Geoadditive hazard regression for interval censored survival times

Thomas Kneib
Department of Statistics, University of Munich

Abstract

The Cox proportional hazards model is probably the most commonly used method when analyzing the impact of covariates on continuous survival times. In its classical form, the Cox model was introduced in the setting of right-censored observations. However, in practice other sampling schemes are frequently encountered and therefore extensions allowing for interval and left censoring or left truncation are clearly desired. Furthermore, many applications require a more flexible modeling of covariate information than the usual linear predictor. For example, effects of continuous covariates are likely to be of nonlinear form or spatial information is to be included appropriately. Further extensions should allow for time-varying effects of covariates or covariates that are themselves time-varying. Such models relax the assumption of proportional hazards. We propose a regression model for the hazard rate that combines and extends the above-mentioned features on the basis of a unifying Bayesian model formulation. Nonlinear and time-varying effects as well as the baseline hazard rate are modeled by penalized splines. Spatial effects can be included based on either Markov random fields or stationary Gaussian random fields. The model allows for arbitrary combinations of left, right and interval censoring as well as left truncation. Estimation is based on a reparameterisation of the model as a variance components mixed model. The variance parameters corresponding to inverse smoothing parameters can then be estimated based on an approximate marginal likelihood approach. As an application we present an analysis on childhood mortality in Nigeria, where the interval censoring framework also allows to deal with the problem of heaped survival times caused by memory effects. In a simulation study we investigate the effect of ignoring the impact of interval censored observations.

Key words: extended Cox model, interval censoring, left truncation, marginal likelihood, mixed models, geoadditive hazard regression, time-varying covariates

1 Introduction

When analyzing continuous survival times, the Cox proportional hazards model is the classical choice, if no parametric form for the distribution of the survival times can be assumed. While allowing for a flexible baseline hazard rate, the Cox model expects a parametric form for all covariate effects, which may be too restrictive in realistically complex applications. For example, in the present analysis of childhood mortality the effect of the mother's body mass index is often assumed to be of nonlinear form due to theoretical considerations. In addition, the data set contains spatial information on the observations and it is of interest to judge whether spatial variation remains unexplained by the covariates considered in the analysis. Furthermore the baseline hazard rate itself is of interest in this specific application and therefore joint estimation of the baseline hazard rate and covariate effects is desirable.

Several proposals for the analysis of such geoadditive survival data have been made in the last years. Henderson, Shimakura & Gorst (2002) propose a Cox model with gamma frailties, where the frailty 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 & Carlin (2003) and Carlin & Banerjee (2002) combine MRF and GRF priors for the spatial component with nonparametric estimation of the baseline hazard rate. Effects of continuous covariates are still assumed to be of linear parametric form. Full and empirical Bayes inference in hazard regression models that can deal with all the aforementioned issues have been developed by Hennerfeind, Brezger & Fahrmeir (2004) and Kneib & Fahrmeir (2004), respectively.

While most of the recent literature on geoadditive survival data deals only with the classical case of right-censored observations, other sampling schemes are often encountered in practice. For example, almost all uncensored survival times in our exemplary data set are given in months because the data were collected using a retrospective questionnaire of the mother. In contrast, censoring times of right-censored observations are available in days. A possible way to deal with this problem, is to treat the survival times as interval censored. In addition, the problem of heaped survival times, caused by memory effects due to the retrospective design of the study, could easily be incorporated in the interval censoring framework.

Cai & Betensky (2003) present a mixed model approach to estimate the baseline hazard rate in the presence of interval censoring based on penalized splines. Their model also allows for the inclusion of parametric covariate effects. An extended class of hazard regression models is described in Kooperberg & Clarkson (1997). The baseline hazard

rate, covariate effects and time-varying effects are approximated by linear splines. Tensor product splines can be used to model interaction surfaces. Smoothness of the estimated curves and surfaces is not ensured via penalization but through a variable selection procedure based on information criteria. A Bayesian approach to correlated interval censored survival times is presented in Komárek et al. (2005). While interval censoring is modeled via data augmentation, frailties are used to incorporate correlations. Transformation models for interval censored survival times in combination with a generalized estimating equations approach to account for correlations are described in Bogaerts et al. (2002).

In this paper, we propose an extended geoadditive Cox model that combines the following features:

- the ability to deal with arbitrary combinations of left, right, and interval censoring as well as left truncation,
- joint estimation of covariate effects and baseline hazard rate,
- the possibility to include (piecewise constant) time-varying covariates,
- relaxation of the proportional hazards assumption via the inclusion of time-varying effects,
- estimation of non-linear effects of continuous covariates based on penalized splines,
- estimation of spatial effects based on Markov random fields, stationary Gaussian random fields, and two-dimensional extensions of penalized splines,
- further model components such as cluster-specific frailties, interaction surfaces or varying coefficient terms.

Inference in this extended Cox model is based on a unified Bayesian formulation of the different model components that supplements all effects with appropriate priors of different degrees of smoothness but one general form. This general form allows to rewrite the model as a variance components model where regression coefficients can be estimated based on penalized likelihood. The smoothing parameters of the original model formulation transform to variance components in the mixed model and are estimated jointly with the regression coefficients using (approximate) marginal likelihood. The presented methodology is implemented in BayesX, a public domain software package for Bayesian inference, available from `http://www.stat.uni-muenchen.de/~bayesx`[1].

Section 2 describes geoadditive hazard regression models and likelihood contributions for different censoring schemes. Section 3 gives details on the mixed model based inferential procedure. A simulation study investigating the effect of ignoring interval censoring

[1]The described methodology will be available in release 1.4 of BayesX

is conducted in Section 4 and Section 5 presents an application that demonstrates the flexibility of geoadditive hazard regression models. The concluding section comments on directions of future research.

2 Geoadditive Hazard Regression

2.1 Hazard Rate Model

Since the publication of the seminal paper of Cox (1972) influences of covariates on survival times are commonly described by a regression model for the hazard rate. The Cox proportional hazards model assumes the multiplicative structure

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

where $\lambda_0(t)$ is an unspecified smooth baseline hazard rate and $v'\gamma$ is a linear predictor formed of covariates v and regression coefficients γ. On the line of additive regression models, the Cox model can be extended to

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

where i is an observation index and $\eta_i(t)$ is a geoadditive predictor of the form

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

Here $g_0(t) = \log(\lambda_0(t))$ is the log-baseline hazard, $g_l(t)$ represent time-varying effects of covariates u_{il}, $f_j(x_{ij})$ are nonlinear effects of continuous covariates, $f_{spat}(s_i)$ is a spatial effect, and $v_i'\gamma$ corresponds to covariate effects that are modeled in the usual parametric way. Nonparametric effects f_j as well as time-varying effects $g_0(t)$ and $g_l(t)$ are estimated based on penalized splines, see Section 2.2.1. Spatial effects can be estimated either based on Markov random field priors or Gaussian random field priors, see Section 2.2.2. A number of further extensions, such as interaction surfaces or cluster-specific frailties can be included in the predictor (3) and are also supported in our implementation (see Section 2.2.3).

To obtain a compact formulation of geoadditive hazard regression models and to ease the description of inferential details in Section 3, we introduce some matrix notation. All different effects in (3) can be cast into one general form, and therefore each vector of function evaluations can be written as the product of a design matrix Z and a possibly high-dimensional vector of regression coefficients β. Thus, after appropriate reindexing, the predictor (3) can be rewritten as

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

where $V\gamma$ represents parametric effects while each of the terms $Z_j\beta_j$ represents a non-parametric, time-varying or spatial effect.

2.2 Priors

From a Bayesian perspective, specification of model (4) is completed by assigning appropriate priors to the regression coefficients γ and β_j. While diffuse priors are assigned to fixed effects γ, priors for the remaining effects can be expressed in the general form of a multivariate Gaussian distribution, i.e.

$$p(\beta_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right). \tag{5}$$

The precision matrix K_j acts as a penalty matrix and shrinks parameters towards zero or penalizes too abrupt jumps between adjacent parameters. The variance parameter τ_j^2 can be interpreted analogously to a smoothing parameter with large (small) values corresponding to rough (smooth) estimates. From a frequentist perspective, assuming prior (5) is equivalent to specifying β_j as a correlated random effect. However, since K_j is in general rank-deficient, the random effects distribution may be partially improper.

2.2.1 Continuous covariates and time-varying effects

Effects of continuous covariates as well as time-varying effects are often assumed to vary smoothly over their codomain. One possibility to express this prior knowledge is the usage of penalized splines (Eilers & Marx 1996), where a function $f_j(x_j)$ or $g_l(t)$ is approximated by a polynomial spline of degree l. Such a polynomial spline can be written as a sum of basis functions B_m defined on a grid of equally spaced knots $x_{min} = \kappa_0 < \kappa_1 < \ldots < \kappa_s = x_{max}$, i.e.

$$f_j(x_j) = \sum_{m=1}^{l+s} \beta_{jm} B_m(x_j). \tag{6}$$

To ensure smoothness of the fitted curve, a moderately large number of knots is used in combination with penalization of adjacent regression coefficients based on a difference penalty. In a Bayesian formulation, the difference penalty can be replaced by the assumption of first or second order random walks, see (Lang & Brezger 2004) for details. The joint distribution of the regression coefficients can then be shown to be of form (5) with $K_j = D'D$, where D is a first or second order difference matrix. Since linear (constant) effects are not penalized by K_j if a second (first) order random walk is employed, the precision matrix has a two-(one-)dimensional null space. The design matrix Z_j contains the B-spline basis functions evaluated at the observed covariate values, i.e. $Z_j[i,m] = B_m(x_{ij})$.

2.2.2 Spatial effects

For spatial effects we distinguish two situations: Either spatial information is given exactly in terms of longitude and latitude or the observations can be assigned to a finite number of regions or sites.

In the first case, spatial effects can be constructed as in classical geostatistical models (kriging) based on zero-mean stationary Gaussian stochastic processes $\{\beta_s^{spat} : s \in \mathbb{R}^2\}$. Due to the assumption of normality, the prior distribution of the spatial effect is completely determined by its variance τ^2_{spat} and a correlation function $\rho(\beta_s^{spat}, \beta_{s'}^{spat})$. In many applications isotropy of the correlation function is a reasonable assumption, i.e. $\rho(\beta_s^{spat}, \beta_{s'}^{spat}) = \rho(||s - s'||)$ depends only on the Euclidean distance of the two sites and not on their direction and location. Kriging terms can be cast into the general form (5) with $K_{spat} = C^{-1}$ and $C[i, j] = \rho(||s_i - s_j||)$. In this case K_{spat} is of full rank and the corresponding prior distribution is proper. The design matrix Z_{spat} is a 0/1-incidence matrix, i.e. its value in the i-th row and the s-th column is 1 if the i-th observation is located at site s, and zero otherwise.

If observations are clustered in geographical regions, Markov random field (MRF) priors can be used to induce spatial correlations among observations. In contrast to GRFs correlations are not modeled explicitly but via an extension of random walks to two dimensions. If δ_s denotes the set of neighbors of region s, a MRF assumes

$$\beta_s|\beta_{s'},\ s' \neq s, \tau^2_{str} \sim N\left(\frac{1}{N_s}\sum_{s'\in\partial_s}\beta_{s'}, \frac{\tau^2_{str}}{N_s}\right). \tag{7}$$

Therefore the expected value of the spatial function at site s is given by the (unweighted) average of the adjacent sites. Extensions of the basic MRF (7) allow for weighted averages but are less often used in practice. Whether two regions are neighbors is most commonly decided by the existence of a common boundary. The design matrix Z_{spat} is again a 1/0-incidence matrix and K_{spat} has the form of an adjacency matrix.

Although presented separately, approaches for exact locations can be used in the case of connected geographical regions too, e.g. based on the centroids of the regions. Conversely, we can also apply MRFs to exact locations if neighborhoods are defined by a distance measure or via discretisation of the observation area. The main difference between GRFs and MRFs, considering their numerical properties, is the dimension of the penalty matrix. For MRFs the dimension of K_{spat} equals the number of different regions and is therefore independent from the sample size. On the other side, for GRFs, the dimension of K_{spat} is given by the number of distinct locations, which in most cases is close or equal to the sample size. To overcome the numerical problems that arise from the large number of regression coefficients involved in a GRF, Kammann & Wand (2003) proposed low-rank

kriging, where a space-filling algorithm is used to reduce the dimension of the estimation problem. Again, low-rank kriging can be cast into the presented general framework.

2.2.3 Extensions

Several extensions of the basic model (3) are conceivable and supported by the presented framework. For example, i.i.d. cluster-specific frailties with Gaussian prior are a special case of (5). Furthermore, interaction surfaces based on two-dimensional extensions of penalized splines or varying coefficient terms with either spatial or continuous effect modifiers can be included in the predictor (3). Note that the time-varying effects in (3) can also be subsumed in the varying coefficients framework. Further details on extended modelling of covariate effects and the inclusion in the presented framework are discussed in Fahrmeir, Kneib & Lang (2004) in the context of regression models for univariate responses from exponential families.

2.3 Likelihood Contributions

Usually, the Cox model and extensions are developed for right-censored observations. More formally spoken, if the true survival time is given by T and C is a censoring time, only $\tilde{T} = \min(T, C)$ is observed along with the censoring indicator $\delta = \mathbb{1}_{(T \leq C)}$. Many applications, however, confront the analyst with more complicated data structures involving more general censoring schemes. For example, interval censored survival times T are not observed exactly but are only known to fall into an interval $[T_{lo}, T_{up}]$. If $T_{lo} = 0$ such survival times are also referred to as being left censored. Furthermore, each of the censoring schemes may appear in combination with left truncation of the corresponding observation, i.e. the survival time is only observed if it exceeds the truncation time T_{tr}. Accordingly, some survival times are not observable and the likelihood has to be adjusted appropriately. Figure 1 illustrates the different censoring schemes we will consider in the following: The true survival time is given by T which is observed for individual 1 and 2. While individual 1 is not truncated, individual 2 is left truncated at time T_{tr}. Similarly, individuals 3 and 4 are right-censored at time C and individuals 5 and 6 are interval censored with interval $[T_{lo}, T_{up}]$ with the same pattern for left truncation.

In a general framework an observation can now be described completely by the quadruple $(T_{tr}, T_{lo}, T_{up}, \delta)$, with

$$
\begin{array}{ll}
T_{lo} = T_{up} = T,\ \delta = 1 & \text{if the observation is uncensored,} \\
T_{lo} = T_{up} = C,\ \delta = 0 & \text{if the observation is right censored,} \\
T_{lo} < T_{up},\ \delta = 0 & \text{if the observation is interval censored.}
\end{array}
$$

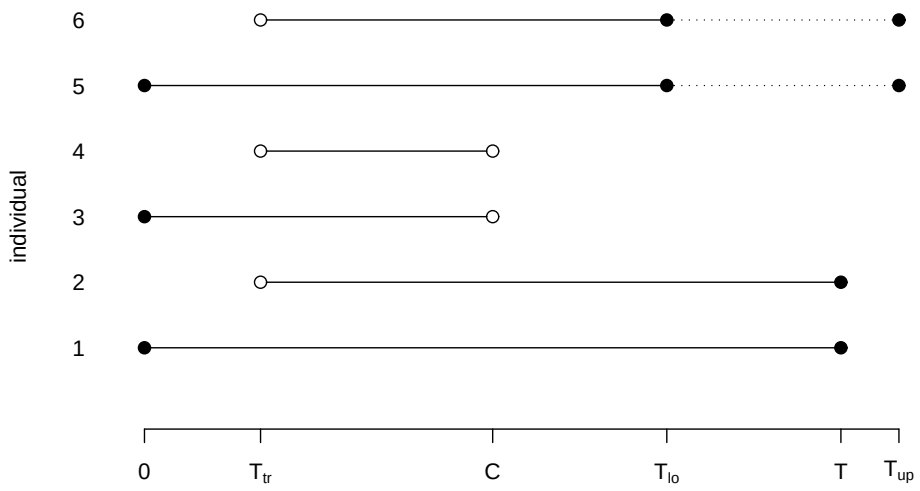


Figure 1: Illustration of different censoring schemes: For individuals 1 and 2 the true survival time T is observed, individuals 3 and 4 are right censored at time C, and individuals 5 and 6 are interval censored, where the interval is given by $[T_{lo}, T_{up}]$. Individuals 2, 4 and 6 are left truncated at time T_{tr}.

For left truncated observations we have $T_{tr} > 0$ and $T_{tr} = 0$ for observations which are not truncated.

Based on these definitions we can now construct the likelihood contributions for the different censoring schemes in terms of the hazard rate (2) and the survivor function $S(t) = \exp(\int_0^t \lambda(u)du)$. Under the common assumption of noninformative censoring and conditional independence, the likelihood for $\beta = (\gamma', \beta_1', \ldots, \beta_p')'$ is given by

$$L(\beta) = \prod_{i=1}^{n} L_i(\beta), \tag{8}$$

where

$$L_i(\beta) = \lambda(T_{up})S(T_{up})/S(T_{tr}) = \lambda(T_{up}) \exp\left(-\int_{T_{tr}}^{T_{up}} \lambda(t)dt\right)$$

for an uncensored observation,

$$L_i(\beta) = S(T_{up})/S(T_{tr}) = \exp\left(-\int_{T_{tr}}^{T_{up}} \lambda(t)dt\right)$$

for a right censored observation and

$$L_i(\beta) = (S(T_{lo}) - S(T_{up}))/S(T_{tr}) = \exp\left(-\int_{T_{tr}}^{T_{lo}} \lambda(t)dt\right)\left(1 - \exp\left(-\int_{T_{lo}}^{T_{up}} \lambda(t)dt\right)\right)$$

for an interval censored observation. Note that for explicit evaluation of the likelihood (8) some numerical integration technique has to be employed, since none of the integrals can in general be solved analytically.

The above notation also allows for the easy inclusion of piecewise constant, time-varying covariates via some data augmentation. Noting that

$$\int_{T_{tr}}^{T} \lambda(t)dt = \int_{T_{tr}}^{t_1} \lambda(t)dt + \int_{t_1}^{t_2} \lambda(t)dt + \ldots + \int_{t_{p-1}}^{t_p} \lambda(t)dt + \int_{t_p}^{T} \lambda(t)dt$$

for $T_{tr} < t_1 < \ldots < t_q < T$, we can replace an observation $(T_{tr}, T_{lo}, T_{up}, \delta)$ by a set of new observations $(T_{tr}, t_1, t_1, 0)$, $(t_1, t_2, t_2, 0)$, $\ldots (t_{p-1}, t_p, t_p, 0)$, $(t_p, T_{lo}, T_{up}, \delta)$ without changing the likelihood. Therefore, observations with time-varying covariates can be split up into several observations, where the values $t_1 < \ldots < t_p$ are defined by the changepoints of the covariate and the covariate is now time-constant on each of the intervals. In theory, other paths for a covariate $x(t)$ than piecewise constant ones are also possible, if $x(t)$ is known for $T_{tr} \leq t \leq T_{lo}$. In this case the the likelihood (8) can also be evaluated numerically but a general path $x(t)$ may lead to complicated data structures.

Figure 2 illustrates the data augmentation step for a left truncated, uncensored observation and a covariate $x(t)$ that takes the three different values x_1, x_2 and x_3 on the three intervals $[T_{tr}, t_1], [t_1, t_2]$ and $[t_2, T_{up}]$. Here, the original observation $(T_{tr}, T_{up}, T_{up}, 1)$ has to be replaced by $(T_{tr}, t_1, t_1, 0)$, $(t_1, t_2, t_2, 0)$ and $(t_2, T_{up}, T_{up}, 1)$.

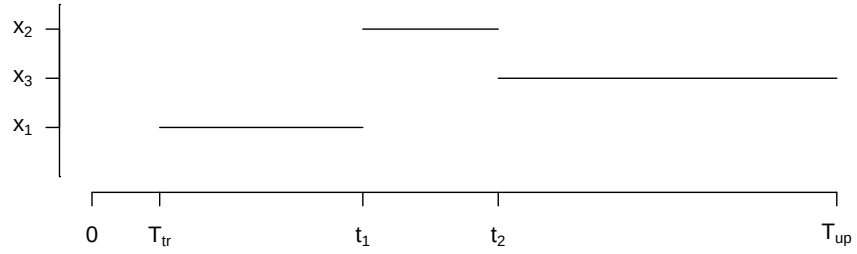


Figure 2: Illustration of time-varying covariates: Covariate $x(t)$ takes the three different values x_1, x_2 and x_3 on the three intervals $[T_{tr}, t_1], [t_1, t_2]$ and $[t_2, T_{up}]$.

Combining prior information and the likelihood contributions given above finally leads to the posterior

$$L_{pen}(\beta) = L(\beta) \prod_{j=1}^{p} p(\beta_j | \tau_j^2), \tag{9}$$

which has to be maximized to obtain posterior mode or empirical Bayes estimates. Note that the form of posterior (9) is similar to a penalized likelihood where penalty terms are based on the log-priors and so posterior mode estimation is closely related to penalized likelihood estimation.

3 Mixed Model based Inference

Since in most cases at least some of the effects in (3) exhibit improper priors, geoadditive hazard regression model can not be directly estimated using mixed model methodology for survival times. Instead we first have to reparameterise the model to obtain proper priors, see the next subsection. Within the obtained proper mixed model, estimates for regression coefficients and variance components can be derived by iterating between the following two steps: Given the current values of the variances, estimates of the regression coefficients are computed via maximization of a penalized likelihood based on a Newton-Raphson-algorithm (subsection 3.2). Conversely, updated estimates for the variances given the regression coefficients are obtained by a Fisher-Scoring-type algorithm (subsection 3.3) that maximizes the (approximate) marginal likelihood of the variances. This way of estimating nonparametric or spatial effects based on mixed models has become quite popular in the context of generalized linear models throughout the last years (compare e.g. Lin & Zhang (1999), Ruppert, Wand & Carroll (2003), Kammann & Wand (2003) or Kneib & Fahrmeir (2004)). While estimation of regression coefficients could also be performed within the original formulation of geoadditive hazard regression, estimation of the variances relies heavily on the reparameterisation, since improper priors do not allow for marginal likelihood estimation.

3.1 Mixed Model Representation

In the following we assume that β_j has dimension d_j and the corresponding penalty matrix has rank $k_j \leq d_j$. To rewrite the geoadditive predictor (3) we proceed as follows: Each vector of regression coefficients β_j is decomposed into two parts, i.e.

$$\beta_j = Z_j^{unp}\beta_j^{unp} + Z_j^{pen}\beta_j^{pen} \tag{10}$$

with a $d_j \times (d_j - k_j)$ matrix Z_j^{unp} and a $d_j \times k_j$ matrix Z_j^{pen}. Choosing appropriate matrices in (10) results in a $(d_j - k_j)$-dimensional vector β_j^{unp} with a flat prior and a k_j-dimensional vector β_j^{pen} with i.i.d. Gaussian prior, i.e.

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

While β_j^{unp} captures the part of f_j that is not penalized by K_j, β_j^{unp} represents the orthogonal deviation from this unpenalized part. Accordingly, the matrices Z_j^{unp} and Z_j^{pen} can be constructed based on the spectral decomposition of the penalty matrix K_j. To be more specific, Z_j^{unp} contains a basis of the null space of K_j and Z_j^{pen} is build from a basis of the orthogonal deviation from this null space (compare Kneib & Fahrmeir (2004) for details).

Applying decomposition (10) to all components of the additive predictor (4) yields

$$\begin{aligned}\eta &= V\gamma + \sum_{j=1}^{p} Z_j Z_j^{unp}\beta_j^{unp} + Z_j Z_j^{pen}\beta_j^{pen} \\ &= X\beta^{unp} + Z\beta^{pen}\end{aligned}$$

where $X = (V, Z_1 Z_1^{unp}, \ldots, Z_p Z_p^{unp})$, $Z = (Z_1 Z_1^{pen}, \ldots, Z_p Z_p^{pen})$, $\beta^{unp} = (\beta_1^{unp}, \ldots, \beta_p^{unp})$ and $\beta^{pen} = (\beta_1^{pen}, \ldots, \beta_p^{pen})$. This is a variance components model with distributional assumptions

$$p(\beta^{unp}) \propto const \qquad \text{and} \qquad \beta^{pen} \sim N(0, \Sigma),$$

where $\Sigma = \text{blockdiag}(\tau_1^2 I, \ldots, \tau_p^2 I)$.

3.2 Regression Coefficients

To construct a Newton-Raphson update step for the regression coefficients, we need first and second derivatives of (9) with respect to β^{unp} and β^{pen}. To ease notation, consider for the moment a hazard rate of the form

$$\lambda(t) = \exp(x(t)'\beta)$$

which essentially reflects the structure of a structured hazard regression model. Defining

$$D_j(t) = -\frac{\partial}{\partial \beta_j}\int_0^t \lambda(u)du = -\int_0^t x_j(u)\lambda(u)du$$

and

$$E_{jk}(t) = -\frac{\partial^2}{\partial \beta_j \partial \beta_k}\int_0^t \lambda(u)du = -\int_0^t x_j(u)x_k(u)\lambda(u)du,$$

first and second derivatives of the log-likelihood contributions for uncensored and right censored observations are given by

$$\delta \cdot x_j(T_{up}) + D_j(T_{up}) - D_j(T_{tr}) \quad \text{and} \quad E_{jk}(T_{up}) - E_{jk}(T_{tr}).$$

For interval censored survival times formulae become more complicated. Here, first and second derivatives of the log-likelihood contributions can be shown to equal

$$D_j(T_{lo}) - D_j(T_{tr}) - \frac{\exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]\left[D_j(T_{lo}) - D_j(T_{up})\right]}{1 - \exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]}$$

and

$$\begin{aligned}E_{jk}(T_{lo}) - E_{jk}(T_{tr}) &- \frac{\exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]^2\left[D_j(T_r) - D_j(T_l)\right]\left[D_k(T_r) - D_k(T_l)\right]}{\left\{1 - \exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]\right\}^2} \\ &- \frac{\exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]\left\{\left[D_k(T_r) - D_k(T_l)\right]\left[D_j(T_r) - D_j(T_l)\right] + \left[E_{jk}(T_r) - E_{jk}(T_l)\right]\right\}}{1 - \exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]}.\end{aligned}$$

Note that for $T_{tr} = 0$ these results are equivalent to those presented in Kooperberg & Clarkson (1997). To evaluate the derivatives, we again have to employ some numerical integration rule. Due to its simplicity, we used the trapezoidal rule based on an equidistant set of knots in our implementation.

3.3 Variance Components

The main benefit of the mixed model representation of structured hazard regression models is the possibility to estimate the variance parameters based on methodology for mixed models. Most commonly, this is achieved via maximization of the marginal likelihood

$$L^{marg}(\Sigma) = \int L_{pen}(\beta^{unp}, \beta^{pen}, \Sigma) d\beta^{unp} d\beta^{pen} \tag{11}$$

with respect to the variances contained in Σ. In Gaussian regression models this is equivalent to restricted maximum likelihood estimation of the variances (Harville 1974).

Direct maximization of (11) is in general intractable, since the high-dimensional integral can not be evaluated, neither analytically nor numerically. Instead we apply a Laplace approximation to the marginal likelihood yielding

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

Assuming that $l(\hat{\beta}^{unp}, \hat{\beta}^{pen})$ and $\hat{\beta}^{pen}$ vary only slowly when changing the variance components allows for a further reduction of the marginal likelihood to

$$l^{marg}(\Sigma) \approx -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|H| - \frac{1}{2}\beta^{pen\prime}\Sigma^{-1}\beta^{pen}, \tag{12}$$

where β^{pen} denotes a fixed value not depending directly on the variances, e.g. a current estimate. This approximation was found to work well for right censored survival times (see Kneib & Fahrmeir (2004)) and also proved to result in reasonable estimates in our general setting (compare the simulation study in the next section).

Since the approximate marginal likelihood (12) is of the same form regardless of the special type of censoring involved, first and second derivatives of (12) can be computed in complete analogy to Kneib & Fahrmeir (2004) to construct a Fisher-Scoring algorithm.

4 Ignoring interval censoring: A simulation Study

To investigate the impact of ignoring interval censoring when analyzing survival data, we conducted a simulation study that mimics a situation frequently found in clinical studies: The survival status of a patient is assessed at fixed dates until the end of the study. Exact survival times were generated from a geoadditive model with hazard rate

$$\lambda(t; x, s) = \exp(g_0(t) + f(x) + f_{spat}(s)),$$

where $g_0(t)$ is the log-baseline hazard rate, $f(x)$ is a function of the continuous covariate x with sinusoidal form and $f_{spat}(s)$ is a spatial function defined by the density of a mixture of two two-dimensional normal distributions. Two different baseline hazard rates were

applied: A bathtub-shaped one with strong variation over the whole time-domain and a relatively flat one. All survival times exceeding 8 were treated as right-censored at $C = 8$. The remaining interval $[0, 8]$ was divided into l equidistant intervals and each observation was assigned to the interval, the corresponding survival time pertained to. To evaluate the impact of interval censoring, we compared three different values for l, namely $l = 8$, $l = 16$ and $l = 32$ corresponding to intervals with length 1, 0.5 or 0.25. The simulation design is summarized in more detail in Figure 3.

The resulting data sets were analyzed based on three different strategies:

- Use the correct censoring mechanisms, i.e. treat all observations with survival time less than 8 as interval censored and all other observations as right censored (IC).
- Use a binary discrete-time survival model with complementary log-log-link. Such a model can be seen as a grouped Cox-model (compare e.g. Fahrmeir & Tutz (2001)) (CLL).
- Treat all observations with survival time less than 8 as uncensored and all other observations as right censored. To account for the interval censoring, uncensored observations are spread randomly across the corresponding interval (UC).

Note that we also tried to treat all survival times less than 8 as uncensored without spreading the observations across the intervals. However, due to numerical problems this strategy could not be routinely applied and is therefore not included in the comparison (for details on the numerical problems see also the application in the next section). Both the log-baseline and the effect of x are modeled by cubic P-splines with second order random walk penalty and 20 inner knots. The spatial effect is estimated using Markov random field prior (7).

The results of the simulation study can be summarized as follows:

- In case of the bathtub-shaped baseline, the interval censoring approach leads to the best estimates for the baseline hazard rate. While the discrete time model performs comparably well for a sufficient large number of intervals, the uncensored approach remains dissatisfying (Figure 4 a)).
- In contrast, in case of the flat baseline, the discrete time model leads to the best estimates for the baseline for a small number of intervals. A medium number of intervals leads to comparable estimates of all approaches and for a large number of intervals the uncensored approach results in the best estimates (Figure 4 b)). A possible explanation for this surprising result will be discussed in the application of the next section.

- Considering covariate effects, both types of baseline hazard rates lead to similar conclusions and we therefore only show results for the bathtub-shaped baseline: For a sufficient large number of intervals, all strategies lead to comparable fit in terms of MSEs. For a smaller number of intervals the interval censoring approach performs better than the uncensored approach, but the discrete time analysis results in even somewhat better point estimates (Figure 4 c) and d)).

- Figure 5 shows a similar results based on the average estimates for the spatial function. While the uncensored approach introduces noticeably more bias for a small number of intervals, discrete time and interval censoring lead to comparable estimates. When increasing the number intervals, differences between the three strategies almost vanish.

Based on these results we come to the conclusion that the impact of interval censoring depends on the structure of the underlying model, especially on the baseline hazard rate. While details of the model may be lost by ignoring interval censoring for highly fluctuating baselines and a relatively small number of large intervals, this effect decreases for an increasing number of intervals. When the baseline is relatively flat, interval censoring does not per se lead to improved estimates as asserted in the next section.

5 Childhood mortality in Nigeria

This analysis is based on data collected within the 2003 Nigeria Demographic and Health Survey (DHS), which is a nationally representative survey concerning the health status of women in reproductive age (13-49 years) and their children. The survival time of the children is obtained from a retrospective interview of the mother and should (in theory) be known in days. However, due to memory effects, only survival times within the first two months are observed exactly while all other survival times are actually given in months. In contrast, right censoring times are given in exact days, since these could be computed from the date of the interview and the child's birth date. Because of this special structure of the data, a model based on interval censored survival times seems to be more appropriate than a classical Cox model including only right censored observations. In particular, all survival times exceeding two months are treated as interval censored, where the interval is determined by the first and the last day of the corresponding month.

An additional challenge of this survival data can be seen from Figure 6, which shows the absolute frequencies of the observed survival times in months. Obviously, a lot of survival times are heaped at the values 12, 18, 24, 36 and 48 while a much smaller number of deaths is recorded between these time points. Such a heaping effect occurs

quite commonly in retrospective studies on survival times and has to be incorporated appropriately to obtain valid estimates. Within the interval censoring framework this can easily be achieved by introducing larger intervals for the heaped observations. In the present analysis we assigned non-overlapping, symmetric intervals around the heaped values of 6 or 12 month length to the heaped survival times.

For the hazard rate we chose the geoadditive predictor

$$\eta = g_0(t) + f_1(bmi) + f_2(age) + f_3(bord) + f_4(size) + f_{spat}(s) + u(t)'\gamma$$

where $g_0(t)$ denotes the log-baseline hazard rate, $f_1, \ldots, f_4$ are functions of the continuous covariates 'body mass index of the mother' (bmi), 'age of the mother at birth' (age), 'number of the child in the birth order' (bord) and 'number of household members' (size). f_{spat} models a spatial effect based on the district s the mother lives in and $u(t)$ comprises fixed effects of numerous categorical covariates describing the economic situation of the family, circumstances at birth, and the breastfeeding behaviour of the mother. While most of these categorical covariates are time invariant, the duration of breastfeeding is described by a time-varying covariate which takes the value one as long as the child is breastfed zero otherwise. Using the findings from subsection 2.3 this can be easily included in the present model using data augmentation.

Both the log-baseline and nonparametric effects are modeled by cubic P-splines with 20 inner knots. The spatial effect is assumed to follow the Markov random field prior (7). Due to missing values, the final number of observations is given by $n = 5323$. 117 children die within the first two months and are therefore treated as uncensored. The 474 children that die within the remaining study time are treated as interval censored as described above.

To shorten discussion, we will not show results for the fixed effects but focus on results of nonparametric and spatial effects (see Figure 8). The effect of the maternal body mass index to be almost linear with a slightly increasing risk for children of a mother with high body mass index. However, since the credible intervals include a horizontal line, the influence of the body mass index might be neglectable. The remaining three nonparametric effects are of nonlinear but almost monotone functional form. While a higher age of the mother could be shown to induce an increased risk, both a higher number of the child in the birth order and a higher number of household members lead to decreased risk. While the former effect may be caused by an increased knowledge about childcare by the mother, the latter may reflect the fact that well-endowed households attract additional members. The range of the estimated spatial effect is very small and a pointwise significance map shows no districts with effects different from zero. It should however be noted, that in an analysis which only comprises a spatial effect and no other

covariates, a highly significant spatial pattern emerges. Therefore observations are clearly spatially correlated but the spatial variations is completely explained by the covariates considered.

Figure 7 shows the estimated log-baseline hazard rate for three different models: The first one (straight line) is exactly the model given above, where all observed death times beyond two months are treated as interval censored and heaping effects are incorporated. In the second model (dashed line), death times are treated as interval censored but the heaping effect is neglected. Finally the third model (dotted line) mimics model 1 but achieves the interval censoring by randomly spreading the death times across the corresponding interval (similarly as in the simulation study in section 4). Note that this model also accounts for heaping effects.

Obviously, ignoring the heaping effect leads to highly implausible results, with risk estimates approximating zero where no deaths are recorded. This problem also occurred in the simulation study when the right interval boundaries were to be used as exact survival times. Incorporating the heaping effect significantly reduces this phenomena but still leaves some fluctuations in the estimate which are not expected to reflect the true temporal development of the hazard rate. Surprisingly, model 3 leads to the most plausible, smooth estimate for the log-baseline. Probably this outcome, which reflects the superiority of the corresponding approach in the simulation study with a flat baseline, is caused by the additional information assumed in this model. Since all observed death times are treated as exactly observed, the model contains much more information than the corresponding model based on interval censoring which is therefore more susceptible to produce artificial behavior.

6 Discussion

We presented a rather general approach for the analysis of continuous survival times, both in terms of the functional form of covariate effects and the supported censoring schemes. Particularly the possibility to combine left truncation, right censoring and interval censoring considerably broadens the applicability of geoadditive hazard regression models. The results of our simulation studies showed that the inclusion of interval censoring can in some situations lead to substantially improved estimates, but also indicated some situations, in which interval censoring may not be the appropriate model choice.

In future work we plan to extend geoadditive regression models to the more general setting of multi-state models. This framework includes a number of well known model classes for the analysis of competing risks or event history analysis. Within such models a similar data structure as with interval censored survival times is frequently encountered: In many

applications exact transition times are not available and the states can only be observed at fixed time points. While the the likelihood of multi-state models can be easily calculated if transition times are observed exactly, the likelihood becomes much more complicated when interval censoring is present and additional numerical problems have to be addressed in order to obtain estimates of the parameters of interest.

Acknowledgement:
I thank Ludwig Fahrmeir and Susanne Heim for helpful discussions and comments, Samson Babatunde Adebayo for assistance with the childhood mortality data and Rudi Eichholz for first analyses of this data. Financial support from the German Science Foundation (DFG), Collaborative research center 386 "Statistical Analysis of Discrete Structures" is gratefully acknowledged.

References

Banerjee, S. & Carlin, B. P. (2003). Semiparametric spatio-temporal frailty modelling. *Environmetrics*, **14**, 523-535.

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

Bogaerts, K., Leroy, R., Lesaffre, E. & Declerck, D. (2002) Modelling tooth emergence data based on multivariate interval-censored data, *Statistics in medicine*, **21** 3775-3787.

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

Carlin, B. P. & 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. & Marx, B.D. (1996). Flexible smoothing using B-splines and penalties. *Statistical Science*, **11**, 89-121.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression: A Bayesian perspective. *Statistica Sinica*, **14**, 715-745.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models*, Springer, New York.

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

Henderson, R., Shimakura, S. & Gorst, D. (2002). Modelling Spatial variation in Leukemia Survival Data. *Journal of the American Statistical Association*, **97**, 965-972.

Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2004). Geoadditive survival models. Under revision for *Journal of the American Statistical Society*.

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

Komárek, A., Lesaffre, E., Härkänen, T., Declerck, D. and Virtanen, J. I. (2005). A Bayesian analysis of multivariate doubly-interval-censored dental data. *Biostatistics*, **6**, 145-155.

Kneib, T. & Fahrmeir, L. (2005) Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics*, to appear.

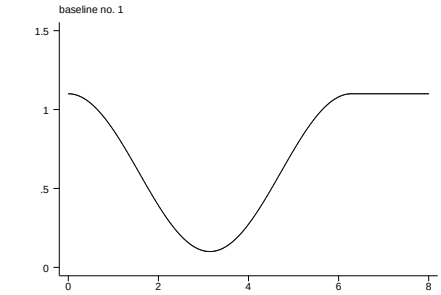
Kneib, T. & Fahrmeir, L. (2004) A mixed model approach for structured hazard regression. SFB 386 Discussion Paper 400, University of Munich.

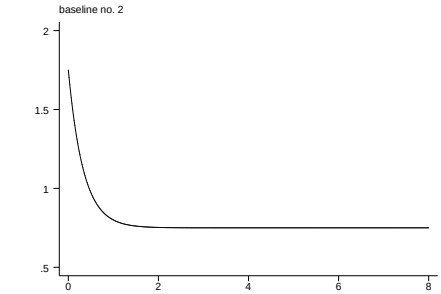
Kooperberg, C. & Clarkson, D. B. (1997) Hazard regression with interval-censored data. *Biometrics*m **53**, 1485-1494.

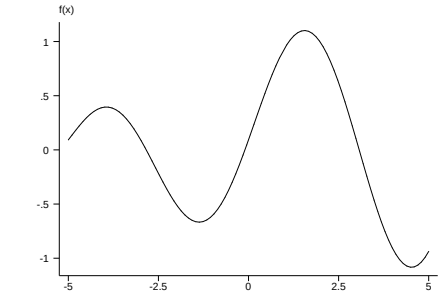
Lang, S. & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183-212.

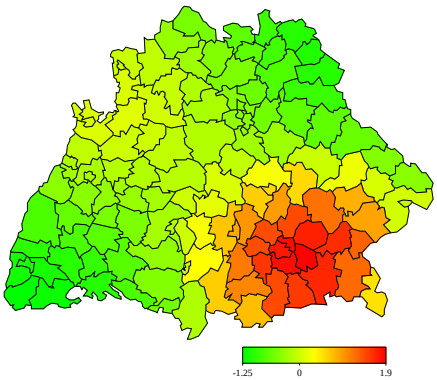
Lin, X. & Zhang, D. (1999) Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society B*, **61**, 381–400.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric Regression*, University Press, Cambridge.









- Hazard rate:

$$\lambda(t; x, s) = \exp(g_0(t) + f(x) + f_{spat}(s))$$

- Baseline no. 1:

$$\exp(g_0(t)) = \begin{cases} 0.5 \cdot [\cos(t) + 1.2] & , t \leq 2\pi \\ 0.5 \cdot [1 + 1.2] & , t > 2\pi \end{cases}$$

- Baseline no. 2:

$$\exp(g_0(t)) = \exp(-3t) + 0.75$$

- $f(x) = \sin(1.05 \cdot x) \cdot \log(x + 6)$

- x is chosen randomly from an equidistant of 100 values between -5 and 5.

- $f_{spat}(s) = N(\mu_1, \Sigma_1, s_x, s_y) + N(\mu_2, \Sigma_2, s_x, s_y) - 1.406$

$$\mu_1 = \begin{pmatrix} 0.75 \\ 0.25 \end{pmatrix} \quad \Sigma_1 = \begin{pmatrix} 0.05 & 0.01 \\ 0.01 & 0.05 \end{pmatrix}$$

$$\mu_2 = \begin{pmatrix} 0.25 \\ 0.75 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 0.1 & 0.01 \\ 0.01 & 0.1 \end{pmatrix}$$

- (s_x, s_y) are the centroids of the 124 districts s of the two southern states of Germany.

- Survival times exceeding 8 are considered as right-censored.

- The interval $[0, 8]$ is divided in $l = 8, 16$ or 32 equidistant parts for interval censoring.

- number of observations per replication: $n = 500$.

Figure 3: Simulation design.

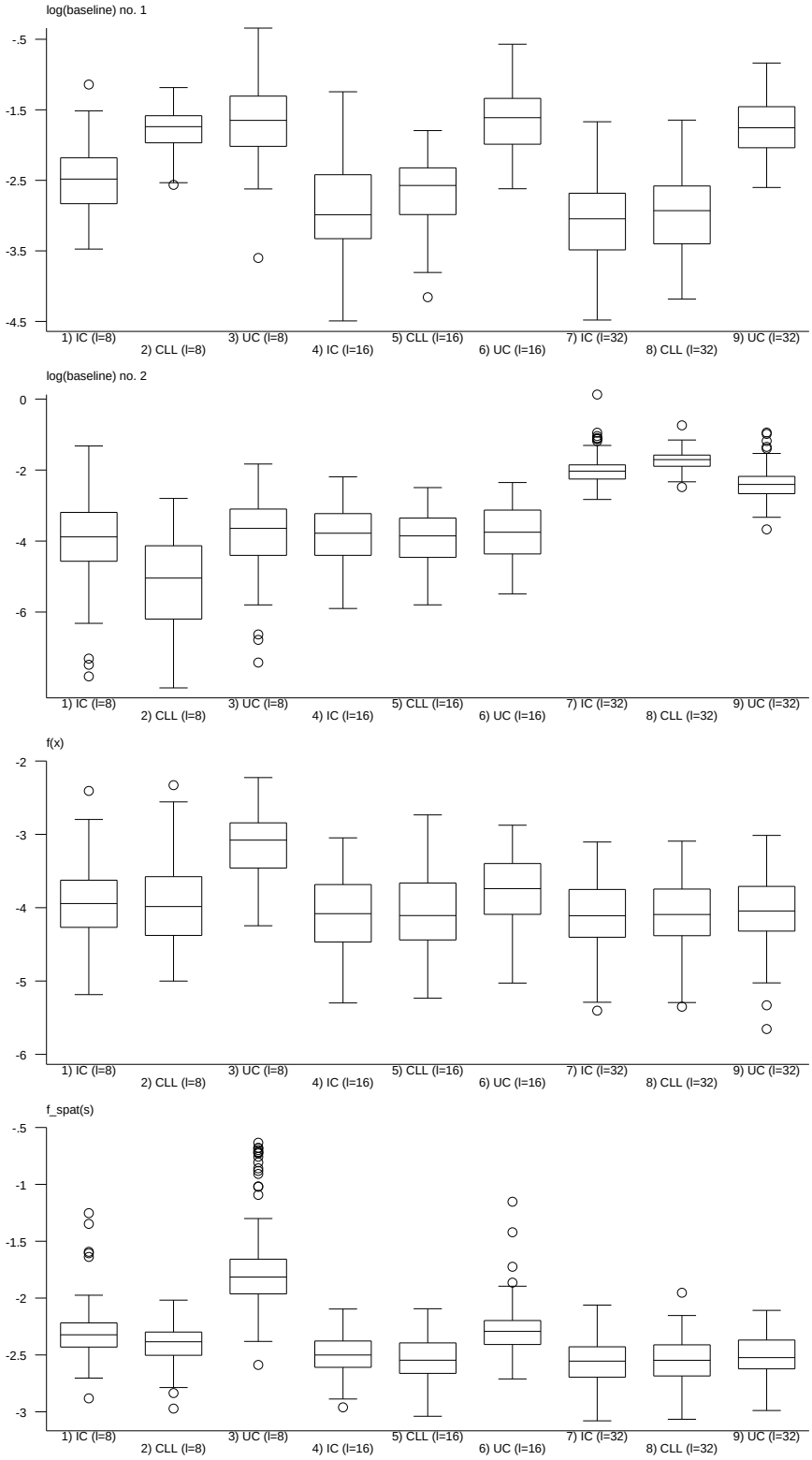


Figure 4: Simulation study: Boxplots of log(MSE) for the two different baselines, the nonparametric effect and the spatial effect. IC denotes results from treating survival times as interval censored, CLL denotes results from the complementary log-log model and UC denotes results from treating the survival times as uncensored. The boxplots are arranged columnwise corresponding to $l = 8$ intervals (left three boxplots), $l = 16$ intervals (middle three boxplots) and $l = 32$ intervals (right three boxplots).

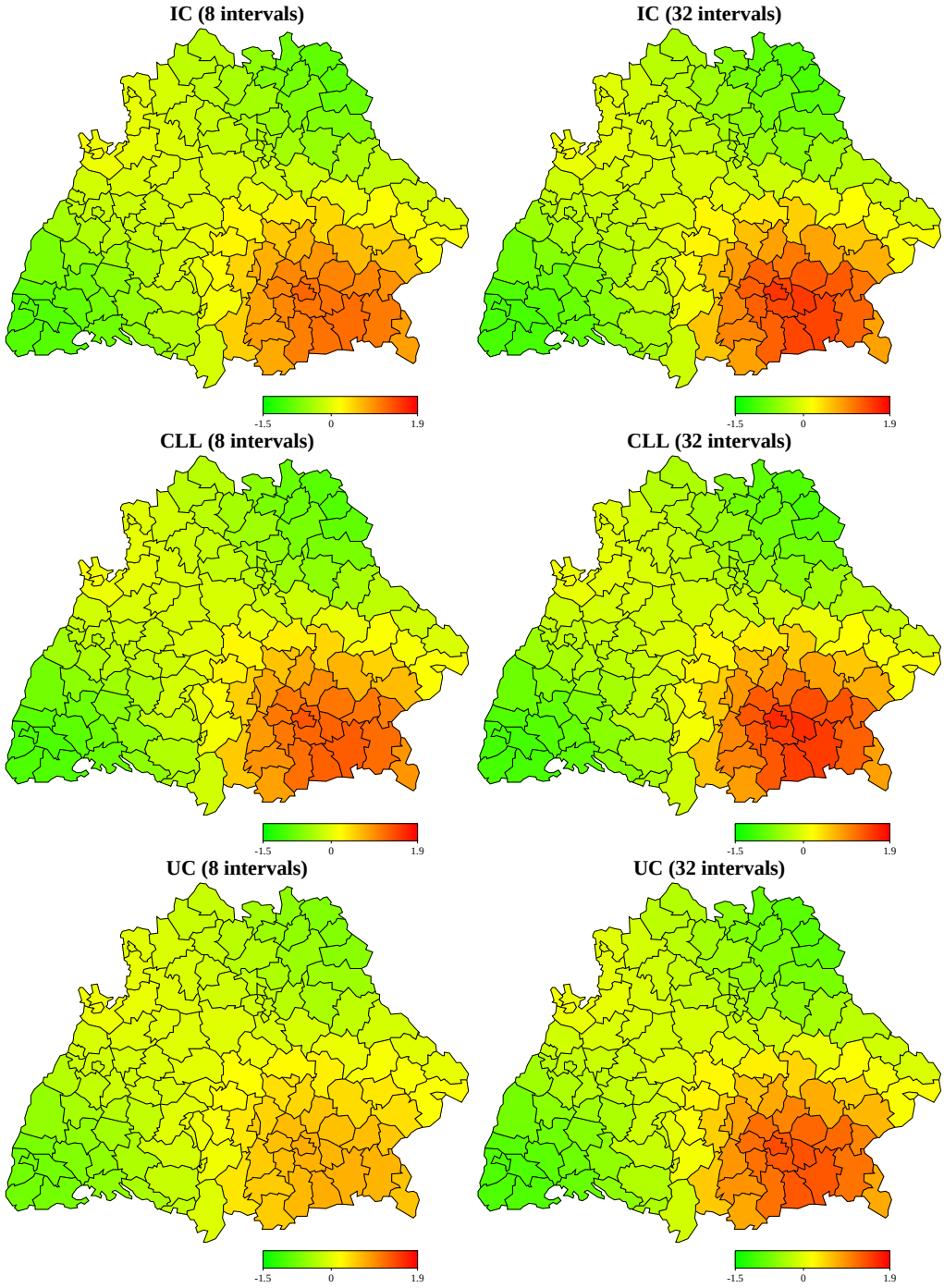


Figure 5: Simulation study: Average estimates for f_{spat}. IC denotes results from treating survival times as interval censored, CLL denotes results from the complementary log-log model and UC denotes results from treating the survival times as uncensored. The left panel shows results obtained for $l = 8$ intervals and the right panel shows results for $l = 32$ intervals.

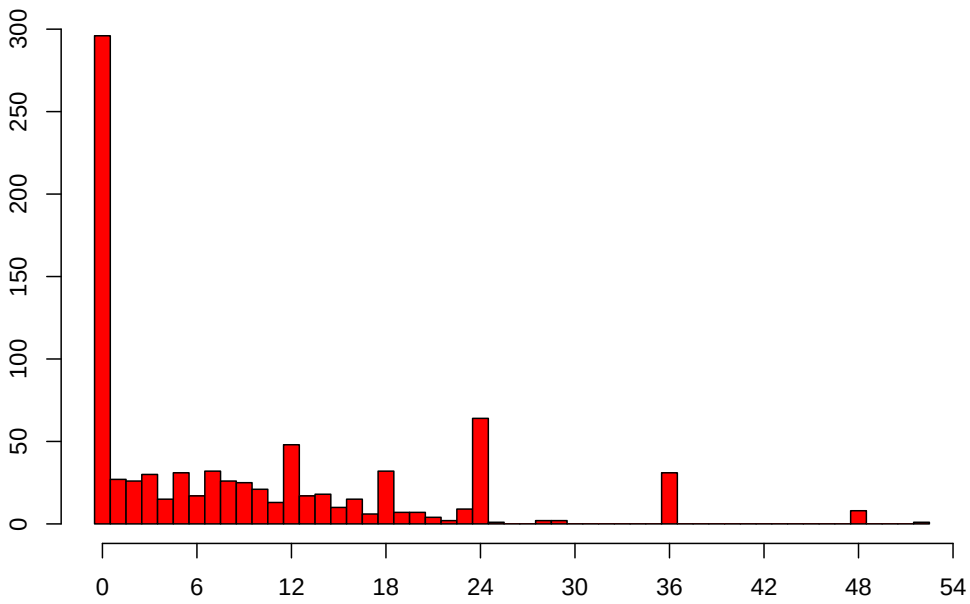


Figure 6: Childhood mortality in Nigeria: Frequencies of observed survival times in months.

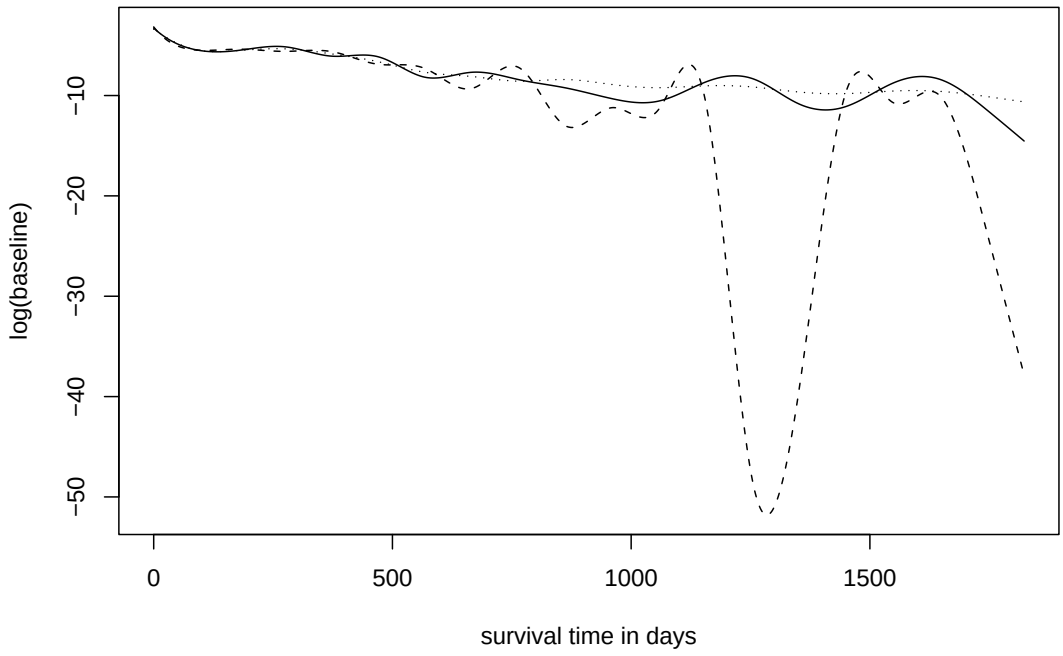


Figure 7: Childhood mortality in Nigeria: Estimated log-baselines based on interval censoring with heaping (straight line), interval censoring without heaping (dashed line) and randomly spread uncensored observations (dotted line).

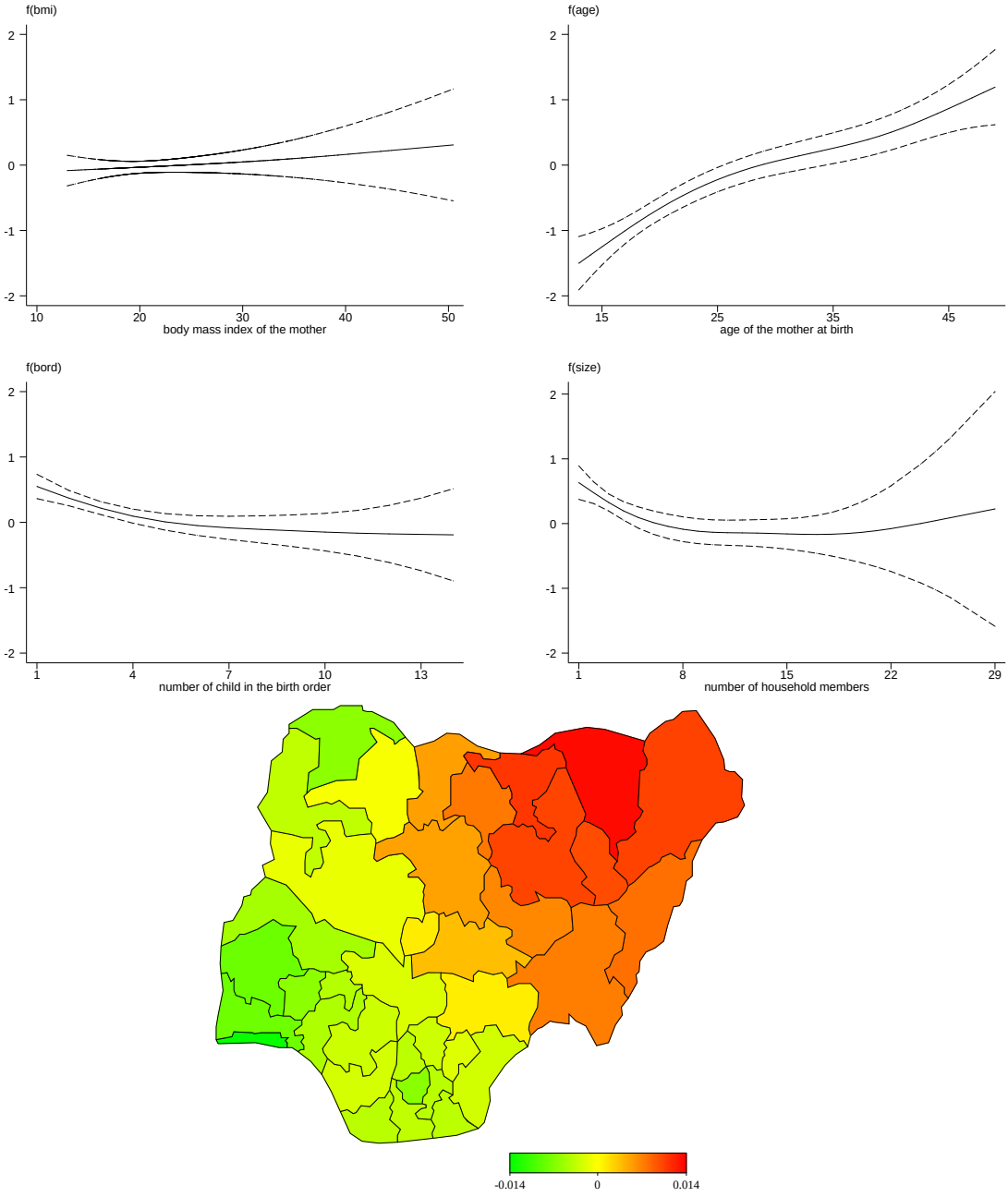


Figure 8: Childhood mortality in Nigeria: Estimates for nonparametric effects (with 95% credible intervals) and for the spatial effect.