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

Ludwig Fahrmeir, Thomas Kneib and Stefan Lang

Abstract

We propose extensions of penalized spline generalized additive models for analyzing space-time regression data and study them from a Bayesian perspective. Non-linear effects of metrical covariates and time trends are modelled through Bayesian versions of penalized splines, while correlated spatial effects follow a Markov random field prior. This allows to treat all functions and effects within the same general framework by assigning appropriate priors with different forms and degrees of smoothness. Inference is based on a generalized linear mixed model representation. This approach 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 marginal quasi-likelihood.

Bayesian structured additive regression

Consider regression situations, where observations $(y_i, x_i, u_i),\ i = 1, \dots, n,$ on a response y, a vector $x = (x_1, \dots, x_p)$ of metrical covariates, time scales or spatial covariates and a vector u of further covariates are given. Generalized additive and semiparametric models (Hastie and Tibshirani, 1990) assume that, given x_i and u_i, the distribution of y_i belongs to an exponential family, with mean $\mu_i = E(y_i|x_i, u_i)$ linked to an additive semiparametric predictor η_i by

$$\mu_i = h(\eta_i),\ \eta_i = f_1(x_{i1}) + \dots + f_p(x_{ip}) + u_i'\gamma. \tag{1}$$

Here h is a known response function, and $f_1, \dots, f_p$ are unknown smooth functions of the covariates.

Priors for a function

Let $f_j = (f(x_{1j}), \dots, f(x_{nj}))$ be the vector of corresponding function evaluations at the observed values of x_j. In the following we will always be able to write f_j as the matrix product of a design matrix X_j and a vector of unknown regression parameters β_j, i.e.

$$f_j = X_j\beta_j.$$

Similarly, priors for β_j can be brought into a general form as well. The general form of the prior is given by

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

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

A particular prior depends on the *type of the covariate* and on *prior beliefs about the smoothness* of f_j. In the following we give some examples:

Metrical covariates and time scales

- A flexible and parsimonious possibility to model non-linear effects of metrical covariates and time scales are P-splines.
- Basic assumption:

$$\begin{aligned} f_j(x) &= \text{spline of degree } l \text{ with equally spaced inner knots} \\ &\quad\ t_1, \dots, t_r \text{ between } x_{min} \text{ and } x_{max} \\ &= \beta_{j1}B_{j1}(x) + \dots + \beta_{j,r+l+1}B_{j,r+l+1}(x), \end{aligned}$$

  where $B_{j1}, \dots, B_{j,r+l+1}$ is a B-spline basis.
- The design matrix X_j consists of the basis functions evaluated at the observations, i.e. $X_j(i,k) = B_{jk}(x_i)$.
- Idea:
  - Define a relatively *large number of inner knots* to guarantee enough flexibility.
  - Assign a *smoothness prior* for $\beta_{j1}, \dots, \beta_{j,r+l+1}$ to penalize too rough functions f_j. This can be achieved through first or second order random walks

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

    with Gaussian errors $u_{jm} \sim N(0, \tau_j^2)$ and diffuse priors for initial values.
  - One obtains $K_j = D_j'D_j$ where D_j is a first or second order difference matrix.
- References: Eilers, Marx (1996), Lang, Brezger (2003), Brezger, Lang (2003).

Spatial covariates

- The values of $x \in \{1, \dots, s, \dots, S\}$ represent the location or site in geographical regions.
- We may estimate only a structured spatially correlated effect or split up the spatial effect into a structured (correlated) and an unstructured (uncorrelated) component:

$$f_j(s) \;=\; f_{j,str}(s) + f_{j,unstr}(s).$$

- Each site is associated with one parameter, i.e.

$$f_{j,str}(s) = \beta_{js}^{str} \text{ and } f_{j,unstr}(s) = \beta_{js}^{unstr}.$$

- A common choice are Markov random fields (Besag, York and Mollié, 1991) for the structured effect, e.g.

$$\beta_{js}^{str}|\beta_{j,-s}^{str}, \tau_j^2 \sim N\left(\sum_{k \in \partial_s} \frac{1}{N_s}\beta_{jk}^{str}, \frac{1}{N_s}\tau_j^2\right), \tag{3}$$

  where

  ∂_s denotes the sites, which are neighbours of site s and
  N_s are the number of neighbours.
- For the unstructured effect f_j^{unstr} we assume that the parameters β_{js}^{unstr} are i.i.d. Gaussian

$$\beta_{js}^{unstr}|\tau_j^2 \sim N(0, \tau_j^2). \tag{4}$$

- For both cases the design matrix X_j is a 0/1 incidence matrix where the number of columns is equal to the number of different sites. If observation i belongs to site s, then the element in the i-th row and the s-th column is one, zero otherwise.

Unordered group indicators

- Suppose x is now a grouping variable with values $x = 1, \dots, m, \dots, M$. The values of x may denote a unit or cluster index, or the location in geographical maps.
- To account for unobserved unit or group specific heterogeneity one possible way is to include an additive random effect into the predictor. Then we assume the unit or group specific effects β_{jm} to be i.i.d. Gaussian,

$$\beta_{jm}|\tau_j^2 \sim N(0, \tau_j^2). \tag{5}$$

- Again the design matrix is a 0/1 incidence matrix and the penalty matrix is the identity matrix I.
- An unstructured spatial effect is a special case of (5) with the regions of a geographical map as the grouping variable.

Possible Extensions

- Interactions between covariates may be modelled through Varying Coefficient Models. This also allows to incorporate random slopes in the model, since models with unordered group indicators as effect modifiers are equivalent to models with random slopes.
- A more flexible approach for modelling interactions between metrical covariates can be based on two dimensional surface fitting. Two dimensional P-splines, defined as the tensor product of two one dimensional B-Splines with a spatial smoothness prior, are described in Lang and Brezger (2003).
- All extensions can be cast into the general form (2) and may therefore be treated using the same methodology as presented here.

Mixed model representation

To rewrite the model (1) as a generalized linear mixed model (GLMM) we proceed as follows:

- Decompose the vectors of regression coefficients β_j into an *unpenalized* and a *penalized* part:

$$\beta_j = X_j^{unp}\beta_j^{unp} + X_j^{pen}\beta_j^{pen}. \tag{6}$$

- X_j^{unp} contains a basis of the nullspace of the penalty matrix K_j and X_j^{pen} is given by $X_j^{pen} = L_j(L_j'L_j)^{-1}$, where L_j is a full column rank matrix with $K_j = L_jL_j'$. A requirement for the choice of X_j^{unp} and X_j^{pen} is that $L_j'X_j^{unp} = (X_j^{unp})'L_j = 0$ holds.
- In general L_j can be obtained from the spectral decomposition $K_j = \Gamma_j\Omega_j\Gamma_j'$ as $L_j = \Gamma_j\Omega_j^{\frac{1}{2}}$, where Ω_j contains the positive eigenvalues of K_j and Γ_j is formed by the corresponding eigenvectors.
- In some cases more favorable decompositions of K_j can be found. For instance, for P-splines one may choose $L_j = D_j'$, where D_j is the first or second order difference matrix.
- For P-splines with first order random walk penalty and Markov random fields X_j^{unp} is simply a vector of ones. For P-splines with second order random walk X_j^{unp} is a two column matrix composed from a vector of ones and the vector of the (equidistant) knots of the spline.
- For unordered group indicators a decomposition is not necessary since $K_j = I$. In this case the unpenalized part vanishes completely.
- From decomposition (6) we get

$$\frac{1}{\tau_j^2}\beta_j'K_j\beta_j = \frac{1}{\tau_j^2}(\beta_j^{pen})'\beta_j^{pen}.$$

  and from the general prior (2) it follows that

$$p(\beta_j^{unp}) \propto const$$

  and

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

- Finally, defining

$$\begin{aligned} \tilde{X} &= (X_1X_1^{pen}\ X_2X_2^{pen}\ \cdots\ X_pX_p^{pen}), \\ \beta^{pen} &= ((\beta_1^{pen})', \dots, (\beta_p^{pen})')' \end{aligned}$$

  and

$$\begin{aligned} \tilde{U} &= (X_1X_1^{unp}\ X_2X_2^{unp}\ \cdots\ X_pX_p^{unp}U), \\ \beta^{unp} &= ((\beta_1^{unp})', \dots, (\beta_p^{unp})', \gamma')'. \end{aligned}$$

  yields a generalized linear mixed model with linear predictor

$$\eta = \tilde{U}\beta^{unp} + \tilde{X}\beta^{pen},$$

  fixed effects β^{unp} and random effects β^{pen} with

$$\beta^{pen} \sim N(0, \Lambda)$$

  and $\Lambda = \text{diag}(\tau_1^2, \dots, \tau_1^2, \dots, \tau_p^2, \dots, \tau_p^2)$.
- The GLMM representation allows to examine the identification problem inherent to nonparametric regression from a different angle: Except for i.i.d. random effects the matrix product $X_jX_j^{unp}$ contains the identity vector and therefore $\tilde{U}$ has not full column rank. So all identity vectors have to be eliminated from $\tilde{U}$ to guarantee identifiability.
- Now we can utilize GLMM methodology for simultaneous estimation of the smooth functions and the variance parameters τ_j^2.
- Especially, variance parameters may be estimated via marginal likelihood. The marginal likelihood of $\tau^2 = (\tau_1^2, \dots, \tau_p^2)$ is defined as

$$l(\tau^2) = \int p(y|\beta^{unp}, \beta^{pen}, \tau^2)p(\beta^{pen})d\beta^{pen}d\beta^{unp}.$$

  For Gaussian response the maximization of $l(\tau^2)$ yields restricted maximum likelihood (REML) estimates. For more general responses a Laplace approximation to $l(\tau^2)$ has to be used.
- Since τ^2 is estimated via marginal likelihood, the estimates $\hat{f}_j$ can be seen as empirical Bayes / posterior mode estimates.
- References: Fahrmeir, Kneib, Lang (2003), Kneib (2003)

Simulation study

We carefully compared the presented empirical Bayes approach with a fully Bayesian approach that uses MCMC techniques for posterior analysis (see Fahrmeir, Lang, 2001a, 2001b, Lang, Brezger, 2003 and Brezger, Lang, 2003) through a simulation study. The results of this simulation study can be summarized as follows:

- In general, the empirical Bayes approach yields better point estimates of the functions f_j in terms of MSE.
- The differences are most noticeably for Bernoulli distributed response and turn out to be smaller for Gaussian, Poisson or Binomial distributed response with at least 3 repeated binary observations.
- The empirical Bayes approach tends to smoother function estimates. This can also be shown theoretically (Kauermann, 2002).
- Coverage probabilities meet the nominal level for smooth functions of metrical covariates. This is not the case for spatial and random effects, where coverages are far from the nominal level.
- Since no problems with coverage probabilities occur in the fully Bayesian analysis a combination of both approaches seems to be promising: The variance components are estimated via marginal likelihood while the function estimates and the credible intervals are obtained from an MCMC analysis, that uses these variance components.
- The combination leads to estimates, that keep the smaller MSE of the empirical Bayes estimates, but inherit the better coverage properties from the fully Bayesian analysis.

Applications

Rents for flats: A spatial study

According to the German law, owners of apartments or flats can base an increase in the amount that they charge for rent on "average rents" for flats comparable in type, size, equipment, quality and location in a community. To provide information about these "average rents", most larger cities publish "rental guides", which can be based on regression analysis. In our first application we use data from the city of Munich collected in 2002 for a sample of approximately 3000 flats.

As response variable we choose the monthly net rent per square meter in German Marks R. Covariates are given as follows:

F floor space,
Y year of construction,
L location of the flat in Munich,
u vector of 25 further (binary) covariates.

For our analysis we choose a geoadditive Gaussian model

$$R = \eta + \varepsilon$$

with predictor

$$\eta = \gamma_0 + f_1(F) + f_2(Y) + f_3(L) + u'\gamma.$$

The effects f_1 and f_2 of floor space and year of construction are modelled by cubic P-splines with 20 knots and a second order random walk penalty. For the spatial effect $f_3(L)$ we choose the Markov random field prior (3).

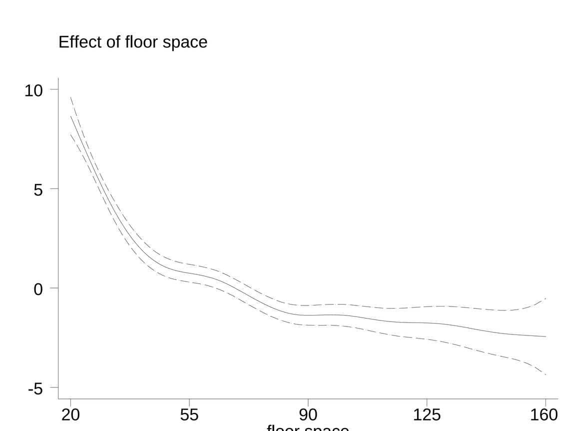

Figure 1: Estimated effect of floor space with pointwise 95% credible intervals.

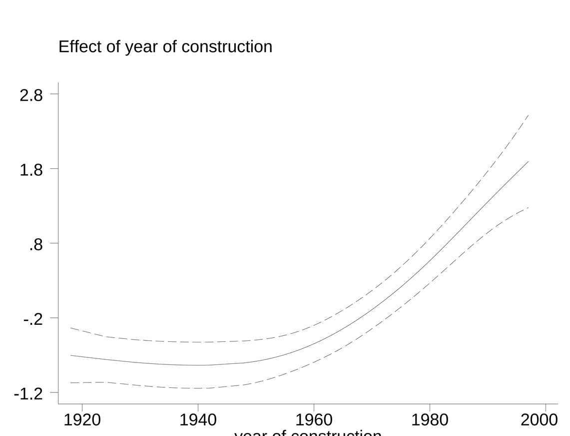

Figure 2: Estimated effect of year of construction with pointwise 95% credible intervals.

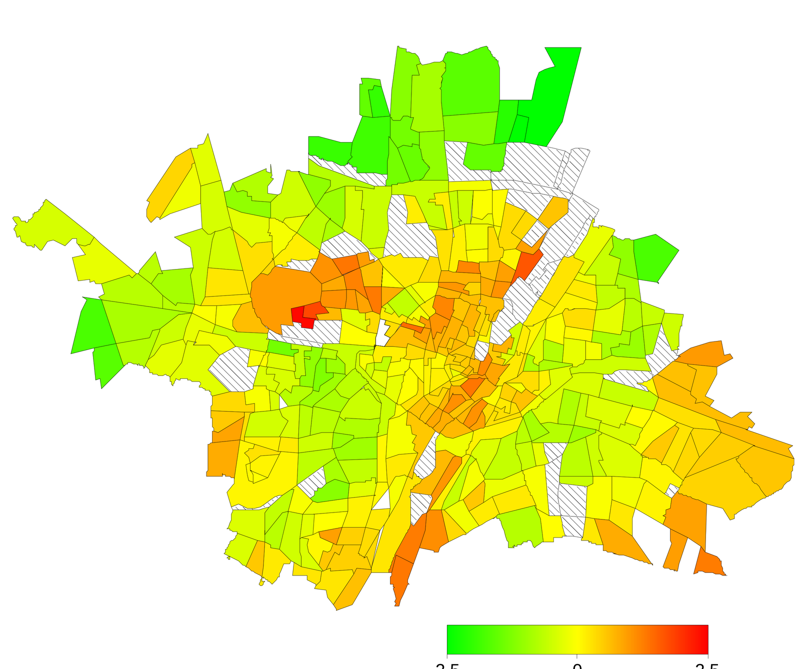

Figure 3: Estimated spatial effect.

Figures 1 and 2 display the estimated effects of floor space and year of construction. Both figures show a monotone but obviously nonlinear dependency of the net rents on the metrical covariates.

The estimated spatial effect, shown in Figure 3, reflects quite well what we know from expert assessments, with an increase of average rents in popular subquarters along the isar river and near to parks.

A space-time study on forest damage

The data used in our second application have been collected in yearly visual forest damage inventories carried out in a forest district around Rothenbuch 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 84 stands of trees as observation points.

Figure 4: Temporal development of the frequency of damaged trees.

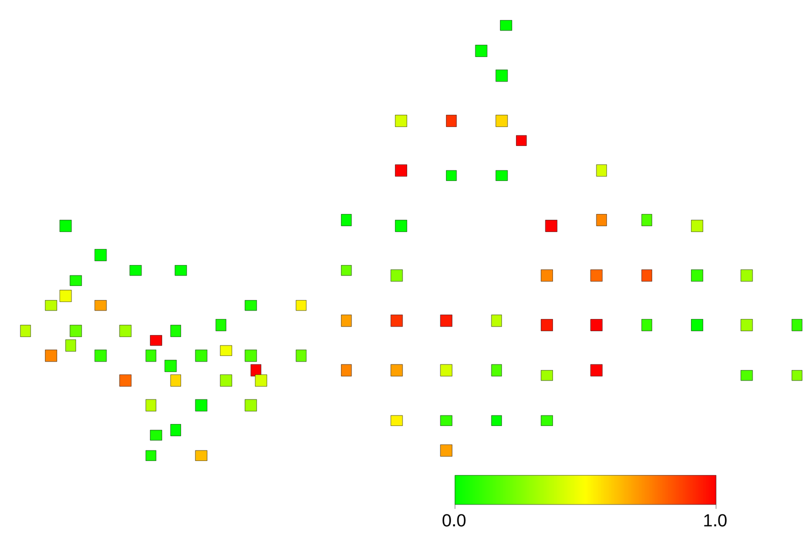

Figure 5: Percentage of damage, averaged over the entire observation period.

As response we consider the damage state of beeches. For each tree, the defoliation degree serves as an indicator of its damage state, resulting in a binary response y_{it} with $y_{it} = 1$ (damage of tree i in year t) and $y_{it} = 0$ (no damage), $i = 1, \dots, 84$, $t = 1983, \dots, 2001$. Figure 4 and 5 show the temporal development of the frequency of damaged trees and the spatial distribution of trees together with the percentage of damage, averaged over the entire observation period.

Our analysis is based on the following covariates:

A age of the tree in years,
C canopy density at the stand, measured in steps of 10%,
t calendar time in years,
S site of the tree.

We model the probability $\mathbf{P}(y_{it} = 1)$ (tree i is damaged in year t) through the following logit model

$$\log \frac{\mathbf{P}(y_{it} = 1)}{\mathbf{P}(y_{it} = 0)} = \gamma_0 + f_1(t) + f_2(A_{it}) + f_3(C_{it}) + f_4(S_i),$$

where the functions f_1, f_2 and f_3 are modelled through cubic P-splines with second order random walk penalty, and the spatial component f_4 follows a Markov random field prior. A pair of trees is considered as neighbours if their distance is less than 1.2 km.

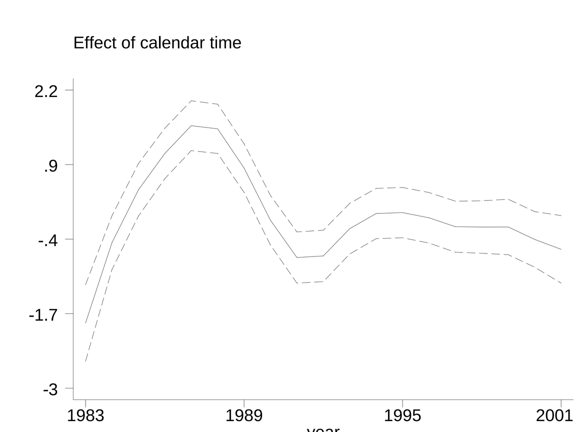

Figure 6: Estimated effect of calendar time with pointwise 95% credible intervals.

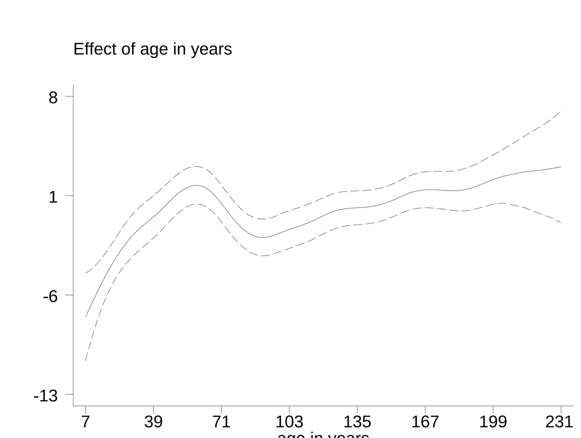

Figure 7: Estimated effect of the age of the trees with pointwise 95% credible intervals.

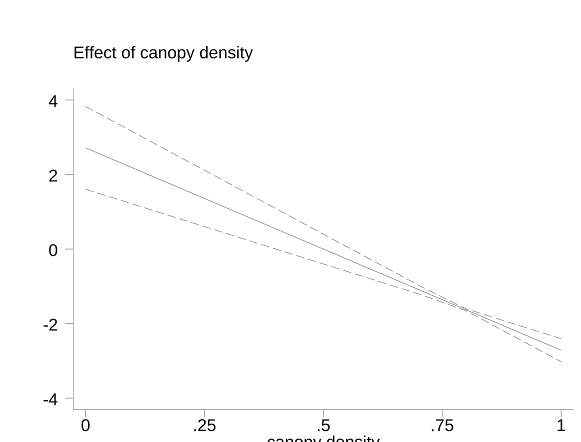

Figure 8: Estimated effect of the canopy density with pointwise 95% credible intervals.

Figures 6 to 8 display the estimated functions $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$. The estimates $\hat{f}_1$ and $\hat{f}_2$ are clearly nonlinear, while the effect of the canopy density seems to be linearly decreasing, leading to a possible model simplification. The shape of the confidence intervals in Figure 8 is caused by the centering of f_3 and the very small variance τ_3^2 that is estimated for f_3 ($\hat{\tau}_3^2 \approx 4 \cdot 10^{-7}$). This leads to an almost linear function with a predetermined value of $f_3(\bar{C})$, where $\bar{C}$ is the mean canopy density.

The estimated effect $\hat{f}_1$ of calendar time reflects the descriptive trend from Figure 4 with a peak in the mid-eighties, recovering thereafter and staying on a more or less constant level in the nineties. Astonishingly, the nonlinear effect of age is not monotone, with a first peak around 65 years.

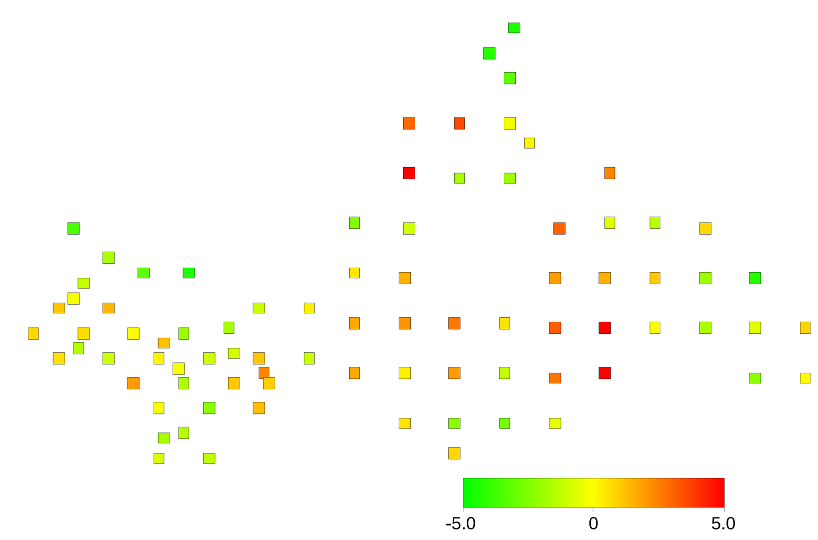

Figure 9: Estimated spatial effect

The estimated spatial effect is given in Figure 9. It reflects the raw spatial effect shown in Figure 5 but also illuminates a spatial pattern with increased damage state around the village of Rothenbuch.

	$\hat{y}_{it}$			$\hat{y}_{it}$	
y_{it}	0	1	y_{it}	0	1
0	900	71	0	846	125
1	113	465	1	207	371

Table 1: Classification table with and without spatial effect

In Table 1 we compare the classification of trees for all years based on the spatio-temporal logit model and, alternatively, on a model without the spatial component f_4. The classification table of the spatio-temporal model shows a clear improvement, confirming that inclusion of the spatial information is substantial. This is also reflected in the misclassification rates 11.9% (with spatial component) and 21.4% (without spatial component).

Software

The presented mixed model approach is implemented in *GGAMM*, a software package that includes several Splus-/R-functions. The program allows the estimation of nonlinear effects of metrical covariates (modelled as P-splines), structured effects of spatial covariates (modelled as Markov random fields) and uncorrelated random effects (random intercepts and random slopes) for Gaussian, gamma, Poisson and Binomial distributed response. *GGAMM* is available from

www.stat.uni-muenchen.de/ kneib

Fully Bayesian analyses have been carried out with *BayesX*, a software for Bayesian inference based on MCMC techniques. *BayesX* is available from

www.stat.uni-muenchen.de/ lang

References

Besag, J., York, J. & 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. & Lang, S. (2003). Generalized structured additive regression based on Bayesian P-splines, *Discussion Paper 321, SFB 386, University of Munich* . Available from www.stat.uni-muenchen.de/∼lang.

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

Fahrmeir, L., Kneib, T. & Lang, S. (2003). Penalized additive regression for space-time data: A Bayesian perspective, *Statistica Sinica (under revision)* . Available from www.stat.uni-muenchen.de/∼kneib.

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

Fahrmeir, L. & Lang, S. (2001b). Bayesian semiparametric regression analysis of multicategorical time-space data, *Annals of the Institute of Statistical Mathematics* **53**: 11–30.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman and Hall, London.

Kauermann, G. (2002). A note on bandwidth selection for penalised spline smoothing, *Technical Report, Department of Statistics, University of Glasgow* . Available from www.wiwi.uni-bielefeld.de/∼kauermann.

Kneib, T. (2003). Bayes-Inferenz in generalisierten geoadditiven gemischten Modellen. Diploma thesis, University of Munich. Available from www.stat.uni-muenchen.de/∼kneib.

Lang, S. & Brezger, A. (2003). Bayesian P-Splines, *Journal of Computational and Graphical Statistics (to appear)* . Preprint available from www.stat.uni-muenchen.de/∼lang.