Appl. Statist. (2007)
56, *Part* 1, *pp.* 29–50

A spatial model for the needle losses of pine-trees in the forests of Baden-Württemberg: an application of Bayesian structured additive regression

Nicole H. Augustin,

University of Bath, UK

Stefan Lang,

Leopold-Franzens-Universität Innsbruck, Austria

Monica Musio

University of Cagliari, Italy

and Klaus von Wilpert

Forest Research Centre Baden-Württemberg, Freiburg, Germany

[Received November 2004. Revised June 2006]

Summary. The data that are analysed are from a monitoring survey which was carried out in 1994 in the forests of Baden-Württemberg, a federal state in the south-western region of Germany. The survey is part of a large monitoring scheme that has been carried out since the 1980s at different spatial and temporal resolutions to observe the increase in forest damage. One indicator for tree vitality is tree defoliation, which is mainly caused by intrinsic factors, age and stand conditions, but also by biotic (e.g. insects) and abiotic stresses (e.g. industrial emissions). In the survey, needle loss of pine-trees and many potential covariates are recorded at about 580 grid points of a 4 km × 4 km grid. The aim is to identify a set of predictors for needle loss and to investigate the relationships between the needle loss and the predictors. The response variable needle loss is recorded as a percentage in 5% steps estimated by eye using binoculars and categorized into healthy trees (10% or less), intermediate trees (10–25%) and damaged trees (25% or more). We use a Bayesian cumulative threshold model with non-linear functions of continuous variables and a random effect for spatial heterogeneity. For both the non-linear functions and the spatial random effect we use Bayesian versions of *P*-splines as priors. Our method is novel in that it deals with several non-standard data requirements: the ordinal response variable (the categorized version of needle loss), non-linear effects of covariates, spatial heterogeneity and prediction with missing covariates. The model is a special case of models with a geoadditive or more generally structured additive predictor. Inference can be based on Markov chain Monte Carlo techniques or mixed model technology.

Keywords: Cumulative threshold model; Defoliation; Generalized additive mixed models; Markov chain Monte Carlo methods; Norway spruce; Ordinal response; *P*-splines; Spatial forestry data; Structured additive predictor

Address for correspondence: Nicole H. Augustin, Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK.
E-mail: n.h.augustin@bath.ac.uk

 0035–9254/07/56029

1. Introduction

During the past 30 years an increase in forest damage has been observed in the forests of Baden-Württemberg, a federal state in the south-western region of Germany. In particular the deterioration of tree crown condition is supposed to be of a chronic nature and likely to be caused by acidification of the soil. This acidification is caused by industrial emissions (Forstliche Versuchs- und Forschungsanstalt Baden-Württemberg, 2001) and affects soil chemistry including the availability of nutrient and metals. The acidification causes the washing out of essential alkaline macronutrients in the root area, and this has a negative influence on tree nutrition which finally causes the loss of needles and leaves. In some of the areas of Baden-Württemberg the soils are already acidic, e.g. in the Black Forest where the geology is mainly siliceous bedrock such as granite and gneiss and does not have a high buffer capacity against the acids. When acid deposition occurs on soils which are not well buffered, alkaline macronutrients including potassium (K), calcium (Ca) and magnesium (Mg) are readily washed out, making them unavailable to the forest as nutrients. In contrast, soils that are rich in Ca, K and Mg are more resistant to the effects of acid deposition since they are naturally alkaline; for example the Schwäbische Alp, which is mostly limestone, is such an area. In these areas a delayed reaction to the acid deposition is to be expected.

Since the 1980s the forest health status has been monitored in Baden-Württemberg, as part of a bigger monitoring scheme at the European Union level, using different schemes. The data that are available here are from the Survey of Emission Impact and Forest Nutrition of 1994, which is characterized by measurements of the nutrients in the needles of the trees. Sampling of forest nutrition involves the felling of trees and a subsequent chemical analysis of the needles. This makes it very time consuming and costly, and hence surveys of this kind have only been undertaken in four years since 1980. The different nutrients that were sampled in the Survey are markers for different processes: nitrogen (N) indicates the supply of trees with this essential plant nutrient, which is limited in natural ecosystems; too much N is caused by anthropogenic input from the air originating for example from cattle breeding as well as by industrial burning processes with high temperatures, which generate NO_x. A surplus of N availability destabilizes the balance of the other nutrients. If too much N is available trees can compensate by growing faster, and hence high N levels can be detected in the soil, but often not in the needles since the concentration of N in the needles tends to be regulated by the plant in a physiologically favourable range. Phosphorus (P), K, Ca and Mg are expected to be low if acidification is taking place. P is fixed by aluminium, which is mobilized during soil acidification at soil acidity levels below pH 4.5–4.2 and thus is not available to the plant. Ca, Mg and K are washed out during soil acidification. Manganese (Mn) becomes mobilized from limited soil pools at an intermediate acidification status which is characterized by soil acidity values around pH 5–4.2 (Hildebrand, 1986). Thus Mn can be used as an indicator for this transient acidification phase, but otherwise it is not thought to play an important role for tree nutrition. Other stand-specific variables are also available (Table 1).

The main scientific objective of the Survey of Emission Impact and Forest Nutrition is to explain the processes causing tree deterioration. In particular the relationships between crown condition (needle losses) and locational attributes, soil condition, geology and tree nutrients are of interest. This objective involves two statistical tasks:

(a) identifying an appropriate statistical tool for such a model and
(b) finding a set of *best* covariates for the defoliation.

The main purpose of the model is to assess the importance of the different nutrients that

Table 1. Available data from the Survey of Emission Impact and Forest Nutrition of 1994 in Baden-Württemberg

Variable	*Type*	*Description*
Response variable		
needle loss	Ordered categorical	Percentage needle loss of the felled Survey of Emission Impact and Forest Nutrition tree estimated by eye (binoculars)
Tree-specific covariates		
age	Continuous	Age of tree (years)
(treespec	Binary	Tree species: 1, spruce; 0, fir)
Mg	Continuous	Mg value (g kg^{-1}) in needles
Ca	Continuous	Ca value (g kg^{-1}) in needles
K	Continuous	K value (g kg^{-1}) in needles
N	Continuous	N value (g kg^{-1}) in needles
P	Continuous	P value (g kg^{-1}) in needles
Mn	Continuous	Mn value (g kg^{-1}) in needles
Zn	Continuous	Zinc value (g kg^{-1}) in needles
N/K	Continuous	N to K ratio
N/P	Continuous	N to P ratio
N/Ca	Continuous	N to Ca ratio
N/Mg	Continuous	N to Mg ratio
Stand-specific covariates		
altitude	Continuous	Altitude (m)
geolnr	Categorical	Geological area: 1, metamorphic crystalline bedrock; 2, pure granite bedrock; 4, Triassic sandstones; 5, Triassic claystone and sands; 6, limestone; 7, tertiary sands; 8, glacial fluviatile sediments, loam and moraine
soiltxt	Categorical	Soil texture: 2, gravel, sand, loamy sand, loamy grus, sand above clay; 6, sandy loam, loam above clay, loam; 10, clay, Nitisole material, chalky Loess, silty loam; 13, peat
soiltype	Categorical	Soil type: 1, Fluvisol; 4, Cabisol brown soil; 5, gley, pseudogley; 7, Luvisol para brown soil; 8, pelosol; 9, podsol; 11, rendzina soil; 12, calcic Nitisol
soildepth	Categorical	Soil depth: 1, shallow; 2, average; 3, deep
soilwbudg	Categorical	Soil water budget: 2, humid, fresh; 3, medium fresh; 4, medium dry, dry; 6, changing humid or dry
nutrbal	Categorical	Nutrient balance: 1, bad; 2, medium; 3, good
humus	Categorical	Humus form: 1, Mull; 2, Mullmoder; 3, Moder; 4, raw humus like Moder; 5, raw humus, peat
slopedir	Ordered categorical	Direction or type of slope: 0, no slope; 1, other; 7, west facing
(slopegrad	Continuous	Gradient of slope (%))
relief	Categorical	Relief type: 1, plateau; 3, upper slope; 4, middle slope; 5, lower slope; 6, valley, river-bed
situat	Binary	Type of situation: 0, well situated (sheltered); 1, exposed
temperature	Continuous	Mean temperature at sample location in 1994
Pr	Continuous	Mean precipitation at sample location in 1994
E	Continuous	Mean of the actual evaporation at sample location in 1994

are sampled as covariates. A further objective is to produce spatial predictions of defoliation in unsampled locations for maps of needle loss. These maps are needed for forest management, e.g. to decide which areas of the forest need some treatment, such as liming to prevent acidification. This objective involves the statistical problem of interpolating between the sampled locations, i.e. prediction. Given that we are dealing with some covariates that are measured at the tree level and hence not available at unsampled locations, this is not straightforward.

Given that the response variable needle loss is recorded as a percentage in steps of 5%, we cannot assume that, conditional on covariate effects, it follows a normal distribution. As is commonly done for the reporting of forest damage (Meining *et al.*, 2003), we categorize the variable into healthy trees (10% or less), intermediate trees (10–25%) and damaged trees (more than 25%). Hence we require a model for the ordinal response variable, the categorized version of needle loss. We also know *a priori* that some of the explanatory variables may have non-linear effects, e.g. age and altitude. Although we have potentially a very large set of variables explaining needle loss (Table 1) it is very likely that there is still some unexplained spatial heterogeneity or correlation. The four requirements of an ordinal response variable, non-linear effects of covariates, spatial heterogeneity and prediction with missing covariates represent a statistical challenge in terms of methodology and software. We use a Bayesian cumulative threshold model with non-linear functions of continuous variables and a random effect for spatial heterogeneity. The model is called a geoadditive model (Kammann and Wand, 2003) which in turn is a special case of structured additive models (Fahrmeir *et al.*, 2004; Brezger and Lang, 2006a) or additive plus interactions models that were discussed in Ruppert *et al.* (2003). Inference can be performed either by using Markov chain Monte Carlo (MCMC) techniques (Fahrmeir and Lang, 2001a,b; Lang and Brezger, 2004) or by utilizing a mixed model representation of structured additive regression models (Lin and Zhang, 1999; Fahrmeir *et al.*, 2004; Kneib and Fahrmeir, 2006). Our model is novel because it allows for an ordinal response variable in a model with non-linear effects of covariates, handles spatial correlation and produces predictions that are useful for monitoring purposes.

There are many applications to forestry data; for example in Fahrmeir and Lang (2001a) an example of a Bayesian generalized additive mixed model fitted to binary forest damage response data is presented. In Preisler *et al.* (1997), Wood and Augustin (2002) and Augustin *et al.* (2005) generalized additive models are used, i.e. the spatial heterogeneity or correlation is dealt with by modelling it as a spatial trend using smooth trend functions such as penalized thin plate regression splines smoothers. Although ordinal responses are very common, none of the applications that were mentioned above fits models to ordinal response variables.

Section 2 describes the motivating data and survey in detail. Section 3 describes Bayesian geoadditive models for ordered categorical responses, including details on assumptions about the priors and on the MCMC sampler that was used. Section 4 presents the results. We finish with a discussion in Section 5.

2. The data

The data come from the Survey of Emission Impact and Forest Nutrition that was carried out by the Forest Research Centre Baden-Württemberg in 1994 in which on each of 576 grid points on a 4 km $\times$ 4 km grid two random coniferous trees are sampled to check the health status of the forest. Fig. 1 shows the sampling locations and the various growth areas. The growth areas tend to have similar landscape and topographic characteristics, e.g. the Black Forest, where we have mainly metamorphic crystalline bedrock, granite and sandstones (Fig. 1). As an indicator of the state of deterioration of the forest, we consider the needle losses which are recorded as a percentage estimated by eye by using binoculars. We define the variable *needleloss* Y as taking the value 1 if the tree is healthy, 2 if it is an intermediate tree and 3 if it is damaged. Only fir (*abies alba*, L.) and Norway spruce (*picea abies*, L.) trees were considered in the survey, of which only 17% were fir. The species of tree is confounded with other covariates, e.g. the age of the tree, the rarer fir trees tending to be older than the spruce trees, with a median age for fir of 103 years compared with a median age of spruce of 81 years. Because of this confounding we restrict the analysis to spruce trees only. The observed needle loss of spruce trees is shown in Fig. 1.

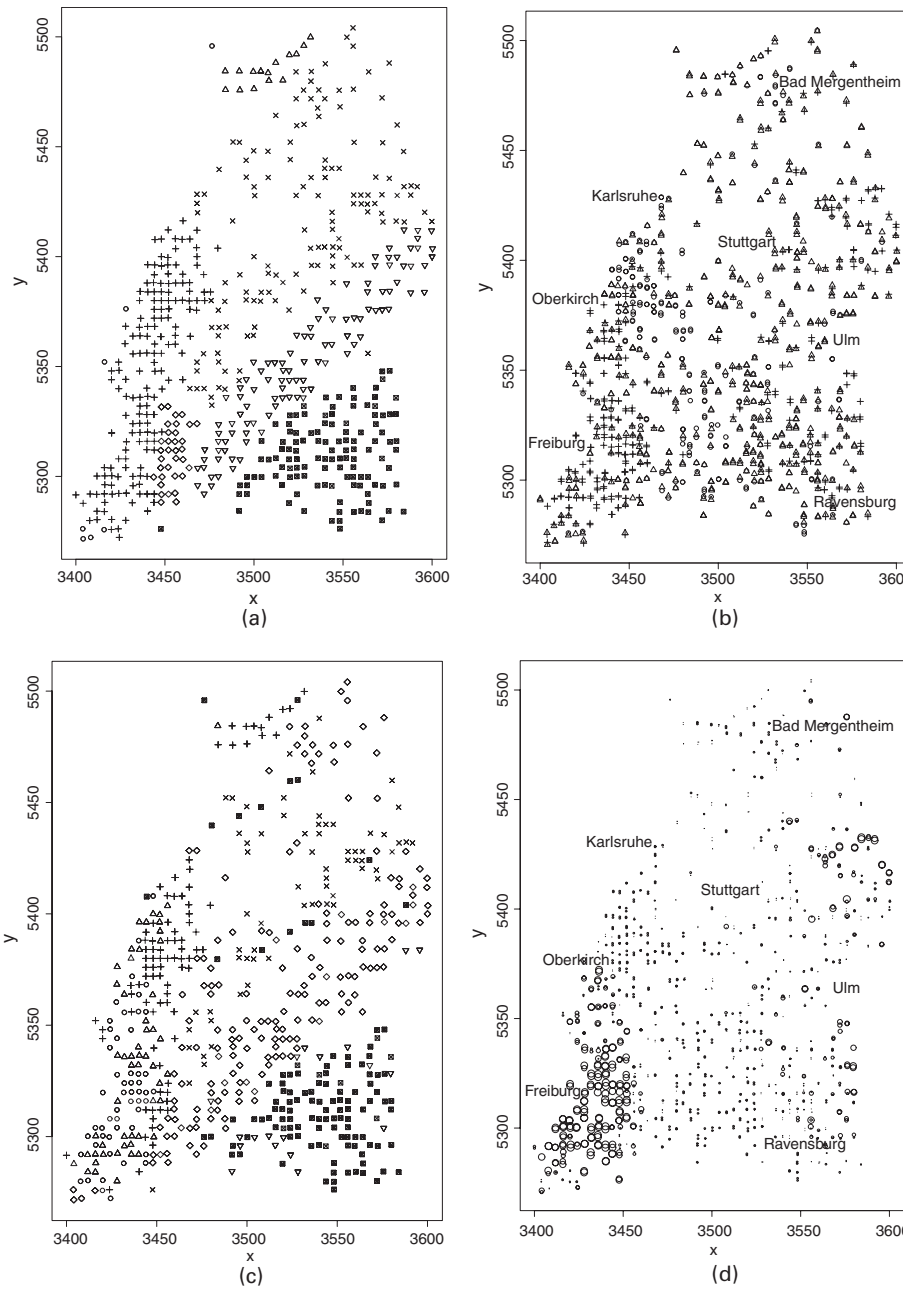

Fig. 1. (a) Growth areas (○, Rhine area; △, Oden Forest; +, Black Forest; ×, around Stuttgart; ◇, Baar Black Forest; ∇, Swabian Alp; ☒, Donau, Lake of Konstanz), (b) observed needle loss of spruce trees by location (○, healthy; △, intermediate; +, damaged; two trees were sampled by location), (c) geological areas (○, metamorphic crystalline; △, pure granite; +, Triassic sandstones; ×, Triassic claystones; ◇, limestone; ∇, Tertiary sands; ☒, glacial fluvial sediments) and (d) spatial map of the predictive distribution for the outcome based on the model fitted to observed and interpolated covariates (○, P(damaged) = 1; •, P(damaged) = 0): shown is the probability (by size of the circle radii) of the tree being damaged as estimated from MCMC samples of the predictive distribution from the spatial model (4.1)

In what follows, we shall consider 28 possible covariates $z = (z_1, \ldots, z_q)'$, $q = 28$, shown in Table 1. These possible influential factors, mostly collected as part of the Survey, can be summarized into the following categories: *nutrients* in the needles such as Mg, Ca, K, Mn, P, N and zinc (Zn), other *tree-specific variables* such as the age, *landscape and topographic characteristics* such as altitude, geological substrate (shown in Fig. 1), direction or type of slope, gradient of slope, relief form, type of situation, *soil characteristics* such as soil texture, soil type, soil depth, soil water budget, trophic class of the soil and humus form, and *meteorological characteristics* such as the mean temperature and precipitation. The supply of N tends to a surplus in the whole region. Hence we expect imbalances between N and other nutrients intensifying the chronic shortage of those nutrients, rather than direct damage caused by a surplus of N. Thus the ratios N/P, N/K, N/Ca and N/Mg are also included in the analysis.

For illustrating the prediction of needle loss at unsampled locations, we use additional data comprising 229 new locations, where we have some covariate information available.

3. Bayesian geoadditive models for ordered categorical responses

3.1. Cumulative threshold models

A common tool for analysing regression data with ordinal responses is the widely used *cumulative threshold model* (see for example Fahrmeir and Tutz (2001)). The model assumes that the response variable Y, here needleloss, is a categorized version of a latent continuous variable,

$$U = \eta + \varepsilon,$$

where η is a predictor depending on covariates and parameters and ε is the error variable. The two variables Y and U are linked by $Y = r$ if and only if $\theta_{r-1} < U \leqslant \theta_r$, $r = 1, 2, 3$, with thresholds $-\infty = \theta_0 < \theta_1 < \theta_2 < \theta_3 = \infty$. It follows immediately that Y is determined by the model $P(Y \leqslant r) = F(\theta_r - \eta)$ where F is the distribution function of the error variable ε of U. In this paper we assume that the errors are Gaussian, i.e. $\varepsilon \sim N(0, 1)$, leading to a cumulative probit model. Alternatively we could use the extreme value distribution for the errors, which yields the cumulative logit model. We prefer Gaussian errors because MCMC inference described in Section 3.3 is considerably easier for probit models.

The covariates $\mathbf{z}$ enter the model through the predictor η. Traditionally, a linear predictor is assumed, i.e.

$$\eta = \gamma_1 z_1 + \ldots + \gamma_q z_q = \mathbf{z}'\boldsymbol{\gamma} \tag{3.1}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)'$ is unknown and must be estimated together with the unknown thresholds from the data. For identifiability, the linear combination does not contain an intercept term γ_0. Otherwise one of the thresholds must be set to 0.

With our forestry data we are facing the following problems.

(a) *Non-linear covariate effects*: the influence of some of the continuous covariates, e.g. the age of the tree, might be non-linear. If the functional form of the influence is known *a priori*, such a relationship can be easily modelled within the traditional framework using the parametric predictor (3.1). In most cases, however, the functional form is completely unknown.
(b) *Spatial heterogeneity*: possible *spatial correlations* between neighbouring observations must to be considered appropriately.

A flexible and practical modelling approach which can deal with these difficulties is that based on Bayesian cumulative threshold models with a *geoadditive* or more generally a *structured*

additive predictor as developed in Fahrmeir and Lang (2001a,b), Lang and Brezger (2004) and Brezger and Lang (2006a). A geoadditive predictor (see below) is sufficiently flexible to deal with non-linear covariate effects as well as spatially correlated observations. The approach is included in the publicly available software package BayesX; see Brezger *et al.* (2005a) for the first steps and the three manuals (Brezger *et al.*, 2005b) for a detailed description. An example of the use of BayesX is also given in Appendix A. The program is available via the Internet at `http://www.stat.uni-muenchen.de/~bayesx/`.

In what follows we describe Bayesian cumulative threshold models with a geoadditive predictor. To consider possible non-linear effects of continuous covariates as well as spatial heterogeneity, we replace the linear predictor (3.1) by the geoadditive predictor

$$\eta = f_1(x_1) + \ldots + f_p(x_p) + f_{\text{spat}}(c_1, c_2) + \mathbf{w}'\boldsymbol{\gamma}. \tag{3.2}$$

Here, $f_1, \ldots, f_p$ are possibly non-linear functions of the continuous covariates $\mathbf{x} = (x_1, \ldots, x_p)'$ in the data set, and f_{spat} is a spatial effect of the Cartesian co-ordinates c_1 and c_2 of the tree location. The term $\mathbf{w}'\boldsymbol{\gamma}$ corresponds to usual linear effects of (usually categorical) covariates $\mathbf{w}$.

3.2. Prior assumptions

For Bayesian inference, the unknown functions $f_1, \ldots, f_p$ and f_{spat} in equation (3.2), or more exactly corresponding vectors of function evaluations, and the fixed effects parameters $\boldsymbol{\gamma}$ are considered as random variables and must be supplemented by appropriate prior assumptions.

3.2.1. Priors for linear effects parameters

Throughout the paper we shall assume independent diffuse priors $p(\gamma_j) \propto$ constant for the linear effects parameters $\boldsymbol{\gamma}$.

3.2.2. Priors for effects of continuous covariates

For continuous covariates, *P*-splines, which were introduced by Eilers and Marx (1996) in a frequentist setting and by Lang and Brezger (2004) in a Bayesian version, will be our standard choice. The basic assumption is that an unknown smooth function f_j of a covariate x_j can be approximated by a polynomial spline of degree l, defined on a set of equally spaced knots $x_{j,\min} = \zeta_{j0} < \zeta_{j1} < \ldots < \zeta_{j,k-1} < \zeta_{jk} = x_{j,\max}$ within the domain of x_j. The spline can be written in terms of a linear combination of $S_j = k + l$ *B*-spline basis functions B_{js}, i.e.

$$f(x_j) = \sum_{s=1}^{S_j} \beta_{js} B_{js}(x_j).$$

Here $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jS_j})'$ corresponds to the vector of unknown regression coefficients. The crucial point with regression splines is the choice of the number and the position of the knots. For a small number of knots, the resulting spline may be not sufficiently flexible to capture the variability of the data. For a large number of knots, estimated curves tend to overfit the data and, as a result, excessively rough functions are obtained. As a remedy for these problems Eilers and Marx (1996) suggested a moderately large number of equally spaced knots (usually between 20 and 40) to ensure enough flexibility and defined a *roughness penalty* based on differences of adjacent *B*-spline coefficients to guarantee sufficient smoothness of the fitted curves. In a Bayesian approach, as considered here, the regression coefficients $\boldsymbol{\beta}_j$ must be supplemented with appropriate prior distributions. The stochastic analogues of difference penalties are first-

and second-order random walks for the coefficients $\beta_{j1}, \ldots, \beta_{jS_j}$ defined by

$$\beta_{js} = \beta_{j,s-1} + u_{js}$$

or

$$\beta_{js} = 2\beta_{j,s-1} - \beta_{j,s-2} + u_{js}$$

with independent and identically distributed noise $u_{js} \sim N(0, \tau_j^2)$.

The variance parameter τ_j^2 is equivalent to the inverse smoothing parameter in a frequentist approach and controls the trade-off between flexibility and smoothness. For full Bayesian inference, the unknown variance parameters τ_j^2 are also considered as random and are estimated simultaneously with the unknown $\boldsymbol{\beta}_j$. Therefore, hyperpriors are assigned to the variances τ_j^2 in a further stage of the hierarchy by highly dispersed (but proper) inverse gamma priors $p(\tau_j^2) \sim \text{IG}(a_j, b_j)$. Common choices for a_j and b_j are $a_j = 1$ and b_j small, e.g. $b_j = 0.005$ or $b_j = 0.0005$. Alternatively we may set $a_j = b_j$, e.g. $a_j = b_j = 0.001$. On the basis of experience from extensive simulation studies we use $a_j = b_j = 0.001$ as our standard choice. In some situations, the estimated non-linear functions f_j may be very sensitive to the particular choice of hyperparameters a_j and b_j. This may be so for very low signal-to-noise ratios and/or small sample sizes. It is therefore highly recommended to estimate all models under consideration by using a (small) number of *different* choices for a_j and b_j to assess the dependence of results on minor changes in the model assumptions.

3.2.3. *Priors for spatial effects*

For the spatial covariate effect f_{spat} several alternative specifications are possible. In principle, any two-dimensional surface estimator might be used. Fahrmeir and Lang (2001b) used a Markov random-field prior (Besag *et al.*, 1991) where two locations are assumed to be neighbours if they are within a certain Euclidean distance. Kammann and Wand (2003) used, in a frequentist setting, Gaussian fields which turn out to be two-dimensional surface estimators based on radial basis functions. Spatial smoothing based on Gaussian fields is also known as kriging.

For the forestry data we use two-dimensional *P*-splines as introduced in Land and Brezger (2004); see also Eilers and Marx (2003). Here we assume that the unknown surface can be approximated by the tensor product of the two one-dimensional *B*-splines, i.e.

$$f_{\text{spat}}(c_1, c_2) = \sum_{\rho=1}^{m} \sum_{\nu=1}^{m} \beta_{\text{spat},\rho\nu} B_{1\rho}(c_1) B_{2\nu}(c_2).$$

A prior for the $S_{\text{spat}} = m^2$-dimensional parameter vector $\boldsymbol{\beta}_{\text{spat}} = (\beta_{\text{spat},11}, \ldots, \beta_{\text{spat},mm})'$ is now based on spatial smoothness priors that are common in spatial statistics (see for example Besag and Kooperberg (1995)). The most commonly used prior specification based on the four nearest neighbours can be defined by

$$\beta_{\text{spat},\rho\nu} | \cdot \sim N\{\tfrac{1}{4}(\beta_{\text{spat},\rho-1,\nu} + \beta_{\text{spat},\rho+1,\nu} + \beta_{\text{spat},\rho,\nu-1} + \beta_{\text{spat},\rho,\nu+1}), \tau_{\text{spat}}^2/4\} \tag{3.3}$$

for $\rho, \nu = 2, \ldots, m-1$, with appropriate changes for corners and edges.

3.2.4. *General form of the priors*

It turns out that both the prior for continuous covariates and that for spatial covariates can be written in a general form. The vectors of function evaluations $\mathbf{f}_j = (f_j(x_{j1}), \ldots, f_j(x_{jn}))'$, $j = 1, \ldots, p, \text{spat}$, can be written as the matrix product of an $n \times S_j$ design matrix $\mathbf{X}_j$ and the

vector of parameters $\boldsymbol{\beta}_j$, i.e.

$$\mathbf{f}_j = \mathbf{X}_j \boldsymbol{\beta}_j. \tag{3.4}$$

The design matrix is composed of the basis functions evaluated at the observations, e.g. $\mathbf{X}_j(i,s) = B_s(x_{ij})$, $s = 1,\ldots,S_j$, $i = 1,\ldots,n$, for univariate P-splines.

For the predictor (3.2) we obtain, in matrix notation,

$$\boldsymbol{\eta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_{\text{spat}}\boldsymbol{\beta}_{\text{spat}} + \mathbf{W}\boldsymbol{\gamma},$$

where $\mathbf{W}$ corresponds to the usual design matrix for fixed effects.

The general form of the prior for $\boldsymbol{\beta}_j$ is given by

$$\boldsymbol{\beta}_j|\tau_j^2 \propto \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\mathbf{K}_j\boldsymbol{\beta}_j\right), \tag{3.5}$$

where $\mathbf{K}_j$ is a *penalty matrix*. For example,

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

for P-splines with a first-order random-walk penalty.

3.3. Bayesian inference based on Markov chain Monte Carlo techniques

Bayesian inference is based on the posterior which is given by

$$\begin{aligned} p(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p,\boldsymbol{\beta}_{\text{spat}},\tau_1,\ldots,\tau_p,\tau_{\text{spat}},\boldsymbol{\gamma},\mathbf{U}|\mathbf{Y}) \propto\ & p(\mathbf{Y}|\mathbf{U})\, p(\mathbf{U}|\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p,\boldsymbol{\beta}_{\text{spat}},\boldsymbol{\gamma}) \\ & \times \prod_{j=1}^{p} p(\boldsymbol{\beta}_j|\tau_j)\, p(\tau_j) \\ & \times p(\boldsymbol{\beta}_{\text{spat}}|\tau_{\text{spat}}^2)\, p(\tau_{\text{spat}}^2)\, p(\boldsymbol{\gamma}), \end{aligned}$$

with

$$p(\mathbf{Y}|\mathbf{U}) = \prod_i p(Y_i|U_i).$$

The conditional likelihood $p(Y_i|U_i)$ is given by

$$p(Y_i|U_i) = \sum_{r=1}^{3} I(\theta_{r-1} < U_i \leqslant \theta_r)\, I(Y_i = r), \tag{3.6}$$

because $p(Y_i|U_i)$ is 1 if U_i obeys the constraint that is imposed by the observed value of Y_i. MCMC simulation is based on drawings from full conditionals of blocks of parameters, given the other parameters and the data. In what follows, we use the blocks U_i, $i = 1,\ldots,n$, $\boldsymbol{\beta}_j, \tau_j^2$, $j = 1,\ldots,p$, $\boldsymbol{\beta}_{\text{spat}}$, τ_{spat}^2 and $\boldsymbol{\gamma}$ with the following full conditionals.

(a) The full conditionals for the U_i are truncated normals with $U_i|\cdot \sim \text{TN}_{t_1,t_2}(\eta_i, 1)$. The truncation points t_1 and t_2 depend on the observed Y_i. For $Y_i = r$ they are given by $t_1 = \theta_{r-1}$ and $t_2 = \theta_r$.

(b) The full conditionals for the regression parameters $\boldsymbol{\beta}_j$, $j = 1,\ldots,p$, spat are multivariate

Gaussian with covariance matrix and mean given by

$$\begin{aligned} \boldsymbol{\Sigma}_j = P_j^{-1} &= \left(\mathbf{X}_j' \mathbf{X}_j + \frac{1}{\tau_j^2} \mathbf{K}_j \right)^{-1}, \\ \boldsymbol{\mu}_j &= \boldsymbol{\Sigma}_j \mathbf{X}_j' (\mathbf{U} - \tilde{\boldsymbol{\eta}}), \end{aligned} \tag{3.7}$$

where $\tilde{\boldsymbol{\eta}}$ is the part of the predictor $\boldsymbol{\eta}$ that is associated with the remaining effects in the model.

(c) The full conditionals for the linear effects parameters $\boldsymbol{\gamma}$ are Gaussian with mean and covariance matrix given by

$$\begin{aligned} \mu_\gamma &= (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}' (\mathbf{U} - \tilde{\boldsymbol{\eta}}), \\ \Sigma_\gamma &= (\mathbf{W}'\mathbf{W})^{-1}. \end{aligned} \tag{3.8}$$

(d) The full conditionals for the variance parameters τ_j^2 are inverse gamma with parameters $a_j' = a_j + \text{rank}(\mathbf{K}_j)/2$ and $b_j' = b_j + \frac{1}{2} \boldsymbol{\beta}_j' \mathbf{K}_j \boldsymbol{\beta}_j$.

(e) The full conditional for threshold θ_r, $r = 1, 2$, is uniform on the interval

$$[\max\{U_i : Y_i = r\}, \min\{U_i : Y_i = r+1\}].$$

Posterior samples from these uniform distributions may exhibit bad mixing, because intervals can become quite small and, as a consequence, the chain moves slowly. Following Chen and Dey (2000) we therefore reparameterized the model. First, inclusion of a constant γ_0 in the predictor allows us to set $\theta_1 = 0$. Secondly, because parameters in the predictor of the latent Gaussian model are identifiable only up to a multiplicative factor, we assume that errors ε_i are $N(0, \sigma^2)$ distributed with unknown variance σ^2. This allows us to set $\theta_1 = 1$. For σ^2 we specify an inverse gamma prior, leading to posterior samples from an inverse gamma full conditional. The results in Section 4 are given in the original parameterization which is obtained by simply dividing sampled parameters in the MCMC simulation by the current value of the standard deviation σ.

Since all the full conditionals are known distributions we can use a Gibbs sampler, drawing successively random numbers from the conditional distributions of the parameters. Numerical efficiency is obtained by utilizing the band matrix structure of the posterior precision matrices $\mathbf{P}_j$ of the regression parameters $\boldsymbol{\beta}_j$. A referee asked why the parameter vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p, \boldsymbol{\beta}_{\text{spat}}$ and $\boldsymbol{\gamma}$ are not updated in one large block. In principle, this is possible because the full conditional is Gaussian. However, the simple band matrix structure of the full conditionals would be lost, leading in most cases to computationally infeasible MCMC algorithms. A more detailed exposition of the methodology can be found in Fahrmeir and Lang (2001b), Lang and Brezger (2004) and Brezger and Lang (2006a). The numerical details of the band matrix algorithms that are used in this paper are described in George and Liu (1981).

The choice between (a small number of) competing models can be done by using the deviance information criterion (DIC); see Spiegelhalter *et al.* (2002). The DIC is defined as $\text{DIC} = \bar{D} + p_D$ where $\bar{D}$ is the posterior mean deviance and p_D is the difference between the posterior mean deviance and the deviance evaluated at the posterior mean of $\mathbf{Y}$. p_D serves as a measure for the effective number of parameters in the model. The DIC can be easily computed as a by-product of an MCMC sampler. Unfortunately the DIC itself is subject to sampling error. Hence, a single MCMC run does not provide credible intervals or other measures of accuracy. Credible intervals could be obtained by running the sampler several times, which turns out to be computationally feasible for the models of this paper.

The DIC has been suggested as a Bayesian analogue to the Akaike information criterion AIC and is now widely used for model choice in complex hierarchical Bayesian models; see for example Jin *et al.* (2005) and Congdon (2006) for recent applications in models that are similar to our setting. As a referee pointed out, there is no rigorous justification for the usage of the DIC in such complex semiparametric models as in our application or in the references that were cited above. Hence, the usage of the DIC is at least debatable. However, some simulation results (Brezger and Lang, 2006b) suggest that the DIC gives reasonable results even in complex nonparametric regression models.

3.4. Prediction based on Markov chain Monte Carlo sampling

An important problem in the analysis of forest health surveys is the prediction of needle losses for locations that are not covered in the survey. A further complication is the fact that the values for the chemical explanatory variables such as Mn, Ca and N/K are also missing for some locations. We proceed as follows.

(a) We first compute predictions for the missing values of the explanatory variables. Since the observed values of the chemicals Mn, Ca and N/K obey a strong spatial pattern, predictions are obtained by spatially smoothing the observed values.
(b) In a second step we replace the missing values of the chemicals by their predictions, estimate a cumulative threshold model for the needle losses and obtain predictions for locations that are not covered in the survey.

Prediction of unobserved responses can be done in a relatively straightforward way by treating them as additional unknown parameters. Suppose that x_i^-, w_i^- and c_{1i}^- and c_{2i}^-, $i=1,\ldots,m$, are the covariate values for m additional locations that are not included in the data. Then the latent variables U_i^- that are associated with these covariate values are given by

$$U_i^- = \eta_i^- + \varepsilon_i^-, \qquad i=1,\ldots,m, \tag{3.9}$$

where the predictor η_i^- depends on the covariate values and the regression parameters. Bayesian inference for the regression parameters is now based on the posterior or predictive density given the observed responses $\mathbf{Y}$. It is given by

$$\begin{aligned} p(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p,\boldsymbol{\beta}_{\text{spat}},\tau_1,\ldots,\tau_p,\tau_{\text{spat}},\boldsymbol{\gamma},\mathbf{U},\mathbf{U}^-|\mathbf{Y}) \\ \propto p(\mathbf{Y}|\mathbf{U})\, p(\mathbf{U}|\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p,\boldsymbol{\beta}_{\text{spat}},\boldsymbol{\gamma})\, p(\mathbf{U}^-|\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p,\boldsymbol{\beta}_{\text{spat}},\boldsymbol{\gamma}) \\ \times \prod_{j=1}^{p} p(\boldsymbol{\beta}_j|\tau_j)\, p(\tau_j)\, p(\boldsymbol{\beta}_{\text{spat}}|\tau_{\text{spat}}^2)\, p(\tau_{\text{spat}}^2)\, p(\boldsymbol{\gamma}), \end{aligned}$$

where $\mathbf{U}^-$ is the vector of latent variables for the unobserved locations. The full conditionals for the latent variables U_i^- are determined by equation (3.9), i.e. $U_i^-|\cdot \sim N(\eta_i^-, \sigma^2)$. The full conditionals for the remaining parameters are identical to those which were given in the previous section. The only difference is that the vector of all latent variables now contains values for observed and unobserved locations. Correspondingly, all design matrices are composed of covariates for observed and missing responses.

In an analogous way predictions for the unobserved covariate values of Mn, Ca and N/K are obtained. We briefly illustrate the approach for covariate Mn. We estimate the Gaussian regression model

$$\text{Mn}_i = f_{\text{spat,Mn}}(c_{i1}, c_{i2}) + \varepsilon_{\text{Mn},i}, \qquad i=1,\ldots,n,$$

i.e. the observed values are spatially smoothed. For the spatial effect a two-dimensional P-spline is assumed. The errors are assumed to be independent and identically distributed Gaussian, i.e. $\varepsilon^-_{\mathrm{Mn},i} \sim N(0, \sigma^2_{\mathrm{Mn}})$. By analogy, we obtain

$$\mathrm{Mn}_i^- = f_{\mathrm{spat,Mn}}(c_{i1}^-, c_{i2}^-) + \varepsilon^-_{\mathrm{Mn},i}, \qquad i = 1, \ldots, m, \tag{3.10}$$

for the missing values. In this model the regression parameters for the spatial effect, the overall variance parameter and the missing values are unobserved 'parameters'. Bayesian inference is done via a Gibbs sampler as described primarily in Lang and Brezger (2004). The missing values Mn_i^- are updated by drawing from equation (3.10). The full conditional for the regression parameters $\boldsymbol{\beta}_{\mathrm{spat,Mn}}$ is Gaussian with mean and covariance matrix analogous to expression (3.7) with the vector of latent variables replaced by the vector of observed and unobserved values of Mn. The full conditional of σ^2_{Mn} is inverse gamma. Once we have estimated the model, the posterior mean of Mn_i^- is taken to replace the missing values in the cumulative threshold model for needleloss.

3.5. Inference based on a mixed model representation

As an alternative to MCMC simulation, mixed model technology could be used for (empirical) Bayesian inference. This is possible because the models that were discussed in Sections 3.1 and 3.2 can be equivalently written as generalized additive mixed models; see Fahrmeir *et al.* (2004) and also Lin and Zhang (1999) for models with univariate responses. The key reference for cumulative threshold models as considered in this paper is Kneib and Fahrmeir (2006). Good introductions to generalized additive mixed models from first principles can be found in Wand (2003), Ruppert *et al.* (2003) and Wood (2006). Given the variance or smoothing parameters, the regression parameters are estimated by maximum likelihood. The smoothing parameters are obtained via (approximate) restricted maximum likelihood. The two estimation steps are iterated until convergence. The advantage of the mixed model representation lies in the unified and simultaneous estimation of linear and smooth covariate effects as well as spatial heterogeneity. Another important aspect is that the smoothing parameters can be estimated simultaneously with the other parameters. Finally, it is (in principle) possible to use existing software for mixed models; see Ngo and Wand (2003). A general procedure for doing this is given in Wood (2004) and is the basis for the `gamm` routine in the R (R Development Core Team, 2004) package `mgcv`. Since the software for mixed models is not designed for estimating generalized additive models, the usage is often rather slow. Moreover, software for fitting cumulative threshold models with a random-effects predictor is not available. The approach for cumulative models that was introduced by Kneib and Fahrmeir (2006) is again included in BayesX. Although state of the art algorithms are used in the current implementation, the computational complexity is of order param^3 where param is the number of regression and random-effects parameters. For this reason full Bayesian inference based on MCMC methods is more efficient in terms of computing time for the models that are considered in this paper. Whereas the fully Bayesian approach used approximately 9 min for estimating the model that is described in the next section, the empirical Bayes approach took more than 84 min.

4. Results

4.1. Model selected

Because of the large number of covariates, we used—as a starting-point—traditional variable selection procedures to restrict covariates to the most relevant ones. The stepwise selection pro-

cedure of Stata was applied to the cumulative threshold model with a probit link. The variable age is known to be an important covariate for defoliation (Federal Research Centre for Forestry and Forest Products, 2001) and was therefore forced into the model at all times. The variable slopegrad was excluded owing to the high correlation with relief. The remaining set of the 25 covariates that were described in Section 2 and given in Table 1 were used in the selection. To consider possible non-linear effects, each continuous variable was entered with a linear, quadratic and cubic polynomial. The levels of significance for including and removing effects was $p_{\mathrm{IN}} = 0.01$ and $p_{\mathrm{OUT}} = 0.05$.

The initial selected model contained the variables age, altitude, Mn, Ca, N/K, geolnr, nutrbal and slopedir. This model was then refitted in a Bayesian framework based on MCMC simulation techniques. An additional spatial random effect as described in equation (3.2) was included and now *P*-splines were assumed for the effects of the continuous covariates. More specifically, we used the geoadditive predictor

$$\begin{aligned}\eta = {} & \gamma_0 + f_1(\text{age}) + f_2(\text{altitude}) + f_3(\text{Mn}) + f_4(\text{Ca}) + f_5(\text{N/K}) + f_{\text{spat}}(c_1, c_2) \\ & + \gamma_1\,\text{geolnr2} + \gamma_2\,\text{geolnr4} + \gamma_3\,\text{geolnr5} + \gamma_4\,\text{geolnr6} + \gamma_5\,\text{geolnr7} + \gamma_6\,\text{geolnr8} \\ & + \gamma_7\,\text{nutrbal1} + \gamma_8\,\text{nutrbal3} + \gamma_9\,\text{slopedir0} + \gamma_{10}\,\text{slopedir7} \end{aligned} \tag{4.1}$$

where we have introduced dummy variables for the categorical covariates.

To assess the dependence of results on the hyperparameters a_j and b_j of variance components τ_j^2, we estimated the model with three different choices: $a_j = 1$ and $b_j = 0.005$, $a_j = b_j = 0.001$ and $a_j = b_j = 0.0001$. For the present model, the differences between the results were very small, so we present results for only our standard choice of $a_j = b_j = 0.001$.

Inspection of the credible intervals and the size of the effects of altitude, Mn, Ca and N/K in Fig. 2 reveal uncertainty about the relevance of these covariates, although all effects can be plausibly interpreted. Therefore, we additionally estimated all possible submodels where one or more of the covariates mentioned are omitted and we compared the results via the posterior distribution of the DIC.

For completeness, we stress that we also tested interactions between covariates. Owing to the high dimensionality of the data, only some very plausible interactions have been considered. Plausible interactions are for example between nutrition balance (nutrbal) and Ca or between nutrition balance and Mn. Modelling and estimation of interactions is well supported by our modelling framework and software. For example a possible interaction between nutrition balance and Ca could be modelled within the varying-coefficient framework that was introduced by Hastie and Tibshirani (1993); see also Brezger and Lang (2006a) for a Bayesian treatment. However, for all tested interaction models the DIC is considerably worse compared with the best fitting models. To keep the paper reasonably brief, models with interactions are therefore not discussed further.

The model with all covariates included except for altitude yields the smallest mean posterior DIC of 1772.99 with a 95% credible interval (1771.94, 1774.32). In comparison the model including all covariates (age, altitude, Mn, Ca, N/K, geolnr, nutrbal and slopedir) yields a mean posterior DIC of 1775.35 with 95% credible interval (1774.32, 1776.36). An obvious reason for the weakness of the effect of the spatially correlated variable altitude is the inclusion of the spatial effect, in which the effect of altitude seems to be subsumed to some extent. To check this confounding we additionally estimate all possible submodels without a spatial effect in the model and compare results via the posterior distribution of the DIC. This confirms that altitude is an important predictor in the model without the spatial effect, since model (4.1) (without the spatial effect) yields the lowest mean posterior DIC of 2076.74 with 95% credible interval (2076.14, 2077.89). The model without altitude in comparison yields a substantially

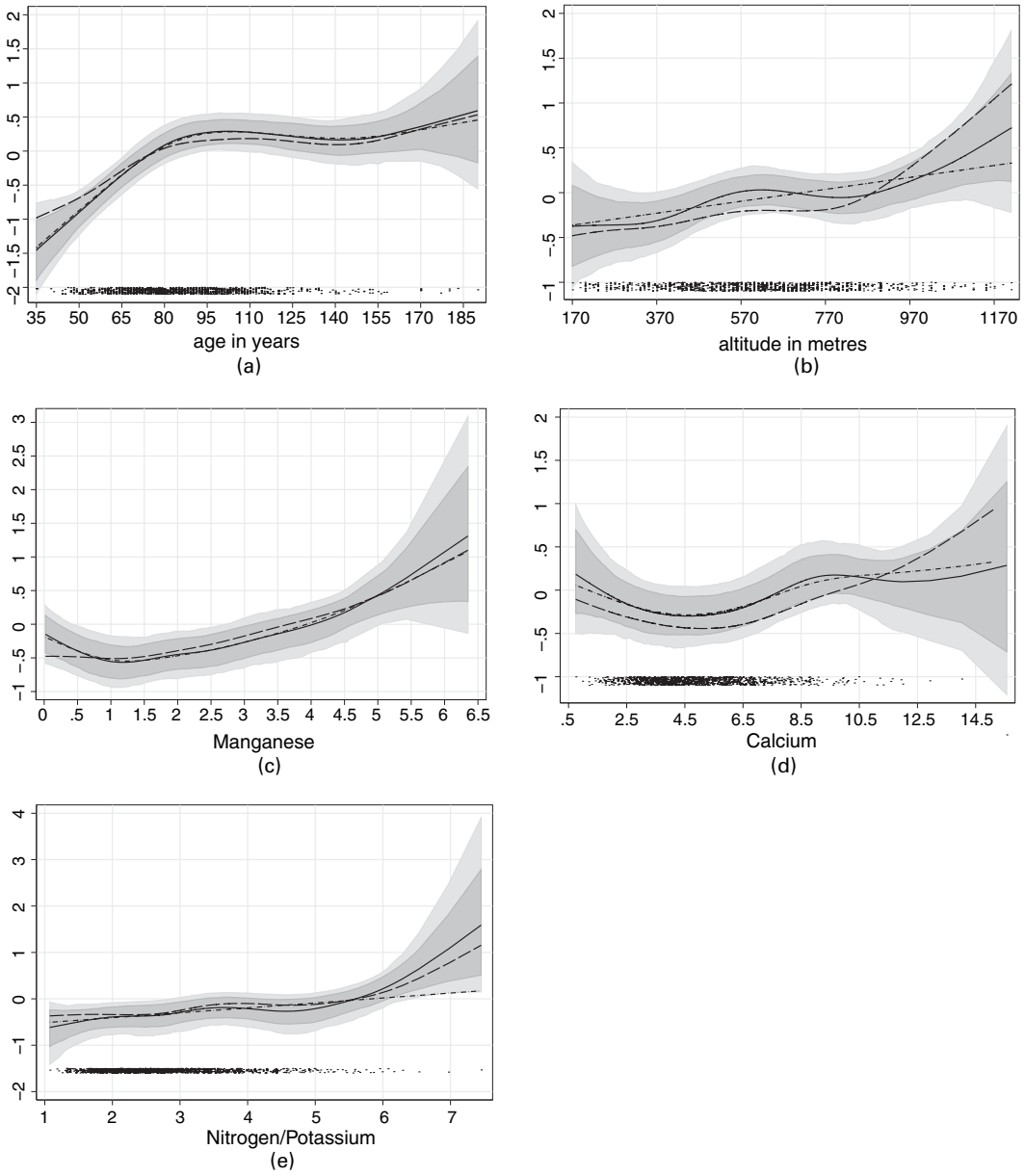

Fig. 2. Non-linear effects of (a) age, (b) altitude, (c) Mn, (d) Ca and (e) N/K with observation points superimposed: shown are the posterior means together with 95% and 80% pointwise MCMC interval estimates for the final model with spatial effect; for comparison, for the same model posterior mode estimates based on mixed model technology (restricted maximum liklihood) are included (-------); in addition, posterior means for the non-spatial model are shown (— —)

higher mean DIC of 2092.29 with the 95% credible interval (2091.5, 2092.85). In fact, dropping either Mn, Ca or N/K from the model selected yields a lower mean DIC than dropping altitude. In addition this exercise shows that the spatial effect improves the model significantly, since the initially selected model yields a significantly lower DIC (1775.35) than the model without spatial effect (DIC 2076.74). This implies that covariates alone do not adequately explain all

Fig. 3. Indirect diagnostic residual plot: representation of the posterior spatial effect f_{spat} in the covariates as offset model—posterior probablities for a nominal level of 95% based on the original data (■, regions with strictly negative interval estimates; □, regions with strictly positive interval estimates)

the spatial correlation in the data. Where this is so can be figured out from the indirect residual diagnostic plot that is shown in Fig. 3. We estimate the spatial effect while keeping the covariate effects fixed via an offset, which is the posterior mean of the linear predictor of the selected model (4.1) *without* the spatial effect. This model will be referred to as the 'covariates as offset' model in what follows. There is a band stretching from north-west to south-east with strictly negative 95% interval estimates of the spatial effect indicating that covariates do not adequately describe the variation in the data. There are also some clustered locations with strictly positive 95% interval estimates of the spatial effect, again indicating that the model without spatial effect does not entirely account for the spatial correlation. Overall, these results indicate that model (4.1) is adequate. In what follows we shall refer to model (4.1) as the 'spatial model' and to model (4.1) without spatial effect as the 'non-spatial model'.

4.2. Interpretation of categorical covariates

Fig. 4 shows summary statistics of the posterior distribution estimated for the spatial and non-spatial model for the categories. For comparison, posterior modes estimated by using restricted maximum likelihood for the spatial model as described in Section 3.5 are added. Both inference approaches give quite similar results here. High values of the posterior effect indicate a high probability of the tree being damaged. Overall, the effects of the spatial and non-spatial model are similar, but weaker for most of the factors. Also, the interval estimates for the non-spatial model are narrower.

4.2.1. Interpretation of geolnr

Here the coefficients are compared with the reference category *geolnr1* (metamorphic crystalline bedrock). Fig. 4(a) shows the (Bayesian) interval estimates. Since the geological categories

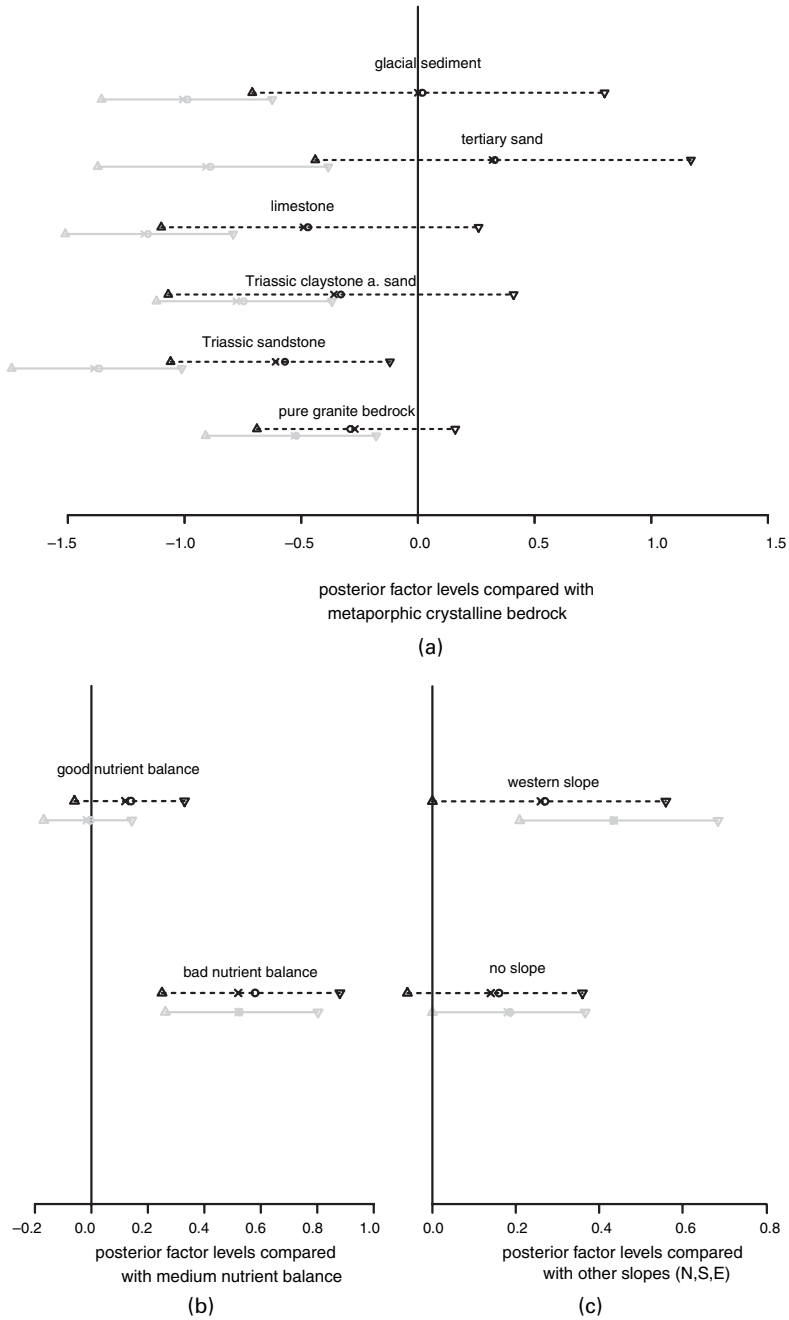

Fig. 4. MCMC interval estimates and median for posterior distributions of factor levels (results from the spatial model (4.1) are shown in black and results from the non-spatial model are shown in grey): (a) geological area compared with the reference category metamorphic crystalline bedrock geol1; (b) nutrient balance compared with the reference category medium nutrient balance; (c) direction of slope compared with the reference category other slopes (northern, southern or eastern) (black indicates estimates from the final model including the spatial effect; grey indicates estimates from the final model excluding the spatial effect; for comparison, posterior mode estimates based on mixed model technology (restricted maximum likelihood) are included in the plots marked by crosses)

split the sampling area into sections (Fig. 1), we have apparent confounding of effects here. The model cannot distinguish between a spatial effect—a proxy for some other unaccounted for effect—and the geological effect. The only posterior coefficient that is significantly different in the spatial model is geolnr4 (Triassic sandstone) with the lowest negative effect, implying that Triassic sandstone has a significantly reduced effect on the probability of damage compared with the reference category, metamorphic crystalline bedrock. The effect is stronger in the case for the non-spatial model. Triassic sandstone mainly occurs in the northern Black Forest, where we find many healthy trees in 1994. It is acidic, and alkaline nutrients are easily washed out, and hence we would naturally expect a result indicating the opposite—a damaging effect. A possible explanation for this surprising result is the fact that, as a countermeasure, forest liming campaigns began in the early 1980s since forest deterioration was first observed mainly in the areas with Triassic sandstone. We have no explicit covariate indicating the liming activity: hence this result. The categories pure granite bedrock, limestone and Triassic claystone and sands also have median negative effects, which are significant for the non-spatial model. These median negative effects imply a reduced probability of damage compared with metamorphic crystalline bedrock. The median posterior effect of tertiary sands is the highest in the spatial model, unlike in the non-spatial model where its median effect is negative.

4.2.2. Interpretation of nutrbal

Fig. 4(b) shows the interval estimates for the posterior effects. Here the results from the spatial and non-spatial model are very similar, indicating that there is no confounding with the spatial effect. A tree growing on soil with bad nutrient balance has a significantly higher effect, implying a higher probability for damage than a tree with medium nutrient balance, the reference category. The effect of a good nutrient balance is not significantly different from that of a medium nutrient balance for the spatial model.

4.2.3. Interpretation of slopedir

Fig. 4(c) shows that trees growing on west facing slopes have a significantly higher effect than trees on slopes facing all other directions. Hence for trees growing on the western 'weather' side, with high exposure to wind and rain, the effect on probability for damage is increased. In contrast, for trees growing on the flat (no slope) there is no significantly higher effect for damage. Again, results from the spatial and non-spatial model are similar.

4.3. Interpretation of non-linear effects

The effects of the continuous covariates estimated by using the spatial model are displayed in Fig. 2. We have also added the effects from the non-spatial model (long broken curves), and in general the effects are similar but stronger for the non-spatial model. The following interpretation refers to the spatial model results. For comparison with the fully Bayesian approach, posterior mode estimates based on restricted maximum likelihood for the spatial model are added (short broken curves). In general, the restricted maximum likelihood results are close to the MCMC results although there are some differences. The effect of the continuous covariate altitude is estimated to be almost linear with the mixed model approach using restricted maximum likelihood. This ties in with the MCMC result, where the posterior means show a slight discrepancy from linearity, but the pointwise 95% MCMC interval estimates do not indicate a significant non-linearity. Nevertheless, pointwise 95% MCMC interval estimates (which are not shown) of the non-spatial model indicate significant non-linearity. Similarly, the effect of the ratio N/K is estimated to be almost linear with the mixed model approach, whereas the estimates that are based on MCMC sampling show a slight discrepancy from linearity. Again

pointwise 95% MCMC interval estimates do not indicate a significant non-linearity of N/K. The continuous covariates age and Mn that are shown in Fig. 2 have a pronounced non-linear effect.

Age is an important predictor for needle loss and, as expected, younger trees have a significantly negative effect on the probability of needle loss and with age the effect on the probability of needle loss increases (Fig. 2). Concerning the altitude, the positive effect on the probability of damage increases considerably above 800 m (Fig. 2). These high regions are mostly in the southern Black Forest (Feldberg), where we have a high proportion of damaged trees.

Now we investigate the posterior effects of the selected nutrients in the needle. The effect on the probability of damage increases sharply with higher values of Mn (Fig. 2). Mn is mobilized in the soil when acidification takes place and hence might be available to the tree in higher quantities when the soil is in a transient state of acidification (Hildebrand, 1986; Feretti *et al.*, 2002). Soil acidification is accompanied by a surplus of Mn. The posterior mean effect of Mn is increasing and positive for Mn levels that are above about 5 g kg^{-1}. This can be interpreted as a damaging effect of Mn in surplus accompanying soil acidification. This is in line with the conventional threshold of 4 g kg^{-1} of needles, an Mn concentration above which it is assumed to be toxic (Mengel, 1991). Thus the increasing effect of Mn on the probability of needle loss could be an indirect indicator for increasing effects of soil acidification. N is an essential plant nutrient, but a surplus of N destabilizes the balance of other nutrients. In the literature the upper threshold for a harmonic N/K ratio is assumed to be 3. In Fig. 2 the effect on the probability of damage is positive above a ratio of 6.5, indicating a strong imbalance between the supply of N and K. Ca is a nutrient for the tree, and for spruce the optimal value in the needles is between 2.0 and 5.0 g kg^{-1} of needles. Fig. 2 indicates that the effect on the probability of damage is lowest in this range.

4.4. Interpretation of the spatial effect

Fig. 5 shows the spatial effect (the model in Section 4.1) with strictly negative 95% interval estimates in a band stretching from the north-western Black Forest south-east to the Lake of Konstanz (Fig. 5(a)). This indicates that, in this area, where most of the healthy trees are observed, the covariates do not adequately describe the variation in the data. It also indicates that the model that is based only on the covariates might lead to excessively high predicted probabilities of tree damage compared with the observed values. Such a finding may indicate the effects of processes which were not included in the form of explanatory variables in the analysis. For example the small band of strictly negative spatial effects coincides with a small area with by far the lowest deposition load in the wind shadow of the Black Forest (Forstliche Versuchs- und Forschungsanstalt Baden-Württemberg, 1994). That could be one possible explanation for the fact that the probabilities of damage that are predicted by the model (without the spatial effect) are in excess of the observed relatively good forest condition. Moreover in the northern Black Forest (the north-western part of that above-mentioned band) most intensive forest liming campaigns had been performed. These are assumed to provide a mitigating effect on forest damage. Thus the spatial effect provides a valuable tool for interpretation and for generating hypotheses for further investigations.

In the area where the interval estimates are entirely positive it is the other way round. The area of the strictly positive credible region is mainly in the southern part of the Black Forest where the majority of damaged trees are observed. The many regions with strictly positive or negative 95% credible intervals show that the spatial effect is needed in the model; the covariates alone cannot describe the observed pattern in needle loss. A comparison with the spatial effect as estimated in the covariates as offset model in Fig. 3 shows a similar but weaker pattern due to the confounding of the spatial effect with some covariates, in particular altitude and geolnr.

Fig. 5. Representation of posterior spatial effect f_{spat} in the selected model (4.1) (spatial model) (■, regions with strictly negative interval estimates; □, regions with strictly positive interval estimates; posterior mode estimates based on mixed model technology have been omitted because they are almost identical to the full Bayesian approach based on MCMC simulations): (a) posterior probabilities for a nominal level of 95% based on the original data; (b) posterior probablities for a nominal level of 95% based on observed and interpolated covariates

4.5. Prediction

For producing spatial predictions of defoliation in unsampled locations, interpolation of the covariates measured at the tree level, and hence not available at unsampled locations, is necessary. The missing tree-specific explanatory variables are the nutrients Mn, Ca and N/K as selected in the model (4.1). All other covariates selected are available for the additional locations. The interpolation was carried out separately for these three covariates by using a two-dimensional P-spline as described in Section 3.4. Once the missing covariate values have been obtained, the selected model (4.1) was applied to the observed and interpolated covariates as described in Section 3.4. Fig. 5(b) shows the spatial effect. This is very similar to the same graph (Fig. 5(a)) for the data without predictions with a strictly negative 95% Bayesian prediction interval in a band stretching from the north-western Black Forest south-east to the Lake of Konstanz. Finally, Fig. 1(d) shows for each location the probability that the category damaged has the highest posterior probability as estimated from the MCMC samples. When this is compared with the observed health status in Fig. 1(b) a good model fit is apparent.

5. Discussion

The Bayesian cumulative threshold model with a geoadditive predictor provides a flexible tool for modelling the needle loss data. The methodology and software even allow the extension to more complex models with a structured additive predictor. Models of this kind can be used to incorporate additionally unit- or cluster-specific heterogeneity effects or complex interactions (e.g. within varying-coefficient terms). The data-driven smoothness parameter selection for the non-linear effects of continuous covariates is also a great advantage.

As indicated in Section 4, in our forestry survey data there may be confounding between the spatial effect, geolnr and altitude. In general, the nature of spatial survey data makes confounding between the spatial effect and covariates inevitable. Such confounding usually concerns covariates which are spatially correlated. It is not clear how this problem can be avoided with spatial survey data. Further work to investigate the problems of confounding with spatial survey data is necessary.

The hierarchical model that is presented here is a very useful tool for prediction. Spatial predictions of defoliation in unsampled locations for maps of needle losses may be required for forest management. However, given that some of the covariates are measured at the tree level and are not available at unsampled locations, this is not straightforward. We have used a two-step procedure to carry out prediction: we first compute predictions for the missing values of the explanatory variables and then replace the missing values in the explanatory variables by their predictions. We plan to extend our model to incorporate simultaneous prediction of explanatory variables. Such a hierarchical model has the positive side-effect that the measurement and prediction error of explanatory variables is automatically incorporated in the prediction of the final response variable. Our model also has a use for planning future surveys. By means of simulation from the model, alterations to the current survey design and their effect on the precision of estimates can be assessed.

The main findings of the model are that stand-specific covariates regarding topography, soil and geology have a strong influence on the probability of needle loss. These covariates include the type of geology, nutrient balance, slope direction and altitude. The effect of altitude on the probability of damage is increasing and positive above approximately 800 m. This agrees with the assumption that the meteorological inversion layer in mountainous regions with its excessively high acid input from rain and fog precipitation is a main cause of needle loss (Schöpfer and Hradetzky, 1984). Also, western slopes have a significantly increased effect on the probability of needle loss compared with other slope directions. Other stand-specific covariates regarding meteorology were not selected. From the tree-specific covariates age was forced into the model because it is known to be a strong predictor. The posterior of the age effect confirms the strong non-linear age effect. From the other tree-specific covariates regarding nutrients in the needles, Mn, Ca and the ratio N/K were selected. It is interesting to see that the posterior assessment of the effects of these variables roughly ties in with disequilibrium thresholds that are conventionally assumed in forestry science (see Section 4.3). The interpretation of the spatial effect indicates that information on the spatial distribution of deposition load as well as an indicator on liming activity should be included in the model. Nevertheless this was not possible because no reliable data are available at the level of resolution of the monitoring scheme in space and time.

Acknowledgements

Part of the work was carried out under German Federal Ministry of Education and Research project 0339985. We are grateful to M. Titterington, the Associate Editor and two referees for useful and constructive comments.

Appendix A: Estimation with BayesX

In this appendix we briefly demonstrate how the cumulative threshold model with geoadditive predictor (4.1) can be estimated with BayesX. The program is command line driven, i.e. the user must enter a number of commands to estimate regression models. The statements can be combined and collected in a batch file which allows us to execute all statements in one step. The following statements are part of a batch file and are required to estimate the cumulative threshold model.

```
delimiter = ;

dataset d;

d.infile using c:\data\forest_spruce.raw;

map m
m.infile using c:\data\map_forest_spruce.bnd;

bayesreg b;
b.regress needleloss = geolnr2 + geolnr4 + geolnr5 + geolnr6 + geolnr7 +
geolnr8 + nutrbal1 + nutrbal3 + west0 + west7 + region(geospline,
map=m) + age(psplinerw2) + altitude(psplinerw2) + Mn(psplinerw2) +
Ca(psplinerw2) + NK(psplinerw2), predict family=cumprobit
iterations=36000 burnin=2000 step=30 using d;

remlreg r;
b.regress needleloss = geolnr2 + geolnr4 + geolnr5 + geolnr6 + geolnr7 +
geolnr8 + nutrbal1 + nutrbal3 + west0 + west7 + region(geospline,
map=m) + age(psplinerw2) + altitude(psplinerw2) + Mn(psplinerw2) +
Ca(psplinerw2) + NK(psplinerw2), family=cumprobit using d;
```

The first statement sets the delimiter to the ';' sign rather than the return key and allows us to split subsequent commands into several lines.

The next two statements create a data set object 'd' and read in the data using the infile command of data set objects. The data are supposed to be stored in the plain text file 'c:\data\forest_spruce.raw', where variables are stored columnwise and observations are separated by blanks or tabs. The first line of the file contains the names of the variables.

In the following two statements a map object 'm' is created and the location of trees is read using the infile command of map objects. This information is required to estimate a spatial effect.

In the next statements we create a `bayesreg` object 'b' and estimate the model with predictor (4.1). The predictor is specified in quite a natural way which is similar to S-PLUS. Several options are specified after the comma. They define the model (a cumulative threshold model) and details about MCMC simulation (the number of iterations, the burn-in period and the thinning parameter). The data set to be used for estimation is given after the keyword 'using'. A couple of post-estimation commands, e.g. to visualize results, are also available. Details can be found in the user manual.

The last two commands re-estimate the model by using mixed model technology and on the basis of restricted maximum likelihood estimates for the variance and smoothing parameters respectively.

If all the statements described above are combined in a batch file which is stored in 'c:\prg\forest.prg' they are executed by typing

```
usefile c:\prg\forest.prg
```

in BayesX.

The latest version of BayesX including three manuals is available at

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

An overview of the capabilities of BayesX is given in Brezger *et al.* (2005a). First steps can be done with a tutorial that can be found at the BayesX home page. Full details are given in the user manuals (Brezger *et al.*, 2005b).

References

Augustin, N., Kublin, E., Metzler, B., Meierjohann, E. and von Wühlisch, G. (2005) Analysing the spread of beech canker. *For. Sci.*, **5**, 438–448.

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

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

Brezger, A., Kneib, T. and Lang, S. (2005a) Bayesx: analysing Bayesian semiparametric regression models. *J. Statist. Softwr.*, **14**, no. 11.
Brezger, A., Kneib, T. and Lang, S. (2005b) Bayesx manuals. *Technical Report*. Department of Statistics, University of Munich, Munich.
Brezger, A. and Lang, S. (2006a) Generalized additive regression based on Bayesian P-splines. *Computnl Statist. Data Anal.*, **50**, 967–991.
Brezger, A. and Lang, S. (2006b) Simultaneous probability statements for Bayesian P-splines. Submitted to *Statist. Modlng*.
Chen, M. and Dey, D. (2000) Bayesian analysis for correlated ordinal data models. In *Generalized Linear Models: a Bayesian Perspective* (eds D. Dey, S. Ghosh and B. Mallick), pp. 133–159. New York: Dekker.
Congdon, P. (2006) A model for non-parametric spatially varying regression effects. *Computnl Statist. Data Anal.*, **50**, 422–455.
Eilers, P. and Marx, B. (1996) Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). *Statist. Sci.*, **11**, 89–121.
Eilers, P. and Marx, B. (2003) Multidimensional calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometr. Intell. Lab. Syst.*, **66**, 159–174.
Fahrmeir, L., Kneib, T. and Lang, S. (2004) Penalized structured additive regression for space-time data: a Bayesian perspective. *Statist. Sin.*, **14**, 731–761.
Fahrmeir, L. and Lang, S. (2001a) Bayesian inference for generalized additive mixed models based on Markov random field priors. *Appl. Statist.*, **50**, 201–220.
Fahrmeir, L. and Lang, S. (2001b) Bayesian semiparametric regression analysis of multi-categorical time-space data. *Ann. Inst. Statist. Math.*, **53**, 11–30.
Fahrmeir, L. and Tutz, G. (2001) *Multivariate Statistical Modelling based on Generalized Linear Models*. New York: Springer.
Federal Research Centre for Forestry and Forest Products (2001) Forest condition in Europe, UNECE and EC. *Technical Report*. Federal Research Centre for Forestry and Forest Products.
Feretti, M., Innes, J., Jalkanen, R., Saurer, M., Schäffer, J., Spieker, H. and von Wilpert, K. (2002) Air pollution and environment chemistry: what role for tree ring studies? *Dendrochronologia*, **20**, 159–174.
Forstliche Versuchs- und Forschungsanstalt Baden-Württemberg (1994) Waldschadensbericht Baden-Württemberg. *Technical Report*. Forstliche Versuchs- und Forschungsanstalt Baden-Württemberg.
George, A. and Liu, J. (1981) *Computer Solution of Large Sparse Positive Definite Systems*. Englewood Cliffs: Prentice Hall.
Hastie, T. and Tibshirani, R. (1993) Varying-coefficient models (with discussion). *J. R. Statist. Soc.* B, **55**, 757–796.
Hildebrand, E. (1986) Zustand und Entwicklung der Austauschereigenschaften von Mineralböden aus Standorten mit erkrankten Waldbeständen. *Forsts Centr.*, **105**, 60–75.
Jin, X., Carlin, B. and Banerjee, S. (2005) Generalized hierarchical multivariate car models for areal data. *Biometrics*, **61**, 950–961.
Kammann, E. E. and Wand, M. P. (2003) Geoadditive models. *Appl. Statist.*, **52**, 1–18.
Kneib, T. and Fahrmeir, L. (2006) Structured additive regression for multicategorical space-time data: a mixed model approach. *Biometrics*, **62**, 109–118.
Lang, S. and Brezger, A. (2004) Bayesian P-splines. *J. Computnl Graph. Statist.*, **13**, 183–212.
Lin, X. and Zhang, D. (1999) Inference in generalized additive mixed models by using smoothing splines. *J. R. Statist. Soc.* B, **61**, 381–400.
Meining, S., Schröter, H. and von Wilpert, K. (2003) *Waldzustandsbericht 2003*. Freiburg: Fortstliche Versuchs- und Forschungsanstalt Baden-Württemberg.
Mengel, K. (1991) *Ernährung und Stoffwechsel der Pflanze*, p. 466. Jena: Fischer.
Ngo, L. and Wand, M. (2003) Smoothing with mixed model software. *J. Statist. Softwr.*, **9**, 2–3.
Preisler, H., Rappaport, N. and Wood, D. (1997) Regression methods for spatially correlated data: an example using beetle attacks in a seed orchard. *For. Sci.*, **43**, 71–77.
R Development Core Team (2004) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Ruppert, D., Wand, M. and Carroll, R. (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.
Schöpfer, W. and Hradetzky, J. (1984) Der Indizienbeweis: Luftverschmutzung massagebliche Ursache der Walderkrankung. *Forsts Centr.*, **103**, 231–248.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Wand, M. P. (2003) Smoothing and mixed models. *Computnl Statist.*, **18**, 223–249.
Wood, S. N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Statist. Ass.*, **99**, 673–686.
Wood, S. (2006) *Generalized Additive Models: an Introduction with R*. Boca Raton: Chapman and Hall–CRC.
Wood, S. N. and Augustin, N. H. (2002) Gams with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecol. Modllng*, **157**, 157–177.