Ann. Inst. Statist. Math.
Vol. 53, No. 1, 11–30 (2001)

BAYESIAN SEMIPARAMETRIC REGRESSION ANALYSIS OF MULTICATEGORICAL TIME-SPACE DATA

LUDWIG FAHRMEIR AND STEFAN LANG

Department of Statistics, University of Munich, Ludwigstrasse 33, D-80539 Munich, Germany

(Received April 17, 2000; revised August 15, 2000)

Abstract. We present a unified semiparametric Bayesian approach based on Markov random field priors for analyzing the dependence of multicategorical response variables on time, space and further covariates. The general model extends dynamic, or state space, models for categorical time series and longitudinal data by including spatial effects as well as nonlinear effects of metrical covariates in flexible semiparametric form. Trend and seasonal components, different types of covariates and spatial effects are all treated within the same general framework by assigning appropriate priors with different forms and degrees of smoothness. Inference is fully Bayesian and uses MCMC techniques for posterior analysis. The approach in this paper is based on latent semiparametric utility models and is particularly useful for probit models. The methods are illustrated by applications to unemployment data and a forest damage survey.

Key words and phrases: Categorical time-space data, forest damage, latent utility models, Markov random fields, MCMC, probit models, semiparametric Bayesian inference, unemployment.

1. Introduction

Multicategorical longitudinal data consist of observations (Y_{it}, x_{it}), $i = 1, \ldots, n$, $t = 1, \ldots, T$, for a population of n units observed across time, where the response variable Y is observed in ordered or unordered categories $r \in \{1, \ldots, k\}$. Covariates may be time-constant or time-varying.

In this paper, we consider multicategorical time-space data, where the spatial location or site s on a spatial array $\{1, \ldots, s, \ldots, S\}$ is given for each unit as an additional information. We also distinguish between metrical covariates $x_t = (x_{t1}, \ldots, x_{tp})'$, whose effects will be modelled and estimated nonparametrically, and a further vector w_t of covariates, whose effects will be modelled parametrically in usual linear form. Multicategorical time-space data on n individuals or units then consists of observations

$$(1.1) \qquad (Y_{it}, x_{it}, w_{it}, s_i), \qquad i = 1, \ldots, n, \quad t = 1, \ldots, T,$$

where $s_i \in \{1, \ldots, S\}$ is the location of individual i.

A typical example are monthly register data from the German Employment Office for the years 1980–1995, where Y_{it} is the employment status (e.g. unemployed, part time job, full time job) of individual i during month t and s_i is the district in Germany where i has its domicile. Data from surveys on forest health are a further example: Damage state Y_{it} of tree i in year t, indicated by the defoliation degree, is measured in ordered categories (severe to none) and s_i is the site of the tree on a lattice map. In both examples, covariates can be categorical or continuous, and possibly time-varying.

In general, time-space data of this kind cannot be analyzed adequately with existing nonparametric or conventional parametric methods. We present a unified semiparametric Bayesian framework for jointly modelling and analyzing effects of time, space and different types of covariates on categorical responses. Trend or seasonal components, spatial effects, metrical covariates with nonlinear effects and usual covariates with fixed effects are all treated within the same general framework by assigning appropriate priors with different forms and degrees of smoothness. This broad class of models contains state space models for categorical time series considered in previous work as a special case, see e.g. Fahrmeir and Tutz (2001). Inference is fully Bayesian and uses recent MCMC techniques. The approach is based on latent variables, where the observable categorical responses are generated through threshold or utility mechanisms. For latent Gaussian variables this leads to multicategorical probit models, see Albert and Chib (1993) for the simpler case of linear predictors, and Yau *et al.* (2000) for nonparametric regression using basis functions. For MCMC inference, Gaussian latent variables are considered as unknown additional "parameters" and are generated jointly with the other parameters in a Gibbs sampling scheme. Efficient methods for sampling from high dimensional Gaussian Markov random fields are incorporated as a major building block.

Section 2 describes our Bayesian semiparametric regression models for categorical responses, observed across time and space, and depending on unknown functions and parameters. MCMC algorithms are presented in Section 3. In Section 4, the methods are applied to reemployment chances based on categorical time-space data on (un-) employment status and to data from a forest health inventory.

2. Semiparametric Bayesian models for multicategorical time-space data

2.1 *Multicategorical response models*

Categorical response models may be motivated from the consideration of latent variables. This is not only useful for construction of models, but also for Bayesian inference, treating latent variables as additional unknown "parameters".

For the case of a *nominal response* Y with unordered categories $1, \ldots, k$, let U_r be a latent variable or utility associated with the r-th category. Assume that U_r is given by

$$U_r = \eta_r + \epsilon_r, \tag{2.1}$$

where η_r is a linear or semiparametric predictor depending on covariates and parameters, and $\epsilon_1, \ldots, \epsilon_k$ are random errors. Following the principle of random utility the observable response Y is determined by

$$Y = r \Leftrightarrow U_r = \max_{j=1,\ldots,k} U_j, \tag{2.2}$$

i.e., in choice situations the alternative is chosen which has maximal utility. Since only differences of utilities are identifiable we may set $U_k = 0$ for the reference category k. In the unemployment example, we choose $Y = 3 =$ "unemployed" as the reference category. Then U_1 and U_2 are latent variables associated with the categories full and part time job. Depending on the distributional assumptions for the error variables ϵ_r, equation (2.2) yields different models. If the ϵ's are i.i.d. normal, one gets the independent probit model. The more general multivariate probit model allows correlated noise variables. Assuming i.i.d. error variables following the extreme value distribution yields the multinomial logit model. For the case of an *ordered response* Y, cumulative models based on a threshold approach are most widely used. It is postulated that Y is a categorized version of a latent variable

$$U = \eta + \epsilon, \tag{2.3}$$

obtained through the threshold mechanism

$$Y = r \Leftrightarrow \theta_{r-1} < U \leq \theta_r, \qquad r = 1, \dots, k, \tag{2.4}$$

with thresholds $-\infty = \theta_0 < \theta_1 < \cdots < \theta_k = \infty$. In the forest damage example, damage state Y is considered as a three-categorical version of the latent continuous variable damage U. If the error variable ϵ has distribution function F, it follows that Y obeys a cumulative model

$$P(Y \leq r) = F(\theta_r - \eta). \tag{2.5}$$

With a linear predictor $\eta = w'\beta$ one gets parametric cumulative models. For identifiability reasons, the linear combination does not contain an intercept term β_0. Otherwise one of the thresholds, for example θ_1, had to be set to zero. The most popular choices for F in (2.5) are the logistic and the standard normal distribution function leading to cumulative logit or probit models.

2.2 *Observation model*

For multicategorical time-space data (1.1), we generally assume more flexible semiparametric predictors. For nominal responses Y_{it}, the general form of a semiparametric additive predictor associated with category r is

$$\eta_{itr} = f^r_{time}(t) + f^r_{spat}(s_i) + \sum_{j=1}^{p} f^r_j(x_{itj}) + w'_{it}\beta_r. \tag{2.6}$$

Here f^r_{time} and f^r_{spat} represent possibly nonlinear effects of time and space, $f^r_1, \dots, f^r_p$ are unknown smooth functions of the metrical covariates $x_1, \dots, x_p$, and $w'_{it}\beta_r$ corresponds to the usual parametric part of the predictor. Note that the latter may also contain a (category specific) intercept. Depending on the analysed dataset, the effect of time f^r_{time} may contain only a nonlinear time trend or may be split up into a trend and a seasonal component, i.e.

$$f^r_{time}(t) = f^r_{trend}(t) + f^r_{season}(t).$$

In analogy we may split up the spatial effect f^r_{spat} into a spatially correlated (structured) and an uncorrelated (unstructured) effect

$$f^r_{spat}(s) = f^r_{str}(s) + f^r_{unstr}(s).$$

A rational is that a spatial effect is usually a surrogate of many unobserved influential factors, some of them may obey a strong spatial structure and others may be present only locally. By estimating a structured and an unstructured effect we aim at separating between the two kinds of influential factors. As a side effect we are able to assess to some extent the amount of spatial dependency in the data by observing which one of the two effects exceeds. If the unstructured effect exceeds, the spatial dependency is smaller and vice versa. With the same arguments we could also divide up the time trend $f^r_{trend}(t)$

into a correlated and an uncorrelated component. Such models are common in spatial epidemiology, see Besag *et al.* (1991) and Knorr-Held and Besag (1998).

A further extension of (2.6) are varying coefficient models, where nonlinear terms $f_j^r(x_{itj})$ are generalized to $f_j^r(x_{itj})z_{itj}$, where z_j may be a component of x or w or a further covariate. Covariate x_j is called the effect modifier of z_j because the effect of z_j varies smoothly over the range of x_j. Of course, time t and even the spatial covariate s are also possible effect modifiers.

For ordered responses Y_{it} following a cumulative model (2.5), we assume semiparametric predictors

$$\eta_{it} = f_{time}(t) + f_{spat}(s_i) + \sum_{j=1}^{p} f_j(x_{itj}) + w_{it}'\beta \tag{2.7}$$

where the terms have the same interpretation as in (2.6), omitting the category-specific index r. Note that the term $w'\beta$ for fixed effects must not contain an intercept to make thresholds identifiable.

2.3 *Prior model*

For Bayesian inference, unknown functions f_{time}, f_{spat}, $f_1, \ldots, f_p$, thresholds $\theta = (\theta_1, \ldots, \theta_{k-1})'$ and all other parameters are considered as random variables. Categorical response models are to be understood conditional upon these random variables and have to be supplemented by appropriate prior distributions. For the "fixed effect" parameters θ and β we assume diffuse priors $p(\theta) \propto$ const, $p(\beta) \propto$ const. Priors for a time trend f_{trend} of time t and functions $f_1, \ldots, f_p$ of metrical covariates are specified by local smoothness priors common in state space modelling of structural time series. We illustrate the approach for the effect of a specific metrical covariate x. Let $x_{(1)} < \cdots < x_{(l)} < \cdots < x_{(m)}$, denote the m *different, ordered* observed values of the metrical covariate x. Define $f(l) := f(x_{(l)})$ and let $f = (f(1), \ldots, f(l), \ldots, f(m))'$ denote the corresponding vector of function evaluations. For equally-spaced values $x_{(1)}, \ldots, x_{(m)}$ we usually assign first or second order random walk models

$$f(l) = f(l-1) + \xi(l) \quad \text{or} \quad f(l) = 2f(l-1) - f(l-2) + \xi(l) \tag{2.8}$$

with Gaussian errors $\xi(l) \sim N(0; \tau^2)$ and diffuse priors $f(1) \propto$ const, or $f(1)$ and $f(2) \propto$ const, for initial values, respectively. Both specifications act as smoothness priors that penalize too rough functions f. The variance τ^2 controls the degree of smoothness of f. Of course, local linear trend models or higher order autoregressive priors are also possible. An example is a time varying seasonal component f_{season} of time t. A flexible seasonal component with period *per* can be defined by

$$f_{season}(t) = - \sum_{j=1}^{per-1} f_{season}(t-j) + \xi(t) \tag{2.9}$$

and once again diffuse priors for initial values and with errors $\xi(t) \sim N(0, \tau^2_{season})$.

For non-equally spaced values $x_{(1)}, \ldots, x_{(m)}$, priors have to be modified to account for nonequal distances $\delta_l = x_{(l)} - x_{(l-1)}$. Random walks of first order are now specified by

$$f(l) = f(l-1) + \xi(l), \qquad \xi(l) \sim N(0, \delta_l \tau^2),$$

and random walks of second order by

$$f(l) = \left(1 + \frac{\delta_l}{\delta_{l-1}}\right) f(l-1) - \frac{\delta_l}{\delta_{l-1}} f(l-2) + \xi(l), \qquad \xi(l) \sim N(0, \gamma_l \tau^2)$$

with appropriate weights γ_l. Based on Fahrmeir and Lang (2001) we choose $\gamma_l = \delta_l(1 + \frac{\delta_l}{\delta_{l-1}})$.

All these priors can equivalently be rewritten in form of a global smoothness prior

$$f \mid \tau^2 \propto \exp\left(-\frac{1}{2\tau^2} f' K f\right), \tag{2.10}$$

with appropriate penalty matrix K. For example,

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

in the simple case of a first order random walk and equidistant observations.

Let us now turn our attention to a spatial covariate s, where the values of s represent the location or site in geographical regions. For the spatially correlated effect $f_{str}(s)$, $s = 1, \ldots, S$, we choose Markov random field priors common in spatial statistics (Besag *et al.* 1991). These priors reflect spatial neighbourhood relationships. For geographical data one usually assumes that two sites or regions s_i and s_j are neighbours if they share a common boundary. Then a spatial extension of random walk models leads to the conditional, spatially autoregressive specification

$$f_{str}(s) \mid f_{str}(u), \ u \neq s \sim N\left(\sum_{u \in \partial_s} \frac{1}{N_s} f_{str}(u), \frac{\tau^2_{str}}{N_s}\right), \tag{2.11}$$

where N_s is the number of adjacent regions, and $u \in \partial_s$ denotes that region u is a neighbour of region s. Thus the (conditional) mean of $f_{str}(s)$ is an average of function evaluations $f_{str}(u)$ of neighbouring regions. Again the variance τ^2_{str} controls the degree of smoothness. This prior will be used in our first application on durations of unemployment. In some applications, as in our second example on forest damage data, a more general prior specification seems to be more appropriate. In this application we assume that two sites s_i and s_j are neighbours if they are within a certain distance, d say. In addition, we assume that the conditional mean of $f_{str}(s)$ is now a weighted average of function evaluations $f_{str}(u)$ of neighbouring sites rather than an unweighted average as in (2.11). The weights are chosen to be proportional to the distance of neighbouring sites to site s. In terms of weights w_{su} a general spatial prior can be defined as

$$f_{str}(s) \mid f_{str}(u), \ u \neq s \sim N\left(\sum_{u \in \partial_s} \frac{w_{su}}{w_{s+}} f_{str}(u), \frac{\tau^2_{str}}{w_{s+}}\right), \tag{2.12}$$

where + denotes summation over the missing subscript. In the forest damage application the weights are set to $w_{su} = c\exp(-d(s,u))$ where d is the Euclidian distance between sites s and u. The normalizing constant c is chosen in such a way that the total sum of weights is equal to the total number of neighbours, which is in analogy to (2.11). Note that the spatial prior (2.11) is a special case of (2.12) with weights $w_{su} = 1$.

As for autoregressive priors, (2.12) can be written in the form (2.10), where the elements of the penalty matrix K are given by $k_{ss} = w_{s+}$ and $k_{su} = -w_{su}$ if u and s are neighbours and otherwise zero.

As mentioned before, we may split up the effect of a spatial covariate into a structured (spatially correlated) and an unstructured (uncorrelated) effect. For an unstructured effect f_{unstr} a common assumption is that the parameters $f_{unstr}(s)$ are i.i.d. Gaussian

$$f_{unstr}(s) \mid \tau^2_{unstr} \sim N(0, \tau^2_{unstr}). \tag{2.13}$$

Note that we are not restricted to an unstructured effect only for the spatial covariate s. An unstructured effect for time t, or with respect to any other grouping variable, is also possible (and already supported in our implementation).

2.4 *Hyperpriors*

For a fully Bayesian analysis, variance or smoothness parameters τ_j^2, $j = 1, \dots, p$, $j = trend, season$, $j = str, unstr$, are also considered as unknown and estimated simultaneously together with corresponding unknown functions f_j. Therefore, hyperpriors are assigned to them in a second stage of the hierarchy by highly dispersed inverse gamma distributions $p(\tau_j^2) \sim IG(a_j, b_j)$ with known hyperparameters a_j and b_j. It turns out that the simultaneous estimation of smooth functions and smoothing parameters is a great advantage of our Bayesian modelling approach. In a frequentist approach smoothing parameters are usually chosen by minimizing some goodness of fit criteria (e.g. AIC) with respect to the smoothing parameters, or via cross validation, see, for example Fahrmeir and Tutz ((2001), Chapter 5). However, if the model contains many nonparametric effects as in the applications of this paper, a multidimensional grid search is required which becomes totally impractical for higher dimensions. This problem gets even worse in multicategorical response models.

The Bayesian model is completed by the following conditional independence assumptions:

(i) For given covariates and parameters observations Y_{it} are conditionally independent.
(ii) Priors for function evaluations, fixed effects parameters and for variances are all mutually independent.

3. Posterior analysis via MCMC

In the following f denotes the vector of all function evaluations including trend and seasonal components of time t and structured and unstructured spatial effects, τ is the vector of all variances, $\gamma = \beta$ for nominal and $\gamma = (\beta, \theta)$ for ordinal models. For a nominal logit model or a cumulative logit model, the contribution of Y_{it} to the likelihood $p(Y \mid f, \gamma)$ of the data given the parameters can be easily calculated. Bayesian inference can then be based on the posterior $p(f, \tau, \gamma \mid Y) \propto p(Y \mid f, \gamma) p(f \mid \tau) p(\tau) p(\gamma)$. MCMC simulation is based on drawings from full conditionals of single parameters or blocks of parameters, given the rest and the data. Single moves update each parameter separately. Convergence and mixing is considerably improved by block moves for the vectors $f_j = (\dots, f_j(l), \dots)'$ of function evaluations, using Metropolis-Hastings (MH) steps with conditional prior proposals as suggested by Knorr-Held (1999). Details of the updating schemes are described in Fahrmeir and Lang (2001) for univariate responses, and can easily be extended to multicategorical responses, see Fahrmeir and Lang (2000).

Useful alternative sampling schemes can be developed on the basis of the latent variable mechanisms (2.2) and (2.4), augmenting the observables Y_{it} by corresponding latent variables $U_{itr} = \eta_{itr} + \epsilon_{itr}$ or $U_{it} = \eta_{it} + \epsilon_{it}$, respectively, with semiparametric predictors as in (2.6) or (2.7). Assuming Gaussian errors, we obtain multicategorical probit models with latent semiparametric Gaussian models. Posterior analysis is now based on

$$p(f, \gamma, \tau, U \mid Y) \propto p(Y \mid U)p(U \mid f, \gamma)p(f \mid \tau)p(\tau)p(\gamma),$$

with $p(Y \mid U) = \prod_{i,t} p(Y_{it} \mid U_{it})$, where $U_{it} = (U_{it1}, \ldots, U_{itk})'$ for nominal responses. The conditional likelihood $p(Y_{it} \mid U_{it})$ is determined by the mechanisms (2.2) or (2.4). For a nominal response, we have

$$p(Y_{it} \mid U_{it}) = \sum_{r=1}^{k} I(\max(U_{it1}, \ldots, U_{itk}) = U_{itr})I(Y_{it} = r). \tag{3.1}$$

For a cumulative model, we get

$$p(Y_{it} \mid U_{it}) = \sum_{r=1}^{k} I(\theta_{r-1} < U_{it} \leq \theta_r)I(Y_{it} = r), \tag{3.2}$$

due to the fact that $p(Y_{it} \mid U_{it})$ is one if U_{it} obeys the constraint imposed by the observed value of Y_{it}. Compared to the direct sampling scheme above, additional drawings from full conditionals for the latent variables U_{it} are necessary. As an advantage, full conditionals for functions and fixed effects parameters become Gaussian, allowing computationally efficient Gibbs sampling. The full conditionals for U_{it} are:

$$p(U_{it} \mid f, \gamma, Y_{it}) \propto p(Y_{it} \mid U_{it})p(U_{it} \mid f, \gamma). \tag{3.3}$$

Since latent variables U_{it} have (conditional) Gaussian distributions with means η_{it} and unit variances, their full conditionals are truncated normals, with truncation points determined by the restrictions (3.1) and (3.2).

To derive full conditionals for functions f_j and fixed effects parameters β it is convenient to rewrite the predictors (2.6) and (2.7) in matrix notation. For example for (2.7) we obtain

$$\eta = X_{time} f_{time} + X_{spat} f_{spat} + \sum_{j=1}^{p} X_j f_j + W\beta. \tag{3.4}$$

Here the X_j are 0/1 matrices where the number of columns is equal to the number of parameters of the respective effect. If for observation i, t the value of covariate x_j (or time t or site s) is l, then the element in the i, t-th row and the l-th column is one, zero otherwise. Now standard calculations show that the full conditional for a function f_j is Gaussian with covariance matrix and mean given by

$$\Sigma_j = P_j^{-1} = \left(X_j' X_j + \frac{1}{\tau_j^2} K_j \right)^{-1}, \quad \mu_j = \Sigma_j X_j'(U - \tilde{\eta}). \tag{3.5}$$

Here $\tilde{\eta}$ is the part of the predictor associated with all remaining effects in the model. Since $X_j'X_j$ is diagonal and the penalty matrix K_j is a bandmatrix (e.g. with bandwidth two for a second order random walk) it follows that the posterior precision P_j is also a bandmatrix with the same bandwith. Following Rue (2000), drawing random numbers from the full conditionals for f_j is as follows:

(i) Compute the Cholesky decomposition $P_j = LL'$.

(ii) Solve $L'f_j = z$, where z is a vector of independent standard Gaussians. It follows that $f_j \sim N(0, \Sigma_j)$.

(iii) Compute the mean μ_j by solving $P_j\mu_j = X_j'(U - \tilde{\eta})$. This is achieved by first solving by forward substitution $L\nu = X_j'(U - \tilde{\eta})$ followed by backward substitution $L'\mu_j = \nu$.

(iv) Set $f_j = f_j + \mu_j$, then $f_j \sim N(\mu_j, \Sigma_j)$.

All algorithms involved take advantage of the bandmatrix structure of the posterior precision P_j.

Finally, the full conditionals for fixed effects parameters β with diffuse priors are Gaussian with mean and covariance matrix given by

$$\mu_{beta} = (W'W)^{-1}W'(U - \tilde{\eta}), \qquad \Sigma_{beta} = (W'W)^{-1}. \tag{3.6}$$

We can now summarize the resulting sampling schemes. For a cumulative probit model, a Gibbs sampling scheme is defined by the following steps.

Sampling scheme 1.

(i) The latent variables U_{it}, $i = 1, \dots, n$, $t = 1, \dots, T$ are sampled as follows. If $Y_{it} = r$, then U_{it} is generated from $N(\eta_{it}, 1)$, with mean η_{it} as in (2.7), evaluated at current values of f_j and β, subject to the constraint $\theta_{r-1} < U_{it} \le \theta_r$.

(ii) Following Albert and Chib (1993), the full conditional for threshold θ_r, $r = 1, \dots, k-1$ is uniform on the interval

$$[\max\{U_{it} : Y_{it} = r\}, \min\{U_{it} : Y_{it} = r+1\}].$$

Posterior samples from these uniform distribution may exhibit bad mixing. A reason is that intervals can become quite small and, as a consequence, the chain moves slowly. In such a case, other parametrizations as suggested for example in Chen and Dey (2000) are a possible alternative. For $k = 3$ such a reparametrization becomes particularly convenient, see Section 4 or our application to forest damage in Section 5.

(iii) Function evaluations f_j are generated from Gaussian full conditionals $p(f_j \mid U, \cdot)$ with covariance matrix and mean in (3.5), using the algorithms for bandmatrices described above.

(iv) Samples for variances τ_j^2 are generated from inverse Gamma posteriors with updated parameters $a_j' = a_j + \frac{rank(K_j)}{2}$ and $b_j' = b_j + \frac{1}{2}f_j'K_jf_j$.

(v) Samples for fixed effects β are drawn from Gaussian full conditionals with mean and covariance matrix in (3.6).

For nominal response, we choose k as the reference category. Since only differences of utilities can be identified (see Section 3.2), we set the latent variable U_{itk} to zero.

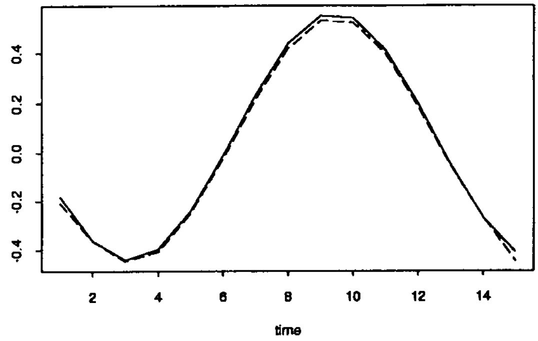

Fig. 1. True time trend (solid line) and average posterior mean estimates (dashed line).

Sampling scheme 2.

(i) Setting $U_{itk} \equiv 0$, latent variables U_{itr}, $r = 1, \dots, k-1$, are generated as follows for each observation Y_{it}, $i = 1, \dots, n$, $t = 1, \dots, T$. If $Y_{it} = r$, $r \neq k$, then U_{itr} is generated first from a normal distribution with mean η_{itr} and variance 1, subject to the constraints $U_{itr} > U_{itl}$, $l \neq k$, and $U_{itr} > 0$ ($\equiv U_{itk}$). Next we generate U_{itl} for $l \neq r$ from a normal distribution with mean η_{itl} and variance 1, subject to the constraint that U_{itl} is less than the U_{itr} generated just before. If $Y_{it} = k$ (the reference category), then we generate $U_{itl}, l = 1, \dots, k-1$, from a normal distribution with mean η_{itl} and variance 1, subject to the constraint $U_{itl} < 0$.

(ii) Posterior samples for functions f_j^r, $r = 1, \dots, k-1$ and all other parameters are generated as in the steps (iii)–(v) of sampling scheme 1.

4. Simulation study

We have analyzed several simulated data sets to see how well our proposed method works. In particular, we have investigated if the procedure provides a reasonal decomposition of temporal and spatial components and how sensitive results depend on the choices of inverse Gamma hyperpriors for variances of Gaussian priors (2.10) and (2.12). In the following we present results for a cumulative probit model defined as in (2.3)–(2.5) with semiparametric predictor

$$\eta_{it} = f_{trend}(t) + f_{str}(s_i) + f_{unstr}(s_i). \tag{4.1}$$

The true trend function was chosen sinusoidal. Evaluations at 15 equidistant time points are displayed in Fig. 1 (solid line). The true structured spatial component $f_{str}(s)$ for the map of West Germany displayed in Fig. 3 a) was constructed from an underlying plane. Observations $f_{str}(s_i)$ for the districts s_i, $i = 1, \dots, 309$ in this map were taken as the values of the plane at the centroids of the districts. The values $f_{unstr}(s_i)$ of the unstructured spatial component were generated by i.i.d. drawings from a $N(0, 0.32)$ distribution. Keeping the "true" predictor fixed, realizations of latent time series $\{U_{it}^l, t = 1, \dots, 15\}$ were generated for each district $i = 1, \dots, 309$ from the latent model $U_{it}^l = \eta_{it} + \epsilon_{it}^l$ for $l = 1, \dots, 100$ simulation runs. Corresponding three-categorical ordered responses y_{it}^l were then obtained via the threshold mechanism (2.4) with $\theta_1 = -0.5$ and $\theta_2 = 0.5$. To investigate the sensitivity of the estimates from the respective choice of hyperparameters

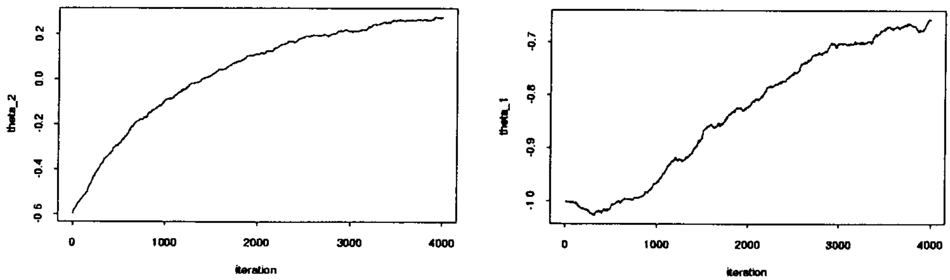

Fig. 2. Sampling paths of thresholds for the first 4000 iterations.

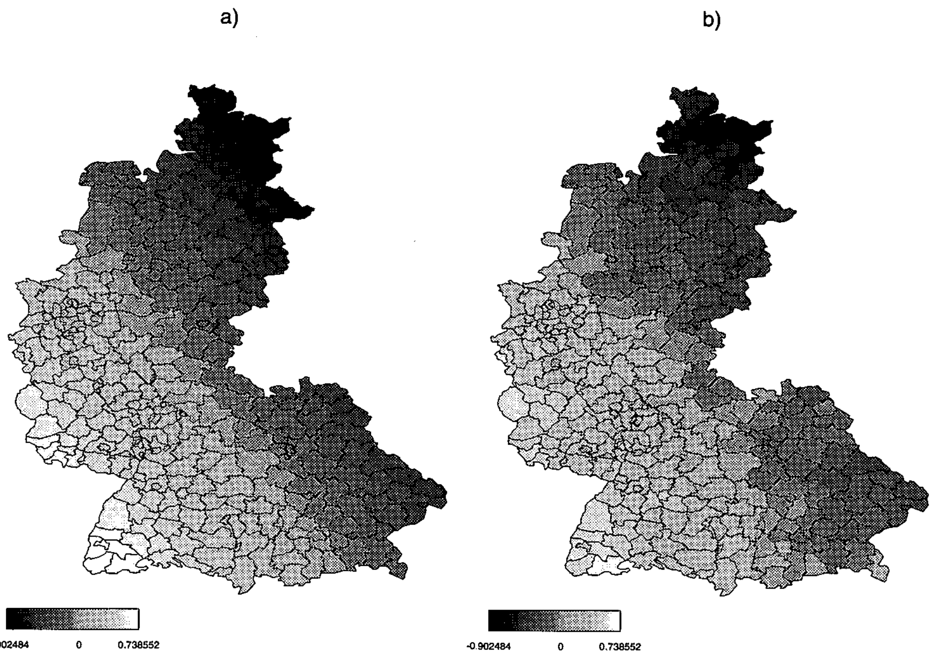

Fig. 3. Map of the "true" structured spatial effect (left) and average posterior mean estimates (right).

for the variances estimation was carried out with three alternative settings for a and b, i.e. $a = 1$, $b = 0.005$, $a = 0.25$, $b = 0.005$ and $a = 0.1$, $b = 0.005$.

We first applied the ordered probit model in standard parametrization. However, in step (ii) of sampling scheme 2 mixing of posterior samples for thresholds θ_1 and θ_2 was not satisfactory. Fig. 2 displays the sampling paths of the sampled threshold parameters for the first 4000 iterations. Obviously, the generated Markov chains show extremly bad mixing properties. Seemingly, the chains are still far away from the equilibrium around the "true" thresholds $\theta_1 = -0.5$ and $\theta_2 = 0.5$. Following Chen and Dey (2000), we therefore reparametrized the model. First, inclusion of a constant β_0 in (4.1) allows to set $\theta_1 = 0$. Secondly, because parameters in the predictor of the latent Gaussian model are only identifiable up to a multiplicative factor, we assume that errors ϵ_{it} are

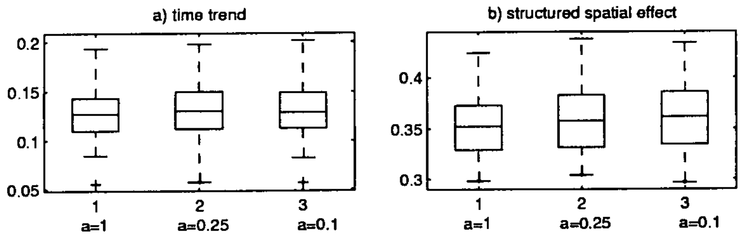

Fig. 4. Relative MSE's for the time trend and the structured spatial component.

$N(0, \sigma^2)$ 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 following results are given in the original parametrization which is obtained by simply dividing sampled parameters in MCMC simulation through the current value of the standard deviation σ.

Figure 1 (dashed line) shows the average $\bar{f}_{time}(t) = \sum_{l=1}^{100} \hat{f}^l_{time}(t)$ of the 100 posterior mean estimates for the values $f(t)$, $t = 1, \ldots, 15$ of the true trend function. It is seen that the true function is estimated almost unbiased. The figure shows only estimation results for the choice of $a = 1$ and $b = 0.005$ for the hyperparameters of variances because results for the three alternative choices of hyperparameters are more or less indistinguishable. Similarly, Fig. 3 b) shows the average $\bar{f}_{str}(s)$ of the 100 posterior mean estimates of the true map $f_{str}(s)$ ($a = 1, b = 0.005$). Although, there is more bias than for the trend component, the map is well recovered. It is however not too surprising that we get more bias for the spatial effect since the number of observations per district is far less compared to the number of observations per time point. Another important aspect is that the smooth component is considerably masked by the superimposed unstructured spatial effect. We found out that the bias becomes smaller if the variability of the unstructured effect is less than in the present simulation experiment, particularly if the unstructured effect is omitted. Similarly to the time trend, the differences between the estimated maps for different choices of a and b are more or less neglegible. We therefore printed only results for $a = 1$ and $b = 0.005$. The robustness with respect to the choice of hyperparameters can also be seen in Fig. 4 which displays boxplots of the relative MSE's

$$\sum_{t=1}^{15} \frac{(\hat{f}_{trend} - f_{trend}(t))^2}{f^2_{trend}(t)}, \quad \sum_{i=1}^{309} \frac{(\hat{f}_{str} - f_{str}(i))^2}{f^2_{str}(i)}$$

for the three choices of hyperparameters. Finally we note that the estimates for the variance component of the unstructured spatial effect are slightly biased. The average of the 100 posterior means of τ^2_{unstr} is 0.26 whereas the true variance is 0.32.

A more detailed simulation study including the case of unordered response can be found in a supplement paper (Lang and Fahrmeir (2001)).

5. Applications

We consider two applications. In a first application on unemployment durations we analyse unemployment data from the German Federal Employment Office. This is a huge dataset with approximately 280000 observations showing the practicability of our methods even for very large datasets. In a second application, we analyse longitudinal data on forest health collected in the forest district of Rothenbuch in northern Bavaria. All computations have been carried out with *BayesX*, a software package for Bayesian inference. The program is available under `http://www.stat.uni-muenchen.de/~lang/`, see also Lang and Brezger (2000).

5.1 *Reemployment chances*

In our first application we analyse monthly unemployment data from the German Federal Employment Office for the years 1980–1995. Our analysis is restricted to data from former West Germany (excluding Berlin) and to women. For each individual the data provides information about the employment status in month t, the district where the individual lives and a number of personal characteristics. Since we are interested in analyzing reemployment chances, distinguishing between full and part time jobs, we define three-categorical response variables Y_{it} as event indicators

$$Y_{it} = \begin{cases} 1, \text{ gets a new full time job in month } t \text{ (calendar time)} \\ 2, \text{ gets a new part time job in month } t \\ 3, i \text{ is unemployed in month } t \text{ (reference category).} \end{cases}$$

Our analysis is based on the following covariates:

- D duration time measured in months
- A age (in years) at the beginning of unemployment
- N nationality, dichotomous with categories "German" and "foreigner" (= reference category)
- U_d unemployment compensation (in month d of duration time), dichotomous with categories "unemployment benefit" (=reference category) and "unemployment assistance"
- P_t number of previous unemployment periods (in month t of calendar time): 1,2,3 and more, 0 (reference category)
- E education, trichotomous with categories "no vocational training" "vocational training" (reference category) and university
- S district in which the unemployed have their domicil

All categorical covariates are coded in effect coding.

We model the probabilities $P(Y_{it} = r \mid \eta_{itr})$, $r = 1, 2$, by an independent probit model with predictors

$$\eta_{itr} = f^r_{trend}(t) + f^r_{season}(t) + f^r_{str}(s_i) + f^r_{unstr}(s_i) + f^r_1(d_{it}) + f^r_2(a_{it}) + w'_{it}\beta_r,$$
$$r = 1, 2,$$

where f^r_{trend} and f^r_{season} are trend and seasonal component of calendar time t, f^r_{str} and f^r_{unstr} are structured and unstructured spatial effects of the district, f^r_1 is the effect of duration D in current unemployment status and f^r_2 is the effect of age A. The priors for f^r_{trend}, f^r_1 and f^r_2 are second order random walk models (2.8) with diffuse priors for initial values. For f^r_{str} and f^r_{unstr} we assign the Markov random field prior (2.11) and the prior (2.13), respectively. For the seasonal component we choose the flexible seasonal prior (2.9). Priors for fixed effects parameters β_r are diffuse. An analysis with similar predictors using a multinomial logit model and with direct drawings from full conditionals can be found in Fahrmeir and Lang (2000).

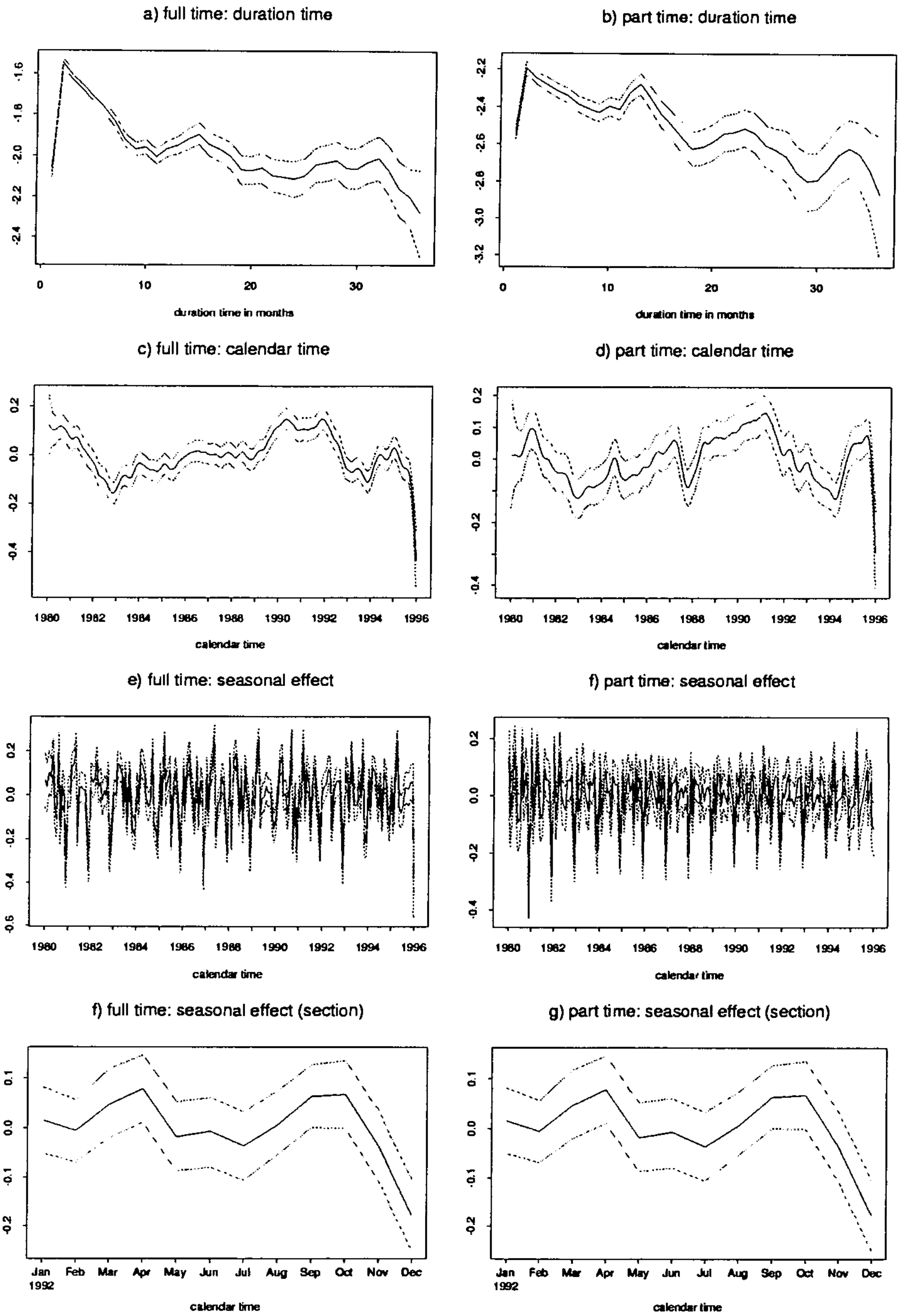

Fig. 5. Estimated nonparametric effects of duration and calendar time. Shown is the posterior mean within 80 % credible regions.

Figure 5 displays estimated effects of duration time, calendar time trend and seasonal component for getting full time jobs (left column) and part time jobs (right column). Duration time effects have the typical pattern also observed in other investigations, with a peak after 2–3 months and sloping downward then. Calendar time trends

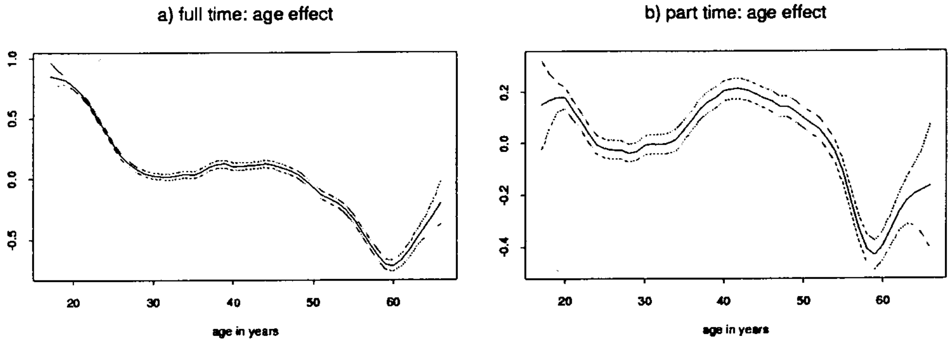

Fig. 6. Estimated nonparametric effects of age. Shown is the posterior mean within 80 % credible regions.

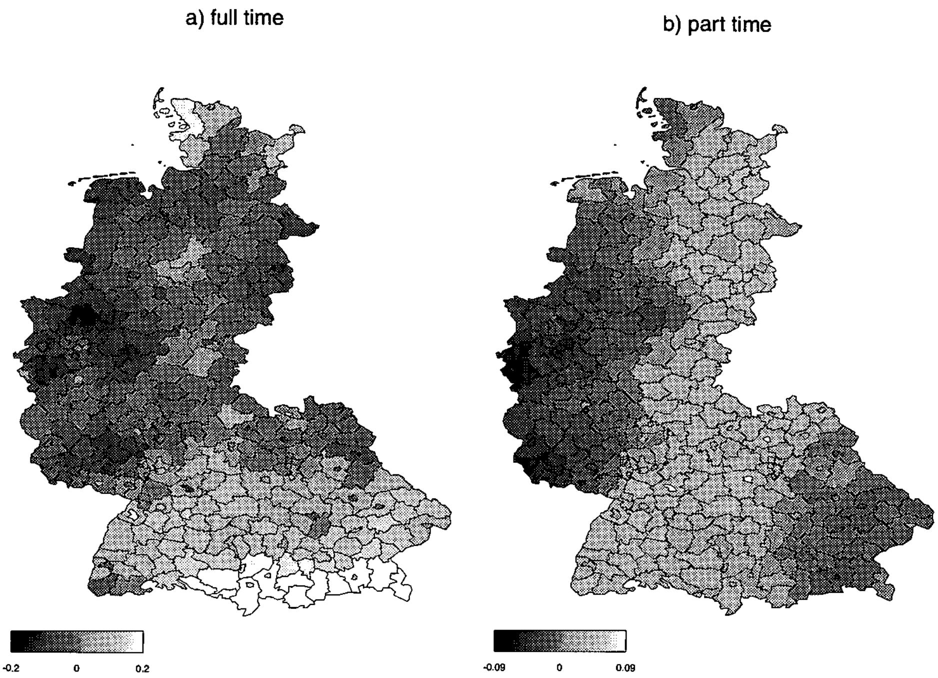

Fig. 7. Posterior mean of the structured spatial effects.

for full and part time jobs show a similar general pattern: declining until the year 1982, then slowly increasing until 1990 (one year after the German reunion), declining distinctly again thereafter, with an intermediate recovery. This corresponds to the observed economic trend of the labor market in Germany during this period. Initial credible regions are larger here for the following reasons: We use diffuse priors $f^r_{trend}(1) \propto const$, $f_{trend}(2) \propto const$, and there is less information provided by the data, since there are comparably few unemployment periods ending already at the beginning of the observation period. Estimated seasonal effects are more or less stable over this period although

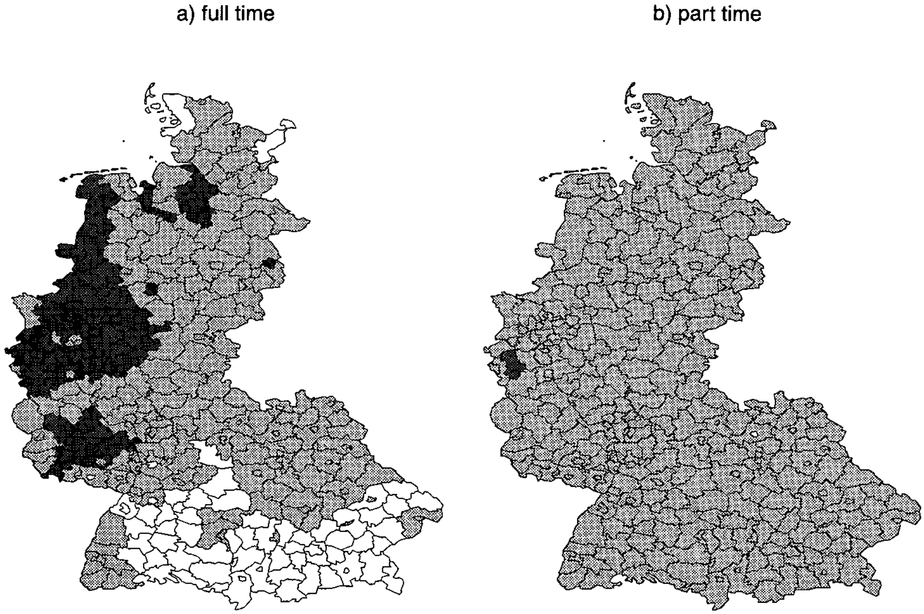

Fig. 8. Posterior probabilities of the structured spatial effects.

varying in size. To gain more insight Fig. 5 f) and g) display a section of the estimated effects for the year 1992 with the typical peaks in spring and autumn, and a global minimum in December. The effect of age can be found in Fig. 6. For the age effect there are local minima for women about 30, which may be a "family" effect. The dramatic decline of unemployment probabilities of people older than 50 years is particularly striking. Again, credible regions near the boundaries are wider because there are much less individuals below 20 and above 60 years in the sample. The increase after 60 years is caused by a very small number of individuals in the sample (reemployed about 65 and 66). We observed this effect for age with other methods as well, showing that this is not a particular problem of our approach. Note that the age effect is much stronger for women seeking full time jobs (Fig. 6 a)) compared to women seeking part time jobs (Fig. 6 b)). Structured regional effects are shown in Figs. 7 and 8. Figure 7 displays the estimated posterior mean and Fig. 8 shows "probability" maps. The levels correspond to "significantly negative" (black colored), "nonsignificant" (grey colored), i.e. zero is within the 80% confidence interval around the estimate, and "significantly positive" (white colored). In order to interpret the structured effects, unstructured effects must be taken into consideration as well. Therefore Table 1 gives a summary of the estimated posterior means of the unstructured spatial effect for the different regions. We observe that the structured effect for getting full time jobs is stronger than for getting part time jobs. Even more important, the unstructured effect for part time jobs cleary exceeds the structured effect which is in constrast to the estimated effects for full time jobs. Although the estimated posterior mean of the structured effect for getting part time jobs in Fig. 7 b) shows some spatial variation, Fig. 8 b) cleary indicates that there is no "significant" variation in terms of posterior probabilities. In the contrary, the structured effect for getting full time jobs displays "significant" variation with improved chances in

Table 1. Summary of the posterior means of the unstructured spatial effect.

	full time	part time
std. dev.	0.0246	0.103
minimum	−0.073	−0.274
10% quantile	−0.0287	−0.120
90% quantile	0.0291	0.134
maximum	0.128	0.501

Table 2. Estimates of constant parameters for the application on duration of unemployment.

	full time			part time		
Variable	mean	10% quant	90% quant	mean	10% quant	90% quant
German	0.05	0.04	0.06	0.11	0.09	0.13
foreign	−0.05	−0.06	−0.04	−0.11	−0.13	−0.09
unemployment ass.	−0.07	−0.08	−0.06	−0.08	−0.10	−0.07
unemployment ben.	0.07	0.06	0.08	0.08	0.07	0.10
no voc. training	−0.03	−0.05	−0.02	−0.10	−0.12	−0.08
voc. training	−0.02	−0.03	−0.01	−0.05	−0.07	−0.035
university	0.05	0.03	0.08	0.15	0.12	0.19
$P = 0$	−0.11	−0.12	−0.10	−0.16	−0.17	−0.14
$P = 1$	−0.04	−0.05	−0.02	−0.05	−0.07	−0.03
$P = 2$	0.00	−0.01	0.02	0.01	−0.02	0.03
$P \geq 3$	0.14	0.13	0.15	0.20	0.18	0.22

the south compared to the middle and the north. The two black spots in the west in Fig. 8 a) mark areas that are known for their structural economical problems during the eighties and nineties.

Estimates of fixed effects for getting full and part time jobs are shown in Table 2. Chances for re-employment are better for Germans and for women with a university degree compared to women with vocational training and no vocational training. Both effects are stronger for women getting part time jobs. The number of previous unemployment periods serves as a surrogate for experience at the labor market: an increase in the number of previous spells increases the probability for shorter unemployment duration. The estimated effect of unemployment assistance is significantly negative. It is positive for unemployment benefits, which seems to contradict the widely-held conjecture about negative side-effects of unemployment benefits. However, it may be that the variable "unemployment benefit" also acts as a surrogate variable for those who have worked, and therefore contributed regularly to the insurance system in the past.

5.2 *Forest health*

In this longitudinal study on the state of trees, we analyse the influence of calendar time, age of trees, canopy density and location of the stand on the defoliation degree of beeches. Data have been collected in yearly forest damage inventories carried out in the forest district of Rothenbuch in northern Bavaria from 1983 to 1997. There are 80 observation points with occurence of beeches spread over an area extending about 15 km

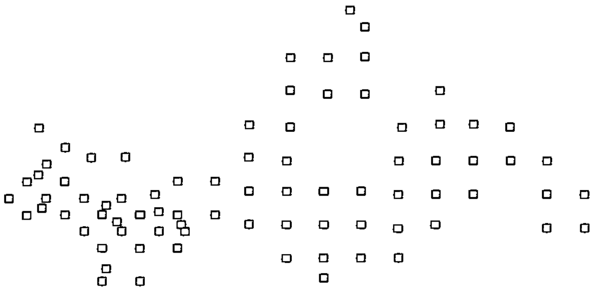

Fig. 9. Map of observation points.

from east to west and 10 km from north to south, see Fig. 9. The degree of defoliation is used as an indicator for the state of a tree. It is measured in three ordered categories, with $Y_{it} = 1$ for "bad" state of tree i in year t, $Y_{it} = 2$ for "medium" and $Y_{it} = 3$ for "good". A detailed data description can be found in Göttlein and Pruscha (1996). Covariates used here are defined as follows:

A age of tree at the beginning of the study in 1983, measured in three effect coded categories a_1 = "below 50 years", a_2 = between 50 and 120 years, and a_3 = above 120 years (reference category);

C Canopy density at the stand measured in percentages 0%, 10%,...,90%, 100%.

The covariate age is time constant by definition, while canopy density is time varying. Based on previous analysis, we use a three-categorical ordered probit model (2.5) based on a latent semiparametric model $U_{it} = \eta_{it} + \epsilon_{it}$ with predictor

$$\eta_{it} = f_{trend}(t) + f_{str}(s_i) + f(c_{it}) + \beta_1 a_{i1} + \beta_2 a_{i2}. \tag{5.1}$$

Here a_{i1} and a_{i2} are the indicators for age categories 1 and 2. The calendar time trend $f_{trend}(t)$ and the effect $f(c)$ of canopy density are modelled by random walks of second order. For the structured spatial effect we assign the Markov random field prior (2.12), with the neighbourhood ∂_s of trees including all trees u with Euclidian distance $d(s,u) \leq 1.2$ km. An unstructured spatial effect is excluded from the predictor for the following two reasons. First, a look at the map of observation points (Fig. 9) reveals some sites with only one neighbour, making the identification of a structured and an unstructured effect difficult if not impossible. The second reason is that for each of the 80 sites only 15 observations on the same tree are available with only minor changes of the response category. In fact, there are only a couple of sites where all three response categories have been observed. Thus, the inclusion of an unstructured effect in our model leads to severe identification problems between the structured and unstructured effect, which can be observed by inspecting sampling paths of parameters.

For interpretation of estimation results note the following: In accordance with our definitions (2.3) to (2.5), higher (lower) values of the predictor (5.1) (or of effects in this predictor) correspond to healthier (worse) state of the trees. Similar to the simulation study we reparametrized the model to obtain better mixing properties for the thresholds (see Section 4), but results are given in the original parametrization. Estimates for β_0 and the effect of age are given in Table 3. As we might have expected younger trees are in healthier state than the older ones. Fig. 10 shows posterior mean estimates for the calendar time trend and for the effect of canopy density. We see that trees recover

Table 3. Estimates of constant parameters in the forest health study.

Variable	mean	10% quant.	90% quant.
a_1	0.55	0.33	0.78
a_2	0.04	−0.14	0.22
a_3	−0.59	−0.78	−0.41

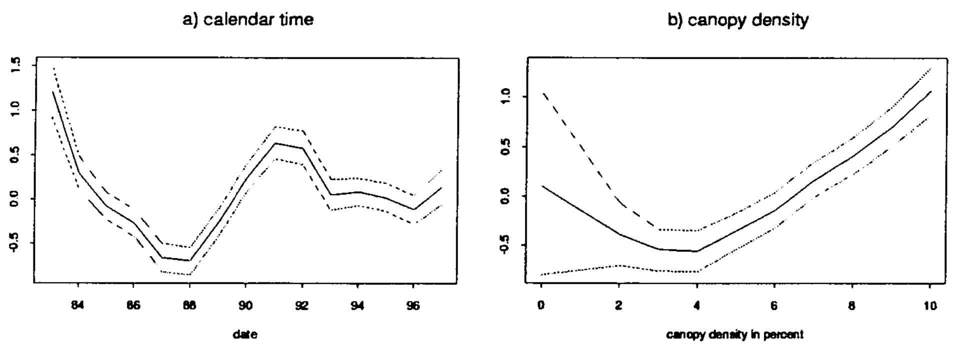

Fig. 10. Estimated time trend and nonlinear effect of canopy density. Shown is the posterior mean within 80 % credible regions.

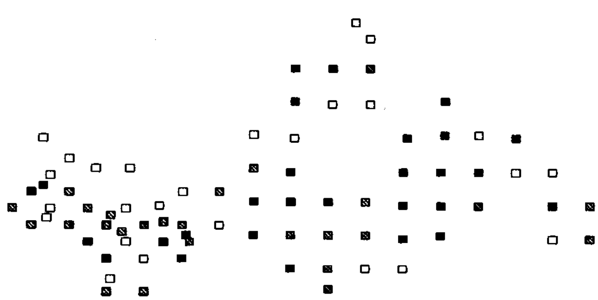

Fig. 11. Posterior probabilities of the spatial effect.

after the bad years around 1986, but after 1994 health status declines to a lower level again. The distinct monotonic increase of the effect of canopy densities $\geq 30\%$ gives evidence that beeches get more shelter from bad environmental influences in stands with high canopy density. Fig. 11 shows the estimated (structured) spatial effect in form of posterior probabilities, where black spots indicate areas with strictly negative credible regions, i.e. areas with more trees in bad state. The black colored sites correspond mostly to areas in the forest district which are located higher above sea level than the other sites. Here the environmental conditions in terms of nutrient quantity and soil quality are worse compared to other areas.

6. Conclusion

The applications demonstrate that the Bayesian methods developed are useful and flexible tools for inference in realistically complex categorical regression models.

A variety of extensions are possible by modifying or generalizing the observation models, predictors and smoothness priors. For example, probit models based on latent utilities can be extended to correlated categorical or mixed continous-categorical responses by considering latent multivariate semiparametric Gaussian models. Predictors can be made more flexible by introducing nonparametric interactions between covariates following suggestions in Clayton (1996) and Knorr-Held (2000). Replacing Gaussian priors by heavy-tail distributions would allow to consider unsmooth regression functions.

Acknowledgements

We thank the editors and two referees for valuable comments that helped to improve a first version. Support from the German National Science Foundation through the Sonderforschungsbereich Statistical Analysis of Discrete Structures is also gratefully acknowledged. Finally we thank Andrea Hennerfeind for computational assistance and Stefan Bender for providing the unemployment data.

References

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *J. Amer. Statist. Assoc.*, **88**, 669–679.

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

Chen, M. H. and Dey, D. K. (2000). Bayesian analysis for correlated ordinal data models, *Generalized Linear Models: A Bayesian Perspective* (eds. D. K. Dey, S. K. Ghosh and B. K. Mallick), Chapter 8, 133–159, Marcel, New York.

Clayton, D. (1996). Generalized linear mixed models, *Markov Chain Monte Carlo in Practice*, (eds. Gilks, W., Richardson S. and Spiegelhalter D), 275–301, Chapman and Hall, London.

Fahrmeir, L. and Lang, S. (2000). Bayesian semiparametric regression analysis of multicategorical time-space data, Proceedings of the International Symposium on Frontiers of Time Series Modeling, Tokyo, 82–91.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors, *Applied Statistics* (to appear).

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd ed., Springer, New York.

Göttlein, A. and Pruscha, H. (1996). Der Einfluss von Bestandskenngrössen, Topographie, Standort und Witterung auf die Entwicklung des Kronenzustandes im Bereich des Forstamtes Rothenbuch, *Forstwissenschaftliches Centralblatt*, **114**, 146–162.

Knorr-Held, L. (1999). Conditional prior proposals in dynamic models, *Scand. J. Statist.*, **26**, 129–144.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk, *Statistics in Medicine*, **19**, 2555–2567.

Knorr-Held, L. and Besag, J. (1998). Modelling risk from a desease in time and space, *Statistics and Medicine*, **17**, 2045–2060.

Lang, S. and Brezger, A. (2000). BayesX—Software for Bayesian inference based on Markov chain Monte Carlo simulation techniques, Discussion Paper 187, Sonderforschungsbereich 386, Ludwigs-Maximilians-Universität München.

Lang, S. and Fahrmeir, L. (2001). Bayesian generalized additive mixed models. A simulation study, Discussion paper 230, Sonderforschungsbereich 386, Ludwigs-Maximilians-Universität München.

Rue, H. (2000). Fast sampling of Gaussian Markov random fields with applications, Tech. Report, No. 1/2000, Norwegian University of Science and Technology, Trondheim, Norway.

Yau, P., Kohn, R. and Wood, S. (2000). Bayesian variable selection and model averaging in high dimensional multinomial nonparametric regression, University of New South Wales (Preprint).