Brezger, Lang:

Generalized structured additive regression based on Bayesian P-splines

Sonderforschungsbereich 386, Paper 321 (2003)

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

Generalized structured additive regression based on Bayesian P-splines

Andreas Brezger and Stefan Lang

Department of Statistics,
University of Munich.
email: andib@stat.uni-muenchen.de,
lang@stat.uni-muenchen.de

June 30, 2003

Abstract

Generalized additive models (GAM) for modelling nonlinear effects of continuous covariates are now well established tools for the applied statistician. In this paper we develop Bayesian GAM's and extensions to generalized structured additive regression based on one or two dimensional P-splines as the main building block. The approach extends previous work by Lang and Brezger (2003) for Gaussian responses. Inference relies on Markov chain Monte Carlo (MCMC) simulation techniques, and is either based on iteratively weighted least squares (IWLS) proposals or on latent utility representations of (multi)categorical regression models. Our approach covers the most common univariate response distributions, e.g. the Binomial, Poisson or Gamma distribution, as well as multicategorical responses. As we will demonstrate through two applications on the forest health status of trees and a space-time analysis of health insurance data, the approach allows realistic modelling of complex problems. We consider the enormous flexibility and extendability of our approach as a main advantage of Bayesian inference based on MCMC techniques compared to more traditional approaches. Software for the methodology presented in the paper is provided within the public domain package *BayesX*.

Key words: geoadditive models, IWLS proposals, multicategorical response, structured additive predictors, surface smoothing

1 Introduction

Generalized additive models (GAM) provide a powerful class of models for modelling nonlinear effects of continuous covariates in regression models with non-Gaussian responses. A huge variety of competing approaches are now available for modelling and estimating nonlinear functions of continuous covariates. Prominent examples are smoothing splines (e.g. Hastie and Tibshirani (1990)), local polynomials (e.g. Fan and Gijbels (1996)), regression splines with adaptive knot selection (e.g. Friedman and Silverman (1989), Friedman (1991), Stone et al. (1997)) and P-splines (Eilers and Marx (1996), Marx and Eilers (1998)). Currently, smoothing based on mixed model representations of GAM's and extensions is extremely popular, see e.g. Lin and Zhang (1999), Currie and Durban

(2002), Wand (2003) and the book by Ruppert et al. (2003). Indeed, the approach is very promising and has several distinct advantages, e.g. smoothing parameters can be estimated simultaneously with the regression functions.

Bayesian approaches are currently either based on regression splines with adaptive knot selection (e.g. Smith and Kohn (1996), Denison et al. (1998), Biller (2000), Di Matteo et al. (2001), Biller and Fahrmeir (2001) and Hansen and Kooperberg (2002)), or on smoothness priors (Hastie and Tibshirani (2000), Fahrmeir and Lang (2001a) and Fahrmeir and Lang (2001b)).

In this paper, we extend previous work by Lang and Brezger (2003) for Gaussian responses based on one or two dimensional Bayesian P-splines as the main building block. Our approach covers univariate GAM's for the most common response distributions (Binomial, Poisson, Gamma) as well as models for multicategorical responses. Inference is fully Bayesian and is based on Markov chain Monte Carlo inference techniques (a nice introduction into MCMC can be found in Green (2001)). We adopt iteratively weighted least squares (IWLS) proposals as proposed by Gamerman (1997) for generalized linear mixed models. Similar proposals have been used by Rue (2001) and Knorr-Held and Rue (2002) primarily in the context of spatial smoothing. A simple alternative are conditional prior proposals (Knorr-Held (1999)) which work surprisingly well in many situations. For most categorical response models, efficiency of MCMC inference can be considerably improved by using latent utility respresentations of such models, see Albert and Chib (1993) and Chen and Dey (2000) for (multicategorical) probit models and Holmes and Knorr-Held (2003) for binary logit models. The advantage of such representations for MCMC inference is, that the full conditionals of the regression coefficients are (multivariate) Gaussian and sampling schemes developed for Gaussian responses in Lang and Brezger (2003) can be utilized with only minor changes. In all cases, numerical efficiency is guaranteed by using matrix operations for band or sparse matrices (George and Liu (1981)).

A main advantage of a Bayesian approach for GAM's is its flexibility and extendability to more complex formulations. Our approach can be well extended to deal with unobserved unit- or cluster specific heterogeneity by incorporating random intercepts or slopes into the predictor. Spatial heterogeneity may be considered by incorporating spatial effects. We will discuss two alternatives, Gaussian Markov random fields (e.g. Besag et al. (1991) and two dimensional P-splines (Lang and Brezger (2003)). Models that can deal simultaneously with nonlinear effects of continuous covariates as well as spatial heterogeneity are called *geoadditive models* (Kammann and Wand (2003)) and are of growing interest in the recent literature, see also Fahrmeir and Lang (2001a) and Fahrmeir and Lang (2001b). In general, we will use models with a *structured additive predictor* including generalized additive mixed models, dynamic models, varying coefficient models and geoadditive models as a special case.

We will present examples of generalized structured additive models in two applications. In our first application we analyse longitudinal data on the health status of beeches in the forest district of Rothenburg in northern Bavaria. Important influencial factors on the health state of trees are e.g. the age of the trees, the canopy density at the stand, calendar time as a surrogate for changing environmental conditions, and the location of the stand. The second application is a space-time analysis of hospital treatment costs based on data from a German private health insurance company.

Another important advantage of inference based on MCMC is easy prediction for unobserved covariate combinations including credible intervals, and the availability of inference for functions of the parameters (again including credible intervals). We will give specific examples in our second application.

The methodology of this paper is included in *BayesX*, a software package for Bayesian inference based on MCMC simulation techniques. *BayesX* is an easy to use public domain program. The program together with a detailed 130 pages manual can be downloaded from `http://www.stat.uni-muenchen.de/~lang/`. A particular advantage of *BayesX* is that it can estimate very complex models and handle large datasets.

The remainder of the paper is organized as follows: The next section describes Bayesian GAM's based on one or two dimensional P-splines and discusses extensions to generalized structured additive regression. Section 3 gives details about MCMC inference. In Section 4 we present two applications on the health status of trees and hospital treatment costs. Section 5 concludes and discusses directions for future research.

2 Bayesian GAM's and extensions

2.1 Bayesian P-splines

Suppose that observations (y_i, x_i, v_i), $i = 1, \dots, n$, are given, where y_i is a response variable, $x_i = (x_{i1}, \dots, x_{ip})'$ is a vector of continuous covariates and $v_i = (v_{i1}, \dots, v_{iq})'$ are further (mostly categorical) covariates. Generalized additive models (Hastie and Tibshirani (1990)) assume that, given x_i and v_i the distribution of y_i belongs to an exponential family, i.e.

$$p(y_i \,|\, x_i, v_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\right) c(y_i, \theta_i) \tag{1}$$

where $b(\cdot)$, $c(\cdot)$, θ_i and ϕ determine the respective distributions. A list of the most common distributions and their specific parameters can be found e.g. in Fahrmeir and Tutz (2001), page 21. The mean $\mu_i = E(y_i|x_i, v_i)$ is linked to a semiparametric additive predictor η_i by

$$\mu_i = h(\eta_i), \quad \eta_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + v_i'\gamma. \tag{2}$$

Here, h is a known response function and $f_1, \dots, f_p$ are unknown smooth functions of the continuous covariates and $v_i'\gamma$ represents the strictly parametric part of the predictor.

For modelling the unknown functions f_j we follow Lang and Brezger (2003), who present a Bayesian version of the P-splines approach introduced in a frequentist setting by Eilers and Marx (1996). Here, we assume that the unknown functions can be approximated by a polynomial spline of degree l and with equally spaced knots

$$\zeta_{j0} = x_{j,min} < \zeta_{j1} < \cdots < \zeta_{j,k-1} < \zeta_{jk_j} = x_{j,max}$$

over the domain of x_j. The spline can be written in terms of a linear combination of $M_j = k_j + l$ B-spline basis functions (De Boor (1978). Denoting the m-th basis function by B_{jm}, we obtain

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

By defining the $n \times M_j$ design matrices X_j with the elements in row i and column m given by $X_j(i, m) = B_{jm}(x_{ij})$, we can rewrite the predictor (2) in matrix notation as

$$\eta = X_1\beta_1 + \cdots + X_p\beta_p + V'\gamma. \tag{3}$$

Here, $\beta_j = (\beta_{j1}, \dots, \beta_{jm})'$, $j = 1, \dots, p$, correspond to the vectors of unknown regression coefficients. The matrix V is the usual design matrix for fixed effects. To overcome the well known difficulties involved with regression splines, Eilers and Marx (1996) suggest a relatively large number of knots (usually between 20 to 40) to ensure enough flexibility, and to introduce a roughness penalty on adjacent regression coefficients to regularize the problem and avoid overfitting. In their frequentist approach they use penalties based on squared r-th order differences. Usually first or second order differences are enough. In our Bayesian approach, we replace first or second order differences with their stochastic analogues, i.e. first or second order random walks defined by

$$\beta_{jm} = \beta_{j,m-1} + u_{jm}, \quad \text{or} \quad \beta_{jm} = 2\beta_{j,m-1} - \beta_{j,m-2} + u_{jm} \tag{4}$$

with Gaussian errors $u_{jm} \sim N(0, \tau_j^2)$ and diffuse priors $\beta_{j1} \propto const$, or β_{j1} and $\beta_{j2} \propto const$, for initial values, respectively. The amount of smoothness is controlled by the variance parameter τ_j^2 which corresponds to the smoothing parameter in the traditional approach. By defining an additional hyperprior for the variance parameters the amount of smoothness can be estimated simultaneously with the regression coefficients. We assign the conjugate prior for τ^2 which is an inverse Gamma prior with hyperparameters a_j and b_j, i.e. $\tau_j^2 \sim IG(a_j, b_j)$. Common choices for a_j and b_j are $a_j = 1$ and b_j small, e.g. $b = 0.005$ or $b_j = 0.0005$. Alternatively we may set $a_j = b_j$, e.g. $a_j = b_j = 0.001$. As a standard choice we use $a_j = 1$ and $b_j = 0.005$. Since the results may considerably depend on the choice of a_j and b_j some sort of sensitivity analysis is strongly recommended. For instance, the models under consideration could be reestimated with (a small) number of different choices for a_j and b_j.

In some situations, a global variance parameter τ_j^2 may be not appropriate, for example if the underlying function is highly oscillating. In such cases the assumption of a global variance parameter τ_j^2 may be relaxed by replacing the errors $u_{jm} \sim N(0, \tau_j^2)$ in (4) by $u_{jm} \sim N(0, \tau_j^2/\delta_{jm})$. The weights δ_{jm} are additional hyperparameters and assumed to follow independent Gamma distributions $\delta_{jm} \sim G(\frac{\nu}{2}, \frac{\nu}{2})$. This is equivalent to a t-distribution with ν degrees of freedom for β_j (see e.g. Knorr-Held (1996) in the context of dynamic models). As an alternative, *locally adaptive dependent* variances as proposed in Lang et al. (2002) could be used as well. Our software is capable of estimating such models, but we do not investigate them in the following to keep the paper in reasonable length. Estimation is, however, straightforward, see Lang and Brezger (2003) and Lang et al. (2002) for details.

2.2 Modelling interactions

In many situations, the simple additive predictor (2) may be not appropriate because of interactions between covariates. In this section we describe interactions between categorical and continuous covariates and between two continuous covariates. In the next section, we also discuss interactions between space and categorical covariates. For notational simplicity, we keep the notation of the predictor as in (2) and assume for the rest of the section that a particular covariate x_j is now two dimensional, i.e. $x_{ij} = (x_{ij}^1, x_{ij}^2)'$.

Interactions between categorical and continuous covariates can be conveniently modelled within the varying coefficient framework introduced by Hastie and Tibshirani (1993). Here, the effect of covariate x_{ij}^1 is assumed to vary smoothly over the range of the second covariate x_{ij}^2, i.e.

$$f_j(x_{ij}) = f_j'(x_{ij}^2) x_{ij}^1. \tag{5}$$

The covariate x_{ij}^2 is called the effect modifier of x_{ij}^1. The design matrix X_j is given by $diag(x_{1j}^1, \dots, x_{nj}^1)X_j^2$ where X_j^2 is the usual design matrix for splines composed of the basis functions evaluated at the observations x_{ij}^2.

If both interacting covariates are continuous, a more flexible approach for modelling interactions can be based on two dimensional surface fitting. Here, we concentrate on two dimensional P-splines described in Lang and Brezger (2003), see also Wood (2003) for a recent approach based on thin plate splines. We assume that the unknown surface $f_j(x_{ij})$ can be approximated by the tensor product of one dimensional B-splines, i.e.

$$f_j(x_{ij}^1, x_{ij}^2) = \sum_{m_1=1}^{M_{1j}} \sum_{m_2=1}^{M_{2j}} \beta_{j,m_1m_2} B_{j,m_1}(x_{ij}^1) B_{j,m_2}(x_{ij}^2). \tag{6}$$

The design matrix X_j is now $n \times M_{1j} \cdot M_{2j}$ dimensional and consists of products of basis functions. Priors for $\beta_j = (\beta_{j,11}, \dots, \beta_{j,M_{1j}M_{2j}})'$ are based on spatial smoothness priors common in spatial statistics (see e.g. Besag and Kooperberg (1995)). Based on previous experience, we prefer a two dimensional first order random walk based on the four nearest neighbours. It is usually defined by specifying the conditional distributions of a parameter given its neighbours, i.e.

$$\beta_{jm_1m_2} \,|\, \cdot \sim N\left(\frac{1}{4}(\beta_{jm_1-1,m_2} + \beta_{jm_1+1,m_2} + \beta_{jm_1,m_2-1} + \beta_{jm_1,m_2+1}), \frac{\tau_j^2}{4}\right) \tag{7}$$

for $m_1 = 2, \dots, M_{1j}-1$,$m_2 = 2, \dots, M_{2j}-1$ and appropriate changes for corners and edges. Again, we restrict the unknown function f_j to have mean zero to guarantee identifiability.

Sometimes it is desirable to decompose the effect of the two covariates x_j^1 and x_j^2 into two main effects modelled by one dimensional functions and a two dimensional interaction effect. Then, we obtain

$$f_j(x_{ij}) = f_j^1(x_{ij}^1) + f_j^2(x_{ij}^2) + f_j^{12}(x_{ij}^1, x_{ij}^2). \tag{8}$$

In this case, additional identifiability constraints have to be imposed on the three functions, see Lang and Brezger (2003).

2.3 Structured additive predictors

A main advantage of Bayesian regression analysis based on MCMC simulation techniques is that models can be easily extended to more complex formulations. So far, we have considered only continuous and categorical covariates in the predictor. In this section, we relax this assumption by allowing that the covariates x_j in (2) or (3) are not necessarily continuous. We still pertain the assumption of the preceeding section that covariates x_j may be one or two dimensional. Based on this assumptions the models can be considerably extended within a unified framework. We are particularly interested in the handling of unobserved unit- or cluster specific and spatial heterogeneity. Models that can deal with spatial heterogeneity are called *geoadditive models* (Kamman and Wand, 2001). We call a predictor with one or two dimensional nonlinear effects of continuous covariates, time scales and unit- or cluster specific and spatial heterogeneity a *structured additive predictor* because it still has an additive structure but is more flexible than the usual predictor in GAM's.

Unit- or cluster specific heterogeneity

Suppose that covariate x_j is a index variable that indicates the unit or cluster a particular observation belongs to. An example are longitudinal data where x_j is an individuum index. In this case, it is common practice to introduce unit- or cluster specific i.i.d. Gaussian random intercepts or slopes, see e.g. Diggle et al. (1994). Suppose x_j can take the values $1, \ldots, M_j$. Then, an i.i.d. random intercept can be incorporated into our framework of structured additive regression by assuming $f_j(m) = \beta_{jm} \sim N(0, \tau_j^2)$, $m = 1, \ldots, M_j$. The design matrix X_j is now a 0/1 incidence matrix with dimension $n \times M_j$. In order to introduce random slopes we assume $x_j = (x_j^1, x_j^2)$ as in Section 2.2. Then, a random slope with respect to index variable x_j^1 is defined as $f_j(x_{ij}) = f_j'(x_{ij}^2)x_{ij}^1$ with $f_j'(x_{ij}^2) = \beta_{jm} \sim N(0, \tau_j^2)$. The design matrix X_j is given by $diag(x_{1j}^1, \ldots, x_{nj}^1)X_j^2$ where X_j^2 is again a 0/1 incidence matrix. Note the close similarity between random slopes and varying coefficient models. In fact, random slopes may be regarded as varying coefficient terms with unit- or cluster variable x_j^2 as the effect modifier.

Spatial heterogeneity

To consider spatial heterogeneity, we may introduce a *spatial effect* f_j of location x_j to the predictor. Depending on the application, the spatial effect may be further split up into a spatially correlated (structured) and an uncorrelated (unstructured) effect, i.e. $f_j = f_{j,str} + f_{j,unstr}$. The correlated effect $f_{j,str}$ aims at capturing spatially dependent heterogeneity and the uncorrelated effect $f_{j,unstr}$ local effects.

For data observed on a regular or irregular lattice a common approach for the correlated spatial effect f_{str} is based on Markov random field priors, see e.g. Besag et al. (1991). Let $s \in \{1, \ldots, S_j\}$ denote the pixels of a lattice or the regions of a geographical map. Then, the most simple Markov random field prior for $f_{str}(s) = \beta_{str,s}$ is defined by

$$\beta_{str,s}|\beta_{str,u}, u \neq s \sim N\left(\sum_{u \in \partial_s} \frac{1}{N_s}\beta_{str,u}, \frac{\tau_{str}^2}{N_s}\right), \tag{9}$$

where N_s is the number of adjacent regions or pixels, and ∂_s denotes the regions which are neighbours of region s. Hence, prior (9) can be seen as a two dimensional extension of a first order random walk. More general priors than (9) are described in Besag et al. (1991). The design matrix X_{str} is a $n \times S_j$ incidence matrix whose entry in the i-th row and s-th column is equal to one if observation i has been observed at location s and zero otherwise.

Alternatively, the structured spatial effect f_{str} could be modelled by two dimensional surface estimators as described in Section 2.2. In most of our applications, however, the MRF random field proves to be superior in terms of model fit.

For the unstructured effect f_{unstr} we may again assume i.i.d Gaussian random effects with the location as the index variable.

Similar to continuous covariates and index variables we can again define varying coefficient terms, now with a spatial covariate as the effect modifier, see e.g. Fahrmeir et al. (2003) for an application.

2.4 General structure of the priors

As we have seen, it is always possible to express the vector of function evaluations $f_j = (f_{j1}, \ldots, f_{jn})$ of a nonlinear effect as the matrix product of a design matrix X_j and a vector of regression coefficients β_j, i.e. $f_j = X_j \beta_j$. It turns out that the smoothness priors for the regresssion coefficients β_j can be cast into a general form as well. It is given by

$$\beta_j | \tau_j^2 \propto \frac{1}{(\tau_j^2)^{rk(K_j)}} \exp(-\frac{1}{2\tau_j^2} \beta_j' K_j \beta_j), \tag{10}$$

where K_j is a *penalty matrix* which depends on the prior assumptions about *smoothness of f_j* and the *type of covariate*. E.g. for a P-spline with a first order random walk penalty K_j is given by

$$K = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}.$$

For the variance parameter an inverse Gamma prior (the conjugate prior) is assumed, i.e. $\tau_j^2 \sim IG(a_j, b_j)$.

The general structure of the priors particularly facilitates the description of MCMC inference in the next section.

3 Bayesian inference via MCMC

Bayesian inference is based on the posterior of the model which is given by

$$\begin{aligned} p(\alpha \,|\, y) \quad \propto \quad & L(y, \beta_1, \tau_1^2, \ldots, \tau_p^2, \beta_p, \gamma) \\ & \prod_{j=1}^{p} \frac{1}{(\tau_j^2)^{rk(K_j)}} \exp(-\frac{1}{2\tau_j^2} \beta_j' K_j \beta_j) \\ & \prod_{j=1}^{p} (\tau_j^2)^{-a_j - 1} \exp(-\frac{b_j}{\tau_j^2}), \end{aligned}$$

where α is the vector of all parameters in the model. The likelihood $L(\cdot)$ is a product of the individual likelihoods (1). Since the posterior is analytically intractable we make use of Markov chain Monte Carlo (MCMC) simulation techniques. Models with Gaussian responses are already covered in Lang and Brezger (2003). Here, the main focus is on methods applicable for general distributions from an exponential family. We first adopt an approach based on iteratively weighted least squares proposals as suggested by Gamerman (1997) in the context of generalized linear mixed models (Subsection 3.1). For (multi)categorical responses more efficient sampling schemes can be developed by considering latent utility representations of the models (Subsection 3.3). In both approaches, MCMC simulation is based on drawings from full conditionals of blocks of parameters, given the rest and the data. Parameters are updated in the order $\tau_1^2, \beta_1, \ldots, \tau_p^2, \beta_p, \gamma$.

3.1 Updating by iteratively weighted least squares (IWLS) proposals

The basic idea is to combine Fisher scoring or IWLS (e.g. Fahrmeir and Tutz (2001)) for estimating regression parameters in generalized linear models, and the Metropolis-

Hastings algorithm. More precisely, the goal is to approximate the full conditionals of regression parameters β_j and γ by a Gaussian distribution, obtained by accomplishing *one* Fisher scoring step in every iteration of the sampler. Suppose we want to update the regression coefficients β_j of the function f_j with current state 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

$$P_j = X_j' W(\beta_j^c) X_j + \frac{1}{\tau_j^2} K_j, \quad m_j = P_j^{-1} X_j' W(\beta_j^c)(\tilde{y}(\beta_j^c) - \tilde{\eta}). \tag{11}$$

Here, $W(\beta_j^c) = diag(w_1, \ldots, w_n)$ is the usual weight matrix for IWLS with weights $w_i^{-1}(\beta_j^c) = b''(\theta_i)\{g'(\mu_i)\}^2$ obtained from the current state β_j^c. The working observations $\tilde{y}_i$ are defined as

$$\tilde{y}_i(\beta_j^c) = \eta_i + (y_i - \mu_i) g'(\mu_i).$$

The vector $\tilde{\eta}$ is the part of the predictor associated with all remaining effects in the model.

Note, that P_j is a symmetric matrix with band structure or which can be at least brought into a band matrix like structure. For one dimensional P-splines, the band size is $max\{$degree of splines l, order of differences $k\}$, for two dimensional P-splines the band size is $M_j \cdot l + l$, and for i.i.d random effects the posterior precision matrix is diagonal. For a Markov random field, the precision matrix is not a priori a band matrix but sparse. It can be transformed into a band matrix (with differing band size in every row) by reordering the regions using the Cuthill Mc-Kee algorithm (see George and Liu (1981) p. 58 ff). Hence, random numbers from the (high dimensional) proposal distributions can be efficiently drawn by using matrix operations, in particular Cholesky decompositions, for sparse matrices. In our implementation we use the *envelope method* for Cholesky decompositions of sparse matrices as described in George and Liu (1981), see also Rue (2001) and Lang and Brezger (2003).

Usually convergence and mixing of Markov chains is excellent with IWLS proposals. However, the following problems occured:

- *Improper starting values:* It turns out that convergence of the algorithm to the stationary distribution is sometimes extremely slow because of improper starting values for the β_j. As a remedy, we initialize the Markov chain with posterior mode estimates which are obtained from a backfitting algorithm with fixed and usually large values for the variance parameters.

- *Convergence problems for the variance parameter:* If the effect of two covariates x_j^1 and x_j^2 is decomposed into main effects and a two dimensional interaction effect as in (8), severe convergence problems for the variance parameter of the interaction effect are the rule. Similar, but less severe problems have been reported in Knorr-Held and Rue (2002) in the context of spatial smoothing with Markov random field priors. To overcome the difficulties, we follow Knorr-Held and Rue (2002) who propose to construct a *joint proposal* for the parameter vector β_j and the corresponding variance parameter τ_j^2, and to simultaneously accept/reject (β_j, τ_j^2). This is done by first sampling $(\tau_j^2)^p$ from a proposal distribution for τ_j^2, and subsequently drawing from the IWLS proposal for the corresponding regression parameters given the proposed $(\tau_j^2)^p$. The proposal distribution for τ_j^2 may depend on the current state $(\tau_j^2)^c$ of the variance, but *must be independent* of β_j^c. As suggested by Knorr-Held and Rue

(2002), we construct the proposal by multiplying the current state $(\tau_j^2)^c$ by a random variable z with density proportional to $1 + 1/z$ on the interval $[1/f, f]$, where $f > 1$ is a tuning constant. Since this proposal is independent of the regression parameters, the joint proposal for (β_j, τ_j^2) is the product of the two proposal densities. Following the advise of Knorr-Held and Rue (2002), we tune f to obtain acceptance probabilities of approximately 30%. The acceptance probability is given by

$$\alpha = \frac{L(y, \dots, \beta_j^p, (\tau_j^2)^p, \dots, \gamma^c)}{L(y, \dots, \beta_j^c, (\tau_j^2)^c, \dots, \gamma^c)} \frac{p(\beta_j^p \mid (\tau_j^2)^p) p((\tau_j^2)^p)}{p(\beta_j^c \mid (\tau_j^2)^c) p((\tau_j^2)^c)} \frac{q(\beta_j^p, \beta_j^c)}{q(\beta_j^c, \beta_j^p)}. \tag{12}$$

Computation of the acceptance probability requires the evaluation of the normalizing constant of the IWLS proposal which is given by $1/(2|P_j^{-1}|)^{0.5}$. The determinant of P_j^{-1} can be computed without significant additional effort as a by-product of the Cholesky decomposition. Note also that the proposal ratio of the smoothing parameter cancels out.

Summarizing, we obtain the following

Sampling scheme based on IWLS proposals:

1. *Initialization:*
   Compute the posterior mode for $\beta_1, \dots, \beta_j$ and γ given fixed variance parameters $\tau_j^2 = c_j$, (e.g. $c_j = 10$). The mode is computed via backfitting with Fisher scoring. Use the posterior mode estimates as the current state β_j^c,$(\tau_j^2)^c$,γ^c of the chain.

2. For $j = 1, \dots, p$:

   - *Propose new τ_j^2:* Sample a random number z with density proportional to $1 + 1/z$ on the interval $[1/f, f]$, $f > 1$. Set $(\tau_j^2)^p = (\tau_j^2)^c$ as the proposed new value for the jth variance parameter.
   - *Propose new β_j:* Sample a random number β_j^p from the multivariate Gaussian distribution with mean and covariance matrix defined in (11).
   - *Accept/reject $\beta_j^p, (\tau_j^2)^p$*: Accept the proposed values β_j^p and $(\tau_j^2)^p$ with acceptance probability (12). If the proposed random numbers are accepted, set $\beta_j^c = \beta_j^p$ and $(\tau_j^2)^c = (\tau_j^2)^p$.

3. *Update fixed effects parameters:* Draw a IWLS proposal γ^p from the Gaussian proposal density $q(\gamma^c, \gamma^p)$. Accept γ^p with probability

   $$\alpha = \frac{L(y, \dots, \gamma^p)}{L(y, \dots, \gamma^c)} \frac{q(\gamma^p, \gamma^c)}{q(\gamma^c, \gamma^p)}.$$

   If the proposal is accepted, set $\gamma^c = \gamma^p$.

4. If the random sample is large enough stop the algorithm, otherwise proceed with 2.

3.2 IWLS versus conditional prior proposals

As an alternative to IWLS proposals, we could use *conditional prior proposals* as suggested by Knorr-Held (1999) in the context of dynamic models and by Fahrmeir and Lang (2001a) for generalized additive mixed models based on simple random walk priors. Here, the parameter vector β_j is divided into smaller blocks $\beta_{j[r,s]} = (\beta_{jr}, \ldots, \beta_{js})$ and a proposal is drawn from the conditional prior distribution of $\beta_{j[r,s]}$ given the remaining parameters, see Fahrmeir and Lang (2001a) for details. The approach is computationally less demanding and distribution free in the sense that no approximation of any characteristics of the posterior (e.g. mode) is needed. A drawback is that careful tuning of block sizes is required to obtain satisfying mixing of the chain and to speed up convergence. To overcome this problem and to avoid several pre-runs in order to find appropriate block sizes, we perform an *automatic tuning*. This means that we check the acceptance rates in the initial burn in period and adjust the block sizes by a rule of thumb, to obtain acceptance rates between about 30% and 70%, which showed to produce the best mixing.

For P-splines and particulary Markov random fields conditionl prior proposals work surprisingly well in many siuations. There are however problems where IWLS proposals may help to substantially improve the mixing of the chains. We illustrate the improvements gained by a large data set ($n = 162548$) from two insurance companies in Belgium. The data have been analyzed in Denuit and Lang (2003) using the methods of this paper. The dependence of covariates on the number of claims reported by car holders was analyzed using a Poisson model. Here we depict the effect of the *bonus-malus score (bm)* indicating the level occupied in the 23-level Belgian bonus-malus scale. The left panel of Figure 1 shows the sampling paths for two particular parameters β_1^{bm} and β_{10}^{bm} obtained by using conditional prior proposals. Obviously, the mixing of the chains is not satisfactorily. The right panel shows that the mixing of the Markov chain can be substantially improved by using IWLS proposals. The mixing of the regression parameters also affects the associated variance parameters, see Figure 2.

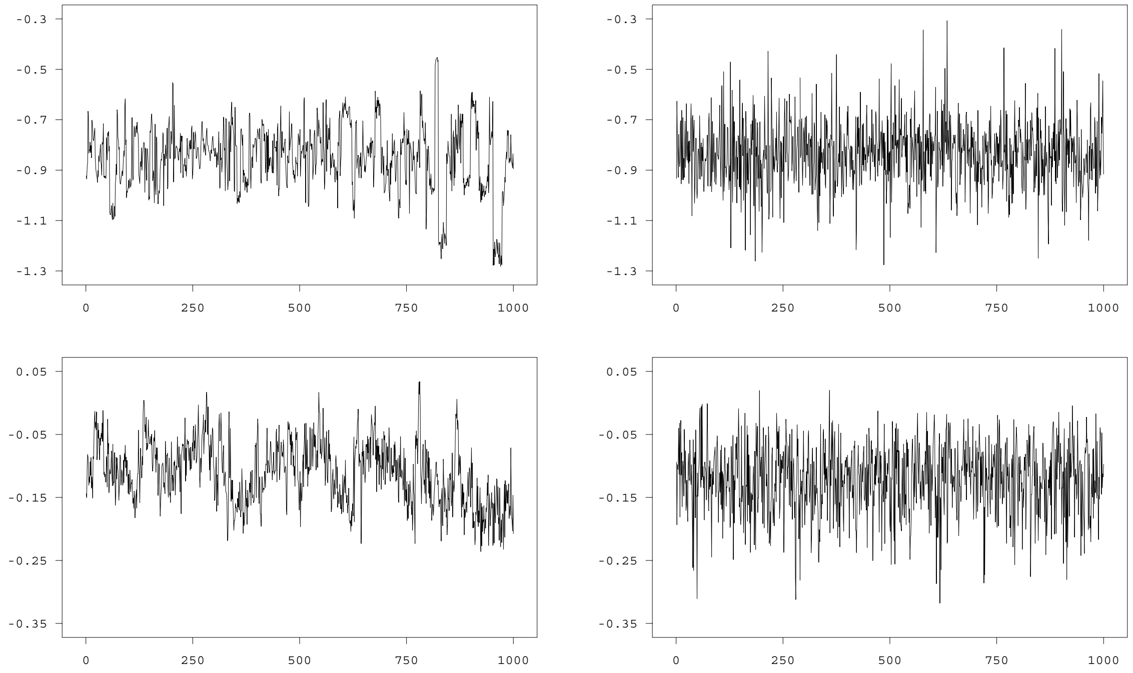


Figure 1: Sampling paths of two particular parameters (β_1^{bm},β_{10}^{bm}) with conditional prior proposal (left) and with IWLS proposal (right)

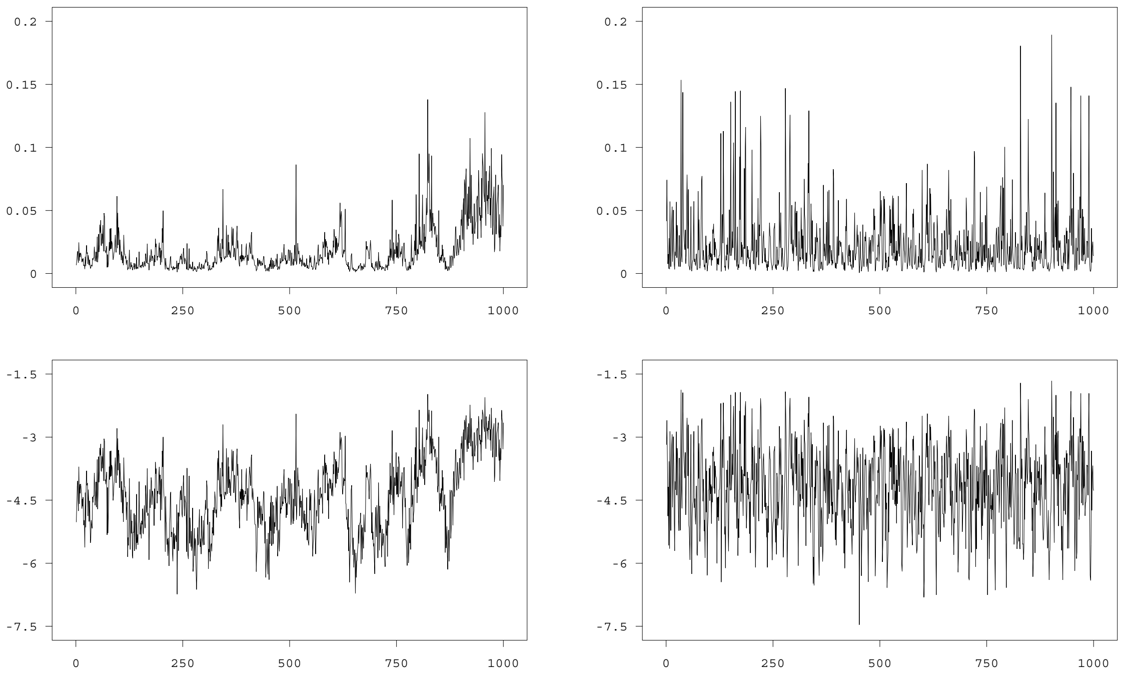


Figure 2: Sampling paths for the variance parameter τ^2_{bm} (top) and $\log(\tau^2_{bm})$ (bottom) resulting from conditional prior proposals for β (left) and from IWLS proposals (right).

3.3 Inference based on latent utility representations of categorical regression models

For most models with categorical responses efficient sampling schemes based on latent utility representations can be developed. The seminal paper by Albert and Chib (1993) develops algorithms for probit models with ordered categorical responses. The case of probit models with unordered multicategorical responses is delt with e.g. in Chen and Dey (2000) or Fahrmeir and Lang (2001b). Recently, another important data augmentation approach for binary logit models has been presented by Holmes and Knorr-Held (2003). The adaption of these sampling schemes to the models discussed in this paper is more or less straightforward. We briefly illustrate the concept for binary data, i.e. y_i takes only the values 0 or 1. We first assume a probit model. Conditional on the covariates and the parameters, y_i follows a Bernoulli distribution $y_i \sim B(1, \mu_i)$ with conditional mean $\mu_i = \Phi(\eta_i)$ where Φ is the cumulative distribution function of a standard normal distribution. Introducing latent variables

$$U_i = \eta_i + \epsilon_i, \tag{13}$$

with $\epsilon_i \sim N(0, 1)$, we define $y_i = 1$ if $U_i > 0$ and $y_i = 0$ if $U_i < 0$. It is easy to show that this corresponds to a binary probit model for the y_i's. The posterior of the model augmented by the latent variables depends now on the additional parameters U_i. Thus, an additional sampling step for updating the U_i's is required. Fortunately, sampling the U_i's is relatively easy and fast because the full conditionals are truncated normal distributions. More specifically, $U_i|\cdot \sim N(\eta_i, 1)$ truncated at the left by 0 if $y_i = 1$ and truncated at the right if $y_i = 0$. Efficient algorithms for drawing random numbers from a truncated normal distribution can be found in Geweke (1991) or Robert (1995). The advantage of defining a probit model through the latent variables U_i is that the full conditionals for the regression parameters β_j (and γ) are Gaussian with precision matrix and mean given by

$$P_j = X_j'X_j + \frac{1}{\tau_j^2}K_j, \quad m_j = P_j^{-1}X_j'(U - \tilde{\eta}). \tag{14}$$

Hence, the efficient and fast sampling schemes developed for Gaussian responses can be used with slight modifications. Updating of β_j and γ can be done exactly as described in Lang and Brezger (2003) using the current values U^c of the latent utilities as (pseudo) responses.

For binary logit models, the sampling schemes become slightly more complicated. A logit model can be expressed in terms of latent utilities by assuming $\epsilon_i \sim N(0, \lambda_i)$ in (13) with $\lambda_i = 4\psi_i^2$, where ψ_i follows a Kolmogorov-Smirnov distribution (Devroye (1986)). Hence, ϵ_i is a scale mixture of normal form with a marginal logistic distribution (Andrews and Mallows (1974)). The full conditionals for the $U_i's$ are still truncated normals with $U_i|\cdot \sim N(\eta_i, \lambda_i)$ but additional drawings from the conditional distributions of λ_i are necessary. Although the distribution has no standard form, sampling may be obtained by Metropolis-Hastings steps with the prior distribution for λ_i as a proposal (Holmes and Knorr-Held (2003)), i.e. draw a random number ψ_i from a Kolmogorov-Smirnov distribution and propose $\lambda_i^p = 4\psi^2$ as the new state of the Markov chain. The proposed new value is then accepted with probability

$$\alpha = \left(\frac{\lambda_i}{\lambda_i^p}\right)^{0.5} \exp\left(\frac{1}{2}(U_i - \eta_i)^2 \left(\frac{1}{\lambda_i} - \frac{1}{\lambda_i^p}\right)\right),$$

where λ_i is the current state of the chain.

4 Applications

4.1 Longitudinal study on forest health

In this longitudinal study on the health status of trees, we analyse the influence of calendar time t, age of trees A (in years), canopy density CP (in percent) and location L of the stand on the defoliation degree of beeches. Data have been collected in yearly forest damage inventories carried out in the forest district of Rothenbuch in northern Bavaria from 1983 to 2001. There are 80 observation points with occurence of beeches spread over an area extending about 15 km from east to west and 10 km from north to south. The degree of defoliation is used as an indicator for the state of a tree. It is measured in three ordered categories, with $y_{it} = 1$ for "bad" state of tree i in year t, $y_{it} = 2$ for "medium" and $y_{it} = 3$ for "good". A detailed data description can be found in Göttlein and Pruscha (1996).

We use a three-categorical ordered probit model based on a latent semiparametric model $U_{it} = \eta_{it} + \epsilon_{it}$ with predictor

$$\eta_{it} = f_1(t) + f_2(A_{it}) + f_{1|2}(t, A_{it}) + f_3(CP_{it}) + f_{str}(L_i). \tag{15}$$

The calendar time trend $f_1(t)$ and the age effect $f_2(A)$ are modelled by cubic P-splines with a second order random walk penalty. The interaction effect between calendar time and age $f_{1|2}(t, A)$ is modelled by a two dimensional cubic P-splines on a 12 by 12 grid of knots. Since canopy density is measrured only in 11 different values (0%, 10%,…,100%) we use a simple second order random walk prior (i.e. a P-spline of degree 0) for $f_3(CP)$. For the spatial effect $f_{str}(L)$ we experimented with both a two dimensional P-spline (model 1) and a Markov random field prior (model 2). Following Fahrmeir and Lang (2001b), the neighbourhood ∂_s of trees for the Markov random field includes all trees u with euclidian distance $d(s,u) \leq 1.2$ km. In terms of the DIC (Spiegelhalter et al. (2002)), the model based on the Markov random field is preferable. An unstructured spatial effect f_{unstr} is excluded from the predictor for the following two reasons. First, a look at the map of observation points (see Figure 5) reveals some sites with only one neighbour, making the identification of a structured and an unstructured effect difficult if not impossible. The second reason is that for each of the 80 sites only 19 observations on the same tree are available with only minor changes of the response category. In fact, there are only a couple of sites where all three response categories have been observed.

The data have been already analysed in Fahrmeir and Lang (2001b) (for the years 1983-1997 only). Here, nonlinear functions have been modelled solely by random walk priors. Also, the modelling of the interaction between calendar time and age is less sophisticated.

Since the results for the two models differ only for the spatial effect, we present for the remaining covariates only estimates based on model 2. Figure 3 shows the nonlinear main effects of calendar time and age of the tree as well as the effect of canopy density. The interaction effect between calendar time and age is depicted in Figure 4. The spatial effect is shown in Figure 5. Results based on a two dimensional P-spline can be found in the left panels, results based on the Markov random field can be found in the right panels. Shown are posterior probabilities based on a nominal level of 80% (top panels) and 95% (bottom panels).

As we might have expected younger trees are in healthier state than the older ones. We also see that trees recover after the bad years around 1986, but after 1994 health status declines to a lower level again. The interaction effect between time and age is relatively strong. It suggests that the health status of young trees was better than average at the beginning of the observation period and considerably worsens in the years after 1986. For very old trees, the interaction effect is always positive. The distinct monotonic increase of the effect of canopy densities $\geq 30\%$ gives evidence that beeches get more shelter from bad environmental influences in stands with high canopy density. The spatial effect based on the two dimensional P-spline and the Markov random field are very similar. The Markov random field is slightly rougher (as could have been expected). Note that the spatial effect is quite strong and therefore not negligible.

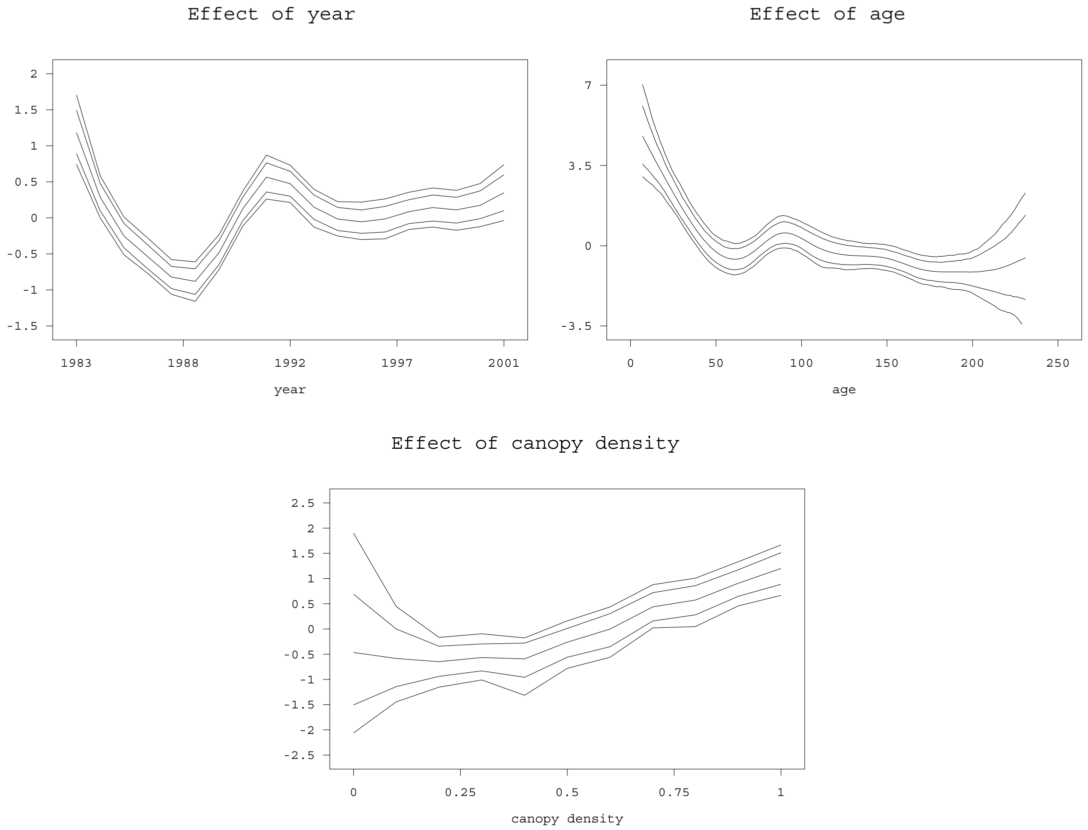


Figure 3: *Forest health data: Nonlinear main effects of calendar time, age of the tree and canopy density. Shown are the posterior means together with 95% and 80% pointwise credible intervals.*

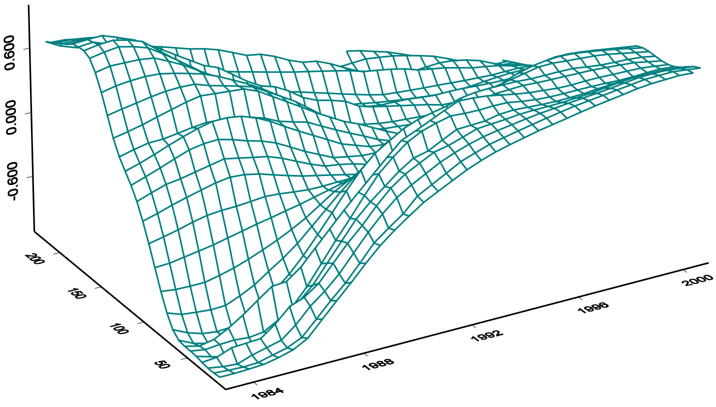


Figure 4: *Forest health data: Nonlinear interaction between calendar time and age of the tree. Shown are the posterior means.*

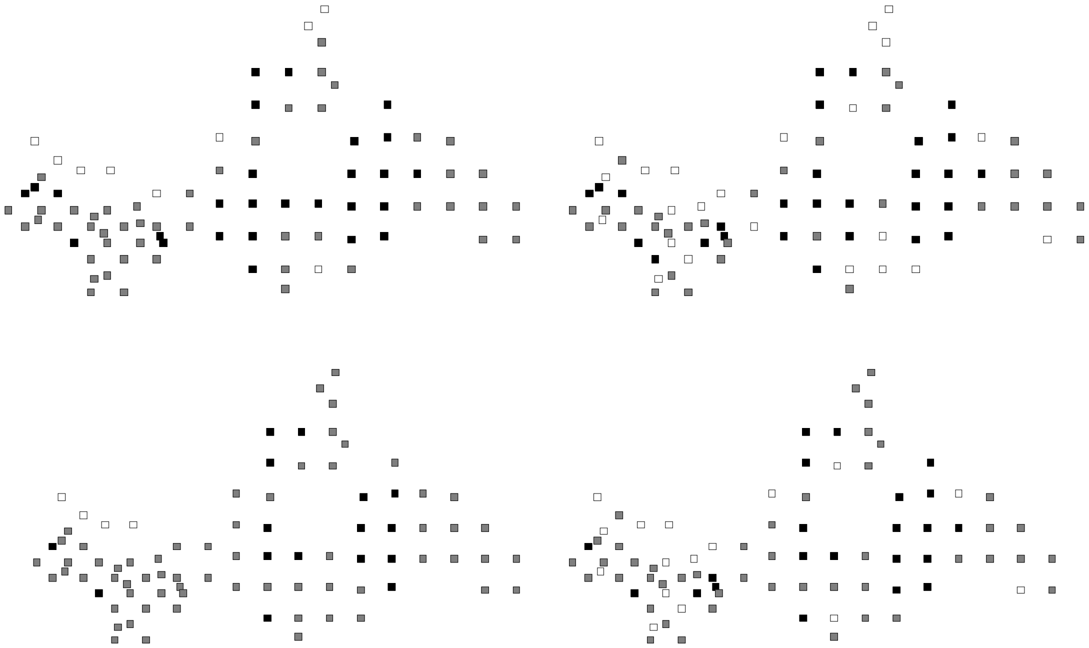


Figure 5: *Forest health data: Spatial effect for model 1 (left panels) and model 2 (right panels). Shown are posterior probabilities for a nominal level of 80% (top panels) and 95% (bottom panels). Black denotes locations with strictly negative credible intervals, white denotes locations with strictly positive credible intervals.*

4.2 Space-time analysis of health insurance data

In this section we analyse space-time data from a German private health insurance company. In a consulting case the main interest was on analysing the dependence of treatment costs on covariates with a special emphasis on modelling the spatio-temporal development. The data set contains individual observations for a sample of 13.000 males (with about 160.000 observations) and 1.200 females (with about 130.000 observations) in West Germany for the years 1991-1997. The variable of primary interest is the treatment cost C in hospitals. Except some categorical covariates characterizing the insured person we analysed the influence of the continuous covariates age (A) and calendar time (t) as well as the influence of the district (D) where the policy holder lives. We carried out separate analysis for men and women. We also distinguish between 3 types of health services, "accomodation", "treatment with operation" and "treatment without operation". In this demonstrating example, we present only results for males and "treatment with operation". Since the treatment costs are nonnegative and considerably skewed we assume that the costs for individual i at time t given covariates x_{it} are Gamma distributed, i.e. $C_{it}|x_{it} \sim Ga(\mu_{it}, \phi)$ where ϕ is a scale parameter and the mean μ_{it} is defined as

$$\mu_{it} = \exp(\eta_{it}) = \exp(\gamma_0 + f_1(t) + f_2(A_{it}) + f_3(D_{it})).$$

For the effects of age and calendar time we assumed cubic P-splines with 20 knots and a second order random walk penalty. To distinguish between spatially smooth and small scale regional effects, we further split up the spatial effect f_3 into a spatially structured and a unstructured effect, i.e.

$$f_3(D_{it}) = f_{str}(D_{it}) + f_{unstr}(D_{it})$$

For the unstructured effect f_{unstr} we assume i.i.d. Gaussian random effects. For the spatially structured effect we tested both a Markov random field prior and a two dimensional P-spline on a 12 by 12 knots grid.

The estimation of the scale parameter ϕ deserves special attention because MCMC inference is not trivial. In analogy to the variance parameter in Gaussian response models, we assume an inverse Gamma prior with hyperparameters a_ϕ and b_ϕ for ϕ, i.e. $\phi \sim IG(a_\phi, b_\phi)$. Using this prior the full conditional for ϕ is given by

$$p(\phi|\cdot) \left(\frac{1}{\Gamma(\phi)\phi^\phi}\right)^n \phi^{a_\phi - 1} \exp(-\phi b'_\phi)$$

with

$$b'_\phi = b_\phi + \sum_{i,t} (\log(\mu_{it}) - \log(C_{it}) + C_{it}/\mu_{it}).$$

This distribution ist not of standard form. Hence, the scale parameter must be updated by Metropolis-Hastings steps. We update ϕ by drawing a random number ϕ^p from an inverse Gamma proposal distribution with a variance s^2 and a mean equal to the current state of the chain ϕ^c. The variance s^2 is a tuning parameter and must be chosen appropriately to guarantee good mixing properties. We choose s^2 such that the acceptance rates are roughly between 30 and 60 percent.

Figure 6 shows the time trend f_1 (panel a) and the age effect f_2 (panel b). Shown are the posterior means together with 80% and 95% pointwise credible intervals. The effect for the year 1999 is future prediction explaining the growing uncertainty for the time effect in this year. Note also the large credible intervals of the age effect for individuals of age 90 and above. The reason are small sample sizes for these age groups. To gain more insight into the size of the effects, panels c) and d) display the marginal effects $f_j^{marginal}$ which are defined as $f_j^{marginal}(x_j) = \exp(\gamma_0 + f_j(x_j))$, i.e. the mean of treatment costs with the values of the remaining covariates fixed such that their effect is zero. The marginal effects (including credible intervals) can be easily estimated in a MCMC sampling scheme by computing (and storing) $f_j^{marginal}(x_j)$ in every iteration of the sampler from the current value of $f_j(x_j)$ and the intercept γ_0. Posterior inference is then based on the samples of $f_j^{marginal}(x_j)$. For the ease of interpretation, a straight line is included in the graphs indicating the marginal effect for $f_j = 0$, i.e. $\exp(\gamma_0) \approx 940DM$. Finally, panels e) and f) show the first derivatives of both effects (again including credible intervals). They may be computed by the usual formulas for derivatives of polynomial splines, see De Boor (1978).

Figure 7 displays the structured spatial effect f_{str} based on a Markov random field prior. The posterior mean of f_{str} can be found in panel a), the marginal effect is depicted in panel b). Panels c) and d) show posterior probabilities based on nominal levels of 80% and 95%. Note the large size of the spatial effect with a marginal effect ranging from 730-1200 DM. It is clear that it is of great interest for health insurance companies to detect regions with large deviations of treatment costs compared to the average. The unstructured spatial effect f_{unstr} is negligible compared to the structured effect and therefore omitted.

Figure 8 shows the respective estimates of f_{str} now based on two dimensional P-splines. The time trend and age effect for this model are almost identical to the effects displayed in Figure 6 and are therefore not displayed. The estimated effects are similar but smoother (as could have been expected) and therefore easier to interpret. However, in terms of the DIC the model based on the MRF prior is preferable.

5 Conclusions

This paper proposes semiparametric Bayesian inference for regression models with responses from an exponential family and with structured additive predictors. The paper can be seen as the last in a series of articles on Bayesian semiparametric regression based on smoothness priors, see Fahrmeir and Lang (2001a), Fahrmeir and Lang (2001b) and Lang and Brezger (2003). It particularly extends the methodology for Gaussian responses in Lang and Brezger (2003) to situations with fundamentally non-gaussian responses. Our approach allows estimation of nonlinear effects of continuous covariates and time scales as well as appropriate consideration of unobserved unit- or cluster specific as well as spatial heterogeneity. Many well known regression models from the literature appear to be special cases of our approach, e.g. dynamic models, generalized additive mixed models, varying coefficient models, geoadditive models or the famous and widely used BYM-model for disease mapping (Besag et al. (1991)). The proposed sampling schemes work well and automatically for the most common response distributions. Software is provided in the public domain package *BayesX*.

Our current research is mainly focused on model choice and variable selection. Presently, model choice is based primarily on pointwise credible intervals for regression parameters and the DIC. A first step for more sohisticated variable selection is to replace pointwise credible intervals by simultaneous probability statements as proposed by Besag et al. (1995) and more rececently by Knorr-Held (2003). For the future, we plan to develop

Bayesian inference techniques that allow estimation and model choice (to some extent) simultaneously.

Acknowledgement:
This research has been financially supported by grants from the German Science Foundation (DFG), Sonderforschungsbereich 386 "Statistical Analysis of Discrete Structures".

References

Albert, J. and Chib, S., 1993: Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88, 669-679.

Andrews, D.F. and Mallows, C.L., 1974: Scale mixtures of normal distributions. *Journal of the Royal Statistical Society B*, 36, 99-102.

Besag, J. E. Green, P. J. Higdon, D. and Mengersen, K., 1995: Bayesian Computation and Stochastic Systems (with Discussion). Statistical Science, 10, 3–66.

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

Besag, J., York, J. and Mollie, A., 1991: Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Biller, C., 2000: Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models. *Journal of Computational and Graphical Statistics*, 9, 122-140.

Biller, C. and Fahrmeir, L., 2001: Bayesian Varying-coefficient Models using Adaptive Regression Splines. *Statistical Modeling*, 2, 195-211.

De Boor, C., 1978: *A Practical Guide to Splines.* Spriner-Verlag, New York.

Denison, D.G.T., Mallick, B.K. and Smith, A.F.M., 1998: Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society B*, 60, 333-350.

Devroye, L., 1986: *Non-Uniform Random Variate Generation.* Springer-Verlag, New York.

Chen, M. H. and Dey, D. K., 2000: Bayesian Analysis for Correlated Ordinal Data Models. In: Dey, D. K., Ghosh, S. K. and Mallick, B. K., 2000: *Generalized linear models: A Bayesian perspective.* Marcel Dekker, New York.

Cleveland, W. and Grosse, E., 1991: Computational Methods for Local Regression. *Statistics and Computing*, 1991, 1, 47-62.

Currie, I. and Durban, M., 2002: Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, 4, 333-349.

Denuit, M. and Lang, S., 2003: Nonlife Ratemaking with Bayesian GAMs. *Preprint.*

Devroye, L., 1986: *Non-Uniform Random Variate Generation.* Springer-Verlag, New York.

Diggle, P., Liang, K.Y. and Zeger, S., 1994: Analysis of Longitudinal Data. London: Chapman and Hall.

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

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

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

Fahrmeir, L. and Lang, S., 2001: Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Annals of the Institute of Statistical Mathematics*, 53, 10-30

Fahrmeir, L., Lang, S.,Wolff, J. and Bender, S. (2003): Semiparametric Bayesian Time-Space Analysis of Unemployment Duration. Journal of the German Statistical Society, to appear.

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

Fan, J. and Gijbels, I., 1996: *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Friedman, J. H., 1991: Multivariate Adaptive Regression Splines (with discussion). *Annals of Statistics*, 19, 1–141.

Friedman, J. H. and Silverman, B. L., 1989: Flexible Parsimonious Smoothing and Additive Modeling (with discussion). *Technometrics*, 1989, 31, 3–39.

Gamerman, D., 1997: Efficient Sampling from the Posterior Distribution in Generalized Linear Models. *Statistics and Computing*, 7, 57–68.

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

Geweke, J. 1991: Efficient Simulation From the Multivariate Normal and Student-t Distribution Subject to Linear Constraints. In: *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface,* 571-578, Alexandria.

Göttlein , A. and Pruscha, H., 1996: Der Einfluß von Bestandskenngrößen, Topographie, Standord und Witterung auf die Entwicklung des Kronenzustandes im Bereich des Forstamtes Rothenbuch. Forstwissenschaftliches Centralblatt, 114, 146–162.

Green, P.J., 2001: A Primer in Markov Chain Monte Carlo. In: Barndorff-Nielsen, O.E., Cox, D.R. and Klüppelberg, C. (eds.), *Complex Stochastic Systems.* Chapmann and Hall, London, 1-62.

Hansen, M. H., Kooperberg, C., 2002: Spline Adaptation in Extended Linear Models. *Statistical Science*, 17, 2–51.

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

Hastie, T. and Tibshirani, R., 1993: Varying-coefficient Models. *Journal of the Royal Statistical Society B*, 55, 757-796.

Hastie, T. and Tibshirani, R., 2000: Bayesian Backfitting. *Statistical Science*, 15, 193–223.

Hastie, T., Tisbshirani, R. and Friedman, J., 2001: *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer–Verlag.

Holmes, C.C., and Knorr-Held, L., 2003: Efficient Simulation of Bayesian Logistic Regression Models. Discussion paper 306, SFB 386, Department of Statistics, University of Munich.

Kamman, E. E. and Wand, M. P., 2003: Geoadditive Models. *Journal of the Royal Statistical Society C (Applied Statistics)*, 52, 1-18.

Knorr-Held, L., 1996: Hierarchical Modelling of Discrete Longitudinal Data. Shaker Verlag.

Knorr-Held, L., 1999: Conditional Prior Proposals in Dynamic Models. *Scandinavian Journal of Statistics*, 26, 129-144.

Knorr-Held, L., 2003: Simultaneous posterior probability statements from Monte Carlo output. *Journal of Computational and Graphical Statistics*, to appear.

Knorr-Held, L. and Rue, H., 2002: On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29, 597-614.

Kohn, R., Smith, M.andChan, D., 2001: Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, 11, 313-322.

Lang, S. and Brezger, A., 2003: Bayesian P-splines. *Journal of Computational and Graphical Statistics*, to appear.

Lang, S., Fronk, E.-M. and Fahrmeir, L., 2002: Function estimation with locally adaptive dynamic models. *Computational Statistics*, 17, 479-500.

Lin, X. and Zhang, D., 1999: Inferene in generalized additive mixed models by using smoothing splines. *Journal auf the Royal Statistical Society B* , 61, 381–400.

Loader, C., 1997: Locfit: An Introduction. *Statistical Computing ang Graphics Newsletter*, 8(1), 11-17.

Marx, B.D. and Eilers, P.H.C., 1998: Direct Generalized Additive Modeling with Penalized Likelihood. Computational Statistics and Data Analysis, 28, 193–209.

Robert, C.P., 1995: Simulation of truncated normal variables. *Statistics and Computing*, 5, 121-125.

Rue, H., 2001: Fast Sampling of Gaussian Markov Random Fields with Applications. *Journal of the Royal Statistical Society B*, 63, 325-338.

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

Smith, M. and Kohn, R., 1996: Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75, 317-343.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A., 2002: Bayesian measures of model complexity and fit., *Journal of the Royal Statistical Society B*, 65, 583 - 639.

Stone, C.J., Hansen, M., Kooperberg, C. and Truong, Y.K., 1997: Polynomial Splines and their Tensor Products in Extended Linear Modeling (with discussion). Annals of Statistics, 25, 1371–1470.

Wand, M.P., 2003: Smoothing and mixed models, *Computational Statistics*, 18, 223-249.

Wang, Y., 1995: GRKPACK: Fitting Smoothing Spline ANOVA Models for Exponential Families. Technical Report No. 942, Department of Statistics, University of Wisconsin.

Wood, S.N., 2000: Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties. *Journal of the Royal Statistical Society B*, 62, 413-428.

Wood, S.N., 2003: Thin plate regression splines. *Journal of the Royal Statistical Society B*, 65, 95-114.

Wood, S.N., Kohn, R., Shively, T. and Jiang, W., 2002: Model selection in spline nonparametric regression. *Journal of the Royal Statistical Society B* 64, Part 1, 119-139.

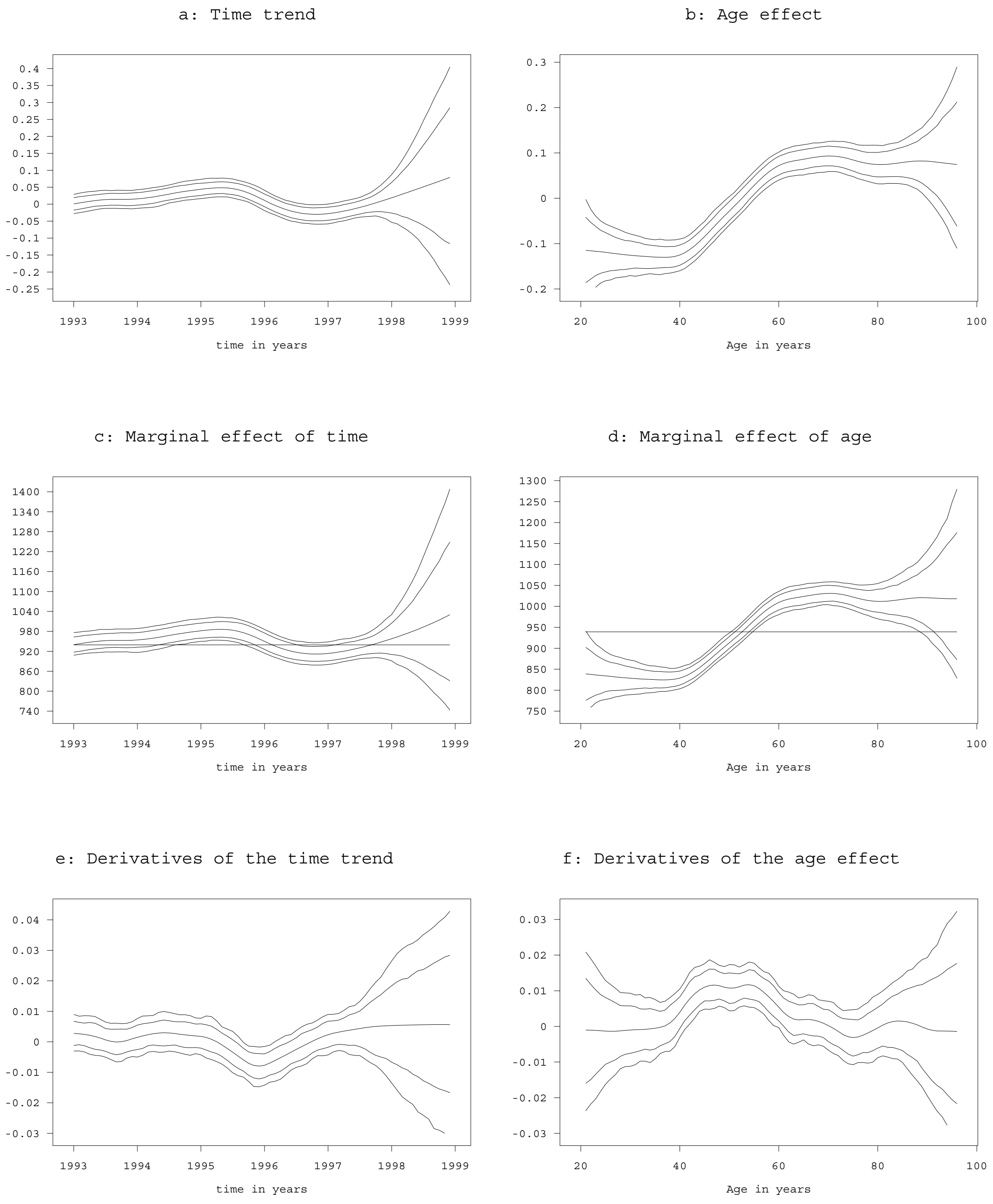
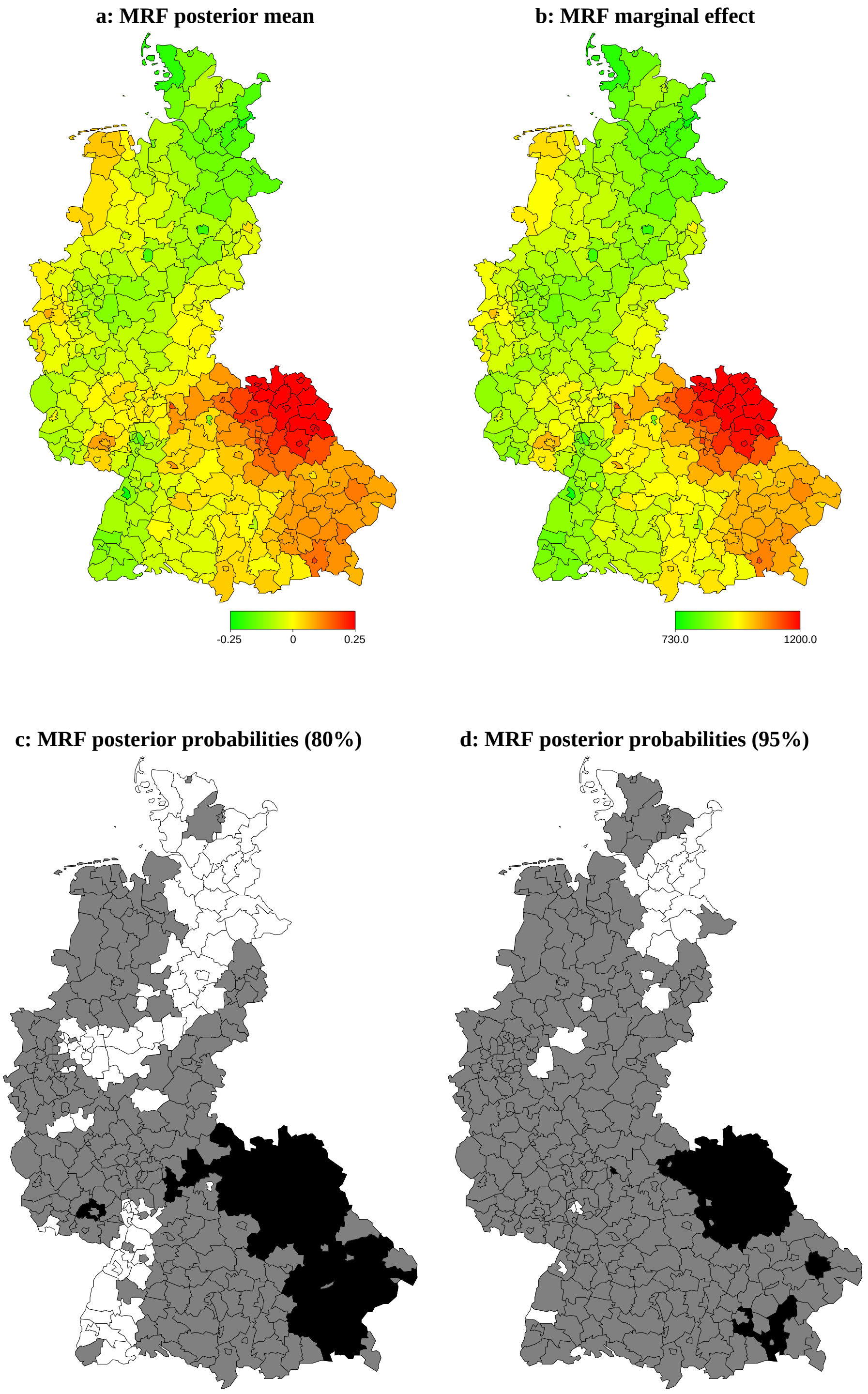


Figure 6: *Health insurance data: Time trend and age effect. Panels a) and b) show the estimated posterior means of functions f_1 and f_2 together with pointwise 80% and 95% pointwise credible intervals. Panels c) and d) depict the respective marginal effects and panels e) and f) the first derivatives f_1' and f_2'.*



a: MRF posterior mean

b: MRF marginal effect

c: MRF posterior probabilities (80%)

d: MRF posterior probabilities (95%)

Figure 7: *Health insurance data: Structured spatial effect* f_{str} *based on Markov random field priors. The posterior mean of* f_{str} *is shown in panel a) and the marginal effect in panel b). Panels c) and d) display posterior probabilities for nominal levels of 80% and 95%. Black denotes regions with strictly positive credible intervals and white regions with strictly negative credible intervals.*

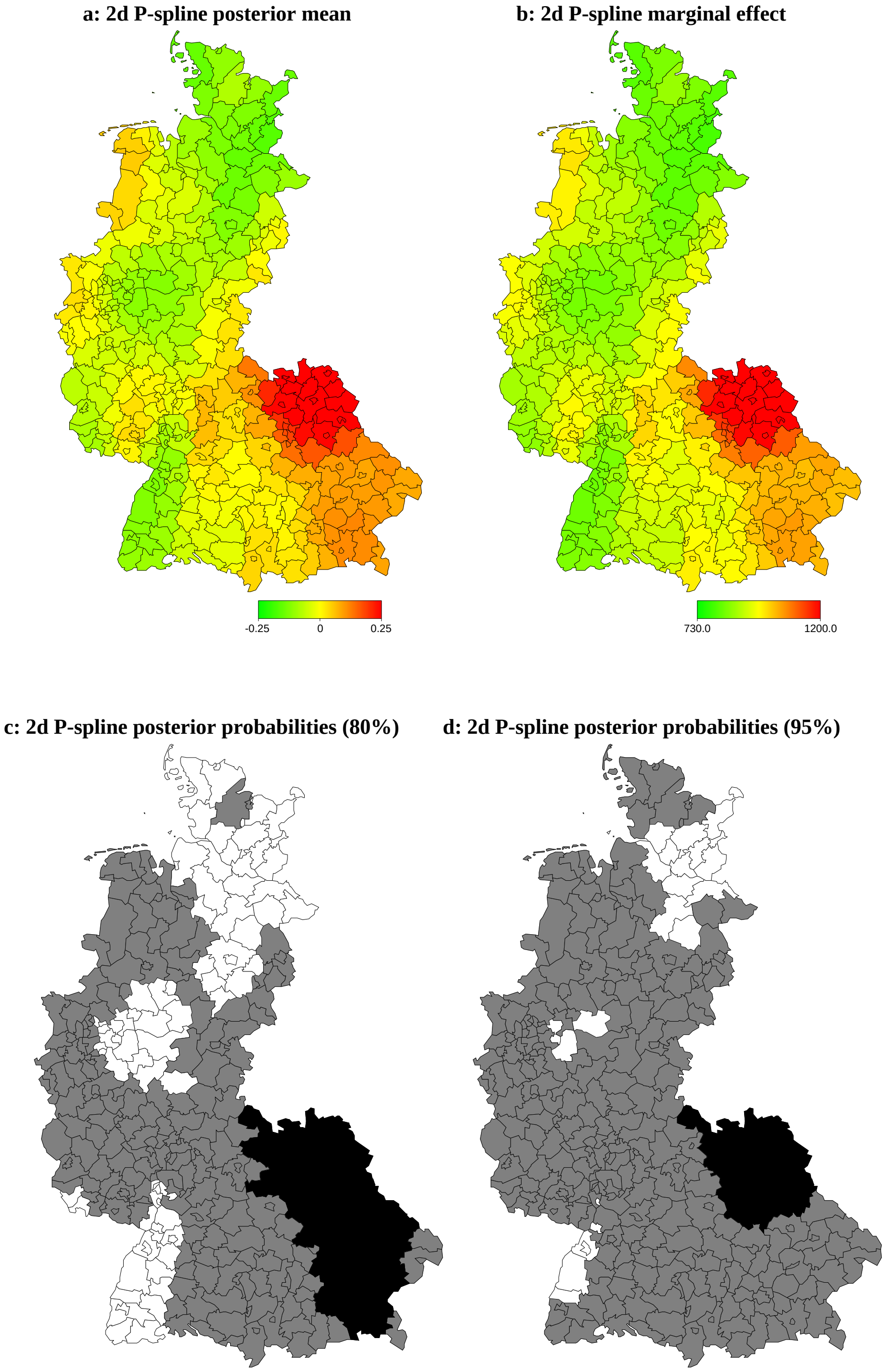

a: 2d P-spline posterior mean
-0.25
0
0.25
b: 2d P-spline marginal effect
730.0
1200.0
c: 2d P-spline posterior probabilities (80%)
d: 2d P-spline posterior probabilities (95%)

Figure 8: *Health insurance data: Structured spatial effect f_{str} based on two dimensional P-splines. The posterior mean of f_{str} is shown in panel a) and the marginal effect in panel b). Panels c) and d) display posterior probabilities for nominal levels of 80% and 95%. Black denotes regions with strictly positive credible intervals and white regions with strictly negative credible intervals.*