Statistica Sinica **14**(2004), 715-745

PENALIZED STRUCTURED ADDITIVE REGRESSION FOR SPACE-TIME DATA: A BAYESIAN PERSPECTIVE

Ludwig Fahrmeir, Thomas Kneib and Stefan Lang

University of Munich

Abstract: We propose extensions of penalized spline generalized additive models for analyzing space-time regression data and study them from a Bayesian perspective. Non-linear effects of continuous covariates and time trends are modelled through Bayesian versions of penalized splines, while correlated spatial effects follow a Markov random field prior. This allows to treat all functions and effects within a unified general framework by assigning appropriate priors with different forms and degrees of smoothness. Inference can be performed either with full (FB) or empirical Bayes (EB) posterior analysis. FB inference using MCMC techniques is a slight extension of previous work. For EB inference, a computationally efficient solution is developed on the basis of a generalized linear mixed model representation. The second approach can be viewed as posterior mode estimation and is closely related to penalized likelihood estimation in a frequentist setting. Variance components, corresponding to inverse smoothing parameters, are then estimated by marginal likelihood. We carefully compare both inferential procedures in simulation studies and illustrate them through data applications. The methodology is available in the open domain statistical package *BayesX* and as an S-plus/R function.

Key words and phrases: Generalized linear mixed models, P-splines, Markov random fields, MCMC, restricted maximum likelihood.

1. Introduction

In longitudinal studies, data usually consist of repeated observations for a population of individuals or units. Response variables may be continuous or discrete as in generalized linear models, and covariates can be metrical or categorical, and possibly time-varying. In various applications, the location or site on a spatial array is given for each unit as additional information, and analyzing its impact on the response simultaneously with the effects of other covariates is of substantive interest.

As a typical example, we analyze data from a forest health survey: each year the damage state of a population of trees is measured as a binary response, and the site of each tree is available on a lattice map. Covariates are age of the tree, canopy density at the stand and calendar time.

If we consider only observations for one time period, then we obtain spatial regression data as a special case. As an application, we consider the 2002 survey

on rents for flats in Munich, where the location of a flat is given by an irregular lattice map of subquarters of Munich together with a large number of covariates characterizing the flat.

In this paper we propose spatio-temporal extensions of generalized additive and varying coefficient models for analyzing such space-time regression data, and we study inference from a Bayesian perspective. This means that usual fixed effects or nonlinear functional effects of covariates, considered as deterministic in a frequentist approach, are interpreted as realizations of random variables or random functions. Based on previous work (Fahrmeir and Lang (2001a), Fahrmeir and Lang (2001b), Lang and Brezger (2004) and Brezger and Lang (2003)), nonlinear effects of continuous covariates as well as smooth time trends are modelled through Bayesian versions of penalized splines (P-splines), introduced in a frequentist setting by Eilers and Marx (1996) and Marx and Eilers (1998). Random walks as special cases of P-splines and more general autoregressive priors for time trends are also included in the model. As for Bayesian versions of smoothing splines (Wahba (1978) and Hastie and Tibshirani (2000)), posterior mode estimates and penalized likelihood estimates coincide for fixed smoothing parameters. Correlated spatial effects are assumed to follow a Gaussian Markov random field prior or are modelled by two dimensional P-splines. Additional uncorrelated random effects may be incorporated as a surrogate for unobserved local small-area, group or individual specific heterogeneity. An advantage of our Bayesian approach is that all unknown functions and parameters can be treated within a unified general framework by assigning appropriate priors with the same general structure but different forms and degrees of smoothness. This broad class of structured additive regression (STAR) models contain several important subclasses as special cases e.g., state-space models for longitudinal data (Fahrmeir and Tutz (2001, Chap. 8)) or geoadditive models, introduced by Kammann and Wand (2003) within a mixed model setting.

Inference for STAR models can be performed either with a full Bayes (FB) or an empirical Bayes (EB) approach. For FB inference, unknown variance or smoothing parameters are considered as random variables with suitable hyperpriors and can be estimated jointly with unknown functions and covariate effects, using computationally efficient extensions of MCMC techniques developed in previous work. For EB inference, variance or smoothing parameters are considered as unknown constants. They are estimated by using (approximate) restricted maximum likelihood (REML). For given or estimated smoothing parameters, unknown functions and covariate effects are obtained as posterior mode estimators by maximizing the posterior density. Our EB approach is based on generalized linear mixed model (GLMM) representations developed in Lin and Zhang (1999) for longitudinal data analysis using smoothing splines, or in Kammann and Wand

(2003) for geoadditive models using stationary Gaussian random fields. Using computationally efficient REML algorithms, we can apply GLMM methodology for EB inference in STAR models even for fairly large data sets. From a more frequentist point of view, EB inference is closely related to penalized likelihood estimation. For the special case of state space models, this close correspondence is also pointed out in Fahrmeir and Knorr-Held (2000). We also suggest a hybrid Bayesian (HB) method, which combines advantages of FB inference with REML estimation of smoothing parameters.

We carefully compare the relative merits of the inferential procedures in simulation studies. A general conclusion is that EB inference performs remarkably well compared to FB inference as long as no problems occur with convergence of REML estimates. Advantages of FB inference are that characteristics and functionals of posteriors can be computed without relying on any large sample normality approximations, and the approach is computationally feasible even for massive data sets with hundreds or even thousands of parameters, because MCMC techniques require only local computations. On the other side, EB estimates are obtained by maximizing an objective function, so that usual questions about convergence of MCMC samples or sensitivity on hyperparameters do not arise. Also, compared to previous implementations, our numerically efficient REML algorithm allows to analyze now even fairly large data sets with the EB approach.

The rest of the paper is organized as follows: models and statistical inference are described in Sections 2 and 3. Performance is investigated through simulation studies in Section 4, and Section 5 contains applications. The concluding Section 6 comments on directions of future research.

The methodology of this paper is available as public domain software. Both the empirical as well as the full Bayesian approach are included in *BayesX*, a software package for Bayesian inference. The program is available at http://www.stat.uni-muenchen.de/~lang/bayesx/bayesx.html. The EB approach is additionally implemented as an S-plus/R function and is available at http://www.stat.uni-muenchen.de/~kneib/software.html.

2. Bayesian Structured Additive Regression

In this section we introduce Bayesian STAR models. They comprise usual generalized additive models, mixed models, varying coefficient models, and extensions to spatial and spatio-temporal models as special cases.

2.1. Observation model

Bayesian generalized linear models (e.g., Fahrmeir and Tutz (2001)) assume that, given covariates u and unknown parameters γ, the distribution of the response variable y belongs to an exponential family, with mean $\mu = E(y|u, \gamma)$

linked to a linear predictor η by

$$\mu = h(\eta) \qquad \eta = u'\gamma. \tag{1}$$

Here h is a known response function, and γ is an unknown regression parameter.

In most practical regression situations, however, we are facing at least one of the following problems.

- For the *continuous covariates* in the data set, the assumption of a strictly linear effect on the predictor may be not appropriate.
- Observations may be *spatially correlated*.
- Observations may be *temporally correlated*.
- Heterogeneity among individuals or units may be not sufficiently described by covariates. Hence, unobserved *unit or cluster specific heterogeneity* must be considered appropriately.

To overcome the difficulties, we replace the strictly linear predictor in (1) by a structured additive predictor

$$\eta_r = f_1(\psi_{r1}) + \cdots + f_p(\psi_{rp}) + u_r'\gamma, \tag{2}$$

where r is a generic observation index, the ψ_j are generic covariates of different types and dimension, and the f_j are (not necessarily smooth) functions of the covariates. The functions f_j comprise usual nonlinear effects of continuous covariates, time trends and seasonal effects, two dimensional surfaces, varying coefficient models, i.i.d. random intercepts and slopes, and temporally or spatially correlated random effects. In order to demonstrate the generality of our approach we point out some special cases of (2) well known from the literature.

- *Generalized additive model (GAM) for cross-sectional data*
  The predictor of a GAM for observation i, $i = 1, \ldots, n$, is given by

  $$\eta_i = f_1(x_{i1}) + \cdots + f_k(x_{ik}) + u_i'\gamma. \tag{3}$$

  Here, the f_j are smooth functions of continuous covariates x_j and, in this paper, they are modelled by (Bayesian) P-splines, see Section 2.2.1. As mentioned in the introduction, these functions are considered as deterministic from a frequentist point of view, while they are interpreted as realizations of random functions within the Bayesian paradigm. We obtain a GAM as a special case of (2) with $r = i$, $i = 1, \ldots, n$, and $\psi_{ij} = x_{ij}$, $j = 1, \ldots, k$.

- *Generalized additive mixed model (GAMM) for longitudinal data*
  Consider longitudinal data for individuals $i = 1, \ldots, n$, observed at time points $t \in \{t_1, t_2, \ldots\}$. For notational simplicity we assume the same time points

for every individual, but generalizations to individual-specific time points are obvious. A GAMM extends (3) by introducing individual specific random effects, i.e.,

$$\eta_{it} = f_1(x_{it1}) + \cdots + f_k(x_{itk}) + b_{1i}w_{it1} + \cdots + b_{qi}w_{itq} + u'_{it}\gamma,$$

where $\eta_{it}, x_{it1}, \ldots, x_{itk}, w_{it1}, \ldots, w_{itq}, u_{it}$ are predictor and covariate values for individual i at time t, and $b_i = (b_{1i}, \ldots, b_{qi})$ is a vector of q i.i.d. random intercepts (if $w_{itj} = 1$) or random slopes. The random effects components are modelled by i.i.d. Gaussian priors, see Section 2.2.3. The functions f_j are nonlinear population "deterministic" effects. Individual specific departures from these population effects and correlations of repeated observations can be modelled through the random effects part of the predictor. As an example, assume that a function represents the population time trend $f(t)$ approximated by a linear combination $f(t) = \sum \beta_j B_j(t)$ of B-spline basis functions $B_j(t)$. Individual specific departures can then be modelled through $f_i(t) = \sum b_{ji}B_j(t)$, where the b_{ji} are i.i.d. random effects, and the design variables w_{itj} are equal to $B_j(t)$. This is in analogy to standard parametric mixed models with, e.g., a linear time trend $\beta_0 + \beta_1 t$ and individual specific random departures $b_{0i} + b_{1i}t$ from this trend. GAMM's can be subsumed into (2) by defining $r = (i,t)$, $\psi_{rj} = x_{itj}$, $j = 1, \ldots, k$, $\psi_{r,k+h} = w_{ith}$, $h = 1, \ldots, q$, and $f_{k+h}(\psi_{r,k+h}) = b_{hi}w_{ith}$. Similarly, GAMM's for cluster data can be written in the general form (2).

- *Space-time main effect model − geoadditive models*
  Suppose we observe longitudinal data with additional geographic information for every observation. A reasonable predictor for such spatio-temporal data (see e.g., Fahrmeir and Lang (2001b)) is given by

$$\eta_{it} = f_1(x_{it1}) + \cdots + f_k(x_{itk}) + f_{time}(t) + f_{spat}(s_{it}) + u'_{it}\gamma, \tag{4}$$

  where f_{time} is a possibly nonlinear time trend and f_{spat} is a spatially correlated (random) effect of the location s_{it} an observation pertains to. Models with a predictor that contains a spatial effect are also called geoadditive models, see Kammann and Wand (2003). The time trend can be modelled by random walk priors, autoregressive process priors, or P-splines (see Section 2.2.1), and the spatial effect by Markov random fields or two dimensional P-splines, see Sections 2.2.2 and 2.2.4. Note that observations are marginally correlated after integrating out the temporally or spatially correlated (random) effects f_{time} and f_{spat}. Individual specific effects can be incorporated as for GAMM's, if appropriate. In the notation of (2) we obtain $r = (i,t)$, $\psi_{rj} = x_{itj}$ for $j = 1, \ldots, k$, $\psi_{r,k+1} = t$ and $\psi_{r,k+2} = s_{it}$.

- *Varying coefficient model (VCM) − Geographically weighted regression*
  A VCM, as proposed by Hastie and Tibshirani (1993), is

$$\eta_i = g_1(x_{i1})z_{i1} + \cdots + g_k(x_{ik})z_{ik},$$

  where the effect modifiers x_{ij} are continuous covariates or time scales and the interacting variables z_{ij} are either continuous or categorical. A VCM can be cast into (2) with $r = i$ and $\psi_{ij} = (x_{ij}, z_{ij})$, by defining the special function $f_j(\psi_{ij}) = f_j(x_{ij}, z_{ij}) = g_j(x_{ij})z_{ij}$. Note that in this paper the effect modifiers are not necessarily restricted to be continuous variables as in Hastie and Tibshirani (1993). For example the geographical location may be used as effect modifiers as well, see Fahrmeir, Lang, Wolff and Bender (2003) for an example. VCM's with spatially varying regression coefficients are well known in the geography literature as *geographically weighted regression*, see e.g., Fotheringham, Brunsdon and Charlton (2002).
- *ANOVA type interaction model*
  Suppose x_{i1} and x_{i2} are two continuous covariates. Then, the effect of x_{i1} and x_{i2} may be modelled by a predictor of the form

$$\eta_i = f_1(x_{i1}) + f_2(x_{i2}) + f_{1|2}(x_{i1}, x_{i2}) + \cdots,$$

  see e.g., Chen (1993). The functions f_1 and f_2 are the main effects of the two covariates and $f_{1|2}$ is a two dimensional interaction surface which might be modelled by two dimensional P-splines, see Section 2.2.4. The interaction can be cast into the form (2) by defining $r = i$, $\psi_{r1} = x_{i1}$, $\psi_{r2} = x_{i2}$ and $\psi_{r3} = (x_{i1}, x_{i2})$. Similarly (4) may be extended to a model incorporating a space-time interaction effect.

At first sight it may look strange to use one general notation for nonlinear functions of continuous covariates, i.i.d. random intercepts and slopes, and spatially correlated random effects as in (2). However, the unified treatment of the different components in our model has several advantages.

- Since we adopt a Bayesian perspective, both "fixed effects" and "random effects" are random variables. They are distinguished by different priors, e.g., diffuse priors for fixed effects and Gaussian priors for i.i.d. random effects, see also the discussion in Hobert and Casella (1996).
- As we will see in Section 2.2, the priors for smooth functions, two dimensional surfaces, i.i.d., serially and spatially correlated random effects can be cast into a general form.
- The general form of the priors allows rather general and unified estimation procedures, see Section 3. As a side effect the implementation and description of these procedures is considerably facilitated.

2.2. Prior assumptions

For Bayesian inference, the unknown functions $f_1, \ldots, f_p$ in (2), more exactly corresponding vectors of function evaluations, and the fixed effects parameter γ are considered as random variables and must be supplemented by appropriate prior assumptions.

Throughout the paper we assume diffuse a prior $p(\gamma) \propto const$ for the fixed effects parameter γ.

Priors for the unknown functions $f_1, \ldots, f_p$ depend on the *type of the covariate* and on the *prior beliefs about smoothness.* In the following we express the vector of function evaluations $f_j = (f_j(\psi_{1j}), \ldots, f_j(\psi_{nj}))'$ of an unknown function f_j as the matrix product of a design matrix Ψ_j and a vector of unknown parameters β_j, i.e.,

$$f_j = \Psi_j \beta_j. \tag{5}$$

Then, we obtain the predictor (2) in matrix notation as

$$\eta = \Psi_1 \beta_1 + \cdots + \Psi_p \beta_p + U\gamma, \tag{6}$$

where U corresponds to the usual design matrix for fixed effects.

A prior for a function f_j is now defined by specifying a suitable design matrix Ψ_j and a prior distribution for the vector β_j of unknown parameters. The general form of the prior for β_j is

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

where K_j is a *penalty matrix* that shrinks parameters towards zero, or penalizes too abrupt jumps between neighboring parameters. In most cases K_j will be rank deficient and therefore the prior for β_j is partially improper.

The variance parameter τ_j^2 is equivalent to the inverse smoothing parameter in a frequentist approach and controls the trade off between flexibility and smoothness. For FB inference, weakly informative inverse Gamma hyperpriors $\tau_j^2 \sim IG(a_j, b_j)$ are assigned to τ_j^2, with $a_j = b_j = 0.001$ as a standard option. For EB inference, τ_j^2 is considered an unknown constant which is determined as a REML estimate.

In the following we describe specific priors for different types of covariates and functions f_j.

2.2.1. Priors for continuous covariates and time scales

Several alternatives have been recently proposed for specifying smoothness priors for continuous covariates or time trends. These are *random walk priors* or more generally *autoregressive priors* (see Fahrmeir and Lang (2001a) and

Fahrmeir and Lang (2001b)), *Bayesian P-splines* (Lang and Brezger (2004)) and *Bayesian smoothing splines* (Hastie and Tibshirani (2000)). In the following we focus on P-splines. The approach assumes that an unknown smooth function f_j of a covariate x_j can be approximated by a polynomial spline of degree l defined on a set of equally spaced knots $x_j^{min} = \zeta_0 < \zeta_1 < \cdots < \zeta_{d-1} < \zeta_d = x_j^{max}$ within the domain of x_j. Such a spline can be written in terms of a linear combination of $M_j = d + l$ B-spline basis functions B_m, i.e.,

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

Here $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_j})'$ corresponds to the vector of unknown regression coefficients. The $n \times M_j$ design matrix Ψ_j consists of the basis functions evaluated at the observations x_{ij}, i.e., $\Psi_j(i,m) = B_m(x_{ij})$. The crucial choice is the number of knots: for a small number of knots, the resulting spline may not be flexible enough to capture the variability of the data; for a large number of knots, estimated curves tend to overfit the data and, as a result, too rough functions are obtained. As a remedy, Eilers and Marx (1996) suggest a moderately large number of equally spaced knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on first or second order differences of adjacent B-Spline coefficients to guarantee sufficient smoothness of the fitted curves. This leads to penalized likelihood estimation with penalty terms

$$P(\lambda_j) = \frac{1}{2}\lambda_j \sum_{m=k+1}^{M_j} (\Delta^k \beta_{jm})^2, \quad k = 1, 2, \tag{8}$$

where λ_j is the smoothing parameter and Δ^k is the difference operator of order k. First order differences penalize abrupt jumps $\beta_{jm} - \beta_{j,m-1}$ between successive parameters and second order differences penalize deviations from the linear trend $2\beta_{j,m-1} - \beta_{j,m-2}$. In a Bayesian approach we use the stochastic analogue of difference penalties, i.e., first or second order random walks, as a prior for the regression coefficients. First and second order random walks are defined by

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

with Gaussian errors $u_{jm} \sim N(0, \tau_j^2)$ and diffuse priors $p(\beta_{j1}) \propto const$, or $p(\beta_{j1})$ and $p(\beta_{j2}) \propto const$, for initial values, respectively. Note that simple first or second order random walks, as proposed in Fahrmeir and Lang (2001a), can be regarded as P-splines of degree $l = 0$ and are therefore a special case. The joint distribution of the regression parameters β_j is easily computed as a product of conditional densities defined by (9) and can be brought into the general form (7).

The penalty matrix is of the form $K_j = D'D$ where D is a first or second order difference matrix. More details about Bayesian P-splines can be found in Lang and Brezger (2004). For time scales, more general autoregressive process priors than the random walk models (9) may be useful, for example to model flexible seasonal patterns, see Fahrmeir and Lang (2001a). Again they can be written in the general form (7).

As an alternative to roughness penalties, approaches based on adaptive knot selection for splines have become very popular, see Friedman (1991) and Stone, Hansen, Kooperberg and Truong (1997) for frequentist versions. Bayesian variants can be found in Denison, Mallick and Smith (1998), Biller (2000), Di Matteo, Genovese and Kass (2001), Biller and Fahrmeir (2001) and Hansen and Kooperberg (2002).

2.2.2. Priors for spatial effects

Suppose that the index $s \in \{1, \ldots, S\}$ represents the location or site in connected geographical regions. For simplicity we assume that the regions are labelled consecutively. A common way to introduce a spatially correlated effect is to assume that neighboring sites are more alike than two arbitrary sites. Thus for a valid prior definition a set of neighbors for each site s must be defined. For geographical data one usually assumes that two sites s and s' are neighbors if they share a common boundary.

The simplest (but most often used) spatial smoothness prior for the function evaluations $f_{spat}(s) = \beta_s$ is

$$\beta_s | \beta_{s'},\ s' \neq s, \tau_j^2 \sim N\Big(\frac{1}{N_s} \sum_{s' \in \partial_s} \beta_{s'}, \frac{\tau_j^2}{N_s}\Big), \tag{10}$$

where N_s is the number of adjacent sites and $s' \in \partial_s$ denotes that site s' is a neighbor of site s. Thus the (conditional) mean of β_s is an unweighted average of function evaluations of neighboring sites. The prior is a direct generalization of a first order random walk to two dimensions and is called a Markov random field (MRF). More general priors based on weighted averages can be found e.g., in Besag York and Mollié (1991). The $n \times S$ design matrix Ψ is now a 0/1 incidence matrix. Its value in the ith row and the sth column is 1 if the ith observation is located in site or region s, and zero otherwise. The $S \times S$ penalty matrix K has the form of an adjacency matrix.

As an alternative to MRF's, we could use two dimensional surface estimators to model spatial effects, see Section 2.2.4 where we propose a two-dimensional version of P-splines. As another alternative, Kammann and Wand (2003) use stationary Gaussian random fields (GRF) which are popular in geostatistics and

can be seen as two-dimensional surface smoothers based on certain basis functions, e.g., radial basis functions, see Ruppert, Wand and Carroll (2003). GRF's may be approximated by MRF's, see Rue and Tjelmeland (2002). From a computational point of view, MRF's and P-splines are preferable to GRF's because their posterior precision matrices are band matrices or can be transformed into a band matrix-like structure. The special structure of the matrices considerably speeds up computations, at least for FB inference, see Section 3.2. In general, it is not clear which of the different approaches leads to the "best" fits. For data observed on a discrete lattice, MRF's seem to be most appropriate. If the exact locations are available, surface estimators may be more natural, particularly because predictions for unobserved locations are available. However, in some situations surface estimators lead to an improved fit compared to MRF's even for discrete lattices and vice versa. A general approach that can handle both situations is given by Müller, Stadtmüller and Tabnak (1997).

2.2.3. Unordered group indicators and unstructured spatial effects

In many situations we observe the problem of heterogeneity among clusters of observations caused by unobserved covariates. Suppose $c \in \{1, \ldots, C\}$ is a cluster variable indicating the cluster a particular observation belongs to. A common approach to overcome the difficulties of unobserved heterogeneity is to introduce additional Gaussian i.i.d. effects $f(c) = \beta_c$ with

$$\beta_c \sim N(0, \tau^2), \quad c = 1, \ldots, C. \tag{11}$$

The design matrix Ψ is again a $n \times C$ 0/1 incidence matrix and the penalty matrix is the identity matrix, i.e., $K = I$. From a classical perspective, (11) defines i.i.d. *random effects*. However, from a Bayesian point of view, all unknown parameters are assumed to be random and hence the notation "random effects" in this context is misleading. We think of (11) more as an approach for modelling an unsmooth function.

Note that we consider *cluster specific* random effects. In GAMM's for longitudinal data, repeated observations for an individual form a cluster with an individual specific random effect. Observation specific random effects are a special case, where each observation is its own cluster. In this case, random effects are not identifiable for Gaussian and binary responses.

The prior (11) may also be used for a more sophisticated modelling of spatial effects. In some situation it may be useful to split up a spatial effect f_{spat} into a spatially correlated (smooth) part f_{str} and a spatially uncorrelated (unsmooth) part f_{unstr}, i.e., $f_{spat} = f_{str} + f_{unstr}$. A rationale is that a spatial effect is usually a surrogate of many unobserved influential factors, some of them may obey a strong spatial structure and others may be present only locally. By estimating

a structured and an unstructured component we aim at distinguishing between the two kinds of influential factors, see Besag York and Mollié (1991). For the smooth spatial part we assume Markov random field priors or two dimensional surface smoothers as described in the next section. For the uncorrelated part we may assume the prior (11).

2.2.4. Modelling interactions

The models considered so far are not appropriate for modelling interactions between covariates. A common approach is based on varying coefficient models introduced by Hastie and Tibshirani (1993) in the context of smoothing splines. Here, the effect of covariate z_{ij} is assumed to vary smoothly over the range of the second covariate x_{ij}, i.e.,

$$f_j(x_{ij}, z_{ij}) = g_j(x_{ij})z_{ij}. \tag{12}$$

In most cases the interacting covariate z_{ij} is categorical whereas the effect modifier may be either metrical, spatial or an unordered group indicator. For the nonlinear function g_j we may assume the priors already defined in Sections 2.2.1 for metrical effect modifiers, 2.2.2 for spatial effect modifiers and 2.2.3 for unordered group indicators as effect modifiers. In Hastie and Tibshirani (1993) only metrical effect modifiers have been considered. Models with spatial effect modifiers are used in Fahrmeir, Lang, Wolff and Bender (2003) and Gamerman, Moreira and Rue (2003) to model space-time interactions. From a classical point of view, models with unordered group indicators as effect modifiers are called models with *random slopes*. In matrix notation we obtain for the vector of function evaluations $f_j = \text{diag}(z_{1j}, \ldots, z_{nj})\Psi_j^*\beta_j$, where Ψ_j^* is the design matrix corresponding to the prior for g_j. Hence the overall design matrix is given by $\Psi_j = \text{diag}(z_{1j}, \ldots, z_{nj})\Psi_j^*$.

Suppose now that both interacting covariates are metrical. In this case, a flexible approach for modelling interactions can be based on (nonparametric) two dimensional surface fitting. Here, we briefly describe an approach based on two dimensional P-splines described in more detail in Lang and Brezger (2004). The assumption is that the unknown surface $f_j(x_{ij}, z_{ij})$ can be approximated by the tensor product of two one dimensional B-splines, i.e.,

$$f_j(x_{ij}, z_{ij}) = \sum_{m_1=1}^{M_j} \sum_{m_2=1}^{M_j} \beta_{j,m_1m_2} B_{j,m_1}(x_{ij}) B_{j,m_2}(z_{ij}).$$

Similar to one-dimensional P-splines, the $n \times M_j^2$ design matrix Ψ_j is composed of products of basis functions. Priors for $\beta_j = (\beta_{j,11}, \ldots, \beta_{j,M_jM_j})'$ are now based on smoothness priors common in spatial statistics, e.g., two-dimensional first-order

random walks, (see Besag and Kooperberg (1995)) which can easily be brought into the general form (7). Details can be found in Lang and Brezger (2004).

2.3. Mixed model representation

In this section, we show how STAR models can be represented by generalized linear mixed models (GLMM) after appropriate reparameterization, see also Lin and Zhang (1999) and Green (1987) in the context of smoothing splines. In fact, model (1) with the structured additive predictor (6) can always be expressed as a GLMM. This provides the key for simultaneous estimation of the functions f_j, $j = 1, \ldots, p$, and the variance (or inverse smoothing) parameters τ_j^2 in an EB approach in Section 3.1 To rewrite the model as a GLMM, the general model formulation is useful again. We proceed as follows.

The vectors of regression coefficients β_j, $j = 1, \ldots, p$, are decomposed into an *unpenalized* and a *penalized part*. Suppose that the jth coefficient vector has dimension $d_j \times 1$ and the corresponding penalty matrix K_j has rank rk_j. Then we define the decomposition

$$\beta_j = \Psi_j^{unp}\beta_j^{unp} + \Psi_j^{pen}\beta_j^{pen}, \tag{13}$$

where the columns of the $d_j \times (d_j - rk_j)$ matrix Ψ_j^{unp} contain a basis of the nullspace of K_j. The $d_j \times rk_j$ matrix Ψ_j^{pen} is given by $\Psi_j^{pen} = L_j(L_j'L_j)^{-1}$ where the full column rank $d_j \times rk_j$ matrix L_j is determined by the decomposition of the penalty matrix K_j into $K_j = L_jL_j'$. A requirement for the decomposition is that $L_j'\Psi_j^{unp} = 0$ and $\Psi_j^{unp}L_j' = 0$ hold. Hence the parameter vector β_j^{unp} represents the part of β_j which is not penalized by K_j whereas the vector β_j^{pen} represents the deviations of the parameters β_j from the nullspace of K_j.

In general, the decomposition $K_j = L_jL_j'$ of K_j can be obtained from the spectral decomposition $K_j = \Gamma_j\Omega_j\Gamma_j'$. The $(rk_j \times rk_j)$ diagonal matrix Ω_j contains the positive eigenvalues ω_{jm}, $m = 1, \ldots, rk_j$, of K_j in descending order, i.e., $\Omega_j = \text{diag}(\omega_{j1}, \ldots, \omega_{j,rk_j})$. Γ_j is a $(d_j \times rk_j)$ orthogonal matrix of the corresponding eigenvectors. From the spectral decomposition we can choose $L_j = \Gamma_j\Omega_j^{1/2}$. In some cases a more favorable decomposition can be found. For instance, for P-splines defined in Section 2.2.1, a more favorable choice for L_j is given by $L_j = D'$ where D is the first or second order difference matrix. Of course, for (the "random effects") prior (11) of Section 2.2.3 a decomposition of $K_j = I$ is not necessary. Also, the unpenalized part vanishes completely.

The matrix Ψ_j^{unp} is the identity vector $\mathbf{1}$ for P-splines with first-order random walk penalty and Markov random fields. For P-splines with second-order random walk penalty, Ψ_j^{unp} is a two column matrix whose first column is again the identity vector and the second column is composed of the (equidistant) knots of the spline.

From (13) we get

$$\frac{1}{\tau_j^2}\beta_j' K_j \beta_j = \frac{1}{\tau_j^2}(\beta_j^{pen})' \beta_j^{pen}.$$

From the general prior (7) for β_j, it follows that $p(\beta_{jm}^{unp}) \propto const$, $m = 1, \ldots, d_j - rk_j$, and

$$\beta_j^{pen} \sim N(0, \tau_j^2 I). \tag{14}$$

Finally, by defining the matrices $\tilde{U}_j = \Psi_j \Psi_j^{unp}$ and $\tilde{\Psi}_j = \Psi_j \Psi_j^{pen}$, we can rewrite the predictor (6) as

$$\eta = \sum_{j=1}^{p} \Psi_j \beta_j + U\gamma = \sum_{j=1}^{p} (\Psi_j \Psi_j^{unp} \beta_j^{unp} + \Psi_j \Psi_j^{pen} \beta_j^{pen}) + U\gamma = \tilde{U}\beta^{unp} + \tilde{\Psi}\beta^{pen}.$$

The design matrix $\tilde{\Psi}$ and the vector β^{pen} are composed of the matrices $\tilde{\Psi}_j$ and the vectors β_j^{pen}, respectively. More specifically, we obtain $\tilde{\Psi} = (\tilde{\Psi}_1\ \tilde{\Psi}_2\ \cdots\ \tilde{\Psi}_p)$ and the stacked vector $\beta^{pen} = ((\beta_1^{pen})', \ldots, (\beta_p^{pen})')'$. Similarly the matrix $\tilde{U}$ and the vector β^{unp} are given by $\tilde{U} = (\tilde{U}_1\, \tilde{U}_2 \cdots \tilde{U}_p\, U)$ and $\beta^{unp} = ((\beta_1^{unp})', \ldots, (\beta_p^{unp})', \gamma')'$.

Finally, we obtain a GLMM with fixed effects β^{unp} and random effects $\beta^{pen} \sim N(0, \Lambda)$ where $\Lambda = \text{diag}(\tau_1^2, \ldots, \tau_1^2, \ldots, \tau_p^2, \ldots, \tau_p^2)$. Hence, we can utilize GLMM methodology for simultaneous estimation of the functions f_j and the variance parameters τ_j^2, see the next section.

The mixed model representation also enables us to examine the identification problem inherent to nonparametric regression from a different angle. Except for i.i.d. Gaussian effects (11), the design matrices $\tilde{U}_j$ for the unpenalized parts contain the identity vector. Provided that there is at least one nonlinear effect and that γ contains an intercept, the matrix $\tilde{U}$ has no full column rank. Hence, all identity vectors in $\tilde{U}$ except for the intercept must be deleted to guarantee identifiability.

3. Inference

Bayesian inference is based on the posterior of the model. The analytic form of the posterior depends on the specific parameterization of the model. If we choose the original parameterization, the posterior for FB inference is given by

$$p(\beta_1, \ldots, \beta_p, \tau_1^2, \ldots, \tau_p^2, \gamma | y) \propto L(y, \beta_1, \ldots, \beta_p, \gamma) \prod_{j=1}^{p} \left(p(\beta_j | \tau_j^2) p(\tau_j^2) \right), \tag{15}$$

where $L(\cdot)$ denotes the likelihood which is the product of individual likelihood contributions. For EB inference, where variances τ_j^2 are considered as constants,

the variances τ_j^2 and the priors $p(\tau_j^2)$ have to be deleted. In terms of the GLMM representation of the model we obtain

$$p(\beta^{unp}, \beta^{pen}|y) \propto L(y, \beta^{unp}, \beta^{pen}) \prod_{j=1}^{p} \left(p(\beta_j^{pen}|\tau_j^2) \right), \tag{16}$$

where $p(\beta^{pen}|\tau_j^2)$ is defined in (14).

3.1. EB inference based on GLMM methodology

Based on the GLMM representation outlined in Section 2.3, regression and variance parameters can be estimated using iteratively weighted least squares (IWLS) and (approximate) restricted maximum likelihood (REML) developed for GLMM's. Estimation is carried out iteratively in two steps.

1. Obtain updated estimates $\hat{\beta}^{unp}$ and $\hat{\beta}^{pen}$ given the current variance parameters as the solutions of the linear equation system

$$\begin{pmatrix} \tilde{U}'W\tilde{U} & \tilde{U}'W\tilde{\Psi} \\ \tilde{\Psi}'W\tilde{U} & \tilde{\Psi}'W\tilde{\Psi} + \Lambda^{-1} \end{pmatrix} \begin{pmatrix} \beta^{unp} \\ \beta^{pen} \end{pmatrix} = \begin{pmatrix} \tilde{U}'W\tilde{y} \\ \tilde{\Psi}'W\tilde{y} \end{pmatrix}. \tag{17}$$

   The $(n \times 1)$ vector $\tilde{y}$ and the $n \times n$ diagonal matrix $W = \text{diag}(w_1, \ldots, w_n)$ are the usual working observations and weights in generalized linear models, see Fahrmeir and Tutz (2001, Chap. 2.2.1).
2. Updated estimates for the variance parameters $\hat{\tau}_j^2$ are obtained by maximizing the (approximate) restricted log likelihood

$$\begin{aligned} l^*(\tau_1^2, \ldots, \tau_p^2) = &-\tfrac{1}{2}\log(|\Sigma|) - \tfrac{1}{2}\log(|\tilde{U}\Sigma^{-1}\tilde{U}|) \\ &-\tfrac{1}{2}(\tilde{y} - \tilde{U}\hat{\beta}^{unp})'\Sigma^{-1}(\tilde{y} - \tilde{U}\hat{\beta}^{unp}) \end{aligned} \tag{18}$$

   with respect to the variance parameters $\tau^2 = (\tau_1^2, \ldots, \tau_p^2)'$. Here, $\Sigma = W^{-1} + \tilde{\Psi}\Lambda\tilde{\Psi}'$ is an approximation to the marginal covariance matrix of $\tilde{y}|\beta^{pen}$.

The two estimation steps are iterated until convergence. We maximize (18) through a computationally efficient alternative to the usual Fisher scoring iterations as described e.g., in Harville (1977), see the second remark below.

Remarks

1. *Credible intervals.* Formula (17) forms the basis for constructing credible intervals of the function estimates $\hat{f}_j$ (Lin and Zhang (1999)). If we denote the coefficient matrix on the left hand side of (17) by H, the approximate covariance matrix of the regression coefficients $\hat{\beta}^{unp}$ and $\hat{\beta}^{pen}$ is given by H^{-1}. Since $\hat{f}_j = \tilde{U}_j\hat{\beta}_j^{unp} + \tilde{\Psi}_j\hat{\beta}_j^{pen}$, we obtain

$$\text{Cov}(\hat{f}_j) = (\tilde{U}_j \ \tilde{\Psi}_j)\text{Cov}\left((\hat{\beta}_j^{unp})' \ (\hat{\beta}_j^{pen})' \right) (\tilde{U}_j \ \tilde{\Psi}_j)' \tag{19}$$

for the covariance matrix of $\hat{f}_j$, where $\text{Cov}\left((\hat{\beta}_j^{unp})' \ (\hat{\beta}_j^{pen})'\right)$ can be obtained from the corresponding blocks in H^{-1}.

2. *Numerically efficient implementation of REML estimates.* The restricted log likelihood (18) is usually maximized by Fisher scoring, i.e.,

$$\hat{\tau}^2 = \tilde{\tau}^2 + F^*(\tilde{\tau}^2)^{-1} s^*(\tilde{\tau}^2), \tag{20}$$

where $\tilde{\tau}^2$ are the variance parameters from the last iteration. The score vector $s^*(\tau^2)$ consists of the elements (compare Harville (1977) or Mc Culloch and Searle (2001))

$$s_j^*(\tau^2) = -\frac{1}{2}\text{tr}\left(P\tilde{\Psi}_j\tilde{\Psi}_j'\right) + \frac{1}{2}(\tilde{y} - \tilde{U}\hat{\beta}^{unp})'\Sigma^{-1}\tilde{\Psi}_j\tilde{\Psi}_j'\Sigma^{-1}(\tilde{y} - \tilde{U}\hat{\beta}^{unp}) \tag{21}$$

$j = 1, \ldots, p$ with

$$P = \Sigma^{-1} - \Sigma^{-1}\tilde{U}(\tilde{U}'\Sigma^{-1}\tilde{U})^{-1}\tilde{U}'\Sigma^{-1}. \tag{22}$$

The elements of the expected Fisher information $F^*(\tau^2)$ are given by

$$F_{jk}^*(\tau^2) = \frac{1}{2}\text{tr}\left(P\tilde{\Psi}_j\tilde{\Psi}_j'P\tilde{\Psi}_k\tilde{\Psi}_k'\right), \tag{23}$$

$j, k = 1, \ldots, p$. The crucial point is that direct use of (21) and (23) is not feasible for more than about $n = 3,000$ observations, since they involve the computation and manipulation of several $n \times n$ matrices including P and Σ. In particular, the determination of Σ^{-1}, which requires $O(n^3)$ computations, makes the direct usage of (21) and (23) impractical.

Inversion may be avoided by changing from the marginal to the conditional view of the GLMM yielding the expressions in Lin and Zhang (1999, p.391)

$$s_j^*(\tau^2) = -\frac{1}{2}\text{tr}\left(P\tilde{\Psi}_j\tilde{\Psi}_j'\right) + \frac{1}{2}||\tilde{\Psi}_j'W(\tilde{y} - \tilde{U}\hat{\beta}^{unp} - \tilde{\Psi}\hat{\beta}^{pen})||^2, \tag{24}$$

$$P = W - W(\tilde{U} \ \tilde{\Psi})H^{-1}(\tilde{U} \ \tilde{\Psi})'W. \tag{25}$$

Using (24) to compute the score vector and (25) in combination with (23) to compute $F^*(\tau^2)$ avoids the inversion of Σ, but there are still $n \times n$ matrices that have to be computed and multiplied in each iteration. To get around this, we first replace P by (25) and use an elementary property of the trace to obtain for the first part of (24):

$$-\frac{1}{2}\text{tr}\left(P\tilde{\Psi}_j\tilde{\Psi}_j'\right) = -\frac{1}{2}\text{tr}\left(\tilde{\Psi}_j'W\tilde{\Psi}_j\right) + \frac{1}{2}\text{tr}\left(\tilde{\Psi}_j'W(\tilde{U} \ \tilde{\Psi})H^{-1}(\tilde{U} \ \tilde{\Psi})'W\tilde{\Psi}_j\right). \tag{26}$$

Note that most matrices used in (26) do not have to be evaluated explicitly, since they are submatrices of the weighted sums of squares and crossproducts (SSCP) matrix

$$\begin{pmatrix} \tilde{U}'W\tilde{U} & \tilde{U}'W\tilde{\Psi} \\ \tilde{\Psi}'W\tilde{U} & \tilde{\Psi}'W\tilde{\Psi} \end{pmatrix}, \tag{27}$$

which may be derived at low computational cost from H by substracting Λ^{-1} from the lower right block. E.g., the matrix $\tilde{\Psi}_j'W\tilde{\Psi}_j$ in (26) is the jth diagonal block in $\tilde{\Psi}'W\tilde{\Psi}$.

Formula (23) may also be reexpressed using the definition of P in (25). Some matrix algebra yields the formula

$$\begin{aligned} F_{jk}^*(\tau^2) = \frac{1}{2}\mathrm{tr}\left(\tilde{\Psi}_k'W\tilde{\Psi}_j\tilde{\Psi}_j'W\tilde{\Psi}_k\right) - \mathrm{tr}\left(\tilde{\Psi}_k'W(\tilde{U}\ \tilde{\Psi})H^{-1}(\tilde{U}\ \tilde{\Psi})'W\tilde{\Psi}_j\tilde{\Psi}_j'W\tilde{\Psi}_k\right) \\ + \frac{1}{2}\mathrm{tr}\left(\tilde{\Psi}_k'W(\tilde{U}\ \tilde{\Psi})H^{-1}(\tilde{U}\ \tilde{\Psi})'W\tilde{\Psi}_j\tilde{\Psi}_j'W(\tilde{U}\ \tilde{\Psi})H^{-1}(\tilde{U}\ \tilde{\Psi})'W\tilde{\Psi}_k\right). \end{aligned}$$

Again most of the matrices involved are submatrices of (27) and may therefore be readily obtained since the SSCP-matrix is available from H.

Now the largest matrix involved in the computation of $s^*(\tau^2)$ and $F_{jk}^*(\tau^2)$ is H^{-1}, which reduces the main computational burden from handling $n \times n$ matrices to the inversion of a matrix whose dimension is given by the number of regression coefficients in the model. Compared to the usual version based on (21) and (23), the current implementation of EB inference significantly speeds up computing time and reduces memory allocation.

3.2. FB inference based on Markov chain Monte Carlo

In the full Bayesian approach, parameter estimates are obtained by drawing random samples from the posterior (15) via MCMC simulation techniques. Variance parameters τ_j^2 can be estimated simultaneously with the regression coefficients β_j by assigning additional hyperpriors to them. The most common assumption is, that the τ_j^2 are independently inverse gamma distributed, i.e., $\tau_j^2 \sim IG(a_j, b_j)$, with hyperparameters a_j and b_j specified a priori. We use $a_j = b_j = 0.001$ as a standard option. In some data situations (e.g., for small sample sizes), the estimated nonlinear functions f_j may depend considerably on the particular choice of hyperparameters. It is therefore good practice to estimate all models under consideration using a (small) number of *different* choices for a_j and b_j to assess the dependence of results on minor changes in the model assumptions.

For updating the parameters in an MCMC sampler, we use an MH-algorithm based on iteratively weighted least squares (IWLS) proposals introduced by

Gamerman (1997) and adapted to the present situation in Brezger and Lang (2003).

Parameters are updated in the order $\beta_1, \ldots, \beta_p, \tau_1^2, \ldots, \tau_p^2, \gamma$. Suppose we want to update the regression coefficients β_j of the jth function f_j with current value β_j^c of the chain. Then, according to IWLS, a new value β_j^p is proposed by drawing a random number from the multivariate Gaussian proposal distribution $q(\beta_j^c, \beta_j^p)$ with precision matrix and mean

$$P_j = \Psi_j' W(\beta_j^c)\Psi_j + \frac{1}{\tau_j^2}K_j, \quad m_j = P_j^{-1}\Psi_j' W(\beta_j^c)(\tilde{y} - \eta_{-j}). \tag{28}$$

Here, W and $\tilde{y}$ are again usual working weights and observations in generalized linear models. The vector $\eta_{-j} = \eta - \Psi_j\beta_j$ is the part of the predictor associated with all remaining effects in the model. The proposed vector β_j^p is accepted as the new state of the chain with probability

$$\alpha(\beta_j^c, \beta_j^p) = \min\Big(1, \frac{p(\beta_j^p|\cdot)q(\beta_j^p, \beta_j^c)}{p(\beta_j^c|\cdot)q(\beta_j^c, \beta_j^p)}\Big),$$

where $p(\beta_j|\cdot)$ is the full conditional for β_j (i.e., the conditional distribution of β_j given all other parameters and the data y).

A fast implementation requires efficient sampling from the Gaussian proposal distributions. Algorithms have to take advantage of the special band or sparse matrix structure of the precision matrices P_j in (28). Rue (2001) uses matrix operations for band matrices to draw random numbers from the high dimensional full conditionals, but the different band sizes in every row are not utilized. In our implementation the different band sizes are exploited by using the *envelope method* for Cholesky decompositions of sparse matrices as described in George and Liu (1981). This implies that the number of calculations required to draw random numbers from the proposal distribution is linear in the number of parameters and observations. Also the computation of the acceptance probabilities is linear in the number of observations. For this reason, the FB approach is able to handle complex models with a larger number of observations and parameters than the alternative based EB methodology discussed in the previous section. Currently, the limit is roughly between 200,000 and 300,000 observations (depending on the complexity of the model).

The full conditionals for the variance parameters τ_j^2 are inverse gamma with parameters $a_j' = a_j + 0.5rank(K_j)$ and $b_j' = b_j + 0.5\beta_j' K_j \beta_j$, and updating can be done by simple Gibbs steps, drawing random numbers directly from the inverse gamma densities.

Convergence of the Markov chains to their stationary distributions is assessed by inspecting sampling paths and autocorrelation functions of sampled parameters. In the majority of cases, however, the IWLS updating scheme has excellent mixing properties and convergence problems do not occur.

3.3. Hybrid Bayesian inference

As a third alternative, we consider a hybrid Bayesian (HB) approach. It is motivated by the fact that our simulation study in Section 4 indicates that REML estimators of variance components are less biased compared to the FB estimators. For HB inference, variance parameters τ^2 are first estimated by REML. Then, instead of drawing from inverse gamma full conditionals as in Section 3.2, FB inference is performed by plugging in the REML estimates for τ^2. This strategy aims at combining advantages of EB and FB inference: stable estimation of variance components, and - on the other side- full posterior analysis for regression functions and parameters of primary interest. This allows, for instance, computation of posteriors of any nonlinear functionals, simultaneous credible intervals (see Knorr-Held (2004)), and of probability statements.

4. Simulation Study

The present simulation study aims at imitating typical spatio-temporal longitudinal data. We investigated performance of FB, EB and HB inference through a number of applications to artificial data. To assess the impact of information contained in different types of responses, the following study is based on binary, binomial (with three repeated binary observations), Poisson and Gaussian regression models. In each case, data were generated from logit, loglinear and additive models using the linear predictor

$$\eta_{it} = f_1(x_{it1}) + f_2(s_{it}) + f_3(i) + f_4(i)x_{it2} + f_5(i)x_{it3} + \gamma_1 x_{it2} + \gamma_2 x_{it3}$$

for $i = 1, \ldots, 24$ individuals and $t = 1, \ldots, 31$ repeated measurements, resulting in 744 observations per simulation run. The function f_1 is a sine function, and the spatial function f_2 is shown in the map of Figure 1a, displaying $s = 1, \ldots, 124$ districts of Bayern and Baden-Württemberg, the two southern states in Germany. The functions $f_3 - f_5$ are i.i.d. individual specific Gaussian (random) effects. From a classical perspective $f_3(i)$ is a random intercept, $f_4(i)$ and $f_5(i)$ represent random slopes. The effects γ_1, γ_2 are usual fixed effects.

For the covariate x_1, values were randomly drawn from 186 equidistant gridpoints between -3 and $+3$. Each gridpoint was randomly assigned four times. Similarly, values for the covariates x_2 and x_3 were drawn from 186 equidistant

gridpoints between -1 and $+1$. The function f_2 has 124 different values; each value was randomly assigned 6 times. The i.i.d. Gaussian (random) effects were obtained as drawings

$$f_3(i) \sim N(0; 0.25), f_4(i) \sim N(0; 0.25), f_5(i) \sim N(0; 0.36), \quad i = 1, \ldots, 24.$$

Keeping the resulting 744 predictor values η_{it}, $i = 1, \ldots, 24$, $t = 1, \ldots, 31$, fixed, binary, binomial, Poisson and Gaussian responses were generated using logit, loglinear Poisson and additive Gaussian models, respectively. For each model, the simulation was repeated over 250 such simulation runs, producing responses $y_{it}^{(l)}$, $l = 1, \ldots, 250$, for the predictor. For additive Gaussian models, the errors are i.i.d. drawings from $N(0; 0.25)$.

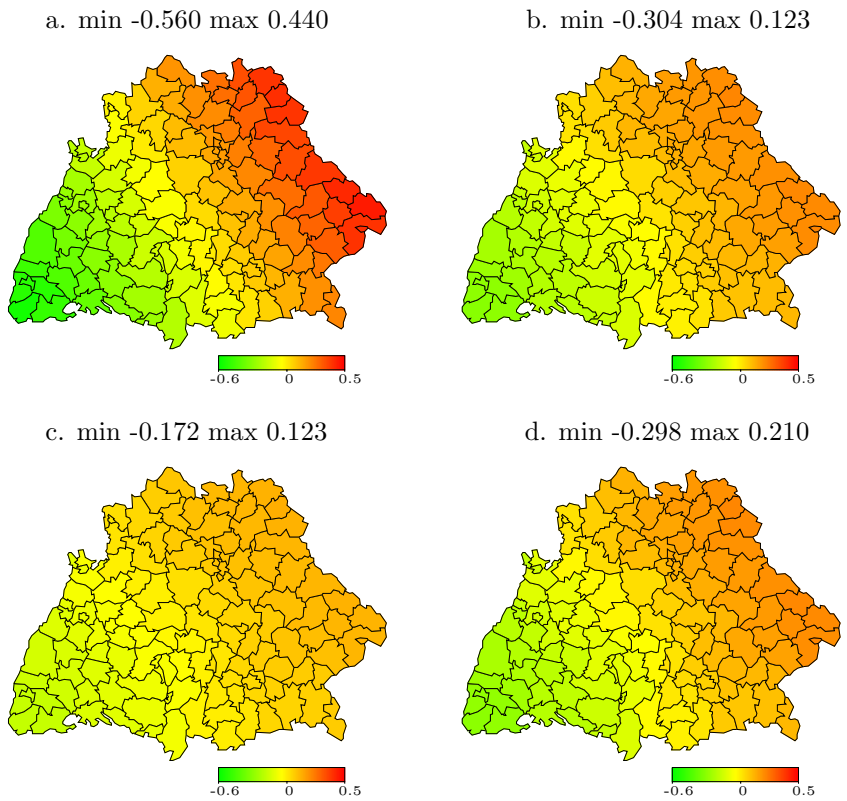

Figure 1. Binary responses: Comparison of average estimates for the spatial effect f_2. Panel (a) shows the true function, panel (b) EB estimates, panel (c) FB estimates with hyperparameters $a = 1$ and $b = 0.005$ and panel (d) FB estimates based on hyperparameters $a = b = 0.001$. Min and max in the titles indicate the range of the true function and the estimated effects.

Using these artificial data, we compared performance in terms of bias, MSE and average coverage properties. For f_1 we assumed a cubic P-spline prior with second order random walk penalty, and for the spatial effect f_2 the MRF prior (10).

A general, but not surprising conclusion is that the bias and MSE tend to decrease with increasing information contained in the responses, i.e., when moving from binary responses to Poisson or Gaussian responses. A further observation is that the REML estimate has convergence problems in about 25% of the analyzed models. In the case of no convergence, usually only one of the variance components switched between two values which were close to each other, while iterations converged for the remaining variance components. A closer inspection of estimates with and without convergence showed that differences in terms of MSE are negligible and the choice of one of the two switching values leads to reasonable estimates. Therefore, it is justified to use the final values after the maximum number of iterations (400) to compute empirical MSE's, bias, average coverage probabilities, etc.

The true sine curve f_1 and the average obtained from all 250 posterior estimates , $l = 1, \ldots, 250$, are hard to distinguish visually for all four observation models, because the bias is very close to zero. Therefore, we only present MSE's in Figure 3.

The true spatial function f_2 and averages of posterior estimates , $l = 1, \ldots,$ 250, are displayed in Figures 1 and 2 for binary, binomial and Poisson observation models. Because EB and HB inference give rather similar results, we do not show HB estimates. We conclude the following: at least for binary observations, the often recommended standard choice $a = 1, b = 0.005$ for hyperparameters of inverse Gamma priors for smoothing parameters results in oversmoothing (Figure 1c), whereas FB inference with $a = b = 0.001$ and EB inference perform considerably better and with comparable bias (Figure 1b and 1d).

For Poisson responses (Figure 2b and 2d), the bias becomes smaller and the true surface is recovered satisfactorily both with full or empirical Bayes estimation. Estimation properties for binomial observations (Figure 2a and 2c) are between results for binary and Poisson models. For Gaussian observations (not shown) we obtain the best results, and EB and FB results are very similar.

For binary and Poisson responses, Figure 3 shows empirical log-MSE's for the sine curve f_1 and the spatial effect f_2, averaged over all covariate values, and for the random effects averaged over $i = 1, \ldots, 24$. From Figure 3 we see that EB (and HB) estimation behaves remarkably well in terms of MSE's when compared to FB inference. For binary responses, the hyperparameter choice $a = 1, b = 0.005$ implies highest MSE for the spatial effect, and also for the random intercept. For Poisson responses FB, EB and HB behave quite similar in terms of MSE's.

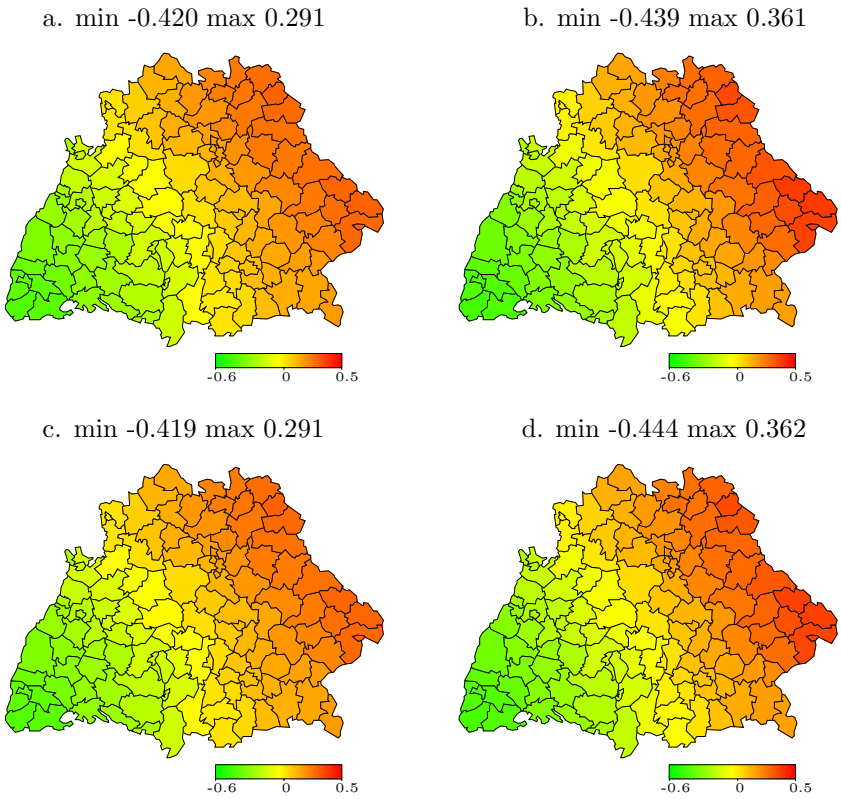

Figure 2. Binomial and Poisson responses: Comparison of average estimates for the spatial effect f_2. Panel (a) shows EB estimates (binomial), panel (b) EB estimates (Poisson), panel (c) FB estimates (binomial, $a = b = 0.001$) and panel (d) FB estimates (Poisson, $a = b = 0.001$). Min and max in the titles indicate the range of the true function and the estimated effects.

Average coverage properties of pointwise credible intervals for a nominal level of 95% are shown in Table 1 for the different effects. For EB inference, credible intervals are computed as described in Section 3.1. In the FB and HB approach pointwise credible intervals are simply obtained by computing the respective empirical quantiles of sampled function values. Table 1 provides some evidence for the following: for EB and HB inference, average coverage probabilities are almost identical in all cases. All four Bayesian approaches have comparable coverage properties for Gaussian and Poisson responses. For binary responses, some difference can be seen. While the average coverage probabilities are still quite acceptable for the nonparametric function f_1, they are partly considerably below the nominal level of 95% for the spatial effect f_2 and the i.i.d. effects f_3, f_4 and f_5. Only FB inference with $a = b = 0.001$ gives satisfactory results.

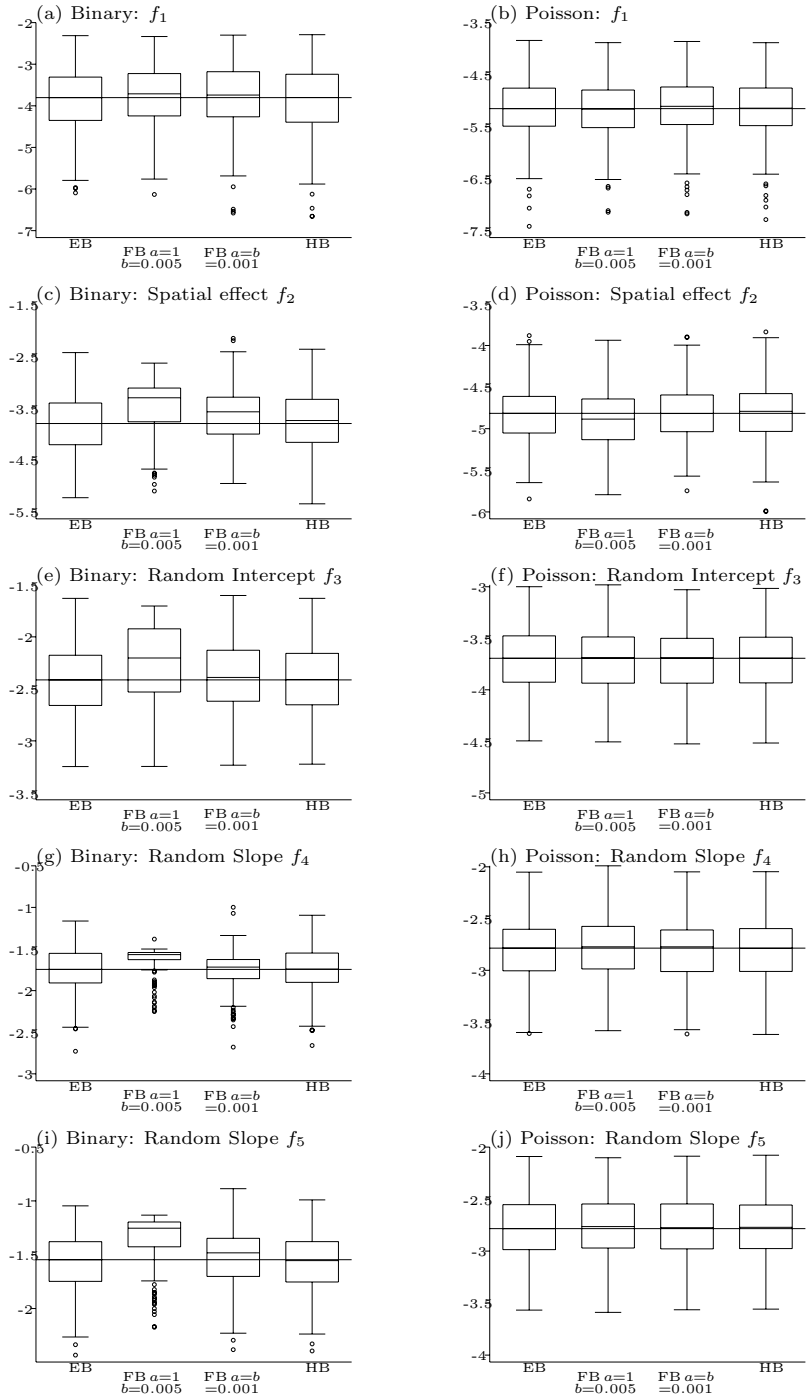

Figure 3. Binary (left panel) and Poisson (right panel) responses: boxplots for log(MSE).

Table 1. Average coverage probabilities for the different effects based on a nominal level of 95%.

	distribution	f_1	f_2	f_3	f_4	f_5
EB	Gaussian	0.993	0.993	0.993	0.976	0.986
	Bernoulli	0.967	0.900	0.915	0.723	0.854
	binomial	0.975	0.990	0.963	0.915	0.947
	Poisson	0.980	0.998	0.972	0.949	0.970
FB	Gaussian	0.971	0.996	0.993	0.975	0.985
$(a = 1, b = 0.005)$	Bernoulli	0.958	0.884	0.856	0.568	0.670
	binomial	0.970	0.984	0.962	0.861	0.932
	Poisson	0.974	0.998	0.973	0.946	0.969
FB	Gaussian	0.973	0.996	0.995	0.978	0.989
$(a = b = 0.001)$	Bernoulli	0.971	0.985	0.935	0.883	0.910
	binomial	0.971	0.995	0.969	0.927	0.959
	Poisson	0.973	0.998	0.976	0.956	0.973
HB	Gaussian	0.970	0.995	0.994	0.975	0.987
	Bernoulli	0.961	0.896	0.915	0.721	0.857
	binomial	0.968	0.986	0.965	0.916	0.949
	Poisson	0.971	0.997	0.972	0.949	0.970

The final comparison concerns estimation of variance components of the random effects f_3, f_4 and f_5. For each type of response, Table 2 compares averages of estimates with the "empirical" variances, obtained from the 24 i.i.d. drawings from the corresponding normals. A comparison with these empirical variances is fairer than with "true" values (given in brackets). For Gaussian responses, FB estimates with $a = b = 0.001$ have larger bias than EB and FB estimates with $a = 1$, $b = 0.005$. For binary responses, on the other side, FB estimates with $a = 1$, $b = 0.005$ have considerable bias. For binomial and Poisson responses, differences between the two FB versions are less distinct, but EB estimates are mostly better. A conclusion emerging from these results is that REML estimates of variance components are preferable in terms of bias.

Table 2. Average bias of the variance components.

emp. value (true value)	bias			
	Gaussian	Bernoulli	binomial	Poisson
0.196 (0.25)	0.010	-0.014	0.003	-0.005
EB 0.226 (0.25)	0.006	-0.047	-0.014	-0.006
0.329 (0.36)	0.017	-0.029	-0.003	0.007
0.196 (0.25)	0.009	-0.066	0.002	-0.001
FB 0.226 (0.25)	0.001	-0.177	-0.070	-0.019
$(a = 1, b = 0.005)$ 0.329 (0.36)	0.013	-0.215	-0.032	-0.004
0.196 (0.25)	0.030	0.024	0.039	0.026
FB 0.226 (0.25)	0.028	-0.019	0.014	0.020
$(a = b = 0.001)$ 0.329 (0.36)	0.051	0.024	0.057	0.048

5. Applications

5.1. Rents for flats: a spatial study

This application illustrates the approaches with a challenging complex geoadditive model. According to the German rental law, owners of apartments or flats can base an increase in the amount that they charge for rent on "average rents" for flats comparable in type, size, equipment, quality and location in a community. To provide information about these "average rents", most larger cities publish "rental guides", which can be based on regression analysis with rent as the dependent variable. We use data from the City of Munich, collected in 2002 by Infratest Sozialforschung for a random sample of approximately 3,000 flats. As response variable we choose

R monthly net rent per square meter in German Marks, that is the monthly rent minus calculated or estimated utility costs.

Covariates characterizing the flat were constructed from almost 200 variables out of a questionnaire answered by tenants of flats. In our reanalysis we use the highly significant metrical covariates "floor space" (F) and "year of construction" (Y) and a vector u of 25 binary covariates characterizing the quality of the flat, e.g., the kitchen and bath equipment, the quality of the heating or the quality of the warm water system. Another important covariate is the location L of the flat in Munich. For the official Munich 2003 rental guide, location in the city was assessed in three categories (average, good, top) by experts. In our reanalysis we focus on a more data-driven assessment of the quality of location by including a spatial effect of the location L into the predictor. So we choose a geoadditive Gaussian model $R = \eta + \varepsilon$ with predictor

$$\eta = \gamma_0 + f_1(F) + f_2(Y) + f_3(L) + u'\gamma. \tag{29}$$

The effects f_1 and f_2 of floor space and year of construction are modelled by cubic P-splines with 20 knots and a second order random walk penalty. For the spatial effect $f_3(L)$ we choose the Markov random field prior (10).

A first analysis was based on the classical assumption of homoscedastic errors $\varepsilon_i \sim N(0, \sigma^2)$. A careful inspection of residuals e_i provides evidence, however, of heteroscedastic errors. We therefore fitted a geoadditive model with log-squared residuals $\log(e_i^2)$ as responses and the same predictor η in (29), and used the predicted responses $\hat{e}_i^2 = \exp(\eta_i)$ as weights for a weighted geoadditive regression with predictor (29). The Figures 4 and 5 show estimated functions f_1, f_2 and the spatial effect f_3 for the (weighted) geoadditive model (29) as well as for the $\log(e_i^2)$ regression, comparing EB and FB results in each case. The discussion and presentation of fixed effects γ is omitted.

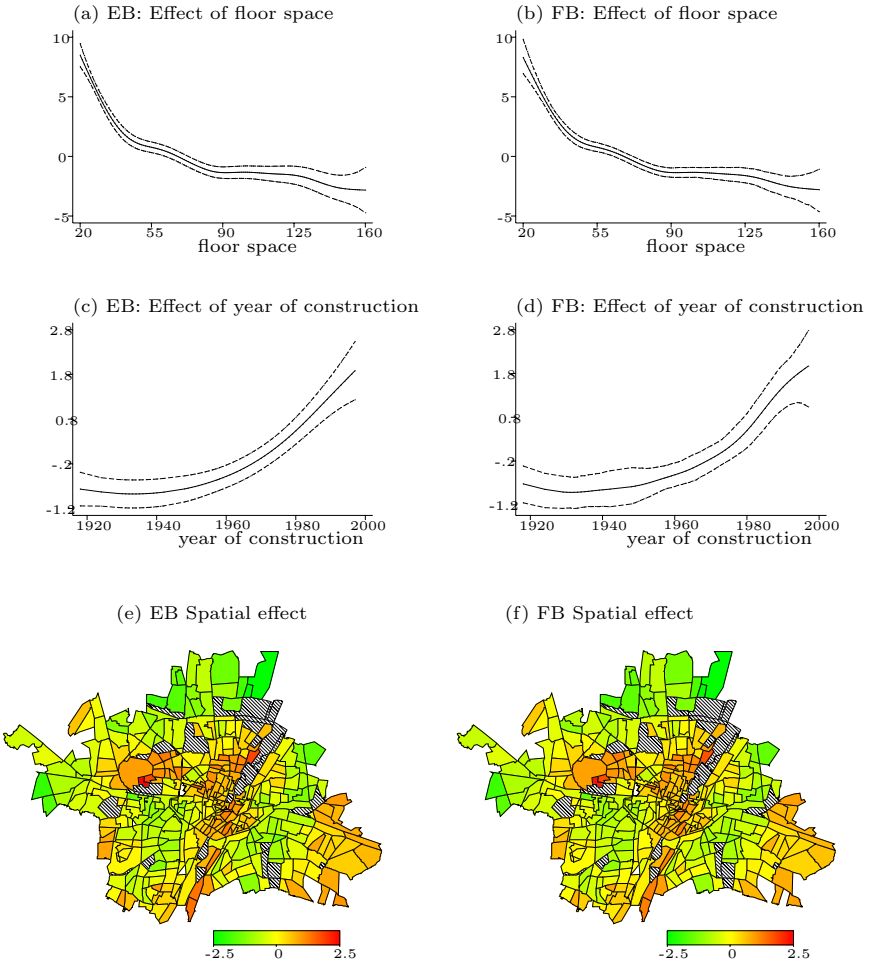

Figure 4. Rent data: Effects of floor space (top), year of construction (middle) and location (bottom) for EB (left panel) and FB (right panel), respectively. Shown are the posterior mode (EB) and mean (FB) estimates. For floor space and year of construction, pointwise 95% credible intervals are included additionally.

The effects of year of construction and floor space in the regression model (29) for rents show the typical nonlinear, monotonically increasing and decreasing curves, respectively. Posterior mode (EB) and posterior mean (FB) estimates are quite similar, in particular for the effect of floor space. The spatial effect of the location in Munich reflects quite well what we know from expert assessments, with an increase of average rents in popular subquarters along the river Isar and near to parks. Again, the differences between posterior modes (EB) and means (FB) are comparably small.

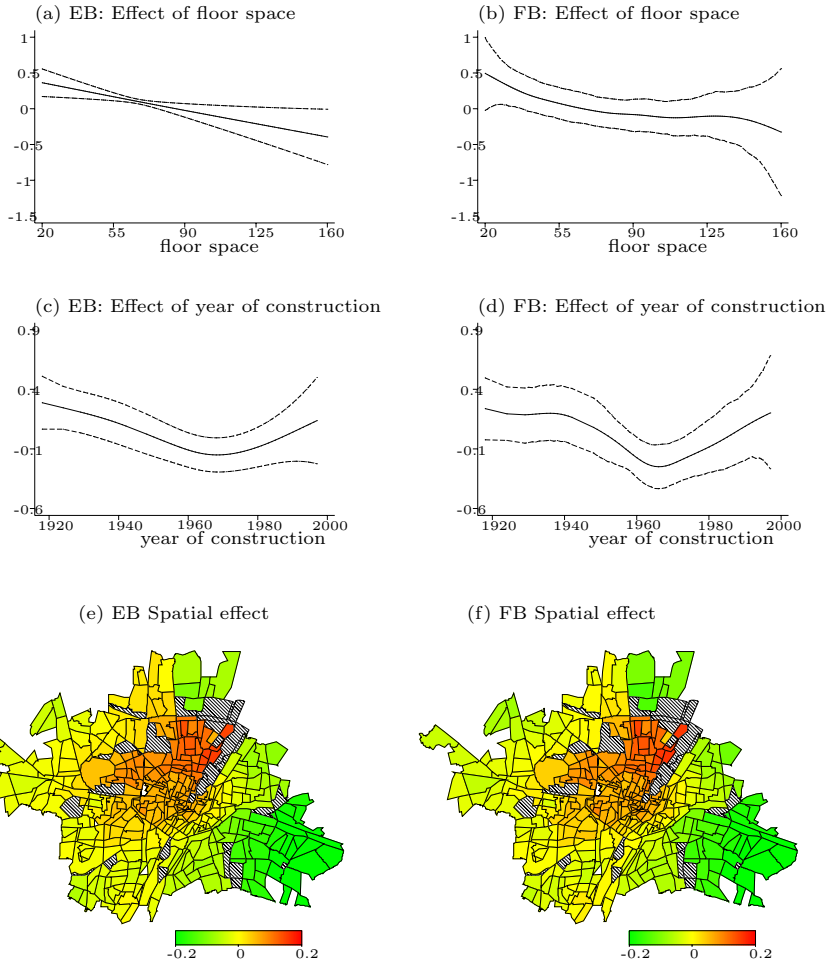

Figure 5. Rent data: Effects of floor space (top), year of construction (middle) and location (bottom) on the log of squared residuals $\log(e_i^2)$ for EB (left panel) and FB (right panel), respectively. Shown are the posterior mode (EB) and mean (FB) estimates. For floor space and year of construction, pointwise 95% credible intervals are included additionally.

The log of squared residuals with predictor (29) is not only useful to construct weights, it is also important for constructing appropriate prediction intervals using $s_i^2 = \exp(\eta_i)$ as an estimate of the error variance $\sigma_i^2 = Var(\varepsilon_i)$. Figure 5 shows that floor space and year of construction have a significant effect on the variance. While the effect of floor space decreases linearly with increasing floor space, the effect of year of construction is lower in the sixties and the seventies compared to other years. This can be explained by a boom in construction building in these years, with flats having comparably homogenous quality. The

shape of the EB confidence interval in Figure 5(a) is caused by centering F about 0. The effect of location also provides interesting evidence of increased variance in the central quarters of Munich, whereas some of the suburban quarters are more homogeneous.

5.2. A space-time study on forest damage

These longitudinal data have been collected in yearly visual forest damage inventories carried out in a forest district in the northern part of Bavaria from 1983 to 2001. The observation area extends 15 km from east to west and 10 km from north to south, with 84 stands of trees as observation points. In the following application, we consider beeches. For each tree, the degree of defoliation serves as an indicator for its damage state, which is given as a binary response, with $y_{it} = 1$ (damage of tree i in year t) and $y_{it} = 0$ (no damage), $i = 1, \ldots, 84$, $t = 1983, \ldots, 2001$. Figure 6 shows the temporal development of the frequency of damaged trees, and the spatial distribution of trees together with the percentage of damage, averaged over the entire observation period. For an illustrative analysis with a spatio-temporal logit model we include age A_{it} (in years) of the tree and canopy density C_{it} at the stand, measured in steps of 10 %, as the most influential covariates. The pH-value of the soil is less important here, because it does not vary a lot within the observation area. Therefore we chose the following logit model

$$\log \frac{P(y_{it} = 1)}{P(y_{it} = 0)} = \gamma_0 + f_1(t) + f_2(A_{it}) + f_3(C_{it}) + f_4(S_i),$$

where the function f_1,f_2 and f_3 are modelled through cubic P-splines with second order random walk penalty, and the spatial component f_4 follows Markov random field prior (10), with S_i denoting the site of tree i. Two trees are considered as neighbors if their distance is less than 1.2 km.

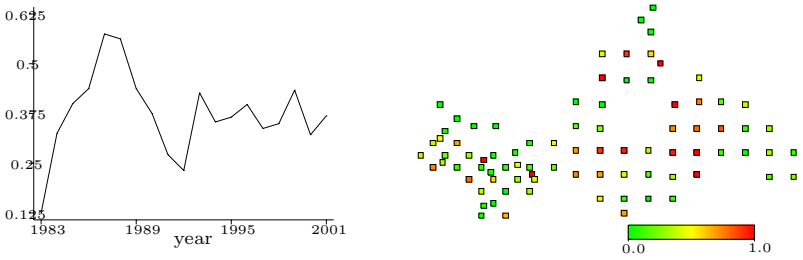

Figure 6. Forest health data: the left panel shows the temporal development of the frequency of damaged trees. The right panel displays the percentage of damage, averaged over the entire observation period.

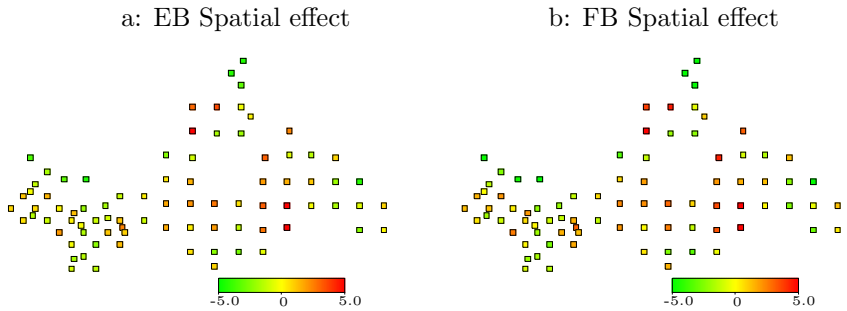

Figure 7. Forest health data: spatial effect for EB (left panel) and FB (right panel). Shown is the posterior mode for EB and the posterior mean for FB.

Figure 8 shows the estimated functions f_1, f_2 and f_3 for EB and FB, respectively. Functions f_1 and f_2 are clearly nonlinear and, again, EB and FB results are rather similar, even for credible intervals. The effect f_1 of calendar time reflects the descriptive trend in Figure 6, with a peak in the mid-eighties, recovering thereafter and staying on a more or less constant level in the nineties. Astonishingly, the nonlinear effect of age is not monotone, with a first peak around 65 years. The effect f_3 of canopy density appears to be linearly decreasing, which means that a dense stand is good for the health of beeches. Note, that this conclusion depends on the type of tree, and can be quite different for other species. The spatial effects in Figure 7 reflect the raw spatial effects in Figure 6, with EB and FB estimates being close to each other again.

In Table 3 we compare the classification of trees for all years based on the spatio-temporal logit model and, alternatively, on a model without the spatial component f_4. The classification table of the spatio-temporal model shows a clear improvement, confirming that inclusion of the spatial information is substantial.

Table 3. Forest health data: Classification tables. Table a) shows the classification including a spatial effect and Table b) shows the classification without spatial effect. In each cell of the tables predictions for EB estimates are shown first and predictions for FB estimates are shown second. Predictions are based on the Bayes rule, i.e., $\hat{y}_{it} = 1$ if $\widehat{P}(y_{it} = 1) > \widehat{P}(y_{it} = 0)$ and $\hat{y}_{it} = 0$ otherwise.

(a) $\hat{y}_{it}$

y_{it}	0	1
0	900 / 896	71 / 75
1	113 / 110	465 / 468

(b) $\hat{y}_{it}$

y_{it}	0	1
0	846 / 848	125 / 123
1	207 / 209	371 / 369

6. Conclusions

We developed empirical Bayesian inference, based on mixed model representations, for a broad class of structured additive regression models. The ap-

proach has been compared to full Bayesian inference using MCMC techniques through simulation studies and applications to spatial and longitudinal regression data. Because we use a computationally efficient modification of the usual version of REML estimation of smoothing parameters, empirical Bayes inference is a promising alternative to full Bayes inference even for fairly large data sets. As the applications to artificial and real data sets show, posterior mode (EB) and mean (FB) estimators are often rather similar, motivating theoretical work to justify this for large samples.

The software provided greatly facilitates applications of the methods in other areas than considered in this paper, and should be of relevance for applied researchers in economics and biostatistics who are confronted with space-time data.

In future research, we aim at extending the methodology to multivariate and multicategorical responses as well as to survival and event history data.

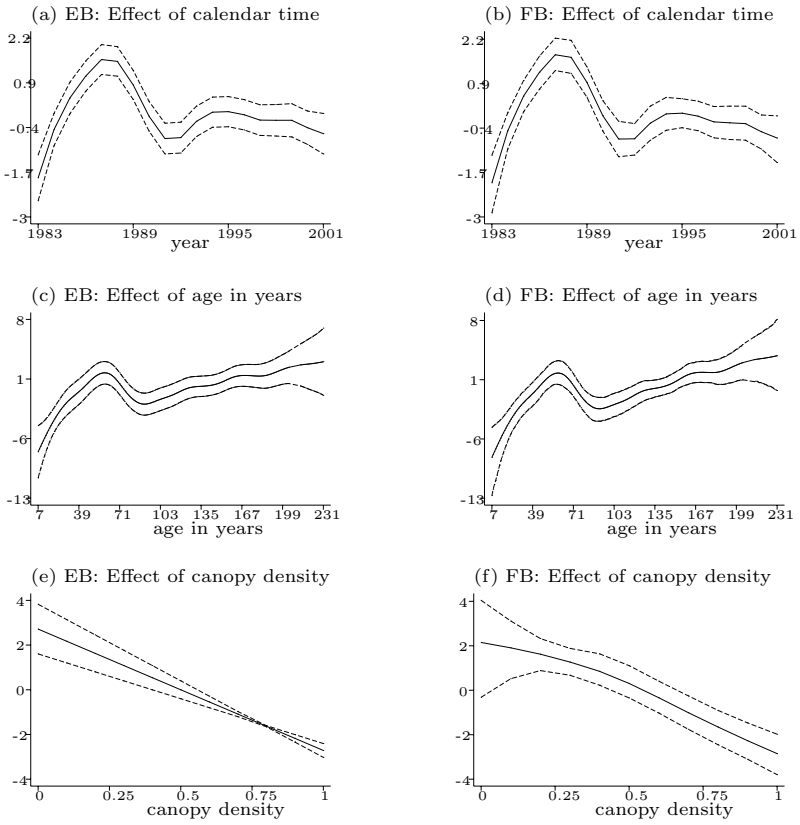

Figure 8. Forest health data: effects of calendar time (top), age of the trees (middle) and canopy density (bottom) for EB (left panel) and FB (right panel), respectively. Shown are the posterior mode (EB) and mean (FB) estimates, together with pointwise 95% credible intervals.

Acknowledgement

We thank an anonymous referee, an associate editor and Leonhard Held for helpful comments. This research has been financially supported by grants from the German Science Foundation (DFG), Sonderforschungsbereich 386 "Statistical Analysis of Discrete Structures".

References

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.* **43**, 1-59.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82**, 733-746.

Biller, C. (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models. *J. Comput. Graph. Statist.* **9**, 122-140.

Biller, C. and Fahrmeir, L. (2001). Bayesian varying-coefficient models using adaptive regression splines. *Statist. Modeling* **2**, 195-211.

Brezger, A. and Lang, S. (2003). Generalized additive regression based on Bayesian P-splines. SFB 386 Discussion paper 321, Department of Statistics, University of Munich.

Chen, Z. (1993). Fitting multivariate regression functions by interaction spline models. *J. Roy. Statist. Soc. Ser. B* **55**, 473-491.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *J. Roy. Statist. Soc. Ser. B* **60**, 333-350.

Di Matteo, I., Genovese, C. R. and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88**, 1055-1071.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing using B-splines and penalties (with comments and rejoinder). *Statist. Sci.* **11**, 89-121.

Fahrmeir, L. and Knorr-Held, L. (2000). Dynamic and semiparametric models. In *Smoothing and Regression: Approaches, Computation and Application* (Edited by M. Schimek). Wiley, New York.

Fahrmeir, L. and Lang, S. (2001a). Bayesian inference for generalized additive mixed models based on Markov random field priors. *J. Roy. Statist. Soc. Ser. C* **50**, 201-220.

Fahrmeir, L. and Lang, S. (2001b). Bayesian semiparametric regression analysis of multicategorical time-space data. *Ann. Inst. Statist. Math.* **53**, 10-30.

Fahrmeir, L., Lang, S., Wolff, J. and Bender, S. (2003). Semiparametric Bayesian time-space analysis of unemployment duration. *J. German Statist. Soc.* **87**, 281-307.

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

Fotheringham, A. S., Brunsdon, C. and Charlton, M. E. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships.* Wiley, Chichester.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19**, 1-141.

Gamerman, D. (1997). Efficient Sampling from the posterior distribution in generalized linear models. *Statist. Comput.* **7**, 57-68.

Gamerman, D., Moreira, A. R. B. and Rue, H. (2003). Space-varying regression models: specifications and simulation. *Comput. Statist. Data Anal.* **42**, 513-533.

George, A. and Liu, J. W. (1981). *Computer Solution of Large Sparse Positive Definite Systems.* Prentice-Hall.

Green, P. J. (1987). Penalized likelihood for general semiparametric regression models. *Internat. Statist. Rev.* **55**, 245-259.

Hansen, M. H. and Kooperberg, C. (2002). Spline adaptation in extended linear models. *Statist. Sci.* **17**, 2-51.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72**, 320-338.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55**, 757-796.

Hastie, T. and Tibshirani, R. (2000). Bayesian backfitting. *Statist. Sci.* **15**, 193-223.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.* **91**, 1461-1473.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *J. Roy. Statist. Soc. Ser. C* **52**, 1-18.

Knorr-Held, L. (2004). Simultaneous posterior probability statements from Monte Carlo output. *J. Comput. Graph. Statist.* **13**, 20-35.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *J. Comput. Graph. Statist.* **13**, 183-212.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *J. Roy. Statist. Soc. Ser. B* **61**, 381-400.

Marx, B. and Eilers, P. (1998). Direct generalized additive modeling with penalized likelihood. *Comput. Statist. Data Anal.* **28**, 193-209.

Mc Culloch, C. E. and Searle, S. R. (2001). *Generalized Linear and Mixed Models.* Wiley, New York.

Müller, H. G., Stadtmüller, U. and Tabnak, F. (1997). Spatial smoothing of geographically aggregated data, with applications to the construction of incidence maps. *J. Amer. Statist. Assoc.* **92**, 61-71.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields with applications. *J. Roy. Statist. Soc. Ser. B* **63**, 325-338.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.* **29**, 31-49.

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

Stone, C. J., Hansen, M., Kooperberg, C. and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25**, 1371-1470.

Wahba, G. (1978). Improper Prior, Spline smoothing and the problem of guarding against model errors in regression . *J. Roy. Statist. Soc. Ser. B* **44**, 364-372.

Department of Statistics, University of Munich, Ludwigstr. 33, D-80539 Munich, Germany.

E-mail: fahrmeir@stat.uni-muenchen.de

Department of Statistics, University of Munich, Ludwigstr. 33, D-80539 Munich, Germany.

E-mail: kneib@stat.uni-muenchen.de

Department of Statistics, University of Munich, Ludwigstr. 33, D-80539 Munich, Germany.

E-mail: lang@stat.uni-muenchen.de

(Received January 2003; accepted March 2004)