Biometrics 62, 109–118
March 2006

DOI: 10.1111/j.1541-0420.2005.00392.x

Structured Additive Regression for Categorical Space–Time Data: A Mixed Model Approach

Thomas Kneib* **and Ludwig Fahrmeir****

Department of Statistics, University of Munich, D-80539 Munich, Germany
**email:* Thomas.Kneib@stat.uni-muenchen.de
***email:* Ludwig.Fahrmeir@stat.uni-muenchen.de

Summary. Motivated by a space–time study on forest health with damage state of trees as the response, we propose a general class of structured additive regression models for categorical responses, allowing for a flexible semiparametric predictor. Nonlinear effects of continuous covariates, time trends, and interactions between continuous covariates are modeled by penalized splines. Spatial effects can be estimated based on Markov random fields, Gaussian random fields, or two-dimensional penalized splines. We present our approach from a Bayesian perspective, with inference based on a categorical linear mixed model representation. The resulting empirical Bayes method is closely related to penalized likelihood estimation in a frequentist setting. Variance components, corresponding to inverse smoothing parameters, are estimated using (approximate) restricted maximum likelihood. In simulation studies we investigate the performance of different choices for the spatial effect, compare the empirical Bayes approach to competing methodology, and study the bias of mixed model estimates. As an application we analyze data from the forest health survey.

Key words: Categorical space–time data; Gaussian random fields; Mixed models; P-splines; Restricted maximum likelihood.

1. Introduction

Space–time regression data consist of repeated observations on a response variable and a set of covariates, where, in addition, the spatial location of each unit in the sample is given. Our methodological development is motivated by a longitudinal study on the health status of trees assessed in ordered categories. In addition to usual covariates, the location of each tree is available on a lattice map. Adequate analysis of such space–time regression data requires flexible semiparametric models that can deal with nonlinear relationships as well as temporal and spatial correlation. In recent years, mixed model approaches for longitudinal or spatial data with univariate responses have gained considerable attention. Lin and Zhang (1999) and Zhang (2004) used smoothing splines and random effects to model longitudinal data with responses from a univariate exponential family, Kammann and Wand (2003) introduced geoadditive models for Gaussian responses based on Gaussian random fields (GRFs) and P-splines, and Fahrmeir, Kneib, and Lang (2004) considered a more general empirical Bayes approach based on Markov random fields (MRF), extending generalized additive mixed models and geoadditive models. A fully Bayesian approach allowing for models of comparable complexity is described in Fahrmeir and Lang (2001a). In contrast, the literature dealing with models for categorical space–time data is rather limited (compare Fahrmeir and Lang [2001b], Brezger and Lang [2005], and Yau, Kohn, and Wood [2003] for notable exceptions based on latent Gaussian utilities and Markov chain Monte Carlo simulation techniques).

In this article we propose a general class of structured additive regression (STAR) models for categorical responses, with a flexible semiparametric predictor accounting for the effects of different types of covariates. Effects of continuous covariates or of time and interaction surfaces are modeled through penalized splines. For spatial effects, we suggest three options: If observations are clustered in connected geographical regions, we follow the MRF approach in Fahrmeir et al. (2004). As an extension for space–time data where exact locations are available, we develop a geostatistical approach based on GRFs. Alternatively, two-dimensional P-spline surface smoothers are available as another option.

We develop our approach from a Bayesian perspective, where inference is based on a categorical linear mixed model representation, and variance components, corresponding to inverse smoothing parameters in a frequentist approach, are estimated via (approximate) restricted maximum likelihood. The resulting empirical Bayes approach is closely related to penalized likelihood estimation with penalty terms corresponding to log priors. The methodology is implemented in BayesX, a public domain software package for Bayesian inference, available at `http://www.stat.uni-muenchen.de/~bayesx`. Currently, the software supports the multinomial logit model as well as cumulative logit and probit models. Further extensions will be available in a future version.

Section 2 describes STAR models for categorical data and the different model components. Inference is presented in Section 3. In Section 4, the performance of the approach is investigated through simulation studies. We compare our approach

to fully Bayesian MCMC inference and the polyclass methodology of Kooperberg, Bose, and Stone (1997), and we study the properties of MRF versus GRF modeling of spatial effects. In addition, we investigate bias of the mixed model estimates in comparison to fully Bayesian MCMC estimates. Section 5 contains an application to the forest health data. The conclusions in Section 6 give comments on directions of future research.

2. Categorical Structured Additive Regression

2.1 *Categorical Response Models*

Depending on the type of response and specific assumptions, various regression models for categorical responses $Y \in \{1, \dots, k\}$ have been proposed; see, for example, Agresti (1990) and Fahrmeir and Tutz (2001, Chapter 3). For the case of a nominal response Y with unordered categories $1, \dots, k$ the multinomial logit model

$$P(Y = r) = \frac{\exp(\eta^{(r)})}{1 + \sum_{s=1}^{q} \exp(\eta^{(s)})}, \quad r = 1, \dots, q = k - 1, \quad (1)$$

where k is the reference category and $\eta^{(r)} = u'\alpha^{(r)}$ is a predictor depending on covariates u and regression coefficients $\alpha^{(r)}$, is the most common choice. For an ordered response Y parametric cumulative logit or probit models $P(Y \le r) = F(\eta^{(r)})$ or, equivalently,

$$P(Y = r) = F(\eta^{(r)}) - F(\eta^{(r-1)}), \quad (2)$$

with linear predictor $\eta^{(r)} = \theta^{(r)} - u'\alpha$ and ordered thresholds $\theta^{(1)} < \cdots < \theta^{(q)}$ are most popular. Here, F is the logistic or standard normal distribution function. Various extensions, such as models with alternative-specific covariates and global parameters, sequential models, two-step models, etc. are described, for example, in Fahrmeir and Tutz (2001, Chapter 3). They are all special cases of the general categorical regression model

$$\pi^{(r)} = P(Y = r) = h^{(r)}(\eta^{(1)}, \dots, \eta^{(q)}), \quad r = 1, \dots, q,$$

where $\eta^{(r)} = v_r'\gamma$ is a predictor with appropriately defined design vector v_r, and γ is the vector of all regression parameters. Defining $\pi = (\pi^{(1)}, \dots, \pi^{(q)})$, $\eta = (\eta^{(1)}, \dots, \eta^{(q)})$, $h = (h^{(1)}, \dots, h^{(q)})$, and the design matrix $V = (v_1, \dots, v_q)'$, the general model is

$$\pi = h(\eta), \quad \eta = V\gamma, \quad (3)$$

with appropriately chosen multivariate response function $h : \mathbb{R}^q \to [0, 1]^q$. More details about special subclasses are described in a supplementary technical report (Kneib and Fahrmeir, 2005).

Categorical space–time data can be seen as longitudinal data for individuals or units $i = 1, \dots, n$, observed at time points $t \in \{t_1, t_2, \dots\}$, where the spatial location or site s on a spatial array $\{1, \dots, S\}$ is given for each unit as additional information. For notational simplicity we assume the same time points for each individual, but generalizations to individual-specific time points are obvious. We distinguish between continuous covariates $x_t = (x_{t1}, \dots, x_{tl})'$, whose effects are assumed to be nonlinear, and a further vector u_t of covariates, whose effects are modeled in usual linear parametric form. Categorical space–time data then consists of observations $\{Y_{it}, x_{it}, u_{it}, s_i\}$, $i = 1, \dots, n$, $t = 1, \dots, T$, where s_i is the location or spatial index of individual i.

2.2 *Observation Models*

STAR models extend models with the common linear predictors to more general semiparametric additive predictors. For example, for nominal responses Y_{it}, the linear predictor in (1) is extended to the space–time main effects model

$$\eta_{it}^{(r)} = u_{it}'\alpha^{(r)} + f_1^{(r)}(x_{it1}) + \cdots + f_l^{(r)}(x_{itl}) + f_{time}^{(r)}(t) + f_{spat}^{(r)}(s_i). \quad (4)$$

Here, $f_{time}^{(r)}$ and $f_{spat}^{(r)}$ represent possibly nonlinear effects of time and space, $f_1^{(r)}, \dots, f_l^{(r)}$ are smooth functions of the continuous covariates $x_1, \dots, x_l$, and $u'\alpha^{(r)}$ corresponds to the usual parametric linear part of the predictor. In complete analogy, we can extend the linear predictor in the general model (3), and in any of its subclasses such as ordinal models to a structured additive predictor.

The following extensions can be included as in Fahrmeir et al. (2004):

1. Individual-specific departures from the population effects in model (4) in form of additional random effects $w_{it}'b_i^{(r)}$ in the predictors as in generalized linear mixed models;
2. Interactions in form of varying coefficient terms, for example, effects $g(t)u$ or $g(s)u$ of a covariate u varying over time or space; and
3. Interactions between two continuous covariates x_1, x_2 in form of a smooth surface $f_{1|2}(x_1, x_2)$ modeled by two-dimensional P-splines as in the application in Section 5; see Lang and Brezger (2004) for details.

For developing and implementing the methodology it is useful to cast all models into a generic form. It turns out in Section 2.3 that all unknown functions as well as extensions can be expressed as the product of appropriately defined design vectors and regression coefficients. Thus, we can always rewrite predictor (4) and extended forms as

$$\eta_{it}^{(r)} = u_{it}'\alpha^{(r)} + \sum_{j=1}^{p} z_{itj}'\beta_j^{(r)}. \quad (5)$$

The general multivariate form (3), with predictor vector $\eta_{it} = (\eta_{it}^{(1)}, \dots, \eta_{it}^{(q)})$ extends to

$$\eta_{it} = V_{it}\gamma + \sum_{j=1}^{p} Z_{itj}\delta_j, \quad (6)$$

with appropriately defined design matrices V_{it} and Z_{itj} and regression coefficients γ and δ_j (see again Kneib and Fahrmeir, 2005). With stacked vectors $\pi = (\pi_{it})$, $\eta = (\eta_{it})$ and design matrices $V = (V_{it})$, $Z = (Z_{it})$, $i = 1, \dots, n$, $t = 1, \dots, T$, the general categorical model for all observations is

$$\pi = h(\eta), \quad \eta = V\gamma + Z_1\delta_1 + \cdots + Z_p\delta_p. \quad (7)$$

In our mixed model approach, γ will comprise fixed effects such as $\alpha^{(r)}$ in (5), while $\delta_1, \dots, \delta_p$ are considered as (correlated) random effects comprising parameters such as $\beta_j^{(r)}$

in (5). Because we adopt a Bayesian point of view, this corresponds to assuming a diffuse prior for γ, and informative priors for $\delta = (\delta_1', \ldots, \delta_p')'$. Because some of these priors are partially improper, a reparameterization will be necessary to apply mixed model methodology for inference; see Section 3.

We make the usual conditional independence assumption: Given unknown functions and parameters, observations Y_{it} are conditionally independent with multinomial distributions defined by the specific models. Then, the likelihood and the log likelihood of all observations are uniquely defined as

$$L(\gamma, \delta) = \prod_{i,t} L_{it}(Y_{it} \,|\, \gamma, \delta) \quad \text{and}$$
$$l(\gamma, \delta) = \sum_{i,t} l_{it}(Y_{it} \,|\, \gamma, \delta), \tag{8}$$

where $L_{it}(Y_{it} \,|\, \gamma,\ \delta)$ and $l_{it}(Y_{it} \,|\, \gamma,\ \delta)$ are the (log-)likelihood contribution of observation Y_{it}.

2.3 *Prior Assumptions on Functions and Parameters*

For fixed effects γ we assume independent diffuse priors $p(\gamma) \propto const$, while informative priors will be assigned to parameters $\beta_j^{(r)}$ in (5) or δ_j in (7), representing certain types of functions. Omitting the index r for notational simplicity, a prior for any type of function is now defined by specifying a suitable design vector z_{itj} and a prior distribution for the vector β_j of unknown parameters in (5). All specific priors defined in the following subsections have the general form

$$p\big(\beta_j \,\big|\, \tau_j^2\big) \propto \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right), \tag{9}$$

where K_j is a penalty matrix that shrinks parameters toward zero or penalizes too abrupt jumps between neighboring parameters and τ_j^2 acts as a smoothing parameter. In most cases K_j will be rank deficient and, therefore, the prior for β_j is partially improper.

For given or known variance parameters, Bayesian inference is then based on the posterior $p(\gamma,\ \delta \,|\, Y)$. For full Bayesian inference, weakly informative inverse gamma hyperpriors are usually assigned to τ_j^2. In our empirical Bayes approach, τ_j^2 is considered an unknown constant. As an alternative to data-driven determination of τ_j^2, for example, by crossvalidation, the Bayesian point of view opens the way to estimate τ_j^2 by restricted maximum likelihood; see Section 3. From a frequentist point of view, the log-prior $\log p(\beta_j \,|\, \tau_j^2)$ in (9) corresponds to a penalty term in a penalized likelihood approach, which is closely related to the empirical Bayes approach.

2.3.1 *Priors for continuous covariates and time scales.* A popular approach to model smooth functions of continuous covariates or time scales are P-splines, introduced by Eilers and Marx (1996). The basic idea is to approximate a function $f_j(x_j)$ by a linear combination of B-spline basis functions B_m of degree l defined on a set of d equally spaced knots, that is,

$$f_j(x_{itj}) = \sum_{m=1}^{M_j} \beta_{jm} B_m(x_{itj}) = z_{itj}'\beta_j,$$

where $M_j = d + l$. The M_j-dimensional design vector z_{itj} consists of the basis functions evaluated at the observation x_{itj}, that is, $z_{itj} = (B_1(x_{itj}), \ldots, B_{M_j}(x_{itj}))'$. Eilers and Marx (1996) suggest to use a moderately large number of knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on first- or second-order differences of adjacent B-spline coefficients to guarantee sufficient smoothness of the fitted curves. In a Bayesian approach we use the stochastic analogue of difference penalties, that is, first- or second-order random walks with Gaussian errors. The joint distribution of the regression parameters β_j is then easily cast into the general form (9) with penalty matrix $K_j = D'D$, where D is a first- or second-order difference matrix. More details about Bayesian P-splines including models for interaction terms can be found in Lang and Brezger (2004).

2.3.2 *Priors for spatial effects.* For the specification of the spatial effect f_{spat} we distinguish two situations:

MRF priors: Suppose first that the index $s \in \{1, \ldots, S\}$ represents the location or site in connected geographical regions. For such data, a spatially correlated effect can be modeled by an MRF, where two sites s and s' are assumed to be neighbors if they share a common boundary. Defining $\beta_s = f_{spat}(s)$, $s = 1, \ldots, S$, the simplest MRF prior for the function evaluations of the spatial effect is

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

where N_s is the number of adjacent sites and $s' \in \partial_s$ denotes that site s' is a neighbor of site s. Thus the (conditional) mean of β_s is an unweighted average of function evaluations of neighboring sites. The S-dimensional design vector $z_{it,spat} = (0, \ldots, 1, \ldots, 0)'$ is now a 0/1 incidence vector. Its sth component is 1 if the corresponding observation is located in region s, and 0 otherwise. The $S \times S$ penalty matrix K has the form of an adjacency matrix.

GRF priors: If exact locations $s = (s_x,\ s_y)$ are available, we can use GRF priors, originating from geostatistics, which have been used by Kammann and Wand (2003) to model the spatial component in Gaussian regression models. The spatial effect $f_{spat}(s) = \beta_s$ is then assumed to follow a zero mean stationary GRF $\{\beta_s : s \in \mathbb{R}^2\}$ with variance τ_{spat}^2 and isotropic correlation function $\operatorname{corr}(\beta_s, \beta_{s+h}) = C(\|h\|)$. This means that correlations between sites that are $\|h\|$ units apart are the same, regardless of direction and location of the sites. For a finite array $s \in \{s_1, \ldots, s_S\}$ of sites as in image analysis or in our application to forest health data, the prior for $\beta_{spat} = (\beta_1, \ldots, \beta_S)'$ is of the general form (9) with $K = C^{-1}$ and $C[i, j] = C(\|s_i - s_j\|)$, $1 \le i, j \le S$. The design vector $z_{it,spat}$ is again a 0/1 incidence vector.

Several proposals for the choice of the correlation function $C(r)$ have been made. In the kriging literature, the Matérn family $C(r;\ \rho,\ \nu)$ is highly recommended (Stein, 1999). For prechosen values $\nu = m + 1/2$, $m = 0, 1, 2, \ldots$, of the smoothness parameter ν, simple correlation functions $C(r;\ \rho)$ are obtained, for example, $C(r;\ \rho) = \exp(-|r/\rho|)(1 + |r/\rho|)$ with $\nu = 1.5$. The parameter ρ controls how fast correlations die out with increasing distance $r = \|h\|$. It can be determined in a preprocessing step, for example, using the simple rule

$$\hat{\rho} = \max_{i,j} \|s_i - s_j\|/c, \tag{11}$$

which ensures scale invariance. The constant $c > 0$ is chosen in such a way, that $C(c)$ is small. Therefore the different values of $\|s_i - s_j\|/\hat{\rho}$ are spread out over the r-axis of the correlation function. This choice of ρ has proved to work well in our experience.

Although we described them separately, approaches for exact locations can also be used in the case of connected geographical regions, for example, based on the centroids of the regions. Conversely, we can also apply MRFs to exact locations if neighborhoods are defined based on a distance measure or via discretization of the observation area. The main difference between GRFs and MRFs, considering their numerical properties, is the dimension of the penalty matrix. For MRFs the dimension of K equals the number of different regions and is therefore independent from the sample size. On the other side, for GRFs, the dimension of K is given by the number of distinct locations, which in most cases is close or equal to the sample size. So the number of regression coefficients used to describe an MRF is usually much smaller than for a GRF and, therefore, the estimation of GRFs is computationally more expensive. To overcome this difficulty, Kammann and Wand (2003) propose low-rank kriging to approximate stationary GRFs. Note first, that we can equivalently define GRFs based on the n-dimensional design vector with entries $z_{it,spat} = (C(\|s_{it} - s_{11}\|), \ldots, C(\|s_{it} - s_{nt}\|))'$ and penalty matrix $K = C$. To reduce the dimensionality of the estimation problem we define a subset of knots $\mathcal{D} = \{\kappa_1, \ldots, \kappa_M\}$ of the set of distinct locations based on a space filling algorithm (compare Nychka and Saltzman, 1998 for details). This allows to define the approximation $f_{spat}(s_i) = z'_{it,spat}\beta$ with the M-dimensional design vector $z_{it,spat} = (C(\|s_{it} - \kappa_1\|), \ldots, C(\|s_i - \kappa_M\|))'$, penalty matrix $K = C$, and $C[j, k] = C(\|\kappa_j - \kappa_k\|)$. The number of knots M controls the tradeoff between the accuracy of the approximation (M close to the sample size) and the numerical simplification (M small).

3. Inference

Inference in STAR models is not performed on the basis of the original parameterization of the regression coefficients, because there is no easy rule on how to choose the variance parameters. For univariate responses, a popular idea is to represent models with penalties as mixed models with i.i.d. random effects. This idea goes back to Green (1987) for smoothing splines and has been used in a variety of settings throughout the last 5 years (e.g., Lin and Zhang, 1999; Kammann and Wand, 2003; Ruppert, Wand, and Carroll, 2003; Wand, 2003; Fahrmeir et al., 2004; or Zhang, 2004). In mixed model representation a variance components model is obtained, and techniques for estimating the variance parameters are already available, for example, via restricted maximum likelihood. In the following we extend this approach to categorical STAR models.

3.1 *Mixed Model Representation*

Assuming that K_j is known and does not depend on further parameters to be estimated, we can express β_j via a one-to-one transformation in terms of a parameter vector β_j^{unp} with flat prior and a parameter vector β_j^{pen} with i.i.d. Gaussian prior. While β_j^{unp} captures the part of a function f_j that is not penalized by K_j, β_j^{pen} captures deviations from this unpenalized part. If K_j has full rank, the unpenalized part vanishes completely, and with $\beta_j^{pen} = K_j^{1/2}\beta_j$ we directly obtain $\beta_j^{pen} \sim N(0, \tau_j^2 I)$. For the general case of rank deficient K_j things are somewhat more complicated. If we assume that the jth parameter vector has dimension d_j and the corresponding penalty matrix has rank k_j the decomposition of β_j into a penalized and an unpenalized part is of the form

$$\beta_j = Z_j^{unp}\beta_j^{unp} + Z_j^{pen}\beta_j^{pen}, \tag{12}$$

with a $d_j \times (d_j - k_j)$ matrix Z_j^{unp} and a $d_j \times k_j$ matrix Z_j^{pen}.

Requirements for decomposition (12) are:

(i) The composed matrix $(Z_j^{unp} Z_j^{pen})$ has full rank to make the transformation in (12) a one-to-one transformation.
(ii) Z_j^{unp} and Z_j^{pen} are orthogonal, that is, $Z_j^{unp\prime} Z_j^{pen} = 0$.
(iii) $Z_j^{unp\prime} K_j Z_j^{unp} = 0$, resulting in β_j^{unp} being unpenalized by K_j.
(iv) $Z_j^{pen\prime} K_j Z_j^{pen} = I$, resulting in an i.i.d. Gaussian prior for β_j^{pen}.

In general, matrices fulfilling these requirements can be obtained from the spectral decomposition of K_j; compare Fahrmeir et al. (2004) for details.

From requirements (i) to (iv) we obtain $\mathrm{p}(\beta_{jm}^{unp}) \propto const$, $m = 1, \ldots, d_j - k_j$ and $\beta_j^{pen} \sim N(0, \tau_j^2 I)$. The decomposition of β_j leads to a similar decomposition for the linear combination $z'_{itj}\beta_j$ representing a function in (5):

$$\begin{aligned} z'_{itj}\beta_j &= z'_{itj} Z_j^{unp}\beta_j^{unp} + z'_{itj} Z_j^{pen}\beta_j^{pen} \\ &= \tilde{z}_{itj}^{unp\prime}\beta_j^{unp} + \tilde{z}_{itj}^{pen\prime}\beta_j^{pen}, \end{aligned} \tag{13}$$

allowing to rewrite the additive predictor (5) as

$$\eta_{it}^{(r)} = u'_{it}\gamma^{(r)} + \sum_{j=1}^{p} \left(\tilde{z}_{itj}^{unp\prime}\beta_j^{unp(r)} + \tilde{z}_{itj}^{pen\prime}\beta_j^{pen(r)}\right).$$

Inserting decomposition (12) into the general model (7), the predictor η can be analogously reparameterized in the form

$$\eta = V\gamma + \sum_{j=1}^{p} \tilde{Z}_j^{unp}\delta_j^{unp} + \sum_{j=1}^{p} \tilde{Z}_j^{pen}\delta_j^{pen}, \tag{14}$$

with appropriately defined design matrices $V, \tilde{Z}_j^{unp}$ and $\tilde{Z}_j^{pen}$. Defining $\delta^{unp} = (\gamma', \delta_1^{unp\prime}, \ldots, \delta_p^{unp\prime})'$, $\delta^{pen} = (\delta_1^{pen\prime}, \ldots, \delta_p^{pen\prime})'$, $Q = (V, \tilde{Z}_1^{unp}, \ldots, \tilde{Z}_p^{unp})$, and $P = (\tilde{Z}_1^{pen}, \ldots, \tilde{Z}_p^{pen})$ we finally obtain

$$\eta = Q\delta^{unp} + P\delta^{pen}, \quad \delta^{pen} \sim N(0, \Lambda),$$

where the diagonal matrix Λ contains all variance components τ_j^2 in appropriately arranged order.

3.2 *Empirical Bayes Inference for Categorical Mixed Models*

Estimation of categorical mixed models now consists of largely two steps: Alternately the regression coefficients are updated given the current values of the variance parameters and vice versa. Posterior mode estimates for δ^{unp} and δ^{pen} given Λ are obtained by maximizing the posterior $p(\delta^{unp}, \delta^{pen} \mid y) \propto L(\delta^{unp}, \delta^{pen})p(\delta^{unp})p(\delta^{pen})$, where $L(\delta^{unp}, \delta^{pen})$ denotes the likelihood of the model, which in fact equals the likelihood in (8). Utilizing the flat prior of δ^{unp} we obtain

$$l_{pen}(\delta^{unp}, \delta^{pen}) = l(\delta^{unp}, \delta^{pen}) - \frac{1}{2}\delta^{pen'}\Lambda^{-1}\delta^{pen}. \qquad (15)$$

In principle, maximization of (15) is carried out through a Fisher scoring type algorithm. Similar to estimation in GLMs, the Fisher scoring algorithm can be rewritten as iteratively weighted least squares (IWLS), yielding the following system of equations

$$\begin{pmatrix} Q'WQ & Q'WP \\ P'WQ & P'WP + \Lambda^{-1} \end{pmatrix} \begin{pmatrix} \delta^{unp} \\ \delta^{pen} \end{pmatrix} = \begin{pmatrix} Q'W\tilde{y} \\ P'W\tilde{y} \end{pmatrix}, \qquad (16)$$

to be solved to obtain updated estimates. The (block diagonal) weight matrix W and the working observations $\tilde{y}$ are defined in complete analogy to usual parametric models for categorical responses; compare Fahrmeir and Tutz (2001, Chapter 3).

Note that iteratively solving the system of equation (16) is equivalent to approximating the likelihood $L(\delta^{unp}, \delta^{pen})$ of a multinomial distribution with the likelihood of a multivariate Gaussian distribution having an iteratively reweighted covariance matrix W^{-1}. This approximation also allows for the construction of pointwise confidence intervals for the regression coefficients and therefore for the function evaluations which are linear combinations of the regression coefficients. Ruppert et al. (2003, Chapter 6.5) give some suggestions on the construction of simultaneous confidence bands, either based on simulation techniques or analytic approximations. The model fit can be assessed via information criteria such as AIC and BIC based on the log likelihood $l(\delta^{unp}, \delta^{pen})$ and the usual degrees of freedom or generalized crossvalidation based on the deviance residuals (see also the application in Section 5).

Restricted maximum likelihood estimates for the variances are obtained by maximizing the marginal likelihood

$$L^*(\Lambda) = \int L(\delta^{unp}, \delta^{pen}, \Lambda)\, d\delta^{pen} d\delta^{unp}, \qquad (17)$$

and, therefore, REML is also termed marginal likelihood estimation in the literature. Because direct evaluation of the integral in (17) is not possible in general, we use a quadratic approximation to $L(\delta^{unp}, \delta^{pen}, \Lambda)$, which is in fact equivalent to the approximation made in IWLS. This yields the restricted log likelihood

$$\begin{aligned} l^*(\Lambda) \approx & -\frac{1}{2}\log(|\Sigma|) - \frac{1}{2}\log(|Q'\Sigma^{-1}Q|) \\ & -\frac{1}{2}(\tilde{y} - Q\hat{\delta}^{unp})'\Sigma^{-1}(\tilde{y} - Q\hat{\delta}^{unp}), \end{aligned} \qquad (18)$$

where $\Sigma = W^{-1} + P'\Lambda P$ is an approximation to the marginal covariance of $\tilde{y} \,|\, \delta^{pen}$. Maximization of (18) can now be conducted by Newton–Raphson or Fisher scoring; compare Harville (1977) or Fahrmeir et al. (2004) for formulae of the score vector and the expected Fisher information. Fahrmeir et al. also derive numerical superior expressions for these formulae, allowing the computation of REML estimates even for fairly large data sets. Although these expressions are derived for univariate responses, they can easily be extended to a categorical setting.

4. Simulation Studies

To investigate performance, we conducted several simulation studies based on a multinomial logit model with three categories and predictors defined to be the sum of a nonparametric effect and a spatial effect (see Figure 1 for a detailed description of the simulation design). The first aim was to compare different parameterizations of the spatial effect and different approaches to the estimation of categorical STAR models. Therefore, 250 simulation runs with $n = 500$ observations were generated. We used cubic P-splines with second-order random walk penalty and 20 knots to estimate effects of the continuous covariate. The spatial effect was estimated either by an MRF, a (full) GRF, or a two-dimensional P-spline (based on 10×10 inner knots). For the competing fully Bayesian approach by Fahrmeir and Lang (2001b) and Brezger and Lang (2005), where inverse Gamma priors $IG(a, b)$ with $a = b = 0.001$ are assigned to the variances, the GRF approach was computationally too demanding due to the inversion of a full precision matrix for the spatial effect in each iteration. Therefore, we excluded the fully Bayesian GRF approach from the comparison. As a further competitor we utilized the R-implementation of the procedure polyclass described in Kooperberg et al. (1997). Here, nonparametric effects and interaction surfaces are modeled by linear splines and their tensor products. Smoothness of the estimated curves is not achieved by penalization but via stepwise inclusion and deletion of model terms corresponding to basis functions based on AIC.

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

1. Generally REML estimates have somewhat smaller median MSE than their fully Bayesian counterparts, with larger differences for spatial effects (see Figure 2a and 2b for exemplary results in category 2).
2. Estimates for the effects of the continuous covariate are rather insensitive with respect to the model choice for the spatial effect (Figure 2a).
3. Two-dimensional P-splines lead to the best fit for the spatial effect although data are provided with discrete spatial information (Figure 2b).
4. Polyclass is outperformed by both the empirical and the fully Bayesian approach and therefore results are deferred to Figure 2c together with REML estimates based on two-dimensional P-splines. Presumably, the poor performance of polyclass is mainly caused by the special choice of linear splines, resulting in rather peaked estimates. Smoother basis functions, for example, truncated cubic polynomials might improve the fit substantially but are not available in the implementation.
5. For some simulation runs with spatial effects modeled by MRFs, no convergence of the REML algorithm could be achieved. This was also the case if the spatial effect was modeled by a two-dimensional P-spline but in a much smaller number of cases. Obviously the same convergence problems as described in Fahrmeir et al. (2004) appear in a categorical setting. However, the arguments given there still hold and so we again used estimates obtained from the final (100th) iteration.

It is frequently argued that results from REML estimation procedures in GLMMs tend to be biased due to the Laplace approximation involved, especially in sparse data situations (compare, e.g., Lin and Breslow, 1996). Therefore, as a

- Predictor:

$$\eta_i^{(r)} = f_1^{(r)}(x_i) + f_2^{(r)}(s_i)$$

- Category 1:

$$f_1^{(1)}(x) = \sin[\pi(2x-1)]$$

$$f_2^{(1)}(s) = -0.75|s_x|(0.5+s_y)$$

- Category 2:

$$f_1^{(2)}(x) = \sin[2\pi(2x-1)]$$

$$f_2^{(2)}(s) = 0.5(s_x+s_y)$$

- x is chosen from an equidistant grid of 100 values between -1 and 1.
- (s_x, s_y) are the centroids of the 124 districts s of the two southern states of Germany (see Figures).

Figure 1. Simulation design.

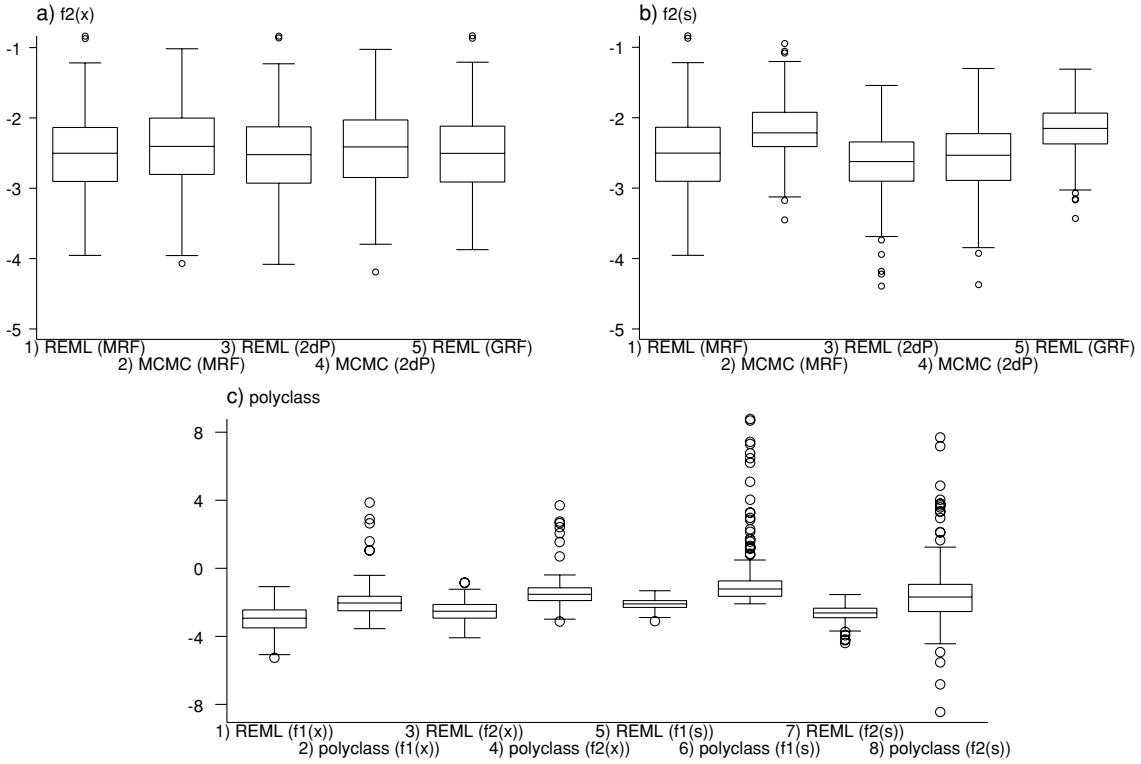

Figure 2. Simulation study: boxplots of log(MSE) for different model choices.

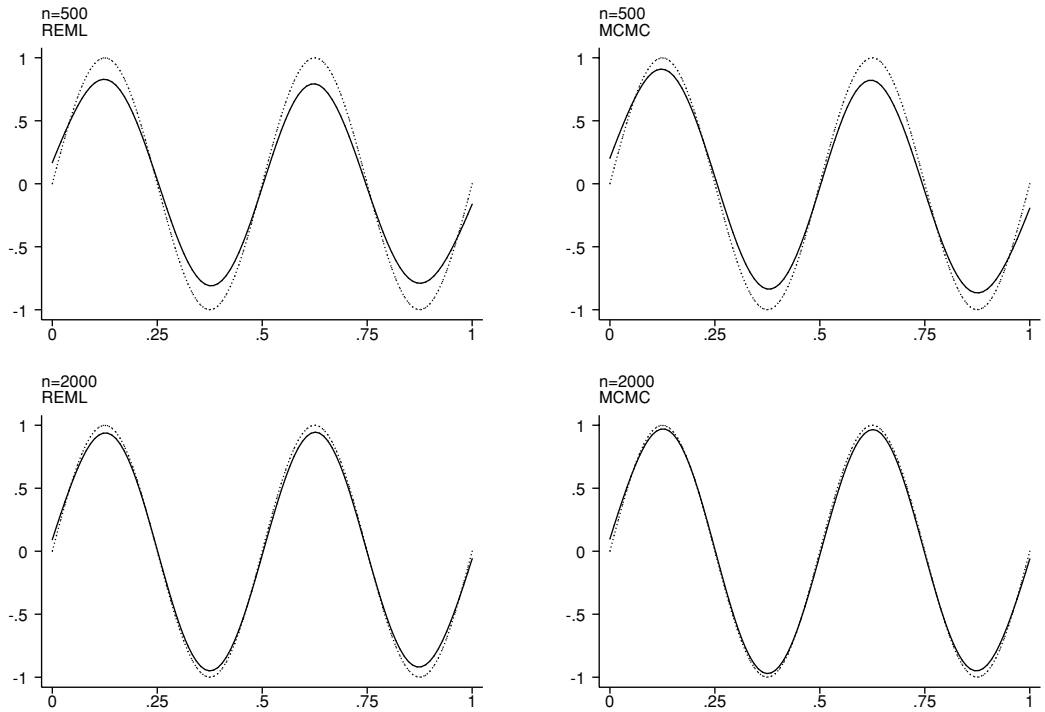

Figure 3. Simulation study: bias for $f^{(2)}(x)$ based on REML estimates (left panels) and MCMC estimates (right panels). Average estimates are indicated by solid lines, true functions by dashed lines.

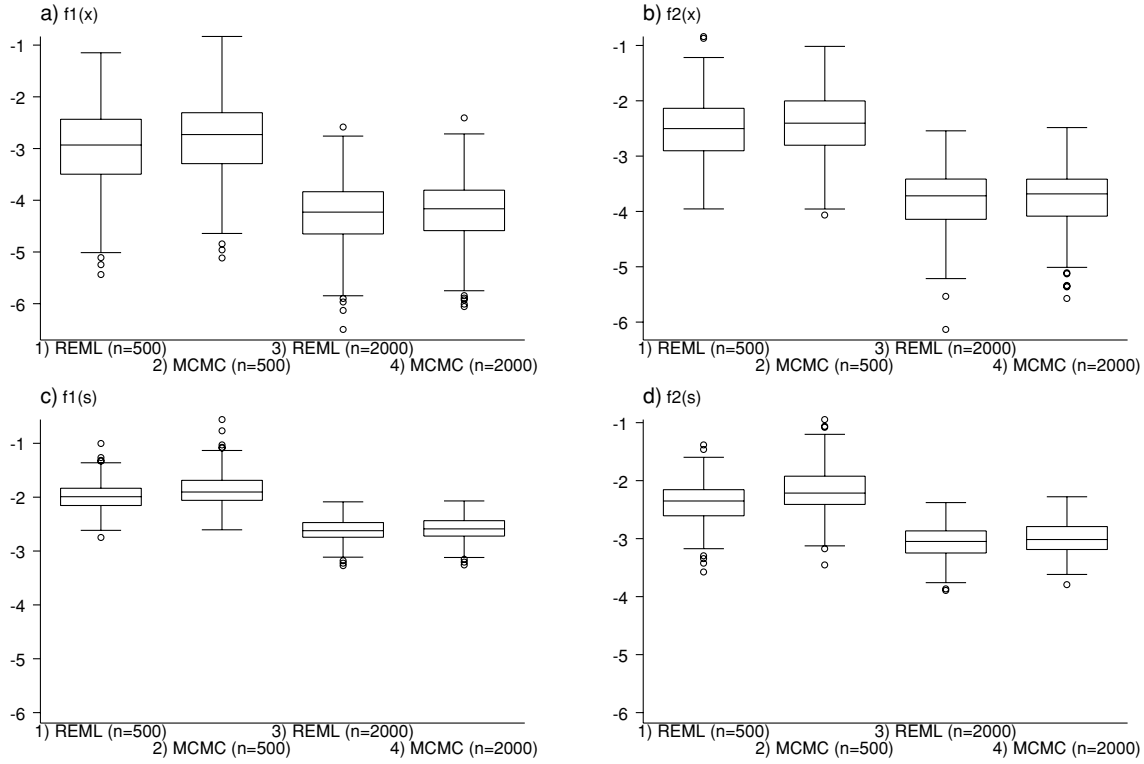

Figure 4. Simulation study: boxplots of log(MSE) for different sample sizes.

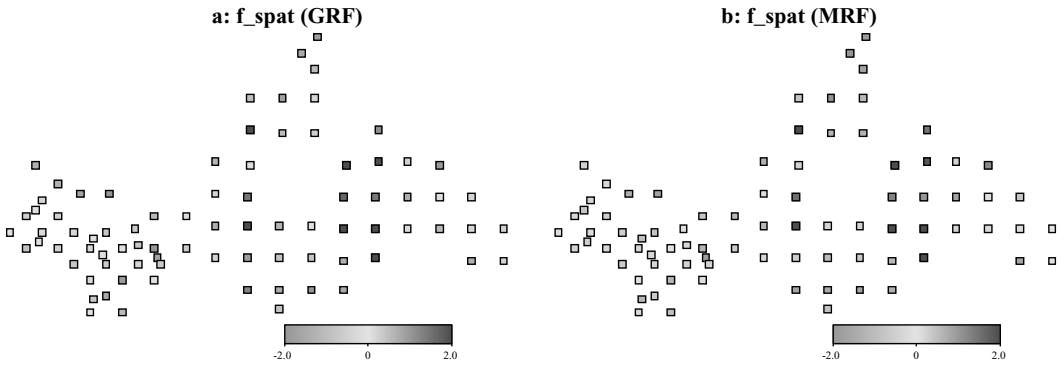

Figure 5. Forest health data: spatial effects based on a GRF and an MRF.

second aim, we investigated whether this observation holds in a categorical setting in a second simulation study, again based on the same model but with different sample sizes, namely $n = 500$, $n = 1000$, and $n = 2000$. Results from the REML estimation procedure were compared to their fully Bayesian counterparts because these estimates do not use any approximations but work with the exact posterior. For both approaches, the spatial effect was estimated by an MRF. The results of the simulation lead to the following conclusions:

1. In general, bias is smaller for MCMC estimates, most noticeably for more wiggly functions. As an example, we show the bias for $f^{(2)}(x)$ with $n = 500$ and $n = 2000$ in Figure 3.
2. For increasing sample sizes, differences almost vanish and both approaches give nearly unbiased estimates.
3. REML estimates perform superior to MCMC estimates in terms of MSE (Figure 4).
4. A further simulation study with ordinal responses led to the same conclusions (results not shown).

5. Application: A Space–Time Study on Forest Health

These space–time data have been collected in yearly visual forest health inventories carried out at 83 observation points with beeches in a forest district in the northern part of Bavaria from 1983 to 2001. For each tree, the degree of defoliation serves as an indicator for its damage state, which is given as an ordered response with categories $Y_{it} = 1$ (no damage of tree i in year t), $Y_{it} = 2$ (medium damage), and $Y_{it} = 3$ (severe damage) $i = 1, \ldots, 83$, $t = 1983, \ldots, 2001$. In addition to temporal and spatial information, the data set includes a number of covariates describing the stand and the site of the

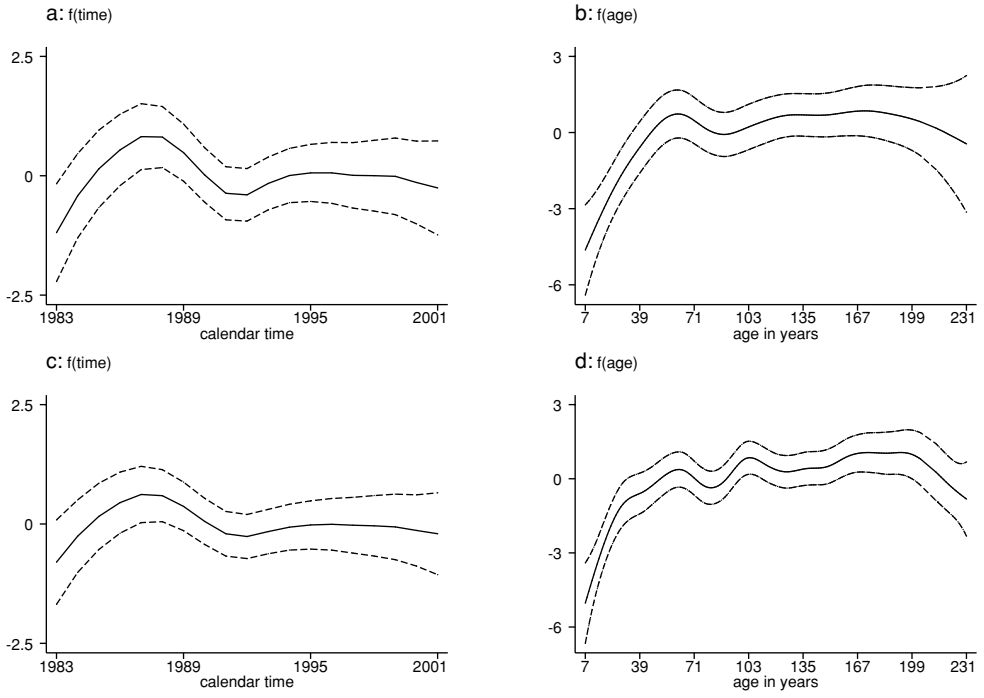

Figure 6. Forest health data: nonparametric main effects for a model with spatial effect modeled by a GRF (upper panels) and for a model excluding spatial effects (lower panels).

tree, and the soil at the stand. Based on exploratory analyses, we chose the cumulative probit model

$$P(Y_{it} \leq r) = \Phi(\theta^{(r)} - [f_1(t) + f_2(a_{it}) + f_3(t, a_{it}) + f_{spat}(s_i) + u'_{it}\gamma]),$$

where Φ denotes the standard normal cumulative distribution function, $f_1(t)$ and $f_2(a_{it})$ are nonparametric effects of calendar time and age of the tree, $f_3(t, a_{it})$ is an interaction surface between these two variables, and $f_{spat}(s_i)$ is a spatial effect. Effects of all further covariates were found to be of parametric form in the exploratory analyses and are therefore subsumed in the vector u_{it}. For interpretation of estimation results note the following: In accordance with definition (2), higher (lower) values of covariate effects correspond to worse (healthier) state of the trees. To shorten the discussion, we do not show any results on covariates with parametric effects.

We examined both MRFs and stationary GRFs for the spatial effect and also estimated a model that neglects spatial correlations and, therefore, has no spatial effect at all. For MRFs two trees were considered as neighbors if their distance was less than 1.2 km. The correlation function of the GRF was chosen to be Matérn with $\nu = 1.5$ and the scale parameter ρ was determined in a preprocessing step using the rule in (11). Both approaches with spatial effects lead to very similar estimates and hence we do not show all results. For the spatial effect, estimates are shown in Figure 5a and 5b. To judge whether a spatial effect is needed and which of the approaches performs better, we computed the Akaike and the Bayesian information criteria as well as the generalized crossvalidation statistic. Table 1 gives the results for the three models. Obviously including a spatial effect improves the model fit dramatically, regardless of the special choice. In terms of AIC and GCV the MRF approach performs better, while the kriging approach is superior when BIC is considered. However, differences are quite small, as are the differences between the estimated effects. Figure 6a and 6b show the estimates of the main effects $f_1(t)$ and $f_2(a)$ for a model with f_{spat} modeled by a GRF. Additionally, we include results for f_1 and f_2 from the model excluding any spatial effect in Figure 6c and 6d. The estimated temporal effect reflects quite well the trends found in descriptive analyses, with an increased frequency of damaged trees in the mid-1980s. For the model without spatial effects, the time trend is less pronounced, but the functional form remains almost the same. For the effect of age differences become more noticeable. Here, estimates without spatial effects are more wiggly with an additional peak around 100 years. Obviously the age effect absorbs some of the effects which are otherwise covered by the spatial component.

Table 1
Forest health data: information criteria and generalized crossvalidation for models with different spatial effects

	No spatial effect	MRF	GRF
$-2 \times$ log likelihood	1411.95	917.86	925.48
Degrees of freedom	67.35	112.87	111.82
AIC	1546.64	1143.61	1149.12
BIC	1906.54	1746.82	1746.69
GCV	0.87	0.55	0.56

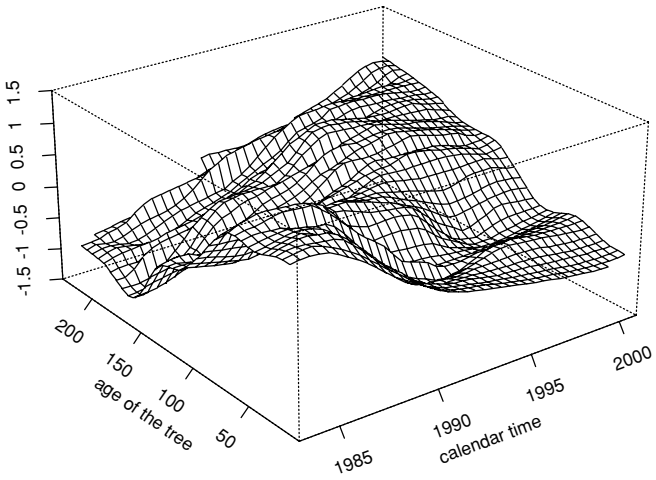

Figure 7. Forest health data: interaction between calendar time and age of the tree for a model with structured spatial effect modeled as GRF.

Finally, Figure 7 shows the interaction between calendar time and age of the tree. Apparently, young trees were in poorer health state in the 1980s but recovered in the 1990s, unlike older trees that showed the contrary behavior. A possible interpretation is that it takes longer until older trees are affected by harmful environmental circumstances, while younger trees are affected nearly at once but manage to accommodate when growing older.

6. Conclusions

Due to the increasing availability of space–time regression data in connection with complex scientific problems, flexible semiparametric regression models of the type considered in this article are of substantial interest in empirical research. Compared to fully Bayesian approaches relying on MCMC sampling techniques, the mixed model approach is a promising alternative and can also be understood as penalized likelihood inference from a frequentist point of view.

For categorical response models, some extensions are desirable. First, we intend to include category-specific effects into ordinal models. For example, the thresholds $\theta^{(r)}$ might be time varying, that is, we have to consider category-specific trend functions $f^{(r)}_{time}(t)$ in the predictors $\eta^{(r)}$. Similarly, inclusion of category-specific covariates in nominal models is often needed in practice. A more challenging extension concerns models for correlated categorical responses. So far we analyze the health status of trees with separate models for beeches, spruces, etc. Instead we might use a joint model in simultaneous analyses for all tree species observed at the stands.

Acknowledgements

We thank two referees and an associate editor for helpful comments, Axel Göttlein for providing the data and for interesting discussions, and we gratefully acknowledge financial support from the German Science Foundation (DFG), Collaborative

Research Center 386 "Statistical Analysis of Discrete Structures."

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Brezger, A. and Lang, S. (2005). Generalized additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, in press.

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

Fahrmeir, L. and Lang, S. (2001a). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society C* **50,** 201–220.

Fahrmeir, L. and Lang, S. (2001b). Bayesian semiparametric regression analysis of multicategorical time-space data. *Annals of the Institute of Statistical Mathematics* **53,** 10–30.

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

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica* **14,** 731–761.

Green, P. J. (1987). Penalized likelihood for general semiparametric regression models. *International Statistical Review* **55,** 245–259.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72,** 320–338.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society C* **52,** 1–18.

Kneib, T. and Fahrmeir, L. (2005). *Supplement to "Structured additive regression for categorical space-time data: A mixed model approach."* Technical Report. Available at `http://www.stat.uni-muenchen.de/~kneib`.

Kooperberg, C., Bose, S., and Stone, C. J. (1997). Polychotomous regression. *Journal of the American Statistical Association* **92,** 117–127.

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

Lin, X. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* **91,** 1007–1016.

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

Nychka, D. and Saltzman, N. (1998). Design of air-quality monitoring networks. *Lecture Notes in Statistics* **132,** 51–76.

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

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics* **18,** 223–249.

Yau, P., Kohn, R., and Wood, S. (2003). Bayesian variable selection and model averaging in high dimensional multinomial nonparametric regression. *Journal of Computational and Graphical Statistics* **12,** 23–54.

Zhang, D. (2004). Generalized linear mixed models with varying coefficients for longitudinal data. *Biometrics* **60,** 8–15.

Received May 2004. *Revised December* 2004.
Accepted March 2005.