Structured additive regression for multicategorical space-time data: A mixed model approach

Thomas Kneib and Ludwig Fahrmeir

Department of Statistics, University of Munich.

May 27, 2004

Abstract

In many practical situations, simple regression models suffer from the fact that the dependence of responses on covariates can not be sufficiently described by a purely parametric predictor. For example effects of continuous covariates may be nonlinear or complex interactions between covariates may be present. A specific problem of space-time data is that observations are in general spatially and/or temporally correlated. Moreover, unobserved heterogeneity between individuals or units may be present. While, in recent years, there has been a lot of work in this area dealing with univariate response models, only limited attention has been given to models for multicategorical space-time data. We propose a general class of structured additive regression models (STAR) for multicategorical responses, allowing for a flexible semiparametric predictor. This class includes models for multinomial responses with unordered categories as well as models for ordinal responses. Non-linear effects of continuous covariates, time trends and interactions between continuous covariates are modelled through Bayesian versions of penalized splines and flexible seasonal components. Spatial effects can be estimated based on Markov random fields, stationary Gaussian random fields or two-dimensional penalized splines.

We present our approach from a Bayesian perspective, allowing to treat all functions and effects within a unified general framework by assigning appropriate priors with different forms and degrees of smoothness. Inference is performed on the basis of a multicategorical linear mixed model representation. This can be viewed as posterior mode estimation and is closely related to penalized likelihood estimation in a frequentist setting. Variance components, corresponding to inverse smoothing parameters, are then estimated by using restricted maximum likelihood. Numerically efficient algorithms allow computations even for fairly large data sets. As a typical example we present results on an analysis of data from a forest health survey.

Key words: Multicategorical space-time data, generalized linear mixed models, restricted maximum likelihood, stationary Gaussian random fields, P-splines

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. These locations can be either exact locations, consisting of longitude and latitude, or locations on a (possibly irregular) spatial array.

In our application, we will analyze data from a forest health survey, where for several years the damage state of a population of trees is measured in ordered categories. In addition to a set of continuous and categorical covariates, the location of each tree is available on a lattice map. A typical approach to deal with such data, ignoring the spatial and temporal correlations, are parametric cumulative regression models, compare for example Fahrmeir and Tutz (2001), ch. 3. However, due to the space-time structure of the data, we have to take temporal as well as spatial correlations into account. Moreover, effects of continuous covariates may be non-linear or complex interactions between covariates might be present. Within a parametric modelling framework, it is virtually impossible to include these aspects.

In recent years, models for space-time data with univariate responses have gained considerable attention. Based on mixed model representations, Lin and Zhang (1999) used smoothing splines and random effects to model longitudinal data with responses from a univariate exponential family and Kammann and Wand (2003) introduced geoadditive models for Gaussian responses based on stationary Gaussian random fields and P-Splines. A more general empirical Bayes approach, extending generalized additive mixed models and geoadditive models, is presented in Fahrmeir et al. (2004). 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 multicategorical space-time data is rather limited (compare Fahrmeir and Lang (2001b) and Brezger and Lang (2003) for a notable exception based on latent Gaussian utilities and Markov Chain Monte Carlo simulation techniques). We propose a general class of structured additive regression models (STAR) for multicategorical responses, allowing for a flexible semiparametric predictor. This class includes models for multinomial responses with unordered categories as well as models for ordinal responses with ordered categories. Our approach is presented from a Bayesian perspective, allowing to treat all functions and effects within a unified general framework by assigning appropriate priors with different forms and degrees of smoothness. This fact greatly facilitates implementation since all effects are treated in a unified way conceptually and also allows to present formulae in a compact way. Note, however, that there is a very close connection to penalized likelihood estimation, with penalty terms in the frequentist setting corresponding to log-priors in the Bayesian approach.

Smooth effects of continuous covariates are modelled by P-splines, introduced by Eilers and Marx (1996) in a frequentist setting and transferred to a full Bayesian formulation by Lang and Brezger (2003). P-splines can also be used to model a flexible time trend. An alternative are more general autoregressive priors, including random walks or flexible seasonal components, to capture temporal correlations of a different form. Brezger and Lang (2003) extend P-splines to two dimensions using tensor products of one dimensional B-spline basis functions together with smoothness priors common in spatial statistics. This allows to model interactions between continuous covariates in a rather flexible way. Another, less flexible possibility to model interactions are varying coefficient models (Hastie and Tibshirani 1993). These are more commonly used if one of the interacting variables is categorical.

For the specification of the spatial effect we distinguish two different situations: The locations can be available exactly in terms of longitude and latitude or observations may be clustered in connected geographical regions. If exact locations are available, we propose to use two-dimensional P-splines to model the spatial effect. As an alternative we consider stationary Gaussian random fields, popular in geostatistics. If the geographical information is not given exactly, but observations can be clustered in connected geographical regions, Markov random fields allow the estimation of smooth spatial effects. Additional uncorrelated random effects may be incorporated as a surrogate for unobserved local small-area heterogeneity. Of course, random effects can also be used to deal with group or individual specific heterogeneity.

Inference for STAR models is performed on the basis of a multicategorical linear mixed model representation. In fact, any model based on smoothness priors or on a penalization approach can be rewritten as a variance components model, and all model components described above are of this type. The variance components, corresponding to inverse smoothing parameters in a frequentist approach, can then be estimated using mixed model methodology, especially restricted maximum likelihood, also termed marginal likelihood in the literature. Given estimates of the variance parameters, regression coefficients are estimated by a modified Fisher-scoring procedure. Since variance components are treated as unknown constants, our approach can be viewed as empirical Bayes/posterior mode estimation. Numerically efficient algorithms, developed in Fahrmeir et al. (2004), allow the computation of the estimates even for fairly large data sets.

Section 2 describes structured additive regression models for multicategorical data and the different model components in greater detail. Inference is presented in Section 3. In Sections 4 and 5 the performance of the approach is investigated through simulation studies and an application to the forest health data mentioned above. The conclusions in Section 6 give comments on directions of future research.

The methodology presented in this paper is implemented in BayesX, a public domain software package for Bayesian inference. The program is available at

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

2 Structured additive regression

2.1 Multicategorical response models

Regression models for a categorical response $Y \in \{1, \dots, k\}$ are mostly based on some latent response mechanism. Depending on the type of response and specific assumptions, various categorical regression models have been proposed, see, e.g., Agresti (1990), and Fahrmeir and Tutz (2001, ch. 3). Here we focus on the most popular models, but our concepts can be extended to other models.

For the case of a nominal response Y with unordered categories $1, \dots, k$ we consider the widely used multinomial logit model

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

where $\eta^{(r)}$ is a predictor depending on covariates and k is chosen as the reference category. For an ordered response Y we consider cumulative logit or probit models

$$P(Y \leq r) = F(\eta^{(r)}) \tag{2}$$

with $\eta^{(r)} = \theta^{(r)} - u'\gamma$ and linear predictor $u'\gamma$, F as the logistic or standard normal distribution function and ordered thresholds $\theta^{(1)} < \dots < \theta^{(q)}$.

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

$$\{Y_{it}, x_{it}, u_{it}, s_i\},\ \ i = 1, \ldots, n, t = 1, \ldots, T_i,$$

where s_i is the location or spatial index of individual i.

In the following we formulate structured additive response models from a Bayesian perspective, but we will also point out the close relationship to penalized likelihood approaches in a frequentist setting.

2.2 Observation models

Structured additive regression models extend the common linear predictors in (1) and (2) to more general semiparametric additive predictors. As an example, we consider models of increasing complexity for ordinal responses. The same extensions can be defined for nominal responses if all effects are assumed to be category-specific and the threshold is dropped from the predictor. At the end of the section we present a generic form of structured additive regression models for both ordinal and nominal responses, which comprises all submodels.

2.2.1 Space-time main effect model

For ordinal responses Y_{it}, the usual linear predictor in (2) can be extended to

$$\eta_{it}^{(r)} = \theta^{(r)} - [f_1(x_{it1}) + \ldots + f_l(x_{itl}) + f_{time}(t) + f_{spat}(s_i) + u_{it}'\gamma]. \tag{3}$$

Here, f_{time} and f_{spat} represent possibly nonlinear effects of time and space, $f_1, \ldots, f_l$ are unknown smooth functions of the continuous covariates $x_1, \ldots, x_l$, and $u'\gamma$ corresponds to the usual parametric linear part of the predictor. The functions $f_1, \ldots, f_l$ will be modelled as P-splines, see Section 2.3.1. Depending on the data, the effect of time may be split up into a trend and a seasonal component, i.e.

$$f_{time}(t) = f_{trend}(t) + f_{season}(t). \tag{4}$$

The trend function can be modelled by random walks or, more generally, by P-splines, and the seasonal component by an autoregressive process, see Section 2.3.1

In analogy the spatial effect can be split up into a spatially correlated (smooth) part f_{str} and a spatially uncorrelated (unsmooth) part f_{unstr}, i.e.

$$f_{spat}(s) = f_{str}(s) + f_{unstr}(s).$$

A rationale is that a spatial effect is usually a surrogate of many unobserved influential factors, some of them may obey a strong spatial structure and others may be present only locally. By estimating a structured and an unstructured component we aim at distinguishing between the two kinds of influential factors, see Besag et al. (1991). For the smooth spatial part we assume Markov random field priors, two-dimensional surface smoothers or stationary Gaussian fields, compare Section 2.3.2. For the uncorrelated part we may assume i.i.d. Gaussian random effects.

In the basic space-time model (3) all effects are population effects. For example, $f_{time}(t)$ is the population time trend, $f_j(x_j)$ is the nonlinear population effect of the covariate x_j, and γ is the vector of "fixed" population effects of the covariate vector u. From a frequentist point of view, these population effects are considered as deterministic. Whether or not a population effect is deterministic is debatable, however. From a Bayesian perspective all effects, including population effects and fixed effects, are interpreted as realizations of random variables or random functions, with appropriate priors assigned to them, see Section 2.3.

This Bayesian perspective also reveals more clearly that temporal or spatial correlation can be taken into account by suitable specification of f_{time} and f_{spat}, through correlated random effects with appropriate priors. Integrating out these random effects from the predictors, observations Y_{it} become marginally correlated.

For nominal responses Y_{it}, category-specific predictors $\eta_{it}^{(r)}$, in (1) are defined in complete analogy to (3), introducing category-specific functions $f_j^{(r)}$, $j = 1, \ldots, l, time, spat$ and parameters $\gamma^{(r)}$.

2.2.2 Models with individual-specific effects

Individual-specific departures from the population effects in model (3) can be specified by introducing additional random effects in the predictors as in generalized linear mixed models. Then, (3) is extended to

$$\eta_{it}^{(r)} = \theta^{(r)} - [f_1(x_{it1}) + \ldots + u_{it}'\gamma + w_{it}'b_i] \tag{5}$$

where w_{it} is a design vector and b_i is a vector of i.i.d. random effects. For the special case $w_{it} = 1$ the random intercept b_i is often introduced to model unobserved heterogeneity. More generally, individual-specific effects b_i and appropriate design vectors w_{it} can be used to model individual-specific departures from population effects as well as correlations of repeated observations. For example, assume that the population time trend is approximated by a linear combination $f(t) = \sum \beta_j B_j(t)$ of basis functions, such as the first terms of a Taylor or Fourier expansion, or of a spline basis. Individual-specific departures can then be modelled through the random effects part of the predictor, i.e. $f_i(t) = \sum b_{ji} B_j(t)$, where b_{ji} are i.i.d. random effects, and the design variables w_{itj} are equal to $B_j(t)$. This is in analogy to standard parametric mixed models with, e.g., a linear time trend $\beta_0 + \beta_1 t$ and individual specific random departures $b_{0i} + b_{1i}t$ from this trend.

2.2.3 Models with interactions

Interactions between variables in the main effect model (3) can be incorporated in various ways. Including an interaction between a categorical or continuous covariate u, say, and time of the form

$$\eta_{it}^{(r)} = \ldots + f_{time}(t) + g(t)u_{it} + \ldots$$

where g is a smooth function, leads to a model with a time-varying effect $g(t)$ of the covariate u. Generally models with interaction terms of the form $g(x)u$ and a continuous

effect modifier x, are called varying coefficient models. We also allow for models with space-varying effects $g(s)u$.
For two continuous covariates x_1 and x_2, say, an interaction effect $f_{1|2}(x_1, x_2)$ may be added to the main effects, leading to predictors of the form

$$\eta_{it}^{(r)} = \ldots + f_1(x_{it1}) + f_2(x_{it2}) + f_{1|2}(x_{it1}, x_{it2}) + \ldots$$

The two-dimensional interaction surface $f_{1|2}$ can be modelled e.g. by two-dimensional P-splines, see Section 2.3.4

2.2.4 Structured additive predictors in generic form

All predictors presented in Sections 2.2.1 to 2.2.3 can be cast into the generic form

$$\eta_{it}^{(r)} = \theta^{(r)} - [f_1(v_{it1}) + \ldots + f_p(v_{itp}) + u_{it}'\gamma] \tag{6}$$

where $f_1, \ldots, f_p$ are different types of functions and $v_1, \ldots, v_p$ are different types of covariates or design variables. For example, $f(v) = g(x)u$ with $v = (x, u)$, represents a varying coefficient term, $f(v = (x_1, x_2))$ denotes an interaction surface, and random effect terms $w_{it}'b_i$ are special linear functions.
The generic form (6) is useful for developing and implementing the methodology in unified and compact form. It turns out in Section 2.3, that we will always be able to express $f_j(v_{itj})$ as the product of a design vector z_{itj} and a (possibly highdimensional) vector β_j of unknown parameters. So we can rewrite the predictor (6) as

$$\eta_{it}^{(r)} = \theta^{(r)} - [z_{it1}'\beta_1 + \ldots + z_{itp}'\beta_p + u_{it}'\gamma]. \tag{7}$$

Similarly, for the multinomial logit model (1), the predictor $\eta^{(r)}$ can be written as

$$\eta_{it}^{(r)} = z_{it1}'\beta_1^{(r)} + \ldots + z_{itp}'\beta_p^{(r)} + u_{it}'\gamma^{(r)}. \tag{8}$$

In the first case the vector of all unknown regression coefficients is given by $\beta = (\theta^{(1)}, \ldots, \theta^{(q)}, \beta_1', \ldots, \beta_p', \gamma')'$, while in the latter case it is $\beta = (\beta_1^{(1)\prime}, \gamma^{(1)\prime}, \ldots, \beta_1^{(q)\prime}, \gamma^{(q)\prime})'$.
Finally, we make the usual conditional independence assumption: Given unknown functions and parameters, the observations Y_{it} are conditionally independent. Together with specific multinomial distributions defined by the logit or probit models (1) and (2), the likelihood of all observations is then uniquely defined as

$$L(\beta) = \prod_{i,t} f(Y_{it}|\beta), \tag{9}$$

and the log-likelihood is

$$l(\beta) = \sum_{i,t} \log f(Y_{it}|\beta). \tag{10}$$

2.3 Prior assumptions on functions and parameters

For simplicity we will restrict the discussion of priors to the case of models for ordinal responses. If multinomial models are considered, all effects have to be treated as category specific and the additional index r has to be added to the vector of regression coefficients, but concepts remain exactly the same. So it should be kept in mind that all results in the following sections apply also for nominal response models. However, no deeper insight is gained if this case is treated separately, only formulae would become more cumbersome.

From a Bayesian point of view, the unknown functions $f_1, \ldots, f_p$ in (6), more exactly corresponding vectors of function evaluations, and the fixed effects parameters γ are considered as random variables and must be supplemented by appropriate prior assumptions.

Throughout the paper we will assume independent diffuse priors $p(\gamma) \propto const$ for the fixed effects parameters γ and the thresholds $\theta^{(r)}$ in cumulative models.

A prior for a function f_j is now defined by specifying a suitable design vector z_{itj} and a prior distribution for the vector β_j of unknown parameters. All specific priors defined in the following subsections have the general form

$$p(\beta_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right), \tag{11}$$

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

For given or known variance parameters, Bayesian inference is then based on the posterior

$$p(\beta|Y) \propto L(\beta) \prod_j \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right). \tag{12}$$

Posterior mode estimates $\hat{\beta}$ for β are obtained by maximizing the right hand side, or, taking logarithms, the penalized log-likelihood

$$l_{pen}(\beta|Y) = l(\beta) - \frac{1}{2}\sum_j \lambda_j \beta_j' K_j \beta_j \tag{13}$$

with smoothing parameters $\lambda_j = 1/\tau_j^2$ and penalty terms $\beta_j' K_j \beta_j$.

From a frequentist point of view, we could start directly from the the penalized likelihood, and penalized likelihood estimates obtained by maximizing (13) are identical to posterior mode estimates. This shows the close connection between both approaches for inference. In particular, all priors below lead to specific penalty terms obtained from the log-priors.

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 as an unknown constant. As an alternative to data driven determination of τ_j^2, e.g. by crossvalidation, the Bayesian point of view opens the way to estimate τ_j^2 by (restricted) maximum likelihood, see Section 3.

2.3.1 Priors for continuous covariates and time scales

Several alternatives have been recently proposed to specify smoothness priors for continuous covariates or time trends. These are *random walk priors* or more generally *autoregressive priors* (see Fahrmeir and Lang (2001a) and Fahrmeir and Lang (2001b)), *Bayesian P-splines* (Lang and Brezger (2003)) and *Bayesian smoothing splines* (Hastie and Tibshirani (2000)). In the following we will focus on P-splines. Commonly used random walk priors for smooth time trends, popular in state space models, result as a special case: they are P-splines of degree 0.

The approach assumes that an unknown smooth function f_j of a covariate x_j can be approximated by a polynomial spline of degree l defined on a set of equally spaced knots $x_j^{min} = \kappa_0 < \kappa_1 < \cdots < \kappa_{d-1} < \kappa_d = x_j^{max}$ within the domain of x_j. Such a spline can be written in terms of a linear combination of $M_j = d + l$ B-spline basis functions B_m, i.e.

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

Here $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_j})'$ corresponds to the vector of unknown regression coefficients. The M_j-dimensional design vector z_{itj} in (7) or (8) consists of the basis functions evaluated at the observation x_{itj}, i.e. $z_{itj} = (B_1(x_{itj}), \ldots, B_{M_j}(x_{itj}))'$.

The crucial point is the choice of the number of knots. For a small number of knots, the resulting spline may not be flexible enough to capture the variability of the data. For a large number of knots, estimated curves tend to overfit the data and, as a result, too rough functions are obtained. As a remedy Eilers and Marx (1996) suggest a moderately large number of equally spaced knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on first or second order differences of adjacent B-Spline coefficients to guarantee sufficient smoothness of the fitted curves. This leads to penalized likelihood estimation with penalty terms

$$P(\lambda_j) = \frac{1}{2}\lambda_j \sum_{m=k+1}^{M_j} (\Delta^k \beta_{jm})^2 = \frac{1}{2}\lambda_j \beta_j' K_j \beta_j, \quad k = 1, 2 \tag{14}$$

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

$$\beta_{jm} = \beta_{j,m-1} + u_{jm} \quad \text{or} \quad \beta_{jm} = 2\beta_{j,m-1} - \beta_{j,m-2} + u_{jm} \tag{15}$$

with Gaussian errors $u_{jm} \sim N(0, \tau_j^2)$ and diffuse priors $p(\beta_{j1}) \propto const$, or $p(\beta_{j1})$ and $p(\beta_{j2}) \propto const$, for initial values, respectively. The joint distribution of the regression parameters β_j is easily computed as a product of conditional densities defined by (15) and can be brought into the general form (11). More details about Bayesian P-splines can be found in Lang and Brezger (2003).

Simple first or second order random walks $\Delta^k \beta_t = u_t$ are often used to model time trends $f(t) =: \beta_t$. They can be regarded as P-splines of degree $l = 0$ and are therefore a special case. More general autoregressive process priors than the random walk models (15) may be useful, for example to model flexible seasonal patterns, see Fahrmeir and Lang (2001a). A flexible seasonal component $f_{season}(t) =: \beta_t$ with period *per* can be defined by

$$\beta_t = -\sum_{j=1}^{per-1} \beta_{t-j} + u_t$$

and once again diffuse priors for initial values and errors $u_t \sim N(0, \tau_{season}^2)$.

2.3.2 Priors for smooth spatial effects

For the specification of the smooth spatial effect f_{str} we can distinguish two different situations: Locations can be available exactly in terms of longitude and latitude or observations may be clustered in connected geographical regions.

Markov random field priors

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

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

$$\beta_s | \beta_{s'},\ s' \neq s, \tau^2_{str} \sim N\left(\frac{1}{N_s}\sum_{s' \in \partial_s} \beta_{s'}, \frac{\tau^2_{str}}{N_s}\right), \tag{16}$$

where N_s is the number of adjacent sites and $s' \in \partial_s$ denotes that site s' is a neighbor of site s. Thus the (conditional) mean of β_s is an unweighted average of function evaluations of neighboring sites. The prior is a direct generalization of a first order random walk to two dimensions and is called a Markov random field (MRF). More general priors can be based on weighted averages rather than an unweighted average as in (16), e.g. with weights being proportional to the distance of neighboring sites to site s. In terms of weights $w_{ss'}$ a general spatial prior can be defined as

$$\beta_s | \beta_{s'},\ s' \neq s, \tau^2_{str} \sim N\left(\sum_{s' \in \partial_s} \frac{w_{ss'}}{w_{s+}} \beta_{s'}, \frac{\tau^2_{str}}{w_{s+}}\right), \tag{17}$$

where $+$ denotes summation over the missing subscript.

For both priors the S-dimensional design vector $z_{str} = (0, \dots, 1, \dots, 0)'$ is now a 0/1 incidence vector. Its value in the s-th row is 1 if the corresponding observation is located in site or region s, and zero otherwise. The $S \times S$ penalty matrix K is given by $k_{ss} = w_{s+}$ and $k_{ss'} = -w_{ss'}$ if s and s' are neighbors and zero otherwise. Therefore K has the form of an adjacency matrix.

Gaussian random field priors

If exact locations $s = (s_x, s_y)$ are available, we can use two-dimensional surface estimators to model spatial effects. One option are two-dimensional P-splines, see Section 2.3.4. Another option are Gaussian random field (GRF) priors, originating from geostatistics. They can be seen as two-dimensional surface smoothers based on special basis functions, e.g. radial basis functions, and have been used by Kammann and Wand (2003) to model the spatial component in Gaussian regression models. The spatial component $f_{str}(s) =: \beta_s$ is then assumed to follow a zero mean stationary Gaussian random field $\{\beta_s : s \in \mathbb{R}^2\}$ with variance τ^2_{str} and isotropic correlation function $\mathrm{cov}(\beta_s, \beta_{s+h}) = C(||h||)$. 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 $s \in \{1, \dots, S\}$ of sites as in image analysis or in our application to forest health data, the prior for $\beta_j = (\beta_1, \dots, \beta_s)'$ is of the general form (11) with $K = C^{-1}$ and

$$C[i,j] = C(||s_i - s_j||), 1 \leq i, j \leq n.$$

The design vector z_{str} 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 (e.g. 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, e.g.

$$C(r; \rho) = \exp(-|r|)(1 + |r|) \tag{18}$$

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 or may be estimated with variance components by restricted maximum likelihood. A simple rule to choose ρ is

$$\hat{\rho} = \max_{i,j} ||s_i - s_j||/c \tag{19}$$

ensuring scale invariance. The constant $c > 0$ is chosen in such a way, that $C(c)$ is small, e.g. 0.001. 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, e.g. 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. Furthermore GRFs may be approximated by MRFs, see Rue and Tjelmeland (2002). In general, it is not clear which of the different approaches leads to the "best" fits. For data observed on a discrete lattice, MRFs seem to be most appropriate. If the exact locations are available, surface estimators may be more natural, particularly because predictions for unobserved locations are available. However, in some situations surface estimators lead to an improved fit compared to MRF's even for discrete lattices and vice versa. A general approach that can handle both situations is given by Müller et al. (1997).

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 S 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 usually is close to or equal to the sample size. So the number of regression coefficients used to describe a 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 Gaussian random fields. Note first, that we can define GRFs equivalently based on a design vector z_{str} with entries $z_{str} = (C(||s - s_1||), \ldots, C(||s - s_n||))'$ 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 $\mathcal{C}$. These knots can be chosen to be "representative" for the set of distinct locations based on a space filling algorithm. Therefore consider the distance measure

$$d(s, \mathcal{D}) = \left(\sum_{\kappa \in \mathcal{D}} ||s - \kappa||^p \right)^{\frac{1}{p}},$$

with $p < 0$, between any location $s \in \mathcal{C}$ and a possible set of knots $\mathcal{D}$. Obviously this distance measure is zero for all knots. Using a simple swapping algorithm to minimize the overall coverage criterion

$$\left(\sum_{s \in \mathcal{C}} d(s, \mathcal{D})^q \right)^{\frac{1}{q}}$$

with $q > 0$ (compare Johnson et al. (1990) and Nychka and Saltzman (1998) for details) yields an optimal set of knots $\mathcal{D}$. Based on these knots we define the approximation $f_{str}(s) = z_{str}'\beta$ with the M-dimensional design vector $z_{str} = (C(||s - \kappa_1||), \ldots, C(||s - \kappa_M||)'$, penalty matrix $K = C$ and $C[i,j] = C(||\kappa_i - \kappa_j||)$. The number of knots M allows us to control the trade-off between the accuracy of the approximation (M close to the sample size) and the numerical simplification (M small).

2.3.3 Group indicators, individual-specific effects and unstructured spatial effects

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

$$\beta_c \sim N(0, \tau^2), \quad c = 1, \ldots, C. \tag{20}$$

The design vector z is again a C-dimensional $0/1$ incidence vector and the penalty matrix is the identity matrix, i.e. $K = I$. From a classical perspective, (20) defines i.i.d. *random effects.*

For longitudinal data, clusters are often defined by the repeated observations on individuals. Then the cluster index c is the individual index $i \in \{1, \ldots, n\}$, and β_i are i.i.d. individual-specific effects.

Another special case are spatial clusters. Identifying a cluster index c with the spatial index s, the unstructured spatial effects $f_{unstr}(s) =: \beta_s$ are assumed to be i.i.d. random effects $\beta_s \sim N(0, \tau^2_{unstr})$.

2.3.4 Interactions

Varying coefficient models are commonly used to incorporate interactions of the form $g(x)u$ between a binary variable u and a continuous covariate x, which may also be a time scale. We also allow models with space-varying effects $g(s)u$. For the smooth nonlinear functions g we assume the same priors defined already in Sections 2.3.1 and 2.3.2.

Suppose now that both interacting covariates are metrical. In this case, a flexible approach for modelling interactions can be based on (nonparametric) two dimensional surface fitting. Here, we follow an approach based on two dimensional P-splines described in more detail in Lang and Brezger (2003). The assumption is that the unknown surface $f_j(x_{j_1}, x_{j_2})$ can be approximated by the tensor product of two one dimensional B-splines, i.e.

$$f_j(x_{j_1}, x_{j_2}) = \sum_{m_1=1}^{M_j} \sum_{m_2=1}^{M_j} \beta_{j,m_1 m_2} B_{j,m_1}(x_{j_1}) B_{j,m_2}(x_{j_2}).$$

Similar to one-dimensional P-splines, the M_j^2-dimensional design vector z_j is composed of products of basis functions. The coefficients $\beta_j = (\beta_{j,11}, \ldots, \beta_{j,M_j M_j})'$ are defined on a two-dimensional regular array of knots in the (x_{j_1}, x_{j_2})-plane. Following the idea of one-dimensional P-splines, we assign two-dimensional random walk priors to enforce smoothness of the surface. These priors are a special case of MRF prior (16) for a regular lattice, consisting of the knots in the (x_{j_1}, x_{j_2})-plane, and with the four next neighbors s' of a knot s defining the neighborhood δ_s.

3 Inference

Inference in structured additive regression models is not performed on the basis of the original parameterization in (7) or (8), since there is no easy rule on how to choose the variance parameters. For univariate responses, a popular idea to get around this 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 five years (e.g. Fahrmeir et al. (2004), Kammann and Wand (2003), Ruppert et al. (2003), Wand (2003) or Lin and Zhang (1999)). In mixed model representation we obtain a variance components model, and techniques for estimating the variance parameters are already available (at least for univariate response). Probably the most common approach is to estimate them via restricted maximum likelihood, also termed marginal likelihood in the literature. In the following we will extend this approach to multicategorical STAR models. Before discussing the estimation of regression coefficients and variance parameters in a multicategorical mixed model in detail (Section 3.2), we will first show how to rewrite our models as variance components models.

3.1 Mixed model representation

To rewrite the models in (7) and (8) as mixed models, we first take a closer look on the general prior (11). To simplify notation, we will again discuss only the case of ordinal responses explicitly. For nominal responses results are summarized briefly at the end of the section.

Prior (11) specifies a multivariate Gaussian distribution for the parameter vector β_j. However, in most cases the precision matrix K_j is rank deficient and therefore (11) is an improper distribution. 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 function f_j that is not penalized by K_j, β_j^{pen} captures the deviation from this unpenalized part. The dimensions of both vectors depend on the rank of the penalty matrix K_j. If K_j had full rank, the unpenalized part would vanish completely and if we choose $\beta_j^{pen} = K_j^{1/2}\beta_j$ 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 j-th 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

$$\beta_j = Z_j^{unp}\beta_j^{unp} + Z_j^{pen}\beta_j^{pen} \tag{21}$$

with a $d_j \times (d_j - rk_j)$ matrix Z_j^{unp} and a $d_j \times rk_j$ matrix Z_j^{pen}. The decomposition of β_j leads to a similar decomposition for $f_j(v_{itj})$ into a penalized and an unpenalized part:

$$f_j(v_{itj}) = z_{itj}' Z_j^{unp}\beta_j^{unp} + z_{itj}' Z_j^{pen}\beta_j^{pen} = \tilde{z}_{itj}^{unp\prime}\beta_j^{unp} + \tilde{z}_{itj}^{pen\prime}\beta_j^{pen}. \tag{22}$$

Requirements for decomposition (21) are:

(i) The composed matrix $(Z_j^{unp}\ Z_j^{pen})$ has full rank to make the transformation in (21) a one-to-one transformation. This also implies that both Z_j^{unp} and Z_j^{pen} have full column rank.

(ii) Z_j^{unp} and Z_j^{pen} are orthogonal, i.e. $Z_j^{unp\prime} Z_j^{pen} = 0$.

(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 the matrices defining (21) can be obtained as follows: Z_j^{unp} contains a $d_j - k_j$ dimensional basis of the null space of K_j. Therefore requirement (iii) is automatically fulfilled. Z_j^{pen} can be obtained by $Z_j^{pen} = L_j(L_j'L_j)^{-1}$ where the full column rank $d_j \times k_j$ matrix L_j is determined by the decomposition of the penalty matrix K_j into $K_j = L_j L_j'$. This ensures requirements (i) and (iv). If we choose L_j such that $L_j' Z_j^{unp} = 0$ and $Z_j^{unp} L_j' = 0$ hold, we finally obtain requirement (ii). The decomposition $K_j = L_j L_j'$ of the penalty matrix can be based on the spectral decomposition $K_j = \Gamma_j \Omega_j \Gamma_j'$. The $(k_j \times k_j)$ diagonal matrix Ω_j contains the positive eigenvalues ω_{jm}, $m = 1, \ldots, k_j$, of K_j in descending order, i.e. $\Omega_j = \text{diag}(\omega_{j1}, \ldots, \omega_{j,rk_j})$. Γ_j is a $(d_j \times k_j)$ orthogonal matrix of the corresponding eigenvectors. From the spectral decomposition we can choose $L_j = \Gamma_j \Omega_j^{1/2}$. Note, that the factor L_j is not unique and in many cases numerical superior factorizations exist.

Although the previous paragraph may sound rather technical, the decomposition is quite intuitive in most cases, as we will show for some specific examples at the end of this section.

We finally obtain

$$p(\beta_{jm}^{unp}) \propto const, \qquad m = 1, \ldots, d_j - k_j$$

and

$$\beta_j^{pen} \sim N(0, \tau_j^2 I). \tag{23}$$

For ordinal responses this allows us to rewrite the additive predictor (7) as

$$\eta_{it}^{(r)} = \theta^{(r)} - \left[\sum_{j=1}^{p} z_{itj}'\beta_j + u_{it}'\gamma\right] = \theta^{(r)} - \left[\sum_{j=1}^{p}(\tilde{z}_{itj}^{unp\prime}\beta_j^{unp} + \tilde{z}_{itj}^{pen\prime}\beta_j^{pen}) + u_{it}'\gamma\right]$$

For nominal responses we obtain an equivalent representation for the predictor in (8):

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

where $\tilde{z}_{itj}^{unp}$ and $\tilde{z}_{itj}^{pen}$ are constructed in complete analogy to the ordinal case.

In both cases the structured additive regression model can now be understood as a multicategorical GLMM with fixed effects β_j^{unp} and $\beta_j^{unp(r)}$, respectively. The random effects β_j^{pen} have distribution $N(0, \Lambda_j)$ with $\Lambda_j = diag(\tau_j^2, \ldots, \tau_j^2)$ for ordinal responses and $\beta_j^{pen(r)} \sim N(0, \Lambda_j^{(r)})$ for nominal responses. Hence, we can utilize methodology for multicategorical GLMMs for simultaneous estimation of the functions $f_j(v_{itj})$ and the variance parameters, see the next section.

Let us now discuss some special cases of (21). As described above, for parameters β_j with proper prior, decomposition (21) is just some kind of standardization such that elements of β_j^{pen} are independent and have a common variance. This case includes stationary Gaussian fields and i.i.d. random effects. For P-splines, the decomposition yields an unpenalized part $\tilde{z}_{itj}^{unp\prime}\beta_j^{unp}$ representing a polynom of degree $k-1$. Therefore the unpenalized part for a P-spline with second order random walk prior is a straight line and for a P-Spline with first order random walk prior it is a horizontal line. The same statement holds for random walk priors themselves. In both cases the factor L_j derived on the basis of a spectral decomposition may be replaced by $L_j = D'$ where D denotes the difference matrix defining the penalty matrix for the corresponding function.

Considering flexible seasonal components, we obtain a fixed seasonal effect as the unpenalized part of $f_j(v_{itj})$. The factor L_j can be derived from a general decomposition of K_j for autoregressive processes, compare e.g. Knorr-Held (1996). In all other cases discussed in Section 2.3 the unpenalized part is simply a constant effect. So we have the general result, that the fixed part $\tilde{z}_{itj}^{unp\prime}\beta_j^{unp}$ equals what we obtain for $f_j(v_{itj})$ if the corresponding smoothing parameter $1/\tau_j^2$ goes to infinity.

The mixed model representation also allows for a different perspective on the identification problem inherent to nonparametric regression models. For all model components with improper prior, $\tilde{z}_{itj}^{unp}$ contains a one representing the mean level of the corresponding function. Therefore, provided that there is at least one such term and that we have an intercept included in the model, which is always the case for ordinal responses, where the intercept is given by the threshold, we observe linear dependencies in the predictor $\eta_{it}^{(r)}$. To get around this, we delete all the ones from the vectors $\tilde{z}_{itj}^{unp}$ which has a similar effect as centering the functions $f_j(v_{itj})$.

3.2 Restricted maximum likelihood inference for multicategorical models

Representing multicategorical structured additive regression models as multicategorical mixed models significantly reduces the complexity of the estimation problem. This is mainly due to the fact that all special cases of structured additive regression can be combined into one single model allowing to apply the same estimation procedure to all estimation problems. To describe algorithms in compact matrix notation we rewrite Y_{it} as a vector of dummy variables $y_{it} = (y_{it}^{(1)}, \dots, y_{it}^{(q)})'$ with

$$y_{it}^{(r)} = \begin{cases} 1 & \text{if } Y_{it} = r, \\ 0 & \text{else.} \end{cases}$$

Therefore we have

$$P(Y_{it} = r) = P(y_{it}^{(r)} = 1) = \pi_{it}^{(r)}.$$

The probabilities $\pi_{it} = (\pi_{it}^{(1)}, \dots, \pi_{it}^{(q)})'$ are connected to the linear predictors $\eta_{it} = (\eta_{it}^{(1)}, \dots, \eta_{it}^{(q)})'$ by the (multivariate) response function $h : \mathbb{R}^q \to [0,1]^q$ via $h(\eta_{it}) = (h^{(1)}(\eta_{it}), \dots, h^{(q)}(\eta_{it}))' = E(y_{it}|\eta_{it}) = \pi_{it}$. The specific form of the response function is derived from expressions (1) and (2), compare Fahrmeir and Tutz (2001), ch. 3. To write the q-dimensional vector η_{it} in matrix notation, define the matrices

$$Q_{it} = \begin{pmatrix} z_{it}^{unp} & & 0 \\ & \ddots & \\ 0 & & z_{it}^{unp} \end{pmatrix} \qquad P_{it} = \begin{pmatrix} z_{it}^{pen} & & 0 \\ & \ddots & \\ 0 & & z_{it}^{pen} \end{pmatrix}$$

for nominal responses and

$$Q_{it} = \begin{pmatrix} 1 & & & z_{it}^{unp} \\ & \ddots & & \vdots \\ & & 1 & z_{it}^{unp} \end{pmatrix} \qquad P_{it} = \begin{pmatrix} z_{it}^{pen} \\ \vdots \\ z_{it}^{pen} \end{pmatrix}$$

for ordinal responses. In both cases the design vectors $\tilde{z}_{it}^{unp}$ and $\tilde{z}_{it}^{pen}$ are composed as follows:

$$\tilde{z}_{it}^{unp} = (z_{it1}^{unp\prime}, \ldots, z_{itp}^{unp\prime}, u_{it}')' \qquad \tilde{z}_{it}^{pen} = (z_{it1}^{pen\prime}, \ldots, z_{itp}^{pen\prime})'$$

Based on these definitions we have $\eta_{it} = \tilde{U}_{it}\beta^{unp} + X_{it}\beta^{pen}$ for both types of multicategorical models. The structure of the vectors of regression coefficients β^{unp} and β^{pen} depends on the specific model and is given by

$$\begin{aligned}
\beta^{unp} &= (\beta_1^{unp(1)\prime}, \ldots, \beta_p^{unp(1)\prime}, \gamma^{(1)\prime}, \ldots, \beta_1^{unp(q)\prime}, \ldots, \beta_p^{unp(q)\prime}, \gamma^{(q)\prime})' \\
\beta^{pen} &= (\beta_1^{pen(1)\prime}, \ldots, \beta_p^{pen(1)\prime}, \ldots, \beta_1^{pen(q)\prime}, \ldots, \beta_p^{pen(q)\prime})'
\end{aligned}$$

for nominal responses and

$$\begin{aligned}
\beta^{unp} &= (\theta^{(1)}, \ldots, \theta^{(q)}, \beta_1^{unp\prime}, \ldots, \beta_p^{unp\prime}, \gamma')' \\
\beta^{pen} &= (\beta_1^{pen\prime}, \ldots, \beta_p^{pen\prime})'
\end{aligned}$$

for ordinal responses. The composed vector of random effects β^{pen} now follows a multivariate Gaussian distribution defined by (23), i.e. $\beta^{pen} \sim N(0, \Lambda)$ with

$$\Lambda = \text{blockdiag}(\Lambda_1^{(1)}, \ldots, \Lambda_p^{(1)}, \ldots, \Lambda_1^{(q)}, \ldots, \Lambda_p^{(q)})$$

for nominal responses and

$$\Lambda = \text{blockdiag}(\Lambda_1, \ldots, \Lambda_p)$$

for ordinal responses.

Finally, we define the stacked vectors and matrices

$$y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{nT_n} \end{pmatrix} \quad \pi = \begin{pmatrix} \pi_{11} \\ \vdots \\ \pi_{nT_n} \end{pmatrix} \quad \eta = \begin{pmatrix} \eta_{11} \\ \vdots \\ \eta_{nT_n} \end{pmatrix} \quad P = \begin{pmatrix} P_{11} \\ \vdots \\ P_{nT_n} \end{pmatrix} \quad Q = \begin{pmatrix} Q_{11} \\ \vdots \\ Q_{nT_n} \end{pmatrix}$$

resulting in

$$\eta = Q\beta^{unp} + P\beta^{pen}$$

and $E(y|\eta) = \pi$. Based on these definitions we are now able to describe estimation in a compact form. Estimating multicategorical mixed models can be performed in largely two steps: Alternately the regression coefficients are updated given the current values of the variance parameters and vice versa.

Posterior mode estimates for the regression coefficients β^{unp} and β^{pen} given the variance parameters in Λ are obtained by maximizing the posterior

$$p(\beta^{unp}, \beta^{pen}|y) \propto L(\beta^{unp}, \beta^{pen})p(\beta^{unp})p(\beta^{pen}),$$

where $L(\beta^{unp}, \beta^{pen})$ denotes the likelihood of the model, which in fact equals the likelihood in (9). The special form of the likelihood depends on the specific model (ordinal or nominal responses) and the choice of the response function. Equivalently we can maximize the log-posterior. Utilizing the flat prior of β^{unp} we obtain

$$l_{pen}(\beta^{unp}, \beta^{pen}) = l(\beta^{unp}, \beta^{pen}) - \frac{1}{2}\beta^{pen\prime}\Lambda^{-1}\beta^{pen} \tag{24}$$

to be maximized with respect to β^{unp} and β^{pen}. Note again, that (24) has the form of a penalized likelihood and therefore – for given variances – posterior mode estimates and penalized likelihood estimates coincide.

In principle, maximization of (24) is carried out through a Fisher scoring type algorithm (compare also Fahrmeir and Tutz (2001), chapter 3). Similar to estimation in GLMs, the Fisher scoring algorithm can be rewritten as iteratively weighted least squares (IWLS), yielding the following system of equations

$$\begin{pmatrix} Q'WQ & Q'WP \\ P'WQ & P'WP+\Lambda^{-1} \end{pmatrix} \begin{pmatrix} \beta^{unp} \\ \beta^{pen} \end{pmatrix} = \begin{pmatrix} Q'W\tilde{y} \\ P'W\tilde{y} \end{pmatrix}. \tag{25}$$

to be solved to obtain updated estimates. The main differences between IWLS for univariate GLMs and (25) are the special structure of the design matrices Q and P and the fact that the weight matrix W is no longer diagonal. Instead $W = D\Sigma^{-1}D$ has a block diagonal structure defined by the block diagonal matrices $D = \text{blockdiag}(D_{11} \ \dots \ D_{nT})$ and $\Sigma = \text{blockdiag}(\Sigma_{11} \ \dots \ \Sigma_{nT})$ and the $q \times q$ matrices

$$D_{it} = \frac{\partial h(\eta_{it})}{\partial \eta} = \begin{pmatrix} \frac{\partial h^{(1)}(\eta_{it})}{\partial \eta^{(1)}} & \dots & \frac{\partial h^{(q)}(\eta_{it})}{\partial \eta^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial h^{(1)}(\eta_{it})}{\partial \eta^{(q)}} & \dots & \frac{\partial h^{(q)}(\eta_{it})}{\partial \eta^{(q)}} \end{pmatrix}$$

and

$$\Sigma_{it} = \text{cov}(y_{it}) = \begin{pmatrix} \pi_{it}^{(1)}(1-\pi_{it}^{(1)}) & -\pi_{it}^{(1)}\pi_{it}^{(2)} & \dots & -\pi_{it}^{(1)}\pi_{it}^{(q)} \\ -\pi_{it}^{(1)}\pi_{it}^{(2)} & \ddots & & \vdots \\ \vdots & & \ddots & -\pi_{it}^{(q-1)}\pi_{it}^{(q)} \\ -\pi_{it}^{(1)}\pi_{it}^{(q)} & \dots & -\pi_{it}^{(q-1)}\pi_{it}^{(q)} & \pi_{it}^{(q)}(1-\pi_{it}^{(q)}) \end{pmatrix}$$

The working observations $\tilde{y}$ are defined by

$$\tilde{y} = \hat{\eta} + (D^{-1})'(y-\pi).$$

Note, that iteratively solving the system of equations (25) is equivalent to approximating the likelihood $L(\beta^{unp}, \beta^{pen})$ of a multinomial distribution with the likelihood of a multivariate Gaussian distribution having an iteratively reweighted covariance matrix W^{-1}. This approximation will also be used to obtain estimates of the variance parameters.

In Gaussian mixed models, a common way to estimate the variance parameters is maximum likelihood. Here, estimates are defined to be the maximizers of the likelihood $L(\beta^{unp}, \Lambda) = \int L(\beta^{unp}, \beta^{pen}, \Lambda) d\beta^{pen}$. For Gaussian responses this likelihood has a closed form, which can be maximized with respect to β^{unp} and the variance parameters iteratively. However, maximum likelihood estimates of the variance parameters do not take into account the loss in degrees of freedom caused by the estimation of β^{unp}. To get around this, Patterson and Thompson (1971) introduced restricted maximum likelihood estimates based on error contrasts of the original data. The distribution of these error contrasts does no longer depend on β^{unp}.

Unfortunately this approach can not be extended to more general responses directly, since in these cases no such error contrasts are available. But, as Harville (1974) showed, the concept of restricted maximum likelihood is equivalent to maximizing the marginal likelihood

$$L^*(\Lambda) = \int L(\beta^{unp}, \beta^{pen}, \Lambda) d\beta^{pen} d\beta^{unp}. \tag{26}$$

This allows to extend REML estimation to generalized mixed models and even to multicategorical mixed models. Because of (26), REML is also termed marginal likelihood in the literature.

Since direct evaluation of the integral in (26) is not possible in general, we use a quadratic approximation to $L(\beta^{unp}, \beta^{pen}, \Lambda)$, which is in fact equivalent to the approximation made in IWLS. This approximation results in the restricted log-likelihood

$$l^*(\Lambda) \approx -\frac{1}{2}\log(|V|) - \frac{1}{2}\log(|Q'V^{-1}Q|) - \frac{1}{2}(\tilde{y} - Q\hat{\beta}^{unp})'V^{-1}(\tilde{y} - Q\hat{\beta}^{unp}), \qquad (27)$$

where $V = W^{-1} + P'\Lambda P$ is an approximation to the marginal covariance of $\tilde{y}|\beta^{pen}$. Maximization of (27) 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 also be used within a multicategorical setting. One should however keep in mind that the weight matrix W is no longer diagonal but blockdiagonal.

Now we are able to define estimates for the function f_j based on $\hat{\beta}^{unp}$ and $\hat{\beta}^{pen}$. Applying (22) to the estimates yields

$$\hat{f}_j(v_{itj}) = z_{itj}^{unp\prime}\hat{\beta}_j^{unp} + z_{itj}^{pen\prime}\hat{\beta}_j^{pen}.$$

This also forms the basis for constructing credible intervals for $\hat{f}_j$. Since the covariance matrix of the regression coefficients $\hat{\beta}^{unp}$ and $\hat{\beta}^{pen}$ is given by H^{-1}, where H denotes the coefficient matrix on the left hand side of (25), we get

$$\text{se}(\hat{f}_j(v_{itj})) = \sqrt{(z_{itj}^{unp\prime}\ z_{itj}^{pen\prime})\text{Cov}\left((\hat{\beta}_j^{unp})'\ (\hat{\beta}_j^{pen})'\right)(z_{itj}^{unp\prime}\ z_{itj}^{pen\prime})'}. \qquad (28)$$

The covariance matrices $\text{Cov}\left((\hat{\beta}_j^{unp})'\ (\hat{\beta}_j^{pen})'\right)$ can be obtained from the corresponding blocks in H^{-1}.

4 Simulation studies

To investigate the performance of the presented approach, we performed simulation studies for models with ordinal and multinomial responses. In both cases the additive predictor was defined to be the sum of a nonparametric effect and a spatial effect and the number of possible response categories is three. Nonparametric effects were estimated by cubic P-splines with second order random walk penalty and 20 inner knots, while spatial effects were assumed to follow the MRF prior (16).

A second approach, allowing to estimate structured additive regression models for multicategorical responses is the fully Bayesian approach presented in Fahrmeir and Lang (2001b) and Brezger and Lang (2003). Here, all unknown parameters, including the variance parameters, are assumed to be random. While priors for nonparametric effects are essentially the same as in section 2.3, priors for the variances are weakly informative inverse gamma distributions $IG(a, b)$. We chose $a = b = 0.001$, which is an often recommended standard choice approximating Jeffrey's prior. This choice has also proved to work better in sparse data situations than the second standard choice $a = 1$ and $b = 0.005$ (compare Fahrmeir et al. (2004)). Estimation in the fully Bayesian approach is based on Markov Chain Monte Carlo simulation techniques, see the references above for a detailed description.

The simulation for ordinal responses is based on a cumulative probit model with predictors

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

and

$$\begin{aligned} f_1(x) &= \sin[\pi(2x-1)], \\ f_2(s) &= 0.5(s_x + s_y). \end{aligned}$$

Function f_1 is a smooth function of the continuous covariate x and can be interpreted as a nonlinear time trend. The values of x where chosen from an equidistant grid of 100 values between -1 and 1. The spatial function f_2 is defined on the centroids (s_x, s_y) of the 124 districts s of the two southern states of Germany. Figure 1 a) and d) display these functions.

For multinomial responses we chose a multinomial logit model with predictors

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

and category specific functions $f_1^{(r)}(x)$ and $f_2^{(r)}(s)$ defined by

$$\begin{aligned} f_1^{(1)}(x) &= \sin[\pi(2x-1)] & f_1^{(2)}(x) &= \sin[2\pi(2x-1)], \\ f_2^{(1)}(s) &= -0.75|s^x|(0.5+s^y) & f_2^{(2)}(s) &= 0.5(s^x+s^y). \end{aligned}$$

Again the values of x were chosen from 100 equidistant points between -1 and 1 and the spatial functions f_2 are defined on the centroids (s_x, s_y) of the 124 districts s of Bayern and Baden-Württemberg. The smooth functions are shown in Figure 1 a) and b), the spatial functions are reproduced in Figure 1 c) and d).

To evaluate the impact of increasing information in the data, we considered three different sample sizes, namely $n = 500$, $n = 1000$ and $n = 2000$. Correspondingly, each value of x was assigned 5, 10 an 20 times. For the districts s, most values were assigned 4, 8 and 16 times, only some (randomly chosen) districts were assigned once more to achieve the total sample size. For each of the different sample sizes the simulation was repeated over 250 runs. Performance of the different approaches was compared in terms of bias, average coverage probabilities, and MSEs.

One first important observation is that the presented mixed model approach failed to converge in several cases. The number of iterations needed for the estimation is displayed in Figure 2. Note, that the estimation procedure terminates after 100 iterations regardless of whether convergence was achieved or not. Obviously the number of iterations reduces with increasing sample size, at least for ordinal responses. For nominal responses this trend is present but less clear cut.

A closer inspection of the convergence problems showed that variances for the nonparametric functions converged to a fixed value in a moderate number of iterations while at least one of the variances for spatial effects kept switching between to values relatively close to each other. This is consistent with the findings in Fahrmeir et al. (2004) for univariate responses. However, using the estimates based on the variances of the last (100th) iteration in the case of no convergence lead to reasonable results. Therefore we used results from all 250 simulation runs to compute MSEs, bias and coverage probabilities, regardless of whether convergence was achieved or not.

The convergence problems seem to be a specific problem of models with spatial effects. In a similar simulation study, where the predictors consisted only of two nonparametric effects of continuous covariates we never observed any problems of this kind. Probably the likelihood is less well behaved in the case of spatial effects because one variance parameter is used for a relatively large number of regression parameters.

For ordinal responses, bias was ignorable for both approaches and all sample sizes, so we do not show any figures here. For nominal responses bias was more obvious for small sample sizes but decreased with increasing information. MCMC estimates performed slightly better in this case because REML estimates tend to oversmooth effects, especially for small samples. As an example we show results on $f_1^{(2)}(x)$ in Figure 3.

Considering MSEs, a general but not surprising results is that the quality of the estimates is improved when sample sizes are extended. Figures 4 and 5 show boxplots of the log(MSE)s for ordinal and nominal responses, respectively. In all cases, REML estimates have a somewhat smaller median MSE, with more obvious differences for the spatial effects. Similar as for the bias, differences almost vanish for larger sample sizes.

The final comparison concentrates on average coverage probabilities. Tables 1 and 2 show average coverages based on nominal levels of 80% and 95%. In most cases, the coverage probabilities are almost identical for both REML and MCMC estimates, with MCMC estimates being slightly closer to the nominal levels. Except for the nonparametric effects in the nominal case with the smallest sample size, coverage probabilities are above the nominal level, indicating a more conservative behavior. This is most clear for coverages of the spatial effects where coverages for a nominal level of 95% are rather close to 1.

5 Application: A space-time study on forest health

These space-time data have been collected in yearly visual forest health inventories carried out in a forest district in the northern part of Bavaria from 1983 to 2001. The observation area extends 15 km from east to west and 10 km from north to south, with 83 stands of trees as observation points. In the following application, we consider beeches. For each tree, the degree of defoliation serves as an indicator for its damage state, which is given as an ordered response with three categories and $Y_{it} = 1$ (no damage of tree i in year t), $Y_{it} = 2$ (medium damage) and $Y_{it} = 3$ (severe damage) $i = 1, \dots, 83$, $t = 1983, \dots, 2001$. Figure 6 shows the temporal development of the frequency of the three damage categories, and the spatial distribution of trees together with the percentage of time points for which a tree was classified to be damaged (damage state 2 or 3), averaged over the entire observation period.

In addition to the temporal and spatial information, the data set includes a number of covariates describing the stand and the site of the tree, and the soil at the stand. In a first exploratory analysis based on an ordinal probit model, all continuous covariates were modelled nonparametrically. Furthermore the predictor contained an interaction term between age of the tree A_{it} and calendar time and different types of spatial effects. These first analyzes suggested some simplifications of the model, especially to model the effects of some continuous covariates in a parametric way. These were mostly variables that do not vary over time and are therefore constant for each tree. This lead to the final model

$$P(Y_{it} \leq r) = \Phi(\theta_r - [u_{it}'\gamma + f_1(t) + f_2(A_{it}) + \ldots + f_3(t, A_{it}) + f_{spat}(s_i)]),$$

where Φ denotes the standard normal distribution function. Covariates modelled in a parametric way are subsumed in the vector u_{it} consisting of both categorical and continuous covariates. Categorical covariates are moisture (3 categories), percentage of alkali (4 categories), thickness of the humus layer (5 categories), type of the forest (deciduous forest or mixed forest) and fertilizing (yes or no). Continuous covariates are the gradient of slope (in %), elevation above sea level, depth of soil (above rock in cm) and the canopy density (in %). To shorten the discussion, we will only show results for the continuous covariates in u_{it}.

We examined various parameterizations for the spatial effect f_{spat} including models with only structured or unstructured effects and the combination of both. In the following we will concentrate the discussion on the latter case. For the structured part of f_{spat} we compared Markov random fields and stationary Gaussian random fields. For Markov random fields two trees were considered as neighbors if their distance was less than 1.2km. 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 (19).

For interpretation of estimation results note the following: In accordance with our definitions (2) – (6), higher (lower) values of covariate effects correspond to worse (healthier) state of the trees.

Figure 7 shows the nonlinear effects f_1 and f_2 of calendar time and age of the tree for both models with MRF or GRF as structured spatial component. Additionally, we include nonlinear effects from a model that neglects spatial correlations and therefore has no spatial effect at all. Obviously, effects for models with MRF and GRF are virtually identical and the temporal effect reflects quite well the trends shown in Figure 6, with an increased frequency of damaged trees in the mid-eighties. 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.

Figure 8 contains the estimated spatial effects. Surprisingly, MRFs and GRFs lead to quite different results for the structured part of the spatial function. While for MRFs structured and unstructured effects are nearly equally pronounced, the unstructured effect almost totally outweighs the structured effect when using a GRF. This is in contrast to the situation, when only a structured effect is considered. Here, both approaches showed very similar results for the spatial effect.

Figure 9 shows the interaction between calendar time and age of the tree for the model with a MRF as structured spatial effect. For all other models the interaction effect looked almost the same. Obviously, young trees were in poorer health state in the eighties but recovered in the nineties unlike the older trees which 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 they grow older.

Table 4 shows estimates for the continuous covariates that were modelled in a parametric way and the thresholds. For all three models, canopy density has a strong negative effect, indicating that a dense stand of the tree decreases the probability of being damaged. This conclusion depends on the type of the tree and can be quite different for other species.

All other parametric effects are not significant, if a spatial component is included in the model. For the model without spatial effect, depth of soil has a small negative effect. Here, depth of soil seems to cover some of the spatial effect, since the covariate itself obeys a spatial structure. This was even more obvious in the exploratory analyzes, where all continuous covariates were modelled nonparametrically. In this case some covariates showed very wiggly effects if the spatial component was excluded from the model, and were therefore no longer interpretable. These effects also seemed to absorb some of the effects covered by the spatial component.

A last comparison of the three models is based on the classification of the trees according to the respective model. Table 3 shows these classifications and also provides misclassification rates. Here, models with MRF and GRF behave quite comparably again with a slight improvement based on a MRF. Both models are clearly superior to the model without spatial effect, confirming that inclusion of the spatial information is substantial.

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 paper 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 multicategorical 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, i.e., we have to consider category-specific trend functions $f_{time}^{(r)}(t)$ 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 analyzes for all tree species observed at the stands.

Acknowledgement:
We thank Axel Göttlein for providing the data and for helpful discussions, and we gratefully acknowledge financial support from the German Science Foundation (DFG), Sonderforschungsbereich 386 "Statistical Analysis of Discrete Structures".

References

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

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

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

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

Fahrmeir, L., Kneib, T. and Lang, S., 2004: Penalized Structured Additive Regression for Space-Time Data: a Bayesian Perspective. *Statistica Sinica*, under revision.

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*, Springer–Verlag, New York.

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

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

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

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

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

Johnson, M.E., Moore, L.M. and Ylvisaker, D., 1990: Minimax and maximin designs. *Journal of Statistical Planning and Inference* , 26, 131–148.

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

Knorr-Held, L., 1996: *Hierarchical Modelling of discrete longitudinal data*, Herbert Utz Verlag, München.

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

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.

Müller, H.G., Stadtmüller, U. and Tabnak, F., 1997: Spatial Smoothing of Geographically Aggregated Data, with Applications to the Construction of Incidence Maps. *Journal of the American Statistical Association*, 92, 61–71.

Nychka, D. and Saltzman, N., 1998: *Design of Air-Quality Monitoring Networks*, Lecture Notes in Statistics, 132, 51–76.

Patterson, H.D. and Thompson, R., 1971: Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545–554.

Rue, H. and Tjelmeland, H., 2002: Fitting Gaussian Markov Random Fields to Gaussian Fields. *Scandinavian Journal of Statistics*, 29, 31–49.

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

Stein, M.L., 1999: *Interpolation of spatial data. Some theory for kriging*, Springer, New York.

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

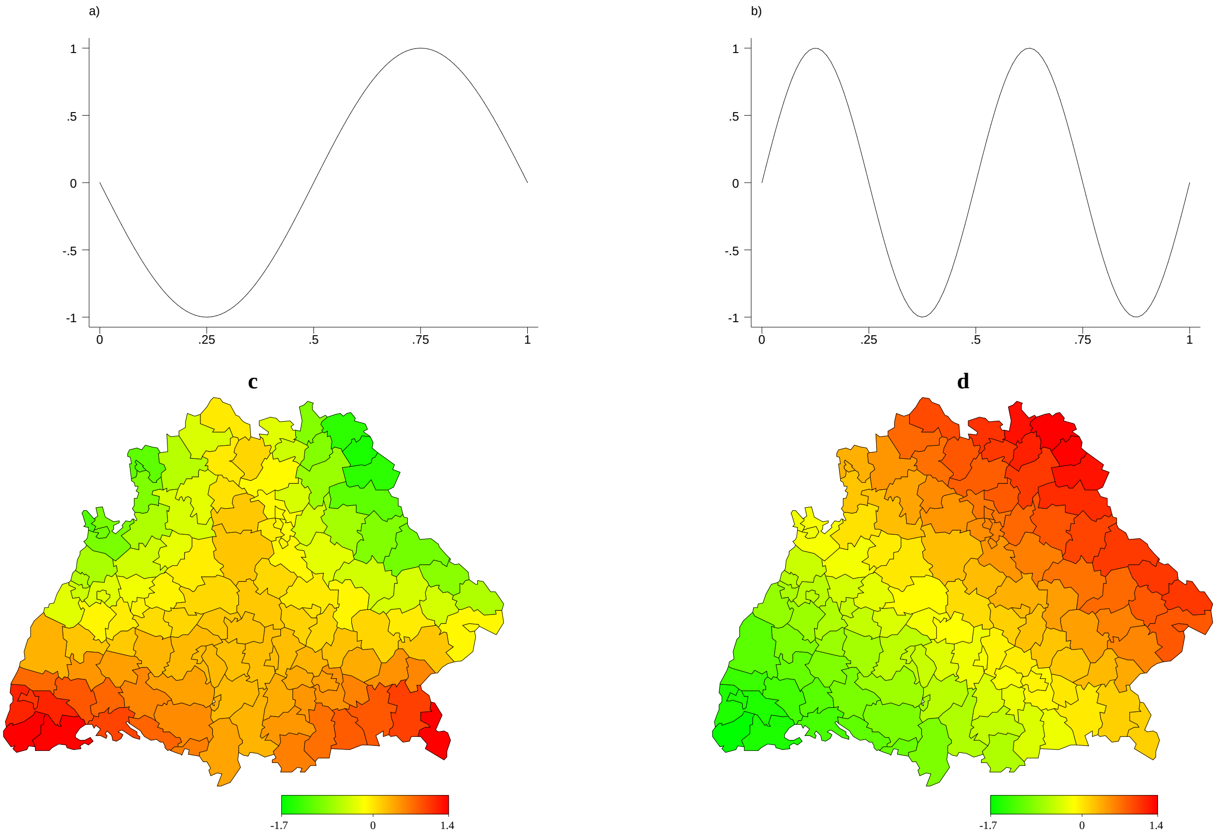

Figure 1: True functions for the simulation studies. Functions a) and d) are used for ordinal response models. Functions a) - d) are used for nominal response models.

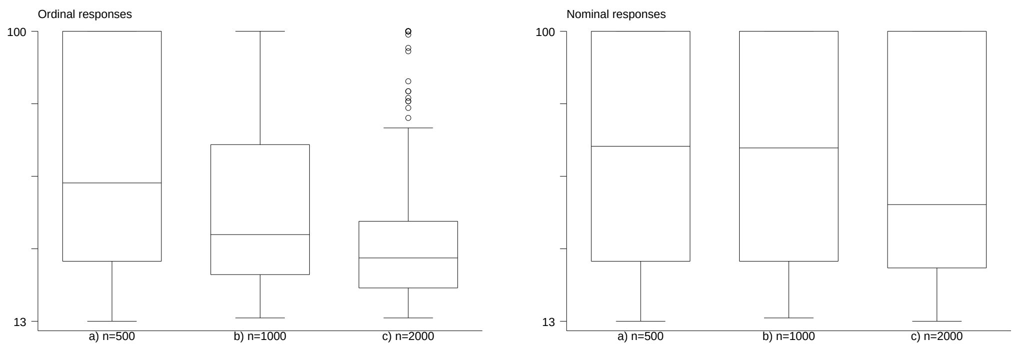

Figure 2: Convergence properties of the REML estimate: Boxplots of the number of iterations needed until convergence for ordinal responses (left panel) and nominal responses (right panel). The estimation procedure was stopped after 100 iterations.

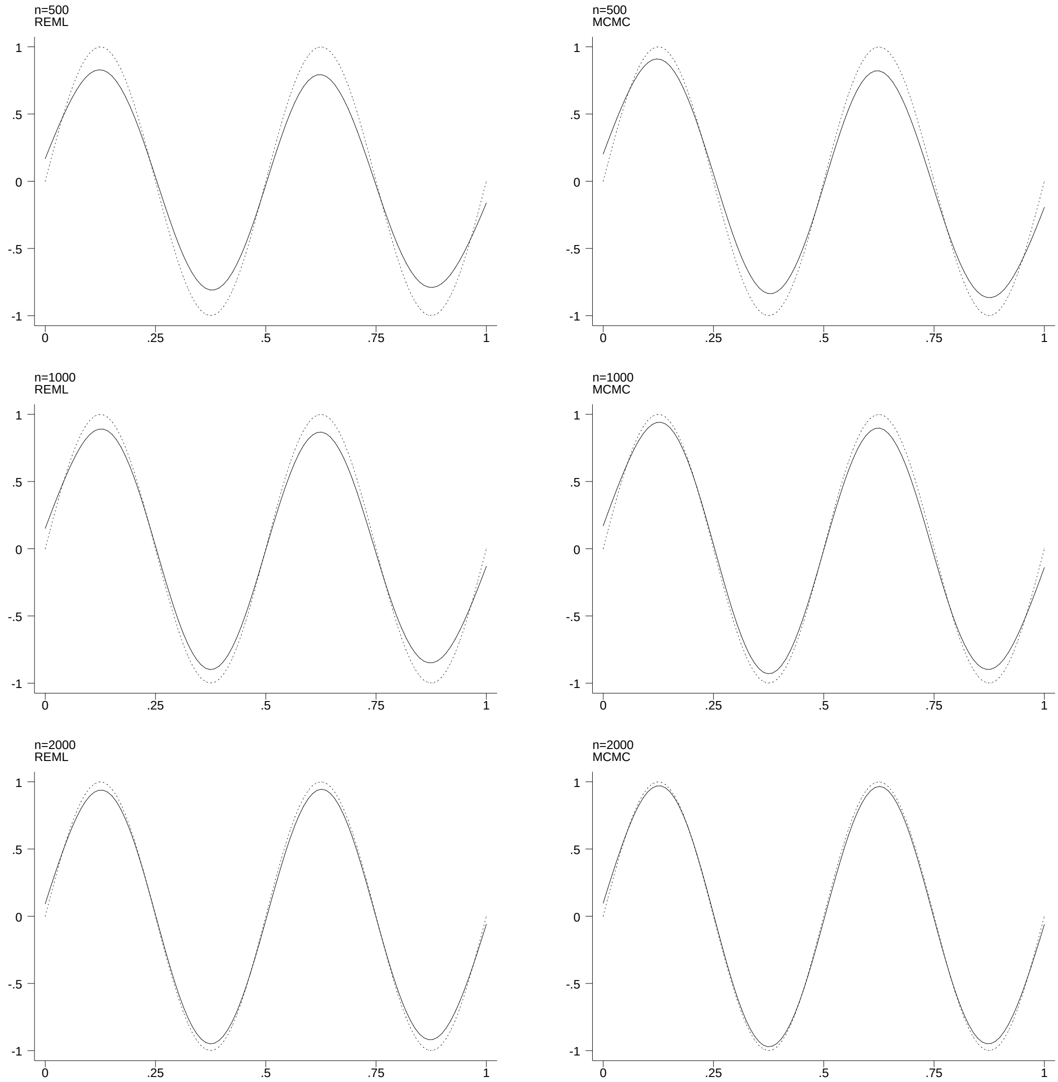

Figure 3: Nominal responses: Bias for $f^{(2)}(x)$ based on REML estimates (left panel) and MCMC estimates (right panel). Average estimates are indicated by solid lines, true functions by dashed lines.

		$f_1(x)$		$f_2(s)$	
		80%	95%	80%	95%
REML	$n = 500$	0.855	0.969	0.939	0.995
	$n = 1000$	0.865	0.976	0.931	0.994
	$n = 2000$	0.870	0.978	0.920	0.991
MCMC	$n = 500$	0.848	0.967	0.940	0.995
	$n = 1000$	0.849	0.968	0.932	0.994
	$n = 2000$	0.849	0.972	0.920	0.991

Table 1: Ordinal responses: Average coverage probabilities based on nominal levels of 80% and 95%.

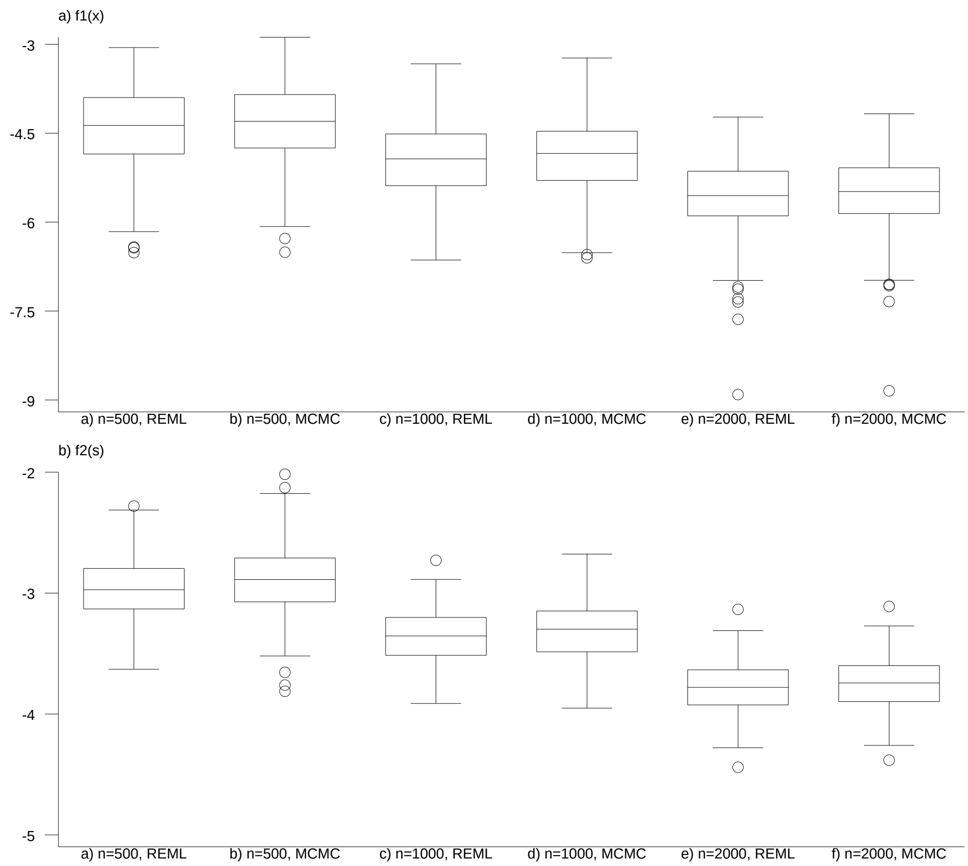

Figure 4: Ordinal Responses: Boxplots of log(MSE).

		$f_1(x)$		$f_2(x)$		$f_1(s)$		$f_2(s)$	
		80%	95%	80%	95%	80%	95%	80%	95%
REML	$n = 500$	0.764	0.899	0.791	0.939	0.890	0.975	0.942	0.994
	$n = 1000$	0.837	0.962	0.833	0.964	0.896	0.983	0.944	0.994
	$n = 2000$	0.866	0.974	0.849	0.973	0.897	0.986	0.946	0.996
MCMC	$n = 500$	0.788	0.947	0.797	0.949	0.896	0.986	0.962	0.998
	$n = 1000$	0.829	0.962	0.826	0.961	0.910	0.989	0.959	0.998
	$n = 2000$	0.855	0.973	0.829	0.964	0.905	0.988	0.949	0.996

Table 2: Nominal responses: Average coverage probabilities based on nominal levels of 80% and 95%.

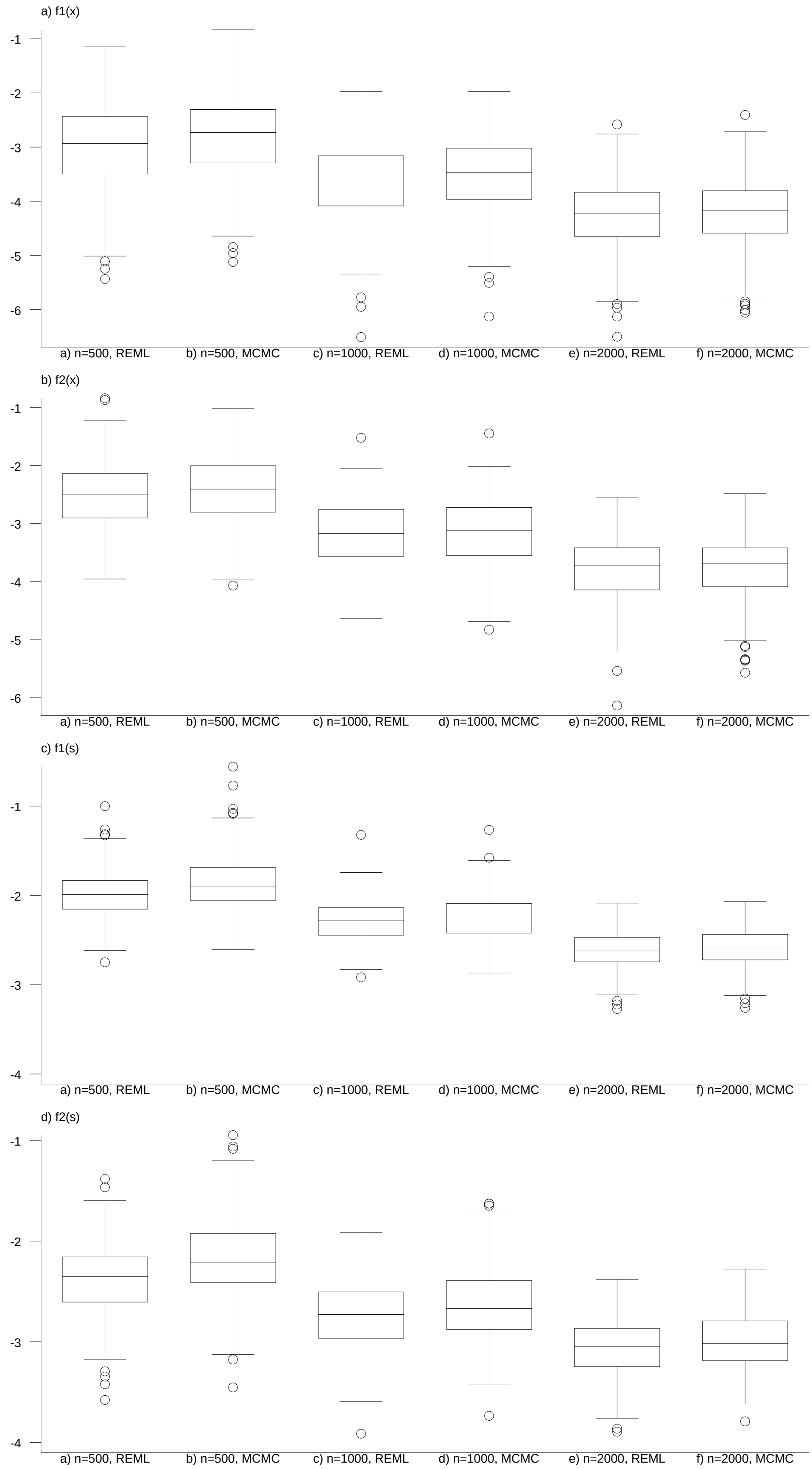

Figure 5: Multinomial Responses: Boxplots of log(MSE)

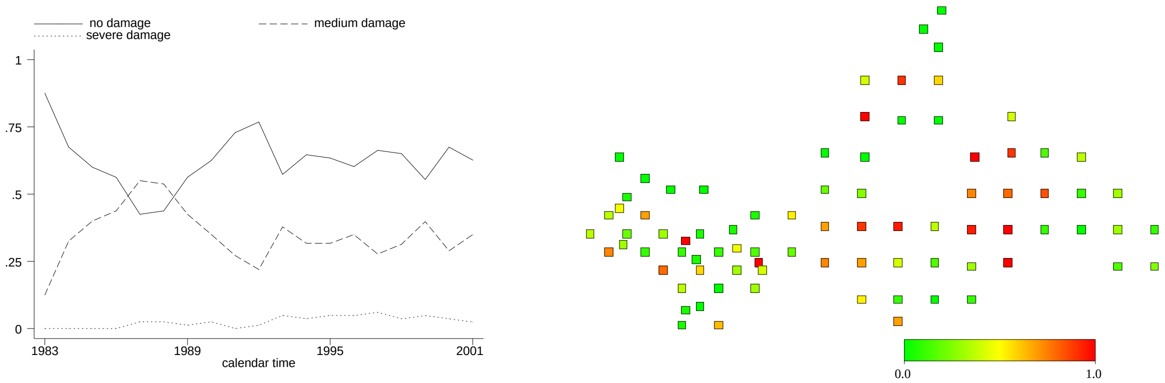

Figure 6: Forest health data: The left panel shows the temporal development of the frequency of the three different damage states. The solid line represents damage state '1' (no damage), the dashed line damage state '2' (medium damage) and the dotted line damage state '3' (severe damage). The right panel displays the percentage of years for which a tree was in damage state '2' or '3', averaged over the entire observation period.

y \ $\hat{y}$	1	2	3
1	906	62	0
2	107	427	5
3	0	16	24

12.3%

y \ $\hat{y}$	1	2	3
1	904	64	0
2	108	426	5
3	0	16	24

12.5%

y \ $\hat{y}$	1	2	3
1	850	118	0
2	150	386	3
3	0	34	6

19.7%

Table 3: Forest health data: Classification tables. The left and the middle table show classifications for models with structured and unstructured spatial effects. The structured spatial effect is modelled as MRF in the left table and as GRF in the middle table. The right table shows classifications in the absence of any spatial effect. Misclassification rates are displayed below the tables.

	MRF + Random Effect			GRF + Random Effect			No spatial effect		
	mode	std	p-val	mode	std	p-val	mode	std	p-val
$\theta^{(1)}$	0.1890	1.5489	0.9033	0.7137	1.4914	0.6318	0.9429	0.4913	0.0545
$\theta^{(2)}$	4.0142	1.5472	0.0098	4.5211	1.4928	0.0029	3.5164	0.4976	0.0000
Gradient of slope	0.0062	0.0156	0.6917	0.0025	0.0158	0.8744	-0.0011	0.0045	0.8128
Elevation	-0.0001	0.0035	0.9748	0.0015	0.0033	0.6450	0.0007	0.0010	0.4451
Depth of soil	-0.0117	0.0155	0.4503	-0.0202	0.0159	0.2050	-0.0111	0.0049	0.0234
Canopy density	-2.3830	0.4181	0.0000	-2.3030	0.4222	0.0000	-2.1756	0.2101	0.0000

Table 4: Forest health data: Estimates for thresholds and parametric effects of continuous covariates.

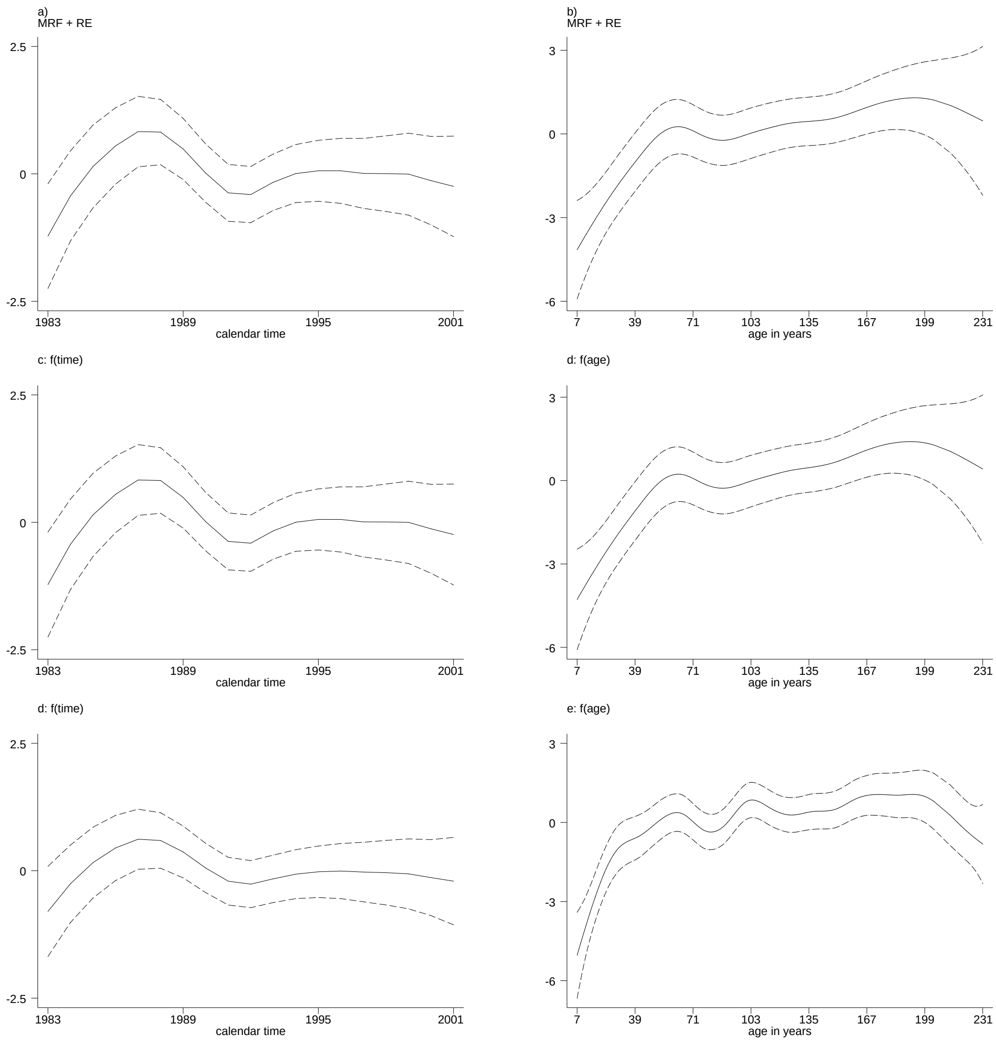

Figure 7: Forest health data: Effects of calendar time (left) and age of the tree (right) together with pointwise 95% credible intervals. The upper and the middle panel show estimates for models with structured and unstructured spatial effects. The structured spatial effect is modelled as MRF in the upper panel and as GRF in the middle panel. The lower panel shows estimates in the absence of any spatial effect.

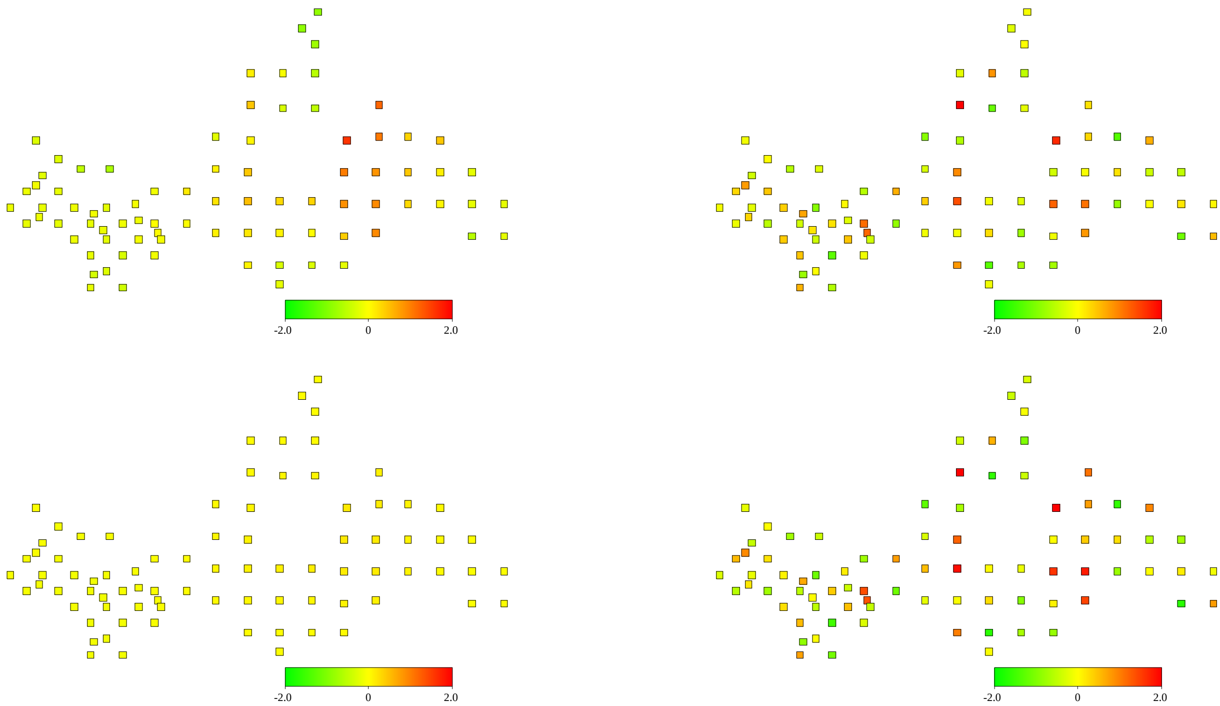

Figure 8: Forest health data: Structured spatial effects (left) and unstructured spatial effects(right). The structured spatial effect is modelled as MRF in the upper panel and as GRF in the lower panel.

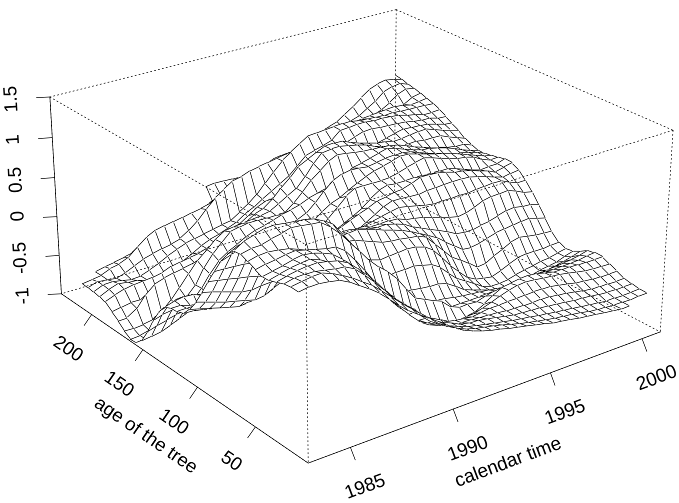

Figure 9: Forest health data: Interaction between calendar time and age of the tree for a model with structured and unstructured spatial effects. The structured spatial effect is modelled as MRF.