the joint distribution of the regional-level log-risk, and base the inference on the resulting regional-level stochastic model, we avoid the approximation by working directly on a lattice representation of the model. We replace the integral expression (2) for the regional level relative risk R_i by a sum over the exponentiated values of the GMRF for the lattice nodes falling within $\mathcal{A}_i$. Due to the conditional independence structure of the GMRF, using a GMRF proxy to the GRF allows for the use of computationally efficient algorithms for sampling based inference. However, the spatial structure is often intuitively easier to specify and interpret using a geostatistical GRF formulation than the conditional formulation represented by the GMRF. Therefore, we specify the spatial structure of the random field in terms of the correlation function for the GRF, using the procedure in Rue and Tjelmeland (2002) to fit the GMRF to the GRF. Thus, our approach relies on the assumptions

that the smooth relative risk surface can be represented on a lattice and that the GRF as defined on this lattice can be well estimated by a GMRF.

Drawing on the routines for fast and exact simulation of GMRF implemented in Rue and Follestad (2002), we develop an efficient block-sampling algorithm for estimating the log-risk surface and the parameters of the model. For each block, the elements of the lattice based log-risk surface are updated using a Metropolis-Hastings step, generating a proposal from a Gaussian approximation to the full conditional distribution. This can be done efficiently after re-formulating the problem of sampling from the proposal distribution to a computationally convenient conditional sampling problem.

The report is organised as follows. In Section 2 we present the statistical model, and an overview of our approach to estimating the log risk surface and the hyper-parameters is presented in Section 3. More details on the estimation algorithm are given in Section 4. In Section 5 we present results for a simulated data set, and results for the German oral cavity cancer data are given in Section 6. The method and the results are summarised and discussed in Section 7.

2 The statistical model

The statistical model is based on disease incidence or mortality data available as aggregated counts y_i in a set of m disjoint regions denoted $\mathcal{A}_i;\ i = 1, \ldots, m$, such that $\mathcal{A} = \cup_i \mathcal{A}_i$ is the overall region of study. Following the approaches of Kelsall and Wakefield (2002), Best et al. (2000) and Diggle et al. (1998) we assume that the geographical variation in the risk of the disease can be represented by a smoothly varying surface $R(s);\ s \in \mathcal{A}$. In this section we specify a lattice based model for the risk surface, first presenting a Gaussian random field model, and then a Gaussian Markov random field proxy to this model.

2.1 A Gaussian random field model on a lattice

The log-risk surface $S(s) = \log R(s)$ is assumed to be a realisation of a Gaussian random field. The basis of our modelling approach is as in Kelsall and Wakefield (2002), but we explicitly define the Gaussian random field model on a lattice overlaying the study region $\mathcal{A}$. Throughout our study, using simulated data as well as the real dataset, we use the 544 districts of Germany for which the German oral cavity data are defined as our region of interest. A map of the study region with an overlaying lattice consisting of $n_{GRF} = 16824$ nodes is given in Figure 2. For a better visual impression of the resolution of the lattice, see the top right panel of Figure 6. The number of lattice nodes within each region is in the range 1 to 136, with a median number of 29.

The multivariate Normal joint prior distribution of the log-risks $S(\boldsymbol{s}_j);\ j = 1, \ldots, n_{GRF}$ is

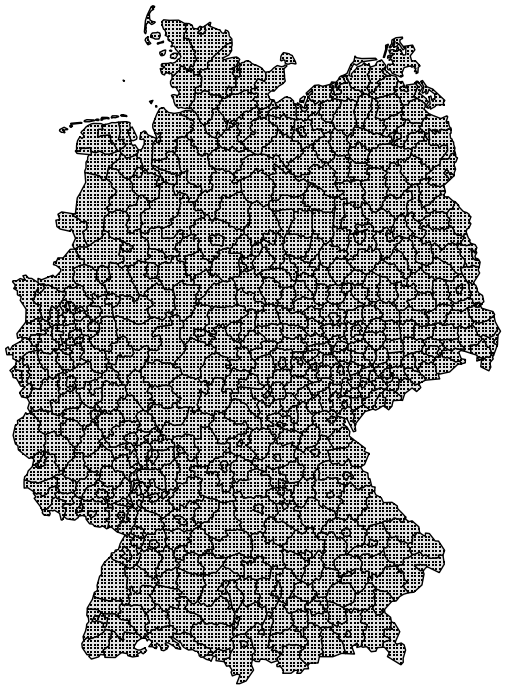


Figure 2: The map of Germany with its 544 districts, overlaid by the lattice for the GRF.

given by the moments

$$\begin{aligned} \mathrm{E}(S(\boldsymbol{s}_j)) &= \mu_j, \\ \mathrm{Var}(S(\boldsymbol{s}_j)) &= \sigma^2, \\ \mathrm{Corr}(S(\boldsymbol{s}_j), S(\boldsymbol{s}_k)) &= \rho(|\boldsymbol{s}_j - \boldsymbol{s}_k|; \boldsymbol{\theta}_\rho). \end{aligned} \tag{3}$$

The correlation function is assumed to be isotropic. The mean vector $\boldsymbol{\mu}$, the marginal variance σ^2 and the parameters $\boldsymbol{\theta}_\rho$ of the correlation function ρ are taken to be unknown and are assigned prior distributions as described in Section 3. Thus, our prior model of the risk surface and the corresponding hyper-parameters is the lattice analogue of the geostatistical model of Kelsall and Wakefield (2002). Given the log-risk surface, the data are assumed to be conditionally independent realisations from Poisson distributions given by

$$y_i \mid R_i \sim \mathrm{Pois}(E_i R_i), \tag{4}$$

where the regional level relative risks R_i are computed from the underlying lattice-based risk surface by

$$R_i = \sum_{j:\, \boldsymbol{s}_j \in \mathcal{A}_i} R(\boldsymbol{s}_j) w(\boldsymbol{s}_j). \tag{5}$$

Here, the population density distributions $f_i(s)$ of the continuous surface analogue (2) are replaced by a set of weights $w(\boldsymbol{s}_j)$ which should satisfy the constraint

$$\sum_{j:\, \boldsymbol{s}_j \in \mathcal{A}_i} w(\boldsymbol{s}_j) = 1; \forall i. \tag{6}$$

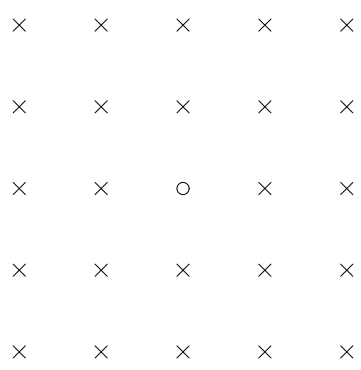


Figure 3: The neighbours ($\times$) of an arbitrary lattice node ($\circ$) using a 5×5 neighbourhood scheme for the GMRF.

In the approach by Kelsall and Wakefield (2002), the GRF prior model for the log-risk surface is used to generate a multivariate log-Normal approximation to the regional-level relative risks. To avoid computing such an approximation, we base inference directly on the risk surface model as defined on the lattice. However, for the lattice model to be a reasonable approximation to the smooth surface, the resolution should be relatively high, and consequently the number of nodes of the lattice will typically be large. Using the Gaussian random field representation of the prior model, estimation routines will be computationally expensive since we need to perform matrix operations on the $n_{GRF} \times n_{GRF}$ covariance matrix $\mathbf{\Sigma}$, which in general is a full matrix. Moreover, using MCMC methods with single-site updating, the convergence will be slow, due to the high correlations inherent in the prior model. Similar problems arise when estimating the hyper-parameters, because of the strong interaction with the elements of the risk surface. On the other hand, due to the high dimensionality and the full structure of the covariance matrix, updating all elements of the surface in one block will be prohibitive. In the next subsection we describe how the GRF can be represented by a more computationally convenient Gaussian Markov random field on the lattice.

2.2 A Gaussian Markov random field proxy to the Gaussian random field model

To reduce the computational cost of the inference, we propose to represent the log-risk surface by a vector variable $\boldsymbol{x} = \{x_j\}_{j=1,\ldots,n_{GRF}}$, which is assumed to be a realisation of a Gaussian Markov random field (GMRF). A GMRF is a GRF with the additional property that the conditional distribution of the GMRF at lattice node j, given the values at all other lattice nodes, only depends on the values at the nodes within a neighbourhood $\delta(j)$ of j. Different definitions of neighbourhoods are possible, but we choose to define the neighbourhood $\delta(j)$ of node j to be an 5×5 square neighbourhood, as illustrated in Figure 3. Since we are dealing with a finite lattice, the number of neighbours of the lattice nodes along the boundary of the lattice will be different from the number given by the 5×5 neighbourhood scheme, see Figure 4. To reduce the impact of any boundary effects induced by using a finite lattice, we extend the support of the GMRF to include a set of nodes outside the region of interest. We will denote the sub-vector corresponding to the nodes falling within the region of interest by

$\boldsymbol{x}_{\mathcal{A}}$, and the nodes external to this region by $\boldsymbol{x}_{-\mathcal{A}}$, such that $\boldsymbol{x} = (\boldsymbol{x}_{\mathcal{A}}^T, \boldsymbol{x}_{-\mathcal{A}}^T)^T$. The extended lattice, consisting of $n = 31089$ nodes, is shown in Figure 5.

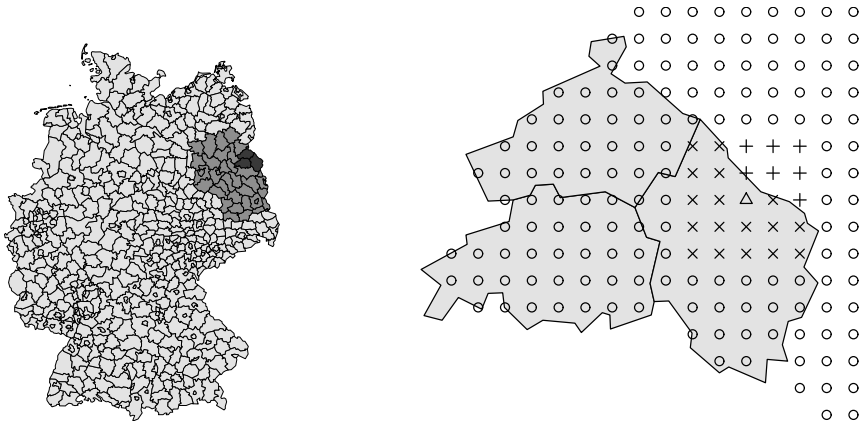


Figure 4: The neighbouring scheme along the boundary of the study region. The neighbours of a node (triangle) at the boundary are partly within the study region (x) and partly outside the region (+).

For a general GMRF $\boldsymbol{x}$, the joint distribution is given by

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}), \tag{7}$$

where the mean vector $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$ and the precision matrix $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\theta})$ both may depend on a set of unknown parameters $\boldsymbol{\theta}$. Because of the conditional independence structure of the GMRF, only the elements Q_{ij} of the precision matrix for which i and j are neighbours are non-zero. The nodes of the lattice can be re-ordered such as to minimise the bandwidth of the corresponding precision matrix (Knorr-Held and Rue, 2002), and due to the band-structure of the matrix, working with a GMRF instead of a GRF can lead to significant reductions in computational cost. This fact is utilised in our sampling based estimation approach described in Sections 3 and 4. There, we make extensive use of efficient algorithms for generating samples from joint and conditional distributions of a GMRF as well as for generating samples conditionally on linear constraints. A sample from the joint distribution of $\boldsymbol{x}$ can be generated by

$$\boldsymbol{x} = \boldsymbol{L}^{-1}\boldsymbol{z} + \boldsymbol{\mu}, \tag{8}$$

where $\boldsymbol{z}$ is a vector of n independent realisation from the standard Normal distribution, and $\boldsymbol{L}$ is the Cholesky factor of the precision matrix $\boldsymbol{Q}$. For a banded symmetric positive definite matrix $\boldsymbol{Q}$ with bandwidth b_w, the Cholesky factorisation

$$\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T \tag{9}$$

can be computed in $\mathrm{O}(nb_w^2)$ flops (Rue, 2001), such that as long as the bandwidth is kept small, efficient samples can be generated from the joint distribution. In our application, we need to generate conditional samples for a subset of the lattice nodes, given the realisation of the GMRF for the remaining nodes. The conditional distribution $\pi(\boldsymbol{x}_{\mathcal{S}}|\boldsymbol{x}_{-\mathcal{S}})$, where $\mathcal{S}$ is a

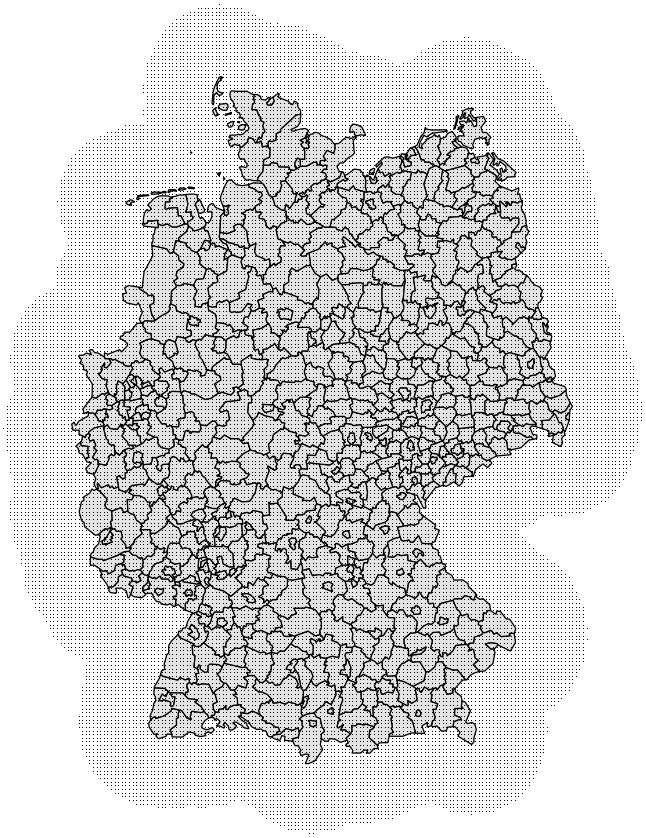


Figure 5: The map of Germany with its 544 districts, overlaid by the lattice GMRF model including a boundary region.

subset of $\mathcal{A}$, is Normal with moments

$$\mathrm{E}(\boldsymbol{x}_{\mathcal{S}} \mid \boldsymbol{x}_{-\mathcal{S}}) = \boldsymbol{\mu}_{\mathcal{S}} - \boldsymbol{Q}_{\mathcal{S},\mathcal{S}}^{-1} \boldsymbol{Q}_{\mathcal{S},-\mathcal{S}} \, (\boldsymbol{x}_{-\mathcal{S}} - \boldsymbol{\mu}_{-\mathcal{S}}), \tag{10}$$

$$\mathrm{Var}(\boldsymbol{x}_{\mathcal{S}} \mid \boldsymbol{x}_{-\mathcal{S}}) = \boldsymbol{Q}_{\mathcal{S},\mathcal{S}}^{-1}. \tag{11}$$

Here, $\boldsymbol{Q}_{\mathcal{S},\mathcal{S}}$ is the $n_{\mathcal{S}} \times n_{\mathcal{S}}$ diagonal block of $\boldsymbol{Q}$ corresponding to the subset $\mathcal{S}$, with bandwidth less than or equal to the bandwidth of $\boldsymbol{Q}$. Each element l of the mean vector $\boldsymbol{\mu}_{\mathcal{S}|-\mathcal{S}} = \mathrm{E}(\boldsymbol{x}_{\mathcal{S}} \mid \boldsymbol{x}_{-\mathcal{S}})$ will only depend on elements of $\boldsymbol{x}_{-\mathcal{S}} - \boldsymbol{\mu}_{-\mathcal{S}}$ at the nodes within the neighbourhood of node l.

We also need to generate samples from a GMRF $\boldsymbol{x}$ conditionally on a linear *soft* constraint $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} + \boldsymbol{\epsilon}$ for a $p \times n$ matrix $\boldsymbol{A}$, a p-vector $\boldsymbol{b}$ and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$. This constraint can be interpreted as a generalisation of the *hard* constraint $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where the quantities representing the linear combinations defining the constraint are observed with noise. In general, a sample conditionally on a soft constraint can be generated by first generating an unconditional sample $\boldsymbol{x}_u$ for $\boldsymbol{x}$ from (7) and an $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, and then computing the conditional sample $\boldsymbol{x}_c$ from

$$\boldsymbol{x}_c = \boldsymbol{x}_u - \boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T + \boldsymbol{\Sigma})^{-1}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{\epsilon} - \boldsymbol{b}). \tag{12}$$

In geostatistics, this result is referred to as conditional simulation using kriging (Cressie,

1993, Section 3.6.2) and the validity of (12) as a sample from $\pi(\boldsymbol{x}|\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} + \boldsymbol{\epsilon})$ follows directly from Normal distribution theory, as shown in Appendix A.3. As long as the number of constraints p is relatively small compared to the number of nodes in the lattice, all computations involved in evaluating (12) can be done efficiently using the Cholesky factorisation (9).

To fully specify the joint distribution (7) we need to specify the non-zero elements of the precision matrix $\boldsymbol{Q}$. However, based on prior information it is often intuitively easier to specify a model for the correlation structure for a Gaussian random field than to specify the elements of the corresponding precision matrix for the GMRF. Rue and Tjelmeland (2002) show how the elements of the precision matrix $\boldsymbol{Q}$ of a GMRF can be estimated from the covariance function that defines the elements of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$. Let $\rho(h; \boldsymbol{\theta}_\rho)$ be the correlation function specifying the correlation between two points of distance h, where h is measured in lattice coordinates. Further, let

$$\boldsymbol{Q} = \tau \boldsymbol{Q}' = \tau \boldsymbol{C}^{-1} \tag{13}$$

where $\boldsymbol{C}$ is the correlation matrix of the GRF and $\tau = 1/\sigma^2$ is the marginal precision. For a given value of the parameter vector $\boldsymbol{\theta}_\rho$, Rue and Tjelmeland (2002) estimate the non-zero elements of the standardised precision matrix $\boldsymbol{Q}'$ by matching the correlation function as defined by these elements to the correlation function ρ of a Gaussian random field. For the exponential, Gaussian, spherical and Matérn classes of correlation functions, they conclude that using a 5×5 neighbourhood the approach gives a good fit to the target correlation function. Among these four classes of functions, the exponential is the one with the best fit.

Using the GMRF prior model for the log relative risk surface, expression (5) for the relative risk R_i at the regional level is replaced by

$$R_i = \sum_{j \in \mathcal{A}_i} \exp(x_j)\ w(\boldsymbol{s}_j), \tag{14}$$

where the sum is taken over the n_i nodes of the lattice falling within region $\mathcal{A}_i$. In what follows we will assume that the weights are constants given by $w(\boldsymbol{s}_j) = 1/n_i;\ j \in \mathcal{A}_i$. This corresponds to an assumption of uniform population density which is often made in disease mapping applications. This does not represent any loss of generality, since the method can easily be modified to allow for non-uniform population distributions by replacing the GMRF $\boldsymbol{x}$ by another GMRF $\boldsymbol{x}'$ with elements $x'_j = x_j + \log(w(\boldsymbol{s}_j))$. In terms of the log-risk surface $\boldsymbol{x}$, the Poisson likelihood model for the incidence counts becomes

$$y_i \mid \boldsymbol{x} \sim \text{Pois}\left(\frac{E_i}{n_i} \sum_{j \in \mathcal{A}_i} \exp(x_j)\right). \tag{15}$$

For notational convenience we define $E'_i = E_i/n_i$, such that in what follows, E'_i is to be interpreted as the expected number of cases per lattice node falling within region $\mathcal{A}_i$.

Let $\boldsymbol{\theta} = (\tau, \boldsymbol{\theta}_\rho^T, \boldsymbol{\theta}_\mu^T)^T$ denote all unknown hyper-parameters of the model, including the precision τ, the parameters $\boldsymbol{\theta}_\rho$ of the correlation function and any parameters $\boldsymbol{\theta}_\mu$ defining the mean vector $\boldsymbol{\mu}$. We take a fully Bayesian approach to parameter estimation, and the Bayesian

Likelihood:	$y_i \mid \boldsymbol{x}$	$\sim$	$\text{Pois}(E_i' \sum_{j \in \mathcal{A}_i} \exp(x_j))$
GMRF prior:	$\boldsymbol{x} \mid \boldsymbol{\theta}$	$\sim$	$N(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{Q}(\boldsymbol{\theta})^{-1})$
Hyper-prior:	$\boldsymbol{\theta}$	$\sim$	$\pi(\boldsymbol{\theta})$

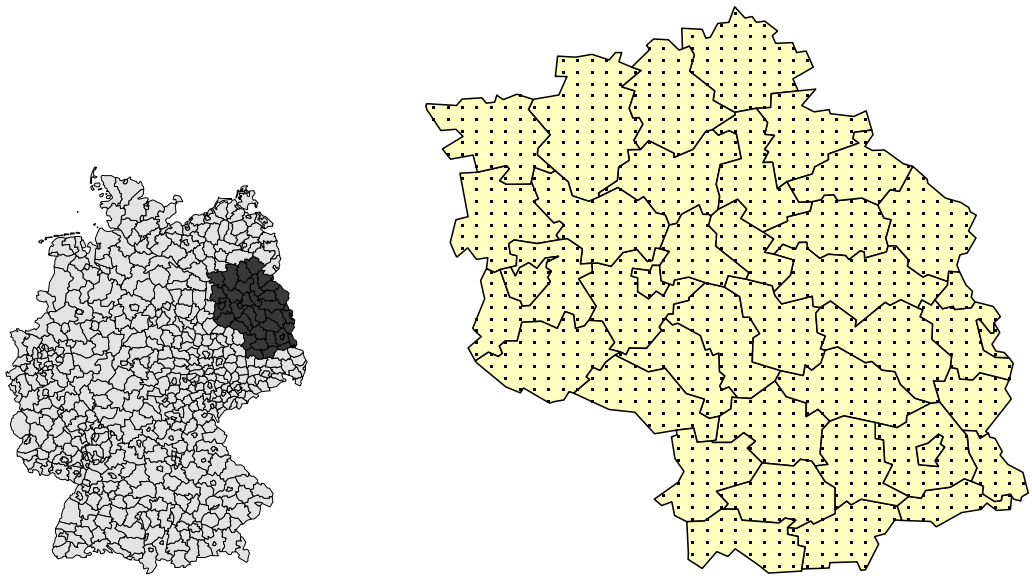
Table 1: A summary of the Bayesian hierarchical model.

hierarchical model for the disease mapping problem is summarised in Table 1. The prior distribution $\pi(\boldsymbol{\theta})$ of the hyper-parameters is specified in Section 3.

We end this section by pointing out some computational pitfalls that are still present using the GMRF representation of the model in combination with an MCMC based approach to parameter estimation. In our approach to estimation of the risk surface and the corresponding hyper-parameters we need to sample from the posterior distribution of the log-risk surface $\boldsymbol{x}$ given count data $\boldsymbol{y}$. In Knorr-Held and Rue (2002) it is illustrated how the use of block-sampling leads to substantial improvement in mixing for MCMC updating schemes for a similar model, but where the GMRF prior for the log-risk is defined on the same level of aggregation as the data, using a common boundary neighbourhood specification. The observations are conditionally independent given the regional risks and the hyper-parameters, and thus the conditional independence structure for the posterior is the same as for the GMRF prior. Since the data in our case are aggregated in regions that in general extends over the size of the local neighbourhoods of the GMRF model for $\boldsymbol{x}$, conditioning on the data will destroy the computationally convenient local neighbourhood structure inherent in the prior. For an illustration, consider the plots in the bottom panels of Figure 6. There, we have visualised the conditional independence structure of the prior model and of the posterior model conditioning on the data for a subset of the study region. The subset is given by the lattice nodes within the shaded region in the top left panel, plotted in larger scale and overlaid by the GMRF lattice in the top right panel. Consequently, to preserve computational efficiency, alternative methods are needed.

We have pointed out the potential problem of slow convergence of the hyper-parameters using MCMC methods with the Gaussian random field. Inference for the hyper-parameters will still be problematic using the lattice based GMRF modelling approach, unless we are able to update the hyper-parameters jointly with a subset of the GMRF. Finally, introducing the additional boundary region nodes to reduce the impact of boundary effects, might slow down the convergence and mixing for the hyper-parameters.

In the next sections we present our sampling based approach to parameter estimation, and discuss how the potential problems listed above can be handled within our sampling algorithm. We first give an overview of the method, and then present the approach in more detail in Section 4.



Conditional independence structure, prior model

Conditional independence structure, posterior model

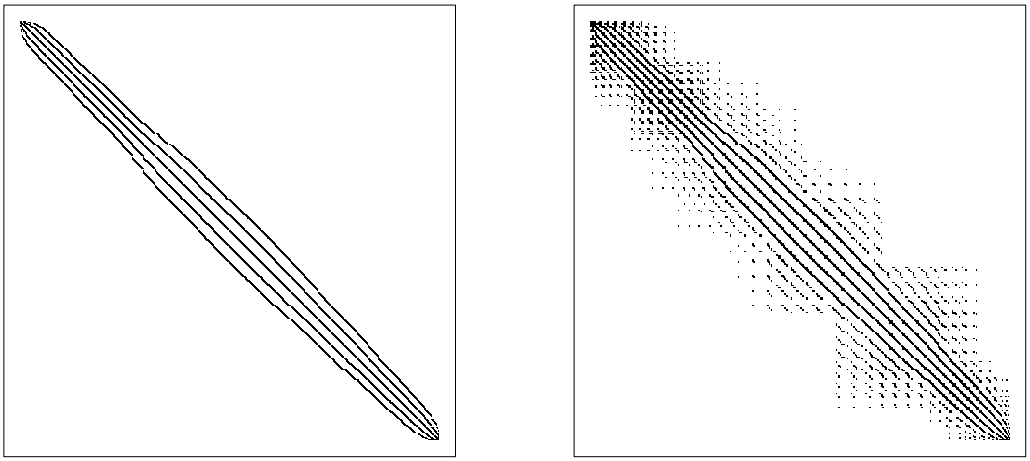


Figure 6: A subset of the region of study (shaded region, top left), with the lattice nodes added (top right). The bottom panels illustrate the conditional independence structure of the prior model (bottom left) and when conditioning on the data (bottom right), for the subset of lattice nodes corresponding to the shaded region.

3 A sampling based estimation approach

In this section we present a sampling based approach to the estimation of the unknown quantities of the model. These are the log risk surface represented by $\boldsymbol{x}$ and the hyper-parameters $\boldsymbol{\theta}$.

Given the Poisson likelihood, the GMRF prior (7) of $\boldsymbol{x}$ and the joint prior distribution $\pi(\boldsymbol{\theta})$ of the parameters $\boldsymbol{\theta}$, the joint posterior distribution of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ given the data $\boldsymbol{y}$ is given by

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \prod_i \{\pi(y_i \mid \boldsymbol{x}, \boldsymbol{\theta})\} \pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}). \tag{16}$$

We will estimate the log risk surface and the hyper-parameters using Markov chain Monte Carlo methods. As pointed out in Section 2, using single-site updating will typically lead to poor mixing and slow convergence due to strong correlations in the prior model. On the other hand, updating $\boldsymbol{x}$ or $\boldsymbol{x}$ and $\boldsymbol{\theta}$ in one block requires careful choice of the proposal distribution to obtain a reasonably high acceptance rate. Knorr-Held and Rue (2002) illustrate the use of block-updating in Markov random field models for disease mapping applications where the relative risk is defined at the regional level. In their case, the Poisson likelihood is given by (4) with $R_i = \exp(x_i)$, where x_i is the log-risk of region i. The joint proposal of the log relative risks $\boldsymbol{x}$ conditional on a proposed value of $\boldsymbol{\theta}$ is generated by using a local quadratic approximation to the posterior. Utilising the band structure of the prior precision matrix $\boldsymbol{Q}$ they obtain computationally efficient samples from the proposal distribution. Applying the approach to a dataset on Insulin dependent Diabetes Mellitus in 366 districts of Sardinia, it is shown that the convergence and mixing of the hyper-parameters are greatly improved by blocking. However, since the spatial model is specified on the regional level, the total number of parameters to be updated are much smaller than in our application, where the Markov random field is specified on a lattice of $n = 31089$ nodes. As a consequence, the acceptance rate of a joint proposal is likely to be reduced compared to the ones reported in their study, and the computational cost of generating the sample is increased. Also, the efficiency of the approach as applied to our problem is reduced because the level of aggregation of the data typically extends the size of the neighbourhood of the GMRF model, as discussed at the end of Section 2.

As a compromise between a full block sampler and the single site Gibbs sampler, we update the hyper-parameters and a *subset* of the elements of $\boldsymbol{x}$ in one block. We split the vector $\boldsymbol{x}$ in the two sub-vectors $\boldsymbol{x}_{\mathcal{A}}$ and $\boldsymbol{x}_{-\mathcal{A}}$, representing the lattice nodes within the region of study $\mathcal{A}$ and in the boundary region respectively, and choose to block-update the hyper-parameters jointly with the subset $\boldsymbol{x}_{-\mathcal{A}}$. As we show in Section 3.3, this is equivalent to updating the hyper-parameters by sampling from the marginal posterior of $\boldsymbol{\theta}$ given $\boldsymbol{x}_{\mathcal{A}}$. Conditionally on $\boldsymbol{x}_{-\mathcal{A}}$ and $\boldsymbol{\theta}$, the elements of $\boldsymbol{x}_{\mathcal{A}}$ are updated in sub-blocks corresponding to the lattice nodes within one or more regions.

In the following subsections we describe our approach to estimation of the log-risk surface as well as the hyper-parameters, and discuss how the potential computational pitfalls pointed out at the end of Section 2 can be handled.

3.1 Step 1: The log risk surface

The sub-vector $\boldsymbol{x}_{\mathcal{A}}$ corresponding the elements of $\boldsymbol{x}$ falling within the study region for which data are available, are updated conditionally on $\boldsymbol{x}_{-\mathcal{A}}$ and the hyper-parameters $\boldsymbol{\theta}$. We update $\boldsymbol{x}_{\mathcal{A}}$ in blocks defined in terms of the regions corresponding to the level of aggregation of the data, using a Metropolis-Hastings step for each block. In this section we give an overview of the approach for a general likelihood, deferring a more detailed description of the sampling procedure for the Poisson likelihood case to Section 4. Also, we first describe the method in terms of m blocks, where each block is made up from the elements of $\boldsymbol{x}_{\mathcal{A}}$ falling within a single region. In Section 3.2 we describe the straightforward extension of the method to blocks made up of from several regions, and discuss briefly considerations to be made by choosing the size of the blocks.

The full conditional distribution for the n_i elements $x_{j_i};\ j_i = 1, \ldots, n_i$ within region $\mathcal{A}_i$ is given by

$$\pi(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}, \boldsymbol{y}) \propto \pi(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}) \pi(y_i \mid \boldsymbol{x}_{\mathcal{A}_i}, \boldsymbol{\theta}). \tag{17}$$

Here we denote by $\boldsymbol{x}_{-\mathcal{A}_i}$ all elements of the vector $\boldsymbol{x}$ except for the elements within region $\mathcal{A}_i$. The posterior distribution is in general non-standard, and we use a Metropolis-Hastings step to generate an update of $\boldsymbol{x}_{\mathcal{A}_i}$. As a proposal distribution for $\boldsymbol{x}_{\mathcal{A}_i}$ we use a quadratic approximation to (17), and we illustrate below that by re-formulating this quadratic approximation to the distribution of a conditional sampling problem, sampling can be done efficiently.

The conditional prior distribution $\pi(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta})$ of $\boldsymbol{x}_{\mathcal{A}_i}$ is Gaussian with mean $\boldsymbol{\mu}_{\mathcal{A}_i|\delta_{\mathcal{A}_i}}$ depending on the values of $\boldsymbol{x}$ in the set of nodes given by $\delta_{\mathcal{A}_i} = \cup_{j \in \mathcal{A}_i} \delta(j)$, and precision matrix $\boldsymbol{Q}_{\mathcal{A}_i}$ given by the $n_i \times n_i$ diagonal block of $\boldsymbol{Q}$ corresponding to the nodes within region $\mathcal{A}_i$. The matrix $\boldsymbol{Q}_{\mathcal{A}_i} = \boldsymbol{Q}_{\mathcal{A}_i}(\boldsymbol{\theta})$ and the vector $\boldsymbol{\mu}_{\mathcal{A}_i|\delta_{\mathcal{A}_i}} = \boldsymbol{\mu}_{\mathcal{A}_i|\delta_{\mathcal{A}_i}}(\boldsymbol{\theta})$ will both in general depend on $\boldsymbol{\theta}$, but for notational convenience we suppress explicit reference to the dependency on $\boldsymbol{\theta}$ in what follows. Thus, the log-posterior distribution corresponding to (17) becomes

$$\begin{aligned}\log(\pi(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}, \boldsymbol{y})) = \\ -\frac{1}{2}(\boldsymbol{x}_{\mathcal{A}_i} - \boldsymbol{\mu}_{\mathcal{A}_i|\delta_{\mathcal{A}_i}})^T \boldsymbol{Q}_{\mathcal{A}_i}(\boldsymbol{x}_{\mathcal{A}_i} - \boldsymbol{\mu}_{\mathcal{A}_i|\delta_{\mathcal{A}_i}}) + h_i(\boldsymbol{x}) + \text{const},\end{aligned} \tag{18}$$

where $h_i(\boldsymbol{x})$ is the log-likelihood of the observed count for region $\mathcal{A}_i$. Introducing the vector $\boldsymbol{d}_i$ given by $\boldsymbol{d}_i = \boldsymbol{Q}_{\mathcal{A}_i}\boldsymbol{\mu}_{\mathcal{A}_i|\delta_{\mathcal{A}_i}}$ and re-arranging terms, the log-posterior distribution (18) can be written in the form

$$\log(\pi(\boldsymbol{x}_{\mathcal{A}_i}|\boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}, \boldsymbol{y})) \quad = \quad -\frac{1}{2}\boldsymbol{x}_{\mathcal{A}_i}^T \boldsymbol{Q}_{\mathcal{A}_i}\boldsymbol{x}_{\mathcal{A}_i} + \boldsymbol{d}_i^T \boldsymbol{x}_{\mathcal{A}_i} + h_i(\boldsymbol{x}) + \text{const}. \tag{19}$$

A Gaussian approximation to (19) can be found by replacing the term $h_i(\boldsymbol{x})$ by a quadratic approximation

$$h_i(\boldsymbol{x}) \approx -\frac{1}{2}\boldsymbol{x}_{\mathcal{A}_i}^T \boldsymbol{B}_i \boldsymbol{x}_{\mathcal{A}_i} + \boldsymbol{b}_i^T \boldsymbol{x}_{\mathcal{A}_i}, \tag{20}$$

where $\boldsymbol{B}_i$ and $\boldsymbol{b}_i$ in general depend on the observation y_i and the parameters $\boldsymbol{\theta}$. We use a second order Taylor expansion of $h_i(\boldsymbol{x})$ to define the quadratic approximation, as will be described in Section 4 for the Poisson likelihood case. Substituting (20) for $h_i(\boldsymbol{x})$ in (19) and

collecting terms that are linear and quadratic in $\boldsymbol{x}_{\mathcal{A}_i}$, a quadratic approximation to the full conditional density (19), which we denote by $\pi_N(\boldsymbol{x}_{\mathcal{A}_i}|\boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{y})$, is

$$\begin{aligned}\log(\pi_N(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}, \boldsymbol{y})) &= -\frac{1}{2}\boldsymbol{x}_{\mathcal{A}_i}^T(\boldsymbol{Q}_{\mathcal{A}_i} + \boldsymbol{B}_i)\boldsymbol{x}_{\mathcal{A}_i} + (\boldsymbol{d}_i + \boldsymbol{b}_i)^T\boldsymbol{x}_{\mathcal{A}_i} + \text{const} \\ &= -\frac{1}{2}\boldsymbol{x}_{\mathcal{A}_i}^T(\boldsymbol{Q}_{\mathcal{A}_i} + \boldsymbol{B}_i)\boldsymbol{x}_{\mathcal{A}_i} + \boldsymbol{c}_i^T\boldsymbol{x}_{\mathcal{A}_i} + \text{const}, \end{aligned} \tag{21}$$

where we have defined $\boldsymbol{c}_i = \boldsymbol{d}_i + \boldsymbol{b}_i$. This Gaussian approximation is to be used as a proposal distribution in a Metropolis-Hastings step for updating $\boldsymbol{x}_{\mathcal{A}_i}$. However, the precision matrix $\boldsymbol{Q}_{\mathcal{A}_i} + \boldsymbol{B}_i$ of the Gaussian distribution defined by (21) is in general a full matrix, such that the computationally convenient band structure of the prior precision matrix $\boldsymbol{Q}_{\mathcal{A}_i}$ is lost. This effect of conditioning on the data was illustrated in Figure 6. If the elements of $\boldsymbol{x}$ are updated for each region $\mathcal{A}_i$ in turn, this does not necessary imply any significant loss of efficiency, since the number of lattice nodes within each region is typically relatively small and not very much larger than the number of neighbours of a lattice node. But in the general case when elements are updated in larger blocks, preserving the band structure might lead to substantial computational savings.

The general idea of our sampling approach is to re-formulate the problem of sampling directly from the Gaussian proposal distribution (21) to a conditional sampling problem for which the band structure of the precision matrix $\boldsymbol{Q}_{\mathcal{A}_i}$ is preserved. The symmetric matrix $\boldsymbol{B}_i$ can be expressed by

$$\boldsymbol{B}_i = \boldsymbol{D}_i + \boldsymbol{A}_i^T\boldsymbol{A}_i, \tag{22}$$

where $\boldsymbol{D}_i$ is a $n_i \times n_i$ diagonal matrix, possibly with zeros on the diagonal, and $\boldsymbol{A}_i$ is a $1 \times n_i$ matrix. Substituting (22) for $\boldsymbol{B}_i$ in (21) and re-arranging terms, we arrive at the expression

$$\log(\pi_N(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}, \boldsymbol{y})) = -\frac{1}{2}\boldsymbol{x}_{\mathcal{A}_i}^T(\boldsymbol{Q}_{\mathcal{A}_i} + \boldsymbol{D}_i)\boldsymbol{x}_{\mathcal{A}_i} + \boldsymbol{c}_i^T\boldsymbol{x}_{\mathcal{A}_i} - \frac{1}{2}\boldsymbol{x}_{\mathcal{A}_i}^T\boldsymbol{A}_i^T\boldsymbol{A}_i\boldsymbol{x}_{\mathcal{A}_i} + \text{const}. \tag{23}$$

As long as the matrix $\boldsymbol{Q}_{\mathcal{A}_i} + \boldsymbol{D}_i$ is positive definite, a requirement that is discussed in Section 4.2, the first two terms on the right define the log-density of a Gaussian variable for which the precision matrix $\boldsymbol{Q}_{\mathcal{A}_i} + \boldsymbol{D}_i$ has the same bandwidth as $\boldsymbol{Q}_{\mathcal{A}_i}$. The last term can be recognised as the log density, up to a constant, of a Gaussian variable with mean $\boldsymbol{A}_i\boldsymbol{x}_{\mathcal{A}_i}$ and covariance matrix $\boldsymbol{I}$, evaluated in $\boldsymbol{0}$. Sampling from (23) is shown in Section 4.2 to be equivalent to sampling from the conditional distribution

$$\pi(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{y}, \boldsymbol{z}^* = \boldsymbol{0}), \tag{24}$$

where $\boldsymbol{z}^* \mid \boldsymbol{x}_{\mathcal{A}_i} \sim N(\boldsymbol{A}_i\boldsymbol{x}_{\mathcal{A}_i}, \boldsymbol{I})$. In Section 4.2 we further illustrate that because of the band structure of $\boldsymbol{Q}_{\mathcal{A}_i} + \boldsymbol{D}_i$, sampling from (24) is computationally much more efficient than sampling directly from the Gaussian approximation as defined by (21), for which the precision matrix is in general a full matrix.

The proposed method for sampling from the full conditional distribution for $\boldsymbol{x}_{\mathcal{A}_i}$, given by $\pi(\boldsymbol{x}_{\mathcal{A}_i}|\boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}, \boldsymbol{y})$ in (19), can be summarised in the following steps.

1. Approximate the likelihood part of the full conditional distribution by a quadratic function in $\boldsymbol{x}_{\mathcal{A}_i}$, obtaining a Normal approximation to the full conditional distribution. The quadratic approximation is computed by Taylor expansion around the conditional mode.

2. Re-formulate the problem of sampling from this Normal approximation as a conditional simulation problem, where the band structure of the precision matrix is preserved.

3. Generate a sample from the Normal approximation based on the re-formulated problem. This can be done using efficient algorithms utilising the band structure of the precision matrix, described in Section 2.2.

4. Use the sample from 3. as a proposed value for $\boldsymbol{x}_{\mathcal{A}_i}$ in a Metropolis-Hastings step, compute the acceptance probability and accept or reject this value.

3.2 Updating blocks of general subsets of $\boldsymbol{x}$

In Section 3.1 we described our approach to updating the elements of $\boldsymbol{x}$ for each region separately. An equivalent approach can be taken to update larger subsets of $\boldsymbol{x}_{\mathcal{A}}$ or all elements of $\boldsymbol{x}_{\mathcal{A}}$ jointly, given the parameters $\boldsymbol{\theta}$ and the boundary elements $\boldsymbol{x}_{-\mathcal{A}}$ of the random field. The full conditional distribution of $\boldsymbol{x}_{\mathcal{S}}$ for a general subset $\mathcal{S}$ of $\mathcal{A}$ given $\boldsymbol{x}_{-\mathcal{S}}$, the data $\boldsymbol{y}$ and the hyper-parameters $\boldsymbol{\theta}$, is given by

$$\pi(\boldsymbol{x}_{\mathcal{S}} \mid \boldsymbol{x}_{-\mathcal{S}}, \boldsymbol{\theta}, \boldsymbol{y}) \propto \pi(\boldsymbol{x}_{\mathcal{S}} \mid \boldsymbol{x}_{-\mathcal{S}}, \boldsymbol{\theta}) \prod_{i:\, \mathcal{A}_i \in \mathcal{S}} \pi(y_i \mid \boldsymbol{x}_{\mathcal{S}}, \boldsymbol{x}_{-\mathcal{S}}, \boldsymbol{\theta}). \tag{25}$$

In analogy to (19), the corresponding log-density can be written as

$$\log(\pi(\boldsymbol{x}_{\mathcal{S}} | \boldsymbol{x}_{-\mathcal{S}}, \boldsymbol{y}, \boldsymbol{\theta})) \quad = \quad -\frac{1}{2}\boldsymbol{x}_{\mathcal{S}}^T \boldsymbol{Q}_{\mathcal{S}} \boldsymbol{x}_{\mathcal{S}} + \boldsymbol{d}_{\mathcal{S}}^T \boldsymbol{x}_{\mathcal{S}} + \sum_{i:\, \mathcal{A}_i \in \mathcal{S}} h_i(\boldsymbol{x}) + \text{const}, \tag{26}$$

where $h_i(\boldsymbol{x})$ is the log-likelihood of the observed count for region $\mathcal{A}_i$, and the sum is taken over the $m_{\mathcal{S}}$ regions corresponding to the subset $\mathcal{S}$. By substituting a quadratic approximation computed by a second order Taylor expansion for $\sum_i h_i(\boldsymbol{x})$ in (26) and re-arranging terms, it is shown in Appendix A.2, using a Poisson likelihood, that the corresponding Gaussian approximation to (26) can be expressed by

$$\log(\pi_N(\boldsymbol{x}_{\mathcal{S}} \mid \boldsymbol{x}_{-\mathcal{S}}, \boldsymbol{\theta}, \boldsymbol{y})) = \\ -\frac{1}{2}\boldsymbol{x}_{\mathcal{S}}^T(\boldsymbol{Q}_{\mathcal{S}} + \boldsymbol{D}_{\mathcal{S}})\boldsymbol{x}_{\mathcal{S}} + \boldsymbol{c}_{\mathcal{S}}^T \boldsymbol{x}_{\mathcal{S}} - \frac{1}{2}\boldsymbol{x}_{\mathcal{S}}^T \boldsymbol{A}_{\mathcal{S}}^T \boldsymbol{A}_{\mathcal{S}} \boldsymbol{x}_{\mathcal{S}} + \text{const}, \tag{27}$$

where $\boldsymbol{D}_{\mathcal{S}}$ is a diagonal matrix and $\boldsymbol{A}_{\mathcal{S}}$ a $m_{\mathcal{S}} \times n_{\mathcal{S}}$ matrix, with $n_{\mathcal{S}}$ equal to the number of lattice nodes within the subset $\mathcal{S}$ of regions. This Gaussian distribution is of the same form as (23) for the single region case, and the approach for generating samples from (27) to be described in Section 4.2 can be applied in the case of general subsets as well.

An extension of single region blocks to larger blocks can be generated by including the lattice nodes corresponding to the neighbours of the region, where we define two regions as neighbours if they share a common boundary, and further extensions can be made by adding the neighbours of the neighbours and so on. In Figure 7 we illustrate the size of the blocks corresponding to different choices of the number of neighbours to include in each block. We define the term 1. order neighbourhood to mean all neighbours of a region, 2. order neighbourhood to mean all neighbours as well as all neighbours of the neighbours and so on. The

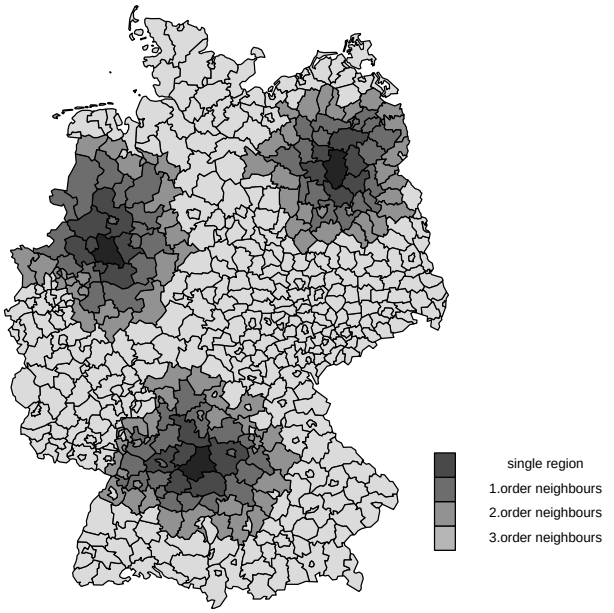


Figure 7: Different possible structures of the blocks for block-updating the log-risk surface. The 1. order neighbours are additional nodes from the 1. order neighbourhood of the single regions, 2. order neighbours are additional nodes in the 2. order neighbourhood and 3. order neighbours the nodes added to the 2. order neighbourhood from the 3. order neighbourhood.

choice of the number of regions to be updated in each sub-block is a trade-off between computational cost and the acceptance probabilities of the Metropolis-Hastings steps. Using the sampling approach outlined in Section 3.1, the problem of reduced computational efficiency due to the fact that the band structure of the precision matrix was not preserved in the Gaussian approximation to the posterior, has been handled. Therefore, the computational cost is expected to be reduced by increasing the size of the blocks and thus reduce the number of blocks. On the other hand, although increased block size might improve mixing due to larger differences between proposed and current values, increasing the number of elements of each block will reduce the quality of the Gaussian approximation, such that the acceptance rate is typically reduced. This should be kept at a reasonable level to ensure proper mixing of the MCMC algorithm.

3.3 Step 2: The hyper-parameters

The hyper-parameters $\boldsymbol{\theta}$ are updated jointly with the remaining elements $\boldsymbol{x}_{-\mathcal{A}}$ of the GMRF. As pointed out in the beginning of this section, updating all element of the GMRF in one block will most likely lead to low acceptance rates, and therefore we block the hyper-parameters with a subset of the elements of $\boldsymbol{x}$.

Using the fact that $\boldsymbol{x}_{-\mathcal{A}}$ is conditionally independent of the data $\boldsymbol{y}$ given $\boldsymbol{x}_{\mathcal{A}}$ and $\boldsymbol{\theta}$, the full

conditional distribution of $(\boldsymbol{\theta}, \boldsymbol{x}_{-\mathcal{A}})$ is

$$\pi(\boldsymbol{x}_{-\mathcal{A}}, \boldsymbol{\theta} \mid \boldsymbol{x}_{\mathcal{A}}, \boldsymbol{y}) = \pi(\boldsymbol{x}_{-\mathcal{A}}, \boldsymbol{\theta} \mid \boldsymbol{x}_{\mathcal{A}}) \propto \pi(\boldsymbol{x}_{\mathcal{A}}, \boldsymbol{x}_{-\mathcal{A}} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{28}$$

This distribution is in general a non-standard distribution, depending on the form of the hyper-prior $\pi(\boldsymbol{\theta})$, and updates are generated using a Metropolis-Hastings step. First, given the current value $\boldsymbol{\theta}$ of the hyper-parameters, a new parameter vector $\boldsymbol{\theta}'$ is sampled from a proposal distribution $q(\boldsymbol{\theta} \to \boldsymbol{\theta}')$, and then a proposed value $\boldsymbol{x}'_{-\mathcal{A}}$ is generated by sampling from the conditional distribution $\pi(\boldsymbol{x}'_{-\mathcal{A}}|\boldsymbol{\theta}', \boldsymbol{x}_{\mathcal{A}}, \boldsymbol{y}) = \pi(\boldsymbol{x}'_{-\mathcal{A}}|\boldsymbol{\theta}', \boldsymbol{x}_{\mathcal{A}})$. Since the joint distribution of $\boldsymbol{x}$ given $\boldsymbol{\theta}$ is Gaussian, it follows that the conditional distribution $\pi(\boldsymbol{x}'_{-\mathcal{A}}|\boldsymbol{\theta}', \boldsymbol{x}_{\mathcal{A}})$ is also Gaussian, with a precision matrix that has the same bandwidth as $\boldsymbol{Q}$. A proposed value can therefore be generated efficiently by sampling directly from this distribution. The proposed value is accepted or rejected according to the acceptance probability

$$\begin{aligned}
\alpha((\boldsymbol{x}_{-\mathcal{A}}, \boldsymbol{\theta}), (\boldsymbol{x}'_{-\mathcal{A}}, \boldsymbol{\theta}')) &= \min\left(1, \frac{\pi(\boldsymbol{x}'_{-\mathcal{A}}, \boldsymbol{\theta}' \mid \boldsymbol{x}_{\mathcal{A}}, \boldsymbol{y})\ \pi(\boldsymbol{x}_{-\mathcal{A}} \mid \boldsymbol{\theta}, \boldsymbol{x}_{\mathcal{A}}, \boldsymbol{y})\ q(\boldsymbol{\theta}' \to \boldsymbol{\theta})}{\pi(\boldsymbol{x}_{-\mathcal{A}}, \boldsymbol{\theta} \mid \boldsymbol{x}_{\mathcal{A}}, \boldsymbol{y})\ \pi(\boldsymbol{x}'_{-\mathcal{A}} \mid \boldsymbol{\theta}', \boldsymbol{x}_{\mathcal{A}}, \boldsymbol{y})\ q(\boldsymbol{\theta} \to \boldsymbol{\theta}')}\right) \\
&= \min\left(1, \frac{\pi(\boldsymbol{x}'_{-\mathcal{A}}, \boldsymbol{\theta}' \mid \boldsymbol{x}_{\mathcal{A}})\ \pi(\boldsymbol{x}_{-\mathcal{A}} \mid \boldsymbol{\theta}, \boldsymbol{x}_{\mathcal{A}})\ q(\boldsymbol{\theta}' \to \boldsymbol{\theta})}{\pi(\boldsymbol{x}_{-\mathcal{A}}, \boldsymbol{\theta} \mid \boldsymbol{x}_{\mathcal{A}})\ \pi(\boldsymbol{x}'_{-\mathcal{A}} \mid \boldsymbol{\theta}', \boldsymbol{x}_{\mathcal{A}})\ q(\boldsymbol{\theta} \to \boldsymbol{\theta}')}\right),
\end{aligned} \tag{29}$$

again utilising the conditional independence between $\boldsymbol{x}_{-\mathcal{A}}$ and $\boldsymbol{y}$ given $\boldsymbol{x}_{\mathcal{A}}$.

Recall the potential pitfall listed at the end of Section 2, pointing at the fact that the inclusion of the boundary region will in general be expected to slow down the convergence of the hyper-parameters. However, writing out the expression for the acceptance probability, we can show that blocking the hyper-parameters and the boundary nodes essentially eliminates this problem. There is still an effect of the boundary nodes on the subset of $\boldsymbol{x}_{\mathcal{A}}$ corresponding to the inner nodes close to the outer boundary of the study region, but this effect is supposed to be minor. Expanding the distribution $\pi(\boldsymbol{x}_{-\mathcal{A}}, \boldsymbol{\theta}|\boldsymbol{x}_{\mathcal{A}})$, (29) can be expressed by

$$\begin{aligned}
\alpha((\boldsymbol{x}_{-\mathcal{A}}, \boldsymbol{\theta}), (\boldsymbol{x}'_{-\mathcal{A}}, \boldsymbol{\theta}')) &= \min\left(1, \frac{\pi(\boldsymbol{x}'_{-\mathcal{A}}, \boldsymbol{\theta}' \mid \boldsymbol{x}_{\mathcal{A}})\ \pi(\boldsymbol{x}_{-\mathcal{A}} \mid \boldsymbol{\theta}, \boldsymbol{x}_{\mathcal{A}})\ q(\boldsymbol{\theta}' \to \boldsymbol{\theta})}{\pi(\boldsymbol{x}_{-\mathcal{A}}, \boldsymbol{\theta} \mid \boldsymbol{x}_{\mathcal{A}})\ \pi(\boldsymbol{x}'_{-\mathcal{A}} \mid \boldsymbol{\theta}', \boldsymbol{x}_{\mathcal{A}})\ q(\boldsymbol{\theta} \to \boldsymbol{\theta}')}\right) \\
&= \min\left(1, \frac{\pi(\boldsymbol{x}'_{-\mathcal{A}} \mid \boldsymbol{\theta}', \boldsymbol{x}_{\mathcal{A}})\ \pi(\boldsymbol{\theta}' \mid \boldsymbol{x}_{\mathcal{A}})\ \pi(\boldsymbol{x}_{-\mathcal{A}} \mid \boldsymbol{\theta}, \boldsymbol{x}_{\mathcal{A}})\ q(\boldsymbol{\theta}' \to \boldsymbol{\theta})}{\pi(\boldsymbol{x}_{-\mathcal{A}} \mid \boldsymbol{\theta}, \boldsymbol{x}_{\mathcal{A}})\ \pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{\mathcal{A}})\ \pi(\boldsymbol{x}'_{-\mathcal{A}} \mid \boldsymbol{\theta}', \boldsymbol{x}_{\mathcal{A}})\ q(\boldsymbol{\theta} \to \boldsymbol{\theta}')}\right) \\
&= \min\left(1, \frac{\pi(\boldsymbol{\theta}' \mid \boldsymbol{x}_{\mathcal{A}})\ q(\boldsymbol{\theta}' \to \boldsymbol{\theta})}{\pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{\mathcal{A}})\ q(\boldsymbol{\theta} \to \boldsymbol{\theta}')}\right).
\end{aligned} \tag{30}$$

Consequently, sampling the hyper-parameters jointly with the elements of $\boldsymbol{x}$ outside the region of interest is equivalent to sampling the hyper-parameters from the marginal posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{\mathcal{A}}, \boldsymbol{y}) = \pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{\mathcal{A}})$, integrated over the outer elements $\boldsymbol{x}_{-\mathcal{A}}$. In effect, using this approach the influence of the boundary nodes on the convergence of the hyper-parameters should be insignificant.

4 Efficient sampling from the full posterior of the log risk surface

In this section we describe in more detail our approach to the generation of samples from the full conditional distribution of $\boldsymbol{x}_{\mathcal{A}_i}$ given by (19), using the Poisson likelihood (15). We

describe the sampling routine in terms of blocks made up from sets of lattice nodes corresponding to single regions, but as we pointed out in Section 3.2, the sampling problem for the general case has the same structure. In Section 4.1 we compute the quadratic approximation to (19) using Taylor expansion, and in Section 4.2 we describe the method for sampling from the resulting Gaussian approximation.

4.1 A Taylor expansion based Gaussian approximation to the posterior

Here, we establish the quadratic approximation to the full conditional distribution of $\boldsymbol{x}_{\mathcal{A}_i}$ analytically by computing the second order Taylor expansion of the log-likelihood part $h_i(\boldsymbol{x})$ of (19). For the Poisson likelihood (15), $h_i(\boldsymbol{x})$ becomes

$$h_i(\boldsymbol{x}) = y_i \log(E_i' \sum_{j \in \mathcal{A}_i} \exp x_j) - E_i' \sum_{j \in \mathcal{A}_i} \exp x_j. \tag{31}$$

The expansion is computed around a point $\boldsymbol{x}^0_{\mathcal{A}_i}$ taken to be the mode of the full conditional distribution (19), found numerically given the current value of $\boldsymbol{\theta}$. Expressing $h_i(\boldsymbol{x})$ in terms of the gradient (first order derivative) $\boldsymbol{g}_i(\boldsymbol{x}_{\mathcal{A}_i})$ and the Hessian (second order derivative) $\boldsymbol{G}_i(\boldsymbol{x}_{\mathcal{A}_i})$ of the Taylor expansion, we get

$$\begin{aligned} h_i(\boldsymbol{x}) &\approx g(\boldsymbol{x}^0_{\mathcal{A}_i}) + \boldsymbol{g}_i^T(\boldsymbol{x}^0_{\mathcal{A}_i})(\boldsymbol{x}_{\mathcal{A}_i} - \boldsymbol{x}^0_{\mathcal{A}_i}) - \frac{1}{2}(\boldsymbol{x}_{\mathcal{A}_i} - \boldsymbol{x}^0_{\mathcal{A}_i})^T(-\boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i}))(\boldsymbol{x}_{\mathcal{A}_i} - \boldsymbol{x}^0_{\mathcal{A}_i}) \quad (32) \\ &= -\frac{1}{2}\boldsymbol{x}^T_{\mathcal{A}_i}(-\boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i}))\boldsymbol{x}_{\mathcal{A}_i} + (\boldsymbol{g}_i^T(\boldsymbol{x}^0_{\mathcal{A}_i}) - (\boldsymbol{x}^0_{\mathcal{A}_i})^T\boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i}))\boldsymbol{x}_{\mathcal{A}_i}, \quad (33) \end{aligned}$$

discarding terms not depending on $\boldsymbol{x}_{\mathcal{A}_i}$. In terms of $\boldsymbol{g}_i(\boldsymbol{x}^0_{\mathcal{A}_i})$ and $\boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i})$, the matrix $\boldsymbol{B}_i$ and the vector $\boldsymbol{b}_i$ in the quadratic approximation $h_i(\boldsymbol{x}) \approx -\frac{1}{2}\boldsymbol{x}^T_{\mathcal{A}_i}\boldsymbol{B}_i\boldsymbol{x}_{\mathcal{A}_i} + \boldsymbol{b}_i^T\boldsymbol{x}_{\mathcal{A}_i}$ are given by $\boldsymbol{B}_i = -\boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i})$ and $\boldsymbol{b}_i = \boldsymbol{g}_i(\boldsymbol{x}^0_{\mathcal{A}_i}) - \boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i})\boldsymbol{x}^0_{\mathcal{A}_i}$, such that the Gaussian approximation (21) is

$$\log(\pi_N(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}, \boldsymbol{y})) = -\frac{1}{2}\boldsymbol{x}^T_{\mathcal{A}_i}(\boldsymbol{Q}_{\mathcal{A}_i} + (-\boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i}))\boldsymbol{x}_{\mathcal{A}_i} + \boldsymbol{c}_i^T\boldsymbol{x}_{\mathcal{A}_i} + \text{const}, \tag{34}$$

with $\boldsymbol{c}_i = \boldsymbol{d}_i + \boldsymbol{g}_i(\boldsymbol{x}^0_{\mathcal{A}_i}) - \boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i})\boldsymbol{x_0}$. Thus, the full conditional distribution for $\boldsymbol{x}_{\mathcal{A}_i}$ is approximated by a Gaussian distribution with mean $\widetilde{\boldsymbol{\mu}}_i$ and precision matrix $\widetilde{\boldsymbol{Q}}_i$, given by

$$\begin{aligned} \widetilde{\boldsymbol{Q}}_i &= \boldsymbol{Q}_{\mathcal{A}_i} - \boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i}) \quad (35) \\ \widetilde{\boldsymbol{\mu}}_i &= \widetilde{\boldsymbol{Q}}_i^{-1}(\boldsymbol{d}_i + \boldsymbol{g}_i(\boldsymbol{x}^0_{\mathcal{A}_i}) - \boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i})\boldsymbol{x}^0_{\mathcal{A}_i}). \quad (36) \end{aligned}$$

As shown in Appendix A.1, the gradient $\boldsymbol{g}_i(\boldsymbol{x})$ and the Hessian $\boldsymbol{G}_i(\boldsymbol{x})$, evaluated in the mode $\boldsymbol{x}^0_{\mathcal{A}_i}$, are given by

$$\begin{aligned} \boldsymbol{g}_i(\boldsymbol{x}^0_{\mathcal{A}_i}) &= (\frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{A}_i})} - E_i')\ \boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{A}_i}) \quad (37) \\ \boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i}) &= (\frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{A}_i})} - E_i')\ \text{diag}(\boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{A}_i})) - \frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{A}_i})^2}\boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{A}_i})\boldsymbol{a}_i^T(\boldsymbol{x}^0_{\mathcal{A}_i}), \quad (38) \end{aligned}$$

where we have used the definitions

$$\boldsymbol{a}_i(\boldsymbol{x}_{\mathcal{A}_i}) = (\exp(x_j))_{j\in\mathcal{A}_i}^T, \tag{39}$$

$$S_i(\boldsymbol{x}_{\mathcal{A}_i}) = \sum_{j\in\mathcal{A}_i}\exp(x_j) = \sum_{k=1}^{n_i} a_{ik}. \tag{40}$$

The mean $S_i(\boldsymbol{x}_{\mathcal{A}_i})/n_i = \frac{1}{n_i}\sum_{j\in\mathcal{A}_i}\exp(x_j)$ is equal to the relative risk in region $\mathcal{A}_i$ conditionally on $\boldsymbol{x}$.

Observe that $\boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i})$ is of the form

$$\boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i}) = -(\boldsymbol{D}_i + \boldsymbol{H}_i) \tag{41}$$

where $\boldsymbol{D}_i$ is a diagonal matrix and $\boldsymbol{H}_i$ a rank one matrix defined by

$$\boldsymbol{D}_i = -(\frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{A}_i})} - E_i')\ \mathrm{diag}(\boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{A}_i})) \tag{42}$$

$$\boldsymbol{H}_i = \frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{A}_i})^2}\boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{A}_i})\boldsymbol{a}_i^T(\boldsymbol{x}^0_{\mathcal{A}_i}) = \boldsymbol{A}_i^T\boldsymbol{A}_i. \tag{43}$$

Here, we have introduced the $1 \times n_i$ matrix $\boldsymbol{A}_i$ given by

$$\boldsymbol{A}_i = \frac{\sqrt{y_i}}{S_i(\boldsymbol{x}^0_{\mathcal{A}_i})}\boldsymbol{a}_i^T(\boldsymbol{x}^0_{\mathcal{A}_i}). \tag{44}$$

Both $\boldsymbol{D}_i$ and $\boldsymbol{H}_i$ depend on the observed count y_i and the point $\boldsymbol{x}^0_{\mathcal{A}_i}$. Substituting the sum $-(\boldsymbol{D_i}+\boldsymbol{H}_i)$ for $\boldsymbol{G}(\boldsymbol{x}^0_{\mathcal{A}_i})$ in (34) using the expression in (43) for $\boldsymbol{H}_i$ and re-arranging terms, we arrive at the expression (23) for the proposal distribution. As we will describe in Section 4.2, this re-formulation can be utilised to reduce the computational cost of sampling from (34).

4.2 Sampling algorithm

In Section 4.1 we re-formulated the problem of sampling from the Normal approximation (34) to the full conditional distribution (19) in terms of the general problem of sampling from a distribution on the form

$$\log(\pi(\boldsymbol{x})) = -\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{Q}+\boldsymbol{D})\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} + \text{const}, \tag{45}$$

where $\boldsymbol{Q}+\boldsymbol{D}$ is a band matrix. Here, we have suppressed the subscripts $\mathcal{A}_i$ and i, the dependency of $\boldsymbol{x}^0_{\mathcal{A}_i}$ and the conditioning on $\boldsymbol{x}_{\mathcal{A}_i}$, $\boldsymbol{y}$ and $\boldsymbol{\theta}$ for notational convenience.

The first two terms of (45) is the log-density function, up to a constant, of a Gaussian vector variable $\boldsymbol{x}^*$ with mean an precision matrix given by

$$\boldsymbol{\mu}^* = (\boldsymbol{Q}^*)^{-1}\boldsymbol{c}, \tag{46}$$

$$\boldsymbol{Q}^* = \boldsymbol{Q}+\boldsymbol{D}. \tag{47}$$

Rewriting the last term of (45) as

$$-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = -\frac{1}{2}(\boldsymbol{0}-\boldsymbol{A}\boldsymbol{x})^T\boldsymbol{I}(\boldsymbol{0}-\boldsymbol{A}\boldsymbol{x}), \tag{48}$$

we observe that this term, up to a constant, is equivalent to the multivariate Normal log-density of another vector variable $\boldsymbol{z}^*$ with mean $\boldsymbol{A}\boldsymbol{x}$ and covariance matrix $\boldsymbol{I}$, that is evaluated in the value $\boldsymbol{z} = \boldsymbol{0}$. Thus, we have introduce two new variables $\boldsymbol{x}^*$ and $\boldsymbol{z}^*$, with distributions given by

$$\begin{aligned} \boldsymbol{x}^* &\sim N(\boldsymbol{\mu}^*, (\boldsymbol{Q}^*)^{-1}) && (49)\\ \boldsymbol{z}^*|\boldsymbol{x}^* = \boldsymbol{x} &\sim N(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{I}), && (50) \end{aligned}$$

where $\boldsymbol{\mu}^*$ and $\boldsymbol{Q}^*$ are defined by (46) and (47). In terms of the variables $\boldsymbol{x}^*$ and $\boldsymbol{z}^*$, the distribution (45) is equivalent to the conditional distribution of $\boldsymbol{x}^*$ given $\boldsymbol{z}^* = \boldsymbol{0}$, evaluated in $\boldsymbol{x}^* = \boldsymbol{x}$. We denote this distribution by $\pi_{\boldsymbol{x}^*|\boldsymbol{z}^*}(\boldsymbol{x}^*|\boldsymbol{z}^* = \boldsymbol{0})$, and it can be expressed by

$$\pi_{\boldsymbol{x}^*|\boldsymbol{z}^*}(\boldsymbol{x}|\boldsymbol{z}^* = \boldsymbol{0}) \propto \pi_{\boldsymbol{x}^*}(\boldsymbol{x})\ \pi_{\boldsymbol{z}^*|\boldsymbol{x}^*}(\boldsymbol{0}|\boldsymbol{x}^* = \boldsymbol{x}), \tag{51}$$

using a compact notation. Consequently, sampling from (45), and thus the Gaussian proposal distribution (21), is equivalent to sampling from the conditional distribution given by (51). Since

$$\boldsymbol{z}^* = \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{\epsilon}^*;\ \ \boldsymbol{\epsilon}^* \sim N(\boldsymbol{0}, \boldsymbol{I}), \tag{52}$$

conditioning on $\boldsymbol{z}^* = \boldsymbol{0}$ is equivalent to conditioning on $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{\epsilon}^*$. In our application, we have that

$$\boldsymbol{A}\boldsymbol{x} = (S_i(\boldsymbol{x}^0_{\mathcal{A}_i})^2/y_i) \sum_{j\in\mathcal{A}_i} \exp(x_{0,j})\ x_j \tag{53}$$

is a weighted sum of the lattice specific log relative risks x_j within each region. Conditioning on $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{\epsilon}^*$ can be interpreted as generating samples for which $\mathrm{E}(\boldsymbol{A}\boldsymbol{x}^*) = \boldsymbol{0}$, and where the elements of $\boldsymbol{A}\boldsymbol{x}^*$ should be independent Gaussian variables with common variance 1.

To generate a sample $\boldsymbol{x}_c$ from the conditional distribution $\pi(\boldsymbol{x}^*|\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{\epsilon}^*)$, we use the approach given by equation (12) in Section 2.2. We first generate an *unconditional* sample $\boldsymbol{x}_u$ from $\pi(\boldsymbol{x}^*)$ and an $\epsilon^* \sim N(\boldsymbol{0}, \boldsymbol{I})$, and then compute $\boldsymbol{x}_c$ by adjusting for the constraint $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{\epsilon}^*$ using the expression

$$\boldsymbol{x}_c = \boldsymbol{x}_u - (\boldsymbol{Q}^*)^{-1}\boldsymbol{A}^T(\boldsymbol{A}(\boldsymbol{Q}^*)^{-1}\boldsymbol{A}^T + \boldsymbol{I})^{-1}(\boldsymbol{A}\boldsymbol{x}_u - \boldsymbol{\epsilon}^*) \tag{54}$$

The precision matrix $\boldsymbol{Q}^*$ has the same bandwidth as the prior precision matrix $\boldsymbol{Q}$ of $\boldsymbol{x}$. Utilising the band structure of $\boldsymbol{Q}^*$, samples from (51) can be generated efficiently using the methods described and implemented in Rue and Follestad (2002).

To compute (54) we need to evaluate the matrix expression $(\boldsymbol{A}(\boldsymbol{Q}^*)^{-1}\boldsymbol{A}^T + \boldsymbol{I})^{-1}$. The matrix $\boldsymbol{A}(\boldsymbol{Q}^*)^{-1}\boldsymbol{A}^T + \boldsymbol{I}$ is in general a full matrix, but it is of dimension m_c by m_c, where m_c is the number of rows in $\boldsymbol{A}$. So far we have considered the sampling problem updating one region at a time, for which $m_c = 1$. But even for generalisations to larger subsets of regions, we usually have that $m_c \ll n$.

As pointed out in Section 3, the matrix $\boldsymbol{Q} + \boldsymbol{D}$ should be positive definite for the Gaussian distribution defined by (46) and (47) to be proper. The matrix $\boldsymbol{Q}^*$ is positive definite iff $\boldsymbol{x}^T\boldsymbol{Q}^*\boldsymbol{x} > 0;\ \forall \boldsymbol{x} > 0$. Substituting $\boldsymbol{Q} + \boldsymbol{D}$ for $\boldsymbol{Q}^*$ we get

$$\begin{aligned} \boldsymbol{x}^T\boldsymbol{Q}^*\boldsymbol{x} &= \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{D}\boldsymbol{x} \\ &= \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} - (\frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{A}_i})} - E_i')\boldsymbol{x}^T \mathrm{diag}(\boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{A}_i}))\boldsymbol{x}. \end{aligned} \tag{55}$$

Since all elements of $\boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{A}_i})$ are strictly positive, the sign of the last term on the right is determined by the sign of the factor

$$\frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{A}_i})} - E_i' = (y_i - S_i(\boldsymbol{x}^0_{\mathcal{A}_i})E_i')/S_i(\boldsymbol{x}^0_{\mathcal{A}_i}) = (y_i - \mathrm{E}(y_i \mid \boldsymbol{x}^0_{\mathcal{A}_i}))/S_i(\boldsymbol{x}^0_{\mathcal{A}_i}). \tag{56}$$

This term can in general be of either sign, such that the precision matrix $\boldsymbol{Q}^*$ is not guaranteed to satisfy the positive definiteness requirement. In the case where the prior variance $1/\tau$ is large, such that $\boldsymbol{x}^T\boldsymbol{Q}^*\boldsymbol{x}$ is small, the second term of (55) will dominate, and the chance is higher that the positive definiteness requirement is not met. However, this problem can in general be dealt with by a slight modification of our sampling algorithm, replacing the diagonal elements $d_{j,j}$ of the matrix $\boldsymbol{D}$ by $\max(d_{j,j}, 0.0)$. The corresponding change in proposal distribution is corrected for by the acceptance probability of the MCMC algorithm.

To summarise the sampling approach, a sample from the conditional distribution (45) can be obtained by the following steps:

1. Sample a value $\boldsymbol{x}_u$ from the unconditional distribution (49).
2. Sample an $\boldsymbol{\epsilon}^* \sim N(\boldsymbol{0}, \boldsymbol{I})$.
3. Compute $\boldsymbol{x}_c$ using (54). Then $\boldsymbol{x}_c$ will be a sample from the posterior distribution given by (51), and consequently from (45).

4.3 Some computational details

The sampling algorithm is implemented in `C`, and is based on the routines for fast and exact simulation of Gaussian Markov random fields implemented in the library `GMRFLib` (Rue and Follestad, 2002). The library provides general algorithm for generating samples from a GMRF, including conditional samples for hard and soft linear constraints, and the algorithms are based on the Cholesky factorisation (9) of the precision matrix $\boldsymbol{Q}$ (Rue, 2001).

When applying the `C`-routines to our problem, we have utilised the structure of the specific problem to further reduce the computational cost. Using the fact that the range of the correlation function is given for a set of discrete values, the normalising constant of the posterior distribution of $\boldsymbol{x}$ can be computed once at the beginning of a run, and stored for later use. Also, the sub-graphs representing the subset of nodes in each block in the block-updates of the log-risk surface $\boldsymbol{x}$, as well as the sub-graph representing the boundary region nodes, are computed only once. A timing of the computer program, running the program for 1000 iterations, reveals that 32% of the time is spent evaluating the elements of the precision matrix

$\boldsymbol{Q}$, and 12% in evaluating the log-likelihood. Further, about 21% of the CPU time is spent on setting up the algorithm, including specifying the neighbourhood structure and computing sub-graphs. Consequently, more CPU time is spent on setting up the problem than actually performing the computations generating the samples.

5 Simulation study

In this section we illustrate the performance of the method by applying the sampling algorithm to two simulated data sets, both generated using the study region of the real data. A plot of the study region made up from the 544 districts of Germany, overlaid by a lattice consisting of $n = 31089$ nodes including boundary nodes, were shown in Figure 5.

The simulated data as well as the real data set are standardised such that the overall risk for the region of study is 1. Therefore, no intercept term is included in the model for the log-risk surface, and the prior mean $\boldsymbol{\mu}$ of $\boldsymbol{x}$ is taken to be $\mathbf{0}$. The remaining hyper-parameters of the GMRF prior are the precision τ and the parameters specifying the spatial structure of the random field. We model the spatial dependency using an isotropic one-parameter exponential correlation function given by

$$\rho(h; r) = \exp(-3h/r). \tag{57}$$

Here h is the distance between two nodes of the lattice, and r is the distance for which the correlation is reduced to 0.05. In the simulated data sets, the range parameter of the exponential correlation function (57) is set equal to $r = 40$ measured in lattice coordinates for both data sets. The precisions are taken to be different, and the chosen values are $\tau = 24$ and $\tau = 8$, corresponding to standard deviations of 0.20 and 0.35 respectively. For each data set, we first generate a realisation of the log-risk surface $\boldsymbol{x}$ from the GMRF prior (7), and conditionally on $\boldsymbol{x}$ a set of regional count data is sampled from the Poisson distribution given by the likelihood (15). We define the expected number of cases E_i to be the ones given in the data set used in Section 6, ranging from 3.0 to 393.1 and with a median of 19. A summary of the two simulated data sets used in the study is given in Table 2, and the realisations of $(\exp(x_j))_{j=1,\ldots,n}$ and the corresponding regional relative risks, given by the mean $(\sum_{j \in \mathcal{A}_i} \exp(x_j))/n_i$ over the n_i lattice nodes within region i, are shown in Figure 8.

The prior distribution for the precision τ and the range parameter r are assumed to be independent. To reduce the computational burden, we use a discrete prior distribution for the range parameter r, such that the determinant of the precision matrix, needed for the evaluation of normalising constants, can be computed once at the beginning of the sampling procedure. The discretisation is done in $n_r = 2001$ steps $r_k;\ k = 1, \ldots, n_r$, where the range at step k is equal to $r_k = (k-1)0.05$ measured in lattice coordinates. The discrete prior distribution is defined on the indexes k, such that

$$\pi(k) \quad \propto \quad \frac{1}{k};\ k = 1, \ldots, n_r. \tag{58}$$

The precision τ, which is constrained to be positive, is assigned a Gamma prior, $\tau \sim \text{Gamma}(\alpha_\tau, \beta_\tau)$. Based on the recommendations in Kelsall and Wakefield (1999) and the discussion of prior

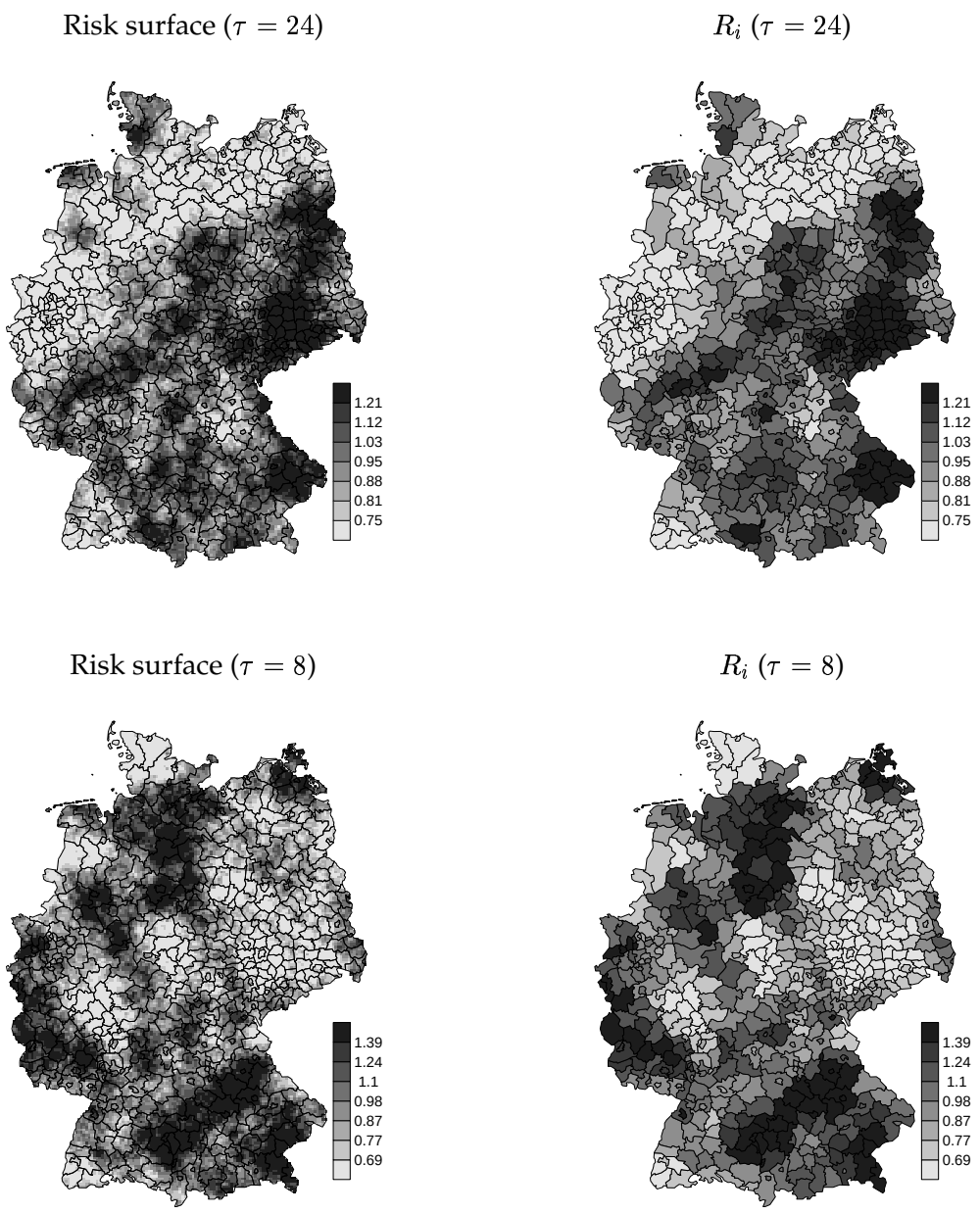


Figure 8: The true risk surfaces and corresponding regional level relative risks for the two simulated data sets.

			Aggregated counts (y_i)				
Data set	τ	r	Min.	2.5% quantile	Median	97.5% quantile	Max.
I	24	40	0	3	19	87	403
II	8	40	1	2	18	111	461
			Relative risk (R_i)				
Data set	τ	r	Min.	2.5% quantile	Median	97.5% quantile	Max.
I	24	40	0.50	0.61	0.98	1.43	1.85
II	8	40	0.45	0.49	095	1.71	2.60

Table 2: The aggregated counts and the true relative risks of the simulated data sets. The quantiles are given as the empirical quantiles of the simulated values.

sensitivity in Pascutto, Wakefield, Best, Richardson, Bernardinelli, Staines and Elliott (2000), we choose the parameters of the Gamma$(\alpha_\tau, \beta_\tau)$-prior for τ such that more weight is given to small variances than the Gamma(ϵ, ϵ)-prior for small ϵ frequently used in this type of applications. Specifically, we choose $\alpha_\tau = 0.2$ and $\beta_\tau = 0.0002$. For the range parameter r we use the discrete prior (58) on the range indexes $k = 1, 2, \ldots, 2001$ corresponding to the values $0.0, 0.05, 0.1, 0.15, \ldots, 100.0$ of the range, as measured in lattice coordinates.

The elements of $\boldsymbol{x}_{\mathcal{A}}$, representing the log relative risk surface within the 544 regions, are updated using the block-sampling approach described in Section 3.2. As pointed out in that section, the optimal choice of block-size can be considered to be a trade-off between computational cost and the acceptance probabilities of the Metropolis-Hastings steps. To study the effect of changing the block-size on the acceptance probabilities, we ran 11000 iterations of the sampler on data set I of Table 2 for four different choices of blocks, keeping the hyperparameters fixed at their true values. The blocks are made up from single regions, 1. order neighbourhoods, 2. order neighbourhoods and 3. order neighbourhoods respectively, using the neighbourhood definitions given in Section 3.2 and Figure 7. The four different block sizes are also illustrated in Figure 9. In our sampling algorithm, the blocks are slightly modified such that the different blocks of one run of the sampler are disjoint, and such that regions with only one neighbour are added to one of the adjacent blocks. This last modification applies to city regions, like two regions within the 2. order neighbourhood block of Figure 9, as well as some of the regions at the boundary. To avoid boundary effects between blocks, we generate the blocks randomly, updating the partition into blocks at every 10th step of the sampler.

The resulting acceptance rates for the four different choices of the block structure are displayed in Figure 10. We observe that the acceptance rates for the single region blocks are very large, with a median acceptance probability of 0.95, indicating that the Gaussian approximation is a good approximation to the posterior distribution (19). For the blocks based on 1., 2. and 3. order neighbourhoods, the median acceptance probabilities are gradually decreased, taking the values 0.66, 0.35 and 0.16 respectively. The acceptance probabilities seem to be independent of the size of the regions, represented by the number of lattice nodes within the region, but they increase as the mean of the regional level risk approaches 1.0. This result is as expected, since the Gaussian approximation (27) to the posterior distribution (25) of $\boldsymbol{x}_{\mathcal{S}}$ is

expected to be better when the values of $\boldsymbol{x}_S$ are small, and thus the corresponding regional level relative risk close to 1.0. Based on these results, we choose to use blocks made up from a region and its 1. order neighbourhood.

The convergence of the MCMC algorithm is assessed by visual inspection of trace plots. The total number of parameters of the risk surface is too large for an inspection of all trace plots to be feasible. So in addition to the hyper-parameters τ and r, we study trace plots of the relative risk R_i of a selected number of regions, and for a subset of the corresponding elements of the log-risk vector $\boldsymbol{x}$. The values of the regional relative risks at iteration k, denoted $R_i^{(k)};\ i = 1, \ldots, m$, are generated from the current values of $\boldsymbol{x}$ by

$$R_i^{(k)} = \frac{1}{n_i} \sum_{j \in \mathcal{A}_i} \exp(x_j^{(k)}), \tag{59}$$

where $x_j^{(k)}$ is the k'th update of x_j. To get an impression of the behaviour of the algorithm for the remaining regions, we compute the mean acceptance probability of the Metropolis-Hastings steps for all regions.

In Figure 11, we show a subset of trace plots for data set I, after running the MCMC algorithm for 101000 iterations. The convergence is fast and the algorithm mixes well for the majority of the relative risk estimates, but the trace plots for region 16 indicate that although the convergence seems to be fast, the mixing is relatively poor for this region. The mean acceptance probability for the corresponding elements of $\boldsymbol{x}$ is 1.9%. The mean acceptance probabilities for all regions are plotted in the top panels of Figure 12, and the acceptance rate for region 16 is seen to be the lowest among the 544 regions, for which the second smallest value is 8.9%. Region 16 corresponds to a region with a small true ($R_{16} = 0.61$) as well as estimated ($\widehat{R}_{16} = 0.65$) relative risk and a large aggregated count ($y_{16} = 202$). From Table 3 we observe that the true risk is relatively similar for region 16 and its neighbours, but the expected and observed aggregated counts are an order of magnitude larger. From Table 2 it is clear that the observed count of region 16 is in the tail of the empirical distribution of the observed counts, and this might explain why the Gaussian approximation is relatively poor for the elements of the log-risk within this region. (Region 16 includes Hamburg, and since we use the expected counts of the German oral cavity cancer data to generate our simulated data set, this explains the high count of this region).

Region no. (i)	y_i	E_i	R_i	SMR
16	**202**	**314.9**	**0.61**	**0.64**
6	17	30.7	0.67	0.55
9	45	50.9	0.62	0.88
13	30	39.9	0.72	0.75
15	22	38.5	0.70	0.57
38	18	38.1	0.63	0.47
44	28	30.6	0.73	0.91

Table 3: The expected (E_i) and observed (y_i) aggregated counts, true relative risks (R_i) and SMR for region 16 and its neighbours for data set I.

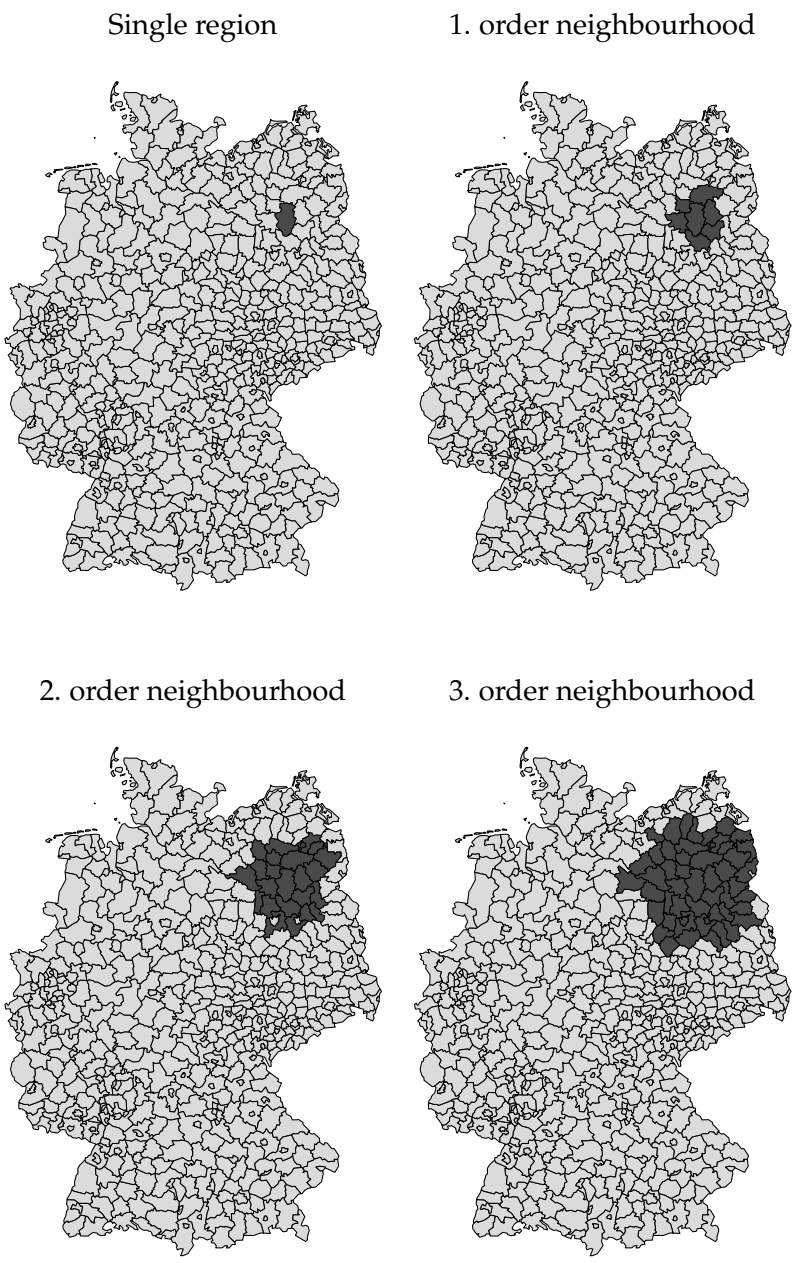


Figure 9: An illustration of the blocks used in the block-sampling of the log-risk surface. The risk surface for the lattice nodes within the dark shaded regions are updated conditionally on the nodes of all remaining regions as well as the boundary nodes.

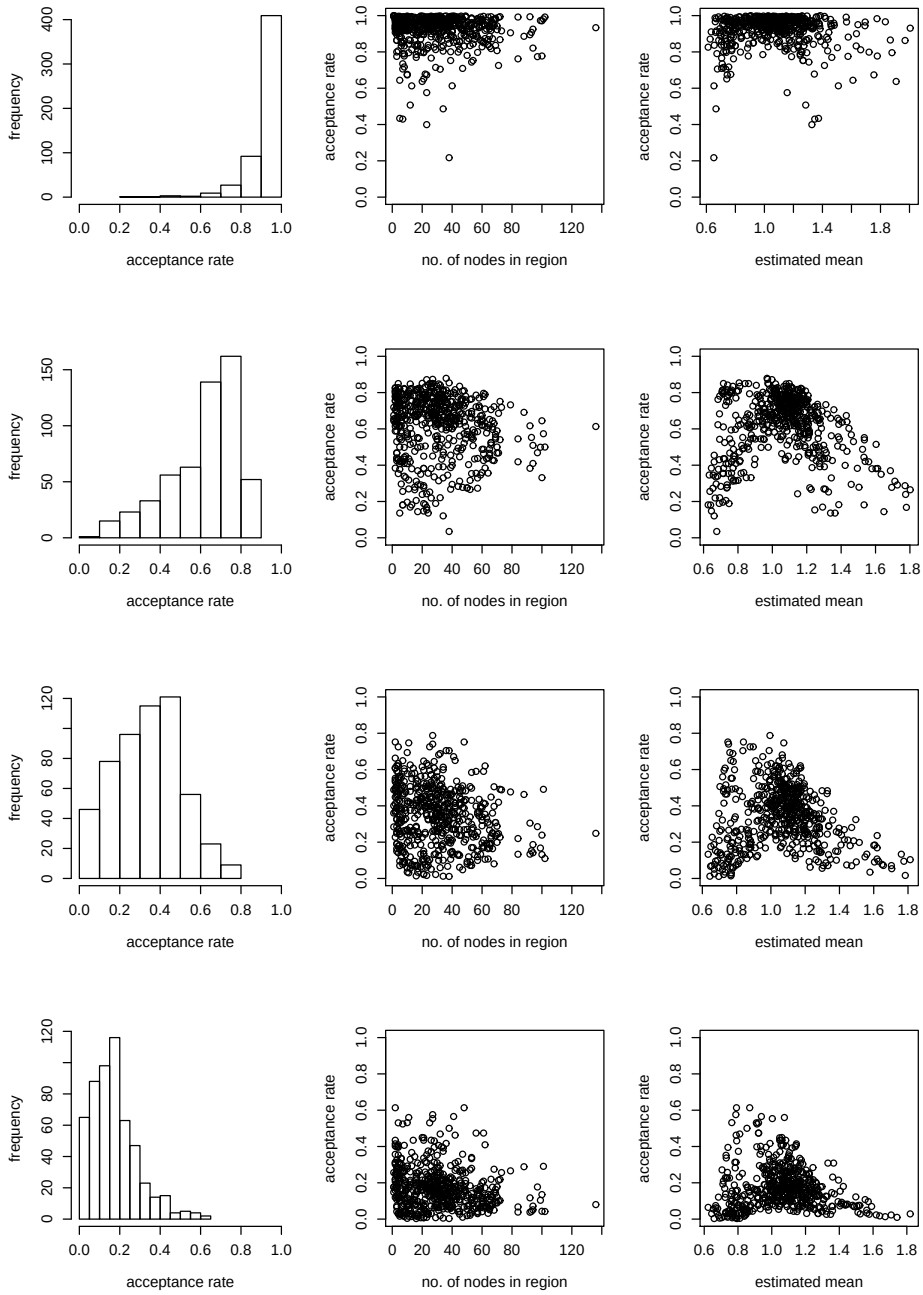


Figure 10: Histograms of the mean acceptance probabilities of the log-risk of the 544 regions using the block-MCMC algorithm (left), the mean acceptance probabilities plotted against the number of nodes within the region (middle), and the same values plotted against the estimated log-risk (left). The plots are given for blocks of single regions and for 1. order, 2. order and 3. order neighbourhoods (from the top and downward), and are based on results from using data set I.

The mixing for τ and r is poorer, but the trace plot of τ indicates that the algorithm has converged for this parameter. For the range parameter r, the mixing is not uniform over the range of possible values. The poor mixing for some neighbouring values for r is due to larger differences between the corresponding neighbouring prior models than the typical differences between neighbouring models over the range of values of r. This is a result of the procedure used for fitting GMRFs to GRFs (Rue and Tjelmeland, 2002). The additional constraint that the coefficients of the precision matrix of the GMRF, computed for each value of the range, should also be near continuous with respect to the range, is not accounted for in the fitting procedure. The effect could be reduced by increasing the resolution for the range r in (58), but probably we need to add explicit smoothing constraints of the parameters with respect to range in the fitting procedure.

To illustrate the effect of slow convergence and mixing of the hyper-parameters on the estimates of the log-risk surface, we have plotted the updated values of some elements of $\boldsymbol{x}$ against corresponding values for τ and r. From the resulting scatter plots shown in the top two rows of Figure 13, we observe that the posterior variance decreases for increasing values of the precision τ, but the posterior means of the elements of $\boldsymbol{x}$ appear to be stable despite the poor mixing of the individual parameters τ and r. Therefore, we proceed by presenting results for the relative risk surface based on estimated posterior means, but the poor mixing of the hyper-parameters should be kept in mind.

We discard the first 1000 iterations and use the remaining 100000 iterations to compute estimated posterior mean values of the relative risk surface and the relative risks within each region. The results for data set I are reported in Figure 14. From the top and middle left panels we observe that the simulation algorithm reproduces the structure of the simulated risk surface well. The corresponding true and estimated values of the regional level risks are plotted in the top and middle right panels. Comparing the estimated values to the standardised mortality ratio (SMR) added in the bottom right panel, the algorithm is seen to smooth the disease map based on the SMR toward the true risk surface. In the bottom left panel we have plotted the estimated probability that the risk R_i exceeds 1.

A selection of trace plots and the estimated posterior mean values for data set II are given in Figures 15 and 16. The general pattern is similar to the results from data set I. The mixing is relatively good and the convergence is fast for the log-risk surface for most regions, but for the hyper-parameters, convergence is not achieved after the 101000 iterations. However, there are more regions for which the mixing for the corresponding elements of $\boldsymbol{x}$ is relatively poor. In Figures 15, we have included trace plots for the regional level relative risk and two corresponding elements of $\boldsymbol{x}$ for a region for which the acceptance rate is extremely low (0.09%). This region is the same as the one with lowest acceptance rate for data set I. In the sampling algorithm we have used the same block-size as for data set I, producing the mean acceptance probabilities illustrated in the middle panels of Figure 12. We observe that the acceptance rates are in general lower than for data set I, and the lowest for the regions with the most extreme values of the risk. This overall decrease in acceptance probabilities corresponds to the fact that the Gaussian approximation is a better fit to the posterior distribution for the data set with the smaller variance. To increase the average acceptance probability, the block-sizes should be reduced to include single regions only for this simulated data set, and in the further discussion of the results, the low acceptance rates for some of the regions should be kept in mind.

As for data set I the posterior mean level of the elements of $\boldsymbol{x}$ seems to be stable despite the convergence problems apparent for the hyper-parameters, as illustrated by the scatter plots in the bottom two rows of Figure 13. For region number 16, there is some indications of negative association between τ and the estimated level of the log-risk, but as we pointed out above, the acceptance rate is very low, and we know from Figure 15 that the mixing is poor for the elements of $\boldsymbol{x}$ within this region. Therefore, the results presented below, based on estimated posterior means, are hoped to be representative despite the poor mixing of the hyper-parameters.

Comparing the differences between the SMR and the estimated risk surface for the two data sets, we observe from Figures 14 and 16 that degree of smoothing is less pronounced for the case $\tau = 8$ than for the data set with $\tau = 24$. Thus, increasing the prior variance of the underlying risk surface seems to reduce the degree of smoothing. This effect is also illustrated in Figure 17, where we have plotted the differences between the estimated and true relative risks together with corresponding differences between the estimated risks and the observed SMR. Some of the larger differences for data set II correspond to regions for which the acceptance probabilities are small and the mixing relatively poor, but comparing similar differences discarding risk estimates for which the acceptance probabilities are all larger than 20%, the tendency is similar.

We end this section by an illustration of how we can assess the validity of the approximation

$$\frac{1}{n_i}\sum_{j\in\mathcal{A}_i}\exp(x_j) \approx \exp(\frac{1}{n_i}\sum_{j\in\mathcal{A}_i}x_j), \tag{60}$$

which is an analogue to the approximation $\log(R_i) = \int_{\mathcal{A}_i}\log R(s)f_i(s)ds$, underlying the geostatistical approach of Kelsall and Wakefield (2002). Let

$$\tilde{R}_i \quad = \quad \exp(\frac{1}{n_i}\sum_{j\in\mathcal{A}_i}x_j), \tag{61}$$

be the approximation to the relative risk $R_i = \frac{1}{n_i}\sum_{j\in\mathcal{A}_i}\exp(x_j)$ of region $\mathcal{A}_i$, and let further $\widehat{\widetilde{R}}_i = \overline{\widetilde{R}_i^{(\cdot)}}$ and $\widehat{R}_i = \overline{R_i^{(\cdot)}}$ be the corresponding posterior mean estimates based on the updates $\widetilde{R}_i^{(k)}$ and $R_i^{(k)}; k = 1,\ldots,100000$ from the Metropolis-Hastings algorithm. By Jensens inequality

$$\exp(\frac{1}{n_i}\sum_{j\in\mathcal{A}_i}x_j) \leq \frac{1}{n_i}\sum_{j\in\mathcal{A}_i}\exp(x_j), \tag{62}$$

such that the estimated posterior means of $\widetilde{R}_i$ should be smaller than or equal to the corresponding values for R_i. In Figure 18 we have plotted $\widehat{\widetilde{R}}_i$ against $\widehat{R}_i$ and $\widehat{\widetilde{R}}_i/\widehat{R}_i$ as a function of the number of lattice nodes within each region, for data sets I and II as well as for the oral cavity cancer data analysed in the next section. We observe that (61) is a good approximation to R_i. As expected, the approximation is better the smaller the number of lattice nodes within the region, and in accordance with Jensens inequality, $\widehat{\widetilde{R}}_i \leq \widehat{R}_i; \forall i$. In the bottom panels of Figure 18, the variability of the updates of the fraction $\widetilde{R}_i^{(k)}/R_i^{(k)}$ is illustrated by plotting histograms of updated values for a region with a relatively large number (53) of lattice nodes. We observe that the variability is largest for data set II, with a minimum value

of about 0.92. The mean acceptance probabilities for the risk updates for this region are 0.56, 0.23 and 0.42 for data set I, data set II and the oral cavity cancer data respectively.

6 Oral cavity cancer data

We apply our estimation approach to a set of data on mortality from oral cavity cancer for males in Germany, over the period 1986-1990. We do not intend to do a thorough analysis of these data, but include the analysis to illustrate the method as applied to a set of real data. The data are given as counts for each of the 544 districts of Germany. The counts range from 1 to 501, with a median count of 19, and the empirical 2.5% and 97.5% quantiles of the observed counts are 3 and 124. The standardised mortality ratios (SMR) for the data were shown in the right panel of Figure 1. The data were analysed by Knorr-Held and Raßer (2000) who identified clusters of elevated or lowered risk using a Bayesian approach based on reversible jump MCMC.

From the bottom panels of Figure 19, we observe that as for the simulated data, the mixing of the hyper-parameters is relatively poor. There is some evidence that the algorithm has converged after about 40000 iterations, but more effort is needed to get reliable estimates of the hyper-parameters using a reasonable amount of computational effort. From trace plots of a selected number of elements of $\boldsymbol{x}$, some of which are plotted in Figure 19, we observe that the mixing is good despite the poor mixing of the hyper-parameters, and the convergence is fast. The acceptance rates are reasonably high for all but a few regions, as illustrated in the bottom panels of Figure 12. The data for the regions for which the mean acceptance probabilities of the log-risk updates are less than 10% are listed in Table 4, and we observe that they all have a relatively large or small SMR or a high observed count, one of which is the maximum observed count (501).

Region no. (i)	y_i	E_i	SMR
197	111	73.0	1.52
322	117	72.9	1.60
324	53	30.8	1.72
328	501	393.1	1.27
414	52	98.5	0.53
443	15	28.1	0.53

Table 4: The expected (E_i) and observed (y_i) aggregated counts SMR for the regions for which the acceptance rates of the log-risk updates for the oral cavity cancer data are less than 10%.

The results from applying our GMRF approach to the data, using blocks made up from regions and their 1. order neighbours, are summarised in Figure 20. The estimated log-risk surface and the corresponding estimated posterior means of the regional relative risks are shown in the upper two panels. The results can be compared to the standardised mortality ratios (SMR) shown in the bottom right panel. We observe that the overall spatial pattern of the estimated relative risk and the SMR are similar, with elevated risk in the north-eastern

and south-western parts, but that the estimated spatial risk surface is smoother. The estimated posterior mean relative risks at the regional level vary between 0.57 and 1.54. The results are similar to the ones obtained by Knorr-Held and Raßer (2000). They reported estimated posterior median relative risks in the range 0.65 and 1.42 using their Bayesian cluster detection approach, and between 0.56 and 1.56 using the method of Besag et al. (1991). The estimated spatial pattern is similar to theirs, but their Bayesian clustering approach leads to a somewhat smoother map. However, the smoothness of the map using our approach will depend on the range parameter r, and since the convergence can be questioned, the result should be interpreted with care.

7 Discussion

We have presented an approach to estimation of a spatially varying risk surface based on aggregated count data, using a Gaussian Markov random field prior defined on a lattice. The method is exact in the sense that the posterior mean estimates are generated on the basis of samples from a Markov chain that converges to the correct posterior distribution. This represents an improvement over the geostatistical approach of Kelsall and Wakefield (2002) using a log-Normal approximation to the regional relative risk, in particular in applications for which the regions representing the level of aggregation of the data vary substantially in size and shape. As illustrated in Section 5, for the regions of our study the approximation gives very similar results, but in general the approximation should be justified for the actual set of regions at hand.

We are still left with the problem of convergence of the MCMC sampling algorithm. For the simulated examples and the data set analysed in Sections 5 and 6 the convergence is fast for the elements of the log-relative risk surface $\boldsymbol{x}$, and the mixing is good except for elements of $\boldsymbol{x}$ corresponding to extremes within the range of the relative risks. The acceptance rates for the Metropolis-Hastings sampler are increased by reducing the size of the blocks in the block-MCMC algorithm, at the expense of increased computational cost. For the hyper-parameters, the mixing turned out to be relatively poor, but the estimated posterior means of the elements of the log-risk surface seemed to be stable despite the poor mixing of the individual hyper-parameters. Using a single site Metropolis-Hastings sampling approach, convergence and mixing is often improved by re-parameterisation, but this will have less effect in our case, since we already accept or reject the proposed values of the range parameter r and the precision τ jointly. We chose to block the hyper-parameters with the boundary nodes, an approach that was shown to be equivalent to sampling τ and r from the marginal posterior distribution of (τ, r), integrating over the boundary nodes. To study the effect of blocking on the mixing of the hyper-parameters, other blocking strategies, like including the nodes corresponding to a random sample of inner regions in the block, could be explored. We ran the sampling algorithm including a randomly chosen inner region and it's 1. order neighbourhood in the block, but no improvement in mixing of r or τ was gained.

We have illustrated our approach using the exponential correlation function to specify the spatial correlation structure. This could be replaced by alternative, more flexible classes of models, like the Matérn class, based on Bessel functions. In Hrafnkelsson and Cressie (2003) a simpler alternative approach to that of Rue and Tjelmeland (2002) is proposed to fit a

GMRF to a geostatistical GRF model using the Matérn class of correlation functions.

The commonly used log-Gaussian random effects model for the regional level relative risk, as given by (1), includes a spatially unstructured as well as spatially structured effect, such that the degree of spatial dependency can be assessed by studying the relative values of the estimated precisions of the two effects. A spatially un-structured effect can also be introduced our model, and the proposed sampling based approach to parameter estimation can be applied to the resulting model after a re-parameterisation adding another level to the hierarchical model (see Knorr-Held and Rue, 2002).

As the methods of Best et al. (2000), using a Poisson-Gamma model with identity link, and Kelsall and Wakefield (2002) using similar distributional assumptions as in our model, our method is aggregation consistent, such that the estimated spatial structure is independent of the level of aggregation of the data. Also, the method can be extended to include covariates observed at different non-nested levels of aggregation, using all covariates at their original level of aggregation. This is an appealing feature, since ethological studies often involves data observed at the individual level, as point observations and as aggregated data.

The results from applying our approach to the German oral cavity cancer data turned out to be very similar to those reported by Knorr-Held and Raßer (2000). A closer look at the resulting risk surface displayed in the top left panel of Figure 20, reveals an apparent difference between the general level of the risk in the former German Democratic Republic (GDR), including Eastern Berlin, and Western Germany (BRD). This could be due to different routines for reporting cases, and the effect could be taken into account by including an indicator variable representing former country (GDR or BRD) as a covariate of the model.

We conclude that using GMRFs as proxies for GRFs on a lattice allows for the development of an aggregation consistent approach to estimating a smoothly varying risk surface based on aggregated count data. Applying the approach to simulated data as well as a set of real data using computationally efficient block-MCMC algorithms for parameter estimation, we have shown that the method reproduces the risk surface well. Despite blocking the hyper-parameters with the boundary nodes, further work seems to be needed to improve mixing and convergence of the hyper-parameters.

Acknowledgements

The authors would like to thank Leonhard Knorr-Held for making the German oral cavity cancer data available and for providing the data routines for generating the map of Germany, and Håkon Tjelmeland for stimulating discussions.

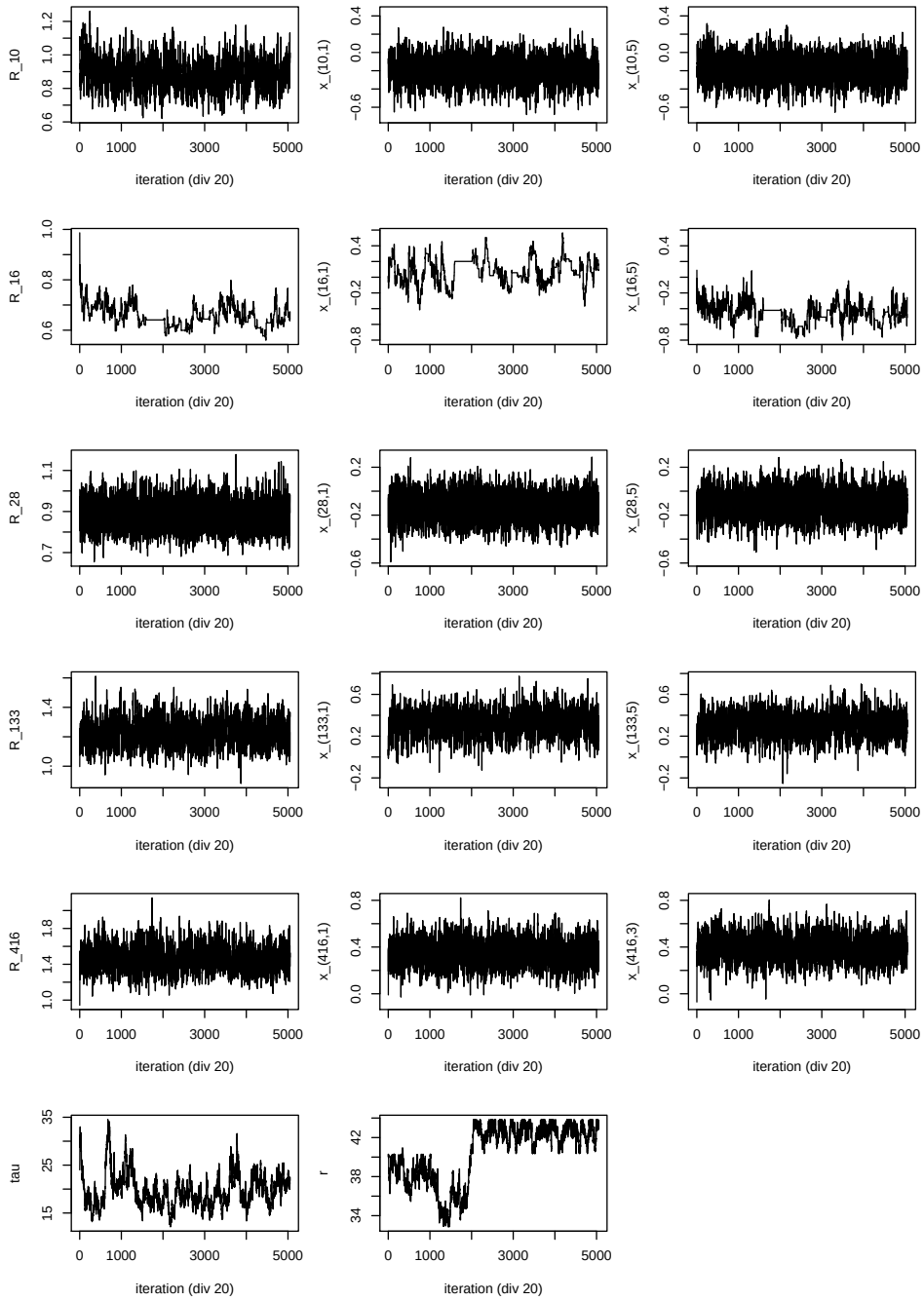


Figure 11: Selected trace plots for the simulated data set I, with $\tau = 24$. The five top rows show trace plots of the regional risk R_i for five regions (left) and of two elements of $\boldsymbol{x}$ falling within each region (middle and right). Every 20th iteration is shown.

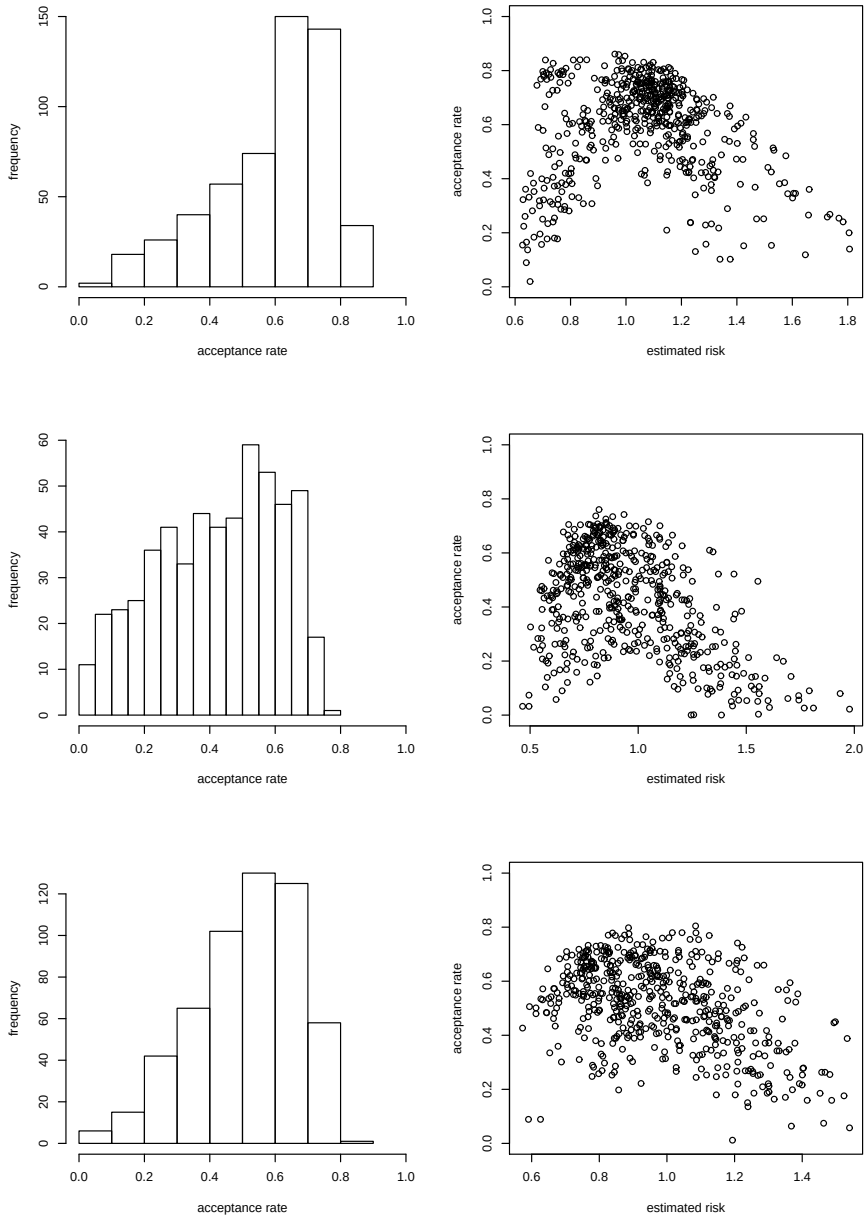


Figure 12: The acceptance rates for the block-MCMC algorithm for (from the top downward) data set I, data set II and the oral cavity cancer data. The left panels show histograms of the acceptance rates of the log-risk $\boldsymbol{x}$ within each region, and in the right panels the acceptance rates are plotted against the estimated risk.

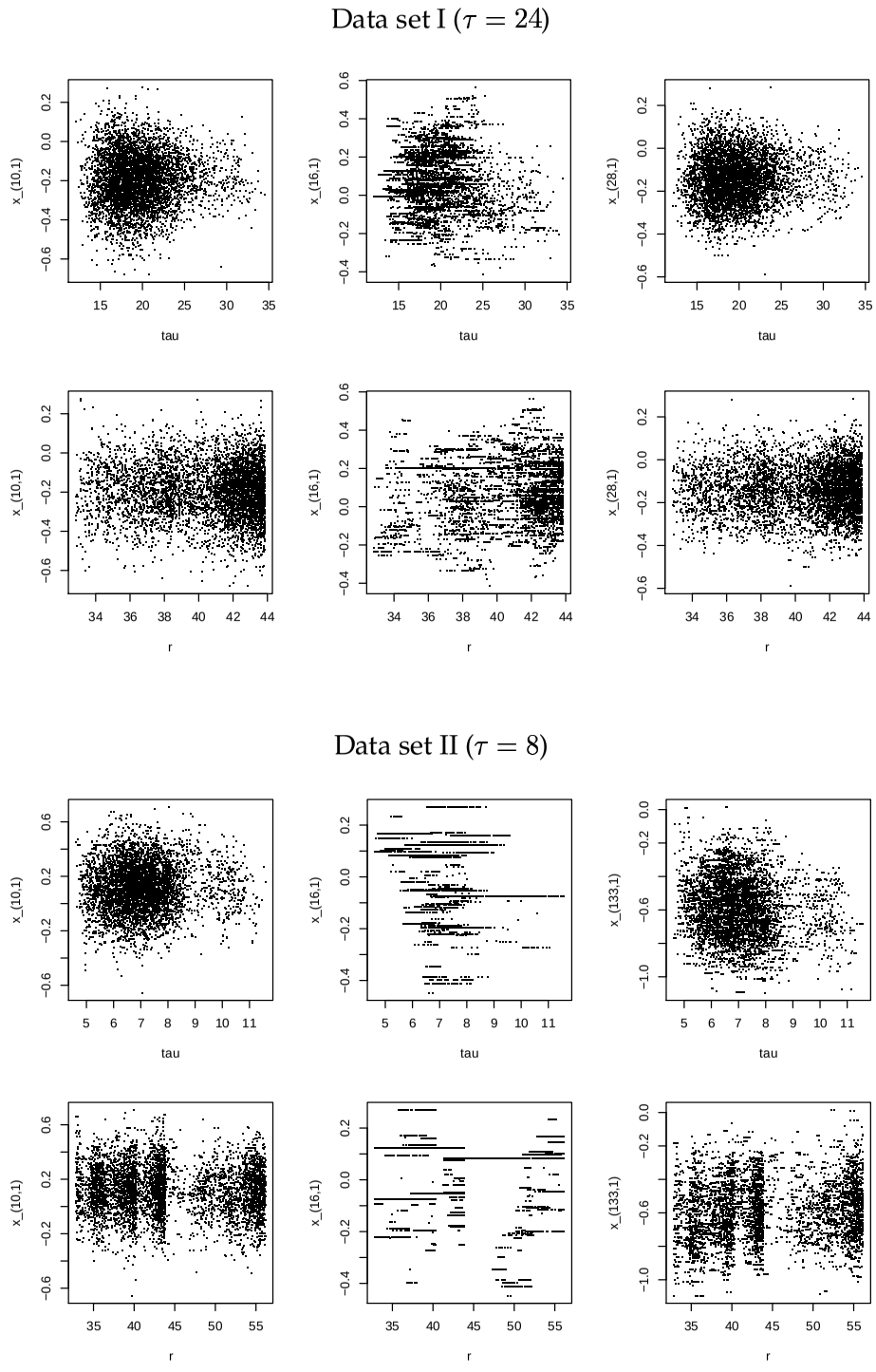


Figure 13: Scatter plots of the hyper-parameters and elements of $\boldsymbol{x}$ for data set I (top rows) and data set II (bottom rows). The numbering of the selected elements of $\boldsymbol{x}$ corresponds to the numbering in Figure 11 and Figure 15, respectively.

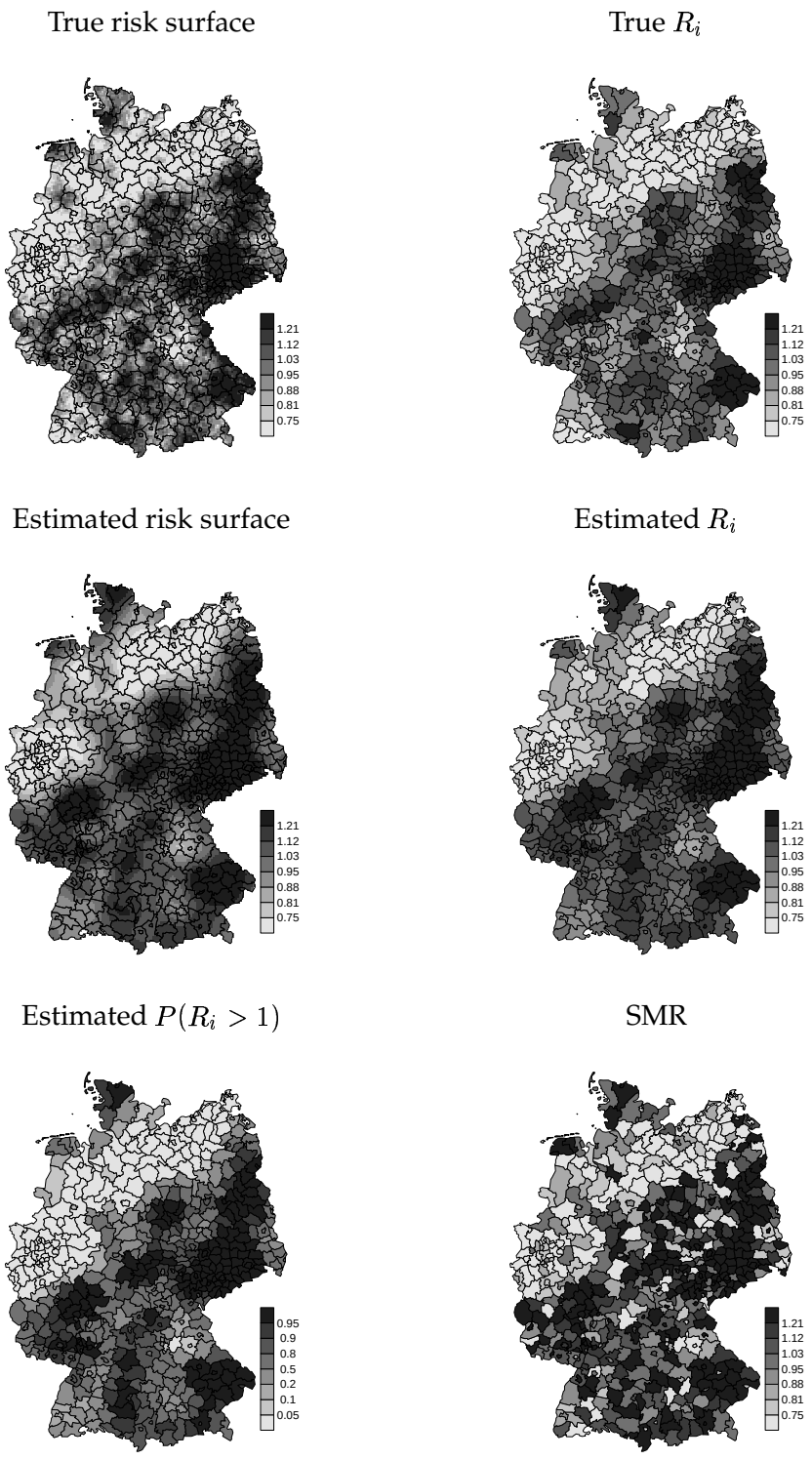


Figure 14: Results for the simulated data set I, with $\tau = 24$ and $r = 40$. The true values of the risk surface and regional risks, also shown in Figure 8, are added for reference.

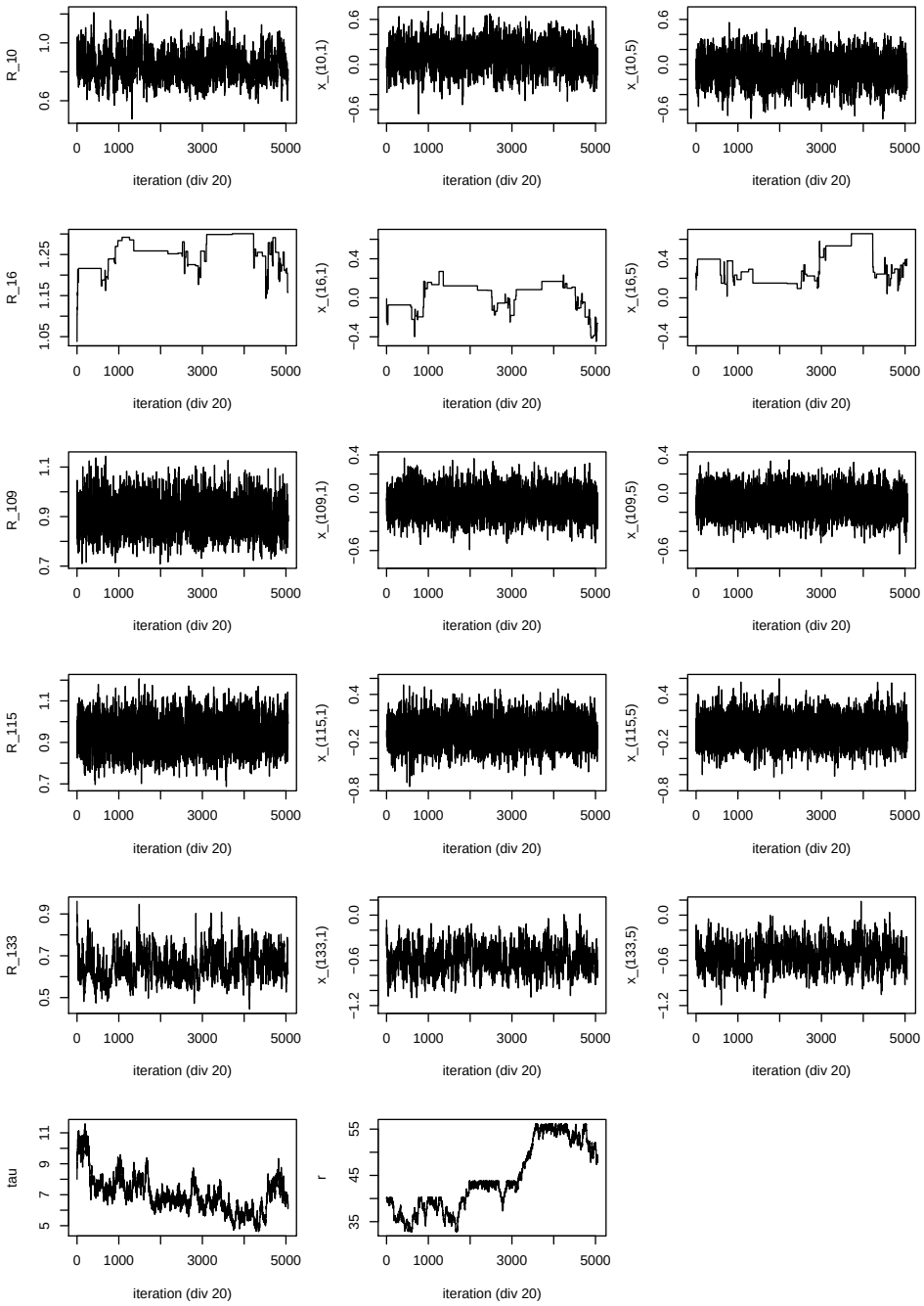


Figure 15: Selected trace plots for the simulated data set II, with $\tau = 8$. The five top rows show trace plots of the regional risk R_i for five regions (left) and of two elements of $\boldsymbol{x}$ falling within each region (middle and right). Every 20th iteration is shown.

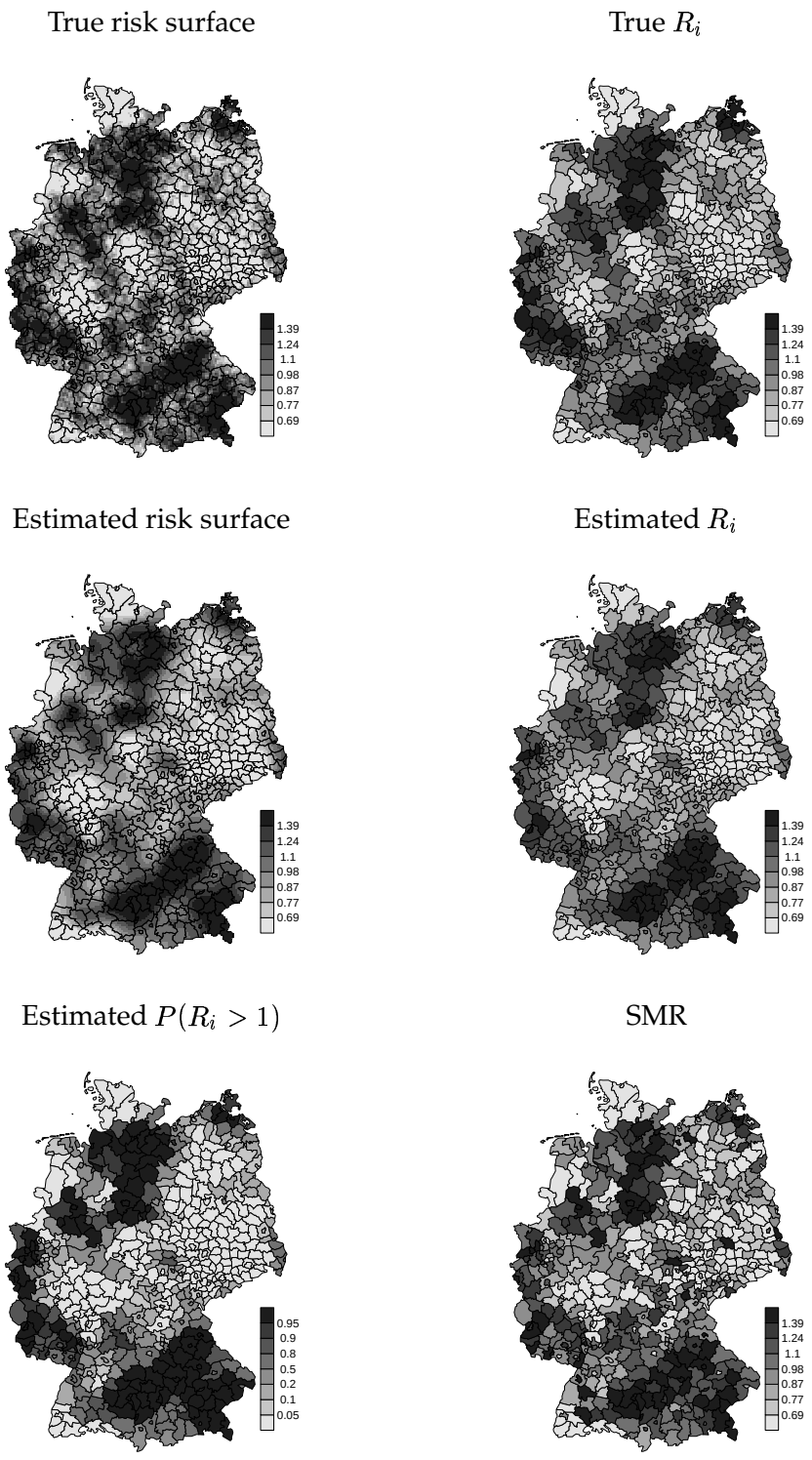


Figure 16: Results for the simulated data set II, with $\tau = 8$ and $r = 40$. The true values of the risk surface and regional risks, also shown in Figure 8, are added for reference.

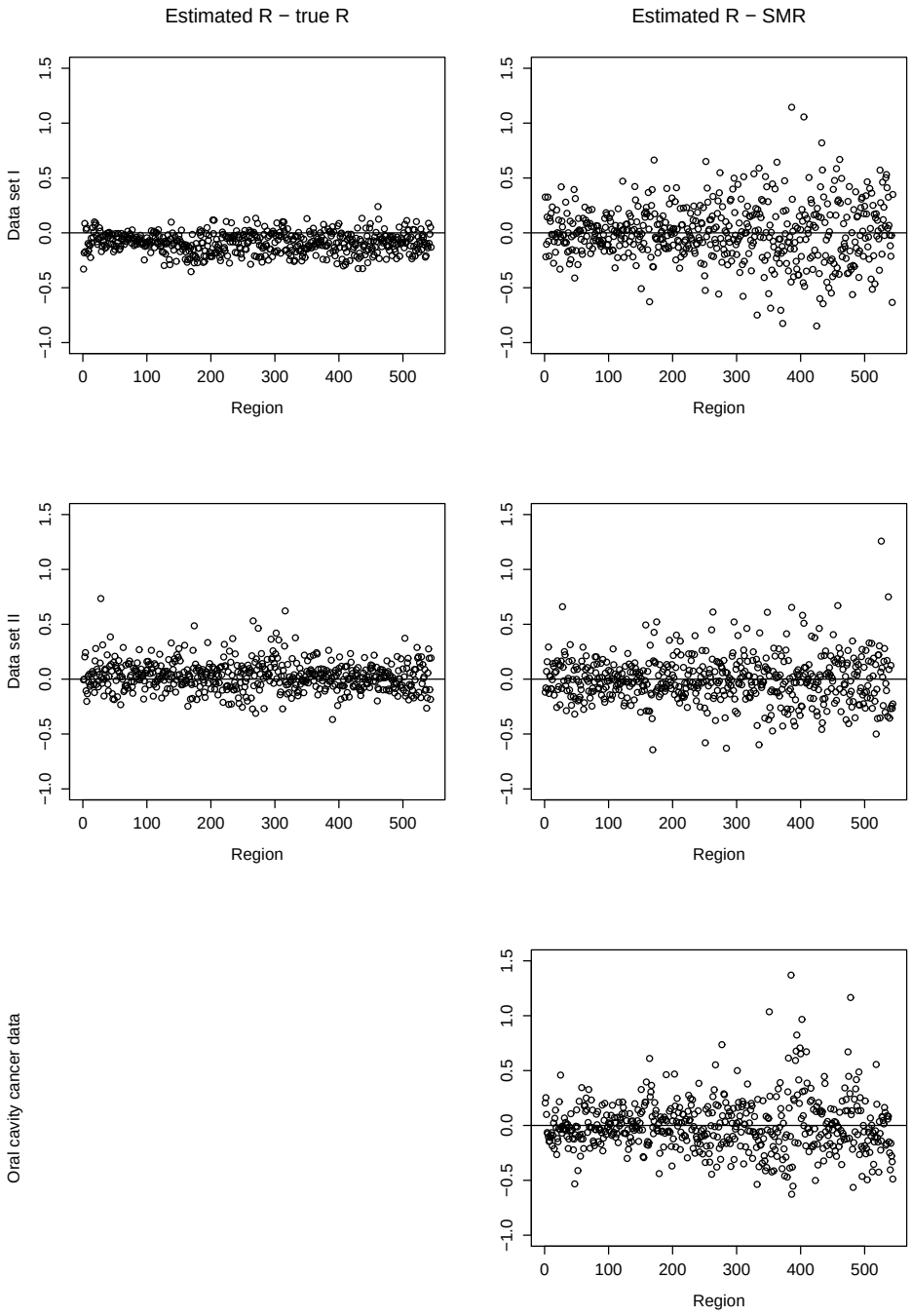


Figure 17: Differences between the estimated and true relative risk (left) and between estimated risk and SMR (right) for the simulated data sets I (top) and II (middle), and differences between estimated risk and SMR for the oral cavity cancer data (bottom).

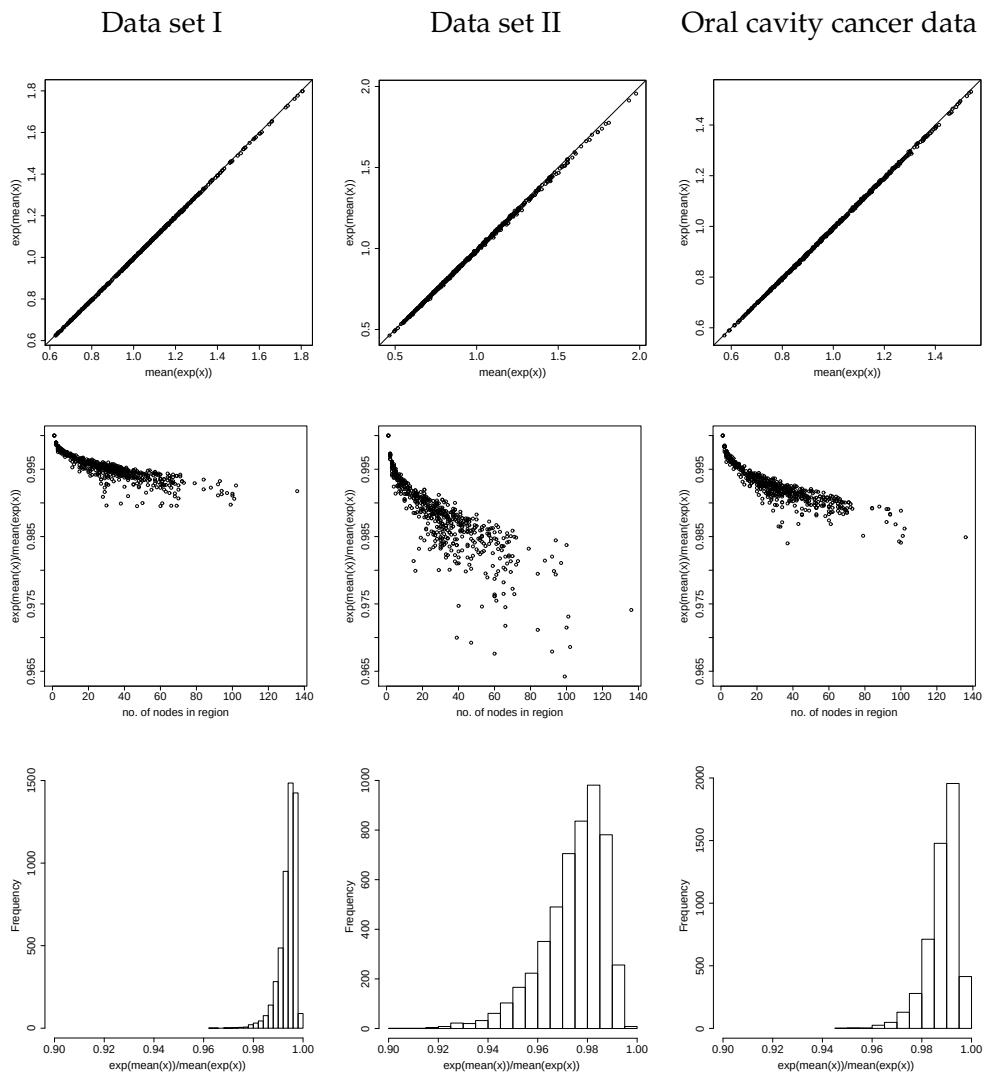


Figure 18: Plots of the posterior mean estimates $\widehat{\widetilde{R}}_i$ against $\widehat{R}_i$ (top panels) and $\widehat{\widetilde{R}}_i/\widehat{R}_i$ as a function of the number of lattice nodes within the region for the two simulated data sets and for the oral cavity cancer data. The bottom panels show histograms of the fractions $\widetilde{R}_{10}^{(k)}/R_{10}^{(k)}$ for region 10, which has 53 lattice nodes.

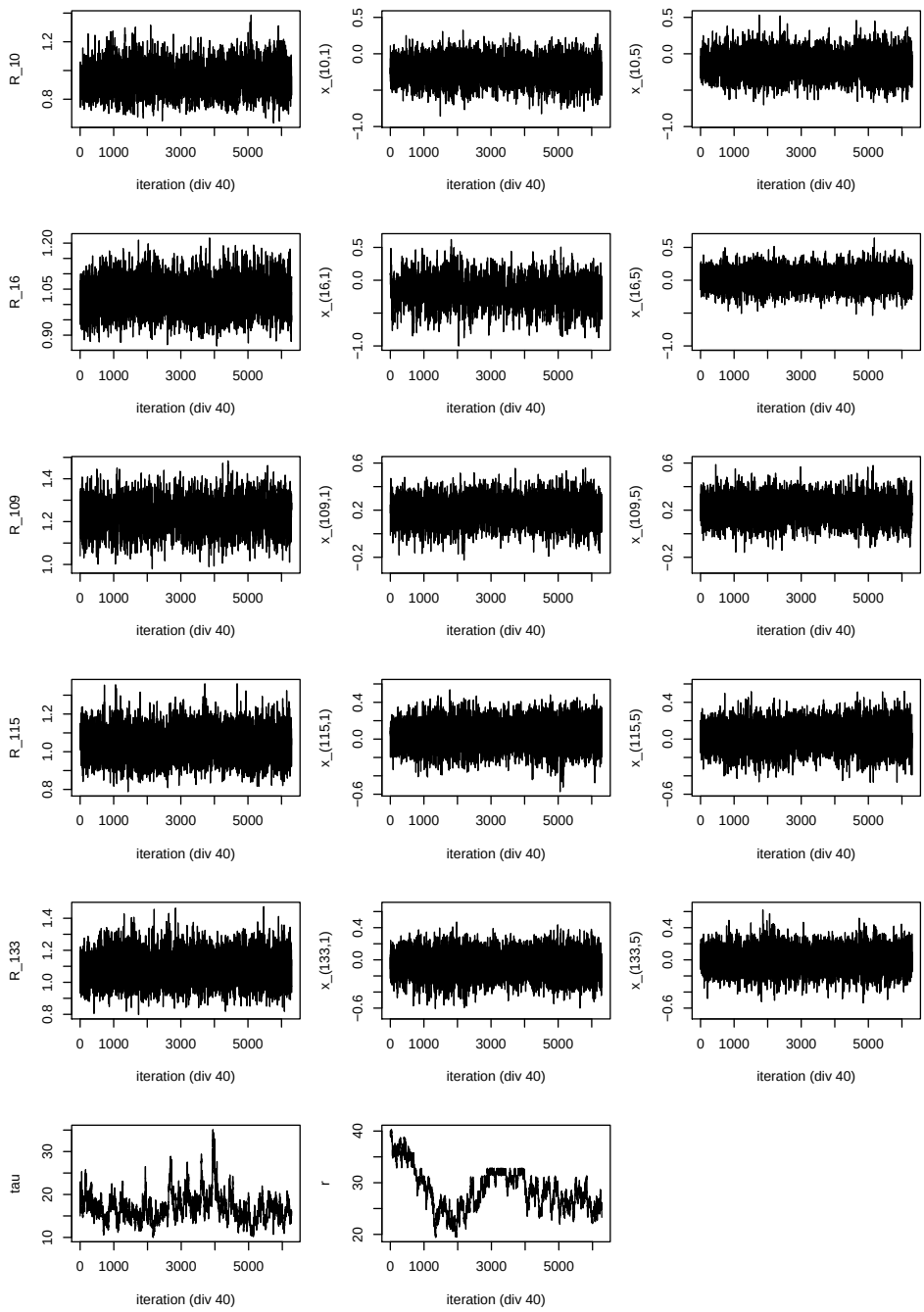


Figure 19: Selected trace plots for the oral cavity cancer data. The five top rows show trace plots of the regional risk R_i for five regions (left) and of two elements of $\boldsymbol{x}$ falling within each region (middle and right). Every 40th iteration is shown.

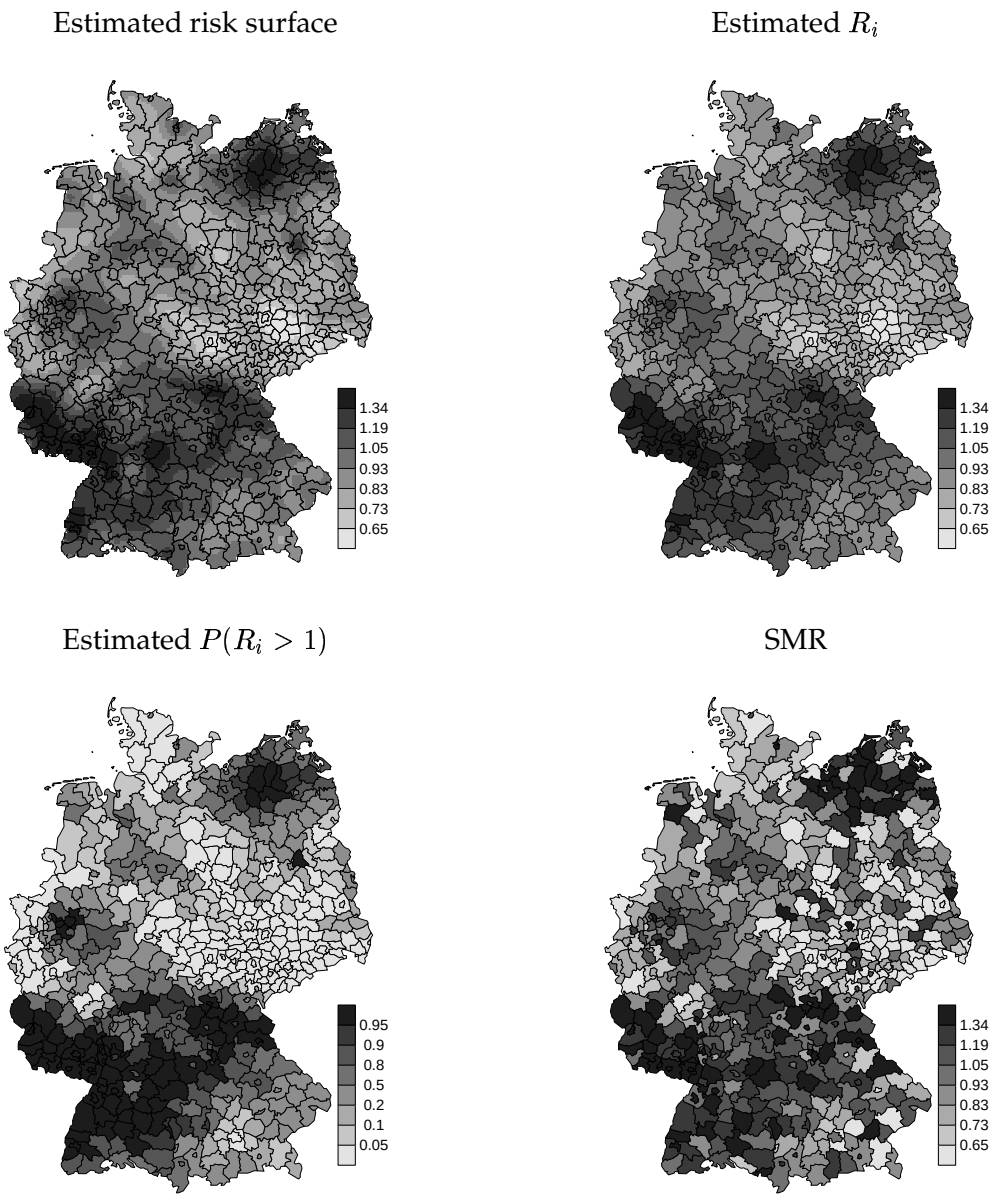


Figure 20: Results for the German oral cavity cancer data.

References

Bernardinelli, L., Pascutto, C., Best, N. G. and Gilks, W. R. (1997). Disease mapping with errors in covariates, *Statistics in Medicine* **16**: 741–752.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics* **43**: 1–59.

Best, N. G., Ickstadt, K. and Wolpert, R. L. (2000). Spatial Poisson regression for health and exposure data measured at disparate resolutions, *Journal of the American Statistical Association* **95**: 1076–1088.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics* **43**: 671–681.

Cressie, N. A. C. (1993). *Statistics for Spatial data*, Second edn, John Wiley & Sons, New York.

Diggle, P. J. (2000). Overview of statistical methods for disease mapping and its relationship to cluster detection, *in* P. Elliott, J. C. Wakefield, N. G. Best and D. J. Briggs (eds), *Spatial Epidemiology. Methods and Applications*, Oxford University Press, New York, pp. 87–103.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics (with discussion), *Applied Statistics* **47**: 299–350.

Fernández, C. and Green, P. J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach, *Journal of the Royal Statistical Society, Series B* **64**: 805–826.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London, UK.

Hrafnkelsson, B. and Cressie, N. (2003). Hierarchical modeling of count data with application to nuclear fall-out, *Journal of Environmental and Ecological Statistics*. To appear.

Ickstadt, K. and Wolpert, R. L. (1999). Spatial regression for marked point processes (with discussion), *in* J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 6*, Oxford University Press, Oxford, UK, pp. 323–341.

Kelsall, J. E. and Wakefield, J. C. (1999). Contribution to the discussion of Best, N. G., Arnold, R. A., Thomas, A.,Waller, L. A. and Conlon, E. M.: "Bayesian models for spatially correlated disease and exposure data", *in* J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 6*, Oxford University Press, Oxford, UK, pp. 131–156.

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

Knorr-Held, L. and Besag, J. (1998). Modelling risk from a disease in time and space, *Statistics in Medicine* **17**: 2045–2060.

Knorr-Held, L. and Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps, *Biometrics* **56**: 13–21.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping, *Scandinavian Journal of Statistics* **29**: 597–614.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Second edn, Chapman & Hall, London, UK.

Mollié, A. (1996). Bayesian mapping of disease, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London, UK, pp. 359–379.

Pascutto, C., Wakefield, J. C., Best, N. G., Richardson, S., Bernardinelli, L., Staines, A. and Elliott, P. (2000). Statistical issues in the analysis of disease mapping data, *Statistics in Medicine* **19**: 2493–2519.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields, *Journal of the Royal Statistical Society, Series B* **63**: 325–338.

Rue, H. and Follestad, T. (2002). GMRFLib: a C-library for fast and exact simulation of Gaussian Markov random fields, *Preprint series in statistics no. 1/2002*, Dept. of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields, *Scandinavian Journal of Statistics* **29**: 31–49.

Wakefield, J. C. and Morris, S. E. (1999). Spatial dependence and errors-in-variables in environmental epidemiology (with discussion), *in* J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 6*, Oxford University Press, Oxford, UK, pp. 657–684.

Wakefield, J. C. and Morris, S. E. (2001). The Bayesian modeling of disease risk in relation to a point source, *Journal of the American Statistical Association* **96**: 77–91.

Wakefield, J. C., Best, N. G. and Waller, L. (2000). Bayesian approaches to disease mapping, *in* P. Elliott, J. C. Wakefield, N. G. Best and D. J. Briggs (eds), *Spatial Epidemiology. Methods and Applications*, Oxford University Press, New York, pp. 104–127.

Waller, L. A., Carlin, B. P., Xia, H. and Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates, *Journal of the American Statistical Association* **92**: 607–617.

Walter, S. D. (2000). Disease mapping: a historical perspective, *in* P. Elliott, J. C. Wakefield, N. G. Best and D. J. Briggs (eds), *Spatial Epidemiology. Methods and Applications*, Oxford University Press, New York, pp. 223–239.

Wolpert, R. L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics, *Biometrica* **85**: 251–267.

A Computational details

A.1 The gradient and the Hessian of the Poisson log-likelihood

Define

$$\boldsymbol{a}_i(\boldsymbol{x}_{\mathcal{A}_i}) = (\exp(x_j))_{j\in\mathcal{A}_i}^T \tag{63}$$

$$S_i(\boldsymbol{x}_{\mathcal{A}_i}) = \sum_{j\in\mathcal{A}_i}\exp(x_j) = \sum_{k=1}^{n_i} a_{ik} \tag{64}$$

where x_j denotes element j of the log-risk surface $\boldsymbol{x}$. We compute the gradient vector $\boldsymbol{g}_i(\boldsymbol{x})$ and the Hessian matrix $\boldsymbol{G}_i(\boldsymbol{x})$ of the Poisson log-likelihood function given by $h_i(\boldsymbol{x}) = y_i \log(E_i' \sum_{j\in\mathcal{A}_i}\exp x_j) - E_i' \sum_{j\in\mathcal{A}_i}\exp x_j$, which define the second order Taylor approximation (33) of the conditional posterior distribution $\pi(\boldsymbol{x}_{\mathcal{A}_i}|\boldsymbol{x}_{-\mathcal{A}_i},\boldsymbol{\theta},\boldsymbol{y})$. The elements of $\boldsymbol{g}_i(\boldsymbol{x})$ and $\boldsymbol{G}_i(\boldsymbol{x})$ are given by

$$\frac{\partial h_i(\boldsymbol{x})}{\partial x_k} = (\frac{y_i}{S_i(\boldsymbol{x}_{\mathcal{A}_i})} - E_i')I_{[k\in\mathcal{A}_i]}\exp(x_k) \tag{65}$$

$$\frac{\partial^2 h_i(\boldsymbol{x})}{\partial x_k \partial x_l} = \begin{cases} \big((\frac{y_i}{S_i(\boldsymbol{x}_{\mathcal{A}_i})} - E_i')I_{[k\in\mathcal{A}_i]} - \frac{y_i}{S_i(\boldsymbol{x}_{\mathcal{A}_i})^2}I_{[k\in\mathcal{A}_i]}\exp(x_l)\big)\exp(x_k) & \text{if } l = k \\ -\frac{y_i}{S_i(\boldsymbol{x}_{\mathcal{A}_i})^2}I_{[k\in\mathcal{A}_i\cap l\in\mathcal{A}_i]}\exp(x_l)\exp(x_k) & \text{if } l \neq k. \end{cases} \tag{66}$$

Consequently, the vector $\boldsymbol{g}_i(\boldsymbol{x}^0_{\mathcal{A}_i})$ and the matrix $\boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i})$, evaluated in the mode $\boldsymbol{x}^0_{\mathcal{A}_i}$ of the posterior distribution of $\boldsymbol{x}_{\mathcal{A}_i}$, can be expressed by

$$\boldsymbol{g}_i(\boldsymbol{x}^0_{\mathcal{A}_i}) = (\frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{A}_i})} - E_i')\ \boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{A}_i}) \tag{67}$$

$$\boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{A}_i}) = (\frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{A}_i})} - E_i')\ \mathrm{diag}(\boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{A}_i})) - \frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{A}_i})^2}\boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{A}_i})\boldsymbol{a}_i^T(\boldsymbol{x}^0_{\mathcal{A}_i}), \tag{68}$$

establishing the expressions (37) and (38) of Section 4.1.

A.2 A Gaussian approximation to the posterior of $\boldsymbol{x}_{\mathcal{S}}$ for general sets of regions $\mathcal{S}$

Here, we establish the Gaussian approximation to the conditional posterior distribution $\pi(\boldsymbol{x}_{\mathcal{S}}|\boldsymbol{x}_{-\mathcal{S}},\boldsymbol{\theta},\boldsymbol{y})$ of the log-risk $\boldsymbol{x}_{\mathcal{S}}$ for blocks of lattice nodes corresponding to a set $\mathcal{S}$ of several regions, given by equation (26) in Section 3.2. In analogy to expression (33) for the single region block case, a Taylor expansion based quadratic approximation to the log-likelihood part $\sum_i h_i(\boldsymbol{x})$ of (26) is given by

$$\sum_i h_i(\boldsymbol{x}) \approx -\frac{1}{2}\boldsymbol{x}_{\mathcal{S}}^T(-\boldsymbol{G}(\boldsymbol{x}^0_{\mathcal{S}}))\boldsymbol{x}_{\mathcal{S}} + (\boldsymbol{g}^T(\boldsymbol{x}^0_{\mathcal{S}}) - (\boldsymbol{x}^0_{\mathcal{S}})^T\boldsymbol{G}(\boldsymbol{x}^0_{\mathcal{S}}))\boldsymbol{x}_{\mathcal{S}}, \tag{69}$$

discarding terms not depending on $\boldsymbol{x}_{\mathcal{S}}$. The vector $\boldsymbol{g}(\boldsymbol{x}^0_{\mathcal{S}})$ and the matrix $\boldsymbol{G}(\boldsymbol{x}^0_{\mathcal{S}})$ are the gradient and the Hessian of $\sum_i h_i(\boldsymbol{x})$ evaluated in the mode $\boldsymbol{x}^0_{\mathcal{S}}$ of the posterior distribution of $\boldsymbol{x}_{\mathcal{S}}$.

We derive the Gaussian approximation using the Poisson likelihood (15). As for the single region case in Appendix A.1, define

$$\boldsymbol{a}_i(\boldsymbol{x}_{\mathcal{S}}) = \left(\exp(x_j) I_{[j\in\mathcal{A}_i]}\right)_{j\in\mathcal{S}}, \text{ and} \tag{70}$$

$$S_i(\boldsymbol{x}_{\mathcal{S}}) = \sum_{j\in\mathcal{S}} (\exp(x_j) I_{[j\in\mathcal{A}_i]}) = \sum_{k=1}^{n_{\mathcal{S}}} a_{ik}. \tag{71}$$

Because of the conditional independence structure of the likelihood, the gradient and the Hessian defining the Taylor expansion of $\sum_i h_i(\boldsymbol{x})$ are given from (67) and (68) by the sums

$$\boldsymbol{g}(\boldsymbol{x}^0_{\mathcal{S}}) = \sum_i \left(\frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{S}})} - E_i'\right) \boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{S}}) \tag{72}$$

$$\boldsymbol{G}(\boldsymbol{x}^0_{\mathcal{S}}) = \sum_i \left(\frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{S}})} - E_i'\right) \operatorname{diag}(\boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{S}})) - \sum_i \frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{S}})^2} \boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{S}}) \boldsymbol{a}_i^T(\boldsymbol{x}^0_{\mathcal{S}}), \tag{73}$$

such that the precision matrix $-\boldsymbol{G}(\boldsymbol{x}^0_{\mathcal{S}})$ of the quadratic approximation to $\sum_i h_i(\boldsymbol{x})$ is block diagonal with blocks $-\boldsymbol{G}_i(\boldsymbol{x}^0_{\mathcal{S}})$ corresponding to (68) for each block $\mathcal{S}$. Define the $m_{\mathcal{S}} \times n_{\mathcal{S}}$ matrix $\boldsymbol{A}_{\mathcal{S}}$ by

$$\boldsymbol{A}_{\mathcal{S}} = \left(\frac{\sqrt{y_i}}{S_i(\boldsymbol{x}^0_{\mathcal{S}})} \boldsymbol{a}_i^T(\boldsymbol{x}^0_{\mathcal{S}})\right)_{i:\ \mathcal{A}_i\in\mathcal{S}}, \tag{74}$$

where $m_{\mathcal{S}}$ and $n_{\mathcal{S}}$ are the number of regions and number of lattice nodes within the block $\mathcal{S}$, respectively. In correspondence with the quadratic approximation for the single region case, the matrix $\boldsymbol{G}(\boldsymbol{x}^0_{\mathcal{S}})$ is of the form

$$\boldsymbol{G}(\boldsymbol{x}^0_{\mathcal{S}}) = -(\boldsymbol{D}_{\mathcal{S}} + \boldsymbol{H}_{\mathcal{S}}) \tag{75}$$

for a diagonal matrix $\boldsymbol{D}_{\mathcal{S}}$ and rank one matrix $\boldsymbol{H}_{\mathcal{S}}$ given by

$$\boldsymbol{D}_{\mathcal{S}} = -\sum_i \left(\frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{S}})} - E_i'\right) \operatorname{diag}(\boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{S}})) \tag{76}$$

$$\boldsymbol{H}_{\mathcal{S}} = \sum_i \frac{y_i}{S_i(\boldsymbol{x}^0_{\mathcal{S}})^2} \boldsymbol{a}_i(\boldsymbol{x}^0_{\mathcal{S}_i}) \boldsymbol{a}_i^T(\boldsymbol{x}^0_{\mathcal{S}}) = \boldsymbol{A}_{\mathcal{S}}^T \boldsymbol{A}_{\mathcal{S}}. \tag{77}$$

Substituting (69) for $\sum_i h_i(\boldsymbol{x})$ in the posterior distribution (26), using the expressions for $\boldsymbol{g}(\boldsymbol{x}^0_{\mathcal{S}})$ and $\boldsymbol{G}(\boldsymbol{x}^0_{\mathcal{S}})$ derived above and collecting terms that are linear and quadratic in $\boldsymbol{x}_{\mathcal{S}}$, we arrive at the Gaussian approximation given by (27).

A.3 Conditioning on a soft linear constraint

Here, we use Normal distribution theory to check the validity of equation (54) as a sample from the GMRF $\boldsymbol{x}$ conditionally on a soft linear constraint. Let

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{Q}) \tag{78}$$

and consider the general problem of sampling from the conditional distribution

$$\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} + \boldsymbol{\epsilon}, \tag{79}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. This is equivalent to sampling from the distribution

$$\boldsymbol{x} \mid \boldsymbol{z} = \boldsymbol{b}, \tag{80}$$

where $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{\epsilon}$. The sampling problem of Section 4 is a special case of (79) for which $\boldsymbol{\mu} = \boldsymbol{\mu}^*$ and $\boldsymbol{Q} = \boldsymbol{Q}^*$ as defined by (46) and (47), and where $\boldsymbol{\Sigma} = \boldsymbol{I}$ and $\boldsymbol{b} = \mathbf{0}$.

Let $\boldsymbol{x}_u$ be an *unconditional* sample for $\boldsymbol{x}$ from (78), and let

$$\boldsymbol{x}_c = \boldsymbol{x}_u - \boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T + \boldsymbol{\Sigma})^{-1}(\boldsymbol{z} - \boldsymbol{b}), \tag{81}$$

where $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}_u - \boldsymbol{\epsilon}$. We will show that $\boldsymbol{x}_c$ has the same distribution as a sample from $\boldsymbol{x}|\boldsymbol{z} = \boldsymbol{b}$, and thus from (79) by comparing the mean and variance of $\boldsymbol{x}_c$ computed by (81) to the moments of (80).

Using multivariate Normal distribution theory, the mean vector and covariance matrix of the distribution $\boldsymbol{x}|\boldsymbol{z} = \boldsymbol{b}$ can be shown to be

$$\begin{aligned} \mathrm{E}(\boldsymbol{x}|\boldsymbol{z} = \boldsymbol{b}) &= \boldsymbol{\mu} + \boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T + \boldsymbol{\Sigma})^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{\mu}) \\ \mathrm{Cov}(\boldsymbol{x}|\boldsymbol{z} = \boldsymbol{b}) &= \boldsymbol{Q}^{-1} + \boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T + \boldsymbol{\Sigma})^{-1}\boldsymbol{A}\boldsymbol{Q}^{-1} \\ &= (\boldsymbol{Q} + \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}. \end{aligned} \tag{82}$$

The mean vector and covariance matrix of $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{\epsilon}$ is

$$\mathrm{E}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{\mu}, \text{ and} \tag{83}$$

$$\mathrm{Cov}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T + \boldsymbol{\Sigma}, \tag{84}$$

such that

$$\begin{aligned} \mathrm{E}(\boldsymbol{x}_c) &= \boldsymbol{\mu} - \boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T + \boldsymbol{\Sigma})^{-1}(\boldsymbol{A}\boldsymbol{\mu} - \boldsymbol{b}) \\ &= \mathrm{E}(\boldsymbol{x}|\boldsymbol{z} = \boldsymbol{b}), \end{aligned} \tag{85}$$

and

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{x}_c) &= \boldsymbol{Q}^{-1} + \boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T + \boldsymbol{\Sigma})^{-1}(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T + \boldsymbol{\Sigma})(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T + \boldsymbol{\Sigma})^{-1}\boldsymbol{A}\boldsymbol{Q}^{-1} \\ &= \boldsymbol{Q}^{-1} + \boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T + \boldsymbol{\Sigma})^{-1}\boldsymbol{A}\boldsymbol{Q}^{-1} \\ &= \mathrm{Var}(\boldsymbol{x}|\boldsymbol{z} = \boldsymbol{b}). \end{aligned} \tag{86}$$

Consequently, $\boldsymbol{x}_c$ computed by (81) has the same first and second order moments as a sample from (79), and since the corresponding distributions are both Gaussian, the validity of (54) as an update of (45) follows.