An intuitive Bayesian spatial model for disease mapping that accounts for scaling

Andrea Riebler,[1] Sigrunn H Sørbye,[2]
Daniel Simpson[3] and Håvard Rue[1]

Abstract

In recent years, disease mapping studies have become a routine application within geographical epidemiology and are typically analysed within a Bayesian hierarchical model formulation. A variety of model formulations for the latent level have been proposed but all come with inherent issues. In the classical BYM (Besag, York and Mollié) model, the spatially structured component cannot be seen independently from the unstructured component. This makes prior definitions for the hyperparameters of the two random effects challenging. There are alternative model formulations that address this confounding; however, the issue on how to choose interpretable hyperpriors is still unsolved. Here, we discuss a recently proposed parameterisation of the BYM model that leads to improved parameter control as the hyperparameters can be seen independently from each other. Furthermore, the need for a scaled spatial component is addressed, which facilitates assignment of interpretable hyperpriors and make these transferable between spatial applications with different graph structures. The hyperparameters themselves are used to define flexible extensions of simple base models. Consequently, penalised complexity priors for these parameters can be derived based on the information-theoretic distance from the flexible model to the base model, giving priors with clear interpretation. We provide implementation details for the new model formulation which preserve sparsity properties, and we investigate systematically the model performance and compare it to existing parameterisations. Through a simulation study, we show that the new model performs well, both showing good learning abilities and good shrinkage behaviour. In terms of model choice criteria, the proposed model performs at least equally well as existing parameterisations, but only the new formulation offers parameters that are interpretable and hyperpriors that have a clear meaning.

Keywords

---

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[2]Department of Mathematics and Statistics, UiT The Arctic University of Norway, Tromsø, Norway
[3]Department of Mathematical Sciences, University of Bath, Bath, UK

Corresponding author:
Andrea Riebler, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.
Email: andrea.riebler@math.ntnu.no

1 Introduction

Over the recent years, there has been much interest in spatial modelling and mapping of disease or mortality rates. Due to inherent sampling variability it is not recommended to inspect crude rates directly, but borrow strength over neighbouring regions to get more reliable region-specific estimates. The state of the art is to use Bayesian hierarchical models, where the risk surface is modelled using a set of spatial random effects, in addition to potentially available covariate information.[1] The random effects shall capture unobserved heterogeneity or spatial correlation that cannot be explained by the available covariates.[2] However, there has been much confusion on the design of the spatial random effects. Besag et al.[3] proposed an intrinsic autoregressive model, often referred to as the CAR prior or Besag model, where the spatial effect of a particular region depends on the effects of all neighbouring regions. They also proposed the commonly known BYM model, where an additional unstructured spatial random effect is included to account for independent region-specific noise. Consequently, various model modifications and alternative approaches have been proposed, see for example Leroux et al.,[4] Stern and Cressie[5] and Dean et al.[6] The appropriateness and behaviour of different latent spatial models have been compared in a full Bayesian context,[2,7–9] an empirical Bayes setting[10] or using penalised quasi likelihood.[4] It is accepted that the Besag model may lead to misleading results in the case where there is no spatial correlation in the data.[4,8] For this purpose, most alternative models propose to account for unstructured variability in space.

By including both structured and unstructured components in the model, potential confounding must be addressed,[11] as otherwise it is not clear how to split the variability over the effects. This problem has motivated development of reparameterised models, in which the precision parameters of the two components are replaced by a common precision parameter and a mixing parameter, which distributes the variability between the structured and unstructured components.[4,6]

However, existing approaches do have two issues. First, the spatially structured component is not scaled. This implies that the precision parameter does not represent the marginal precision but is confounded with the mixing parameter. Thus, the effect of any prior assigned to the precision parameter depends on the graph structure of the application. This has the additional effect that a given prior is not transferable between different applications if the underlying graph changes.[12] Second, the choice of hyperpriors for the random effects is not straightforward. Approaches to design epidemiologically sensible hyperpriors include, among others, Wakefield[8] and Bernardinelli et al.[11]

Recently, Simpson et al.[13] proposed a new BYM parameterisation that accounts for scaling and provides an intuitive way to define priors by taking the model structure into account. This new model provides a new way to look at the BYM model. The primary goal is not to optimise model choice criteria, such as deviance information criterion (DIC) values, but to provide a sensible model formulation where all parameters have a clear meaning. The model structure is similar to the Dean model,[6] with the crucial modification that the precision parameter is mapped to the marginal standard deviation. This makes the parameters of the model interpretable and facilitates assignment of meaningful hyperpriors. The framework of penalised complexity (PC) priors is applied to formulate prior distributions for the hyperparameters. The spatial model is thereby seen as a flexible extension of two simpler base models that it will shrink towards, if not indicated otherwise by the data.[13] The upper level base model assumes a constant risk surface, while the lower level model assumes a varying risk surface over space without spatial autocorrelation. In this paper, we investigate systematically the model behaviour under different simulation scenarios. Furthermore, we compare this new model formulation to commonly used model formulations in disease mapping. We point out differences, focusing on parameter

interpretability and prior distribution assignment. For completeness we also compare commonly used model choice criteria.

We have chosen to implement all models using integrated nested Laplace approximations (INLA),[14] available to the applied user via the R-package INLA (see www.r-inla.org). INLA is straightforward to use for full Bayesian inference in disease mapping and avoids any Markov chain Monte Carlos techniques.[15,16]

The plan for the paper is as follows. Section 2 gives an introduction to disease mapping and motivates the use of Bayesian hierarchical models. Commonly used spatial models for the latent level are reviewed in Section 3, before the need of scaling and consequently the new modified spatial model is presented. Section 4 uses the PC-prior framework to design sensible hyperpriors for the precision and mixing parameter. In Section 5, we investigate the properties and behaviour of the new model in a simulation study, including comparisons to commonly used models. This section also contains an application to insulin-dependent diabetes mellitus (IDDM) data in Sardinia. Discussion and concluding remarks are given in Section 6.

2 Disease mapping

Disease mapping has a long history in epidemiology and one major goal is to investigate the geographical distribution of disease burden.[8] In the following, we assume that our region of interest is divided into n non-overlapping areas. Let E_i denote the number of persons at risk in area i ($i = 1, \ldots, n$). These expected numbers are commonly calculated based on the size and demographic characteristics of the population living in area i.[2] Further, let y_i denote the number of cases or deaths in region i. When the disease is non-contagious and rare, it is usually reasonable to assume that

$$y_i \,|\, \theta_i \sim \text{Poisson}(E_i \,\theta_i)$$

where θ_i, $i = 1, \ldots, n$, denotes the underlying true area-specific relative risk.[11]

The maximum likelihood estimator of θ_i is given by $\hat{\theta}_i = y_i/E_i$ and corresponds to the standardised mortality ratio (SMR) if the counts represent deaths. Mapping the SMRs directly can, however, be misleading. Areas with extreme SMR values often have small expected numbers, so that sampling variability is more pronounced. This is reflected in large values for $\text{Var}(\hat{\theta}_i) = \theta_i/E_i$, see Bernardinelli et al.[17] or Wakefield[8] for more details.

Due to these potential instabilities, methods have been proposed to smooth the raw estimates over space. Bayesian hierarchical models are commonly used in this context, where the latent model involves random effects that borrow strength over neighbouring regions to achieve local smoothness.[17] We refer to Lawson[1] and Banerjee et al.[18] for a detailed description of the class of Bayesian hierarchical models.

A general model formulation is to assume that the log risk $\eta_i = \log(\theta_i)$ has the decomposition

$$\eta_i = \mu + \mathbf{z}_i^\top \beta + b_i$$

Here, μ denotes the overall risk level, $\underline{\mathbf{z}}_i^\top = (z_{i1}, \ldots, z_{ip})^\top$ a set of p covariates with corresponding regression parameters $\beta = (\beta_1, \ldots, \beta_p)^\top$ and b_i a random effect. For μ and β we assign weekly informative Gaussian distributions with mean zero and large variance. The random effects $\mathbf{b} = (b_1, \ldots, b_n)^\top$ are used to account for extra-Poisson variation or spatial correlation due to latent or unmeasured risk factors. In the following, we will review four models that are commonly used for $\mathbf{b}$.

3 Modelling the spatial dependency structure

In general, it seems reasonable to assume that areas that are close in space show more similar disease burden than areas that are not close. To define what 'close' means we need to set up a neighbourhood structure. A common assumption is to regard areas i and j as neighbours if they share a common border, denoted here as $i \sim j$. This seems reasonable if the regions are equally sized and arranged in a regular pattern.[8] We further denote the set of neighbours of region i by δi and its size by $n_{\delta i}$.

3.1 Review of common models

3.1.1 The Besag model

One of the most popular approaches to model spatial correlation is an intrinsic Gaussian Markov random field (GMRF) model,[3] here referred to as the Besag model. The conditional distribution for b_i is

$$b_i | \mathbf{b}_{-i}, \tau_b \sim \mathcal{N}\left(\frac{1}{n_{\delta i}} \sum_{j \in \delta i} b_j, \frac{1}{n_{\delta i} \tau_b} \right) \tag{1}$$

where τ_b is a precision parameter and $\mathbf{b}_{-i} = (b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n)^\top$. The mean of b_i equals the mean of the effects over all neighbours, and the precision is proportional to the number of neighbours. The joint distribution for $\mathbf{b}$ is

$$\pi(\mathbf{b} \mid \tau_b) \propto \exp\left(-\frac{\tau_b}{2} \sum_{i \sim j} (b_i - b_j)^2 \right) \propto \exp\left(-\frac{\tau_b}{2} \mathbf{b}^\top \mathbf{Q} \mathbf{b} \right)$$

where the precision matrix $\mathbf{Q}$ has the entries

$$Q_{ij} = \begin{cases} n_{\delta i} & i = j, \\ -1 & i \sim j, \\ 0 & \text{else} \end{cases} \tag{2}$$

The model is intrinsic in the sense that $\mathbf{Q}$ is singular, i.e. it has a non-empty null-space $\mathbf{V}$, see Rue and Held (Section 3) for details.[19] For any map, $\mathbf{1}$, denoting a vector of length n with only ones, is an eigenvector of $\mathbf{Q}$ with eigenvalue 0. Hence the density is invariant to the addition of a constant. If the spatial map has islands the definition of the null-space is more complex and depends on how islands are defined, i.e. whether they form a new unconnected graph or are linked to the main land, see Section 5.2.1 in Hodges.[20] The rank deficiency of the Besag model is equal to the number of connected subgraphs. The rank deficiency is equal to 1 if all regions are connected. Hence, the density for the Besag model is

$$\pi(\mathbf{b} \mid \tau_b) = K \tau_b^{(n-I)/2} \exp\left(-\frac{\tau_b}{2} \mathbf{b}^\top \mathbf{Q} \mathbf{b} \right)$$

where I denotes the number of connected subgraphs and K is a constant.[20] To prevent confounding with the intercept, sum-to-zero constraints are imposed on each connected subgraph.

3.1.2 The BYM model

The Besag model only assumes a spatially structured component and cannot take the limiting form that allows for no spatially structured variability. Hence, unstructured random error or pure overdispersion within area i, will be modelled as spatial correlation, giving misleading parameter estimates.[21]

To address this issue, the Besag-York-Mollié (BYM) model[3] decomposes the regional spatial effect $\mathbf{b}$ into a sum of an unstructured and a structured spatial component, so that $\mathbf{b} = \mathbf{v} + \mathbf{u}$. Here, $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \tau_v^{-1}\mathbf{I})$ accounts for pure overdispersion, while $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \tau_u^{-1}\mathbf{Q}^-)$ is the Besag model (Section 3.1.1), whereby $\mathbf{Q}^-$ denotes the generalised inverse of $\mathbf{Q}$. The resulting covariance matrix of $\mathbf{b}$ is

$$\mathrm{Var}(\mathbf{b} \mid \tau_u, \tau_v) = \tau_v^{-1}\mathbf{I} + \tau_u^{-1}\mathbf{Q}^-$$

3.1.3 The Leroux model

In the BYM model, the structured and unstructured components cannot be seen independently from each other and are thus not identifiable.[9] An additional challenge, which makes the choice of hyperpriors more difficult, is that τ_v and τ_u do not represent variability on the same level. While τ_v^{-1} can be interpreted as the marginal variance of the unstructured random effects, τ_u^{-1} controls the variability of the structured component u_i, conditional on the effects in its neighbouring areas.[8,11] Leroux et al.[4] proposed an alternative model formulation to make the compromise between unstructured and structured variation more explicit. Here, $\mathbf{b}$ is assumed to follow a normal distribution with mean zero and covariance matrix

$$\mathrm{Var}(\mathbf{b} \mid \tau_b, \phi) = \tau_b^{-1}\big((1 - \phi)\,\mathbf{I} + \phi\mathbf{Q}\big)^{-1} \tag{3}$$

where $\phi \in [0, 1]$ denotes a mixing parameter. The model reduces to pure overdispersion if $\phi = 0$, and to the Besag model when $\phi = 1$.[9,10] The conditional expectation of b_i, given all other random effects, results as a weighted average of the zero-mean unstructured model and the mean value of the Besag model (1). The conditional variance is the weighted average of $1/\tau_b$ and $1/(\tau_b n_{\delta i})$.

3.1.4 The Dean model

Dean et al.[6] proposed a reparameterisation of the BYM model where

$$\mathbf{b} = \frac{1}{\sqrt{\tau_b}}\left(\sqrt{1 - \phi}\,\mathbf{v} + \sqrt{\phi}\,\mathbf{u}\right) \tag{4}$$

having covariance matrix

$$\mathrm{Var}(\mathbf{b} \mid \tau_b, \phi) = \tau_b^{-1}\big((1 - \phi)\,\mathbf{I} + \phi\mathbf{Q}^-\big) \tag{5}$$

Equation (4) is a reparameterisation of the original BYM model, where $\tau_u^{-1} = \tau_b^{-1}\phi$ and $\tau_v^{-1} = \tau_b^{-1}(1 - \phi)$.[9] The additive decomposition of the variance is then on the log relative risk scale. This is in contrast to the Leroux model (3), where the precision matrix of $\mathbf{b}$ resulted as a weighted average of the precision matrices of the unstructured and structured spatial components. As a consequence, the additive decomposition of variance in the Leroux model happens on the log relative risk scale, conditional on $b_j, j \in \delta i$.[4]

3.2   Why do we need scaling?

One issue inherent to all the models described in Section 3.1 is that the spatially structured component is not scaled. However, as discussed in Sørbye and Rue,[12] scaling is crucial to facilitate hyperprior assignment and guarantee that hyperpriors used in one application have the same interpretation in another application.

The Besag model is an intrinsic GMRF and penalises local deviation from its null space, which is a constant level in the case of one connected component.[19] The hyperprior will control this local deviation and, as such, influence the smoothness of the estimated spatial effects. If the estimated field is too smooth, the precision is large and potential spatial variation might be blurred. On the other hand, if the precision is too small the model might overfit due to large local variability.[12]

In the following, we consider only the structured effect $\mathbf{b}$ as introduced in the Besag model in Section 3.1.1. In general, the marginal variances $\tau_b^{-1}[\mathbf{Q}^-]_{ii}$ depend on the graph structure reflected by the structure matrix $\mathbf{Q}$.[12] This can be illustrated by calculating a generalised variance, computed as the geometric mean (or some other typical value) of the marginal variances

$$\sigma_{\mathrm{GV}}^2(\mathbf{u}) = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{\tau_b}[\mathbf{Q}^-]_{ii}\right)\right) = \frac{1}{\tau_b}\exp\left(\frac{1}{n}\sum_{i=1}^{n}\log\left([\mathbf{Q}^-]_{ii}\right)\right) \qquad (6)$$

The generalised variance for two different graph structures is typically not equal, even when the graphs have the same number of regions.

As an example, Figure 1 shows two administrative regions of Germany, Mittelfranken and Arnsberg, that are both fully connected. Assume that we fit models to these two regions separately, including a spatially structured effect in the models. As both regions have 12 districts, we may be tempted to use identical priors for the precision τ_b in these two cases. However, the prior will penalise the local deviation from a constant level, differently. Specifically, if $\tau_b = 1$ in equation (6), it follows that $\sigma_{\mathrm{GV}}^2(\mathbf{b}) \approx 0.30$ for Mittelfranken and $\sigma_{\mathrm{GV}}^2(\mathbf{b}) \approx 0.40$ for Arnsberg. If we fit a similar model to all the 544 districts in Germany, we obtain $\sigma_{\mathrm{GV}}^2(\mathbf{b}) \approx 0.56$. These differences, reflecting a characteristic level for the marginal variances, might be even more striking for other applications. In order to unify the interpretation of a chosen prior for τ_b and make it transferable between applications, the structured effect needs to be scaled so that $\sigma_{\mathrm{GV}}^2(\mathbf{b}) = 1/\tau_b$. This implies that τ_b represents the precision of the (marginal) deviation from a constant level, independently of the underlying graph.

This issue of scaling applies to all intrinsic GMRFs.[12] In practice, the scaling might be costly, as it involves inversion of a singular matrix. The generalised inverse can be used, but the relative tolerance threshold to detect zero singular values will be crucial. Knowing the rank deficiency r of the matrix, the r smallest eigenvalues could be removed. Intrinsic GMRFs have much in common with GMRFs conditional on linear constraints, see Rue and Held[19] (Section 3.2). Hence, if we know the null space $\mathbf{V}$ of $\mathbf{Q}$, we can use sparse matrix algebra to compute $[\mathbf{Q}^-]_{ii}$. The marginal variances can be computed based on $\mathbf{b}\,|\,\mathbf{V}^\top\mathbf{b} = \mathbf{0}$. First, $\mathbf{Q}$ is made regular by adding a small term ϵ to the diagonal and then $\mathrm{Diag}\big(\mathrm{Var}(\mathbf{b}\,|\,\mathbf{V}^\top\mathbf{b} = \mathbf{0})\big)$ is computed according to Rue and Held[19] (Section 3.2, equation (3.17)). In this way, we take advantage of the sparse matrix structure of $\mathbf{Q}$. Extracting the variance components, we compute σ_{GV}^2 and use this as a factor to scale $\mathbf{Q}$. The R-package INLA, available from www.r-inla.org, offers a function inla.scale.model which takes as argument any singular precision matrix $\mathbf{Q}$ and the linear constraints spanning the null-space of $\mathbf{Q}$. The scaled
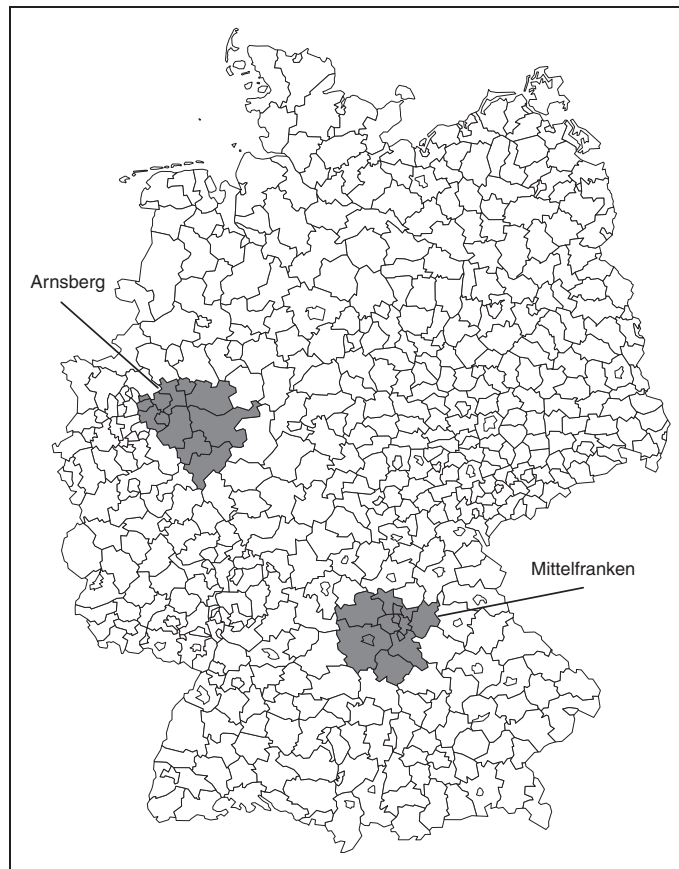
Figure 1. Map of Germany separated in 544 districts. Marked in grey are the administrative districts Arnsberg in the federal state Nordrhein-Westfalen and Mittelfranken in Bayern. Both areas consist of 12 districts.

precision matrix, where the geometric mean of the marginal variances is equal to one, is returned. For the Besag model on a connected graph we have $\mathbf{V} = \mathbf{1}$, so that we can scale the matrix in R using

```
R > Q = inla.scale.model(Q,
    constr = list(A = matrix(1, nrow = 1, ncol = n), e = 0))
```

We would like to note that scaling is also crucial if the goal is hypothesis testing involving an intrinsic GMRF, as for example in Dean et al.,[6] as otherwise the critical region would be naturally dependent on the scale of the random effect. This also applies outside a Bayesian setting.

3.3 A modified BYM model accounting for scaling

Simpson et al.[13] propose a modification of the Dean model which addresses both the identifiability and scaling issue of the BYM model. The model uses a scaled structured component $\mathbf{u}_\star$, where $\mathbf{Q}_\star$ is

the precision matrix of the Besag model, scaled according to Section 3.2. This gives a modified version of the random effect

$$\mathbf{b} = \frac{1}{\sqrt{\tau_b}}\left(\sqrt{1-\phi}\,\mathbf{v} + \sqrt{\phi}\,\mathbf{u}_\star\right) \tag{7}$$

having covariance matrix

$$\mathrm{Var}(\mathbf{b}\,|\,\tau_b,\phi) = \tau_b^{-1}\left((1-\phi)\,\mathbf{I} + \phi\mathbf{Q}_\star^-\right) \tag{8}$$

Breslow et al.[21] criticised the BYM model for obscuring the conditional mean and conditional variance. Using the new parameterisation, $\mathrm{Var}(\mathbf{b}\,|\,\tau_b,\phi)$ has a clear and intuitive interpretation, and interpretations in terms of the conditional distribution of b_i given b_j, $j \in \delta i$, are avoided.

Similarly to the Leroux and the Dean model, equation (7) emphasises the compromise between pure overdispersion and spatially structured correlation, where $0 \le \phi \le 1$ measures the proportion of the marginal variance explained by the structured effect. The model reduces to pure overdispersion for $\phi = 0$ and to the Besag model, i.e. only spatially structured risk, when $\phi = 1$. More importantly, using the standardised $\mathbf{Q}_\star^-$ the marginal variances will be approximately equal to $(1 - \phi)/\tau_b + \phi/\tau_b$. Approximation holds since $\mathbf{Q}^-$ by construction does not give the same marginal variances for all regions, so that the standardisation only holds on the overall level of the generalised variance σ_{GV}^2. With the scaled structured effect $\mathbf{u}_\star$, the random effect $\mathbf{b}$ is also scaled and the prior imposed on τ_b has the same interpretation, being transferred between graphs. Furthermore, the hyperparameters τ_b and ϕ are now interpretable and no longer confounded. It should be noted that in contrast to the Dean model, the Leroux model cannot be scaled since by construction the scaling would depend on the value of ϕ.

3.4 Parameterisation preserving sparsity

A challenge of the new model formulation in equations (7) and (8) is that the covariance matrix involves the generalised inverse $\mathbf{Q}_\star^-$. The precision matrix, defined as the inverse of equation (8), is then no longer sparse. Sparsity properties are important to gain computational efficiency, both implementing the model in the INLA framework,[14] or using block updating in Markov chain Monte Carlo methods. To keep the sparsity of the precision matrix, we propose an augmented parameterisation.

Recall that $\mathbf{u}_\star \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_\star^-)$, where $\mathbf{Q}_\star$ is scaled, and $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ in equation (7). Let $\mathbf{w} = \left(\mathbf{w}_1^\top \;\; \mathbf{w}_2^\top\right)^\top$, where $\mathbf{w}_2 = \mathbf{u}_\star$ and $\mathbf{w}_1 = \mathbf{b}$. Thus

$$\mathbf{w}_1|\mathbf{w}_2 \sim \mathcal{N}\left(\sqrt{\frac{\phi}{\tau_b}}\mathbf{w}_2, \frac{1-\phi}{\tau_b}\mathbf{I}\right)$$

As $\pi(\mathbf{w}) = \pi(\mathbf{w}_1|\mathbf{w}_2)\pi(\mathbf{w}_2)$, it follows that $\mathbf{w}$ is normally distributed with mean $\mathbf{0}$ and precision matrix

$$\begin{pmatrix} \dfrac{\tau_b}{1-\phi}\mathbf{I} & -\dfrac{\sqrt{\phi}\tau_b}{1-\phi}\mathbf{I} \\[2ex] -\dfrac{\sqrt{\phi}\tau_b}{1-\phi}\mathbf{I} & \mathbf{Q}_\star + \dfrac{\phi}{1-\phi}\mathbf{I} \end{pmatrix}$$

The marginal of $\mathbf{w}_1$ then has the correct distribution. Working with this parameterisation, sparsity is warranted and the structured component is given directly by $\mathbf{w}_2$, i.e. the second half of the vector $\mathbf{w}$.

4 Hyperprior choice: Penalised-complexity priors

In applying the original BYM model formulation, priors are usually specified separately for τ_u and τ_v and the prior choices are frequently rather ad hoc.[11] A reasonable alternative is to express prior beliefs in terms of the total variability of the model,[8] which also facilitates clear interpretation of the hyperparameters.

Here, we follow Simpson et al.[13] and apply the framework of PC priors, which allows us to distribute the total variance to the components of the modified BYM model. The construction of PC priors is based on four principles: (1) *Occam's razor* – simpler models should be preferred until there is enough support for a more complex model. (2) *Measure of complexity* – The Kullback-Leibler distance (KLD) is used to measure increased complexity. (3) *Constant-rate penalisation* – The deviation from the simpler model is penalised using a constant decay rate. (4) *User-defined scaling* – The user has an idea of a sensible size of the parameter (or any transformed variation) of interest. For ease of readability, we provide the main ideas and results using the PC framework, and refer to Simpson et al.[13] Section 6, Appendices A.2 and A.3 for further mathematical details.

Hyperpriors are formulated in two levels. First, a prior for τ_b is defined, which will control the marginal variance contribution of the weighted sum of $\mathbf{v}$ and $\mathbf{u}_\star$. Second, the marginal variance is distributed to the components $\mathbf{v}$ and $\mathbf{u}_\star$ by defining a suitable prior for the mixing parameter ϕ. Note that testing for the risk decomposition, as proposed in MacNab,[9] is implicated using the PC-prior framework at this level. The PC prior will automatically shrink towards $\phi = 0$, which refers to no spatial smoothing. The modified BYM model is then seen as a flexible extension of two dependent base models, which it will shrink to if not differently supported by the data.

4.1 Controlling the spatial variance contribution

The simplest model, referred to as a first-level base model, is to have a constant relative risk over all areas, i.e. no spatial variation or equivalently infinite precision. All n regions have the same disease burden. Following the principles of the PC prior framework, we penalise model complexity in terms of a measure of information-theoretic deviation from the flexible model to the base model, which is a function of the Kullback–Leibler divergence

$$\text{KLD}(\tau_b) = \text{KLD}\big(\mathcal{N}(0, \Sigma_{\text{flex}}(\tau_b))|\mathcal{N}(0, \Sigma_{\text{base}})\big) \tag{9}$$

KLD measures the information lost using the base model to approximate the more flexible model. To facilitate interpretation, this divergence is transformed to a unidirectional measure of distance defined by

$$d(\tau_b) = \sqrt{2\text{KLD}(\tau_b)}$$

In our case, $\Sigma_{\text{flex}}(\tau_b) = \text{Var}(\mathbf{b} \,|\, \tau_b, \phi)$ as defined in equation (8), where ϕ is arbitrary and treated as fixed, while the covariance matrix of the base model is $\Sigma_{\text{base}} = \mathbf{0}$, reflecting infinite precision. Details regarding the computation of equation (9) are presented in Simpson et al.[13] and for completeness provided in Appendix 2. On the distance scale the meaning of a prior is much more clear. The prior should have mode at zero, i.e. at the base model, and decreasing density with increasing distance. There are different priors that fulfil these properties, but differ in the way they decrease with increasing distance. Since we are not in a variable selection setting, we follow the recommendation of Simpson et al.[13] and use the exponential distribution, which supports

constant-rate penalisation. Applying the ordinary change of variable transformation, a type-2 Gumbel distribution[13]

$$\pi(\tau_b) = \frac{\theta}{2}\tau_b^{-3/2}\exp\left(-\theta\,\tau_b^{-1/2}\right)$$

is obtained as prior for τ_b.

The given prior corresponds to an exponential prior with rate θ for the standard deviation $1/\sqrt{\tau_b}$. Hence, to infer θ we can use the probability statement $\text{Prob}\left((1/\sqrt{\tau_b}) > U\right) = \alpha$, which gives $\theta = -\log(\alpha)/U$. It is not obvious how to interpret $1/\sqrt{\tau_b}$ epidemiologically. However, in the BYM2 model, τ_b represents the marginal precision, and as such, can be related to the total residual relative risk for which an intuitive interpretation is available.[8] For example, an assumed probability of 0.99 of having residual relative risks smaller than 2 corresponds approximately to $\text{Prob}\left((1/\sqrt{\tau_b}) > 1\right) = 0.01$. Figure 2 (left panel) shows the resulting prior compared to two gamma distributions, which will be used in the simulation study in Section 5. Please note that the use of gamma priors is questionable as the base model, i.e. distance equal to zero, has no density mass and can therefore never be obtained. We refer to Simpson et al.[13] for further discussions about overfitting of priors.

4.2 Distributing the spatial variance contribution

Assume now that the relative risk has spatial variation. This implies that the disease burden over the regions is not constant. Conditioned on a fixed marginal variance, the second-level base model is defined as having only unstructured spatial noise and no spatial dependency, i.e. $\phi = 0$.
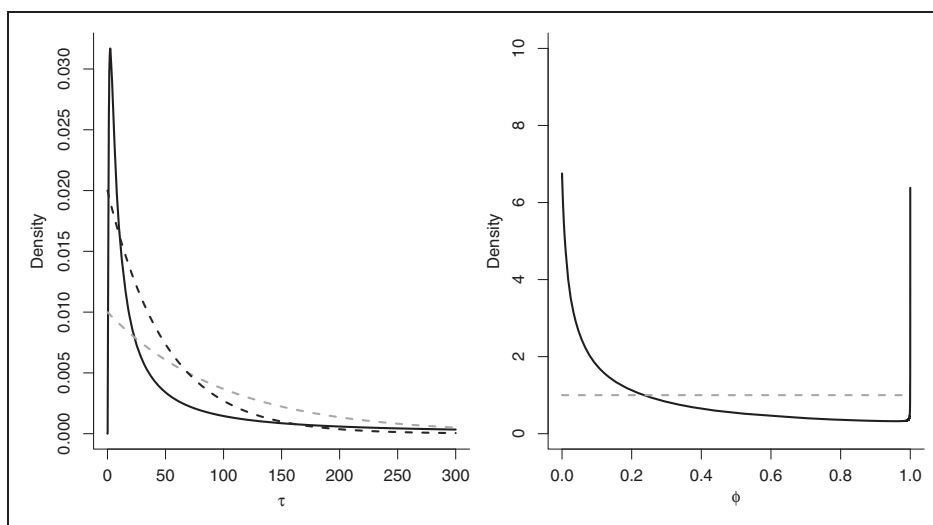

Figure 2. Left panel: PC prior for τ_b (solid) with parameters $U = 1$ and $\alpha = 0.01$, and two gamma priors where for both the shape parameter is equal to 1 and the rate parameter is either 0.01 (dashed, light grey) or 0.02 (dashed, dark grey). Right panel: PC prior for the mixing parameter ϕ with $U = 0.5$ and $\alpha = 2/3$ (solid), and a uniform prior on (0,1) (dashed).

By increasing the value of ϕ, spatial dependency is gradually blended into the model and the model component explaining most of the variance shifts from $\mathbf{v}$ to $\mathbf{u}_\star$.[13]

To derive the PC prior for ϕ, we first compute the KLD of the base model from the flexible model, see Appendix A2.2. The covariance matrices for the base and flexible models are $\Sigma_{\text{base}} = \mathbf{I}$ and $\Sigma_{\text{flex}}(\phi) = (1 - \phi)\mathbf{I} + \phi\mathbf{Q}_\star^-$, respectively. Analogously as for τ_b, we assign an exponential prior with parameter λ to the distance scale $d(\phi) = \sqrt{2\text{KLD}(\phi)}$. For simplicity, this prior is transformed to give a prior on $\text{logit}(\phi)$, as ϕ is defined on a bounded interval. Reasonable values of λ can be inferred in a similar way as in Section 4.1, using a probability statement $\text{Prob}(\phi < U) = \alpha$. In this way it is straightforward to specify both strong or weak prior information. The parameter ϕ has a more direct interpretation than the precision parameter τ_b, representing a fraction of the total variance which can be attributed to spatial dependency structure. A reasonable formulation might be $\text{Prob}(\phi < 0.5) = 2/3$, which gives more density mass to values of ϕ smaller than 0.5. This can be seen as a conservative choice, assuming that the unstructured random effect accounts for more of the variability than the spatially structured effect.

In contrast to the precision parameter τ_b, the resulting PC prior for ϕ is not available in closed form. Within INLA, the function `INLA:::inla.pc.bym.phi` can be used to compute the prior for ϕ (on logit-scale) for a specific Besag neighbourhood matrix $\mathbf{Q}$. The scaling of the graph is implemented by default. Thus, the prior can be obtained in tabulated form and also incorporated in other software packages than INLA. Figure 2 (right panel) shows the prior for ϕ with $U = 0.5$ and $\alpha = 2/3$, generated in R using

```
R > log.prior = INLA:::inla.pc.bym.phi(Q = Q, rankdef = 1,
R > u = 0.5, alpha = 2/3)
R > phis = 1/(1 + exp(-seq(-10, 10, len = 10000)))
R > plot(phis, exp(log.prior(phis)), type = ''l'',
R > lwd = 2, bty = ''l'', ylim = c(0,10),
R > xlab = expression(phi), ylab = ''Density'')
```

For comparison, the figure also displays the commonly used uniform prior for ϕ.

5 Applications

5.1 Does the modified BYM model shrink to the respective base models?

Various types of cancer data, observed in the 544 districts of Germany (Figure 1), have been widely analysed in the literature, both to study the epidemiology[22] and to test methodological developments.[23] Simpson et al.[13] re-analyse a larynx cancer dataset, assigning almost the same PC priors as shown in Figure 2 to the hyperparameters. The modified BYM model (7), here referred to as the BYM2 model, is clearly seen to learn from the data, leading to a posterior marginal concentrated around 1 for ϕ.

5.1.1 Results

Here, we perform a simulation study to investigate whether the BYM2 model shrinks towards the two different base models described in Section 4. Furthermore, we assess the behaviour under a perfectly structured risk surface. The simulations are based on the neighbourhood structure of 366 districts of Sardinia, for which a real case study will be analysed in Section 5.2. Two hundred datasets were simulated under nine different scenarios, using three different types of risk surfaces and three different disease prevalences. The model we simulate from assumes that the log relative

risk is given as $\log(\theta_i) = \mu + \frac{1}{\sqrt{\tau_b}} b_i$, $i = 1, \ldots, 366$, where $\mathbf{b}$ is a zero-mean normal distributed random effect controlled by the precision parameter τ_b. In the simulations, we assume either a constant risk over all regions, i.e. $1/\sqrt{\tau_b} = 0$, an independent area-specific risk setting with $1/\sqrt{\tau_b} = 0.5$ and $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, or a spatially structured area-specific risk setting with $1/\sqrt{\tau_b} = 0.5$ and $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_\star^-)$. We alter the expected number of cases E_i to analyse whether the model performance changes when applied to diseases with different underlying disease prevalence. We set all E_i equal to either 15, 60 or 200 and analyse all datasets using the INLA package in R.

Table 1 shows the posterior mean estimates for the intercept and the two hyperparameters, also including the average standard deviations of the estimates in parentheses.

The constant risk surface scenario corresponds to the first-level base model described in Section 4. The resulting estimates are reasonable for all parameters. Specifically, the standard deviation $1/\sqrt{\tau_b}$ is close to zero and decreases with higher prevalence. As these datasets give no information about the mixing parameter, the prior will dominate the posterior estimates of ϕ. The assumption of independent, area-specific spatial variation of the risk surface corresponds to the second-level base model in Section 4. Again, the posterior mean parameter estimates are seen to be close to the true values. In particular, the model is seen to shrink towards the base model as the estimated mixing parameters are very close to zero, distributing all of the spatial variation to the unstructured component. In the case of a spatially structured surface risk, the intercept and $1/\sqrt{\tau_b}$ are well estimated. The mixing parameter ϕ is estimated slightly lower when $E_i = 15$ compared to $E_i = 60$ and $E_i = 200$, where the estimates are closer to 1. This illustrates the learning ability of the model to support a more complex model if indicated by the data.

We repeated the simulation study using a uniform prior for the mixing parameter ϕ. The obtained results are comparable, see online supplementary material. Differences occur in the case of a simulated constant risk surface where the estimates of the mixing parameter are now close to 0.5, as the estimate is dominated by the new prior. In the case of a spatially unstructured or structured

Table 1. Mean values of the intercept and hyperparameters, for the BYM2 model for varying disease prevalence, using either a constant, spatially varying unstructured risk surfaces or spatially varying structured risk surfaces. PC priors are used for the weighting parameter ϕ and for the precision parameter τ_b. Standard deviations are provided in parentheses. Results are based on 200 simulations for each setting, where the true parameter values are provided in the table.

	μ	$1/\sqrt{\tau}$	ϕ
Constant risk $\left(1/\sqrt{\tau_b} = 0, \phi = 0\right)$:			
$E = 15$	$-3.78E{-}04$ (0.01)	0.02 (0.02)	0.26 (0.05)
$E = 60$	$-9.15E{-}04$ (0.01)	0.01 (0.01)	0.26 (0.05)
$E = 200$	$-4.36E{-}04$ (0.004)	0.01 (0.01)	0.26 (0.04)
Area-specific independent risk $\left(1/\sqrt{\tau_b} = 0.5, \phi = 0\right)$:			
$E = 15$	$1.62E{-}03$ (0.03)	0.50 (0.02)	0.05 (0.03)
$E = 60$	$-1.94E{-}03$ (0.03)	0.51 (0.03)	0.04 (0.02)
$E = 200$	$1.21E{-}03$ (0.03)	0.50 (0.02)	0.04 (0.02)
Area-specific correlated risk $\left(1/\sqrt{\tau_b} = 0.5, \phi = 1\right)$:			
$E = 15$	$2.14E{-}03$ (0.01)	0.47 (0.04)	0.90 (0.06)
$E = 60$	$9.11E{-}06$ (0.01)	0.49 (0.05)	0.93 (0.04)
$E = 200$	$4.70E{-}05$ (0.004)	0.49 (0.07)	0.93 (0.04)

risk, the obtained parameter estimates are seen to be robust to the prior choice for the mixing parameter.

5.1.2 Comparison to other spatial latent models

The main benefit of the BYM2 model is that it allows for an intuitive parameter interpretation and facilitates prior assignment. Performance in terms of model choice criteria is regarded secondary and will be assessed in this section. To compare the BYM2 model with existing approaches, we fit the spatial latent models presented in Section 3.1 to the simulated data. We also include a model having only an unstructured spatial component, referred to as the iid model. For the BYM2 model, we use either a PC prior with $U = 0.5$ and $\alpha = 2/3$, denoted as BYM2 (PC), or a uniform prior, denoted as BYM2 (unif), for the mixing parameter.

Table 2 illustrates the resulting posterior parameter estimates, assuming a constant risk surface and $E = 60$. The table also reports three model choice criteria for this scenario: The root mean squared error, computed as the average of $\sqrt{\frac{1}{366} \sum_{i=1}^{366} \left(y_i/E_i - \hat{\theta}_i \right)^2}$ over all 200 simulations with $\hat{\theta}_i$ denoting the posterior mean fitted risk, the average DIC[24] and the average proper logarithmic scoring rule.[25] For all three criteria, lower values imply better model properties. Note that the parameter estimates of the BYM model are not included here as it is not parameterised in terms of one precision and one mixing parameter. However, we will consider it in terms of model choice, see figures in online supplementary material.

The estimated parameter values in Table 2 have to be interpreted with great care. Even though the different models look comparable and the parameters have the same name, their interpretations are not equivalent. A first issue is that we use different priors for the hyperparameters. Specifically, we assume that $\phi \sim \text{Unif}(0, 1)$ for the Leroux and Dean model, as this represents a popular non-informative prior choice in the literature.[8,9,26] Further, we assume a $\Gamma(\text{shape} = 1, \text{rate} = 0.01)$ distribution for the precision parameter in the iid model. For the spatially structured term, we took the graph structure of Sardinia into account and use a $\Gamma(\text{shape} = 1, \text{rate} = 0.02)$[11] prior. However, also the different parameterisation of the Leroux and Dean model implies that the parameters are not comparable. It is only when $\phi = 0$ or $\phi = 1$, these models reduce to the same model (assuming the same prior for the precision parameter).

Table 2. Mean values of the intercept and hyperparameters, root mean squared error (RMSE) for the fitted risk, DIC and logarithmic score (LS) when using different spatial latent models. Standard deviations are provided in parentheses. Results are based on 200 simulations where the true values are $\mu = 0$, $1/\sqrt{\tau} = 0$, $\phi = 0$, $E = 60$. Be aware that parameter estimates might not be directly comparable between models.

	μ	$1/\sqrt{\tau}$	ϕ	RMSE	DIC	LS
iid model	−1.65E−03 (0.01)	0.05 (0.004)	0.00 (−)	0.11	2545	3.47
Besag	−1.26E−03 (0.01)	0.07 (0.004)	1.00 (−)	0.12	2545	3.48
Leroux	−1.51E−03 (0.01)	0.07 (0.01)	0.56 (0.08)	0.11	2545	3.48
Dean	−1.69E−03 (0.01)	0.06 (0.004)	0.60 (0.12)	0.11	2547	3.48
BYM2 (unif)	−8.39E−04 (0.01)	0.01 (0.01)	0.46 (0.05)	0.12	2540	3.47
BYM2 (PC)	−9.15E−04 (0.01)	0.01 (0.01)	0.26 (0.05)	0.12	2540	3.47

The second issue is that none of the models are properly scaled, except the BYM2 model. Specifically, the precision parameter in the Leroux and Dean model does not represent the marginal precision. It is still confounded with ϕ and hence not comparable to the BYM2 model. The only parameter estimates that are comparable are those from the iid model and the BYM2 model. The Besag model can be scaled, and so can the Dean model leading to the parameterisation of the BYM2 model. However, the Leroux model cannot be scaled as any scaling would depend on the mixing parameter ϕ. Hence, the precision parameter cannot be interpreted as marginal precision as it depends on the underlying graph structure and is in addition confounded with the mixing parameter.

Even though the parameters of the models have different interpretations and priors are chosen differently, the model choice criteria give quite similar average values, for all the models and only the DIC values seem to slightly favour the two BYM2 models. This means that the BYM2 performs at least as well as the other models and changing the prior for the mixing parameters leads to the same results. This is also confirmed inspecting the logarithmic scores and DIC values over all simulations, in terms of boxplots, see figures in online supplementary material, respectively. Results for these two model choice criteria coincide well and indicate that the Besag model performs worse than the other models when simulating an independent area-specific risk surface, where the iid model is naturally but only marginally favoured. In contrast, the iid model performs worse when simulating a spatially structured surface, where the Besag model is marginally favoured for low values of E_i. The other models perform almost identically in both of these simulation settings. Simulating a constant-risk surface, all models behave similarly for $E_i = 15$. For increasing numbers of expected cases, the BYM2 model seems to be slightly beneficial. The figures only include the boxplots obtained by the BYM2 model using a PC prior for ϕ, as results using a uniform prior are visually identical.

5.2 Analysis of IDDM in Sardinia

We re-analyse IDDM counts in Sardinia, see Bernardinelli et al.[17] and Knorr-Held and Rue[23] for previous analyses. This dataset has low counts. Only 619 cases were registered for $n = 366$ districts in the period 1989–1992, for the population aged 0–29 years and there are no covariates. Spatial smoothing is essential to reduce the large variance in the estimates and to get a realistic picture of the underlying risk surface. Held[27] has demonstrated that a constant risk is well supported by the posterior distribution, and based on this we would expect shrinkage towards the first-level base model.

We use $U = 0.5$ and $\alpha = 2/3$ for the prior of ϕ. For τ_b we use $U = 0.2/0.31$ and $\alpha = 0.1$ which corresponds to a marginal standard deviation of 0.2. The R-code to fit the BYM2 model using the INLA-package is explained in Appendix 1. The posterior estimates are shown in Table 3 (first quarter). We see that the precision estimate is rather high supporting a constant log relative risk as a zero log relative risk is well within all the posterior credible intervals for $\{\eta_i\}$. The posterior marginal of ϕ is mostly uniform giving essentially no information on which component is describing the variance. Figure 3 shows the posterior mean for the relative risk over all 366 regions.

MacNab et al.[28] noted considerable sensitivity regarding the prior choice for ϕ in the Leroux model. For the Sardinia data, we empirically check sensitivity by changing the parameters of the PC prior for ϕ to $U = 0.5$ and $\alpha = 0.5$, $U = 0.5$ and $\alpha = 0.1$. That means a priori we expect that either 50 or 90% of the variation is explained by the spatially structured component, respectively. As a fourth alternative we consider a uniform prior for ϕ. The rest of Table 3 shows the obtained posterior quantities. The estimates for the intercept stay unchanged. We see a slight change in the posterior estimates for τ_b and ϕ, whereby the posterior marginal of ϕ still supports almost the whole range for all priors. The posterior precision slightly decreases, whereby the 2.5% quantile and the median stay

Table 3. Posterior quantities for the intercept and hyperparameters under three different PC priors or a uniform prior for the mixing parameter ϕ for IDDM counts in Sardinia.

	Mean	SD	2.5% quant.	Median	97.5 quant.	Mode
	Prior for ϕ: $U = 0.5$, $\alpha = 2/3$					
μ	0.01	0.05	−0.09	0.01	0.10	0.01
τ_b	262.79	751.60	10.14	45.36	2549.91	17.40
ϕ	0.48	0.31	0.01	0.48	0.98	1.00
	Prior for ϕ: $U = 0.5$, $\alpha = 0.5$					
μ	0.01	0.05	−0.09	0.01	0.10	0.01
τ_b	233.51	686.28	10.00	42.78	2241.48	17.37
ϕ	0.54	0.31	0.02	0.57	0.99	1.00
	Prior for ϕ: $U = 0.5$, $\alpha = 0.1$					
μ	0.01	0.05	−0.08	0.01	0.11	0.01
τ_b	201.56	616.57	9.83	40.11	1880.89	17.27
ϕ	0.63	0.29	0.03	0.69	1.00	1.00
	Prior for ϕ: Unif (0,1)					
μ	0.01	0.05	−0.08	0.01	0.11	0.01
τ_b	189.46	591.67	9.77	39.07	1734.34	17.21
ϕ	0.64	0.26	0.08	0.69	0.99	1.00

fairly constant. The upper quantile and therefore the mean change more clearly. As in MacNab et al.,[28] the effect on the posterior risk estimates is negligible (Figure 4).

The table in the online supplementary material shows model choice criteria when using different models for analysing IDDM counts in Sardinia. Furthermore, results using the four different priors for the mixing parameter in the BYM2 model are shown. As in the simulation study there are no clear differences between the models.

6 Discussion

Throughout this paper we have stressed the importance of scaling in Bayesian disease mapping. Without proper scaling, hyperparameters have no clear meaning and may be falsely interpreted. To be more specific, the precision parameter of the commonly used Leroux and Dean models tends to be falsely interpreted as marginal precision controlling the deviation from a constant disease risk over space. However, the parameter depends on the underlying graph structure and is confounded with the mixing parameter if the structured spatial component is not appropriately scaled. In addition, it is not clear how to choose a prior distribution for this parameter. First, due to the lack of scaling, a fixed hyperprior for the precision parameter gives different amount of smoothing if the graph on which given disease counts are observed is changed. Second, commonly used hyperprior distributions induce overfitting[13] and will not allow to reduce to simpler models such as a constant risk surface or uncorrelated noise over space.

Simpson et al.[13] proposed a new modification of the commonly known BYM model, termed BYM2 model, which addresses the aforementioned issues, and applied it to one case study. The BYM2 model consists of one precision parameter and one mixing parameter. The precision parameter represents the marginal precision and controls the variability explained by a spatial effect. The mixing parameter distributes existing variability between an unstructured and
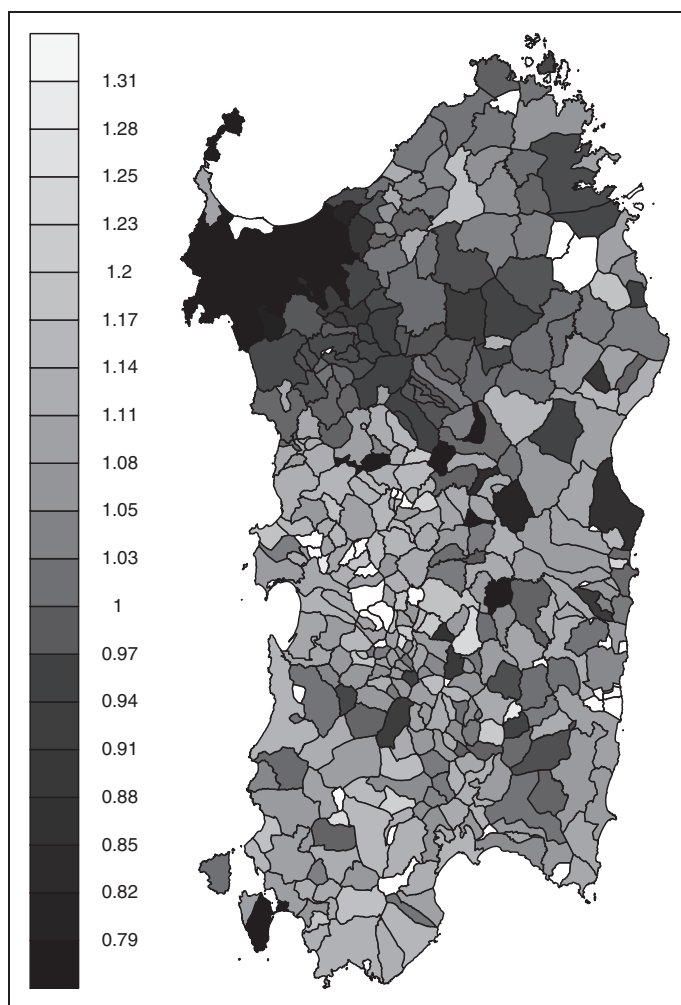
Figure 3. Posterior mean of the relative risk for IDDM counts in Sardinia using the BYM2 model.

structured component. Importantly, the structured component is scaled,[12] which facilitates prior assignment. PC priors, i.e. principle-based priors, are used to favour simpler models until a more complex model is supported. Furthermore, these priors allow epidemiologically intuitive specification of hyperparameters based on the relative risks.

In this paper we have systematically assessed the behaviour of the BYM2 model formulation. By means of a simulation study based on the neighbourhood structure of 366 regions in Sardinia, we have shown that the model is able to shrink towards both a constant risk surface or a spatially unstructured risk for different disease prevalences. This shows that the model does not overfit. When the disease risk is spatially structured, the model shows good learning abilities. In contrast to other model formulations, the posterior estimates of all hyperparameters can be directly and intuitively interpreted. In terms of model comparison, we have found that the BYM2 model is slightly
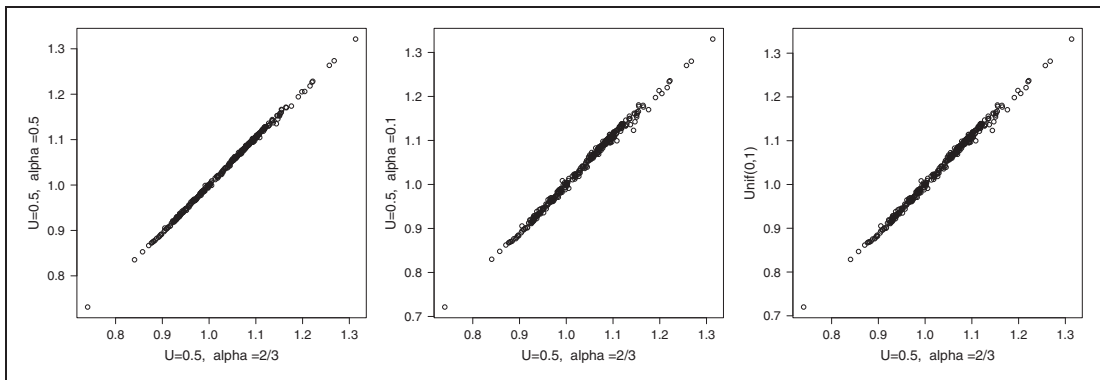
Figure 4. Comparison of the posterior mean of the relative risk under four different prior distributions for the mixing parameter ϕ for IDDM counts in Sardinia.

favoured compared to the Leroux and Dean model in terms of DIC and logarithmic score in the case of a constant risk. In the case of an unstructured risk, the used model choice characteristics are almost indistinguishable. We think that the practical benefits in terms of interpretability and prior assignment make the BYM2 model advantageous compared to existing models and we recommend its usage also since its model criteria performance is at least as good as for existing methods.

All models were implemented using INLA,[14] which provides efficient Bayesian inference without the need of MCMC sampling. This facilitated the simulation study considerably and allowed us to investigate different simulation settings. The user-friendly implementation is illustrated by the R-code in Appendix 1.

MacNab et al.[28] noted sensitivity in choosing a prior distribution for the mixing parameter. Here, we investigated this issue empirically in both the simulation study and an application to IDDM in Sardinia. However, more effort needs to be placed into a detailed prior sensitivity analysis following the theoretical framework of Roos et al.[29]

The BYM2 model can be naturally combined with covariate information, see Simpson et al.[13] for an application or integrated into a space–time context. However, it will require further work to distribute the variance not only within the spatial components but over all model parameters in the linear predictor. It should be noted that the BYM2 model is not only interesting within disease mapping but also within other fields, such as genetics, where genes can be regarded as regions and the neighbourhood structure represents which genes are linked.

Declaration of conflicting interests

Funding

References

1. Lawson AB. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. 2nd Edition, Boca Raton, FL: Chapman and Hall/CRC press, 2013.

2. Lee D. A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial Spatio-Temporal Epidemiol* 2011; **2**: 79–89.

3. Besag J, York J and Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* 1991; **43**: 1–20.

4. Leroux BG, Lei X and Breslow N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: Halloran ME and Berry D (eds) *Statistical models in epidemiology, the environment, and clinical trials*. New York: Springer, 2000, pp.179–191.

5. Stern HS and Cressie N. Posterior predictive model checks for disease mapping models. *Stat Med* 2000; **19**: 2377–2397.

6. Dean CB, Ugarte MD and Militino AF. Detecting interaction between random region and fixed age effects in disease mapping. *Biometrics* 2001; **57**: 197–202.

7. Best N, Richardson S and Thomson A. A comparison of Bayesian spatial models for disease mapping. *Stat Methods Med Res* 2005; **14**: 35–59.

8. Wakefield J. Disease mapping and spatial regression with count data. *Biostatistics* 2007; **8**: 158–183.

9. MacNab YC. On Gaussian Markov random fields and Bayesian disease mapping. *Stat Methods Med Res* 2011; **20**: 49–68.

10. Lee DJ and Durbán M. Smooth-CAR mixed models for spatial count data. *Comput Stat Data Anal* 2009; **53**: 2968–2979.

11. Bernardinelli L, Clayton D and Montomoli C. Bayesian estimates of disease maps: How important are priors? *Stat Med* 1995; **14**: 2411–2431.

12. Sørbye SH and Rue H. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Stat* 2014; **8**: 39–51.

13. Simpson DP, Rue H, Martins TG, et al. Penalising model component complexity: a principled, practical approach to constructing priors. arXiv preprint arXiv:1403.4630v4, 2015, http://arxiv.org/pdf/1403.4630v4.pdf (accessed 17 July 2016).

14. Rue H, Martino S and Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J R Stat Soc Ser B (Stat Methodol)* 2009; **71**: 319–392.

15. Schrödle B and Held L. A primer on disease mapping and ecological regression using INLA. *Comput Stat* 2011; **26**: 241–258.

16. Schrödle B and Held L. Spatio-temporal disease mapping using INLA. *Environmetrics* 2011; **22**: 725–734.

17. Bernardinelli L, Pascutto C, Best N, et al. Disease mapping with errors in covariates. *Stat Med* 1997; **16**: 741–752.

18. Banerjee S, Carlin BP and Gelfand AE. *Hierarchical modeling and analysis for spatial data*. 2nd Edition, Boca Raton, FL: CRC Press, 2014.

19. Rue H and Held L. *Gaussian Markov random fields: theory and applications*. Boca Raton, FL: Chapman and Hall/CRC press, 2005.

20. Hodges JS. *Richly parameterized linear models: additive, time series, and spatial models using random effects*. Boca Raton, FL: CRC Press, 2014.

21. Breslow N, Leroux B and Platt R. Approximate hierarchical modelling of discrete data in epidemiology. *Stat Methods Med Res* 1998; **7**: 49–62.

22. Natário I and Knorr-Held L. Non-parametric ecological regression and spatial variation. *Biometr J* 2003; **45**: 670–688.

23. Knorr-Held L and Rue H. On block updating in Markov random field models for disease mapping. *Scand J Stat* 2002; **29**: 597–614.

24. Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B (Stat Methodol)* 2002; **64**: 583–639.

25. Gneiting T and Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 2007; **102**: 359–378.

26. Ugarte MD, Adin A, Goicoa T, et al. On fitting spatio-temporal disease mapping models using approximate Bayesian inference. *Stat Methods Med Res* 2014; **23**: 507–530.

27. Held L. Simultaneous posterior probability statements from Monte Carlo output. *J Comput Graph Stat* 2004; **13**: 20–35.

28. MacNab YC, Farrell PJ, Gustafson P, et al. Estimation in Bayesian disease mapping. *Biometrics* 2004; **60**: 865–873.

29. Roos M, Martins TG, Held L, et al. Sensitivity analysis for Bayesian hierarchical models. *Bayesian Anal* 2015; **10**: 321–349.

30. Erisman A and Tinney W. On computing certain elements of the inverse of a sparse matrix. *Commun ACM* 1975; **18**: 177–179.

Appendix 1. INLA R-code to implement the BYM2 model

In the following, we illustrate how the BYM2 model can be implemented within the R-package INLA.

```
 1  # load the R-package
 2  library(INLA)
 3
 4  # provide the path to the file storing the neighbourhood structure
 5  g = "sardinia.graph"
 6  # read the data
 7  sardinia = read.table("sardinia.dat", col.names=c("y", "E", "SMR"))
 8
 9  # show the first lines
10  head(sardinia)
11  #    y           E        SMR
12  #1   1   0.5986411  1.6704500
13  #2   0   0.6055964  0.0000000
14  #3  10  13.3658700  0.7481743
15  #4   0   0.2346664  0.0000000
16
17  # get the number of regions
18  n = nrow(sardinia)
19  # define a region index with values 1, 2, ..., n
20  sardinia$region = 1:n
21
22  # specify the latent structure using a formula object
23  formula = y ~ 1 + f(region,
24                      model = "bym2",
25                      graph=g,
26                      scale.model = TRUE,
27                      constr = TRUE,
28                      hyper=list(
29                          phi = list(
30                             prior = "pc",
31                             param = c(0.5, 2/3),
32                             initial = -3),
33                          prec = list(
34                             prior = "pc.prec",
35                             param = c(0.2/0.31, 0.01),
36                             initial = 5)))
37
38  # call the inla function
39  result = inla(formula, data = sardinia, family = "poisson", E=E,
40          control.predictor = list(compute=TRUE))
41  # get improved estimates for the hyperparameters
42  result = inla.hyperpar(result, dz = 0.2, diff.logdens=20)
```

First we have to load the INLA package which can be installed by typing

```
 1  install.packages("INLA", repos="http://www.math.ntnu.no/inla/R/testing")
```

in the R-terminal. In this paper, we used the INLA version 0.0-1445883880. In order to define the model incorporating a spatially structured component, we need to provide the neighbourhood structure for the regions we analyse. In INLA this is done by providing the path to the file which stores this information. The graph file for Sardinia is named sardinia.graph and has the following

structure:

```
366
0 5 13 61 69 73 81
1 5 8 16 40 46 94
2 4 47 59 63 77
3 3 11 17 43
4 5 24 41 51 56 67
...
363 5 292 307 324 344 346
364 5 291 310 317 329 345
365 3 289 303 308
```

The first line provides the number of regions, here 366. The following 366 lines correspond to each of the regions, where the first entry always specifies a unique region index. The second entry specifies the number of neighbouring regions, and then the corresponding region indices for those neighbours are listed. INLA transforms this information into the neighbourhood matrix $\mathbf{Q}$ with entries as specified in equation (2).

Lines 23–36 specify the model structure in terms of a formula object. Here, we have an intercept and a latent Gaussian model given by equation (7). The f(.) function is used to specify a random effect in INLA. The first argument is always the variable name to which it applies, here region. We specify the use of the BYM2 model in line 24. Then, we provide the graph and specify that we want to scale the model in line 26. This will scale the graph as described in Section 3.2. To avoid confounding with the intercept a sum-to-zero constraint is specified in line 27. The next lines specify the hyperprior for the two hyperparameters in form of a list of length two. The first list entry specifies the prior for the mixing parameter ϕ, named phi in the INLA implementation, the second entry for the precision parameter τ_b, named prec. Each entry is itself a list. Using the argument prior we specify the prior we would like to use. The parameters are set in the argument param. They coincide with the values used in Section 5.2. Using the argument initial a starting value for the numerical optimisation of the INLA algorithm can be provided. This has to be given on the log scale for the precision and the logit scale for ϕ, as these are the scales INLA works on internally.

After defining the model, we call the inla function providing the formula, the data and the likelihood. Further, we specify the expected number of counts and specify that we not only would like to get posterior estimates for the components of η_i but also η_i and θ_i (line 40). Improved hyperparameter estimates can be obtained by calling inla.hyperpar afterwards. Here, dz and diff.logdens refer to options in the numerical integration algorithm, see help page.

Results can be inspected using summary(result) or plot(result). Posterior marginals for the fixed effects and hyperparameters can be consequently accessed via result\$summary.fixed or result\$summary.hyperpar, respectively.

Appendix 2. PC prior derivation

A2.1 Precision parameter τ_b

In deriving the PC prior for τ_b, we start by assuming a large fixed precision τ_0 for the base model, where $\tau_0 \gg \tau_b$. The resulting $n-1$-dimensional stochastic part of the KLD is

$$\text{KLD}(\tau_b) = \frac{n-1}{2}\frac{\tau_0}{\tau_b}\left(1 + \frac{\tau_b}{\tau_0}\log\left[\frac{\tau_b}{\tau_0}\right] - \frac{\tau_b}{\tau_0}\right) \to \frac{n-1}{2}\frac{\tau_0}{\tau_b} \tag{10}$$

which implies $d(\tau_b) \to \sqrt{(n-1)\tau_0/\tau_b}$.

The next step is to assign a prior to the distance $d(\tau_b)$ and transform this by a change of variables. We use the exponential prior with decay rate λ, which supports the constant-rate penalisation and apply the ordinary change of variable transformation

$$\pi(\tau_b) = \lambda\exp\left(-\lambda d(\tau_b)\right)\left|\frac{\partial d(\tau_b)}{\partial \tau_b}\right|$$

Formulating this expression out and setting $\theta = \lambda\sqrt{(n-1)\tau_0}$, where λ is chosen so that θ is kept constant in the limit $\tau_0 \to \infty$, we obtain a type-2 Gumbel distribution[13]

$$\pi(\tau_b) = \frac{\theta}{2}\tau_b^{-3/2}\exp\left(-\theta\tau_b^{-1/2}\right)$$

as prior for τ_b.

A2.2 Mixing parameter ϕ

To derive the PC prior for ϕ, we assume that $\mathbf{Q}_\star > 0$ and define the covariance matrices for the base and flexible models as $\Sigma_{\text{base}} = \mathbf{I}$ and $\Sigma_{\text{flex}}(\phi) = (1-\phi)\mathbf{I} + \phi\mathbf{Q}_\star^{-1}$, respectively. The resulting KLD is

$$\text{KLD}(\mathcal{N}(0, \Sigma_{\text{flex}}(\phi))|\mathcal{N}(0, \Sigma_{\text{base}})) = \frac{1}{2}[\text{trace}(\Sigma_{\text{flex}}(\phi)) - n - \log|\Sigma_{\text{flex}}(\phi)|]$$

$$= \frac{1}{2}\left[n\phi\left(\frac{1}{n}\text{trace}(\mathbf{Q}_\star^{-1}) - 1\right) - \log|(1-\phi)\mathbf{I} + \phi\mathbf{Q}_\star^{-1}|\right]$$

The computation of the trace is quick when $\mathbf{Q}_\star$ is sparse.[30] The determinant can be computed as $|(1-\phi)\mathbf{I} + \phi\mathbf{Q}_\star^{-1}| = \prod_{i=1}^{n}(1 - \phi + \phi\tilde{\gamma}_i)$, where $\{\gamma_i\}$ denote the eigenvalues of $\mathbf{Q}_\star$, and $\tilde{\gamma}_i = 1/\gamma_i$.[13] As the given precision matrix $\mathbf{Q}_\star$ is singular with rank deficiency 1, one of the eigenvalues is zero. This implies that $\tilde{\gamma}_i = 1/\gamma_i$ if $\gamma_i > 0$, or else $\tilde{\gamma}_i = 0$. We can avoid calculating eigenvalues if the dimension is high, see Appendix A3 of Simpson et al.[13] for details.