Geoadditive Survival Models

Andrea Hennerfeind Andreas Brezger Ludwig Fahrmeir

Ludwig-Maximilians-Universität
Department of Statistics
Ludwigstr. 33
80539 Munich
Germany

Geoadditive Survival Models

Andrea Hennerfeind, Andreas Brezger, and Ludwig Fahrmeir*

ABSTRACT

Survival data often contain small–area geographical or spatial information, such as the residence of individuals. In many cases the impact of such spatial effects on hazard rates is of considerable substantive interest. Therefore, extensions of known survival or hazard rate models to spatial models have been suggested recently. Mostly, a spatial component is added to the usual linear predictor of the Cox model. We propose flexible continuous–time geoadditive models, extending the Cox model with respect to several aspects often needed in applications: The common linear predictor is generalized to an additive predictor, including nonparametric components for the log–baseline hazard, time–varying effects and possibly nonlinear effects of continuous covariates or further time scales, and a spatial component for geographical effects. In addition, uncorrelated frailty effects or nonlinear two–way interactions can be incorporated. Inference is developed within a unified fully Bayesian framework. We prefer to use penalized regression splines and Markov random fields as basic building blocks, but geostatistical (kriging) models are also considered. Posterior analysis uses computationally efficient MCMC sampling schemes. Smoothing parameters are an integral part of the model and are estimated automatically. Propriety of posteriors is shown under fairly general conditions, and practical performance is investigated through simulation studies. We apply our approach to data from a case study in London and Essex that aims to estimate the effect of area of residence and further covariates on waiting times to coronary artery bypass graft (CABG). Results provide clear evidence of nonlinear time–varying

*Andrea Hennerfeind (email: andreah@stat.uni-muenchen.de), Andreas Brezger (email: andib@stat.uni-muenchen.de) and Ludwig Fahrmeir (email: fahrmeir@stat.uni-muenchen.de) are Statisticians, Ludwig-Maximilians-Universität, Department of Statistics, Ludwigstr. 33, 80539 Munich, Germany. Financial support of the German Science Foundation DFG, Sonderforschungsbereich 386 "Statistische Analyse Diskreter Strukturen" is gratefully acknowledged. The authors thank Dr. H. Hemingway and A. Crook for providing the data for the CABG study, Stefan Lang for helpful discussions and computational assistance with *BayesX*, and Leonhard Held for helpful comments on the CABG data. We also thank the associate editor and the referees for their very useful and constructive comments, helping to substantially improve a first version of this article.

effects, and considerable spatial variability of waiting times to bypass graft.

Key words: Bayesian hazard rate model, Markov random field, penalized spline, MCMC, semiparametric modelling, spatial survival data.

1. INTRODUCTION

In epidemiological, economic or social science applications, survival data often contain geographical or spatial information such as the district or postal code of the residence of individuals in the study. Analyzing and modelling geographical patterns for survival or waiting times, in addition to the impact of other covariates, is of obvious interest in many studies. For example, Henderson, Shimakura and Gorst (2002) model spatial variation in survival of acute myeloid leukemia patients in northwest England, Banerjee, Wall and Carlin (2003) apply a spatial frailty model to infant mortality in Minnesota, and Li and Ryan (2002) analyze the effect of risk factors on the onset of childhood asthma with spatial data from the East Boston Asthma Study. In a study on unemployment duration in Germany, Fahrmeir, Lang, Wolff and Bender (2003) investigate the impact of small area labor market regions and other covariates, such as calendar time, age and unemployment benefits. Because unemployment duration is given in months, they apply a geoadditive discrete–time probit model. In Section 5 of this paper, we will apply our approach to data from the Appropriateness of Coronary Revascularisation (ACRE) study. Spatial survival data from this study have been recently analyzed within a discrete–time setting by Crook, Knorr–Held and Hemingway (2003).

In this paper, we propose geoadditive survival models as a flexible spatial and spatio-temporal generalization of Cox-type models. Within a unified framework, we extend the common linear predictor of the Cox model to an additive predictor, including a spatial component for geographical effects and nonparametric terms for modelling and exploring unknown functional forms of the (log-) baseline hazard rate, of nonlinear effects of continuous covariates and further time scales, such as calendar time, and of time-varying coefficients. The incorporation of such nonparametric components and their simultaneous estimation with the baseline hazard and the spatial effects is a feature which is not considered in recent other proposals for survival models with spatial components. This motivates the term ”geoadditive”, originally introduced by Kammann and Wand

(2003) in a mixed model approach to semiparametric Gaussian regression. In addition, uncorrelated frailty effects or nonlinear two–way interactions can be incorporated if appropriate.

Modelling and inference is developed from a Bayesian perspective, using information from the full likelihood rather than from a partial likelihood. A particular advantage of our approach is that all unknown functions and parameters can be treated within a unified general framework by assigning appropriate priors with the same structure but different forms and degrees of smoothness. Based on previous work (Fahrmeir and Lang, 2001; Lang and Brezger, 2004) on semiparametric regression, nonlinear effects of unknown functions of time, in particular of the log-baseline hazard rate, and of continuous covariates or further time scales are modelled through Bayesian versions of penalized splines (P-splines) introduced by Eilers and Marx (1996), Marx and Eilers (1998) for generalized additive models in a frequentist setting. Basically, time is treated in the same way as a continuous covariate, but the degree and amount of smoothness may be different. For example, simple random walk priors for the log–baseline effect in a piecewise exponential model are P–splines of degree zero. The spatial component is modelled by Gaussian Markov random field (MRF) priors, as common in disease mapping, by two–dimensional penalized tensor–product splines, or by a geostatistical (kriging) stationary Gaussian random field (GRF) model. From a computational point of view, MRF's and P–splines are clearly preferable to GRF's because their posterior precision matrices are band matrices or can be transformed into a band matrix-like structure. This special structure considerably speeds up computations and enhances numerical stability compared to the full precision matrices arising from the GRF approach.

For data observed on a irregular discrete lattice, MRF's seem to be most appropriate. If exact locations are available, P–spline or GRF surface smoothers seem to be more natural, but they can also be applied to discrete lattices after computing centroids of regions.

Our unified general framework also has computational and theoretical advantages for posterior analysis. Extending previous results for mixed models in Sun, Tsutakawa and Speckman (1999) and Speckman and Sun (2003), we can show propriety of posteriors under regularity conditions. This is important, because some of our priors are diffuse or partially improper. The Appendix on propriety in mixed models should be of general value. From the computational point of view, full conditionals of blocks of parameters have similar structure, and lead to efficient MCMC techniques. Smoothing parameters are an integral part of the model and can be estimated jointly with unknown functions and other parameters. Inferential procedures have been implemented in C++ as part of *BayesX* (Brezger, Kneib and Lang 2003).

Non- and semiparametric Bayesian survival models have become quite popular in recent years, and some previous work deals with special or related cases of our approach. For models without a spatial component Ibrahim, Chen and Sinha (2001) provide a good introduction and overview. Joint estimation of the baseline hazard and usual linear covariate effects in the Cox model has been considered by several authors. Gamerman (1991) proposes a Gaussian random walk model for the log–baseline hazard in the piecewise exponential model, and Sinha (1993) suggests a joint Gaussian smoothness prior, and Cai, Hyndman and Wand (2002) and Cai and Betensky (2003) use a mixed model representation of linear basis regression splines to estimate the baseline hazard. In all these approaches, however, effects of continuous covariates are assumed to be of the usual linear parametric form, and no spatial component is present.

Survival models with a spatial component have recently been suggested in several publications. The approaches differ in the specification of the baseline hazard rate and in the model chosen for the spatial component, but the remaining part of the predictor is still of linear parametric form. Thus, non-parametric terms for flexible modelling and estimation of the effects of continuous covariates, further time scales and time-varying coefficients are not considered in these approaches. Li and Ryan (2002) add a spatial component in form of a stationary Gaussian process to the linear predictor of the Cox model. Treating the baseline hazard as a nuisance parameter, inference for the linear predictor and for correlation function parameters is based on a marginal rank likelihood. No procedure for estimating the spatial (random) effects is provided. Henderson et al. (2002) propose a Cox model with conditionally independent gamma frailties, with means following either a geostatistical model or a Markov random field. For inference they use MCMC methods, except the baseline hazard estimate. For this they plug in the Breslow estimator at each iteration of the chain. Banerjee et al. (2003) assume a parametric Weibull baseline hazard and geostatistical or MRF priors for the spatial component. In comparison they prefer MRF priors, since computing times for geostatistical GRF models are much larger. This is in agreement with our own findings. Banerjee and Carlin (2003) develop Bayesian spatio–temporal survival models, modelling baseline hazard functions nonparametrically through a beta mixture approach and assuming MRF or CAR (conditionally autoregressive) priors for spatial effects, and Carlin and Banerjee (2003) extend this approach to multivariate MRF models, with applications to cancer survival data from Iowa.

The rest of the paper is organized as follows. In Section 2 we describe models, likelihood, and priors for unknown functions and parameters. MCMC inference is outlined in Section 3.1, and Section 3.2 provides results on the propriety of posteriors in geoadditive survival models. Performance is studied in Section 4 through a simulation study. An application to the CABG study in Section 5 illustrates the method. The concluding section contains some proposals for future research. The Appendix provides lemmas and corollaries on the propriety of posteriors in mixed models.

2. MODELS, LIKELIHOOD AND PRIORS

2.1 OBSERVATION MODEL AND LIKELIHOOD

Consider survival data in usual form, i.e., it is assumed that each individual i in the study has a lifetime T_i and a censoring time C_i that are independent random variables. The observed lifetime is then $t_i = \min(T_i, C_i)$, and δ_i denotes the censoring indicator. The data are then given by

$$(t_i, \delta_i; \boldsymbol{v}_i), \quad i = 1, \dots, n \tag{1}$$

where $\boldsymbol{v}_i$ is the vector of covariates. Covariates may also be time–dependent, but we restrict discussion to time–constant covariates for simplicity.

In Cox's proportional model the hazard rate for individual i is assumed as the product

$$\lambda_i(t; \boldsymbol{v}_i) = \lambda_0(t) \exp(\gamma_1 v_{i1} + \ldots + \gamma_r v_{ir}) = \lambda_0(t) \exp(\boldsymbol{v}_i' \boldsymbol{\gamma}). \tag{2}$$

The baseline hazard rate is unspecified, and, through the exponential link function, the covariates $\boldsymbol{v} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_r)$ act multiplicatively on the hazard rate. As pointed out in the introduction, in a number of applications there is a need for extending this basic model with respect to several aspects. We propose novel nonparametric Bayesian survival models that can deal with these issues in a flexible and unified framework. Reparametrizing the baseline hazard rate through $\exp\{g_0(t)\}$, $g_0(t) = \log\{\lambda_0(t)\}$ and partitioning the vector of covariates into groups of covariates $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{s}$ and $\boldsymbol{v}$, we extend model (2) to the nonparametric multiplicative observation model

$$\lambda_i(t) := \lambda_i(t; \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{s}_i, \boldsymbol{v}_i) = \exp\{\eta_i(t)\} \tag{3}$$

with geoadditive predictor

$$\eta_i(t) = g_0(t) + \sum_{j=1}^{p} g_j(t) z_{ij} + \sum_{j=1}^{q} f_j(x_{ij}) + f_{spat}(s_i) + \boldsymbol{v}_i' \boldsymbol{\gamma} + b_{g_i}. \tag{4}$$

Here $g_0(t) = \log\{\lambda_0(t)\}$ is the log–baseline effect, $g_j(t)$ are time–varying effects of covariates $\boldsymbol{z}_j$, $f_j(x_j)$ is the nonlinear effect of a continuous covariate x_j, $f_{spat}(s)$ is the (structured) effect of the spatial covariate s, with $s_i = s$ if unit i is from area s, $s = 1, \ldots, S$, $\boldsymbol{\gamma}$ is the vector of usual linear fixed effects, and b_g is a unit– or group–specific frailty or random effect, with $b_{gi} = b_g$ if unit i is in group g, $g = 1, ..., G$. For $G = n$, we obtain individual–specific frailties, for $G < n$, b_g might be the effect of center g in a multicenter study or the unstructured (uncorrelated random) spatial effect of an area (i.e. $b_g = b_s$), for example. As an extension, random slopes could be introduced in (4), but we omit this here. Several other extensions of the model, such as choice of other link functions, inclusion of interactions and competing risks, are possible. We discuss this in the concluding section. For identifiability reasons, we center all unknown functions about zero, and include an intercept term in the parametric linear term.

Under the usual assumption about noninformative censoring, the likelihood is given by

$$\begin{aligned} L &= \prod_{i=1}^{n} \lambda_i(t_i)^{\delta_i} \cdot \exp\left(-\int_0^{t_i} \lambda_i(u) du\right) \\ &= \prod_{i=1}^{n} \lambda_i(t_i)^{\delta_i} \cdot S_i(t_i), \end{aligned} \tag{5}$$

inserting (3) and (4).

To obtain a unified and generic notation, we rewrite the observation model in general matrix notation. This is useful for defining priors in the next subsection and for developing posterior analysis in Section 3 as well as for describing and proving results on propriety of posteriors for mixed models in the Appendix.

Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_i, \ldots, \eta_n)'$ denote the predictor vector, where $\eta_i := \eta_i(t_i)$ is the value of predictor (4) at the observed lifetime $t_i, i = 1, \ldots, n$. Correspondingly, let $\boldsymbol{g}_j = (g_j(t_1), \ldots, g_j(t_n))'$ denote the vector of evaluations of the functions $g_j(t), j = 0, \ldots, p$, $\boldsymbol{f}_j = (f_j(x_{1j}), \ldots, f_j(x_{nj}))'$ the vector of evaluations of the functions $f_j(x_j), j = 1, \ldots, q$, $\boldsymbol{f}_{spat} = (f_{spat}(s_1), \ldots, f_{spat}(s_n))'$ the vector of spatial effects, and $\boldsymbol{b} = (b_{g_1}, \ldots, b_{g_n})'$ the vector of uncorrelated random effects.

In the following, we express all vectors $\boldsymbol{g}_j$, $\boldsymbol{f}_j$, $\boldsymbol{f}_{spat}$ and $\boldsymbol{b}$ as the matrix product of an appropriately defined design matrix $\boldsymbol{Z}$, say, and a (possibly high-dimensional) vector $\boldsymbol{\beta}$ of parameters, e.g. $\boldsymbol{g}_l = \boldsymbol{Z}_l\boldsymbol{\beta}_l$, $\boldsymbol{f}_j = \boldsymbol{Z}_j\boldsymbol{\beta}_j$, etc. Then, after reindexing, we can represent the predictor vector $\boldsymbol{\eta}$ in generic notation as

$$\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{Z}_0\boldsymbol{\beta}_0 + \ldots + \boldsymbol{Z}_m\boldsymbol{\beta}_m. \tag{6}$$

2.2 PRIORS FOR PARAMETERS AND FUNCTIONS

The Bayesian model formulation is completed by assumptions about priors for parameters and functions. For fixed effect parameters $\boldsymbol{\gamma}$ in (6) we assume diffuse priors $p(\boldsymbol{\gamma}) \propto const.$ A weakly informative normal prior would be another choice. Uncorrelated random effects are assumed to be i.i.d. Gaussian, $b_g \sim N(0, \tau_b^2)$. Priors for functions and spatial components are defined by a suitable design matrix $\boldsymbol{Z}_j$, $j = 1, \ldots, m$, and a prior for the parameter vector $\boldsymbol{\beta}_j$. The general form of a prior for $\boldsymbol{\beta}_j$ in (6) is

$$p(\boldsymbol{\beta}_j|\tau_j^2) \propto \tau_j^{-r_j} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right), \tag{7}$$

where $\boldsymbol{K}_j$ is a precision or penalty matrix of rank$(\boldsymbol{K}_j) = r_j$, shrinking parameters towards zero or penalizing too abrupt jumps between neighboring parameters. For P–splines and MRF priors, $\boldsymbol{K}_j$ will be rank deficient, i.e., $r_j < d_j$ =dim$(\boldsymbol{\beta}_j)$, and the prior is partially improper.

For *unknown functions* $f_j(x_j)$ or $g_j(t)$, we assume Bayesian P–spline priors as in Lang and Brezger (2004). Random walk priors, which have been suggested in Fahrmeir and Lang (2001) and may be used as smoothness priors for the baseline effect and time–varying covariate effects in a piecewise exponential model, appear as a special case. The basic idea of P-spline regression (Eilers and Marx 1996) is to approximate a function $f_j(x_j)$ as a linear combination of B-spline basis functions B_m, i.e.

$$f_j(x_j) = \sum_{m=1}^{d_j} \beta_{jm} B_m(x_j). \tag{8}$$

The basis functions B_m are B–splines of degree l defined over a grid of equally spaced knots $x_{min} = \xi_0 < \xi_1 < \ldots < \xi_s = x_{max}$, $d_j = l + s$. The number of knots is moderate, but not too small, to maintain flexibility, but smoothness of the function is encouraged by difference penalties for neighboring coefficients in the sequence $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_j})'$. The Bayesian analogue are first or second order random walk smoothness priors

$$\beta_{jm} = \beta_{j,m-1} + u_{jm} \qquad \text{or} \qquad \beta_{jm} = 2\beta_{j,m-1} - \beta_{j,m-2} + u_{jm} \tag{9}$$

with i.i.d. Gaussian errors $u_{jm} \sim N(0, \tau_j^2)$ and diffuse priors $p(\beta_{j1}) \propto const$, or $p(\beta_{j1})$ and $p(\beta_{j2}) \propto const$, for initial values. A first order random walk penalizes abrupt jumps $\beta_{jm} - \beta_{j,m-1}$, and a second order random walk penalizes deviations from a linear trend. The amount of smoothness or penalization is controlled by the variance τ_j^2, which acts as a smoothness parameter.

The joint prior of the regression parameters $\boldsymbol{\beta}_j$ is Gaussian and can be easily computed as a product of conditional densities defined by (8) as

$$\boldsymbol{\beta_j} \mid \tau_j^2 \propto \tau_j^{-r_j} \exp\left(-\frac{1}{2\tau_j^2} \boldsymbol{\beta_j}' \mathbf{K_j} \boldsymbol{\beta_j} \right), \tag{10}$$

which is the generic form (7).

The penalty matrix $\boldsymbol{K}_j$ is of the form $\boldsymbol{K}_j = \boldsymbol{D'D}$, where $\boldsymbol{D}$ is a first or second order difference matrix. The matrix $\boldsymbol{K}_j$ has band structure which is very useful for computationally efficient MCMC updating schemes. It has rank $r_j = d_j - 1$ and $r_j = d_j - 2$ for first and second order random walk priors, respectively. The $n \times d_j$ design matrix $\boldsymbol{Z}_j$ consists of the basis functions evaluated at the observations x_{ij}, i.e., $\boldsymbol{Z}_j(i, m) = B_m(x_{ij})$. Priors for the unknown functions $g_j(t)$ are defined in complete analogy as in (8), (9) and (10).

A common choice for approximating smooth curves are quadratic or cubic B-splines. Computationally, linear splines are simpler. The simplest choice are B–splines of degree zero, i.e. $B_m(x) \equiv 1$ over the m-th interval, and $B_m(x) \equiv 0$ elsewhere. Then the effect is approximated by a piecewise constant function, and the function values follow a random walk model as in Fahrmeir and Lang (2001). This special choice, with time t as covariate, is the easiest way to smooth the baseline in the piecewise exponential model; moreover the integral in the likelihood (5) reduces to a sum, see the next section. With P–splines of higher degree, however, estimation of smooth baseline effects is improved in terms of MSE's, see Section 4.

For the *structured spatial effect* $f_{spat}(s)$ we assume either Markov random field (MRF) priors, two dimensional tensor product P–spline priors, or Gaussian random field (GRF) priors, common in geostatistics (kriging).

In the case of *MRF priors* we define areas as neighbors if they share a common boundary and assume that the effect of an area s is conditionally Gaussian, with the mean of the effects of neighboring areas as expectation and a variance that is inverse proportional to the number of neighbors of area s, i.e.

$$f_{spat}(s) := \beta_s^{spat} = \frac{1}{N_s} \sum_{s' \in \delta_s} \beta_{s'}^{spat} + u_s, \quad u_s \sim N\left(0, \frac{\tau_{spat}^2}{N_s}\right)$$

where N_s is the number of neighbors of area s, and $s' \in \delta_s$ denotes that area s' is a neighbor of area s. This prior is a generalization of a first order random walk to two dimensions and is also called a conditionally autoregressive (CAR) prior. The $n \times S$ design matrix $\boldsymbol{Z}_{spat}$ is now a 0/1 incidence matrix. Its value in the i-th row and s-th column is 1 if observation i is located in site or region s, and zero otherwise. The $S \times S$ penalty matrix $\boldsymbol{K}_{spat}$ has the form of an adjacency matrix with $\text{rank}(\boldsymbol{K}_{spat}) = r_{spat} = S - 1$.

Our second approach is based on *two–dimensional P–splines*, a rather parsimonious, but flexible method for modelling interactions between continuous covariates described in Lang and Brezger (2004) for Gaussian regression. Considering the x– and y–coordinates of the geographical center of each area, the spatial effect can be seen as an interaction between two continuous covariates x_s and y_s. The assumption is that the unknown structured spatial effect $f_{spat}(s)$ can be approximated by the tensor product of one–dimensional B–splines, i.e.

$$f_{spat}(s) = f_{spat}(x_s, y_s) = \sum_{m_1=1}^{d_{spat}} \sum_{m_2=1}^{d_{spat}} \beta^{spat}_{m_1 m_2} B_{spat,m_1}(x_s) B_{spat,m_2}(y_s).$$

Now the B–splines of degree l are defined over a regular two–dimensional grid of a moderate, but not too small number of equally spaced knots $\xi_{\rho\nu}$, $\rho, \nu = 1, \ldots, d_{spat} - 1$. We restrict ourselves to an equal number of knots for each direction. Knots are equally spaced within each direction, but the distance may differ between direction x_s and y_s. Priors for $\boldsymbol{\beta}^{spat} = (\beta^{spat}_{11}, \ldots, \beta^{spat}_{1d_{spat}}, \ldots, \beta^{spat}_{d_{spat}1}, \ldots, \beta^{spat}_{d_{spat}d_{spat}})'$ are based on spatial smoothness priors common in spatial statistics (see e.g. Besag and Kooperberg, 1995). Since there is no natural ordering of parameters, priors have to be defined by specifying the conditional distributions of $\beta^{spat}_{m_1 m_2}$ given neighboring parameters and the variance component τ^2_{spat}. The most commonly used prior specification based on the four nearest neighbors can be defined by

$$\beta^{spat}_{m_1 m_2} | \cdot \sim N\left(\frac{1}{4}(\beta^{spat}_{m_1-1,m_2} + \beta^{spat}_{m_1+1,m_2} + \beta^{spat}_{m_1,m_2-1} + \beta^{spat}_{m_1,m_2+1}), \frac{\tau^2_{spat}}{4}\right) \qquad (11)$$

for $m_1, m_2 = 2, \ldots, d_{spat} - 1$ and appropriate changes for corners and edges. For example, for the upper left corner we obtain $\beta^{spat}_{11} | \cdot \sim N(\frac{1}{2}(\beta^{spat}_{12} + \beta^{spat}_{21}), \frac{\tau^2_{spat}}{2})$. For the left edge, we get $\beta^{spat}_{1m_2} | \cdot \sim N(\frac{1}{3}(\beta^{spat}_{1,m_2+1} + \beta^{spat}_{1,m_2-1} + \beta^{spat}_{2,m_2}), \frac{\tau^2_{spat}}{3})$.

The prior (11) is a direct generalization of a first order random walk in one dimension. Its conditional mean can be interpreted as a least squares locally linear fit at knot position $\xi_{\rho\nu}$ given the neighboring parameters.

More details can be found in Lang and Brezger (2004). Defining $\boldsymbol{K}_{spat} = \boldsymbol{D}_1'\boldsymbol{D}_1 + \boldsymbol{D}_2'\boldsymbol{D}_2$, where $\boldsymbol{D}_1 = \boldsymbol{I} \otimes \boldsymbol{D}$ and $\boldsymbol{D}_2 = \boldsymbol{D} \otimes \boldsymbol{I}$, the prior can again be expressed in the general form (7). Here, $\boldsymbol{D}$ is the first order difference matrix known from the one–dimesional case, and $\boldsymbol{D}_1'\boldsymbol{D}_1$ corresponds to the penalization in the direction of x and $\boldsymbol{D}_2'\boldsymbol{D}_2$ corresponds to the penalization in the direction of y.

Our third option are *stationary Gaussian random field* (GRF) priors, which can be seen as two-dimensional surface smoothers based on special basis functions, e.g. radial basis functions, and have been used by Kammann and Wand (2003) for modelling the spatial component in Gaussian regression models. The spatial component $f_{spat}(s) = \beta_s^{spat}$ is assumed to follow a zero mean stationary Gaussian random field $\{\beta_s^{spat} : s \in \mathbb{R}^2\}$ with variance τ_{spat}^2 and use an isotropic covariance function $\text{cov}(\beta_s^{spat}, \beta_{s'}^{spat}) = C(\|s - s'\|)$ as proposed by Stein (1999). For a finite array $s \in \{1, \dots, S\}$ of sites as in our application the prior can be brought in the general form

$$\boldsymbol{\beta}^{spat} \mid \tau_{spat}^2 \propto \exp\left(-\frac{1}{2\tau_{spat}^2}(\boldsymbol{\beta}^{spat})'\boldsymbol{K}_{spat}\boldsymbol{\beta}^{spat}\right)$$

with penalty matrix $\boldsymbol{K}_{spat} = \boldsymbol{C}^{-1}$, where $C[k,l] = C(\|s_k - s_l\|), 1 \leq k, l \leq n$, and design matrix $\boldsymbol{Z}_{spat} = \boldsymbol{C}$. For the covariance function $C(r)$ we follow again recommendations of Stein (1999) and use the Matérn family of covariance functions $C(r; \rho, \nu)$. For the special case $\nu = 1.5$ for the smoothness parameter the covariance functions simplify to

$$C(r; \rho, \nu) = \tau_{spat}^2(1 + |r|/\rho)e^{-|r|/\rho},$$

which is the simplest member of the Matérn family that results in differentiable surface estimates as Kammann and Wand (2003) point out. The parameter ρ controls how fast covariances die out with increasing distance r. We choose ρ according to the rule

$$\hat{\rho} = \max_{k,l} \|s_k - s_l\|/c$$

to ensure scale invariability. This rule proved to work well in practice. The constant c is chosen in such a way that $C(c)$ is small, e.g. $C(c) = 0.001$.

While the dimension of the penalty matrix in a MRF equals the number of different regions S, in a GRF the dimension corresponds to the number of distinct locations which is likely to be close to or equal to the sample size. To overcome this computational burden Kammann and Wand (2003) propose low–rank kriging to approximate stationary Gaussian random fields. Therefore they define a 'representative' subset of knots $\mathcal{D} = \{\kappa_1, \dots, \kappa_M\}$ of the set of distinct locations by applying a space filling algorithm (compare Johnson et al. (1990) and Nychka and Saltzman (1998) for details). Based on these knots, we obtain the approximation $f_{spat}(s) = \boldsymbol{z}'_{spat}(s)\boldsymbol{\beta}^{spat}$ with the M-dimensional design vector $\boldsymbol{z}_{spat}(s) = (C(\|s - \kappa_1\|), \dots, C(\|s - \kappa_M\|))'$ and penalty matrix $\boldsymbol{K}_{spat} = \tilde{\boldsymbol{C}}$ and $\tilde{C}[k, l] = C(\|\kappa_k - \kappa_l\|)$. The number of knots controls the trade-off between accuracy of the approximation and numerical simplification. Details on GRF and (low-rank) kriging can be found in Kammann and Wand (2003) or Kneib and Fahrmeir (2004).

The main drawback of this approach is the computational effort involved. Since the penalty matrix $\tilde{\boldsymbol{C}}$ has no longer band structure it is not possible to employ efficient matrix algorithms for sparse matrices like the Cholesky decomposition in order to draw samples from our multivariate normal proposal density and to compute the inverse of the precision matrix, which is needed to calculate the acceptance probability of the MH-step in every iteration (compare Section 3). For the application in Section 5, e.g., this means that the required CPU time multiplies approximately by the factor 20, even if we use low–rank kriging with a moderate number of 100 knots.

In general, it is not clear which of the different approaches leads to the best fit. For data observed on a discrete lattice or on the level of geographical regions as in our application, MRFs seem to be most adequate, while surface smoothers as 2d P–splines of kriging may be more natural in situations where exact locations are available. However, in applications sometimes surface estimators outperform MRFs even for discrete data and vice versa.

Again, in all described approaches the amount of smoothness is controlled by a smoothing parameter τ^2_{spat} that is estimated jointly with the unknown parameters $\boldsymbol{\beta}^{spat}$.

When applying our model to real data we do not know how much of the spatial variation is explained by structured, spatially correlated effects and how much by unstructured, uncorrelated effects. Therefore we usually fit an additional (unstructured) area–specific random effect in a first step and possibly remove one (or both) of these spatial effects, if it does not improve the fit. When fitting a structured and an unstructured spatial effect, we interpret the sum of the two effects, since identifiability is not given in that case.

Variances τ^2_j routinely follow inverse Gamma priors $IG(a_j; b_j)$. The hyperparameters a_j, b_j are chosen such that this prior is weakly informative. We use $a_j = b_j = 0.001$ as a standard choice. From our experience results are rather insensitive to the choice of $a_j > 0$ and $b_j > 0$ for moderate to large data sets and the posterior distribution is proper in any case (see Subsection 3.2 and Appendix for a proof). However, since the limiting case, when a_j and b_j are zero, leads to an improper posterior distribution, we present a sensitivity analysis in Section 4 and compare the results to those we obtain with an alternative prior specification that does not depend on further hyperparameters and is proposed in Gelman (2004), who imposes an uniform prior on the standard deviations τ_j.

We also routinely assume an inverse Gamma prior for the variance τ_b^2 of the Gaussian random effects b_g and the variance τ_{spat}^2 of the spatial effect (or an uniform prior on τ_b and τ_{spat}, respectively).

The Bayesian model specification is completed by assuming that all priors for parameters are conditionally independent, and that all priors are mutually independent.

3. MARKOV CHAIN MONTE CARLO INFERENCE

In what follows, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_0', ..., \boldsymbol{\beta}_m')'$ denote the vector of all regression coefficients in the generic notation (6), $\boldsymbol{\gamma}$ the vector of fixed effects, and $\boldsymbol{\tau}^2 = (\tau_0^2, ..., \tau_m^2)$ the vector of all variance components.

Full Bayesian inference is based on the entire posterior distribution

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2 \mid data) \propto L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2)\, p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2).$$

Due to the (conditional) independence assumptions, the joint prior factorizes into

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2) = \left\{ \prod_{j=0}^{m} p(\boldsymbol{\beta}_j \mid \tau_j^2) p(\tau_j^2) \right\} p(\boldsymbol{\gamma}),$$

where the last factor can be omitted for diffuse fixed effect priors.

The likelihood $L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2)$ is given by inserting (3),(4) into (5), but the integral requires integration over all terms depending on survival time t, i.e. terms of the form

$$I_i = \int_0^{t_i} \exp\left(g_0(u) + \sum_{j=1}^{p} g_j(u) z_{ij} \right) du,$$

where $g_j(t) = \sum \beta_{jm} B_m(t)$. Apart from B–splines $B_m(t)$ of degree zero, i.e. random walk models, and linear B–splines, these integrals are not available in closed form. The first case leads to the piecewise exponential model: The time axis is divided into a grid

$$0 = \xi_0 < \xi_1 < ... < \xi_{t-1} < \xi_t < ... < \xi_s = t_{max},$$

and $g_j(t)$ is assumed to be a piecewise constant function, i.e.

$$g_j(t) = \beta_{jt}$$

in time interval $(\xi_{t-1}, \xi_t]$, $t = 1, ..., s$. In this case, the integral reduces to a sum, and, after some simple calculations, the log–likelihood contribution of observation i in the interval $(\xi_{t-1}, \xi_t]$ can be expressed as

$$l_{it} = y_{it}\eta_{it} - \exp\left(\delta_{it} + \eta_{it}\right)$$

where

$$y_{it} = \begin{cases} 1 & t_i \in (\xi_{t-1}, \xi_t], \delta_i = 1 \\ 0 & \text{else.} \end{cases}$$

$$\Delta_{it} = \begin{cases} \xi_t - \xi_{t-1}, & \xi_t < t_i \\ t_i - \xi_{t-1}, & \xi_{t-1} < t_i \leq \xi_t \\ 0, & \xi_{t-1} \geq t_i \end{cases}$$

$$\delta_{it} = log\Delta_{it} \quad (\delta_{it} = -\infty \text{ if } \Delta_{it} = 0).$$

This likelihood is proportional to a Poisson–likelihood, with the predictor η_{it} containing an additional offset term δ_{it}, see Fahrmeir and Tutz (2001, Section 9.1) or Ibrahim et al. (2001, Section 3.1) for details.

For linear B–splines, the integrals can still be solved analytically, but expressions are rather messy and the computational effort is quite high, see Cai et al. (2002, Appendix). Following their suggestion, we use simple numerical integration in form of the trapezoidal rule for linear B–splines as well as for the commonly used cubic B–splines, where analytical integration is not possible anyway.

Full Bayesian inference via MCMC simulation is based on updating full conditionals of single parameters or blocks of parameters, given the rest of the data.

For updating the parameter vectors $\boldsymbol{\beta}_j$, which correspond to the time–independent functions $f_j(x_j)$, as well as spatial effects $\boldsymbol{\beta}^{spat}$, fixed effects $\boldsymbol{\gamma}$ and random effects $\boldsymbol{b}$, we use a slightly modified version of an MH–algorithm based on iteratively weighted least squares (IWLS) proposals, developed for fixed and random effects by Gamerman (1997) and adapted to generalized additive mixed models in Brezger and Lang (2003). More precisely, the goal is to approximate the posterior by a Gaussian distribution, obtained by accomplishing *one* IWLS step in every iteration of the sampler. Then, random samples have to be drawn from a high dimensional multivariate Gaussian distribution with precision matrix and mean

$$\boldsymbol{P}_j = \boldsymbol{X}_j'\boldsymbol{W}(\boldsymbol{\beta}_j^c)\boldsymbol{X}_j + \frac{1}{\tau_j^2}\boldsymbol{K}_j, \quad \boldsymbol{m}_j = \boldsymbol{P}_j^{-1}\boldsymbol{X}_j'\boldsymbol{W}(\boldsymbol{\beta}_j^c)(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{\eta}}).$$

Here, $\tilde{\eta_i} = \eta_i(t_i) - f_j(x_{ij})$, $\boldsymbol{W}(\boldsymbol{\beta}_j^c) = diag(w_1, \ldots, w_n)$ is the weight matrix for IWLS with weights

$$w_i = \exp\left(\sum_{j=1}^{q} f_j(x_{ij}) + f_{spat}(s_i) + \boldsymbol{v}_i'\boldsymbol{\gamma} + b_{g_i}\right) \cdot I_i$$

obtained form the current state $\boldsymbol{\beta}_j^c$. The working observations $\tilde{y}_i$ are given by

$$\tilde{y}_i = \eta_i(t_i) + \frac{\delta_i}{w_i} - 1.$$

Random numbers from the high dimensional proposal distributions can be efficiently drawn by using matrix operations for sparse matrices.

Suppose we want to update $\boldsymbol{\beta}_j$, with current value $\boldsymbol{\beta}_j^c$ of the chain. Then a new value $\boldsymbol{\beta}_j^p$ is proposed by drawing a random vector from a (high–dimensional) multivariate Gaussian proposal distribution $q(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p)$, which is obtained from a quadratic approximation of the log–likelihood by a second order Taylor expansion with respect to $\boldsymbol{\beta}_j^c$, in analogy to IWLS iterations in generalized linear models. The proposed vector $\boldsymbol{\beta}_j^p$ is accepted as the new state of the chain with probability

$$\alpha(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p) = \min\left(1, \frac{p(\boldsymbol{\beta}_j^p \mid \cdot)q(\boldsymbol{\beta}_j^p, \boldsymbol{\beta}_j^c)}{p(\boldsymbol{\beta}_j^c \mid \cdot)q(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p)}\right)$$

where $p(\boldsymbol{\beta}_j \mid \cdot)$ is the full conditional for $\boldsymbol{\beta}_j$ (i.e. the conditional distribution of $\boldsymbol{\beta}_j$ given all other parameters and the data).

For a fast implementation, we use the fact that the precision matrices of the Gaussian proposal distributions are banded, so that Cholesky decompositions can be performed efficiently.

For the parameters $\boldsymbol{\beta}_j$ corresponding to the functions $g_0(t), ..., g_p(t)$ depending on time t, the IWLS–MH algorithm requires considerably more computational effort, because the integrals in the log–likelihood as well as first and second derivatives are involved now. Therefore, we adopt a computationally faster MH–algorithm based on conditional prior proposals, although IWLS–MH has better mixing properties. This algorithm was first developed by Knorr–Held (1999) for state space models and extended for generalized additive mixed models in Fahrmeir and Lang (2001). It requires only evaluation of the log–likelihood, not of derivatives. However, draws are not performed for the entire vector $\boldsymbol{\beta}_j$, but iteratively for blocks of subvectors, see Fahrmeir and Lang (2001) for details.

The full conditionals for the variance parameters τ_j^2 are inverse Gamma with parameters

$$a_j' = a_j + \frac{1}{2} r_j \quad \text{and} \quad b_j' = b_j + \frac{1}{2} \boldsymbol{\beta}_j' \boldsymbol{K}_j \boldsymbol{\beta}_j$$

for inverse Gamma priors on τ_j^2 and

$$a_j' = \frac{r_j - 1}{2} \quad \text{and} \quad b_j' = \frac{1}{2} \boldsymbol{\beta}_j' \boldsymbol{K}_j \boldsymbol{\beta}_j$$

for uniform priors on τ_j. Updating can be done by simple Gibbs steps, drawing random numbers directly from the inverse Gamma densities. In complete analogy, the full conditional for a variance component τ_{spat}^2 of the spatial effect and τ_b^2 of a random intercept or slope is again an inverse gamma distribution, and updating is straightforward.

For model comparison we suggest to use the Deviance Information Criterion (DIC) developed in Spiegelhalter, Best, Carlin and van der Linde (2002). It is given as

$$DIC = D(\overline{\boldsymbol{\theta}}) + 2p_D = \overline{D(\boldsymbol{\theta})} + p_D,$$

where $\boldsymbol{\theta}$ is the vector of parameters, $D(\overline{\boldsymbol{\theta}})$ is the deviance of the model evaluated at the posterior mean estimate $\overline{\boldsymbol{\theta}}$, $\overline{D(\boldsymbol{\theta})}$ is the posterior mean of the deviance and $p_D = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}})$ is the effective number of parameters. Since it is at least unclear, how the saturated model should be defined in the case of survival data when the baseline hazard and other nonparametric functions are parameters of interest, we use the unstandardized deviance $D(\boldsymbol{\theta}) = -2 \cdot$log–likelihood instead of the saturated deviance.

3.2 PROPRIETY OF POSTERIORS IN GEOADDITIVE SURVIVAL MODELS

Consider a geoadditive survival model with predictor

$$\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{Z}_1\boldsymbol{\beta}_1 + ... + \boldsymbol{Z}_m\boldsymbol{\beta}_m + \boldsymbol{Z}_0\boldsymbol{\beta}_0$$

in generic form, where $\boldsymbol{Z}_0\boldsymbol{\beta}_0$ corresponds to an effect with prior

$$\boldsymbol{\beta}_0 \sim \tau_0^{-r_0} \exp\left(-\frac{1}{2\tau_0^2}\boldsymbol{\beta}_0'\boldsymbol{K}_0\boldsymbol{\beta}_0\right),$$

such that

$$\dim(\boldsymbol{\beta}_0) = d_0 \geq d_j, \quad \mathrm{rank}(\boldsymbol{K}_0) = r_0 \geq r_j, \qquad j = 1, ..., m.$$

This assumption is usually fulfilled for the spatial component or for a high–dimensional vector of group– or individual–specific uncorrelated random effects.

Denote by $\boldsymbol{\eta}_u$, $\boldsymbol{V}_u$, $\boldsymbol{Z}_u = (\boldsymbol{Z}_{1u}, ..., \boldsymbol{Z}_{mu})$, $\boldsymbol{Z}_{0u}$ the (sub–) predictor and sub–design matrices corresponding to uncensored observations. Assume that the following conditions hold:

(C1) $\mathrm{rank}(\boldsymbol{V}_u) = \mathrm{rank}(\boldsymbol{V}) = p = \dim(\boldsymbol{\gamma})$,

$\mathrm{rank}(\boldsymbol{Z}_{ju}) = \mathrm{rank}(\boldsymbol{Z}_j) = d_j = \dim(\boldsymbol{\beta}_j), \quad j = 0, ..., m$

$\mathrm{rank}(\boldsymbol{Z}_u'\boldsymbol{R}\boldsymbol{Z}_u + \boldsymbol{K}) = d$

where $d = d_1 + ... + d_m, \quad \boldsymbol{K} = \mathrm{diag}(\boldsymbol{K}_1, ..., \boldsymbol{K}_m), \quad \boldsymbol{R} = \boldsymbol{I} - \boldsymbol{V}_u(\boldsymbol{V}_u'\boldsymbol{V}_u)^{-1}\boldsymbol{V}_u'$

(C2) The priors $p(\tau_j^2)$, $j = 1, ..., m$, are proper, and $\int p(\tau_0^2)\tau_0^{-(r_0-p-(d-r))}d\tau_0^2 < \infty$,

where $r = r_1 + ... + r_m$.

Theorem: If conditions (C1), (C2) hold then the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\tau}^2, \tau_0^2 \mid \boldsymbol{y})$, where $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_m^2)'$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)'$, is proper.

Corollary: Assume proper inverse Gamma priors for τ_j^2 with

$$a_j > 0,\ b_j > 0, \qquad j = 0, ..., m,$$

and

$$r_0 - p - (d - r) - (d_0 - r_0) > 0.$$

If condition (C1) holds, then the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\tau}^2, \tau_0^2 \mid \boldsymbol{y})$ is proper.

Remark: Condition (C1) is equivalent to

$$\text{rank} \begin{pmatrix} \boldsymbol{V}'_u \boldsymbol{V}_u & \boldsymbol{V}'_u \boldsymbol{Z}_u \\ \boldsymbol{Z}'_u \boldsymbol{V}_u & \boldsymbol{Z}'_u \boldsymbol{Z}_u + \boldsymbol{K} \end{pmatrix} = p + d$$

Proof: We first show that the conditions $(*)$, $(**)$ of Lemma A2 (Appendix) are fulfilled for right–censored survival data $(t_i = \text{min}(T_i, U_i), \delta_i), \quad i = 1, ..., n.$ The density of observation i is given by

$$f_i(t_i \mid \eta_i(t_i)) = \lambda_i(t_i)^{\delta_i} S_i(t_i),$$

where

$$\lambda_i(t_i) = \exp(\eta_i(t_i)), \quad S_i(t_i) = \exp\left(-\int_0^{t_i} \lambda_i(s) ds\right)$$

For censored observations ($\delta_i = 0$), we have $f_i(t_i \mid \eta_i(t_i)) = S_i(t_i) \leq 1$, so that condition $(**)$ of Lemma A2 holds.

For uncensored observations ($\delta_i = 1$)

$$f_i(t_i \mid \eta_i(t_i)) = \lambda_i(t_i) S_i(t_i).$$

Setting $\eta_i := \eta_i(t_i), \quad \lambda_i := \lambda_i(t_i)$, we obtain

$$\int_0^\infty f_i(t_i \mid \eta_i) d\eta_i = \int_0^\infty \lambda_i S_i(t_i) \lambda_i^{-1} d\lambda_i = \int_0^\infty S_i(t_i) d\lambda_i,$$

so that assumption $(*)$ is equivalent to

$$\int_0^\infty S_i(t_i) d\lambda_i < \infty. \tag{13}$$

We factorize the multiplicative hazard rate $\lambda_i(t)$ into

$$\lambda_i(t) = c_i l_i(t),$$

where $c_i > 0$ is the time–constant part. Then

$$\int_0^\infty S_i(t_i) d\lambda_i = \int_0^\infty \exp\left\{-c_i \int_0^{t_i} l_i(s) ds\right\} d\lambda_i.$$

Consider first the case where $\eta_i(t)$ is piecewise constant (on the intervals I_k, $k = 1, 2, ...$ defined by the knots of B–splines of degree 0). Then

$$\lambda_i(t) = c_i \lambda_{ik} \text{ for } t \in I_k, \ k = 1, 2, ...$$

For $t_i \in I_k$, say, we have $\lambda_i = \lambda_i(t_i) = c_i \lambda_{ik}$, and

$$\begin{aligned} \int_0^\infty S_i(t_i) d\lambda_i &\propto \int_0^\infty \exp\left(- \left(c_i \sum_{j=1}^{k-1} \Delta_j \lambda_{ij} \right) - c_i \int_{\xi_{k-1}}^{t_i} \lambda_{ik} d\lambda_{ik} \right) d\lambda_{ik} \\ &\propto C_i \int_0^\infty \exp(-c_i(t_i - \xi_{k-1})\lambda_{ik}) d\lambda_{ik} < \infty, \end{aligned}$$

for $t_i - \xi_{k-1} > 0$, which is valid a.s. for continuous T_i.

Consider now the case, where the time–varying part of $\eta_i(t)$ is defined by B–splines of higher degree. Let

$$\lambda_{ik} = \min_{t \in I_k} l_i(t) > 0, \quad k = 1, 2, ...$$

be the minimum of the time–varying part of $\lambda_i(t)$ on I_k.

Then

$$\begin{aligned} \int_0^\infty \exp\left\{ -c_i \int_0^{t_i} l_i(s) ds \right\} d\lambda_i &\leq C_i \int_0^\infty \exp\left\{ -c_i \int_{\xi_{k-1}}^{t_i} \lambda_{ik} d\lambda_{ik} \right\} d\lambda_{ik} \\ &= C_i \int_0^\infty \exp(-c_i(t_i - \xi_{k-1})\lambda_{ik}) d\lambda_{ik} < \infty, \end{aligned}$$

so that assumption (13) is fulfilled.

Remark: We have tacitly made the assumption that $\lambda_i(t) > 0$ for any choice of covariates and parameters. This is valid because of our parametrization

$$\lambda_i(t) = \exp(\eta_i(t)).$$

4. SIMULATION STUDY

We investigate performance through a simulation study. Life times T_i, $i = 1, ..., 1236$, were generated according to the hazard model

$$\begin{aligned} \lambda_i(t) &= \lambda_0(t) \exp(f_1(x_i) + f_{spat}(s_i) + \gamma v_i) \\ &= \exp(log(3t^2) + sin(x_i) + sin(x_{s_i} \cdot y_{s_i}) - 0.3 v_i). \end{aligned} \tag{14}$$

In this model, the baseline hazard rate $\lambda_0(t)$ is set to $3t^2$, which is a Weibull hazard rate, so that $g_0(t) = log(3t^2)$. The covariate v is binary, with the v_i´s randomly drawn from a Bernoulli $B(1; 0.5)$ distribution, and the covariate x is continuous, with the x_i´s randomly drawn from a uniform $U[-3,3]$ distribution. The spatial covariate s_i denotes one of the $s = 1, \ldots, S = 309$ counties of the former Federal Republic of Germany and x_{s_i} and y_{s_i} are the centered coordinates of the geographic center of county s_i. We simulated four observations per county. Censoring variables C_i, $i = 1, ..., 1236$, were generated as i.i.d. draws from a uniform $U[0,5]$ distribution, resulting in a proportion of 15–20 % of censored observations.

Keeping the predictor fixed, 100 replications $\{T_i^{(r)}, C_i^{(r)}, i = 1, ..., 1236\}$ resp. $\{(t_i^{(r)}, \delta_i^{(r)}),\ i = 1, ..., 1236\}$, $r = 1, ..., 100$ of censored survival times were generated.

The log–baseline hazard $g_0(t)$ was modelled by second order random walk priors, corresponding to a piecewise exponential model (with grid length $\triangle = 0.1$), and – alternatively – as a cubic P–spline, with 20 knots. A cubic P–spline prior with 20 knots was chosen for $f_1(x) = sin(x)$. The spatial effect was modelled as a MRF and alternatively as a two dimensional cubic P–spline with 12x12 knots. Hyperparameters of inverse Gamma priors for variance components were set to $a = 0.001$, $b = 0.001$, the standard choice.

For each replication $r = 1, ..., 100$, we computed the mean square errors

$$MSE_r(g_0) = \frac{1}{1236} \sum_{i=1}^{1236} (\widehat{g}_0^{(r)}(t_i^{(r)}) - g_0(t_i^{(r)}))^2,$$

for the log–baseline hazard $g_0(t)$,

$$MSE_r(f_1) = \frac{1}{1236} \sum_{i=1}^{1236} (\widehat{f}_1^{(r)}(x_i) - f_1(x_i))^2$$

for $f_1(x) = sin(x)$, and

$$MSE_r(f_{spat}) = \frac{1}{1236} \sum_{i=1}^{1236} (\widehat{f}_{spat}^{(r)}(s_i) - f_{spat}(s_i))^2$$

for the spatial effect $f_{spat}(s) = sin(x_c \cdot y_c)$, where $\widehat{g}_0^{(r)}$ and $\widehat{f}_k^{(r)}$, $k = 1, spat$, are posterior mean estimates for simulation run r.

The $MSE(\gamma)$ was computed in the usual way.

Table 1 summarizes the results, displaying $meanMSE = \frac{(\sum_{r=1}^{100} MSE_r)}{100}$ as well as $min_r MSE_r$ and $max_r MSE_r$ in each cell. As was to be expected, the P–spline model has smaller MSE´s for g_0 when compared to the piecewise exponential model. Interestingly, the MSE´s for $\gamma = -0.3$, $f_1(x)$ and $f_{spat}(s)$ are more or less unaffected by the choice of the smoothness prior for the log–baseline $g_0(t)$. Estimated functions of replication r, with r chosen such that MSE_r is the median of $MSE_1, \ldots, MSE_{100}$, for $g_0(t)$,

$f_1(x)$ and $f_{spat}(s)$ are displayed in Figures 1-3.

In order to analyze the behavior of the Markov chains when a and b approach zero (and the prior for the hyperparameters thus approaches the IG(0,0) distribution, that leads to an improper posterior), we exemplary single out the P–spline model with MRF-prior and alternatively set $a = b = 0.0001$, $a = b = 0.00001$ and $a = b = 0.00000001$. We additionally run the simulation study with uniform priors on the standard deviations τ_0, τ_1 and τ_{spat} that act as smoothing parameters for the log-baseline, the nonlinear effect of x and the spatial effect, respectively. We did not face problems with mixing or convergence of Markov chains with any of these prior distributions. Figure 4 shows kernel density estimators of the posterior mean of the variance parameters based on $\widehat{\tau_j^2}^{(r)}, r = 1, \ldots, 100$ for $j = 0, 1, spat$. Obviously the different choices of the hyperparameters a and b of the inverse Gamma prior do not seem to have much effect, whereas the uniform prior on the standard deviations tends to result in larger estimates for the variance parameters and thus in less smooth effects. The posterior distribution of the variance parameter of the spatial effect is less sensitive to the different choices of priors, as the full conditional is dominated by the values of r_j =rank($\boldsymbol{K}_j$) and $\boldsymbol{\beta}_j' \boldsymbol{K}_j \boldsymbol{\beta}_j$ at this. Table 2 summarizes the MSE´s, that are computed and displayed as before. While the MSE´s are quite unaffected by the choice of the hyperparameters a and b of the inverse Gamma prior, the uniform prior results in a slightly smaller MSE for $g_0(t)$, but a slightly bigger MSE for $f_1(x)$. Altogether we come to the conclusion that (at least with this model) it does not seem to be crucial, which one of these weakly informative priors is assumed for the variance parameters.

5. APPLICATION: WAITING TIMES TO CABG

We illustrate our methods by an application to data from a study in London and Essex that aims to analyze the effects of area of residence and further individual specific covariates on waiting times to coronary artery bypass graft (CABG). The data comprise observations for 3015 patients with definite coronary artery disease who were referred to one cardiothoracic unit from five contiguous health authorities. Waiting times from angiography to CABG are given in days. Covariates are, among others, sex, age (in years), number of diseased vessels (1, 2, 3), and the area of residence (one of 488 electoral wards).

The data were previously analyzed by Crook et al. (2003) who classified waiting times in months and applied discrete–time survival methodology as described for example in Fahrmeir and Tutz (2001, chap. 9). They analyzed and compared a hierarchy of models, with model comparison based on the deviance information criterion (DIC), developed in Spiegelhalter et al. (2002). Here we apply continuous–time geoadditive survival models, with waiting times given in days as in the original data set, and predictors based on models 8 and 12 in Crook et al. (2003), which were two of the best in terms of DIC. Model 8 corresponds to a continuous–time model with hazard rate

$$\lambda(t) = \exp(g_0(t) + f_{age}(age) + f_{spat}(ward) + \gamma_1 sex + \gamma_2 dv2 + \gamma_3 dv3), \tag{15}$$

where $g_0(t)$ is the log–baseline rate, $f_{age}(age)$ is the nonlinear effect of age and $f_{spat}(ward)$ is the structured spatial effect. The remaining covariates are dummy–coded: $sex = 1$ for female, and $sex = 0$ for male, $dv2 = 1$ if the number of diseased vessels equals 2, $dv2 = 0$ else, and $dv3 = 1$ if the number of diseased vessels equals 3, $dv3 = 0$ else. We did not add an unstructured spatial effect here, since the DIC would not be improved substantially (compare model 10 in Crook et al. (2003)).

The (log–) baseline prior was assumed as a (log–) piecewise exponential model with grid length $\triangle = 50$ days and, alternatively, as a cubic P–spline model with 20 knots. For f_{age} we assumed a cubic P–spline prior with 20 knots. Since the distribution of the values of age is quite skew, it is an interesting alternative to chose the knot positions according to quantiles, but we used equidistant knots here, which is our standard choice. The spatial effect $f_{spat}(ward)$ is once modelled through a MRF prior and in the case of the P–spline model through a GRF prior with 100 knots, alternatively. Since the data augmentation that has to be accomplished for the p.e.m. results in an "observation number" of more than 30000, a GRF prior would lead to a computation time of several days, which is not very viable.

Model (16) is a modification of (15), where the fixed effects γ_2 and γ_3 of $dv2$ and $dv3$ are replaced by time varying effects:

$$\lambda(t) = \exp(g_0(t) + f_{age}(age) + f_{spat}(ward) + \gamma_1 sex + g_1(t) dv2 + g_2(t) dv3). \tag{16}$$

This model is a nonproportional hazard model and can be compared to the geoadditive proportional hazard rate model (15).

Table 3 contains estimation results for the fixed effects in model (15). While the effect of sex is nonsignificant, the effects of two or three diseased vessels are clearly significant and show that waiting times are decreasing with increasing number of vessels. These results correspond to the findings of Crook et al. (2003). The baseline effects in Figure 5 show an initially high, but strongly decreasing chance of CABG immediately after diagnosis, followed by a slow increase between 150 - 450 days. Later, the chance of being operated decreases. The overall pattern is similar to the results in Crook et al. (2003), obtained with a discrete–time model. However, with the P–spline prior we get a distinctly smoother curve. The effect of age (Figure 6) is almost constant between 40 and 80 years and does not have significant influence on the waiting time. Also, the estimates under a piecewise exponential and a cubic P–spline baseline prior are visually indistinguishable - regardless of which prior is chosen for the structured spatial effect.

The maps in Figure 7 show the estimates for the structured spatial effects and give an impression of the spatially varying chance of CABG with light (dark) areas indicating an increased (decreased) effect. Again, the estimates under a piecewise exponential and a cubic P–spline baseline prior are nearly visually indistinguishable in the case of a MRF prior. Predictably, the GRF prior results in a smoother estimated spatial effect than the MRF prior does, but besides that the results are quite alike. Areas with increased chances are Chelmsford and Malden in North Essex, while in areas around Harlow in North Essex and Walthamstow and Chingford in North East London chances are lower, that means patients have to wait longer for surgery. The maps in Figure 8 show posterior probabilities of these spatial effects. White (black) areas indicate that at least 80 % of the sample estimates were positive (negative). Remaining grey areas are considered as 'nonsignificant'. Striped areas denote wards, where no patient was observed.

Model (16) with time–varying effects $g_1(t)$ and $g_2(t)$ of $dv2$ and $dv3$ can be interpreted as a model with three separate baseline effects $g_0(t)$, $g_0(t) + g_1(t)$, $g_0(t) + g_2(t)$ for patients with one, two or three diseased vessels, respectively. The corresponding estimated curves are displayed in Figure 9 and indicate that the proportional hazards assumption is violated, because the baseline effect of patients with three diseased vessels crosses the two other curves.

6. CONCLUSION

Spatial extensions of statistical models for analyzing survival and, more general, event history data, will be of increasing relevance because spatial small–area information is often available. Assessment of spatial effects on hazard or survivor functions is not only of interest in its own but can be quite useful for detecting unobserved covariates which carry spatial information. In this work, we have developed a flexible class of nonparametric geoadditive survival models within a unified Bayesian framework for modelling and inference.

Several extensions could be considered in future research. More general event history models and censoring mechanisms including spatial components can be embedded in the counting process framework (Andersen, Borgan, Gill and Keiding 1993). Practical issues are the development of numerical alternatives for evaluating the likelihood in the presence of time–varying effects of covariates and of low–rank kriging approximations to improve computational efficiency of geostatistical approaches.

References

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993), *S*tatistical models based on counting processes, New York: Springer.

Arjas, E., and Gasbarra, D. (1994), "Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler," *Statistica Sinica* 4, 505–524.

Banerjee, S., and Carlin, B.P. (2002), "Semiparametric Spatiotemporal Frailty Modelling," *Environmetrics*, 14, 523–535.

Banerjee, S., Wall, M. M., and Carlin, B. P. (2003), "Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota," *Biostatistics*, 4, 123–142.

Besag, J. and Kooperberg, C. (1995), "On Conditional and Intrinsic Autoregressions," *Biometrika*, 82, 733–746.

Brezger, A., Kneib, T., and Lang, S. (2003), "BayesX - Software for Bayesian Inference based on Markov Chain Monte Carlo simulation techniques," open domain software available from `http://www.stat.uni-muenchen.de/~lang/`.

Brezger, A., and Lang, S. (2003), "Generalized structured additive regression based on Bayesian P-splines," Discussion Paper 321, SFB 386, Ludwig-Maximilians-Universität München.

Cai, T., and Betensky, R. A. (2003), "Hazard Regression for Interval Censored Data with Penalized Spline," *Biometrics*, 59, 570–9.

Cai, T., Hyndman, R., and Wand, M. (2002), "Mixed model-based hazard estimation," *Journal of Computational and Graphical Statistics*, 11, 784–798.

Carlin, B. P., and Banerjee, S. (2002), "Hierarchical Multivariate CAR Models for Spatio–Temporally Correlated Data," In:*Bayesian Statistics 7*, eds. J.M. Bernardo et al., Oxford: Oxford University Press.

Crook, A., Knorr-Held, L., and Hemingway, H. (2003), "Measuring spatial effects in time to event data: a case study using months from angiography to coronary artery bypass graft (CABG)," *Statistics in Medicine*, 22, 2943–2961.

Eilers, P.H.C., and Marx, B.D. (1996), "Flexible smoothing using B-splines and penalized likelihood" (with comments and rejoinder), *Statistical Science*, 11 (2), 89–121.

Fahrmeir, L., and Lang, S. (2001), "Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors," *Journal of the Royal Statistical Society*,Ser. C, 50, 201–220.

Fahrmeir, L., Lang, S.,Wolff, J., and Bender, S. (2003), "Semiparametric Bayesian Time-Space Analysis of Unemployment Duration," *Journal of the German Statistical Society (Allgemeines Statistisches Archiv)*, 87, 281–307.

Fahrmeir, L., and Tutz, G. (2001), *Multivariate Statistical Modelling based on Generalized Linear Models*, Springer–Verlag, New York.

Gamerman, D. (1991), "Dynamic Bayesian models for survival data," *Applied Statistics*, 40, 63–79.

——(1997), "Efficient Sampling from the Posterior Distribution in Generalized Linear Models," *Statistics and Computing*, 7, 57–68.

Gelman, A. (2004), "Prior distributions for variance parameters in hierarchical models," provided by *Economics Working Paper Archive at WUSTL* in its series *Econometrics* with number 0404001.

Henderson, R., Shimakura, S., and Gorst, D. (2002), "Modeling Spatial Variation in Leukemia Survival Data," *Journal of the American Statistical Assosiation*, 97, 965–972.

Ibrahim, J.G., Chen, M.H., and Sinha, D. (2001), *Bayesian Survial Analysis.* Springer Series in Statistics, New York.

Johnson, M.E., Moore, L.M. and Ylvisaker, D. (1990), "Minimax and maximin designs," *Journal of Statistical Planning and Inference*, 26, 131–148.

Kammann, E.E., and Wand, M.P. (2003), "Geoadditive models," *Journal of the Royal Statistical Society*, Ser. C, 52, 1–18.

Kneib, T. and Fahrmeir, L. (2004), "Structured additive regression for multicategorical space-time data: A mixed model approach," *SFB 386 Discussion Paper 377*, University of Munich.

Knorr–Held, L. (1999), "Conditional Prior Proposals in Dynamic Models," *Scandinavian Journal of Statistics*, 26, 129–144.

Lang, S., and Brezger, A. (2004), "Bayesian P-splines," *Journal of Computational and Graphical Statistics*, 13, 183–212.

Li, Y., and Ryan, L. (2002), "Modeling Spatial Survival Data Using Semiparametric Frailty Models," *Biometrics*, 58, 287–297.

Magnus, J.R. and Neudecker, H. (1991), *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley Series in Probability and Statistics.

Marx, B.D., and Eilers, P. (1998), "Direct Generalized Additive Modeling with Penalized Likelihood," *Computational Statistics and Data Analysis*, 28, 193–209.

Nychka, D. and Saltzman, N. (1998), "Design of Air-Quality Monitoring Networks," *Lecture Notes in Statistics*, 132, 51–76.

Sinha, D. (1993), "Semiparametric Bayesian analysis of multiple event time data," *Journal of the American Statistical Association*, 92, 1195–1212.

Speckman, P.L. and Sun, D. (2003), "Fully Bayesian spline smoothing and intrinsic autoregressive priors," *Biometrika*, 90, 2, 289–302.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002), "Bayesian measures of model complexity and fit" (with discussion and rejoinder), *Journal of the Royal Statistical Society*, Ser. B, 64, 583–639.

Stein, M.L. (1999), *Interpolation of spatial data. Some theory for kriging*, Springer, New York.

Sun, D., Tsutakawa, R.K. and Speckman, P.L. (1999), "Posterior distribution of hierarchical models using CAR(1) distributions," *Biometrika*, 86, 341–350.

Table 1: Summary of MSE´s

	p.e.m		P–spline–model	
MSE–type	MRF	geospline	MRF	geospline
$meanMSE(g_0)$	0.154	0.155	0.126	0.127
$minMSE(g_0)$	0.049	0.044	0.033	0.027
$maxMSE(g_0)$	0.497	0.496	0.450	0.453
$meanMSE(f_1)$	0.0068	0.0061	0.0070	0.0063
$minMSE(f_1)$	0.0006	0.0006	0.0009	0.0006
$maxMSE(f_1)$	0.0193	0.0182	0.0209	0.0178
$meanMSE(f_{spat})$	0.042	0.022	0.043	0.022
$minMSE(f_{spat})$	0.028	0.010	0.029	0.010
$maxMSE(f_{spat})$	0.068	0.039	0.071	0.039
$meanMSE(\gamma)$	0.0045	0.0038	0.0046	0.0038
$minMSE(\gamma)$	≈ 0	≈ 0	≈ 0	≈ 0
$maxMSE(\gamma)$	0.0268	0.0197	0.0297	0.0202

Table 2: Summary of MSE´s

prior	IG,a=b=0.001	IG,a=b=0.0001	IG,a=b=1e-05	IG,a=b=1e-08	uniform
$meanMSE(g_0)$	0.126	0.126	0.127	0.126	0.122
$minMSE(g_0)$	0.033	0.032	0.032	0.032	0.034
$maxMSE(g_0)$	0.450	0.451	0.455	0.452	0.451
$meanMSE(f_1)$	0.0070	0.0071	0.0071	0.0070	0.0076
$minMSE(f_1)$	0.0009	0.0008	0.0008	0.0009	0.0010
$maxMSE(f_1)$	0.0209	0.0208	0.0212	0.0201	0.0213
$meanMSE(f_{spat})$	0.043	0.043	0.043	0.043	0.044
$minMSE(f_{spat})$	0.029	0.028	0.029	0.028	0.029
$maxMSE(f_{spat})$	0.071	0.068	0.070	0.070	0.070
$meanMSE(\gamma)$	0.0046	0.0046	0.0046	0.0046	0.0047
$minMSE(\gamma)$	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0
$maxMSE(\gamma)$	0.0297	0.0269	0.0282	0.0282	0.0281

Table 3: Posterior mean estimates and standard deviations for the fixed effects on time to CABG

effect	P–spline model, GRF		P–spline model, MRF		p.e.m., MRF	
sex	-0.041	(0.083)	-0.038	(0.085)	-0.037	(0.082)
dv2	1.479	(0.098)	1.485	(0.099)	1.495	(0.099)
dv3	1.793	(0.094)	1.804	(0.093)	1.817	(0.094)

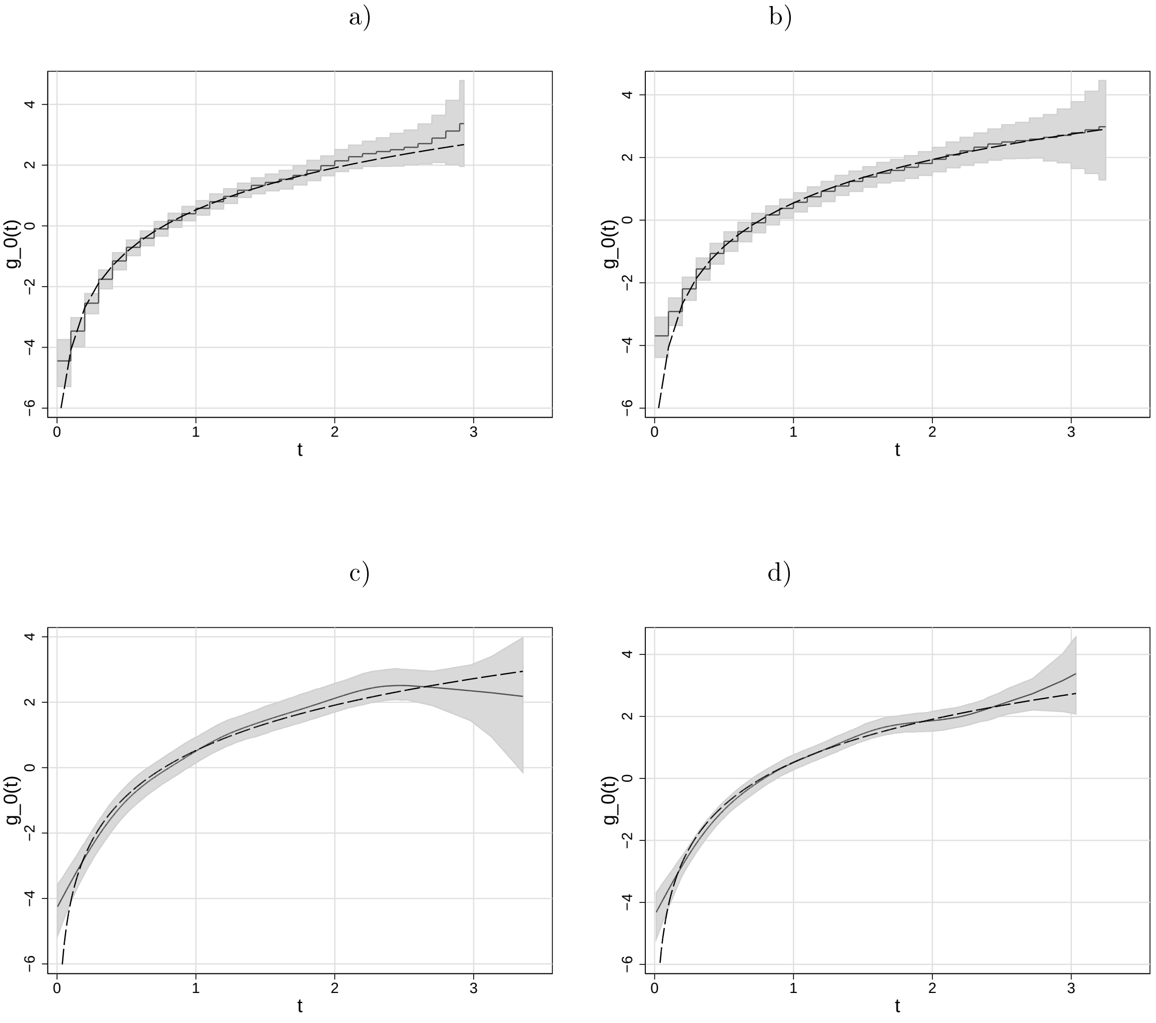


Figure 1: (log–)Baseline effects for the various model specifications; displayed are posterior mean estimates and 95% credible intervals of run r, with r chosen such that MSE_r is the median of $MSE_1, \ldots, MSE_{100}$ (solid line and grey shaded area), and the true (log–)baseline effect (dashed line). a) p.e.m., MRF, r=59, MSE=0.138 b) p.e.m., geospline, r=4, MSE=0.140 c) P–spline model, MRF, r=32, MSE=0.106 d) P–spline model, geospline, r=24, MSE=0.112

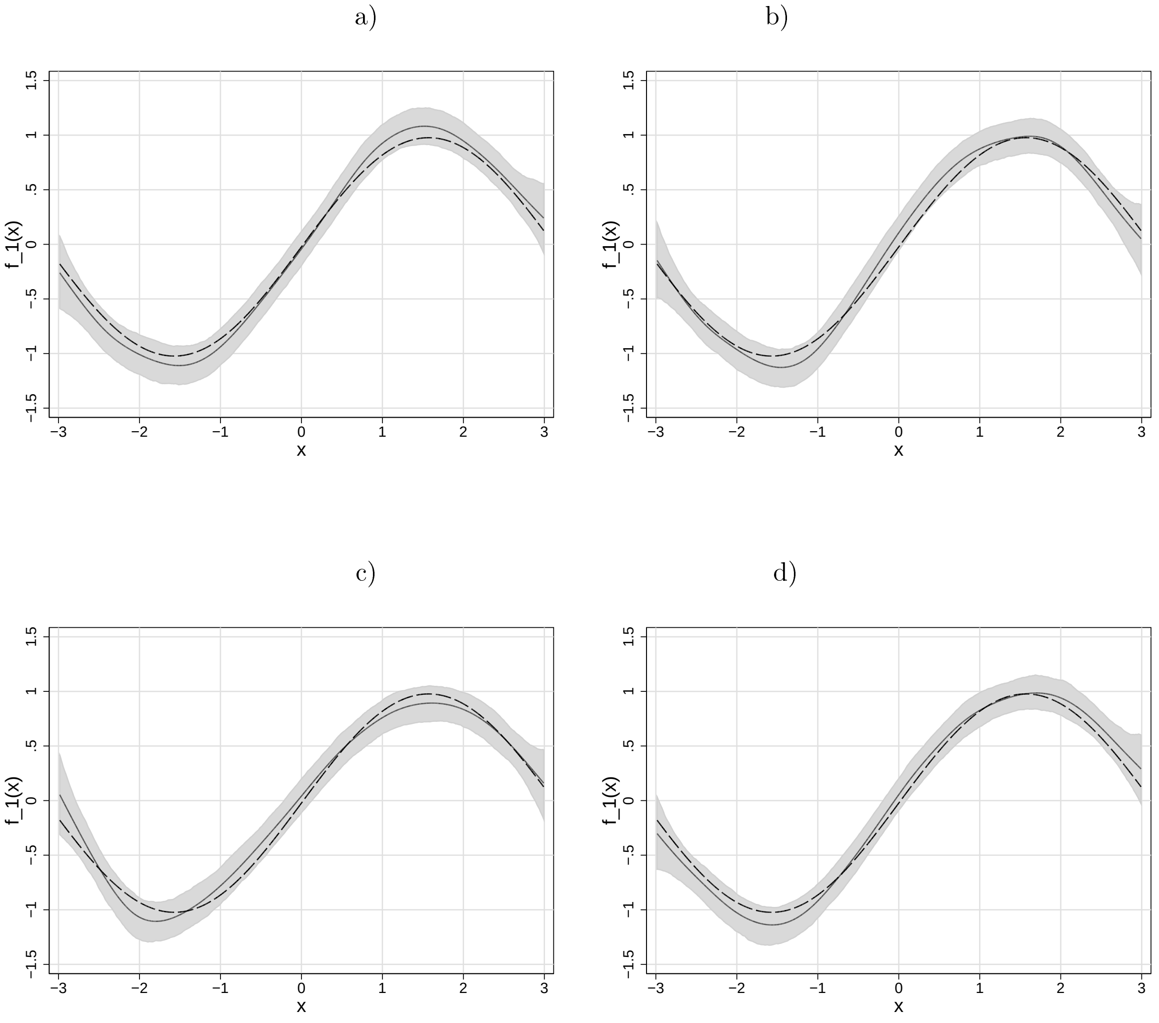


Figure 2: Nonparametric effects for the various model specifications; displayed are posterior mean estimates and 95% credible intervals of run r, with r chosen such that MSE_r is the median of $MSE_1, \ldots, MSE_{100}$ (solid line and grey shaded area), and the true function (dashed line). a) p.e.m., MRF, r=9, MSE=0.0059 b) p.e.m., geospline, r=70, MSE=0.0056 c) P–spline model, MRF, r=26, MSE=0.0061 d) P–spline model, geospline, r=74, MSE=0.0057

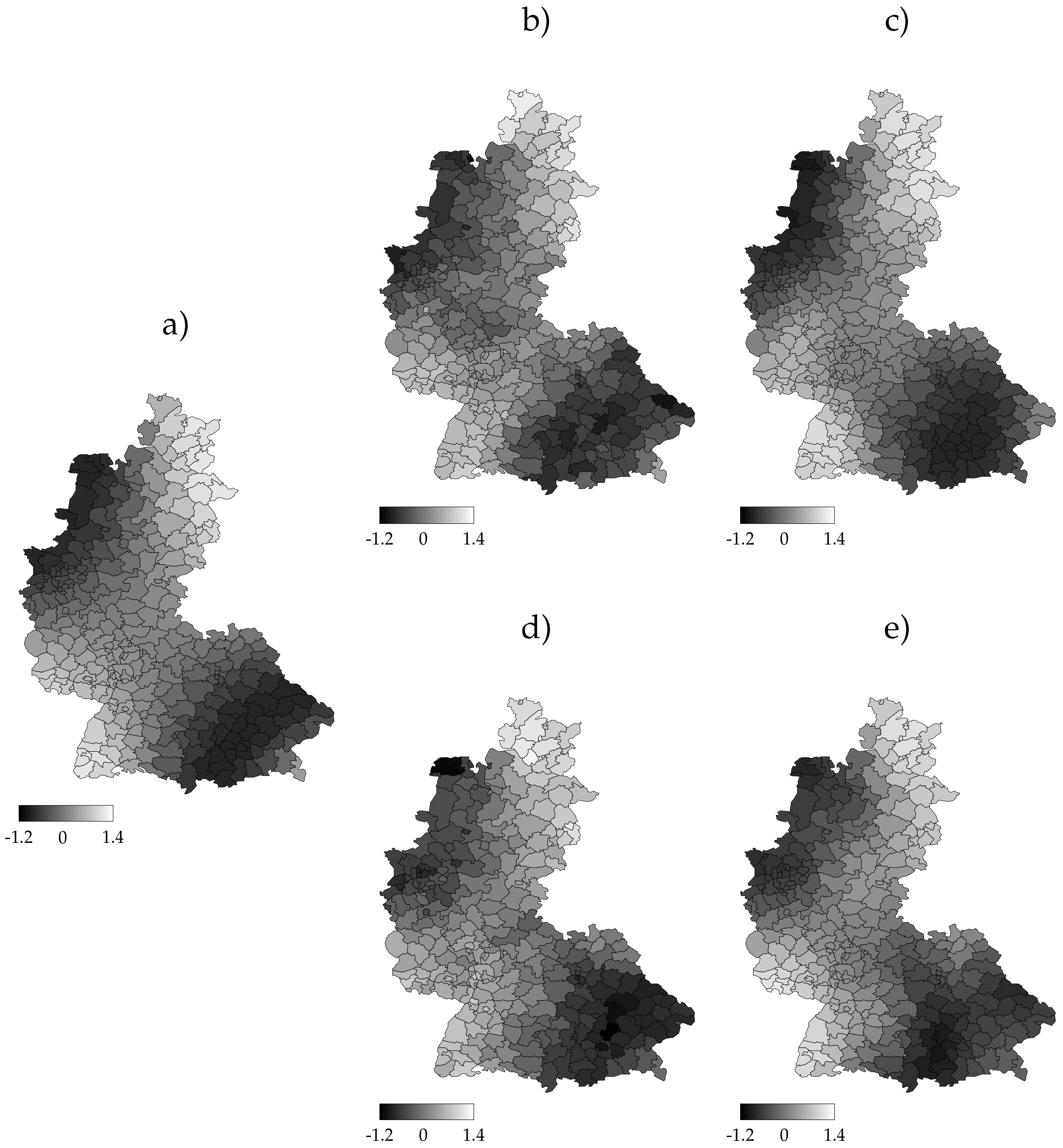


Figure 3: Spatial effects for the various model specifications; displayed are posterior mean estimates of run r, with r chosen such that MSE_r is the median of $MSE_1, \ldots, MSE_{100}$ a) true function b) p.e.m., MRF, r=65, MSE=0.041 c) p.e.m., geospline, r=83, MSE=0.021 d) P–spline model, MRF, r=67, MSE=0.042 e) P–spline model, geospline, r=22, MSE=0.021

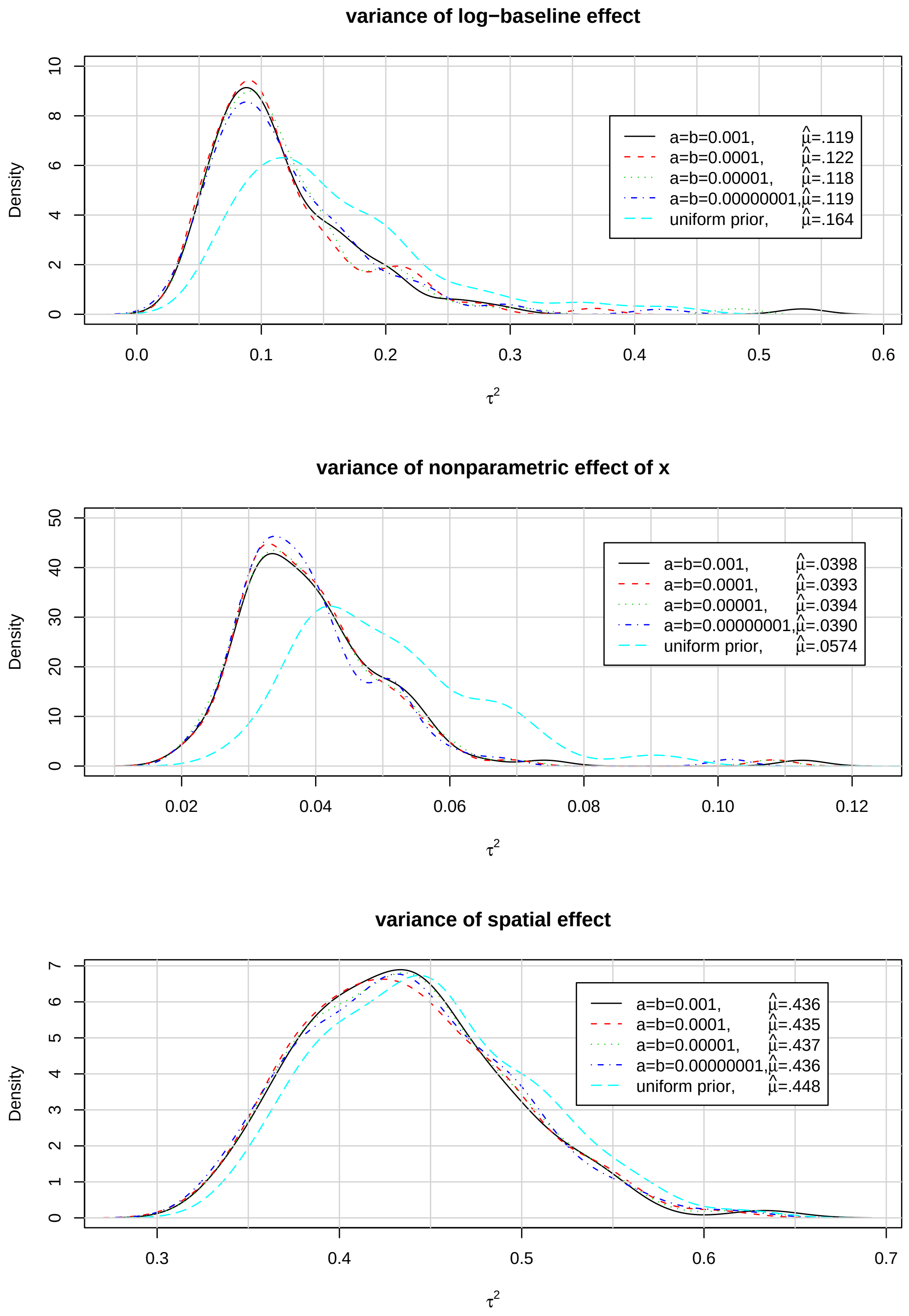
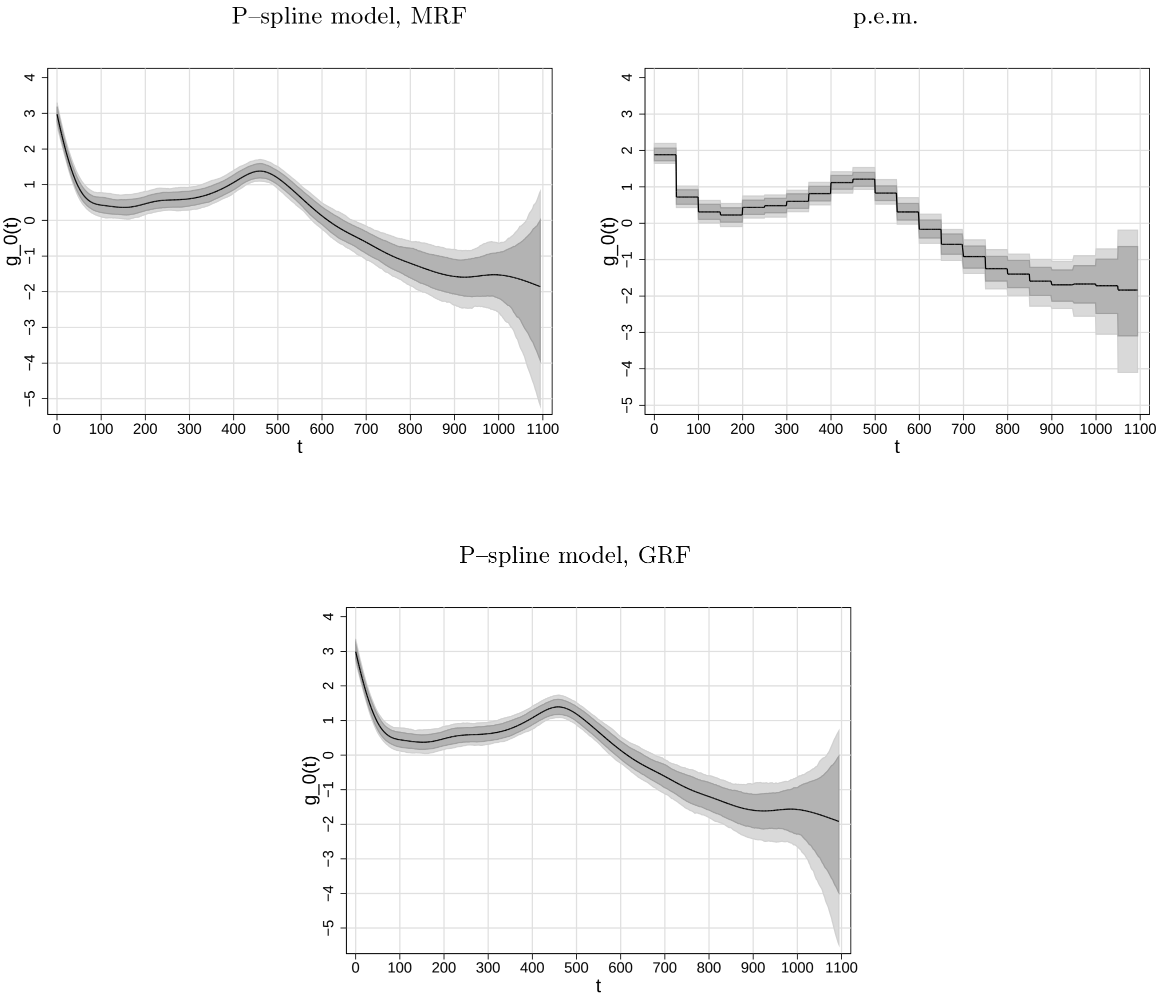


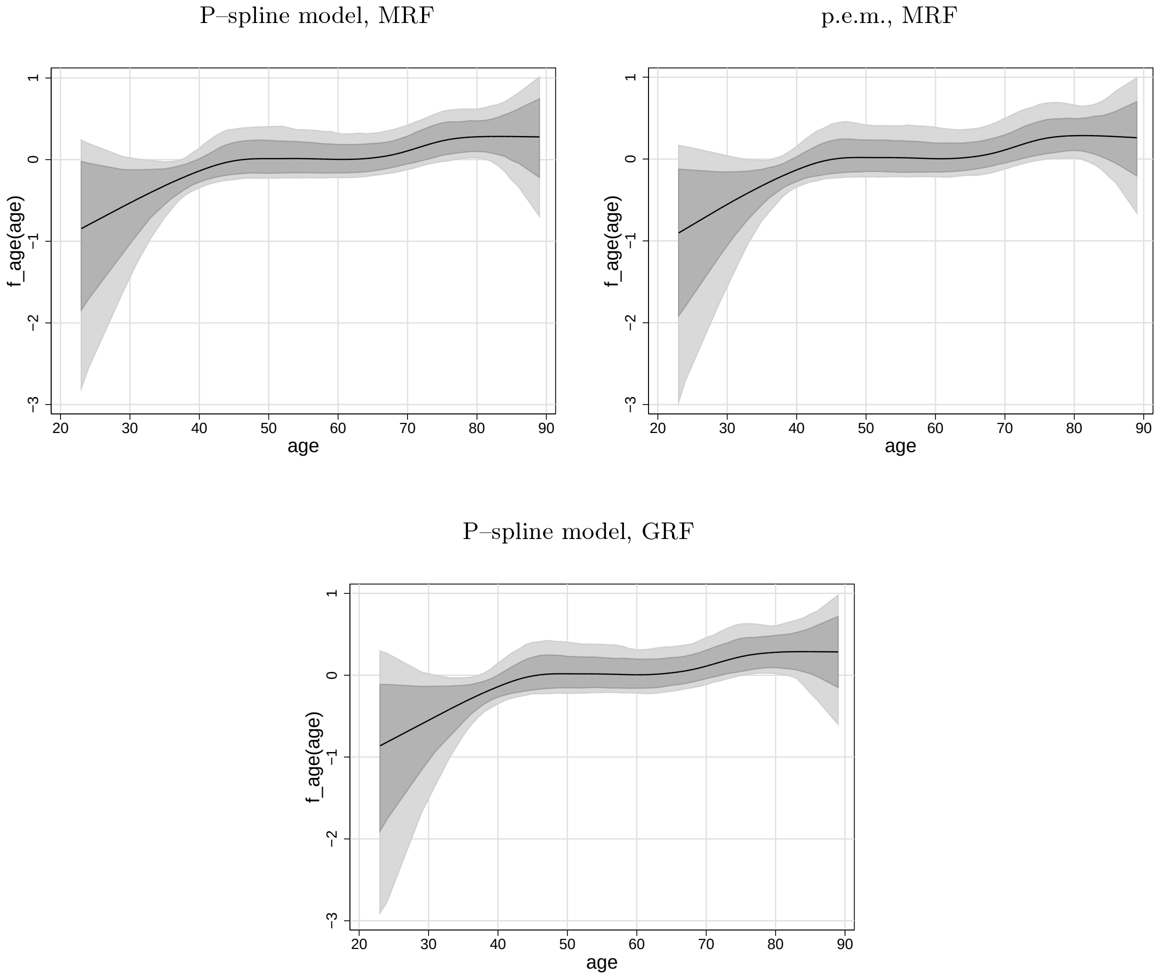
Figure 4: Kernel density estimates based on $\widehat{\tau_j^2}^{(r)}, r = 1, \dots, 100$ for $j = 0, 1$ and *spat*, respectively. $\left(\hat{\mu} = \frac{1}{100}\sum_{r=1}^{100}(\hat{\tau_j^2})^{(r)}, j = 0, 1, spat\right)$


P–spline model, MRF

p.e.m.

P–spline model, GRF

Figure 5: Posterior mean estimate for the (log-)baseline effect on time to CABG and 80% and 95% credible intervals


P–spline model, MRF

p.e.m., MRF

P–spline model, GRF

Figure 6: Posterior mean estimates of the effect of age on time to CABG and 80% and 95% credible intervals

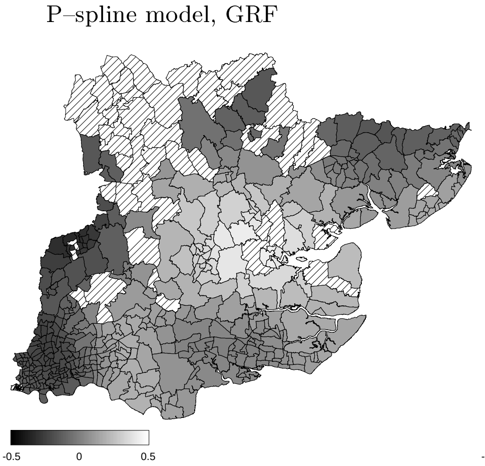


Figure 7: Posterior mean estimates of the structured spatial effect on time to CABG; the estimates under the p.e.m. with a MRF prior are visually indistinguishable from those of the P–spline model with MRF prior, and are therefore not shown here

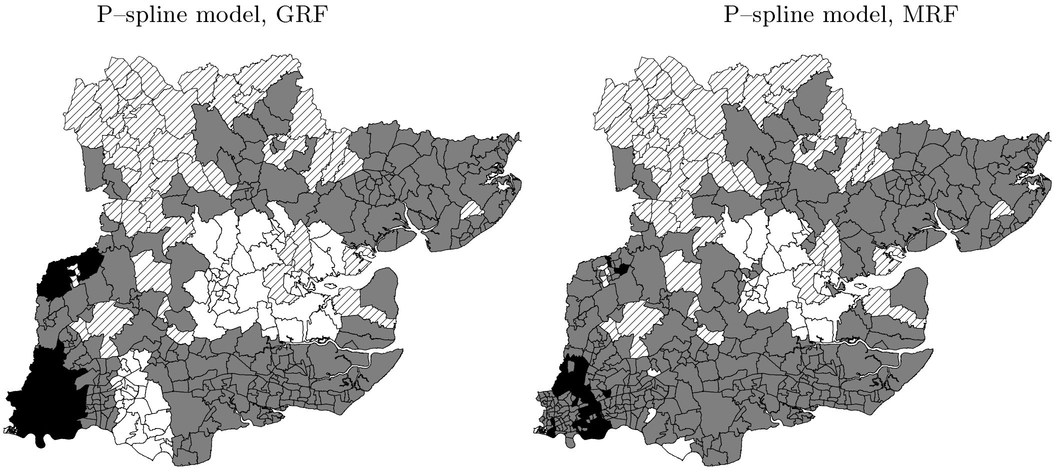


Figure 8: Posterior probabilities of the structured spatial effects, with white (black) areas indicating that at least 80% of the sample estimates were positive (negative)

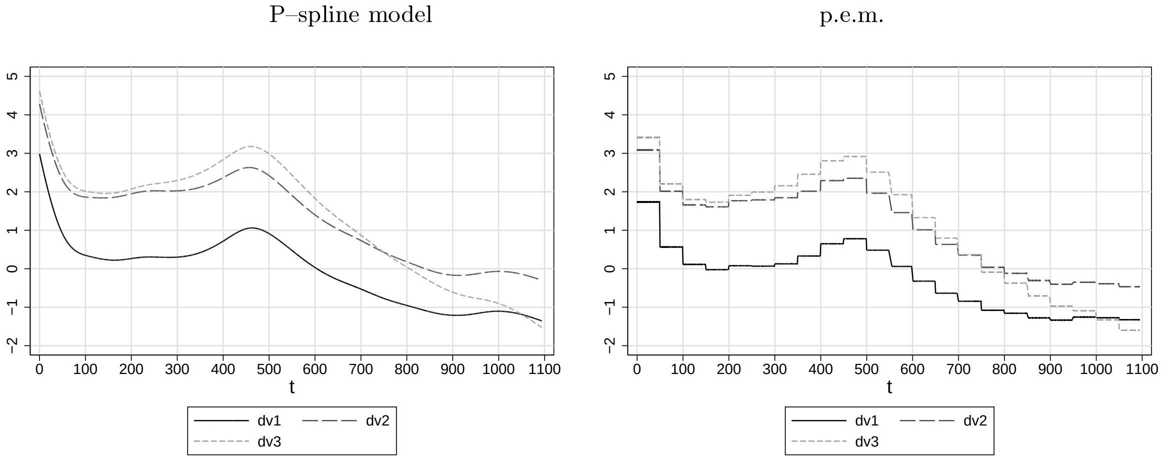


Figure 9: (log–)baseline effects on time to CABG: posterior mean estimates for 1 diseased vessel ($dv1$), 2 diseased vessels ($dv2$) and 3 diseased vessels ($dv3$)

Appendix: Propriety of posteriors in mixed models

We first consider Gaussian linear mixed models

$$\boldsymbol{y} = \boldsymbol{V\gamma} + \boldsymbol{Z}_1\boldsymbol{\beta}_1 + ... + \boldsymbol{Z}_m\boldsymbol{\beta}_m + \boldsymbol{\varepsilon} \tag{17}$$

for observations $\boldsymbol{y} = (y_1, ..., y_n)'$, with an additive predictor, and a Gaussian error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n) \sim N(\boldsymbol{0}, \tau_0^2\boldsymbol{I})$. For identifiability reasons, the predictor must not contain individual–specific uncorrelated random effects in addition to $\boldsymbol{\varepsilon}$. The prior assumptions for the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_j$, $j = 1, ..., m$, are the same as in Section 2, i.e., a flat prior

$$p(\boldsymbol{\gamma}) \equiv 1 \tag{18}$$

for the vector $\boldsymbol{\gamma}$ of 'fixed' effects, and

$$p(\boldsymbol{\beta_j}) \propto \tau_j^{-r_j} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right), \tag{19}$$

with $d_j = \dim(\boldsymbol{\beta}_j)$ and $r_j = \text{rank}(\boldsymbol{K}_j)$. For $r_j < d_j$, the prior for $\boldsymbol{\beta_j}$ is partially improper.

Priors for hyperparameters $\boldsymbol{\tau}^2 = (\tau_0^2, ..., \tau_m^2)'$ are $p(\boldsymbol{\tau}^2) = \prod_{j=0}^m p(\tau_j^2)$. An important special case are inverse Gamma priors

$$p(\tau_j^2) \propto \frac{1}{(\tau_j^2)^{a_j+1}} \exp\left(-\frac{b_j}{\tau_j^2}\right), \tag{20}$$

which are proper for $a_j > 0$, $b_j > 0$.

Defining $\boldsymbol{Z} = (\boldsymbol{Z}_1, ..., \boldsymbol{Z}_m)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_m')'$, the model (17) is

$$\boldsymbol{y} = \boldsymbol{V\gamma} + \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}.$$

Further, with $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{Z})$, let $(\hat{\gamma}', \hat{\beta}')' = (\boldsymbol{X'X})^{-}\boldsymbol{X'y}$ be the least squares estimator, and

$$SSE = \boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-}\boldsymbol{X'})\boldsymbol{y}$$

be the sum of squared errors, which is invariant for any choice of the generalized inverse $(\boldsymbol{X'X})^{-}$.

Lemma A1

Consider the Gaussian mixed model defined by (17), (18) and (19), and assume that the following conditions hold:

(i) rank($\boldsymbol{V}$)=p, rank($\boldsymbol{Z'RZ}+\boldsymbol{K}$)=$d$

where $p=\dim(\boldsymbol{\gamma})$, $d=d_1+...+d_m=\dim(\boldsymbol{\beta})$, $\boldsymbol{K}=\text{diag}(\boldsymbol{K_1},...,\boldsymbol{K_m})$, $\boldsymbol{R}=\boldsymbol{I}-\boldsymbol{V}(\boldsymbol{V'V})^{-1}\boldsymbol{V'}$.

(ii) the priors $p(\tau_j^2)$, $j=1,...,m$ are proper, and

$$\int p(\tau_0^2)\tau_0^{-(n-p-(d-r))}\exp\left(-\frac{SSE}{2\tau_0^2}\right)d\tau_0^2<\infty,$$

where $r=r_1+...+r_m$.

Then the posterior distribution $p(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\tau^2}\mid\boldsymbol{y})$ is proper.

Corollary A1

For a linear mixed model (17) with prior (18) and (20), the posterior $p(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\tau^2}\mid\boldsymbol{y})$ is proper if condition (i) of Lemma A1 and

$$a_j>0,\, b_j>0,\quad j=1,...,m,$$

$$n-p-(d-r)+2a_0>0,\quad SSE+2b_0>0$$

hold.

Remark: Condition (i) of Lemma A1 is equivalent to

$$\text{rank}\begin{pmatrix}\boldsymbol{V'V} & \boldsymbol{V'Z}\\ \boldsymbol{Z'V} & \boldsymbol{Z'Z}+\boldsymbol{K}\end{pmatrix}=p+d.$$

Proof of Lemma A1 and Corollary A1

The proof extends arguments in Sun, Tsutakawa and Speckman (1999), see also Speckman and Sun (2003), using a theorem on eigenvalues in Magnus and Neudecker (1991). From the model assumptions we have

$$p(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\tau^2}\mid\boldsymbol{y})\propto\tau_0^{-n}\tau_1^{-r_1}\cdot...\cdot\tau_m^{-r_m}\cdot\exp\left\{-\frac{(\boldsymbol{y}-\boldsymbol{V\gamma}-\boldsymbol{Z\beta})'(\boldsymbol{y}-\boldsymbol{V\gamma}-\boldsymbol{Z\beta})}{2\tau_0^2}-\sum_{j=1}^{m}\frac{\boldsymbol{\beta_j'K_j\beta_j}}{2\tau_j^2}\right\}p(\boldsymbol{\tau^2})$$

Following Sun et al. (1999), we rewrite

$$(\boldsymbol{y}-\boldsymbol{V\gamma}-\boldsymbol{Z\beta})'(\boldsymbol{y}-\boldsymbol{V\gamma}-\boldsymbol{Z\beta})=SSE+(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}-\boldsymbol{c_1})'\boldsymbol{V'V}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}-\boldsymbol{c_1})+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{Z'RZ}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}),$$

where $\boldsymbol{c_1} = \boldsymbol{(V'V)^{-1}V'Z(\hat{\beta} - \beta)}$.

Integrating the right hand side with respect to $\boldsymbol{\gamma}$, we get

$$\int p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\tau^2} \mid \boldsymbol{y}) d\boldsymbol{\gamma} \propto \frac{(2\pi)^{p/2} \mid \boldsymbol{V'V} \mid^{-1/2}}{\tau_0^{n-p} \prod_{j=1}^{m} \tau_j^{r_j}} \cdot \exp\left(-\frac{SSE}{2\tau_0^2} - \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{Z'RZ}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{2\tau_0^2} - \frac{1}{2}\boldsymbol{\beta' K_{\tau^2} \beta} \right) p(\boldsymbol{\tau^2}),$$

where $\boldsymbol{K}_{\tau^2} = \text{diag}\left(\boldsymbol{K_1}/\tau_1^2, ..., \boldsymbol{K_m}/\tau_m^2\right)$.

Define $\boldsymbol{R_1} = \tau_0^{-2}\boldsymbol{Z'RZ} + \boldsymbol{K_{\tau^2}}$. Then for any $\tau_j^2 > 0$, $j = 0, ..., m$, $\boldsymbol{R_1^{-1}}$ exists by assumption (i) of Lemma A1. Set

$$\begin{aligned} \boldsymbol{c_2} &= \tau_0^{-2}\boldsymbol{R_1^{-1}Z'RZ\hat{\beta}} \\ \boldsymbol{R_2} &= \boldsymbol{Z'RZ} - \tau_0^{-2}\boldsymbol{Z'RZR_1^{-1}Z'RZ}. \end{aligned}$$

Then

$$\frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\boldsymbol{Z'RZ}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{\tau_0^2} + \boldsymbol{\beta' K_{\tau^2} \beta} = (\boldsymbol{\beta} - \boldsymbol{c_2})'\boldsymbol{R_1}(\boldsymbol{\beta} - \boldsymbol{c_2}) + \frac{\hat{\boldsymbol{\beta}}'\boldsymbol{R_2}\hat{\boldsymbol{\beta}}}{\tau_0^2}.$$

Integrating out $\boldsymbol{\beta}$, we get

$$\int p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\tau^2} \mid \boldsymbol{y}) d\boldsymbol{\gamma} d\boldsymbol{\beta} \propto \frac{(2\pi)^{(p+r)/2} \mid \boldsymbol{V'V} \mid^{-\frac{1}{2}} \mid \boldsymbol{R_1} \mid^{-\frac{1}{2}}}{\tau_0^{n-p} \prod_{j=1}^{m} \tau_j^{r_j}} \cdot \exp\left\{ -\frac{SSE + \hat{\boldsymbol{\beta}}'\boldsymbol{R_2}\hat{\boldsymbol{\beta}}}{2\tau_0^2} \right\} p(\boldsymbol{\tau^2}). \tag{21}$$

Since $\boldsymbol{R_2}$ is nonnegative definite, the second factor is bounded by $\exp\left\{-SSE/(2\tau_0^2)\right\}$.

For an upper bound of the first factor, we first derive a lower bound for $\mid \boldsymbol{R_1} \mid$, applying Theorem 9 in Magnus and Neudecker (1991, ch. 11, p. 208) to the eigenvalues of

$$\boldsymbol{R_1} = \tau_0^{-2}\boldsymbol{Z'RZ} + \boldsymbol{K_{\tau^2}}.$$

Note that the $d - r$ smallest eigenvalues of $\boldsymbol{K}$ and $\boldsymbol{K_{\tau^2}}$ are zero, while the eigenvalues $\lambda_l(\boldsymbol{K_{\tau^2}})$, $l = d - r + 1, ..., r$, are positive. Application of the theorem to the positive eigenvalues of $\boldsymbol{R_1}$ gives

$$\lambda_l(\tau_0^{-2}\boldsymbol{Z'RZ} + \boldsymbol{K_{\tau^2}}) \geq \lambda_l(\boldsymbol{K_{\tau^2}}) = \lambda(\boldsymbol{K_j})\tau_j^{-2} \geq \lambda_j \tau_j^{-2},$$

where $\lambda(\boldsymbol{K_j})$ is a positive eigenvalue of one of the precision matrices $\boldsymbol{K_j}$ and $\lambda_j > 0$ is the smallest positive eigenvalue of $\boldsymbol{K_j}$.

Application of the theorem to the eigenvalues $\lambda_l(\boldsymbol{K_{\tau^2}}) = 0$, $l = 1, ..., d - r$, of $\boldsymbol{K_{\tau^2}}$ gives

$$\lambda_l(\tau_0^{-2}\boldsymbol{Z'RZ} + \boldsymbol{K_{\tau^2}}) \geq \lambda_l(\tau_0^{-2}\boldsymbol{Z'RZ}) \geq \tau_0^{-2}\lambda_0,$$

where $\lambda_0 > 0$ is the smallest eigenvalue of $\boldsymbol{Z'RZ}$.

Taken together, we get

$$\mid \boldsymbol{R_1} \mid = \prod_l \lambda_l(\boldsymbol{R_1}) \geq \tau_0^{-2(d-r)} \prod_{j=1}^{m} \tau_j^{-2r_j} \cdot L$$

where $L = \lambda^{d-r} \prod_{j=1}^{m} \lambda_j^{r_j} > 0$, and

$$\mid \boldsymbol{R_1} \mid^{-1/2} \leq \frac{1}{L^{1/2}} \tau_0^{d-r} \prod_{j=1}^{m} \tau_j^{r_j}.$$

Inserting in (21), we obtain

$$\int p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta} \mid \boldsymbol{y}) d\boldsymbol{\gamma} d\boldsymbol{\beta} \leq C \frac{1}{\tau_0^{n-p-(d-r)}} \cdot \exp\left\{-\frac{SSE}{2\tau_0^2}\right\} \prod_{j=0}^{m} p(\tau_j^2).$$

Thus, if condition (ii) in Lemma A1 holds, the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\tau^2} \mid \boldsymbol{y})$ is proper.

Corollary A1 follows immediately, because then

$$\frac{1}{\tau_0^{n-p-(d-r)}} \exp\left\{-\frac{SSE}{2\tau_0^2}\right\} \frac{1}{(\tau_0^2)^{a_0+1}} \exp\left\{-\frac{b_0}{\tau_0^2}\right\} = \frac{1}{\tau_0^{n-p-(d-r)+2(a_0+1)}} \exp\left\{-\frac{SSE/2 + b_0}{\tau_0^2}\right\}.$$

We recognize a proper inverse Gamma density for $(n-p-(d-r))/2 + a_0 > 0$ and $SSE/2 + b_0 > 0$.

Propriety of the posterior for generalized (geo–) additive models

The following Lemma A2 gives sufficient conditions for the propriety of the posterior in generalized linear and additive mixed models. The lemma and its proof rest heavily on Theorem 4 in Sun et al. (1998), who considered models with densities $f_i(y_i \mid \eta_i)$ for the observations y_i given a predictor η_i and predictors $\boldsymbol{\eta} = (\eta_1, ..., \eta_i, ..., \eta_n)$ given by

$$\boldsymbol{\eta} = \boldsymbol{V\gamma} + \boldsymbol{Z_1\beta_1} + \boldsymbol{\varepsilon},$$

with partially improper prior for $\boldsymbol{\beta_1}$, and individual–specific random effects $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_i, ..., \varepsilon_n)' \sim N(\boldsymbol{0}, \tau_0^2 \boldsymbol{I})$. We extend their theorem in two directions: First, we allow for several random effects with different degree and type of smoothness priors, and, second, we do not necessarily assume that individual–specific random effects ε_i are included in the predictor.

We consider models with predictor

$$\boldsymbol{\eta} = \boldsymbol{V\gamma} + \boldsymbol{Z_1\beta_1} + ... + \boldsymbol{Z_m\beta_m} + \boldsymbol{Z_0\beta_0}, \tag{22}$$

where $\boldsymbol{\gamma}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m$ have priors as in (18) and (19). The term $\boldsymbol{Z}_0\boldsymbol{\beta}_0$ represents a random effect with a $n \times d_0$ design matrix $\boldsymbol{Z}_0$, with $\mathrm{rank}(\boldsymbol{Z}_0) = d_0 = \dim(\boldsymbol{\beta}_0)$, and a (possibly partially improper) prior

$$p(\boldsymbol{\beta}_0) \propto \tau_0^{-r_0} \exp\left(-\frac{1}{2\tau_0^2}\boldsymbol{\beta}_0'\boldsymbol{K}_0\boldsymbol{\beta}_0\right), \tag{23}$$

with $r_0 = \mathrm{rank}(\boldsymbol{K}_0)$, such that

$$d_0 \geq d_j, \quad r_0 \geq r_j, \quad j = 1, ..., m.$$

Setting $\boldsymbol{Z}_0 = \boldsymbol{I}$, $\boldsymbol{\beta}_0 = \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \tau_0^2\boldsymbol{I})$, the predictor (22) also covers the case of individual–specific random effects $\boldsymbol{Z}_0\boldsymbol{\beta}_0 = \boldsymbol{\varepsilon}$. In geoadditive models $\boldsymbol{Z}_0\boldsymbol{\beta}_0$ will usually represent a spatial effect with a MRF or kriging prior, or an unstructured spatial effect.

Lemma A2

Consider a generalized linear mixed model with observation densities $f_i(y_i \mid \eta_i)$, predictor (22), and priors (18), (19), (23). Suppose that (after a reordering of observations)

$$(*) \qquad \int f_i(y_i \mid \eta_i) d\eta_i < \infty$$

holds for observations $i = 1, ..., n^*$, and

$$(**) \qquad f_i(y_i \mid \eta_i) \leq M, \quad i = n^* + 1, ..., n$$

holds for the remaining observations.

Denote the corresponding submatrices of $\boldsymbol{V}$, $\boldsymbol{Z}$ and $\boldsymbol{Z}_0$ by $\boldsymbol{V}^*$, $\boldsymbol{Z}^* = (\boldsymbol{Z}_1^*, ..., \boldsymbol{Z}_m^*), \boldsymbol{Z}_0^*$, and assume:

(iii) $\mathrm{rank}(\boldsymbol{Z}_0^*) = d_0$,

the rank conditions (i) in Lemma A1 hold for $\boldsymbol{V}^*$, $\boldsymbol{Z}^*$,

condition (ii) in Lemma A1 holds with r_0 replacing n and SSE replaced by SSE^*.

Then the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \tau_0^2, ..., \tau_m^2 \mid \boldsymbol{y})$ is proper.

The following corollary is easier to check.

Corollary A2

Assume that conditions $(*)$, $(**)$ and the rank conditions for $\boldsymbol{V}^*$, $\boldsymbol{Z}^*$, $\boldsymbol{Z_0}^*$ in Lemma A2 hold, and that

$$r_0 - p - (d - r) > 0$$

with $d = d_0 + ... + d_m$, $r = r_0 + ... + r_m$, and

$$a_j > 0, \quad b_j > 0, \quad j = 0, ..., m$$

hold for the inverse Gamma priors (23).

Then the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \tau_0^2, ..., \tau_m^2 \mid \boldsymbol{y})$ is proper.

Proofs: We consider first the simpler case of individual–specific random effects $\boldsymbol{\beta_0} {\equiv} \boldsymbol{\varepsilon} {\sim}\ N(\mathbf{0}, \tau_0^2 \boldsymbol{I})$. Using the one–to–one relation $\boldsymbol{\eta}{=}\ \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ between $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$, we consider propriety of $p(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tau_0^2, \boldsymbol{\tau}^2 \mid \boldsymbol{y})$ instead of $p(\boldsymbol{\varepsilon}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tau_0^2, \boldsymbol{\tau}^2 \mid \boldsymbol{y})$. Proceeding as in Sun et al. (1998), one starts from

$$p(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tau_0^2, \boldsymbol{\tau}^2 \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) p(\boldsymbol{\beta}) p(\tau_0^2) p(\boldsymbol{\tau}^2).$$

Using $(**)$ and integrating out $\boldsymbol{\eta}^{**} = (\eta_{n^*+1}, ..., \eta_n)$, one arrives at

$$\begin{aligned} p(\boldsymbol{\eta}^*, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tau_0^2, \boldsymbol{\tau}^2 \mid \boldsymbol{y}) \quad &\propto \quad \prod_{i=1}^{n^*} f_i(y_i \mid \eta_i) \{p(\boldsymbol{\eta}^* \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) p(\boldsymbol{\beta}) p(\tau_0^2) p(\boldsymbol{\tau}^2)\} \\ &\propto \quad \prod_{i=1}^{n^*} f_i(y_i \mid \eta_i) \{p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \tau_0^2, \boldsymbol{\tau}^2 \mid \boldsymbol{\eta}^*)\}. \end{aligned}$$

Applying Lemma A1 (or Corollary A1) to

$$\boldsymbol{\eta}^* = \boldsymbol{V}^* \boldsymbol{\gamma} + \boldsymbol{Z}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}^*, \quad \boldsymbol{\varepsilon}^* \sim N(\mathbf{0}, \tau_0^2 \boldsymbol{I}),$$

gives

$$p(\boldsymbol{\eta}^* \mid \boldsymbol{y}) \propto \prod_{i=1}^{n^*} f_i(y_i \mid \eta_i),$$

and propriety follows from $(*)$.

For the general case $\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{Z}_0 \boldsymbol{\beta}_0$, with prior (23) for $\boldsymbol{\beta}_0$, we first decompose $\boldsymbol{\beta}_0$ into a $(d_0 - r_0)$–dimensional subvector $\boldsymbol{\beta}_0^{fl}$ with flat prior $p(\boldsymbol{\beta}_0^{fl}) \equiv 1$ and a r_0–dimensional subvector $\boldsymbol{\beta}_0^{pr}$ with a proper prior $\boldsymbol{\beta}_0^{pr} \sim N(\mathbf{0}, \tau_0^2 \boldsymbol{I})$:

$$\boldsymbol{\beta}_0 = \boldsymbol{Z}_0^{fl} \boldsymbol{\beta}_0^{fl} + \boldsymbol{Z}_0^{pr} \boldsymbol{\beta}_0^{pr},$$

where the $d_0 \times (d_0 - r_0)$ matrix $\boldsymbol{Z}_0^{fl}$ contains a basis of the nullspace of $\boldsymbol{K}_0$. The matrix $\boldsymbol{Z}_0^{fl}$ is the identity vector $\mathbb{1}$ for P–splines with first–order random walk prior, Markov–random fields and 2d–P–splines with

MRF prior for the coefficients. For P–splines with second–order random walk prior it is a two column matrix whose first column is the identity vector and the second column is composed of the (equidistant) knots of the spline.

The $d_0 \times r_0$ matrix $\boldsymbol{Z}_0^{pr}$ is given by

$$\boldsymbol{Z}_0^{pr} = \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{L})^{-1},$$

where $\boldsymbol{L} = \boldsymbol{S'\Lambda^{1/2}}$ is obtained from the spectral decomposition $\boldsymbol{K}_0 = \boldsymbol{S\Lambda S'}$ of $\boldsymbol{K}_0$. It follows that

$$\boldsymbol{\beta}_0^{pr} \sim N(\boldsymbol{0}, \tau_0^2 \boldsymbol{I}).$$

Defining $\tilde{\boldsymbol{V}} = (\boldsymbol{V}, \boldsymbol{Z}_0 \boldsymbol{Z}_0^{fl})$, $\tilde{\gamma}' = (\boldsymbol{\gamma}, \boldsymbol{\beta}_0^{fl})'$, $\tilde{\boldsymbol{Z}}_0 = \boldsymbol{Z}_0 \boldsymbol{Z}_0^{pr}$, we can rewrite the predictor as

$$\boldsymbol{\eta} = \tilde{\boldsymbol{V}}\tilde{\gamma} + \boldsymbol{Z}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}_0 \boldsymbol{\beta}_0^{pr}.$$

For identifiability reasons, the columns of $\boldsymbol{Z}_0 \boldsymbol{Z}_0^{fl}$ are not contained in the $(d_0 - r_0)$ column space of $\boldsymbol{V}$, so that $\text{rank}(\tilde{\boldsymbol{V}}) = p + (d_0 - r_0)$. Defining $\boldsymbol{\varepsilon}_0 = \tilde{\boldsymbol{Z}}_0 \boldsymbol{\beta}_0^{pr}$, we have an additive mixed model

$$\boldsymbol{\eta} = \tilde{\boldsymbol{V}}\tilde{\gamma} + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0 \tag{24}$$

for the predictor $\boldsymbol{\eta}$, with singular covariance matrix $\text{cov}(\boldsymbol{\varepsilon}_0) = \tilde{\boldsymbol{Z}}_0 \tilde{\boldsymbol{Z}}_0' \tau_0^2$ of the 'error term' $\boldsymbol{\varepsilon}_0$. Let

$$\tilde{\boldsymbol{S}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{S}}' = \tilde{\boldsymbol{Z}}_0 \tilde{\boldsymbol{Z}}_0'$$

be the spectral decomposition of $\tilde{\boldsymbol{Z}}_0 \tilde{\boldsymbol{Z}}_0'$, with $\tilde{\boldsymbol{\Lambda}} = \text{diag}(\lambda_1, ..., \lambda_{r_0})$ containing the r_0 positive eigenvalues, and set

$$\boldsymbol{T} = \tilde{\boldsymbol{\Lambda}}^{-1/2} \tilde{\boldsymbol{S}}'.$$

Multiplying equation (24) by $\boldsymbol{T}$, we obtain the reduced model

$$\tilde{\boldsymbol{\eta}} = \boldsymbol{T}\tilde{\boldsymbol{V}}\tilde{\gamma} + \boldsymbol{T}\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \tau_0^2 \boldsymbol{I}),$$

where $\tilde{\boldsymbol{\eta}} = \boldsymbol{T}\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon} = \boldsymbol{T}\boldsymbol{\varepsilon}_0$ have dimension r_0.

Altogether, we obtain a (linear) one–to–one transformation between $\boldsymbol{\beta}_0^{pr}$ and $\tilde{\boldsymbol{\eta}}$, and proving propriety of $p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\tau}^2, \tau_0^2 \mid \boldsymbol{y})$ is equivalent to proving propriety of $p(\tilde{\boldsymbol{\eta}}, \tilde{\gamma}, \boldsymbol{\beta}, \boldsymbol{\tau^2}, \tau_0^2 \mid \boldsymbol{y})$.

Thus, we can repeat the arguments of the first part of the proof, replacing $\boldsymbol{\eta}$ by $\tilde{\boldsymbol{\eta}}$, $\boldsymbol{V}$ by $\boldsymbol{T}\tilde{\boldsymbol{V}}$, $\boldsymbol{Z}$ by $\boldsymbol{T}\boldsymbol{Z}$, and n by r_0.

From Magnus, Neudecker (1991, p. 273) it follows that

$$\operatorname{rank}(\boldsymbol{T}\tilde{\boldsymbol{V}}) = \operatorname{rank}(\tilde{\boldsymbol{V}}) = p + d_0 - r_0,$$

$$\operatorname{rank}(\boldsymbol{T}\boldsymbol{Z}_j) = \operatorname{rank}(\boldsymbol{Z}_j) = r_j.$$

Applying now Lemma A1 (or Corollary A1) to the model for $\tilde{\boldsymbol{\eta}}$, we obtain Lemma A2 and Corollary A2.