Kneib, Fahrmeir:

A mixed model approach for structured hazard regression

Sonderforschungsbereich 386, Paper 400 (2004)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

A mixed model approach for structured hazard regression

Thomas Kneib and Ludwig Fahrmeir

Department of Statistics, University of Munich.

October 27, 2004

Abstract

The classical Cox proportional hazards model is a benchmark approach to analyze continuous survival times in the presence of covariate information. In a number of applications, there is a need to relax one or more of its inherent assumptions, such as linearity of the predictor or the proportional hazards property. Also, one is often interested in jointly estimating the baseline hazard together with covariate effects or one may wish to add a spatial component for spatially correlated survival data. We propose an extended Cox model, where the (log-)baseline hazard is weakly parameterized using penalized splines and the usual linear predictor is replaced by a structured additive predictor incorporating nonlinear effects of continuous covariates and further time scales, spatial effects, frailty components, and more complex interactions. Inclusion of time-varying coefficients leads to models that relax the proportional hazards assumption. Nonlinear and time-varying effects are modelled through penalized splines, and spatial components are treated as correlated random effects following either a Markov random field or a stationary Gaussian random field. All model components, including smoothing parameters, are specified within a unified framework and are estimated simultaneously based on mixed model methodology. The estimation procedure for such general mixed hazard regression models is derived using penalized likelihood for regression coefficients and (approximate) marginal likelihood for smoothing parameters. Performance of the proposed method is studied through simulation and an application to leukemia survival data in Northwest England.

Key words: extended Cox model, structured hazard regression, mixed models, marginal likelihood

1 Introduction

A standard tool for analyzing the impact of covariates v on continuous survival times is the Cox proportional hazards model (Cox 1972) where the multiplicative structure

$$\lambda(t, v) = \lambda_0(t) \exp(v'\gamma) \tag{1}$$

is assumed for the hazard rate and γ is a vector of regression coefficients. The baseline hazard rate $\lambda_0(t)$ remains unspecified and estimation of the regression coefficients is based

on the partial likelihood. In a second step the baseline hazard can be approximated by a step function using Breslow's estimate. However, it is often desirable to estimate $\lambda_0(t)$ in a smooth way simultaneously with covariable effects, for example if we are interested in predicting survival times for new observations or if we are interested in analytic properties of the baseline. Furthermore the linear predictor in (1) is often not flexible enough to describe data situations of realistic complexity in an adequate way. Considering the data set on leukemia survival times, the effect of the age of a patient or other continuous covariates may supposed to be of an unknown nonlinear form. In addition, the data set contains information on the residence of the patient and survival times are likely to be spatially correlated. Further questions might be, whether there are interactions between two continuous or continuous and categorical covariates or whether some covariate effects are time-varying.

Non- and semiparametric Bayesian or related penalized likelihood approaches that can deal with these issues through extensions of the basic Cox model (1) have been suggested by several researchers. Fully Bayesian models, for estimating the baseline hazard rate and possibly time-varying covariate effects, are described in Ibrahim, Chen and Sinha (2001). Survival models which add a spatial component to the linear predictor in (1) have been developed recently. Li and Ryan (2002) model the spatial component through a stationary Gaussian random field. The baseline hazard rate, however, is treated as a nuisance parameter, and no procedure for estimating the spatial effects is provided. Henderson, Shimakura and Gorst (2002) propose a Cox model with gamma frailties, where the means follow either a Markov random field (MRF) or a stationary Gaussian random field (GRF) kriging model. They use a kind of hybrid MCMC scheme, plugging in the Breslow estimator for the baseline hazard at each updating step. Banerjee, Wall and Carlin (2003) assume a parametric Weibull baseline hazard rate and MRF or GRF priors for the spatial component. Banerjee and Carlin (2003) and Carlin and Banerjee (2002) extend this work by including nonparametric estimation of the baseline hazard rate. Effects of continuous covariates are still assumed to be of linear parametric form as in (1). A semiparametric fully Bayesian extension to Cox-type models that can deal with all the issues mentioned has been developed by Hennerfeind, Brezger and Fahrmeir (2004). An empirical Bayes or penalized likelihood approach based on a mixed model representation of linear regression splines to estimate the baseline hazard rate has been suggested by Cai, Hyndman and Wand (2002). Cai and Betensky (2003) extended this work by including estimation of linear covariate effects and considering interval-censored data.

In this paper we propose an extended Cox-type model that allows for the simultaneous estimation of the baseline hazard and a structured additive predictor acting multiplicatively on the baseline. Both the log-baseline hazard and effects of continuous covariates or further time scales such as calendar time are weakly parameterized using penalized splines. If observations are clustered in connected geographical regions, spatial effects can be estimated using the MRF approach commonly known from disease mapping. If exact spatial locations are available, we propose to use GRFs or two-dimensional P-spline surface smoothers, which can also be used to model flexible interaction surfaces between two continuous covariates. Our approach also supports cluster-specific random effects (or frailties) and varying coefficient terms both with continuous and spatial effect modifiers. Time-varying effects can be subsumed into the varying coefficients framework, where survival time acts as effect modifier and is again modelled as a P-spline.

The estimation procedure is based on a variance components mixed model representation

of the structured additive predictor, which has become popular in a generalized additive model context (compare Fahrmeir, Kneib and Lang (2004) or Ruppert, Wand and Carroll (2003) and the references therein). We extend existing methods for the estimation of Cox models with uncorrelated cluster-specific frailties or random effects to more general mixed models for survival times. Variance components of the mixed model, corresponding to inverse smoothing parameters, are estimated using marginal or restricted maximum likelihood. Since the marginal likelihood can not be derived analytically, certain approximations are proposed. Performance of these approximations is investigated by comparing results from the mixed model approach to results from its fully Bayesian counterpart by Hennerfeind et al. (2004), where posterior estimates for the smoothing parameters are available without approximations. The presented methodology is implemented in BayesX, a public domain software package available from

`http://www.stat.uni-muenchen.de/~lang/bayesx`[1]

Section 2 describes structured hazard regression models for survival times including a discussion of the different model terms and priors. Inference is outlined in Section 3. In Section 4 we analyze the data set on leukemia survival times and Section 5 further investigates properties of the presented approach through simulation. The concluding Section 6 gives comments on future research.

2 Structured hazard regression

2.1 Hazard rate model and likelihood

Consider right-censored survival data given by observed lifetimes t_i with censoring indicator δ_i, $i = 1, \ldots, n$, and additional covariate observations. As pointed out in the introduction, in a number of applications there is a need for extending the basic Cox model (1) to hazard rate models which can incorporate flexible nonparametric terms for time scales and continuous covariates, time-varying effects, and a spatial component. Reparametrizing the baseline hazard rate through $g_0(t) = \log(\lambda_0(t))$ and partitioning covariates into groups of different type, we extend the Cox model to a semiparametric hazard rate model

$$\lambda_i(t) = \exp(\eta_i(t)), \quad i = 1, \ldots, n, \tag{2}$$

with structured additive predictor

$$\eta_i(t) = v_i'\gamma + g_0(t) + \sum_{l=1}^{L} g_l(t)u_{il} + \sum_{j=1}^{J} f_j(x_{ij}(t)) + f_{spat}(s_i) + b_{g_i}. \tag{3}$$

In (3), $g_0(t)$ is the log-baseline effect, $g_l(t)$ are time-varying effects of covariates u_l, $f_j(x_j(t))$ is the nonlinear effect of a possibly time-varying covariate $x_j(t)$, and $f_{spat}(s)$ is the spatial effect at site or in region $s \in \{1, \ldots, S\}$. The vector γ contains the usual linear effects, and b_g, $g \in \{1, \ldots, G\}$ are uncorrelated individual- or group-specific frailties, with $b_{g_i} = b_g$ if individual i is in group g. Several extensions of the predictor (3), such as inclusion of

[1] The new version of BayesX containing structured hazard regression within a mixed model approach will be available in the beginning of November 2004.

interactions $f_{jk}(x_{ij}(t), x_{ik}(t))$ between two continuous covariates and random slopes $z_i' b_{g_i}$, are possible and included in our implementation.

To obtain a general mixed model formulation of (3) we introduce some matrix notation. Let $\eta = (\eta_1, \ldots, \eta_i, \ldots, \eta_n)'$ denote the predictor vector, where $\eta_i := \eta_i(t_i)$ is the value of predictor (3) at the observed lifetime t_i, $i = 1, \ldots, n$. Correspondingly, let $g_l = (g_l(t_1), \ldots, g_l(t_n))'$ denote the vector of evaluations of the functions $g_l(t)$, $l = 0, \ldots, L$, $f_j = (f_j(t_1), \ldots, f_j(t_n))'$ the vector of evaluations of the functions $f_j(t)$, $j = 1, \ldots, J$, $f_{spat} = (f_{spat}(s_1), \ldots, f_{spat}(s_n))'$ the vector of spatial effects, and $b = (g_{g_1}, \ldots, b_{g_n})'$ the vector of uncorrelated random effects.

In the following, we express all vectors g_l, f_j, f_{spat} and b as the matrix product of an appropriately defined design matrix Z, say, and a (possibly high-dimensional) vector β of regression coefficients, e.g. $g_l = Z_l \beta_l$, $f_j = Z_j \beta_j$, etc. Then, after reindexing, we can represent the predictor vector η in generic notation as

$$\eta = V\gamma + Z_1\beta_1 + \ldots + Z_p\beta_p. \tag{4}$$

Under the usual assumptions about noninformative censoring, the likelihood, given all parameters γ and $\beta = (\beta_1', \ldots, \beta_p')$, is

$$L(\gamma, \beta) = \prod_{i=1}^{n} \lambda_i(t_i)^{\delta_i} \exp\left(- \int_0^{t_i} \lambda_i(t) dt \right), \tag{5}$$

inserting (2) and (3) for $\lambda_i(t_i)$. let $\delta = (\delta_1, \ldots, \delta_n)'$ denote the vector of censoring indicators and $\Lambda = (\Lambda_1(t_1), \ldots, \Lambda_i(t_i), \ldots, \Lambda_n(t_n))'$ the vector of cumulative hazard rates $\Lambda_i(t_i) = \int_0^{t_i} \lambda_i(t) dt$. Then the log-likelihood can be written as

$$l(\gamma, \beta) = \delta'\eta - \mathbb{1}'\Lambda. \tag{6}$$

2.2 Priors for random effects

In our mixed model approach to structured hazard regression, the parameters γ are considered as fixed while $\beta_1, \ldots, \beta_p$ are random effects. Specification of the model is completed by appropriate distributional assumptions. From a Bayesian point of view, we have to specify priors for $\beta_1, \ldots, \beta_p$. The general form of a prior or random effects distribution for β_j in (4) is Gaussian,

$$p(\beta_j) \propto \exp\left(-\frac{1}{2\tau_j^2} \beta_j' K_j \beta_j \right), \tag{7}$$

where K_j is a precision or penalty matrix. Uncorrelated random effects $b = (b_1, \ldots, b_G)'$ are a special case, with $K_j = I$. Generally, the random effects β_j representing a function are correlated, and K_j shrinks parameters towards zero or penalizes too abrupt jumps between neighboring components of β_j. In the case of P-splines and MRF models for the spatial component, K_j will be rank deficient, i.e. $\text{rank}(K_j) < \dim(\beta_j)$, and the Gaussian distribution is partially improper. The variance τ_j^2 in (7) acts as an inverse smoothing parameter: A small (large) value of τ_j^2 corresponds to an increase (decrease) of the penalty or shrinkage. We consider these variances as unknown fixed constants, which are estimated through a marginal likelihood approach, see Section 3.3.

In the following, we outline specification of K_j in (7) and the design matrices Z_j for functions g_l, f_j and the spatial component f_{spat}. For more details, especially on how to

include interactions and varying coefficient terms in the general framework, we refer to Fahrmeir et al. (2004) and Lang and Brezger (2004).

Unknown functions g_l or f_j are modeled through P-splines. The basic idea (Eilers and Marx 1996) is to a 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 $x_{min} = \kappa_0 < \kappa_1 < \ldots < \kappa_s = x_{max}$, $d_j = l + s$. The number of knots is moderate, but not too small, to maintain flexibility, and smoothness of the functions is encouraged by quadratic difference penalties for neighboring coefficients in the sequence $\beta_j = (\beta_{j1}, \ldots, \beta_{jd_j})'$. In a penalized log-likelihood setting, the difference penalty can be expressed as $\lambda_j \beta_j' K_j \beta_j$, where the penalty matrix is of the form $K_j = D'D$, with D a first or second order difference matrix and $\lambda_j = 1/2\tau_j^2$ a smoothing parameter. In a mixed model or Bayesian approach this is equivalent to a prior (7). The matrix K_j has rank $\dim(\beta_j) - 1$ or $\dim(\beta_j) - 2$ for first or second order difference penalties, respectively and therefore prior (7) is partially improper. The $n \times \dim(\beta_j)$ design matrix Z_j consists of the basis functions evaluated at the observations x_{ij}, i.e., $Z_j[i, m] = B_m(x_{ij})$. Priors for the unknown functions $g_j(t)$ are defined in complete analogy.

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

$$f_{spat}(s) := \beta_s^{spat} = \frac{1}{N_s} \sum_{s' \in \delta_s} \beta_{s'}^{spat} + u_s, \quad u_s \sim N\left(0, \frac{\tau_{spat}^2}{N_s}\right) \tag{9}$$

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. The $n \times S$ design matrix 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 K_{spat} has the form of an adjacency matrix with $\text{rank}(K_{spat}) = S - 1$.

Our second 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) in their mixed model approach for 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 an isotropic correlation function $\text{corr}(\beta_s^{spat}, \beta_{s'}^{spat}) = \rho(||s - s'||)$. For a finite array $s \in \{1, \ldots, S\}$ of sites as in our application the prior can be brought in the general form (7) with penalty matrix $K_{spat} = C^{-1}$ and

$$C[k, l] = \rho(||s_k - s_l||),\ 1 \leq k, l \leq n.$$

For the correlation function $\rho(r)$ we use the Matérn family of correlation functions $\rho(r; \alpha, \nu)$. For the special case $\nu = 1.5$ the correlation function simplifies to

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

which is the simplest member of the Matérn family that results in differentiable surface estimates. The parameter α controls how fast correlations die out with increasing distance r. We choose α in a preprocessing step according to the rule

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

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

To decrease the computational burden and to enhance numerical stability, we suggest to use low-rank kriging instead of a full kriging approach where the dimension of β^{spat} is equal or close to the number of observations. Applying a space filling algorithm to the locations s_i yields a set of knots $\kappa_1, \ldots, \kappa_M$. Based on these knots we can approximate f_{spat} as $Z_{spat}\beta^{spat}$, where the $n \times M$ design matrix Z_{spat} consists of elements $Z[i,j] = \rho(||s_i - \kappa_j||)$ and the penalty matrix is given by $K_{spat} = \tilde{C}$ with $\tilde{C}[k,l] = \rho(||\kappa_k - \kappa_l||)$. The number of knots controls the trade-off between accuracy of the approximation and numerical simplification. Details on GRFs and (low-rank) kriging can be found in Kammann and Wand (2003) or Kneib and Fahrmeir (2004).

A third alternative approach is based on two-dimensional tensor product 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, the spatial effect can be seen as an interaction between two continuous covariates x and y. The corresponding spatial prior for the array of two-dimensional B-splines can again be expressed in the general form (7), see Lang and Brezger (2004).

3 Mixed Model based Inference

Since regression parameters describing nonparametric and spatial effects are assumed to have a random effects or prior distribution, inference is not based on the log-likelihood itself but on the penalized log-likelihood

$$l_{pen}(\gamma, \beta) = l(\gamma, \beta) - \sum_{j=1}^{p} \frac{1}{2\tau_j^2} \beta_j' K_j \beta_j. \tag{10}$$

From a Bayesian viewpoint (10) is equivalent to the log-posterior and therefore maximizing (10) with respect to the regression coefficients yields either penalized likelihood or posterior mode estimates. Though direct maximization of (10) is possible, marginal likelihood estimates for the variance parameters τ_j^2 cannot be derived from this penalized likelihood, since some of the random effects distributions (7) for the correlated effects in (4) are improper. We therefore propose to estimate structured hazard regression models via the following two steps:

1. Reparametrize the general mixed model (4) in a classical variance components model formulation to obtain uncorrelated random effects with proper priors.

2. Iteratively update regression coefficients given the current variance parameters and variance components given current regression coefficients through Newton-Raphson- / Fisher-Scoring-steps.

The different parts of the estimation procedure will now be described in further detail.

3.1 Mixed model representation

In the following we assume that β_j has dimension d_j and the corresponding penalty matrix has rank k_j. To rewrite the structured additive predictor (4) in variance components formulation the vectors of regression coefficients β_j, $j = 1, \ldots, p$, are decomposed into an unpenalized part (with flat prior) and a penalized part (with i.i.d. Gaussian prior), i.e.

$$\beta_j = Z_j^{unp}\beta_j^{unp} + Z_j^{pen}\beta_j^{pen} \tag{11}$$

with a $d_j \times (d_j - k_j)$ matrix Z_j^{unp} and a $d_j \times k_j$ matrix Z_j^{pen}. We expect decomposition (11) to satisfy the following conditions:

(i) The composed matrix $(Z_j^{unp}\ Z_j^{pen})$ has full rank to make the transformation in (11) a one-to-one transformation. This also implies that both Z_j^{unp} and Z_j^{pen} have full column rank.

(ii) Z_j^{unp} and Z_j^{pen} are orthogonal, i.e. $Z_j^{unp\prime}Z_j^{pen} = 0$.

(iii) $Z_j^{unp\prime}K_jZ_j^{unp} = 0$, resulting in β_j^{unp} being unpenalized by K_j.

(iv) $Z_j^{pen\prime}K_jZ_j^{pen} = I$, resulting in an i.i.d. Gaussian prior for β_j^{pen}.

In general, matrices fulfilling these requirements can be obtained as follows: Z_j^{unp} contains a $d_j - k_j$ dimensional basis of the null space of K_j. Therefore requirement (iii) is automatically fulfilled. Z_j^{pen} can be obtained by $Z_j^{pen} = L_j(L_j'L_j)^{-1}$ where the full column rank $d_j \times k_j$ matrix L_j is determined by the decomposition of the penalty matrix K_j into $K_j = L_jL_j'$. This ensures requirements (i) and (iv). If we choose L_j such that $L_j'Z_j^{unp} = 0$ holds, we finally obtain requirement (ii). The decomposition $K_j = L_jL_j'$ of the penalty matrix can be based on the spectral decomposition $K_j = \Gamma_j\Omega_j\Gamma_j'$, where the $(k_j \times k_j)$ diagonal matrix Ω_j contains the positive eigenvalues ω_{jm}, $m = 1, \ldots, k_j$, of K_j in descending order, i.e. $\Omega_j = \text{diag}(\omega_{j1}, \ldots, \omega_{j,k_j})$. Γ_j is a $(d_j \times k_j)$ orthogonal matrix of the corresponding eigenvectors. From the spectral decomposition we can choose $L_j = \Gamma_j\Omega_j^{1/2}$. Note, that the factor L_j is not unique and numerically superior factorizations may exist (compare Fahrmeir et al. (2004) for a more detailed discussion of special model terms).

From decomposition (11) and requirements (i) to (iv) it follows that

$$p(\beta_j^{unp}) \propto const \quad \text{and} \quad \beta_j^{pen} \sim N(0, \tau_j^2 I). \tag{12}$$

Now, defining the matrices $Q_j = Z_jZ_j^{unp}$ and $P_j = Z_jZ_j^{pen}$, allows us to rewrite the structured additive predictor (4) as

$$\eta = \sum_{j=1}^{p} Z_j\beta_j + V\gamma = \sum_{j=1}^{p}(Z_jZ_j^{unp}\beta_j^{unp} + Z_jZ_j^{pen}\beta_j^{pen}) + V\gamma = Q\beta^{unp} + P\beta^{pen},$$

which is the predictor of a variance components mixed model with fixed effects β^{unp}, random effects $\beta^{pen} \sim N(0, \Sigma)$ and $\Sigma = \text{diag}(\tau_1^2, \ldots, \tau_1^2, \ldots, \tau_p^2, \ldots, \tau_p^2)$. The design matrix P and the vector β^{pen} are composed of the matrices P_j and the vectors β_j^{pen}, respectively. More specifically, we obtain $P = (P_1\ P_1\ \cdots\ P_p)$ and the stacked vector $\beta^{pen} = ((\beta_1^{pen})', \ldots, (\beta_p^{pen})')'$. Similarly the matrix Q and the vector β^{unp} are given by $Q = (Q_1\ \cdots\ Q_p\,U)$ and $\beta^{unp} = ((\beta_1^{unp})', \ldots, (\beta_p^{unp})', \gamma')'$.

In variance components model representation the penalized likelihood (10) transforms to

$$l_{pen}(\beta^{unp}, \beta^{pen}) = \delta'\eta - \mathbb{1}'\Lambda - \frac{1}{2}\beta^{pen\prime}\Sigma^{-1}\beta^{pen}. \tag{13}$$

3.2 Estimation of regression coefficients

The main difficulty in obtaining derivatives of (13) with respect to the regression coefficients is to derive expressions for the derivatives of the cumulative hazard function $\Lambda(t)$. For simplicity consider for the moment a cumulative hazard of the form

$$\Lambda(t_i) = \int_0^{t_i} \exp(z_i(t)'\beta)dt,$$

where $z_i(t)$ is a (possibly time-dependent) vector of covariates and β is a vector of regression coefficients. This setting essentially reflects the structure of $\Lambda(t)$ in a structured hazard regression model. Now, first and second derivatives are given by

$$\frac{\partial}{\partial \beta_j}\Lambda(t_i) = \int_0^{t_i} z_{ij}(t) \exp(z_i(t)'\beta)dt \tag{14}$$

and

$$\frac{\partial}{\partial \beta_j \partial \beta_k}\Lambda(t_i) = \int_0^{t_i} z_{ij}(t) z_{ik}(t) \exp(z_i(t)'\beta)dt. \tag{15}$$

Both expressions include integrals, which, in general, can not be solved analytically. Therefore we have to apply a numerical integration procedure such as the trapezoidal rule to approximate them. In our implementation we use a quantile-based grid on the time axis instead of the observed t_i to reduce the gridsize without loosing accuracy of the approximation. Note, that expression (14) can be somewhat simplified if some covariates are constant over time, i.e. $z_{ij}(t) \equiv z_{ij}$, since in this case (14) reduces to $z_{ij}\Lambda(T_i)$ and $\Lambda(T_i)$ has to be computed only once. Similar simplifications can be used when computing (15) if both $z_{ij}(t)$ and $z_{ik}(t)$ do not depend on t.

In a mixed model, the score-vector can be partitioned into two parts defined by the derivatives with respect to the unpenalized and the penalized vector of regression coefficients, i.e.

$$s = \begin{pmatrix} s^u \\ s^p \end{pmatrix} = \begin{pmatrix} \dfrac{\partial l_p}{\partial \beta^{unp}} \\ \dfrac{\partial l_p}{\partial \beta^{pen}} \end{pmatrix} = \begin{pmatrix} \delta' Q - \mathbb{1}' \dfrac{\partial \Lambda}{\partial \beta^{unp}} \\ \delta' P - \mathbb{1}' \dfrac{\partial \Lambda}{\partial \beta^{pen}} - \Sigma^{-1}\beta^{pen} \end{pmatrix}.$$

Similarly, the observed Fisher-information is partitioned into four blocks:

$$F = \begin{pmatrix} F^{uu} & F^{up} \\ F^{pu} & F^{pp} \end{pmatrix} = \begin{pmatrix} \mathbb{1}' \dfrac{\partial^2 \Lambda}{\partial \beta^{unp} \partial \beta^{unp\prime}} & \mathbb{1}' \dfrac{\partial^2 \Lambda}{\partial \beta^{unp} \partial \beta^{pen\prime}} \\ \mathbb{1}' \dfrac{\partial^2 \Lambda}{\partial \beta^{pen} \partial \beta^{unp\prime}} & \mathbb{1}' \dfrac{\partial^2 \Lambda}{\partial \beta^{pen} \partial \beta^{pen\prime}} + \Sigma^{-1} \end{pmatrix}.$$

Both quantities are now easy to calculate based on expressions (14) and (15) and allow the computation of updated estimates for the regression coefficients given the variances via a Newton-Raphson-step.

3.3 (Approximate) Marginal likelihood for variance components

In Gaussian linear mixed models, a well established method for the estimation of variance components is restricted maximum likelihood (REML), which - in contrast to ordinary

maximum likelihood - takes into account the loss of degrees of freedom due to the estimation of the regression coefficients. As Harville (1974) showed, REML estimation is equivalent to maximizing the marginal likelihood for the variance components

$$L^{marg}(\Sigma) = \int L_{pen}(\beta^{unp}, \beta^{pen}, \Sigma) d\beta^{unp} d\beta^{pen}. \tag{16}$$

This equivalence allows to generalize REML estimation to more general situations including regression models for survival times. Up to now marginal likelihood estimation has mostly been used in the context of subject-specific frailty models based on the partial likelihood (compare e.g. Therneau and Grambsch (2000) or Ripatti and Palmgren (2000)). Cai et al. (2002) use marginal likelihood estimates for the smoothing parameter of the baseline hazard but do not provide estimation equations. Instead they maximize the marginal likelihood numerically, which may become quite computerintensive if the model includes more variance components for a structured additive predictor. Furthermore, in their model effects of covariates are assumed to have parametric form.

In the following we describe a possibility to estimate variances in a structured hazard regression model based on the full marginal likelihood (not the partial marginal likelihood). Two approximation steps allow to use a Fisher-Scoring-algorithm for the maximization of (16), yielding estimation equations which are numerically simple to evaluate. First, applying a Laplace approximation to the marginal log-likelihood results in

$$l^{marg}(\Sigma) \approx l(\hat{\beta}^{unp}, \hat{\beta}^{pen}) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\hat{\beta}^{pen\prime}\Sigma^{-1}\hat{\beta}^{pen} - \frac{1}{2}\log|H|.$$

If we can assume that both $l(\hat{\beta}^{unp}, \hat{\beta}^{pen})$ and $\hat{\beta}^{pen}$ vary only slowly when changing the variance components we can further reduce the marginal likelihood to

$$l^{marg}(\Sigma) \approx -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|H| - \frac{1}{2}\beta^{pen\prime}\Sigma^{-1}\beta^{pen}, \tag{17}$$

where β^{pen} denotes a fixed value not depending directly on the variances, e.g. a current estimate. The second approximation seems to be reasonable since - at least in generalized additive models - it is well known that small changes in the smoothing parameters do not affect the estimates of the regression coefficients very much. A similar argument is given by Breslow and Clayton (1993) to simplify the marginal likelihood for variance components of generalized linear mixed models. Although the approximation steps may look rather crude at first sight, they proved to work well in simulations as well as in real data applications.

First and second derivatives of (17) can be easily derived, based on differentiation rules for matrices. Since expressions for general covariance matrices Σ become quite lengthy, we make use of the block diagonal structure of Σ in variance components models to obtain simpler formulae. For the score function this yields

$$s_j^* = \frac{\partial l^{marg}(\Sigma)}{\tau_j^2} = -\frac{k_j}{2\tau_j^2} + \frac{1}{2\tau_j^4}\operatorname{tr}\left(G_{jj}^{pp}\right) + \frac{1}{2\tau_j^4}\beta_j^{pen\prime}\beta_j^{pen},$$

where $k_j = \operatorname{rank}(K_j)$, $G = F^{-1}$ denotes the inverse Fisher information (for the regression coefficients) and G_{jj}^{pp} is the diagonal block of F^{-1} corresponding to β_j^{pen}. The expected Fisher-information is given by

$$F_{jk}^* = E\left(-\frac{\partial^2 l^{marg}(\Sigma)}{\partial\tau_j^2\partial\tau_k^2}\right) = \frac{1}{2\tau_j^4\tau_k^4}\operatorname{tr}\left(G_{jk}^{pp}G_{kj}^{pp}\right).$$

Here, G_{jk}^{pp} denotes the off-diagonal block of F^{-1} corresponding to β_j^{pen} and β_k^{pen}. Both expressions are numerically simple to evaluate since H^{-1} and β^{pen} are direct byproducts from the estimation of the regression coefficients. Based on the score-function and the Fisher-information we can compute updated variances via a Fisher-scoring step.

4 Application: Leukemia survival data

To illustrate the usefulness and flexibility of structured hazard regression, we reanalyze a data set from Henderson et al. (2002) on leukemia survival times in Northwest England. Their analysis concentrated on the detection of spatial variation in survival times but retained the assumption of a linear predictor for covariate effects. Modeling such covariates as penalized splines allows to check whether this assumption is appropriate or whether a more flexible modeling improves the model fit.

The data set contains information on all 1,043 cases of acute myeloid leukemia in adults that have been diagnosed between 1982 and 1998 in Northwest England. Almost 16% of the cases are censored. Continuous covariates include the age of the patient, the white blood cell count (WBC) at diagnosis and the Townsend deprivation index (TPI) which measures the deprivation for the enumeration district of residence. Higher values of this index indicate poorer regions while smaller values correspond to wealthier regions. Since the observation area consists of 8,131 enumeration districts, the Townsend index can be considered a subject-specific covariate. The sex of a patient is included in dummy-coding (1=female, 0=male). Spatial information is available in both ways described in Section 2: The exact location of the residence of a patient is given in terms of longitude and latitude, but of course we can also aggregate this information to district-level. Figure 1 shows the district boundaries together with the exact locations of the observed cases. Comparing results from district-level and individual-level analyzes allows to judge the loss of information caused by the aggregation of observations within districts.

In both situations the structured additive predictor is given by

$$\eta_i(t) = \gamma_0 + \gamma_1 sex_i + g_0(t) + f_1(age_i) + f_2(wbc_i) + f_3(tpi_i) + f_{spat}(s_i)$$

where g_0 is the (centered) log-baseline, f_1, f_2 and f_3 are smooth functions of the continuous covariates and f_{spat} is a spatial effect. Both g_0 and the f_j will be modeled as cubic P-splines with second order difference penalty and 20 knots. In an individual-level analysis $s_i = (s_i^x, s_i^y)$ is the exact location of the patient's residence while in a district-level analysis s_i denotes the district the patient lives in.

4.1 District-level analysis

In a district level analysis a natural choice to model the spatial effect is the Markov random field (9). With this specification of the spatial effect, we obtain the estimates shown in Figure 2a-d for the log-baseline $g_0(t)$ and the nonparametric effects f_j. The log-baseline decreases monotonically over nearly the whole observation period, alternating between relatively steep decreasing periods and almost flat periods. At the end of the observation period there is a strong increase in $g_0(t)$. However, only 26 individuals survived more than 10 years and therefore this increase should not be over-interpreted.

Obviously the effects f_1 and f_2 of age and white blood cell count are almost linear and could therefore be modeled parametrically to reduce model complexity. Both effects are

quite similar to those found by Henderson et al. (2002) as is the effect of sex ($\gamma_1 = 0.076$). Note, that Henderson et al. modelled sex in effect-coding.

In contrast, the effect of the deprivation index is clearly nonlinear with lowest values for the developed enumeration districts with a low value of the TPI. Moving to the right on the TPI-scale first increases the risk to die from leukemia but remains almost constant when reaching a value of about zero. Although both effects of age and WBC are nearly linear, the flexible modeling is a clear improvement over a purely parametric approach since it allows to check for the linearity of some effects but also allows more flexible functional forms when needed.

Looking at the estimated spatial effect in Figure 3a, we find several districts with low risk in the western part of the map, surrounded by districts of increased risk. In the southern part of the map there are also some districts with lower risk but the spatial effect is less pronounced here. This structure can be seen even more clearly from the significance map in Figure 3b, where black denotes districts with strictly negative credible intervals and white denotes districts with strictly positive credible intervals. Again, estimation results are quite similar to those found by Henderson et al. (2002).

4.2 Individual-level analysis

Of course, performing a district-level analysis when more detailed information is available is questionable. Therefore we replaced the MRF with a stationary Gaussian random field based on the exact locations of the residences. Using a complete kriging term would require the computation and inversion of an approximately 1,100 times 1,100 matrix, since there is a total number of about 1,100 regression parameters in this model. Such computations are rather time-consuming and we used the low-rank kriging approach described in Section 2.2 with a much smaller number of knots instead. We tried 50, 100 and 200 knots with essentially the same results indicating that the approach is rather insensitive to the number of knots. We also fitted a two-dimensional P-spline for the individual-level spatial effect and obtained comparable estimates.

Effects for continuous and categorical covariates are almost the same as in the district-level model and we do not display them again but concentrate on the spatial effect. Figure 4 shows this spatial effect for a low-rank kriging term with 50 knots. In general, results are comparable to those from the district-level analysis but the kriging approach shows a more detailed spatial pattern and also finds a somewhat larger spatial variation. In particular, there is considerable variation of the spatial effect within most of the districts. When performing a district-level analysis, such information is lost since a constant risk level is assumed for each district. This assumption may be problematic because district boundaries are political constructs and usually do not reflect factors relevant for the risk of patients. Therefore, the computationally feasible low-rank kriging approach seems to be preferable to the Markov random field approach when individual-level information is available.

4.3 Inclusion of time-varying effects

To check the proportional hazards assumption for males and females, we included a time-varying effect of sex in our model. Although the estimated effect is somewhat increasing over time (see Figure 2e), it is almost equal to a horizontal line and has rather wide

credible intervals including such a horizontal line. Therefore we may conclude that the proportional hazard assumptions is valid for the sub-populations of males and females.

5 Simulation study

To gain deeper insight in the statistical properties of the presented mixed model approach, especially compared to the fully Bayesian alternative of Hennerfeind et al. (2004), we performed a simulation study. We generated 250 data sets, each with 750 observations based on the following structured additive predictor:

$$\eta_i(t) = g_0(t) + f(x_i) + f_{spat}(s_i). \tag{18}$$

The baseline hazard $\lambda_0(t) = \exp(g_0(t))$ (shown in Figure 5a) is chosen to reflect a situation where the risk for an event is initially high, decreasing for some time and rising again at the end of the observation period. Such bathtub-shaped hazard rates are quite common in studies on survival times, as we have already seen in Section 4, but can hardly be handled within a regression approach assuming a parametric form of the baseline, e.g. a Weibull distributed baseline. The nonparametric effect $f(x)$ is given by a sine curve, i.e. $f(x) = 0.6 \cdot \sin(\pi(2x - 1))$, where x is chosen randomly from an equidistant grid of 75 values within the interval $[0, 1]$. The spatial function f_{spat} is defined based on the centroids of the 124 districts of the two southern states of Germany (Bavaria and Baden-Württemberg) and is shown in Figure 5b. Again the value s is randomly assigned to the observations. Three different amounts of censoring were considered: No censoring at all, moderate censoring (10-15% censored observations) and high censoring (20-25% censored observations). To obtain censored observations, we generated independent censoring times $C_i \sim \text{Exp}(0.2)$ (medium censoring) and $C_i \sim \text{Exp}(0.6)$ (high censoring) and defined the observed survival time to be $t_i = \min(T_i, C_i)$, where T_i is generated according to the hazard rate $\lambda_i(t) = \exp(\eta_i(t))$.

In general it is not clear, how to simulate survival times from a Cox-type model with such a hazard rate since $\lambda_i(t)$ does not correspond to a known distribution. We used a technique proposed in Bender, Augustin and Blettner (2004) that allows the simulation of Cox-models with arbitrary baseline hazard as long as the cumulative baseline hazard $\Lambda_0(t)$ and its inverse $\Lambda_0^{-1}(t)$ are available (at least for numerical evaluation). In this case uncensored survival times T_i can be simulated via $T_i = \Lambda_0^{-1}(-\log(U_i))$, where U_i is a random variable uniformly distributed on $[0, 1]$, i.e. $U_i \sim U[0, 1]$.

To compare the accuracy of the point estimates produced by the mixed model approach and the fully Bayesian MCMC-approach by Hennerfeind et al. (2004), we computed the empirical MSEs shown in Figure 6. A first important observation is that results for the covariate effects f_1 and f_{spat} are estimated with approximately the same precision regardless of the amount of censoring. For the log-baseline the median MSE also remains almost unchanged but its variability increases with increasing censoring. In general (except for the log-baseline in the case of high censoring) the mixed model approach performs better than the MCMC-approach when considering the median MSE but differences are quite small. For the spatial effect there are some outliers for the mixed model estimates caused by replications where the variances of the spatial effect were estimated to be close to zero resulting in very flat curves. This phenomena of underestimating effects in a moderate number of replications has already been observed in the context of structured additive regression with responses belonging to exponential families (Fahrmeir et al. 2004).

Figure 7 shows the average estimates and the bias for the covariate effects f_1 and f_{spat}. Since average estimates were roughly the same for all censoring mechanisms, we only show results for the highest amount of censoring. Both approaches perform comparably in terms of bias. While bias is almost negligible for the nonparametric effect f_1, it becomes more distinct for the spatial effect.

The last comparison concerns average coverage probabilities summarized in Table 1. In general the MCMC-approach produces more conservative credible intervals than the mixed model approach. Both approaches meet the nominal level in most cases, only the mixed model approach is somewhat below the nominal 95%-level for the spatial effect without censoring. This is caused by the outliers discussed earlier, were the estimated spatial effect was close to zero.

6 Conclusions

In this paper we presented a flexible possibility to extend the traditional Cox model to allow for the simultaneous estimation of a smooth hazard rate and a complex structured additive predictor acting multiplicatively on the baseline. The approach proved to be useful in a real data example on leukemia survival times and showed satisfactory statistical properties in a simulation study. The approximation steps outlined in Section 3.3 allow for a fast optimization of the marginal likelihood of the variance parameters even in fairly complex models.

Some further extensions of the proposed method might be desirable and will be investigated in future work. Two main issues are: First, generalizing our mixed model approach for survival data to more general censoring schemes, such as left or interval censoring as often encountered in practice, and secondly, extending it to more complex event history data, such as competing risks and recurrent event data.

Acknowledgement:
We thank Silvia Shimakura for providing the data, and we gratefully acknowledge financial support from the German Science Foundation (DFG), Collaborative research center 386 "Statistical Analysis of Discrete Structures".

References

Banerjee, S., Wall, M. M., Carlin, B. P., 2003: Frailty modelling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*, 4, 123-142.

Banerjee, S. and Carlin, B. P., 2003: Semiparametric spatio-temporal frailty modelling. *Environmetrics*, 14, 523-535.

Bender, R., Augustin, T. and Blettner, M., 2004: Generating Survival Times to Simulate Cox Proportional Hazards Models. *Statistics in Medicine*, to appear.

Breslow, N. E. and Clayton, D. G., 1993: Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

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

Cai, T. and Betensky, R., 2003: Hazard Regression for Interval Censored Data with Penalized Splines. *Biometrics*, 59, 570-579.

Carlin, B. P. and Banerjee, S., 2002: Hierarchical multivariate CAR models for spatio-temporally correlated data. In *Bayesian Statistics 7*, Bernardo et al. (eds.), University Press, Oxford.

Cox, D. R., 1972: Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, 34, 187-220.

Eilers, P.H.C. and Marx, B.D., 1996: Flexible smoothing using B-splines and penalties. *Statistical Science*, 11, 89-121.

Fahrmeir, L., Kneib, T. and Lang, S.: Penalized structured additive regression: A Bayesian perspective. *Statistica Sinica*, 14, 715-745.

Harville, D. A., 1974: Bayesian inference for variance components using only error contrasts. *Biometrika*, 61, 383-85.

Henderson, R., Shimakura, S. and Gorst, D.: Modelling Spatial variation in Leukemia Survival Data. *Journal of the American Statistical Association*, 97, 965-972.

Hennerfeind, A., Brezger, A. and Fahrmeir, L., 2004: Geoadditive survival models. SFB Discussion paper 333, University of Munich.

Ibrahim, J., Chen, M. H. and Sinha, D., 2001: *Bayesian Survival Analysis*, Springer, New York.

Kammann, E. E. and Wand, M. P., 2003: Geoadditive Models. *Journal of the Royal Statistical Society C*, 52, 1-18.

Kneib, T. and Fahrmeir, L., 2004: Structured additive regression for categorical space-time data: A mixed model approach. SFB 386 Discussion paper 377, University of Munich.

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

Li, Y. and Ryan, L., 2002: Modelling Spatial Survival Data Using Semiparametric Frailty Models. *Biometrics*, 58, 287-297.

Ripatti, S. and Palmgren, J., 2000: Estimation of Multivariate Frailty Models Using Penalized Likelihood. *Biometrics*, 56, 1016-1022.

Ruppert, D., Wand, M.P. and Carroll, R.J., 2003: *Semiparametric Regression*, University Press, Cambridge.

Therneau, T. M. and Grambsch, P. M., 2000: *Modeling Survival Data: Extending the Cox Model*, Springer, New York.

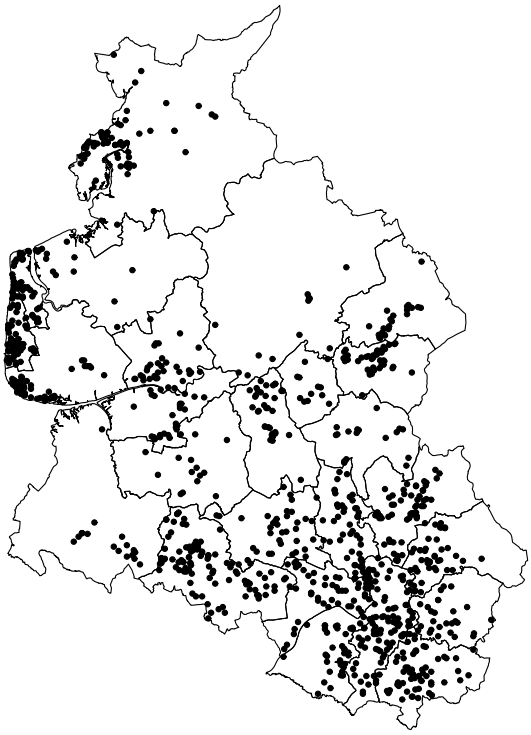

Figure 1: Leukemia survival data: Districts of Northwest England and locations of the observations.

		$g_0(t)$		$f_1(x)$		$f_{spat}(s)$	
		80%	95%	80%	95%	80%	95%
REML	no censoring	0.911	0.963	0.854	0.976	0.892	0.931
	medium censoring	0.915	0.960	0.860	0.976	0.937	0.979
	high censoring	0.851	0.944	0.864	0.976	0.943	0.993
MCMC	no censoring	0.925	0.975	0.847	0.975	0.949	0.996
	medium censoring	0.944	0.978	0.853	0.976	0.952	0.997
	high censoring	0.973	0.984	0.856	0.977	0.940	0.994

Table 1: Simulation Study: Average coverage probabilities.

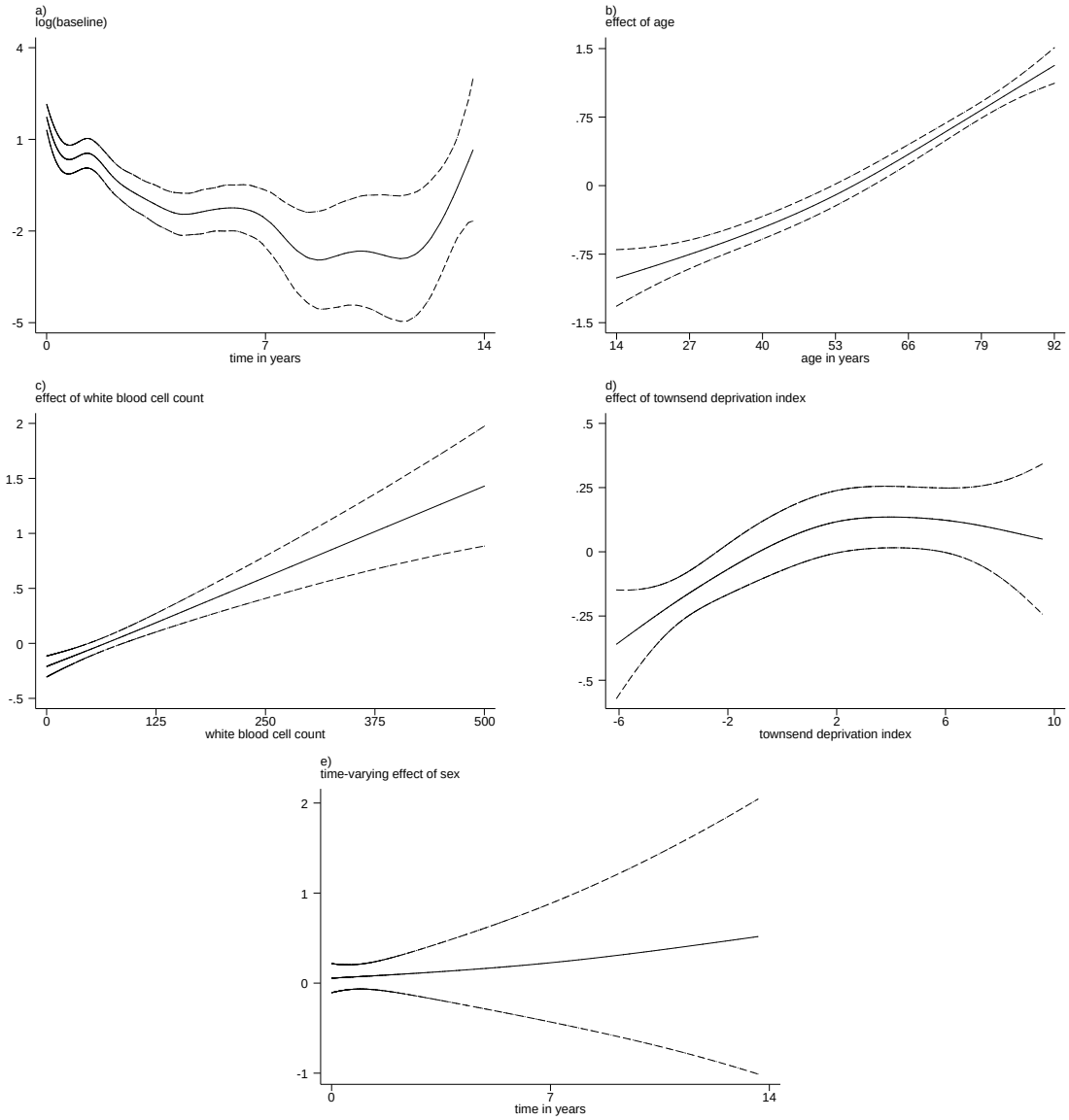

Figure 2: Leukemia Survival data: Estimates of the log-baseline, the effects of age, wbc and tpi, and of a time-varying effect of sex.

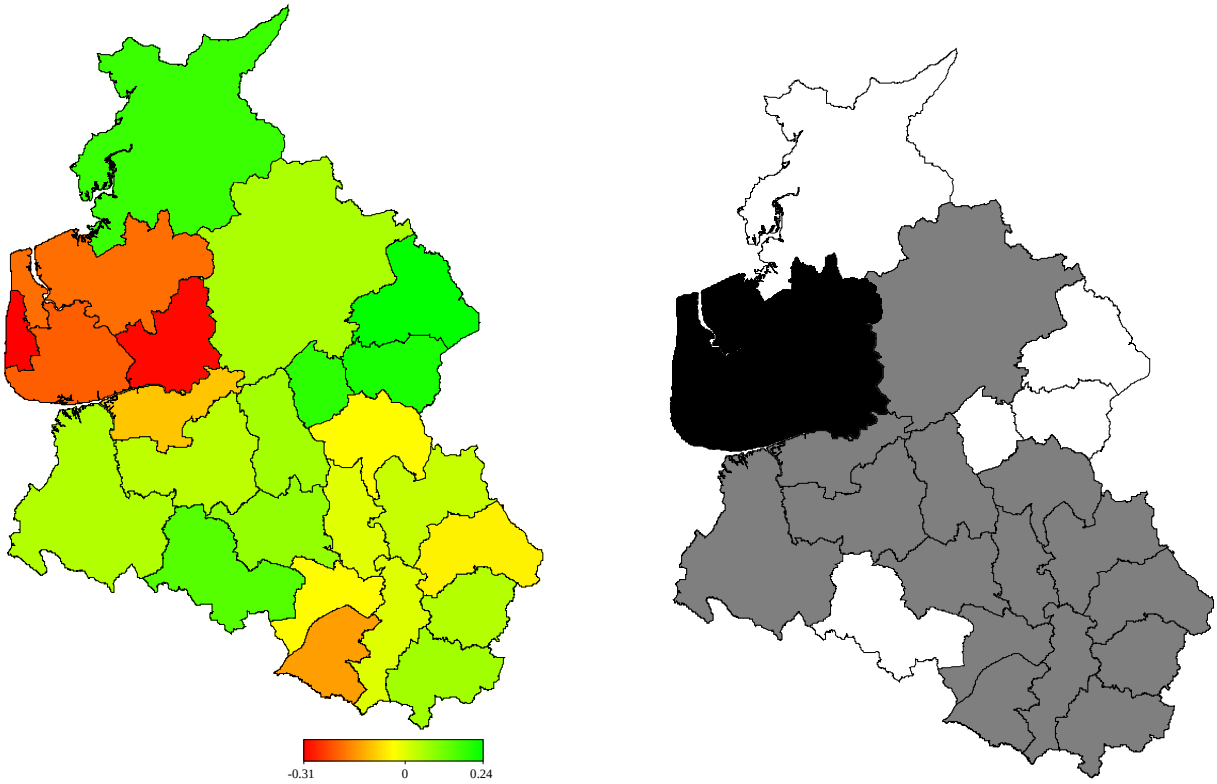

Figure 3: Leukemia Survival data: Spatial effect based on a district-level analysis and pointwise significance map. Black denotes districts with strictly negative credible intervals, white denotes districts with strictly positive credible intervals.

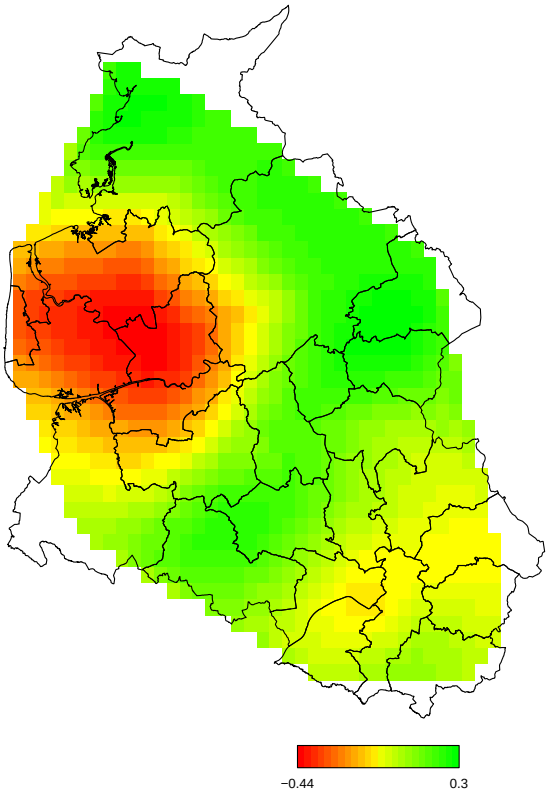

Figure 4: Leukemia Survival data: Spatial effect based on an individual-level analysis.

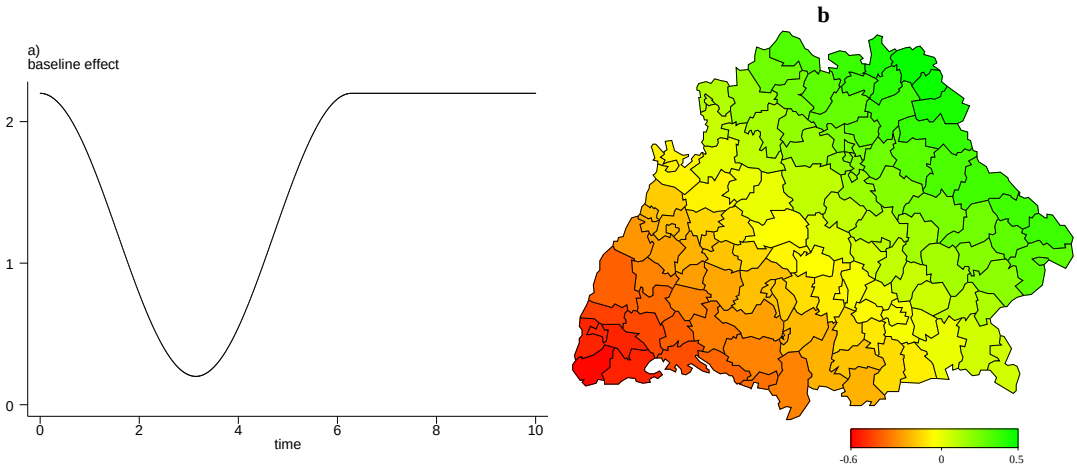

Figure 5: Simulation Study: True baseline hazard and true spatial effect.

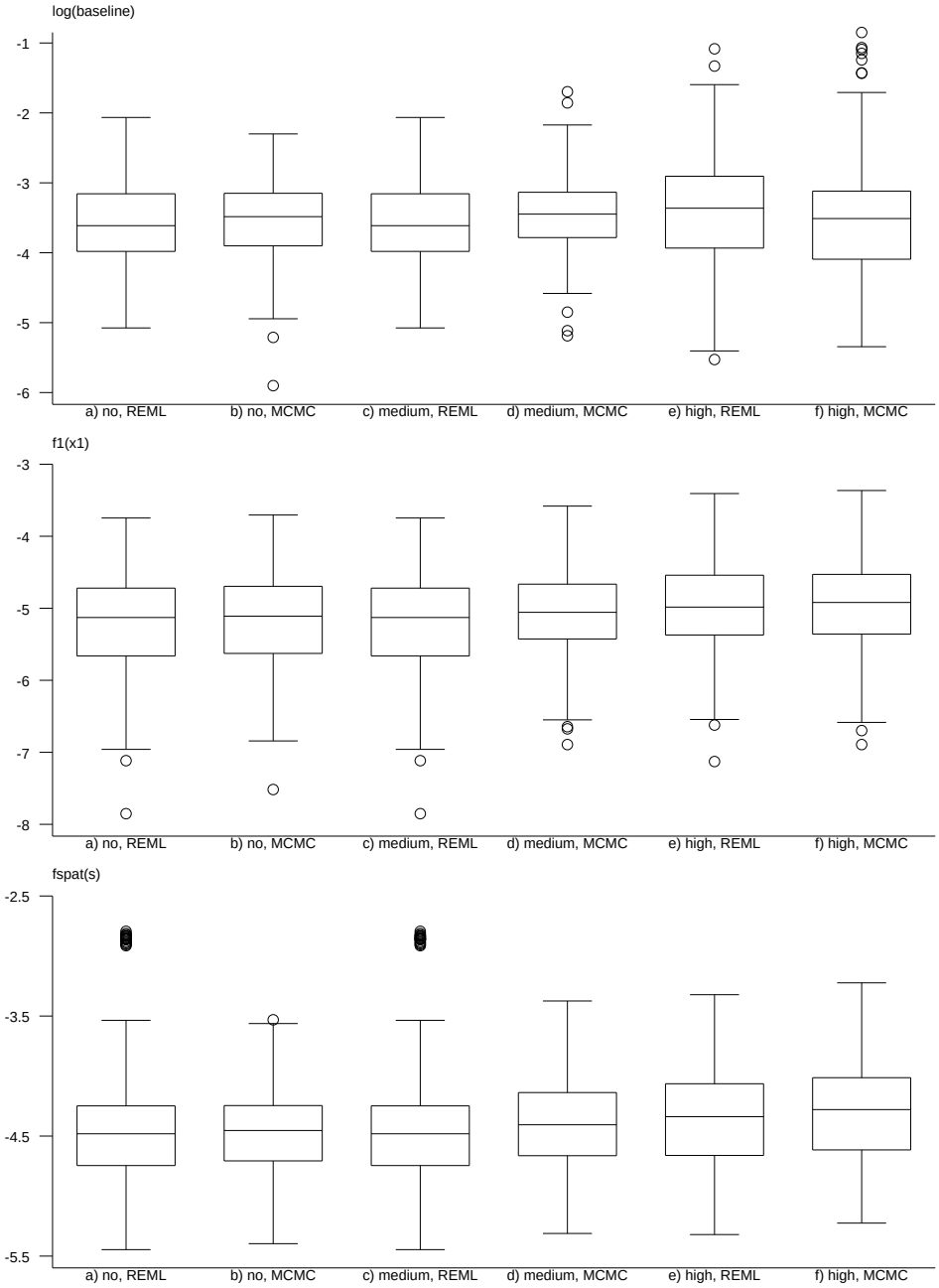

Figure 6: Simulation Study: Boxplots of log(MSE) in the case of no censoring (left two boxplots), medium censoring (third and fourth boxplot) and high censoring (right two boxplots).

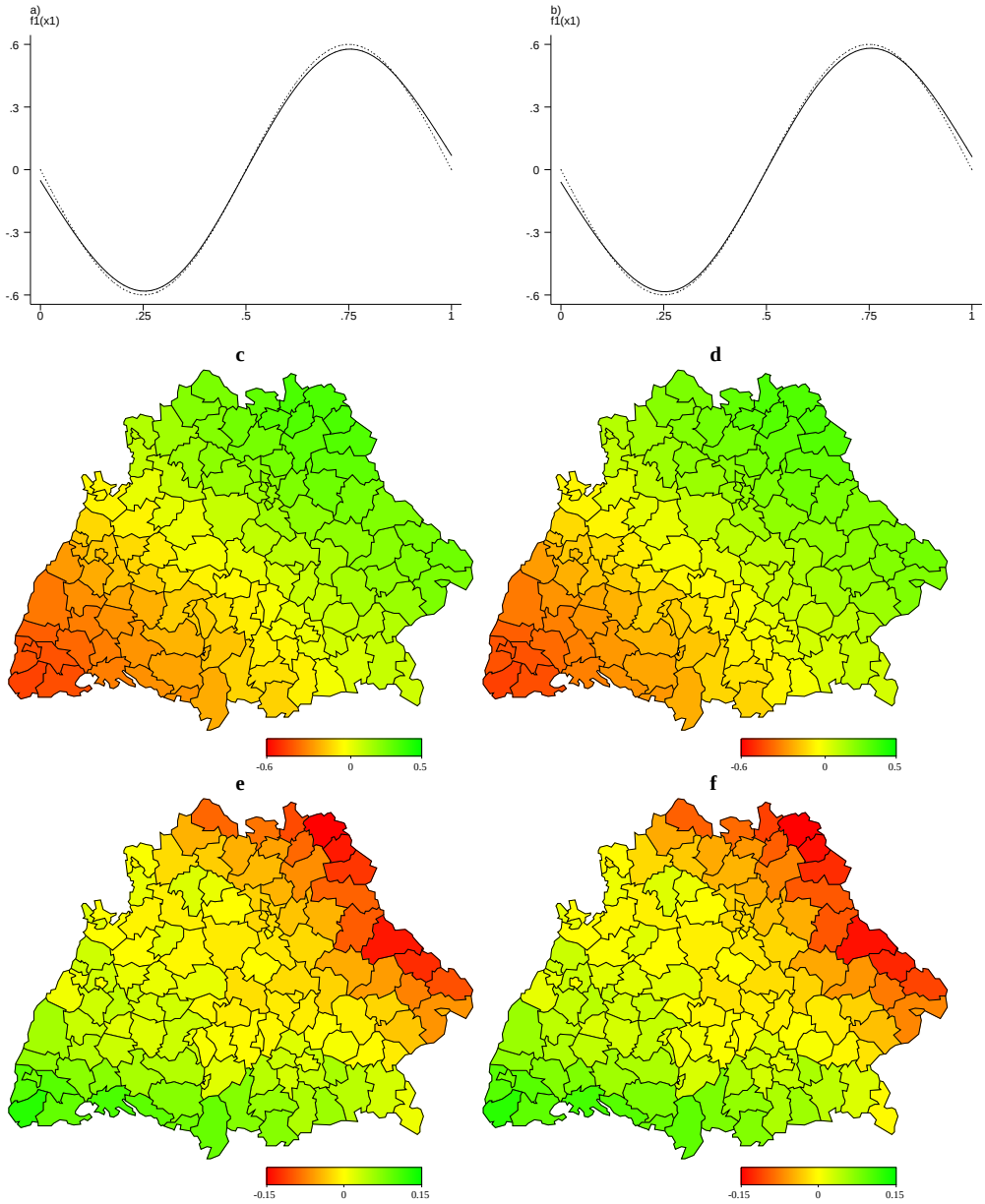

Figure 7: Simulation Study: Bias for $f_1(x)$ (upper panel), average estimates for f_{spat} (middle panel) and bias for f_{spat} (lower panel) in the case of high censoring. Results for the mixed model approach are displayed in the left panel, results for the MCMC approach in the right panel.