Appl. Statist. (2001)
50, *Part* 2, *pp.* 201–220

Bayesian inference for generalized additive mixed models based on Markov random field priors

Ludwig Fahrmeir and Stefan Lang

University of Munich, Germany

[Received August 1999. Final revision October 2000]

Summary. Most regression problems in practice require flexible semiparametric forms of the predictor for modelling the dependence of responses on covariates. Moreover, it is often necessary to add random effects accounting for overdispersion caused by unobserved heterogeneity or for correlation in longitudinal or spatial data. We present a unified approach for Bayesian inference via Markov chain Monte Carlo simulation in generalized additive and semiparametric mixed models. Different types of covariates, such as the usual covariates with fixed effects, metrical covariates with non-linear effects, unstructured random effects, trend and seasonal components in longitudinal data and spatial covariates, are all treated within the same general framework by assigning appropriate Markov random field priors with different forms and degrees of smoothness. We applied the approach in several case-studies and consulting cases, showing that the methods are also computationally feasible in problems with many covariates and large data sets. In this paper, we choose two typical applications.

Keywords: Generalized semiparametric mixed models; Markov chain Monte Carlo sampling; Random effects; Spatial and spatiotemporal data; State space and Markov random field models; Varying coefficients

1. Introduction

Our work has been motivated by various non-standard regression problems that resulted from consulting cases. In this paper, we focus on two applications with longitudinal and spatiotemporal structure of the data.

The first application is a regression analysis of longitudinal data on forest damage. The data are collected in a yearly visual inventory carried out in a forest district in the northern part of Bavaria. In our analysis, we shall focus on the state of damage of beeches. There are 80 observation points with the occurrence of beeches spread over the whole area. The dependent variable is the degree of defoliation on a binary indicator for the state of damage, with categories 'light or distinct damage' and 'no damage' assessed for each tree after inspection in year t. In addition, tree-specific covariates are measured. Some are categorical, like the age of the trees which is given in three age groups; others are metrical and time varying, like the acidity (pH value) of the soil or density of the canopy at the stand. A detailed description of the data can be found in Göttlein and Pruscha (1996). For a thorough analysis with modern regression techniques, a model is needed that simultaneously incorporates the possibly non-linear effects of the metrical covariates, possibly time-varying effects of categorical covariates and tree-specific random effects to account for unobserved heterogeneity or correlation.

Address for correspondence: Ludwig Fahrmeir, Institut für Statistik, Ludwig-Maximilians-Universität München, Ludwigstrasse 33, D-80539 München, Germany.
E-mail: fahrmeir@stat.uni-muenchen.de

 0035–9254/01/50201

The second application is an investigation of monthly unemployment data based on a large spatiotemporal sample from the German Federal Employment Office for the years 1980–1995. Here, the dependent variable is binary, indicating whether an individual is unemployed or employed in month t during this period. The covariates are the duration of current unemployment (in months), age (in years) and binary or categorical personal characteristics like nationality and financial support by unemployment benefits. In addition the district where the individual lives is given. Now a semiparametric binary regression model is needed that allows the simultaneous nonparametric modelling and estimation of non-linear effects of the metrical covariates calendar time, duration of unemployment and age as well as spatially correlated effects of districts or labour market regions.

We have been confronted with similar data structures and regression problems in other applications. In a regression analysis of rents paid for flats or appartments for a sample of more than 3000 households, the monthly net rent per square metre was chosen as the dependent variable. The covariates include the highly influential metrical variables floor space and age of the building, and a large number of binary indicators for equipment and quality of the flat. In addition, spatial information is given by the location of the flat in one of the 382 subquarters in Munich. The effect of location was modelled by spatially correlated random effects. In analyses of health and car insurance data, the dependence of the number and amount of insurance claims of individuals on calendar time, district and personal characteristics had to be investigated. In each of these applications, we found that adequate statistical inference was at best limited with the existing methodology and software.

As a common feature in all cases, we have metrical or spatially correlated covariates $x_1, \ldots, x_p$, say, with unknown, possibly non-linear, effects and a vector w of further covariates, whose influence on the predictor is assumed to be linear. In the forest damage problem, we shall add unstructured tree-specific random effects b to account for unobserved heterogeneity or correlation. Therefore an additive predictor of the form

$$\eta = f_1(x_1) + \ldots + f_p(x_p) + w'\beta + b \tag{1}$$

seems reasonable. Together with an exponential family observation model and a suitable link function, equation (1) defines a generalized additive mixed model (GAMM) or a semi-parametric mixed model. A further extension, that we consider in the forest damage study, leads to a varying-coefficient mixed model (VCMM).

In the duration of unemployment analysis we include district-specific spatially correlated random effects, to account for spatial heterogeneity that is not explained by other covariates. From a classical perspective, spatial effects are considered as correlated random effects with an appropriate prior reflecting neighbourhood relationships. For our Bayesian approach, it is more useful to consider them as spatially correlated effects $f_j(x_j)$ of a spatial covariate x_j, and to incorporate them formally as a component of the nonparametric additive terms in equation (1). This will lead to a unified treatment of metrical covariates and spatially correlated random effects or, in other words, spatial covariates. Nevertheless, we shall call such models GAMMs or VCMMs. Inference for these models with existing methodology has some drawbacks and limitations, in particular for non-Gaussian responses. For GAMMs, Lin and Zhang (1999) proposed approximate inference using smoothing splines and double penalized quasi-likelihood, extending previous work of Breslow and Clayton (1993) and Breslow and Lin (1995) for generalized linear mixed models (GLMMs). As they pointed out in the discussion, similarly to approximate inference in GLMMs, there are bias problems, especially with binary data or correlated random effects, so Markov chain Monte Carlo

(MCMC) methods may be an attractive alternative. S-PLUS has an option for exploratory data analysis in GAMMs with splines or LOESS functions, but there is no possibility for estimating variance components jointly with smoothing parameters.

In this paper we propose a general Bayesian approach via MCMC sampling for inference in generalized additive and varying-coefficients models, including mixed models with structured or unstructured random effects. As an important feature, all different types of covariates, or effects, are considered from a unified viewpoint. The usual covariates with fixed effects, metrical covariates, such as timescales with nonparametric trend or seasonal effects, unstructured random effects and spatial covariates, are treated within the same general framework by assigning Markov random field smoothness priors with a common structure but different degrees of smoothness to corresponding effects. A data-driven choice of smoothing parameters is automatically included. Although models with Gaussian responses are also covered by the framework, our main interest lies in models for fundamentally non-Gaussian responses, such as binary or other discrete-valued responses. The data examples in the applications section and experience with other applications mentioned before show the practical feasibility in situations with many covariates and with large data sets. The MCMC procedure provides samples from all posteriors of interest and permits the estimation of posterior means, medians, quantiles confidence bands and predictive distributions. No approximations based on conjectures of asymptotic normality have to be made, and no 'plug-in' procedures are needed.

For Gaussian responses, Gibbs sampling can be used for fully Bayesian analysis based on smoothness priors; see for example Wong and Kohn (1996), who used state space or dynamic model representations of splines in additive models without any inclusion of unstructured or spatial random effects, or Hastie and Tibshirani (2000), who derive the Gibbs sampler as a Bayesian version of back-fitting. Bayesian basis function approaches with regression splines or a more general class of piecewise polynomials have been proposed by Smith and Kohn (1996) and Denison *et al.* (1998), again without random effects. A more direct approach, that we use in our implementation, makes use of banded posterior precision matrices; see Rue (2000) or Lang and Brezger (2000).

For fundamentally non-Gaussian models, more general MCMC techniques than Gibbs sampling are needed, and there is still a lack of methods and of practical experience with existing suggestions. Hastie and Tibshirani (2000) sketch a Metropolis–Hastings-type algorithm for generalized additive models but do not give any examples or statements about performance. Mallick *et al.* (2000) propose a Bayesian multivariate adaptive regression splines method for generalized linear models as an extension of the Bayesian curve fitting method of Denison *et al.* (1998), but, as they state, the sampler has slow convergence. The more recent reversible jump MCMC algorithm developed by Biller (2000) for adaptive regression splines in semiparametric generalized linear models seems to have better convergence properties, and an extension to GAMMs and VCMMs might be promising.

Our approach is based on Gaussian smoothness priors derived from state space and Markov random field models. With non-Gaussian observation models as considered in this paper, Gibbs sampling cannot be applied for sampling from high dimensional posteriors. Therefore, we incorporate the Metropolis–Hastings algorithm with conditional prior proposals, developed by Knorr-Held (1999) in the context of dynamic generalized linear models, for drawing block move samples from posteriors of non-linear effects $f_j(x_j)$ of metrical covariates x_j. For spatial covariates with Markov random field priors, this block move sampler is extended in a computationally efficient way for drawing from posteriors of spatial effects.

The rest of the paper is organized as follows: Bayesian semiparametric mixed models are described in Section 2, whereas Section 3 contains details about the chosen MCMC techniques. In Section 4 we demonstrate the practical feasibility of our approach through the data applications mentioned at the beginning. All computations are carried out with BayesX, a program for Bayesian inference with MCMC simulation techniques (Lang and Brezger, 2000). The software contains the approach described here and is available from

`http://www.stat.uni-muenchen.de/~lang/`

2. Bayesian semiparametric mixed models

2.1. Observation model

Consider now regression situations, where observations $(y_i, x_{i1}, \ldots, x_{ip}, w_i)$, $i = 1, \ldots, n$, on a response y, a vector $x = (x_1, \ldots, x_p)$ of metrical or spatial covariates and a vector w of further covariates are given.

In longitudinal studies, as in our applications to forest damage or to the duration of unemployment in Section 4, the covariate vector will typically include one or more timescales, such as duration and calendar time. In some studies, as in the study on unemployment, a spatial covariate, such as the district in which the unemployed live, may be considered and appropriately incorporated into the model.

Generalized additive and semiparametric models (Hastie and Tibshirani, 1990) assume that, given $x_i = (x_{i1}, \ldots, x_{ip})$, and w_i, the distribution of y_i belongs to an exponential family, with mean $\mu_i = E(y_i|x_i, w_i)$ linked to an additive semiparametric predictor η_i by

$$\begin{aligned} \mu_i &= h(\eta_i), \\ \eta_i &= f_1(x_{i1}) + \ldots + f_p(x_{ip}) + w_i'\beta. \end{aligned} \tag{2}$$

Here h is a known link or response function, and $f_1, \ldots, f_p$ are unknown smooth functions of the covariates. Note that spatially correlated (random) effects of a spatial covariate x_j are included as nonparametric component $f_j(x_j)$. For identifiability reasons, unknown functions are centred appropriately. Observation models of the form (2) may be inappropriate if heterogeneity between units is not sufficiently described by covariates or for correlated data in a longitudinal study. A common way to deal with this problem is the inclusion of additive random effects in the predictor. This leads to GAMMs with a predictor of the form

$$\eta_i = f_1(x_{i1}) + \ldots + f_p(x_{ip}) + w_i'\beta + b_{g_i}, \tag{3}$$

where b_{g_i} is a unit- or group-specific random effect, with $b_{g_i} = b_g$ if unit i is in group g, $g = 1, \ldots, G$. For example, in our analysis of the forest damage data, b_g is an additional tree-specific effect to account for correlation and unobserved heterogeneity. Owing to the large number of trees a fixed effect approach will not be feasible, and a random-effects model is chosen instead. A further extension leads to VCMMs

$$\eta_i = f_1(x_{i1})z_{i1} + \ldots + f_p(x_{ip})z_{ip} + w_i'\beta + b_{g_i}; \tag{4}$$

see Hastie and Tibshirani (1993) for the case without random effects. The design vector $z = (z_1, \ldots, z_p)$ may contain components of x or w as well as some additional covariates. If a design variable is identically 1, e.g. $z_j = 1$, then the corresponding function f_j is the main effect of x_j, whereas terms like $f_j(x_{ij})z_{ij}$ model an effect of z_j that varies over x_j or, in other words, interaction between x_j and z_j. We shall use a VCMM in the forest damage study, with time-varying effects and a tree-specific random effect.

2.2. Prior distributions

For Bayesian semiparametric inference, the unknown functions $f_1, \ldots, f_p$, or more exactly corresponding vectors of function evaluations, the parameters $\beta = (\beta_1, \ldots, \beta_r)$ and the random effects $b = (b(1), \ldots, b(G))$ are all considered as random variables. The observation models (2), (3) or (4) are understood to be conditional on these random variables and must be supplemented by appropriate prior distributions.

Priors for the unknown functions $f_1, \ldots, f_p$ depend on the type of the covariates and on prior beliefs about the smoothness of f_j. Priors for timescales and metrical covariates are based on Gaussian smoothness priors that are common in dynamic generalized linear models; see, for example, Fahrmeir and Tutz (1997), chapter 8. For spatial covariates, priors are based on (Gaussian) Markov random fields; see for example Besag (1974), Besag *et al.* (1991) or Besag and Kooperberg (1995).

2.2.1. *Metrical covariates and timescales*

Let us first consider the case of a metrical covariate x with *equally spaced observations* x_i, $i = 1, \ldots, m$, $m \leqslant n$. Then the ordered sequence $x_{(1)} < \ldots < x_{(t)} < \ldots < x_{(m)}$ defines an equidistant grid on the x-axis. The typical case for this situation arises if the covariate x corresponds to time t, and the grid points correspond to time units such as weeks, months or years, but generally x can be any ordered dimension. Define $f(t) := f(x_{(t)})$ and let

$$f = (f(1), \ldots, f(t), \ldots, f(m))'$$

denote the vector of function evaluations. Then, just as for the time trends example in Section 1, common priors for smooth functions are first- or second-order random walk models

$$f(t) = f(t-1) + u(t) \tag{5a}$$

or

$$f(t) = 2f(t-1) - f(t-2) + u(t) \tag{5b}$$

with Gaussian errors $u(t) \sim N(0; \tau^2)$ and diffuse priors $f(1) \propto$ constant and $f(1)$ and $f(2) \propto$ constant, for initial values, respectively. Both specifications act as smoothness priors that penalize too rough functions f. A first-order random walk penalizes abrupt jumps $f(t) - f(t-1)$ between successive states and a second-order random walk penalizes deviations from the linear trend $2f(t-1) - f(t-2)$. Note that a second-order random walk is derived by computing second differences, i.e. the differences of neighbouring first-order differences. The difference in practice between the two specifications is that estimated functions tend to be somewhat smoother for second-order random walk priors. Of course, higher order difference priors are also possible. For example if the covariate x is time t, measured in months, then a common smoothness prior for a seasonal component $f(t)$ is

$$f(t) + f(t-1) + \ldots + f(t-11) = u(t) \sim N(0, \tau^2). \tag{6}$$

Next we consider the general case with non-equally spaced observations. Let

$$x_{(1)} < \ldots < x_{(t)} < \ldots < x_{(m)}$$

denote the $m \leqslant n$ strictly ordered different observations of the covariate x, and

$$f = (f(1), \ldots, f(t), \ldots, f(m))'$$

with $f(t) := f(x_{(t)})$, the vector of function evaluations.

Random walk or autoregressive priors must be modified to account for non-equal distances $\delta_t = x_{(t)} - x_{(t-1)}$ between observations. Random walks of first order are now specified by

$$\begin{aligned} f(t) &= f(t-1) + u(t), \\ u(t) &\sim N(0;\ \delta_t \tau^2), \end{aligned} \tag{7}$$

i.e. by adjusting error variances from τ^2 to $\delta_t \tau^2$. Random walks of second order are

$$f(t) = \left(1 + \frac{\delta_t}{\delta_{t-1}}\right) f(t-1) - \frac{\delta_t}{\delta_{t-1}} f(t-2) + u(t), \tag{8}$$

$u(t) \sim N(0;\ w_t \tau^2)$, where w_t is an appropriate weight. Several possibilities are conceivable for the weights. The simplest is $w_t = \delta_t$ as for a first-order random walk. Another choice is

$$w_t = \delta_t \left(1 + \frac{\delta_t}{\delta_{t-1}}\right),$$

which also takes into account the distance δ_{t-1}. It can be derived from the difference

$$\frac{f(t) - f(t-1)}{\delta_t} - \frac{f(t-1) - f(t-2)}{\delta_t - 1}$$

of first-order differences and by treating them as if they were independent. A related, yet different, proposal for a second-order autoregressive prior is given by Berzuini and Larizza (1996). Another possibility would be to use state space representations of stochastic differential equation priors based on the work by Kohn and Ansley (1987). Biller and Fahrmeir (1997) followed this idea, but significant problems are associated with the convergence and mixing behaviour of posterior samples, compared with the priors chosen here.

Owing to the Markovian specification by random walks or other autoregressive models, priors for the vector f of function evaluations are seemingly defined in an asymmetric directed way. However, these priors can always be rewritten in an undirected symmetric form. This follows from the fact that any discrete Markov process can be formulated in the undirected form of a Markov random field by conditioning not only on previous variables $f(t-1), f(t-2)$, etc. but also on future variables $f(t+1), f(t+2)$, etc. This also becomes evident from the joint Gaussian prior for the entire vector f: for each of the priors (7), (8) or (6) f has a partially improper Gaussian prior $f \sim N(0;\ \tau^2 K^-)$, where K^- is a generalized inverse of a band diagonal precision or penalty matrix K. For example, it is easy to see that for a random walk of first order (7) the penalty matrix is

$$K = \begin{pmatrix} \delta_2^{-1} & -\delta_2^{-1} & & & & & \\ -\delta_2^{-1} & \delta_2^{-1} + \delta_3^{-1} & -\delta_3^{-1} & & & & \\ & -\delta_3^{-1} & \delta_3^{-1} + \delta_4^{-1} & -\delta_4^{-1} & & & \\ & & \ddots & & \ddots & & \\ & & -\delta_{m-2}^{-1} & \delta_{m-2}^{-1} + \delta_{m-1}^{-1} & -\delta_{m-1}^{-1} & & \\ & & & -\delta_{m-1}^{-1} & \delta_{m-1}^{-1} + \delta_m^{-1} & -\delta_m^{-1} \\ & & & & -\delta_m^{-1} & \delta_m^{-1} \end{pmatrix},$$

which reduces to

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

for equidistant x-values.

2.2.2. *Spatial covariates*

Let us now turn our attention to a spatial covariate x, where the values of x represent the location or site in connected geographical regions. For example in the unemployment study x represents the district in which the unemployed have their home. A common way to deal with spatial covariates is to assume that neighbouring sites are more alike than two arbitrary sites. Thus for a valid prior definition a set of neighbours for each site x_t must be defined. For geographical data as considered in this paper we usually assume that two sites x_t and x_j are neighbours if they share a common boundary. However, more sophisticated definitions of neighbourhoods are possible; see for example Besag *et al.* (1991). We assume the following spatial smoothness prior for the function evaluations $f(t)$, $t = 1, \ldots, m$, of the m different sites x_t:

$$f(t)|f(j)\; j \neq t,\, \tau^2 \sim N\left\{ \sum_{j \in \partial_j} f(j)/N_t,\, \tau^2/N_t \right\}, \tag{9}$$

where N_t is the number of adjacent sites and $j \in \partial_t$ denotes that site x_j is a neighbour of site x_t. Thus the (conditional) mean of $f(t)$ is an unweighted average of function evaluations of neighbouring sites. For spatial data, conditioning is undirected since there is no natural ordering of different sites x_t as is the case for timescales or metrical covariates.

A more general prior including expression (9) as a special case is given by

$$f(t)|f(j)\; j \neq t,\, \tau^2 \sim N\left\{ \sum_{j \in \partial_t} w_{tj}/w_{t+} f(j),\, 1/w_{t+}\tau^2 \right\}, \tag{10}$$

where w_{tj} are known (not necessarily) equal weights and the plus sign denotes summation over the missing subscript.

Such a prior is called a Gaussian intrinsic autoregression; see Besag *et al.* (1991) and Besag and Kooperberg (1995). Prior (9) is obtained as a special case by specifying $w_{tj} = 1$, resulting in equal weights for each neighbour of site x_t. Unequal weights are for example based on the common boundary length of neighbouring sites or the distance of the centroids of two sites; see Besag *et al.* (1991) for details. However, the applications in this paper are restricted to prior (9) based on adjacency weights.

As in the case of the autoregressive priors (7), (8) or (6), definition (10) may equivalently be written in terms of a penalty matrix K, i.e.

$$f|\tau^2 \propto \exp\left(-\frac{1}{2\tau^2} f' K f \right) \tag{11}$$

where the elements of K are given by

$$k_{tt} = w_{t+}$$

and

$$k_{tj} = \begin{cases} -w_{tj} & j \in \partial_t, \\ 0 & \text{otherwise}. \end{cases}$$

Usually this prior is improper since K is rank deficient and thus not invertible.

The close formal similarity of our priors for the function f of a metrical or a spatial covariate allows a unified MCMC algorithm that depends essentially on the penalty matrix K and is virtually independent of the type of covariate and definition of smoothness. This is described in detail in the next section.

2.2.3. *Further prior assumptions*

For a fully Bayesian analysis, hyperpriors for variances are introduced in a further stage of the hierarchy. This allows for a simultaneous estimation of the unknown function and the amount of smoothness. Common choices are highly dispersed inverse gamma priors

$$p(\tau^2) \sim \mathrm{IG}(a;\, c).$$

A possible choice for a and c is very small $a = c$, e.g. $a = c = 0.0001$, leading to almost diffuse priors for the variance parameters. An alternative proposed, for example, in Besag *et al.* (1995) is $a = 1$ and a small value for c, such as $c = 0.005$. The choice of such a highly dispersed but proper prior avoids problems arising with improper priors. Such problems are discussed in Hobart and Casella (1996) for linear mixed models. However, since estimation results tend to be sensitive to the choice of hyperpriors, especially in situations when data are sparse, some kind of sensitivity analysis should always be performed.

For the fixed effect parameters $\beta_1, \ldots, \beta_r$, we shall assume independent diffuse priors $p(\beta_j) \propto$ constant, $j = 1, \ldots, r$. Another choice would be highly dispersed Gaussian priors.

For random effects, we make the usual assumption that the b_g are independent and identically distributed Gaussian,

$$b_g|\upsilon^2 \sim N(0,\, \upsilon^2), \qquad g = 1, \ldots, G,$$

and define again a highly dispersed hyperprior for υ^2. In what follows, let

$$f = (f_1, \ldots, f_p), \qquad \tau^2 = (\tau_1^2, \ldots, \tau_p^2), \qquad \beta = (\beta_1, \ldots, \beta_r), \qquad b = (b_1, \ldots, b_G)$$

denote parameter vectors for function evaluations, variances and fixed and random effects. Then the Bayesian model specification is completed by the following *conditional independence assumptions*.

(a) For given covariates and parameters f, β and b, observations y_i are conditionally independent.
(b) Priors $p(f_j|\tau_j^2)$, $j = 1, \ldots, p$, are conditionally independent.
(c) Priors for fixed and random effects, and hyperpriors τ_j^2, $j = 1, \ldots, p$, are mutually independent.

3. Markov chain Monte Carlo inference

Full Bayesian inference is based on the entire posterior distribution

$$p(f, \tau^2, \beta, b|y) \propto p(y|f, \tau^2, \beta, b)\, p(f, \tau^2, \beta, b).$$

By assumption (a), the conditional distribution of observed data y is the product of individual likelihoods:

$$p(y|f, \tau^2, \beta, b) = \prod_{i=1}^{n} L_i(y_i; \eta_i), \tag{12}$$

with $L_i(y_i; \eta_i)$ determined by the specific exponential family distribution and the form chosen for the predictor η.

Together with the conditional independence assumptions (b) and (c), we have

$$p(f, \tau^2, \beta, b|y) \propto \prod_{i=1}^{n} L_i(y_i; \eta_i) \prod_{j=1}^{p} \{p(f_j|\tau_j^2)\, p(\tau_j^2)\} \prod_{k=1}^{r} p(\beta_k) \prod_{g=1}^{G} p(b_g|v^2)\, p(v^2)$$

for the posterior.

Bayesian inference via MCMC simulation is based on updating full conditionals of single parameters or blocks of parameters, given the rest and the data. Single-move steps, as in Carlin *et al.* (1992), which update each parameter $f(t)$ separately, suffer from problems with convergence and mixing. For Gaussian models, Gibbs sampling with so-called multimove steps can be applied; see, for example, Carter and Kohn (1994) and Wong and Kohn (1996). For non-Gaussian responses Gibbs sampling is no longer feasible and more general Metropolis–Hastings algorithms are needed. We adopt and extend a computationally very efficient Metropolis–Hastings algorithm with conditional prior proposals developed recently by Knorr-Held (1999) for dynamic generalized linear models. Convergence and mixing are considerably improved by block moves, where blocks $f[r, s] = (f(r), \ldots, f(s))$ of parameters are updated instead of single parameters $f(t)$. Suppressing conditioning parameters and data notationally, the full conditionals for the blocks $f[r, s]$ are

$$p(f[r, s]|\cdot) \propto L(f[r, s])\, p\{f[r, s]|f(l), l \notin [r, s], \tau^2\}.$$

The first factor $L(f[r, s])$ is the product of all likelihood contributions in equation (12) that depend on $f[r, s]$. The second factor, the conditional distribution of $f[r, s]$ given the rest $f(l)$, $l \notin [r, s]$, is a multivariate Gaussian distribution. Its mean and covariance matrix can be written in terms of the precision matrix K of f. Let $K[r, s]$ denote the submatrix of K, given by the rows and columns numbered from r to s, and let $K[1, r-1]$ and $K[s+1, m]$ denote the matrices to the left and right of $K[r, s]$. Then the (conditional) mean $\mu[r, s]$ and the covariance matrix $\Sigma[r, s]$ are given by

$$\mu[r, s] = \tau^2 \begin{cases} -K[r, s]^{-1}\, K[s+1, m]\, f[s+1, m] & r = 1, \\ -K[r, s]^{-1}\, K[1, r-1]\, f[1, r-1] & s = m, \\ -K[r, s]^{-1}\, (K[1, r-1]\, f[1, r-1] + K[s+1, m]\, f[s+1, m]) & \text{otherwise} \end{cases}$$

and

$$\Sigma[r, s] = \tau^2\, K[r, s]^{-1}$$

respectively. This result may be obtained by applying the usual formulae for conditional Gaussian distributions. Metropolis–Hastings block move updates for $f[r, s]$ are obtained by drawing a conditional prior proposal $f^*[r, s]$ from the conditional Gaussian $N(\mu[r, s], \Sigma[r, s])$ and accepting it with probability

$$\min\left\{1, \frac{L(f^*[r, s])}{L(f[r, s])}\right\}.$$

A fast implementation of MCMC updates requires efficient computing of the mean $\mu[r, s]$. For that reason we define the matrices

$$K[r, s]^l = -K[r, s]^{-1} K[1, r-1]$$

and

$$K[r, s]^r = -K[r, s]^{-1} K[s+1, m].$$

Then the conditional mean may be rewritten as

$$\mu[r, s] = \tau^2 \begin{cases} K[r, s]^r f[s+1, m] & r = 1, \\ K[r, s]^l f[1, r-1] & s = m, \\ K[r, s]^l f[1, r-1] + K[r, s]^r f[s+1, m] & \text{otherwise.} \end{cases}$$

Knorr-Held's (1999) conditional prior proposal refers to autoregressive priors in dynamic models, where the precision matrix K has non-zero diagonal bands, implying sparse structures for $K[r, s]^r$ and $K[r, s]^l$ where most elements are 0. For example for a first-order random walk only the first column of $K[r, s]^r$ and the last column of $K[r, s]^l$ contain non-zero elements. For spatial covariates, the precision matrix is no longer band diagonal but it is still sparse, with non-zero entries reflecting neighbourhood relationships, so sparse matrix operations can be used to improve computational efficiency. Now computing of $\mu[r, s]$ and $\Sigma[r, s]^{1/2}$ required for drawing the conditional prior proposal $f^*[r, s]$ is as follows: for every block $[r, s]$ we compute and store the matrices $K[r, s]^{-1/2}$, $K[r, s]^l$ and $K[r, s]^r$ in advance, where the last two are stored as sparse matrices. In every iteration of our MCMC algorithm, computing $\mu[r, s]$ and $\Sigma[r, s]^{-1/2}$ requires at most two (sparse) matrix multiplications with a column vector and multiplication of the resulting matrices with the scalar τ^2.

From a computational point of view, another main advantage is the simple form of the acceptance probability. Only the likelihood must be computed; no first or second derivatives etc. are involved, thus considerably reducing the number of calculations.

An important point, besides computational aspects, is the choice of block sizes. In our experience, the best results, in terms of mixing and autocorrelations of sampled parameters, are obtained with block sizes between 1 and 40. This corresponds to acceptance rates between 30% and 80%. In these cases it is typical for autocorrelations to become negligible for lags greater than 25–50. However, in some exceptional cases, autocorrelations for some parameters persist until lag 100. Consequently, our final runs for MCMC simulation contain between 27 000 and 105 000 iterations, with a burn-in period between 2000 and 5000 iterations.

To avoid correlation problems between successive blocks, several strategies are possible; see for example Knorr-Held (1999). In our implementation a minimum and maximum block size are specified and then the block size is chosen randomly in every iteration of the MCMC simulation. To obtain an impression of how our blocking strategy works, Fig. 1 shows the sampling paths for one of the parameters of the forest damage study, which will be discussed in more detail in Section 4.1. Fig. 1(a) shows the sampling path for iterations 2001–5000 for a single-move updating strategy (corresponding to block size 1). Fig. 1(b) shows the sampling path for the same parameter but with block sizes chosen randomly between 20 and 35. It is clear that the single-move strategy shows extremely poor mixing of the Markov chain—obviously the algorithm did not even converge. In contrast, Fig. 1(b) shows that the blocking strategy has improved the mixing of the Markov chain dramatically.

The full conditional for a variance parameter τ^2 is still an inverse gamma distribution

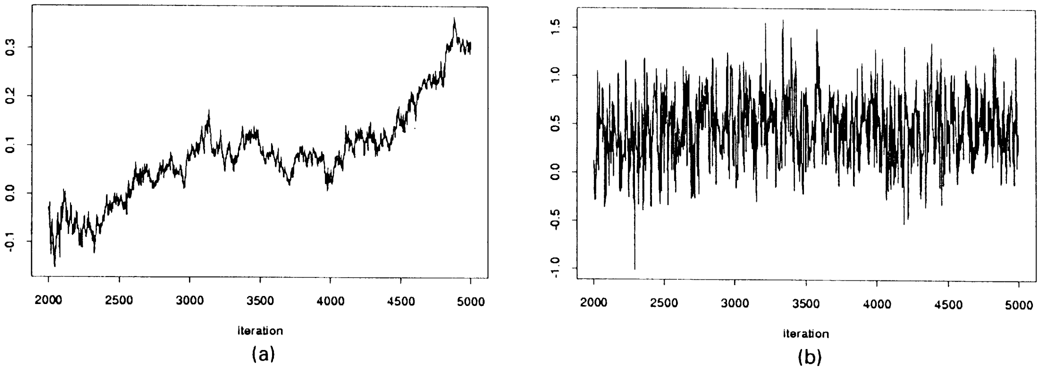


Fig. 1. Sampling path for one parameter of the forest damage study with different blocking strategies: (a) block size 1; (b) block sizes 20–35

$$p(\tau^2|\cdot) \propto \mathrm{IG}(a', c'). \tag{13}$$

Its parameters a' and c' again can be expressed in terms of the penalty matrix K and are given by

$$a' = a + \tfrac{1}{2}\mathrm{rank}(K)$$

and

$$c' = c + \tfrac{1}{2} f' K f.$$

Thus the updating of variance parameters can be done by simple Gibbs steps, drawing directly from the inverse gamma densities (13). Once again a fast implementation requires sparse matrix multiplications of $f'Kf$.

With a diffuse prior $p(\beta_j) \propto$ constant for the fixed effects parameters, the full conditional for β is

$$p(\beta|\cdot) \propto \prod_{i=1}^{n} L_i(y_i; \eta_i).$$

Updating of β can in principle be done by Metropolis–Hastings steps with a random walk proposal $q(\beta, \beta^*)$, but a serious problem is tuning, i.e. specifying a suitable covariance matrix for the proposal that guarantees high acceptance rates and good mixing. Especially when the dimension of β is high, with significant correlations between components, tuning interactively by trial and error is no longer feasible. An alternative is the weighted least squares proposal suggested by Gamerman (1997). Here a Gaussian proposal is used with mean $m(\beta)$ and covariance matrix $C(\beta)$, where β is the current state of the chain. The mean $m(\beta)$ is obtained by making one Fisher scoring step to maximize the full conditional $p(\beta|\cdot)$ and $C(\beta)$ is the inverse of the expected Fisher information, evaluated at the current state β of the chain. In this case the acceptance probability of a proposed new vector β^* is

$$\min\left\{1, \frac{p(\beta^*|\cdot)\, q(\beta^*, \beta)}{p(\beta|\cdot)\, q(\beta, \beta^*)}\right\}. \tag{14}$$

q is not symmetric because the covariance matrix C of q depends on β. Some computer time can be saved by omitting the Fisher scoring step when computing the mean of Gamerman's proposal, and simply taking the current state of the chain as the mean.

For an additional random intercept, the full conditional for parameter b_g is given by

$$p(b_g|\cdot) \propto \prod_{i \in \{j: g_j = g\}} L_i(y_i; \eta_i)\, p(b_g|\upsilon^2).$$

Here a simple Gaussian random walk proposal with mean b_g and variance υ^2 works well in most cases. To improve mixing, tuning is sometimes required by multiplying the prior variance υ^2 in the proposal with a constant factor, which is obtained by preliminary runs. An alternative is, again, Gamerman's weighted least squares proposal or a slight modification. This becomes especially attractive when the observation model contains one or more random slope parameters in addition to the random intercept. By analogy with the variance parameters τ_j^2 of nonparametric terms the full conditional of υ^2 is again an inverse gamma distribution, so updating is straightforward.

4. Applications

We consider two applications. In the first application on the state of forest damage of trees, we use a VCMM with unstructured tree-specific random effects. In our second application on the duration of unemployment, we apply a GAMM with spatial random effects for the district in which the unemployed have their home. All computations were carried out with BayesX.

To assess the sensitivity of the estimated functions to the hyperparameters a and c of the prior for the variance parameter, in both applications models were (re-)estimated with various choices of a and c. We tested four different choices for the hyperparameters, i.e. $a = c = 0.01$, $a = c = 0.0001$, $a = 1$, $c = 0.01$ and $a = 1$, $c = 0.005$. However, since the results differ only for one effect in our first application they are presented only for our standard choice for the hyperparameters, i.e. $a = 1$ and $c = 0.005$. We give a comparison of relative changes of estimated functions with respect to the different hyperparameters only for the forest damage application.

4.1. Forest damage

In this longitudinal study, we analyse the influence of calendar time, age of trees, pH value of the soil and density of the canopy of the stand on the state of damage of beeches at the stand. Data have been collected in yearly forest damage inventories carried out in the forest district of Rothenbuch in northern Bavaria from 1983 to 1997. There are 80 observation points with occurrences of beeches spread over the whole area. We use the degree of defoliation as a binary indicator for the state of damage, with $y_{it} = 1$ for light or distinct damage of tree i in year t and $y_{it} = 0$ for no damage. Fig. 2 shows relative frequencies of the response 'damage state' over the years. There is a clear pattern with a maximum of damage around 1986, whereas trees seem to recover in the following 5 years.

A detailed description of the data can be found in Göttlein and Pruscha (1996). The covariates used here are defined as follows: A is the age of the trees at the beginning of the study in 1983, measured in three categories 'below 50 years' (1), between 50 and 120 years (2), and above 120 years (the reference category); CD is the density of the canopy at the stand measured in percentages 0%, 10%, . . ., 90%, 100%; pH is the pH value of the soil near the surface, measured as a metrical covariate with values ranging from a minimum of 3.3 to a maximum of 6.1.

The covariate pH and CD are time varying, whereas A is time constant by definition. On

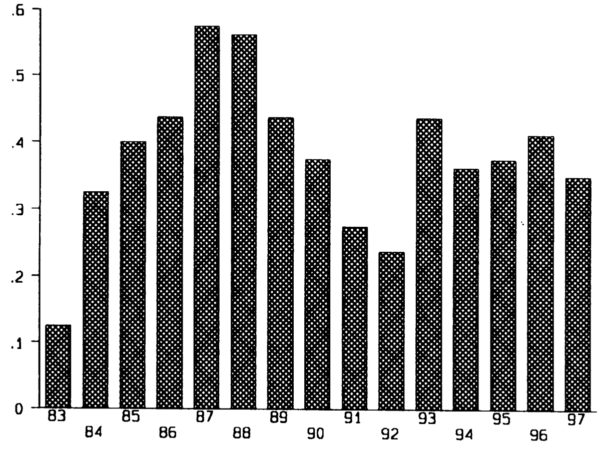


Fig. 2. Relative frequencies of the response 'damage state' over the years 1983–1997

the basis of the results of Gieger (1998) with a marginal model, we use a logistic VCMM with predictor

$$\eta = f_1(t) + f_2(t)A^{(1)} + f_3(t)A^{(2)} + f_4(\text{pH}) + f_5(\text{CD}) + b,$$

where t is the calendar time in years and $A^{(1)}$ and $A^{(2)}$ are dummy variables for A. The effect of calendar time is modelled by a base-line effect $f_1(t)$ and time-varying effects $f_3(t)$ and $f_4(t)$ of age categories, and the possibly non-linear effects of pH and CD are also modelled nonparametrically, using random walk RW(2) smoothness priors. The tree-specific random effect b accounts for correlation and unobserved heterogeneity. The incorporation of these individual random effects is important in this application to avoid biased estimation of the other effects. We noticed this in a reanalysis where the random effect was omitted.

Fig. 3 shows posterior mean estimates for the base-line effect $f_1(t)$, the time-varying effects $f_2(t)$ and $f_3(t)$ of age and the effects of the pH value $f_4(\text{pH})$ and canopy density $f_5(\text{CD})$. The base-line effect $f_1(t)$ corresponds to the time trend of trees over 120 years old. Trees in this age group recovered somewhat after 1987, but then the probability of damage increases again. This is in contrast with the effect of the age groups $A^{(1)}$ (young trees) and $A^{(2)}$ (medium age). The estimated curve $f_2(t)$ is significantly negative and decreases, i.e., in comparison with old trees, younger trees have lower probabilities of being damaged and they seem to recover better over the years. The effect for trees of medium age in Fig. 3(c) is similar but less pronounced. As we might have expected, low pH values, i.e. acid soil, have a positive effect on the probability of damage. At first sight the effect seems to increase for pH values above 4.6. Note, however, that all credible intervals in that area contain 0, so this increase is not significant. The distinct monotonic decrease of the effect of canopy densities of 30% or greater gives evidence that beeches have more shelter from bad environmental influences in stands with high canopy densities.

The effect of the pH value seems to be very small, because the credible intervals are very large (in particular near the boundaries), indicating a strong uncertainty with the estimated effect. In addition, there seems to be some uncertainty about the effect of canopy density for values below 30%. This becomes even more obvious if we compare the estimation results for various choices of hyperparameters of the variance parameters. Table 1 compares the relative

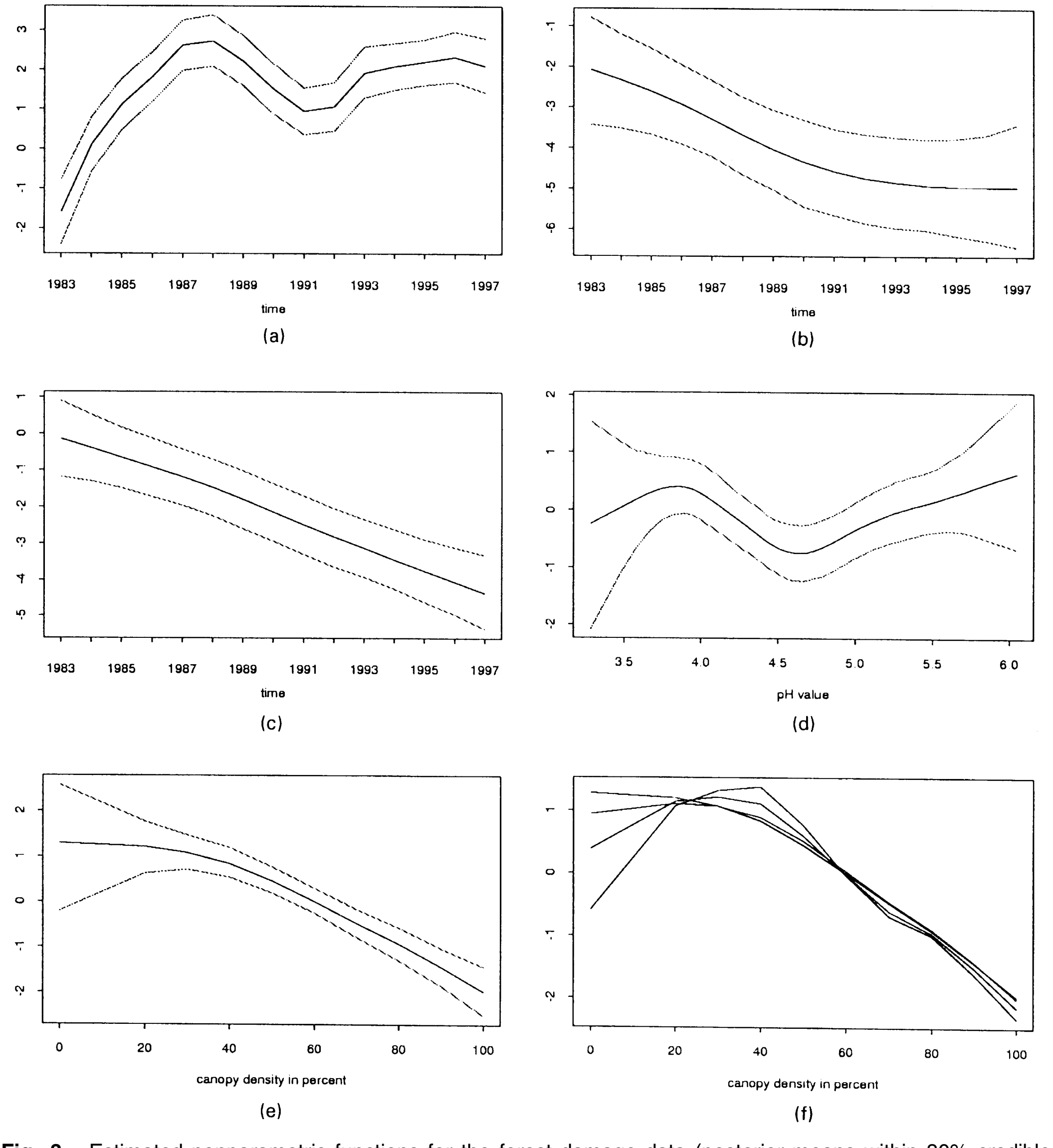


Fig. 3. Estimated nonparametric functions for the forest damage data (posterior means within 80% credible regions): (a) calendar time; (b) varying effect of age category 1; (c) varying effect of age category 2; (d) pH value; (e) canopy density; (f) canopy density (for various hyperpriors)

changes of the estimated posterior means for various choices of hyperpriors with respect to our standard choice $a = 1$ and $c = 0.005$. As can be seen, the estimates for the effect of the pH value and in particular for canopy density change considerably for different hyperpriors whereas the remaining effects are virtually unaffected by the choice of hyperparameters. However, plotting the four estimated functions for pH value onto one graph (not shown) shows no qualitative differences between the different estimates. This is in contrast with the effect of canopy density where the various estimates are quite different for values below 30%. See Fig. 3(f), which shows the estimated posterior means for the various choices of

Table 1. Relative changes of estimated functions for various choices of hyperparameters

a	c	$f_1(t)$	$f_2(A^{(1)})$	$f_3(A^{(2)})$	$f_4(pH)$	$f_5(CD)$	b
1	0.005	0	0	0	0	0	
1	0.01	0.0068	0.0200	0.0122	0.1898	0.1065	0.0259
0.01	0.01	0.0969	0.0822	0.0304	0.3228	0.5881	0.0873
0.0001	0.0001	0.0549	0.0208	0.0106	0.1408	0.2842	0.0491

hyperparameters. However, the effect for values above 30% is virtually unaffected by the choices of hyperparameters.

4.2. Duration of unemployment

In the duration-of-unemployment application, we analyse official unemployment data from the German Federal Employment Office. Our analysis is based on a sample of approximately 12 000 males with one or more spells of unemployment during the observation period from 1980 to 1995. Since the data set must be augmented to apply an ordinary logistic regression model (see below), we end up with more than 192 000 observations, which is a very large data set.

Typical questions that arise in studies on durations of unemployment are as follows. How can the base-line effect (the dependence of duration) be modelled? How can trend and seasonal effects of calendar time be flexibly incorporated? What effect has age? Are there regional differences for the probability of leaving unemployment and seeking a new job? An important problem in connection with persistent unemployment in the 1990s in Europe is the effect of compensation for unemployment and social welfare. Are there negative side-effects of public compensation for unemployment?

Our analysis is based on the following covariates: D, the calendar time measured in months; A, the age (in years) at the beginning of unemployment; N, nationality, dichotomous with categories 'German' and 'foreigner' (the reference category); U, compensation for unemployment, dichotomous with categories 'unemployment benefit' (the reference category) and 'unemployment assistance'; P, the number of previous unemployment periods, 1, 2, 3 and more, and 0 (the reference category); C, the district in which the unemployed have their home.

Note that the calendar time D and compensation for unemployment U are both duration time dependent covariates. Effect coding is used for all the categorical covariates. Since the duration of unemployment is measured in months, we use a discrete time duration model as described in Fahrmeir and Tutz (1997), chapter 9. Let $T = t \in \{1, \ldots, q+1\}$ denote the end of a duration in month t after the beginning of unemployment, and let $x_t^* = (x_1, \ldots, x_t)$ denote the history of covariates up to month t. Then the discrete hazard function is given by

$$\lambda(t;\, x_t^*) = \mathrm{pr}(T = t \mid T \geqslant t,\, x_t^*), \qquad t = 1, \ldots, q.$$

We assume that censoring is non-informative and occurs at the end of the interval, so that the risk set R_t includes all individuals who are censored in interval t. We define binary event indicators y_{it}, $i \in R_t$, $t = 1, \ldots, t_i$, by

$$y_{it} = \begin{cases} 1 & \text{if } t = t_i \text{ and } \delta_i = 1, \\ 0 & \text{otherwise,} \end{cases}$$

where δ_i is the censoring indicator with $\delta_i = 1$ if no censoring occurs. Then the duration process of individual i can be considered as a sequence of binary decisions between remaining unemployed, $y_{it} = 0$, or leaving for the absorbing state $y_{it} = 1$, i.e. the end of unemployment at t. For $i \in R_t$, the hazard function for individual i can be modelled by binary response models

$$\text{pr}(y_{it} = 1 | x_{it}^*) = h(\eta_{it}), \tag{15}$$

with appropriate predictor η_{it} and response function h: $\mathbf{R} \rightarrow (0, 1)$. We choose a logit model with semiparametric predictor

$$\eta = f_1(t) + f_2^{\text{Tr}}(D) + f_3^{\text{S}}(D) + f_4(A) + f_5(C) + \beta_1 N + \beta_2 U + \beta_3 P_1 + \beta_4 P_2 + \beta_5 P_3.$$

The base-line effect $f_1(t)$, the calendar time trend $f_2^{\text{Tr}}(D)$ and the effect of age $f_4(A)$ are estimated nonparametrically by using second-order random walks. For the seasonal component $f_3^{\text{S}}(D)$ we choose the smoothness prior (6) to allow for a flexible time-varying seasonal effect. For the spatial covariate 'district' we choose the (Gaussian) Markov random field prior (9). The influences of the categorical covariates nationality, compensation for unemployment and number of previous unemployment spells are modelled as fixed effects. The results of estimation of the nonparametric terms and the seasonal component are shown in Fig. 4. The base-line effect (Fig. 4(a)) is increasing for the first 4 months and then slopes downwards. Therefore, the probability of finding a new job is best in the first few months of unemployment. The effect of age in Fig. 4(b) is slowly declining until age 52 years and declines dramatically for people older than 53 years. Fig. 4(c) displays the calendar time trend. For comparison with the estimated trend, the absolute number of unemployed people in Germany from 1980 to 1995 is shown in Fig. 4(d). Not surprisingly, a declining calendar time trend corresponds to an increase in the unemployment rate, and vice versa. So the estimated calendar time trend accurately reflects the economic trend of the labour market in Germany. The estimated seasonal pattern (Fig. 4(e)) is relatively stable over the observation period. To gain a better insight, a section of the seasonal pattern for 1990 is displayed in Fig. 4(f). It shows typical peaks in the spring and autumn, a global minimum in the winter and a local minimum in July and August. Low rates of hiring labour in summer can be explained by the distribution of holidays. In Fig. 5(a) the estimated posterior mean of the district-specific effect $f_5(C)$ is displayed, showing a strong spatial pattern, with better chances of finding a new job in the southern part of West Germany, and lower chances in the middle and in the north. The dark spots in the map mark areas that are known for their economic problems during the 1980s and 1990s. This becomes even more obvious in Fig. 5(b), which shows areas with strictly positive (white) and strictly negative (black) credible intervals. Areas where the credible interval contains 0 are grey. Table 2 gives the results of the remaining effects. Germans have improved job chances compared with foreigners, but the effect is not overwhelmingly large. The estimate of -0.37 for unemployment assistance is significantly negative, implying a significantly positive effect of unemployment benefits. At first sight, this result seems to contradict the widely held conjecture about the negative side-effects of unemployment benefits. However, it may be that the unemployment benefit variable also acts as a surrogate variable for those who have worked, and therefore contributed regularly to the insurance system in the past. Further substantive research will be necessary to give definite answers. The number of previous unemployment periods serves as a surrogate for experience in the labour market: an increase in the number of previous spells increases the probability for shorter durations of unemployment.

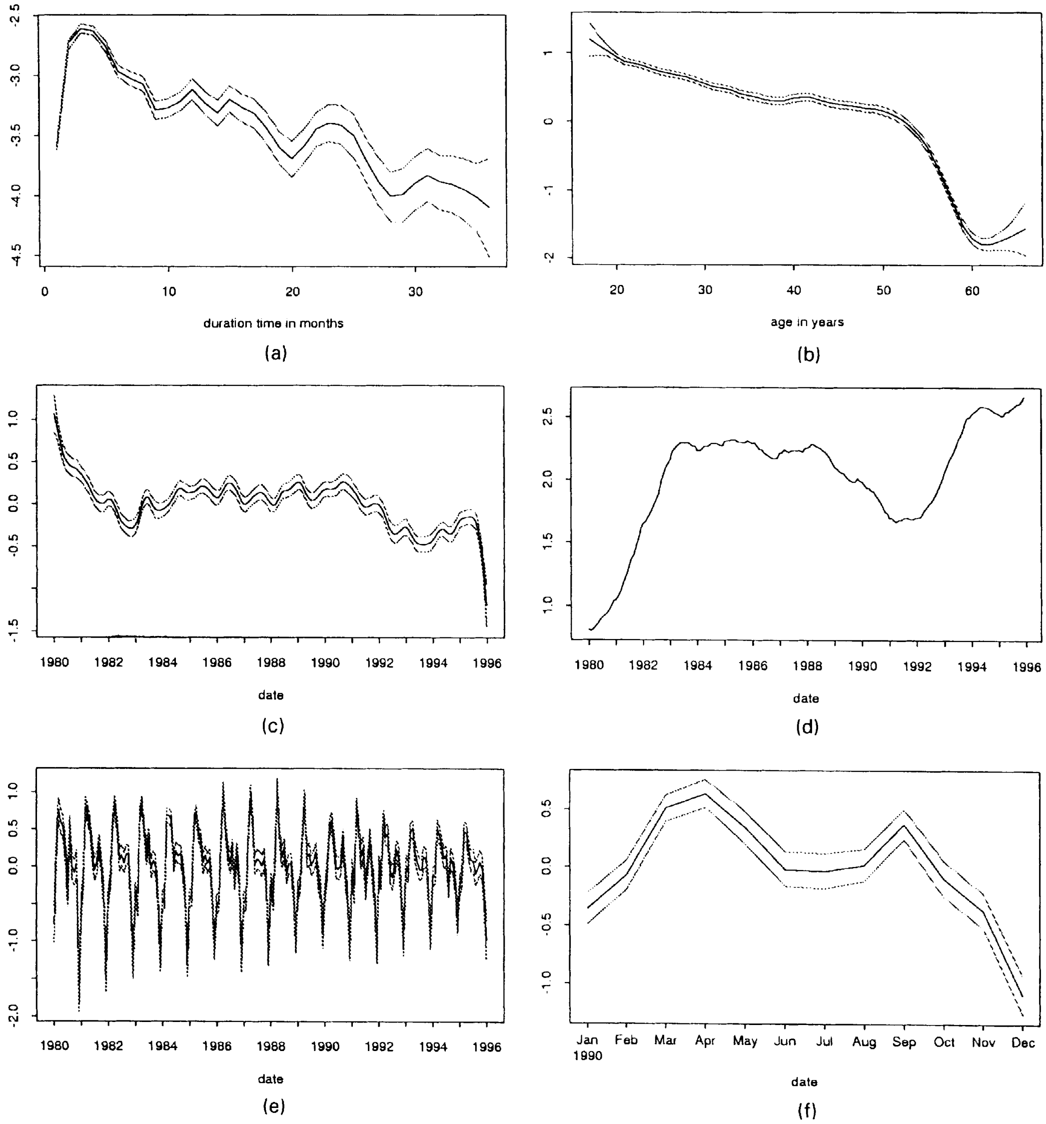
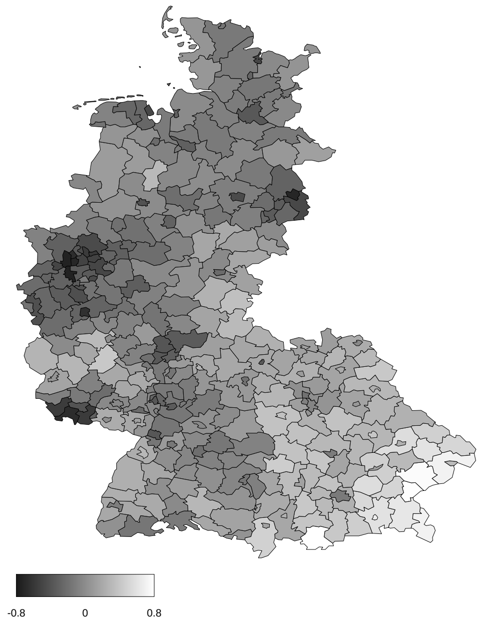


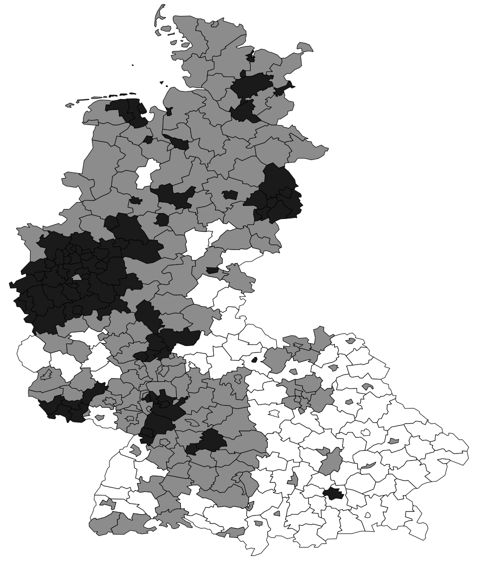
Fig. 4. Estimated nonparametric functions and seasonal effect for the unemployment data (posterior means within 80% credible regions): (a) duration of unemployment; (b) age; (c) calendar time; (d) unemployment rate; (e) seasonal effect; (f) section of the seasonal effect

5. Conclusions

Nonparametric and semiparametric Bayesian regression are useful tools for practical data analysis. They provide posterior mean or median estimates, confidence bands and estimates of other functionals, without having to rely on approximate normality of estimators. A data-driven choice of the smoothing parameters is also incorporated as part of the model. Many recent approaches based on smoothness priors or basis functions considered the case of Gaussian or related responses; our method is particularly useful for nonparametric regression



(a)



(b)

Fig. 5. (a) Posterior means and (b) posterior probabilities of the district-specific effect

Table 2. Estimates of constant parameters in the unemployment data

Covariate	*Mean*	*10% quantile*	*90% quantile*
N	0.15	0.13	0.17
U	−0.37	−0.39	−0.35
P_1	−0.04	−0.06	−0.01
P_2	−0.02	−0.05	0.00
P_3	0.23	0.21	0.25

with fundamentally non-Gaussian responses. The main advantage of hierarchical Bayesian models for nonparametric regression is their modular structure and flexibility. By appropriate modifications of observation models or priors, further generalizations and extensions to other settings are conceptually simple.

For example, the inclusion of interactions between metrical covariates in the observation model can be based on the suggestion of Clayton (1996) for the interaction effects between two timescales. More generally, let x_1 and x_2 be two covariates of possibly different type, with smoothness priors leading to penalty matrices K_1 and K_2 for the unknown main effects $f_1(x_1)$ and $f_2(x_2)$. Then the smoothness prior for the interaction effect $f_{12}(x_1, x_2)$ is defined by the penalty matrix $K_{12} = K_1 \otimes K_2$, the Kronecker product of the precision matrices of the main effects. This idea can be considered as the Bayesian analogue of modelling interactions by tensor product splines in a penalized log-likelihood framework.

To fit unsmooth functions $f(x)$, i.e. functions with discontinuities, edges or rather volatile curvature, the Gaussian prior for the errors in random walk or autoregressive models might be replaced by heavy-tail distributions, or by Gaussian distributions with locally varying variances $u(t) \sim N(0, \tau_t^2)$, $\tau_t^2 = \tau^2 \exp(h_t)$, with h_t obeying a random walk model in a further stage of the hierarchy. We intend to investigate these possibilities in future research.

Acknowledgements

This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 386 ‘Statistische Analyse diskreter Strukturen’. We thank Leo Knorr-Held for helpful discussions, Stefan Bender for providing the unemployment data and Murray Smith for help with the final preparation of the manuscript. Last but not least we thank Dieter Gollnow and our students Andi Brezger, Andrea Hennerfeind and Petra Kragler for computational assistance.

References

Berzuini, C. and Larizza, Ch. (1996) A unified approach for modeling longitudinal and failure time data, with application in medical monitoring. *IEEE Trans. Pattn Anal. Mach. Intell.*, **18**, 109–122.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 192–236.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion). *Statist. Sci.*, **10**, 3–66.

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

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.

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

Biller, C. and Fahrmeir, L. (1997) Bayesian spline-type smoothing in additive generalized regression. *Comput. Statist.*, **12**, 135–151.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.
Breslow, N. E. and Lin, X. (1995) Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.
Carlin, B. P., Polson, N. G. and Stoffer, D. S. (1992) A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. Am. Statist. Ass.*, **87**, 493–500.
Carter, C. K. and Kohn, R. (1994) On Gibbs sampling for state space models. *Biometrika*, **81**, 541–553.
Clayton, D. (1996) Generalized linear mixed models. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 275–301. London: Chapman and Hall.
Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998) Automatic Bayesian curve fitting. *J. R. Statist. Soc.* B, **60**, 333–350.
Fahrmeir, L. and Tutz, G. (1997) *Multivariate Statistical Modelling based on Generalized Linear Models*. New York: Springer.
Gamerman, D. (1997) Efficient sampling from the posterior distribution in generalized linear models. *Statist. Comput.*, **7**, 57–68.
Gieger, Ch. (1998) *Marginale Regressionsmodelle mit Variierenden Koeffizienten für Kategoriale Zielvariablen*. Aachen: Shaker.
Göttlein, A. and Pruscha, H. (1996) Der Einfluss von Bestandskenngrössen, Topographie, Standord und Witterung auf die Entwicklung des Kronenzustandes im Bereich des Forstamtes Rothenbuch. *Forstwissens. Zent.*, **114**, 146–162.
Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.
——— (1993) Varying-coefficient models (with discussion). *J. R. Statist. Soc.* B, **55**, 757–796.
——— (2000) Bayesian backfitting. *Statist. Sci.*, to be published.
Hobart, J. P. and Casella, G. (1996) The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Am. Statist. Ass.*, **91**, 1461–1473.
Knorr-Held, L. (1999) Conditional prior proposals in dynamic models. *Scand. J. Statist.*, **26**, 129–144.
Kohn, R. and Ansley, C. F. (1987) A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM J. Sci. Statist. Comput.*, **8**, 33–48.
Lang, S. and Brezger, A. (2000) BayesX — software for Bayesian inference based on Markov Chain Monte Carlo simulation techniques. *Discussion Paper 187*. Sonderforschungsbereich 386, Ludwig-Maximilians-Universität München, Munich. (Available from `http://www.stat.uni-muenchen.de/sfb386/publikation.html`.)
Lin, X. and Zhang, D. (1999) Inference in generalized additive mixed models by using smoothing splines. *J. R. Statist. Soc.* B, **61**, 381–400.
Mallick, B. K., Denison, D. G. T. and Smith, A. F. M. (2000) Semiparametric generalized linear models: Bayesian approaches. In *Generalized Linear Models: a Bayesian Perspective* (eds D. K. Dey, S. K. Ghosh and B. K. Mallick). New York: Dekker.
Rue, H. (2000) Fast sampling of Gaussian Markov random fields with applications. *Technical Report*. Norwegian University for Science and Technology, Trondheim. (Available from `http://www.math.ntnu.no/preprint/statistics/2000/s1-2000.ps`.)
Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *J. Econometr.*, **75**, 317–343.
Wong, C. and Kohn, R. (1996) A Bayesian approach to estimating and forecasting additive nonparametric autoregressive models. *J. Time Ser. Anal.*, **17**, 203–220.