Structured count data regression

Ludwig Fahrmeir & Leyre Osuna
Department of Statistics, Ludwig–Maximilian–Universitt Mnchen

Overdispersion in count data regression is often caused by neglection or inappropriate modelling of individual heterogeneity, temporal or spatial correlation, and nonlinear covariate effects. In this paper, we develop and study semiparametric count data models which can deal with these issues by incorporating corresponding components in structured additive form into the predictor. The models are fully Bayesian and inference is carried out by computationally efficient MCMC techniques. In a simulation study, we investigate how well the different components can be identified with the data at hand. The approach is applied to a large data set of claim frequencies from car insurance.

Keywords: Bayesian semiparametric count data regression, negative binomial distribution, Poisson–Gamma distribution, Poisson–Log–Normal distribution, MCMC, spatial models.

1 Introduction

Count data regression has become a very active topic in methodological and applied research, see Cameron and Trivedi (1998) and Winkelmann (2000) for recent surveys. In applications, one is often confronted with one or several of the following issues, preventing use of standard Poisson regression: individual unobserved heterogeneity caused by omitted covariates, temporal or spatial correlation, and possibly nonlinear effects of metrical covariates or time scales. This situation arises for example in our application to car insurance data where we analyze the effects of some categorical covariates, of metrical covariates such as age of the policyholder and age of the car, and of the residence of the policyholder on claim frequencies. While the effects of categorical covariates may be modelled in usual linear parametric form, it is usual very difficult if not impossible to specify nonlinear effects of metrical covariates or of time scales and, in particular, spatial effects a priori through conventional parametric functional forms.

Neglection or inappropriate modelling of these issues will often result in biased estimates and in the problem of overdispersion. Statistical modelling of unobserved heterogeneity and serially correlated data is well developed, see, for example, the books by Cameron and Trivedi (1998) and Winkelmann (1997), as well as Wooldridge (1997), Chib, Greenberg and Winkelmann (1998) and Toscas and Faddy (2003). Chib and Winkelmann (2001) develop a Markov Chain Monte Carlo (MCMC) approach for analyzing correlated count data by introducing latent effects which follow a multivariate Gaussian distribution with full unrestricted covariance matrix. While this approach is useful in the case of several correlated responses for the same subject, it is not feasible for high–dimensional spatial data. Also, in all this previous work the remaining part of the predictor models the effect of covariates in usual linear parametric form.

In this paper, we develop hierarchical Bayesian semiparametrically structured count data regression to deal with these aspects within a joint model and a unified framework for inference. Parametric and nonparametric components for covariate effects, temporal or spatial effects, and unobserved heterogeneity are included in an additive predictor. Given this predictor, observations are assumed to be conditionally independent, with Poisson, negative binomial, Poisson–Log–Normal or Poisson–Gamma distributions. Inference is fully Bayesian and extends previous work by Fahrmeir and Lang (2001), Lang and Brezger (2003) and Brezger and Lang (2003) within the exponential family framework. It is based on MCMC techniques, implemented in BayesX. For model comparison, we routinely use the deviance information criterion (DIC) developed in Spiegelhalter, Best, Carlin and van der Linde (2002).
In a simulation study, we investigate performance and how well the different components, which are theoretically identifiable through different types of priors, can be separated in practice from the data at hand. Finally, we apply our approach to a massive set of claim frequency data from a car insurance company.

The rest of the paper is organized as follows: Structured count data models are introduced in Section 2, while Section 3 outlines inference through MCMC simulation. Section 4 contains the simulation study, and the application is described in Section 5. The conclusions point out some goals of future research.

2 Bayesian structured count data models

2.1 Observation models

Consider regression count data (y_i, z_i), $i = i, ..., n$, where y_i are observations on a counting variable y, such as claim frequency, and z_i are observed values of a covariate vector z. In addition, known offsets r_i, such as the exposure time of individual i within the time period under consideration, may be given. The basic regression model is a loglinear Poisson model, where the observations $y_i \mid z_i$ are (conditionally) independent

$$y_i \mid z_i \sim Po(r_i \lambda_i), \tag{1}$$

with rate $\lambda_i = \exp(\eta_i)$ and linear predictor $\eta_i = z_i' \beta$. Equivalently, we can rewrite (1) as

$$y_i \mid z_i \sim Po(\mu_i), \quad \mu_i = \exp(o_i + \eta_i), \tag{2}$$

with known offset o_i. We extend this basic loglinear Poisson model in two ways: First, we consider negative binomial, Poisson–Gamma and Poisson–Log–Normal distributions, which can account for individuals–specific unobserved heterogeneity. Secondly, we generalize the parametric linear predictor to a semiparametric structured additive predictor.

Given the predictor η_i and a scale parameter $\delta > 0$, the *negative binomial (NB) model* assumes conditionally independent observations

$$y_i | \eta_i, \delta \quad \sim \quad NB(\mu_i, \, \delta), \tag{3}$$

with probability function given by

$$P(y_i | \eta_i, \delta) \quad = \quad \frac{\Gamma(y_i + \delta)}{\Gamma(y_i + 1)\Gamma(\delta)} \left(\frac{\mu_i}{\mu_i + \delta} \right)^{y_i} \left(\frac{\delta}{\mu_i + \delta} \right)^{\delta}$$

for $y_i \in \mathbb{N} \cup \{0\}$. The mean and the variance are

$$\begin{aligned} E[y_i | \eta_i, \delta] &= \mu_i, \\ V[y_i | \eta_i, \delta] &= \mu_i + \frac{\mu_i^2}{\delta}. \end{aligned}$$

In a Bayesian approach, priors have to be assigned to all unknowns. Priors for the components of the predictor η_i are defined in the next subsection. For the scale parameter $\delta > 0$, we assume a Gamma prior

$$\delta \sim G(a, b) \tag{4}$$

with density

$$g(\delta) = \frac{b^a}{\Gamma(a)} \delta^{a-1} \exp(-b\delta), \tag{5}$$

and with mean $E[\delta] = \frac{a}{b}$ and variance $V[\delta] = \frac{a}{b^2}$.

The parameters a and b can be chosen such that the Gamma distribution has a flat prior. As a more data driven alternative, we consider them as hyperparameters and introduce a hyperprior in a further stage of the hierarchy. For reasons discussed in Section 4, we set $a = 1$ and assume a flat Gamma hyperprior

$$b \sim G(1, 0.005) \tag{6}$$

for b.

The NB model admits overdispersion. This is made more explicit in the *Poisson–Gamma (PoGa) model*. To account for individual–specific effects, uncorrelated i.i.d. random effects ν_i, $i = 1, ..., n$, are introduced in multiplicative form by assuming

$$y_i|\eta_i, \nu_i \quad \sim \quad Po(\nu_i\, \mu_i). \tag{7}$$

The probability function, mean and variance, given μ_i and ν_i, are

$$\begin{aligned} P(y_i|\eta_i, \nu_i) &= \frac{\exp(-\nu_i\mu_i)\,(\nu_i\mu_i)^{y_i}}{y_i!} \quad \text{for} \quad y_i \in I\!N \cup \{0\}, \\ E[y_i|\eta_i, \nu_i] &= V[y_i|\eta_i, \nu_i] = \nu_i\, \mu_i. \end{aligned}$$

The PoGa model is obtained, if the ν_i follow a Gamma distribution with mean 1 and scale parameter $\delta > 0$:

$$\begin{aligned} \nu_i|\delta &\sim G(\delta, \delta) \\ g(\nu_i|\delta) &= \frac{\delta^\delta}{\Gamma(\delta)} \nu_i^{\delta-1} \exp(-\delta\, \nu_i) \\ E[\nu_i|\delta] &= 1 \\ V[\nu_i|\delta] &= \frac{1}{\delta}. \end{aligned} \tag{8}$$

Here, δ can be considered as a hyperparameter, one level beneath in the model hierarchy, and also has a Gamma hyperprior with flat density function as described in (5) in the NB–model. The assumptions made in (6) for the hyperparameter b are also valid here.

Integrating out the ν_i parameters from $P(y_i \mid \eta_i, \nu_i)$, that means computing the marginal

$$P(y_i \mid \eta_i, \delta) = E_{\nu_i}[P(y_i \mid \eta_i, \nu_i)] = \int_0^\infty P(y_i \mid \eta_i, \nu_i) g(\nu_i \mid \delta) d\nu_i, \tag{9}$$

leads to the NB model. As a consequence, marginal means $E[y_i \mid \eta_i, \delta]$ and variances $V[y_i \mid \eta_i, \delta]$ are the same as in the NB model. The advantage of the PoGa model is that latent variables are

made explicit and can be generated in MCMC sampling techniques.
As alternatives to Gamma random effects ν_i, we also considered *Poisson–Log–Normal (PoLN) models* and *Poisson–Inverse–Gauss (PoIG) models*. In a PoLN model, a log–normal distribution is assumed for the i.i.d. random effects ν_i. This can be rewritten as

$$\mu_i = \exp(o_i + \widetilde{\nu}_i + \eta_i)$$

where $\widetilde{\nu}_i$ are i.i.d. $N(0, \widetilde{\delta})$ Gaussian random effects. For the variance $\widetilde{\delta}$, we make the usual assumption of a weakly informative inverse Gamma prior

$$\widetilde{\delta} \sim IG(a, b),$$

with $a = 1$, $b = 0.005$ as a standard choice.

We also experimented with PoIG models, where the random effects ν_i follow an inverse Gaussian distribution, which has heavier tails than a Gamma or log–normal prior. However, performance in MCMC posterior inference was less reliable than for the other models, so that we will not consider it in this paper.
In our second generalization, we extend the linear predictor η_i to a much more flexible additive predictor. As in the application to car insurance data we split up the covariate vector z into a vector $x = (x_1, ..., x_p)'$ including all metrical covariates and time scales (for longitudinal data) as well as group indicators with many values such as car type, a vector w of categorial covariates, and a spatial covariate s, which denotes the district or postal code where individual i lives. The basic structured additive predictor η_i has the form

$$\eta_i = \sum_{j=1}^{p} f_j(x_{ij}) + f_{spat}(s_i) + w_i'\gamma, \tag{10}$$

$i = 1, \ldots, n$. The unknown functions $f_j(x_j)$ are the nonlinear effects of a metrical covariate x_j or $f_j(x_{ij})$ are the (random) effects of a group indicator like car type with 31 groups. The function $f_{spat}(s)$ represents the effect of district $s \in \{1, \ldots, S\}$, with S= 327 districts in west Germany in our application. We further split up this spatial effect into the sum

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

of structured (spatially correlated) and unstructured (uncorrelated) effects. A rationale for this decomposition is that a spatial effect is usually a surrogate of many underlying unobserved influential factors. Some of them may be present only locally, while others are correlated with neighboring effects. By estimating a structured and an unstructured effect we aim at taking into account both kinds of influential factors, and we may be able to assess to some extent the amount of spatial dependency in the data by observing which one of the two effects exceeds. The last term in (10) is the usual linear part of the predictor, with fixed effects. To ensure identifiability, an intercept is always included into w_i, and the unknown functions are centered about zero.

The basic additive predictor (10) can be extended further to include varying coefficient terms $w_i f(x_i)$, where w is a categorical covariate and x a metrical covariate, or interactions $f(x_\ell, x_k)$ between two metrical covariates x_ℓ and x_k, say.

2.2 Priors for the predictor

For Bayesian inference all functions and parameters in the predictor are regarded as random variables with suitable priors. To formulate priors in compact and unified notation, we express the predictor vector $\eta = (\eta_i)$ in matrix notation by

$$\eta = \sum_{j=1}^{p} f_j + f_{str} + f_{unstr} + W\gamma, \tag{11}$$

where $f_1, \ldots, f_p, f_{str}, f_{unstr}$ are the vectors of corresponding function values and $W = (w_i)$ is the design matrix for fixed effects. It turns out that each function vector can always be expressed as the product of a design matrix X and a (high-dimensional) parameter vector β. Using $f = X\beta$ as a generic notation for functions, (10) becomes

$$\eta = \ldots + X\beta + \ldots + W\gamma.$$

For fixed effects γ, we generally choose a diffuse prior $p(\gamma) \propto const$, but a (weakly) informative normal prior is also possible. Constructions of the design matrix X and priors for β depend on the type of the function and on the degree of smoothness. For metrical covariates random penalized regression (P-)splines, P-splines and smoothing splines are suitable choices, structured spatial effects are modelled through Markov random field priors, and unstructured random effects through i.i.d. normal random effects. In any case, priors for the vectors β have the same general Gaussian form

$$p(\beta|\tau^2) \propto \exp\left(-\frac{\beta' K \beta}{2\tau^2}\right). \tag{12}$$

The penalty matrix K penalizes roughness of the function. Its structure depends on the type of covariate and on smoothness of the function. The variance τ^2 corresponds to the inverse of a smoothing parameter in a frequentist setting and controls the trade–off between data fit and smoothness. We outline this below, but refer to Fahrmeir and Lang (2001) and Lang and Brezger (2003) for details.

2.2.1 P-splines

For metrical covariates, P-splines introduced by Eilers and Marx (1996) in a frequentist setting, will be our standard choice. It is assumed that an unknown smooth function f of a covariate x can be approximated by a polynomial spline of degree l defined on a set of equally spaced knots $x_{min} = \zeta_0 < \zeta_1 < \ldots < \zeta_{r-1} < \zeta_r = x_{max}$ within the domain of x. It is well known that such a spline can be written in terms of a linear combination of $m = r + l$ B-spline basis functions B_j, i.e.

$$f(x) = \sum_{j=1}^{m} \beta_j B_j(x).$$

Here $\beta = (\beta_1, \ldots, \beta_m)$ corresponds to the vector of unknown regression coefficients. The $n \times m$ design matrix X now consists of the basis functions evaluated at the observations x_i, i.e. $X(i,j) = B_j(x_i)$. 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 flexible enough to capture the variability of the data. For a large number of knots, estimated curves tend to overfit the data and, as a result, too rough functions are obtained. As a remedy to these problems Eilers and Marx (1996) suggest a moderately large number of equally spaced knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on differences of adjacent B-Spline coefficients to guarantee sufficient smoothness of the fitted curves.

In a Bayesian approach, as considered here, the regression coefficients β have to be supplemented with appropriate prior distributions. The stochastic analogue of difference penalties are first and second order random walks for the coefficients $\beta_1, \ldots, \beta_m$ defined by

$$\beta_j = \beta_{j-1} + u_j \quad or \quad \beta_j = 2\beta_{j-1} - \beta_{j-2} + u_j$$

with i.i.d. noise $u_j \sim N(0, \tau^2)$ and diffuse priors $p(\beta_1) \propto const$ or $p(\beta_1) \propto const$, $p(\beta_2) \propto const$ for initial values. First order random walks penalize abrupt jumps between successive parameters, and second order differences penalize deviations from a linear trend. The variance τ^2 controls the amount of penalization. This prior for $\beta = (\beta_1, \dots, \beta_m)$ can be expressed in the form (12). For a first order random walk the penalty matrix K is defined by

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

2.2.2 Spatial covariates

For a spatial covariate s, the values of s represent the locations or districts in connected geographical regions. For simplicity we assume $s \in \{1, \dots, S\}$, i.e. the regions are labelled by the numbers $1, \dots, S$. A common way to deal with spatial covariates is to assume that neighbouring sites are more alike than two arbitrary sites. Thus for a valid prior definition a set of neighbours for each site s must be defined. For geographical data, one usually assumes that two sites s and j are neighbours if they share a common boundary.

The simplest spatial smoothness prior for the function values $f_{str}(s) = \beta_s$ is

$$\beta_s | \beta_j \, j \neq s, \tau^2 \sim N\left(\sum_{j \in \partial_s} \frac{1}{N_s} \beta_j, \frac{\tau^2}{N_s} \right), \tag{13}$$

where N_s is the number of adjacent sites and $j \in \partial_s$ denotes that site j is a neighbour of site s. Thus the (conditional) mean of β_s is an unweighted average of function evaluations of neighbouring sites. Such a prior is called a Gaussian intrinsic autoregression, see Besag, York and Mollie (1991) and Besag and Kooperberg (1995). The vector $\beta = (\beta_1, ..., \beta_s, ..., \beta_S)'$ of spatial effects has a joint–distribution of the form (12) with the elements of K defined by $k_{ss} = w_{s+}$, $k_{sj} = -w_{sj}$ for $j \in \delta_s$, where $w_{sj} = 1$ and $+$ denotes summation over the missing subscript, and 0 else, .

A more general prior including (13) as a special case is given by

$$\beta_s | \beta_j \, j \neq s, \tau^2 \sim N\left(\sum_{j \in \partial_s} \frac{w_{sj}}{w_{s+}} \beta_j, \frac{\tau^2}{w_{s+}} \right), \tag{14}$$

where w_{sj} are known weights and $+$ denotes summation over the missing subscript.

If we express the effect f_{str} of the spatial covariate s as the product of a design matrix X and the vector of unknown parameters β, then X is a $0/1$ incidence matrix.

2.2.3 Unordered group indicators

Suppose $x = g \in \{1, \ldots, G\}$ is now a group variable indicating the group a particular observation i belongs to. We treat this type of covariate by introducing random effects $f(g) = \beta_g$. We assume that the β_g's are i.i.d. Gaussian, i.e.

$$\beta_g \sim N(0, \tau^2), \quad g = 1, \ldots, G. \tag{15}$$

Formally, this prior can be once again brought into the general form (12). The design matrix is again a $n \times G \quad 0/1$ incidence matrix, and $K = I$.

Apparently, there is only a slight difference to the smoothness priors described above for metrical covariates. In fact, instead of specifying for example P-splines for a function f, the i.i.d random effects prior (15) may also be specified. The main difference between the two specifications is the amount of smoothness allowed for a function f. With the i.i.d. random effects specification (15), successive parameters are allowed to vary more or less *unrestricted*, whereas random walk priors guarantee that successive parameters vary smoothly over the range of x. If s is a spatial covariate it can be even useful to incorporate both, a spatially correlated smooth effect $f_{str}(s)$ as well as a spatially uncorrelated effect $f_{unstr}(s)$ as in (15).

To distinguish group indicators from individual specific random effects, we have to assume $G < n$. Otherwise, the prior (15) is the same as for $\log(\nu_i)$ in the PoLN normal and quite close the prior for individual random effects on a logscale for the PoGa model.

3 Posterior inference

Bayesian inference is based on the posterior distribution of the usually very high–dimensional vector of all parameters. In the following, we split up this vector into the subvector ξ containing all parameters defining the predictor η, and remaining parameters specifying the count data distributions for given η. MCMC inference is carried out by repeatedly drawing from full conditionals of (blocks of) parameters given the remaining parameters and the data. Because drawings from full conditionals for components of ξ are identical or similar to sampling schemes in Fahrmeir and Lang (2001) and Brezger and Lang (2003), we focus on full conditionals for the remaining parameters.

For the NB model, the posterior is defined by

$$p(\xi, \delta, b \mid y) \propto P(y \mid \xi, \delta) g(\delta \mid b) g(b) p(\xi),$$

where $P(y \mid \xi, \delta)$ is the likelihood of the NB model, $g(\delta \mid b)$ ist the prior of the scale parameters δ given by (5), $g(b)$ is the hyperprior (6) for b and $p(\xi)$ is defined by the prior assumptions in subsection 2.2. Then the full conditional

$$\begin{aligned} p(\delta|\ldots) &\propto P(y|\eta,\delta)\, g(\delta|b) \\ &\propto \prod_{i=1}^{n} \left\{ \frac{\Gamma(y_i+\delta)}{\Gamma(y_i+1)\Gamma(\delta)} \left(\frac{\mu_i}{\mu_i+\delta}\right)^{y_i} \left(\frac{\delta}{\mu_i+\delta}\right)^{\delta} \right\} \frac{b^a}{\Gamma(a)} \delta^{a-1} \exp(-b\delta) \\ &\propto \prod_{i=1}^{n} \left\{ \frac{\Gamma(y_i+\delta)}{(\mu_i+\delta)^{\delta+y_i}} \right\} \Gamma(\delta)^n\, \delta^{\,n\delta+a-1} \exp(-b\delta) \end{aligned}$$

This expression has no analytical closed form, so that we implement a MH algorithm with a random walk proposal. Let δ^* denote the proposed value for δ in an iteration step. We choose a Gamma proposal

$$\delta^*|\delta \quad \sim \quad G\left(\frac{\delta^2}{p_\delta}, \frac{\delta}{p_\delta}\right) \tag{16}$$

with

$$\begin{aligned} E[\delta^*|\delta] &= \delta \\ V[\delta^*|\delta] &= p_\delta. \end{aligned}$$

The parameter p_δ is a tuning parameter, that allows us to control the acceptance probability for the MH algorithm. It is adapted in the burn in period to achieve acceptance probabilities for δ between 0.4 and 0.6.

For the hyperparameter b the full conditional is calculated as follows:

$$\begin{aligned} p(b|\ldots) &\propto g(\delta|b)\, g(b) \\ &= \frac{b^a}{\Gamma(a)} \delta^{a-1} \exp(-b\delta) \frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} b^{\theta_1-1} \exp(-\theta_2 b) \\ &\propto b^{\theta_1+a-1} \exp(-(\delta+\theta_2)b) \\ &\propto G(\theta_1+a, \delta+\theta_2), \end{aligned} \tag{17}$$

where our standard choice is $a = 1,\ \theta_1 = 1,\ \theta_2 = 0.005$. Therefore b can be updated in a Gibbs step.

For the PoGa model, the posterior is defined by

$$p(\xi, \nu, \delta, b \mid y) \propto P(y \mid \xi, \nu) g(\nu \mid \delta) g(\delta \mid b) g(b) p(\xi),$$

where ν is the vector of all individual–specific effects $\nu_i,\ i = 1, ..., n$, with likelihood and priors defined in Section 2. The full conditional for $\nu_i,\ i = 1, ..., n$, is a Gamma distribution:

$$\begin{aligned} p(\nu_i| \ldots) &\propto P(y_i|\eta_i, \nu_i)\, g(\nu_i|\delta) \\ &= \frac{\exp(-\nu_i\, \mu_i)\, (\nu_i\, \mu_i)^{y_i}}{y!} \frac{\delta^\delta}{\Gamma(\delta)} \nu_i^{\delta-1} \exp(-\nu_i\, \delta) \\ &\propto \nu_i^{y_i+\delta-1} \exp\left(-\nu_i\, (\mu_i + \delta)\right) \\ &\propto G(y_i + \delta, \mu_i + \delta). \end{aligned} \tag{18}$$

Therefore, updates are obtained from direct Gibbs steps. For δ the full conditional is proportional to

$$\begin{aligned} p(\delta| \ldots) &\propto g(\nu|\delta) g(\delta|b) \\ &= \prod_{i=1}^{n} \left\{ \frac{\delta^\delta}{\Gamma(\delta)} \nu_i^{\delta-1} \exp(-\nu_i\, \delta) \right\} \frac{b^a}{\Gamma(a)} \delta^{a-1} \exp(-\delta\, b) \\ &\propto \frac{\delta^{\,n\delta+a-1}}{\Gamma(\delta)^n} \left(\prod_{i=1}^{n} \nu_i \right)^{\delta-1} \exp\left(-\delta \left(b + \sum_{i=1}^{n} \nu_i \right) \right). \end{aligned}$$

The MH algorithm implemented for its update uses the same proposal as for the NB model. Sampling from full conditionals for the variance $\widetilde{\delta}$ of the random effects in the PoLN model as well as for the components of the parameter vector ξ specifying the priors in the predictor are performed along the lines detailed in Fahrmeir and Lang (2001) and Brezger and Lang (2003), and implemented in BayesX.

4 Simulation study

The aim of this study is to explore the performance of the proposed methodology for complex predictor structures similar to those which will be used in the real data application in the next section. In particular, we will investigate how well different components in the predictor can be identified and separated from each other.

We generate data sets from PoGa and PoLN models $\mu_i = \nu_i \exp(\eta_i)$ and $\mu_i = \nu_i \exp(\eta_i) = \exp(\eta_i + \widetilde{\nu}_i)$, respectively, with Gaussian random effects $\widetilde{\nu}_i$ in the PoLN case. The predictor η_i is the same for both models and is defined by

$$\eta_i = o_i + \gamma_0 + \gamma_1 d_i + sin(x_i) + f(g_i) + f_{str}(s_i) + f_{unstr}(s_i), \qquad i = 1, ..., 1920.$$

The offsets o_i are obtained by i.i.d. sampling from a uniform distribution on the interval [3,6]. The values d_i are obtained as i.i.d. samples from a binary random variable $d \sim B(1; 0.5)$. The intercept and slope are $\gamma_0 = -5$ and $\gamma_1 = 0.5$.

The realizations of the metrical covariate x are the 26 knots of an equidistant grid on the interval [-3,3]. The observations x_i, $i = 1, ..., 1920$, are generated by systematically repeating these 26 values until 1920 observations are reached. The nonlinear effect $f(x)$ of x is assumed to be a sine–curve $f(x) = sin(x)$ on the interval [-3,3].

The covariate g represents a group indicator, as for the covariate type class of car in our car insurance application. It has 7 levels $g = 1, ..., 7$, with 7 equidistant effects

$$f(1) = -0.3,\ f(2) = -0.2, ..., f(6) = 0.2,\ f(7) = 0.3.$$

The observations $f(g_i)$, $i = 1, ..., 1920$, are generated as a random sample from these values.

The structured spatial effects $f_{str}(s_i)$ are evaluations of the function $f_{str} : \mathbb{R}^2 \to \mathbb{R}$

$$f_{str}(s = (u, v)) = c_0 sin(5u \cdot v) - c_1$$

at the coordinates $s_i = (u_i, v_i), i = 1, ..., 96$, of the standardized centroids of the 96 districts in Bavaria. The normalizing constants c_0 and c_1 are chosen so that the function values are centered about 0 and have approximate empirical variance 0.25. These structured spatial effects $f_{str}(s)$, $s = 1, ..., 96$, are visualized in Figure 5. For each district i, we assign $f_{str}(s_i)$ to 20 observations.

To generate the unstructured effects $f_{unstr}(s_i)$, we draw $f_{unstr}(s)$, $s = 1, ..., 96$, as an i.i.d. sample from $N(0, \tau^2)$. Then these values are assigned to the same 20 observations per district as in the case of structured spatial effects. To investigate the impact of unstructured effects, we generate data for three values

$$\tau^2 = 0,\ \tau^2 = 0.01,\ \tau^2 = 0.25$$

of the variance τ^2, corresponding to no, small and large unstructured effects. For $\tau^2 = 0.25$, the unstructured effects have the same variability as the structured effects. A particular reason for this choice is that we want to see whether f_{str} and f_{unstr} can be separately identified in the sum

$$f_{spat} = f_{str} + f_{unstr}$$

of total spatial effects.

The random effects ν_i, $i = 1, ..., 1920$, for the PoGa model are obtained as i.i.d. samples from a $G(\delta, \delta)$ distribution with

$$E[\nu_i] = 1, \, Var[\nu_i] = \frac{1}{\delta}.$$

As for f_{unstr}, we generate data for

$$\delta = 0.5, \; \delta = 1, \; \delta = 2$$

corresponding to large, medium and small individual–specific effects. Random effects in the PoLN model are obtained as i.i.d. samples

$$\widetilde{\nu}_i = \log(\nu_i) \sim N\left(0; \sigma^2\right), \, with \quad \sigma^2 = \log\left(1 + \frac{1}{\delta}\right).$$

Then the log–normal effects have

$$Var[\nu_i] = \frac{1}{\delta},$$

just as the Gamma random effects. Combining the possible values of the variance τ^2 of the unstructured spatial effects and for the scale parameter δ, we obtain data for 9 different NB, PoGa and PoLN models. For the discussion of simulation results, we denote them by $M(\tau^2, \delta)$. For example, $M(0, 1)$ is a (NB, PoGa or PoLN) model without ($\tau^2 = 0$) unstructured spatial effects and individual random effects with medium ($\delta = 1$) variability, and $M(0.25, 2)$ is a model with high variability ($\tau^2 = 0.25$) of unstructured spatial effects and low ($\delta = 2$) variability of individual random effects. With this simulation design, we can assess the impact of the relative magnitude of spatial and individual random effects on estimation of the various components.

For each model, we generate counts

$$\{y_i^{(r)}, \, i = 1, ..., 1920\}, \; r = 1, ..., R = 100,$$

for simulation runs $r = 1, ..., R$. For each simulation run r, we calculate posterior means, standard deviations, quantiles and the DIC criterion. From $R = 100$ simulation runs, we then obtain overall empirical bias, MSE, boxplots etc. for the estimates of all unknown parameters and functions.

Four important messages arise from this simulation study. First, results for NB and PoGa models, applied to the same data sets are virtually indistinguishable. Therefore, if one is not interested in the latent individual random effects, a NB model may be preferable. Also, computation time and storage requirements may be an issue, depending on the same size.

Second, unknown fixed effects γ_0, γ_1 and the scale parameter δ are estimated very well, regardless of the specific model. This is illustrated for a sample of models in Table 1. This is also true for estimating the nonlinear sine curve $f(x) = sin(x)$, see Figures 1 and 2. A reason for this obviously quite stable identification of fixed effects and the nonlinear effect of the metrical covariate x is that the priors are rather different from the priors for the remaining effects, which supports separation from the latter ones.

Third, the effects $f(g)$ of the group indicator g can still be estimated quite well, but they seem to be more sensitive to the specific model. Figure 3 displays boxplots of mean square errors for the 9 models. It seems that variation of the scale parameters has some impact, while results are comparably insensitive to variations in variability of unstructured spatial effects. Figure 4 shows true effects and (averaged) posterior mean effects for selected models. We can observe a shrinkage effect towards zero, which becomes larger for smaller δ, i.e. larger individual random effects.

Fourth, separation of structured and unstructured spatial effects is generally very unreliable. In particular, unstructured spatial effects are always underestimated, partly to a large extent. Obviously their effects are already captured by structured spatial and by individual effects. This can be particularly well recognized in the "diagonal plots" of Figures 5 to 8, where true and estimated individual specific random effects are plotted against each other. Ideally, the scatter plots should be near to the diagonal; but they are almost horizontal! For models with no ($\tau^2 = 0$) or small ($\tau^2 = 0.01$) unstructured effects, the structured spatial effects are still recovered satisfactorily (Figures 5 and 6). For models $M(0.25, \delta)$, where variability of structured and unstructured effects is the same, most of unstructured spatial variability is captured by overestimating structured spatial effects, see Figure 7, 8 and 9.

However, as Figure 10 shows, it makes always sense to include structured and unstructured effects, because the sum

$$f_{spat} = f_{str} + f_{unstr}$$

has always the lowest MSE. Of course, then only the total spatial effects f_{spat} can be interpreted.

5 Application to car insurance

We apply structured count data regression models to a data set of 171288 individuals claim frequencies of a sample of policyholders with full comprehensive car insurance for one year. Among others, the following covariates were included in the predictor:

k: kilometers driven per year in thousands;

g: car classification, measured by $G = 31$ scores from 10-40;

s district in Germany ("Zulassungsbezirk" resp. "Landkreis"), with $S = 327$ districts.

To make the data source anonymous, some additional covariates used for the analysis are not described in the paper. The first covariate will be considered as metrical, car classification g as a group indicator, and s as a spatial covariate. Among others, the vector w of categorical covariates comprises garage (yes/no). Claim frequencies were analyzed with structured additive NB, PoGa and PoLN models. Here only part of the results for the NB model are shown. The predictor is defined by

$$\eta_i = \ldots + f_1(k_i) + f_2(g_i) + f_{spat}(s_i) + w_i'\gamma. \tag{19}$$

The dots indicate that the model comprises more than the metrical covariate k. The spatial effect $f_{spat}(s)$ is further split up into the sum

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

The vector w contains the categorical covariates and an intercept term. The effect f_1 of the metrical covariate is modelled by cubic P-splines, the effect f_2 of car classification and the unstructured spatial effect f_{unstr} are treated as i.i.d. random effects, and for the structured spatial effect f_{str} a Markov random field prior is used.

The posterior mean estimates of the functions f_1 and f_2, together with 80% and 95% pointwise confidence bands are displayed in Figure 11. The confidence intervals are constructed by computing the lower and upper posterior quantiles corresponding to the respective nominal level, e.g. 10% and 90% quantiles for a nominal level of 80%. Note that the functions are centered about zero.

The effect of kilometers driven per year shows a distinct, almost linear increase until about 20 000 km/year. Thereafter, the increase becomes much smaller. Looking at the confidence bounds, even a constant effect cannot be rejected. A possible explanation is that these frequently used cars are driven by experienced persons and, probably, to a larger extent on a "Autobahn" than others. Because the covariate car classification was considered as a group indicator with a random effects assumption, the estimated function looks considerably rougher than the other function. It shows an increasing trend until about category 33 that is coherent with the intended definition of the groups. The decreasing trend after this category may be due to sparse data in these last categories. Let us now turn to the geographical, district-specific effects which are displayed in Figures 12-14. The top panel of Figure 12 shows the posterior means for the structured effects f_{str} displaying a smooth but very clear regional pattern: There is a clear decline from south to northwest, perhaps with the exception of parts in the southeast. This is confirmed by the 95% and 80% "significance maps" in the bottom panels of Figure 12. White colored regions correspond to strictly negative confidence intervals (i.e. a "significant negative effect") and blue colored regions to strictly positive confidence intervals (i.e. a "significant positive effect"). Districts with confidence intervals containing zero are colored in light blue. The top map in Figure 13 shows the posterior means of the unstructured effects f_{unstr}. We cannot observe any typical pattern, and, moreover, the unstructured, local effects are much smaller than the corresponding structured effects. This is confirmed by the significance maps in the lower part: no district has significant effect neither for a nominal level of 80% nor for a level of 90%. The maps for the sum f_{spat} of structured and unstructured effects in Figure 14 resemble the maps in Figure 12, but are less smooth.

6 Conclusion

Structured count data models allow to analyze the effects of covariates of various types in much more detail and with higher resolution than with traditional parametric GLM approaches. The results permit a close look at characteristic features of specific effects and can be of great help for monitoring premiums for risk classes of interest. In particular, regional geographical effects can be investigated on a district-specific level within a joint model that adjusts for the presence of other covariates, and they can be compared with regional ratings suggested by experts.

Extensions of our approach to other count data models, such as zero–inflated or truncated models, hurdle models etc. are a topic of future research.

Acknowledgement
We thank Stefan Lang for his valuable advice how to implement the models in BayesX . We also gratefully acknowledge financial support from the Sonderforschungsbereich 386 ”Statistical Analysis of Discrete Structures” and from the Graduiertenkolleg ”Angewandte Algorithmische Mathematik”, both funded by the German National Science Foundation DFG.

References

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

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

Brezger, A. and Lang, S. (2003), 'Generalized structured additive regression based on bayesian p–splines', *SFB 386 Discussion Paper 321, Deparment of Statistics, University of Munich* .

Cameron, A. and Trivedi, P. (1998), *Regression Analysis of Count Data*, Cambrigde University Press, New York.

Chib, S., Greenberg, E. and Winkelmann, R. (1998), 'Posterior simulation and bayes factors in panel count data models', *Journal of Econometrics* pp. 33–54.

Chib, S. and Winkelmann, R. (2001), 'Markov chain monte carlo analysis of correlated count data', *Journal of Business and Economic Statistics* pp. 428–435.

Eilers, P. H. and Marx, B. D. (1996), 'Flexible smoothing using b–splines and penalized likelihood (with comments and rejoinder)', *Statistical Science* pp. 89–121.

Fahrmeir, L. and Lang, S. (2001), 'Bayesian inference for generalized additive mixed models based on markov random field priors', *Applied Statistics, (JRSS C)* pp. 201–220.

Lang, S. and Brezger, A. (2003), 'Bayesian p–splines', *Journal of Computational and Graphical Statistics* .

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society, Series B* pp. 1–34.

Toscas, P. J. and Faddy, M. J. (2003), 'Likelihood-based analysis of longitudinal count data using a generalized poisson model', *Statistical Modelling* pp. 99–108.

Winkelmann, R. (1997), *Econometric Analysis of Count Data*, Verlag, Heidelberg.

Winkelmann, R. (2000), 'Seemingly unrelated negative binomial regression', *Oxford Bulletin of Economics and Statistics* pp. 553–560.

Wooldridge, J. M. (1997), 'Multiplicative panel data models without the strict exogeneity assumption', *Econometric Theory* pp. 667–679.

	NB			PoGa			PoLN		
	M(0.01;1)	M(0.25;1)	M(0.25;2)	M(0.01;1)	M(0.25;1)	M(0.25;2)	M(0.01;1)	M(0.25;1)	M(0.25;2)
γ_0	-5.02	-5.008	-5.001	-5.019	-5.009	-5.002	-4.986	-5.003	-4.994
	(0.0543)	(0.0551)	(0.0473)	(0.0543)	(0.0558)	(0.0475)	(0.0457)	(0.0471)	(0.043)
γ_1	0.511	0.518	0.504	0.512	0.515	0.502	0.486	0.498	0.496
	(0.0706)	(0.0716)	(0.0591)	(0.0705)	(0.0719)	(0.0591)	(0.0604)	(0.0623)	(0.0543)
δ	1.028	0.979	2.013	1.029	0.981	2.013	0.668	0.701	0.397
	(0.0764)	(0.0726)	(0.1846)	(0.0771)	(0.0718)	(0.1841)	(0.0496)	(0.0512)	(0.036)

Table 1: Posterior means and standard deviations (in brackets) for selected models

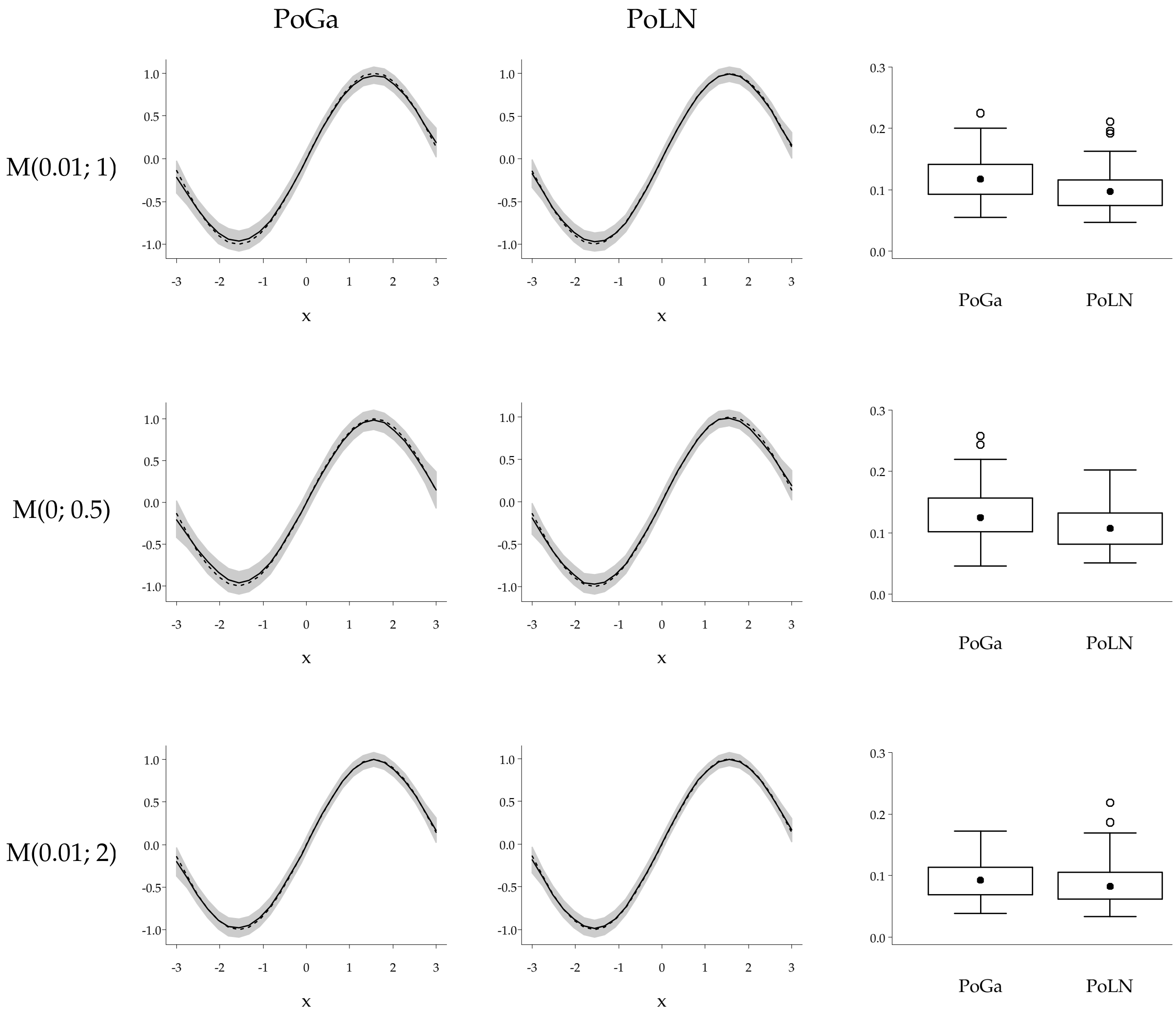

Figure 1: Average posterior mean estimates and MSE Box Plots for models M(0.01;1), M(0;0.5) and M(0.01;2)

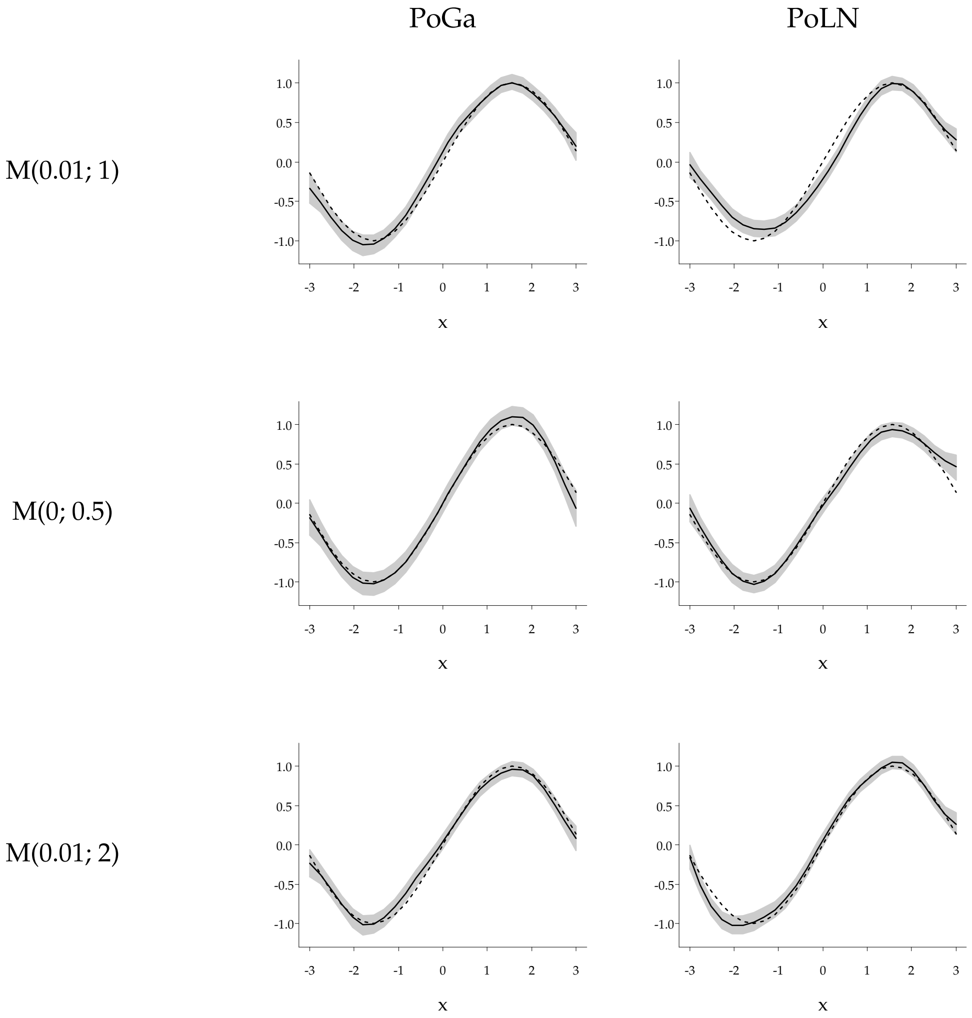

Figure 2: Selected estimates for models M(0.01;1), M(0;0.5) and M(0.01;2)

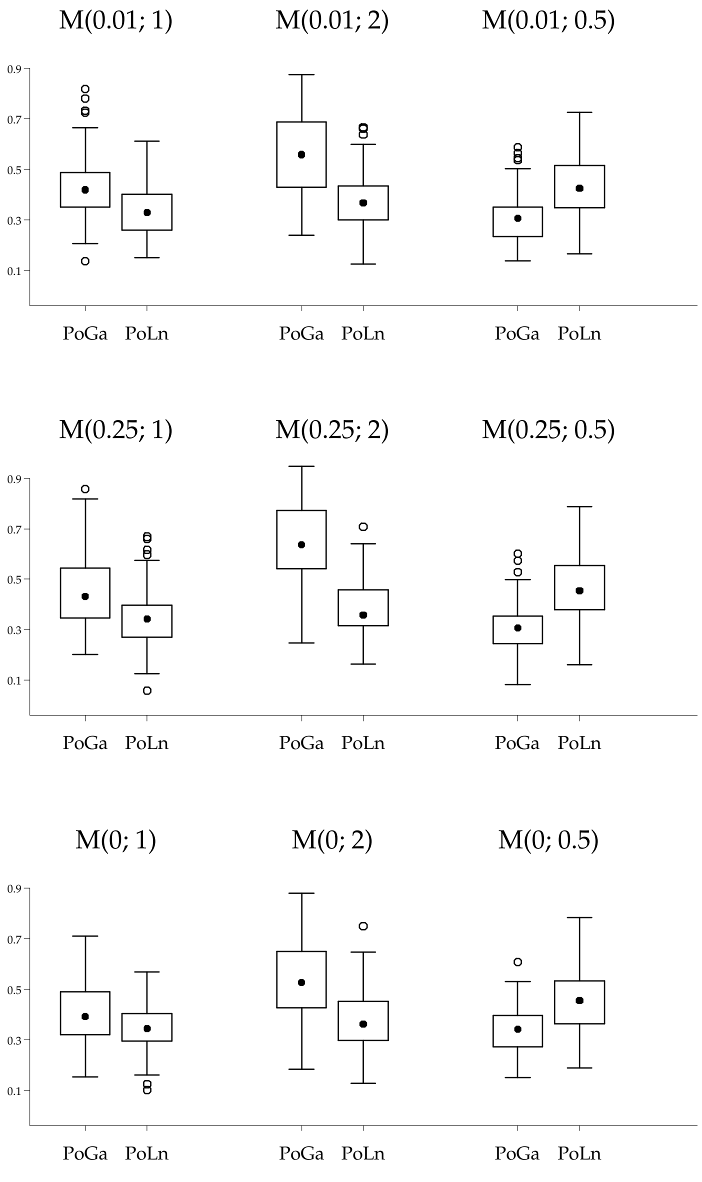

Figure 3: MSE Box Plots for posterior mean estimates of group indicator effects

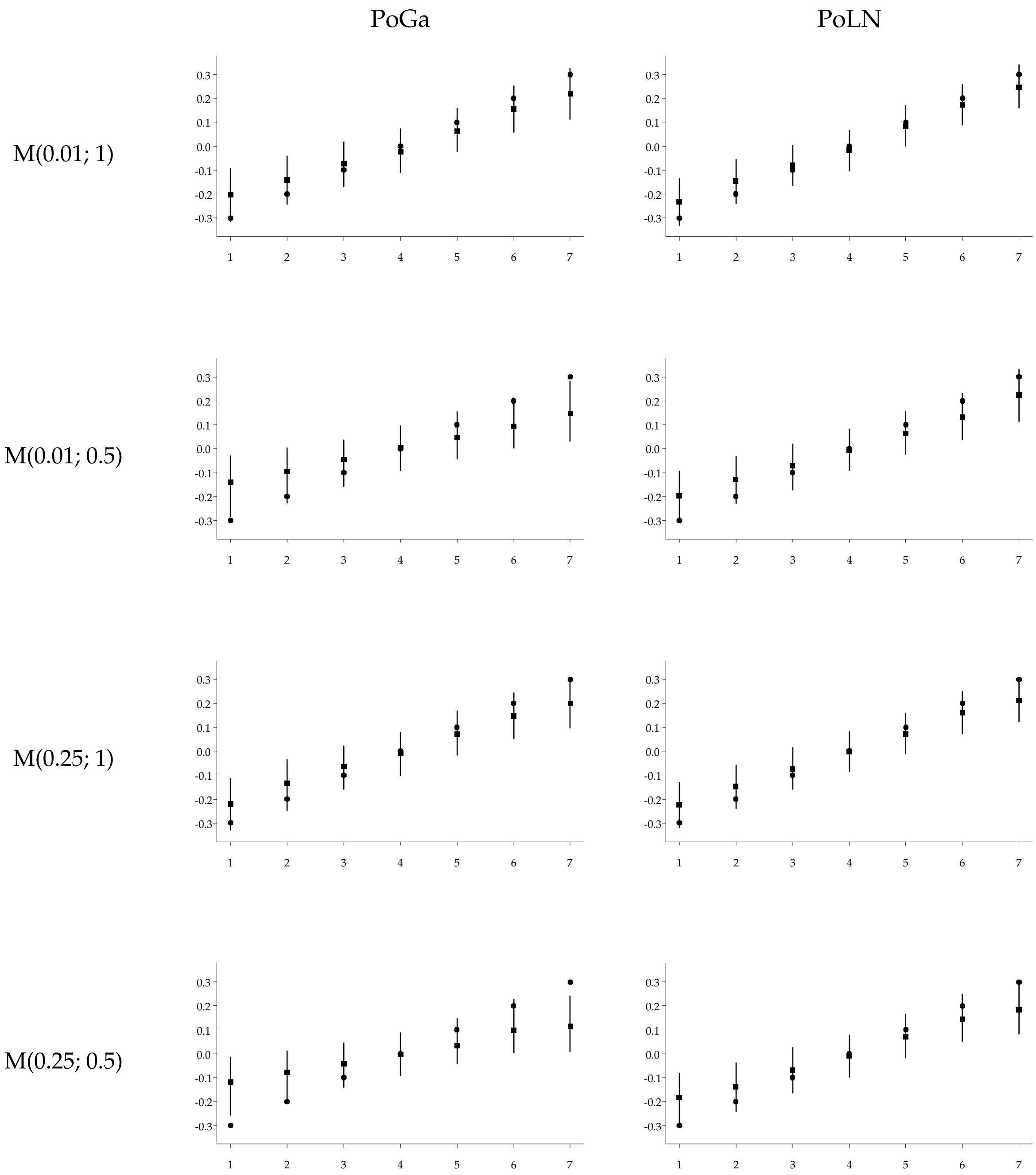

Figure 4: True effects and average posterior means of group indicator effects for selected models

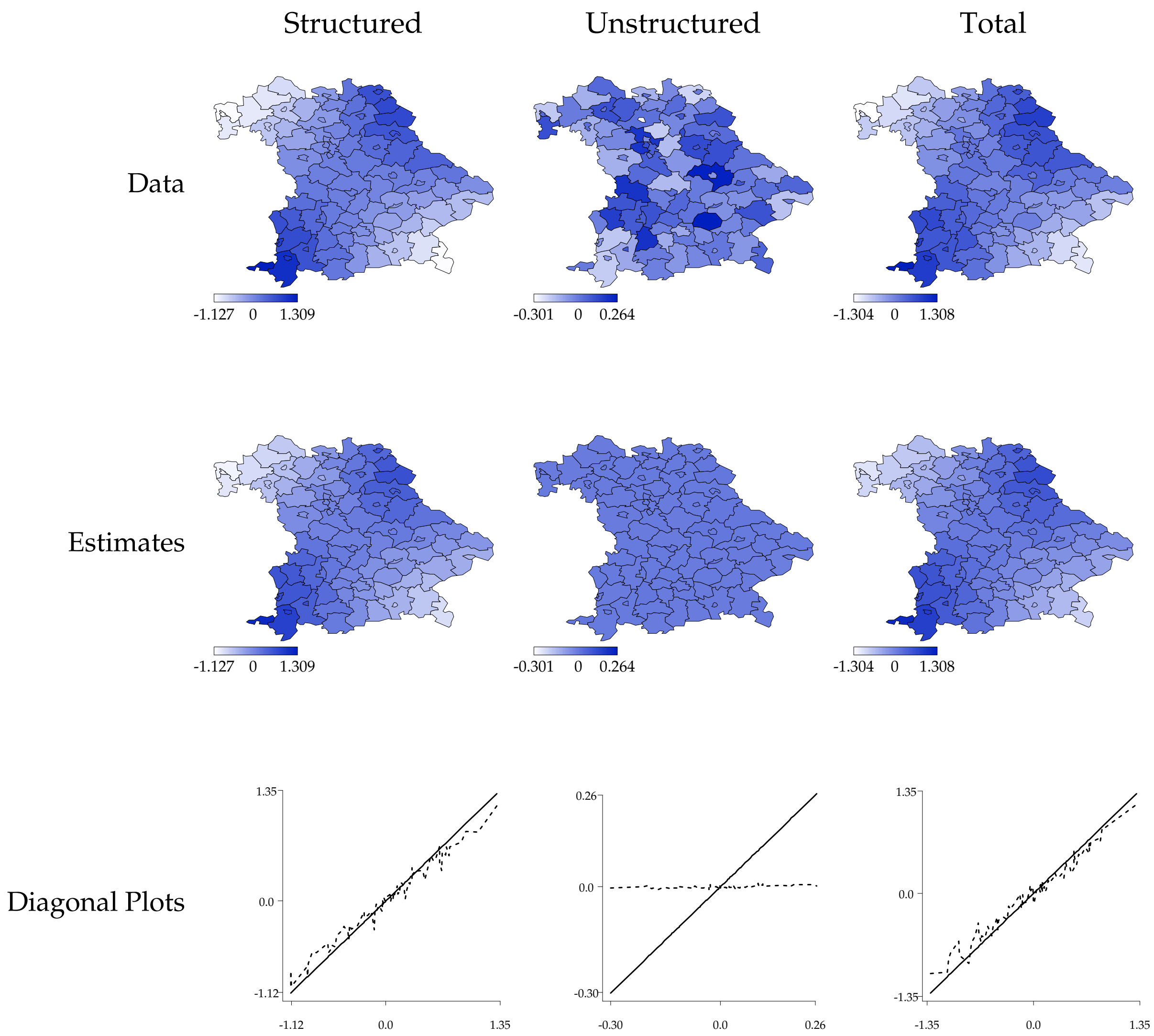

Figure 5: True and estimated structured (left row), unstructured (middle) and total (right) spatial effects toghether with diagonal plots (true versus estimated effects) obtained for the PoGa model M(0.01, 1)

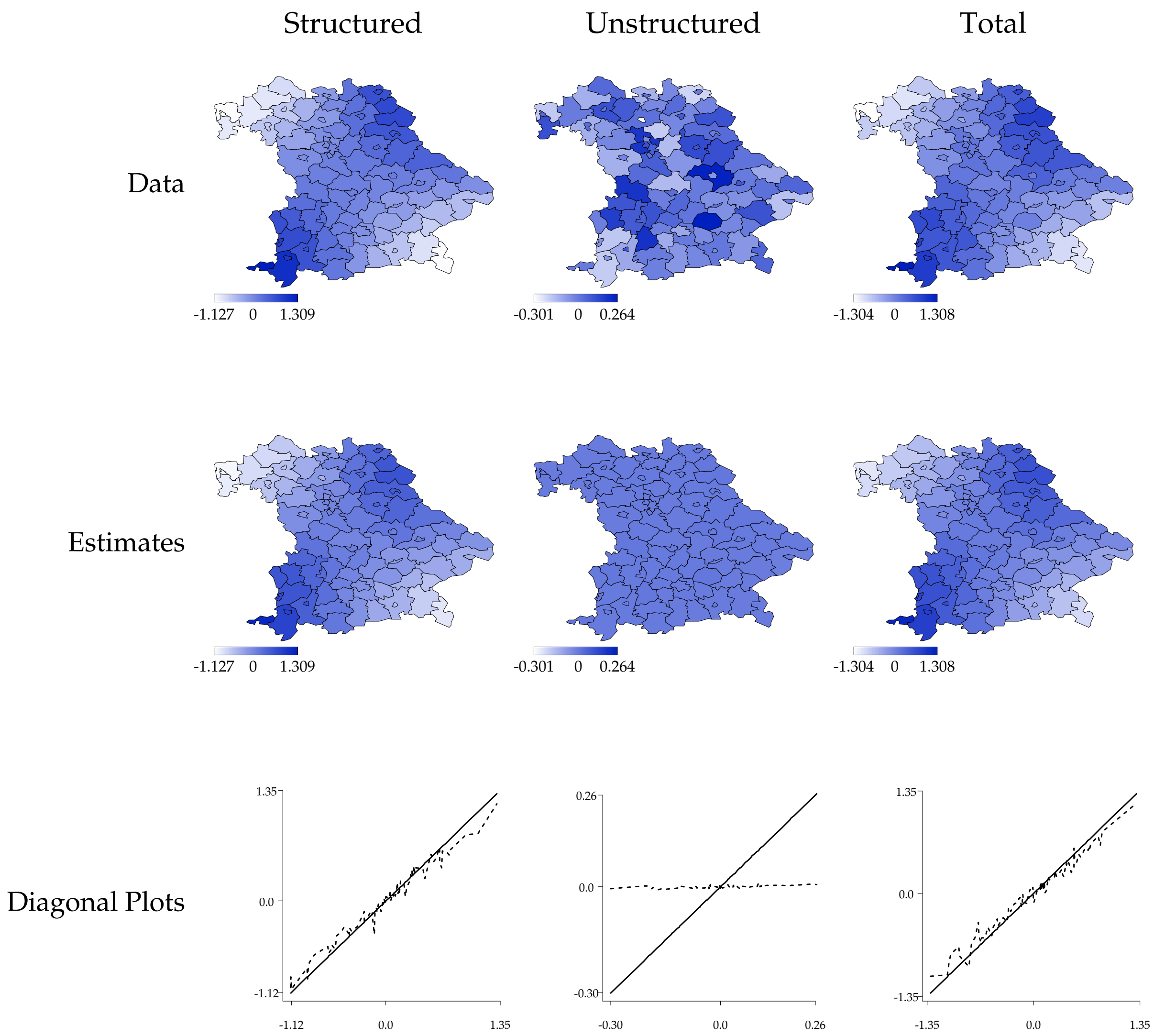

Figure 6: True and estimated structured (left row), unstructured (middle) and total (right) spatial effects toghether with diagonal plots (true versus estimated effects) obtained for the PoLN model M(0.01, 1)

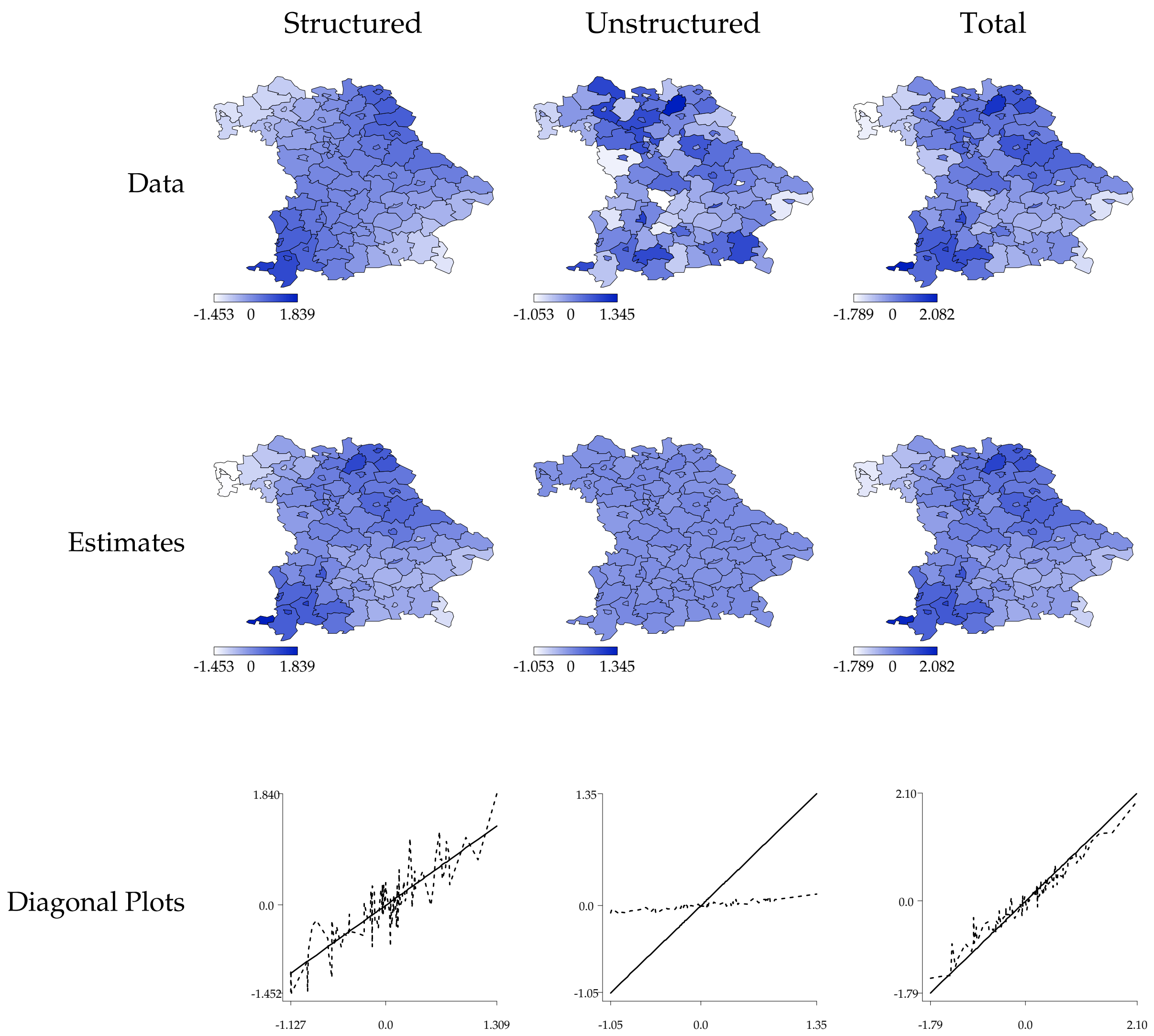

Figure 7: True and estimated structured (left row), unstructured (middle) and total (right) spatial effects toghether with diagonal plots (true versus estimated effects) obtained for the PoGa model M(0.25, 1)

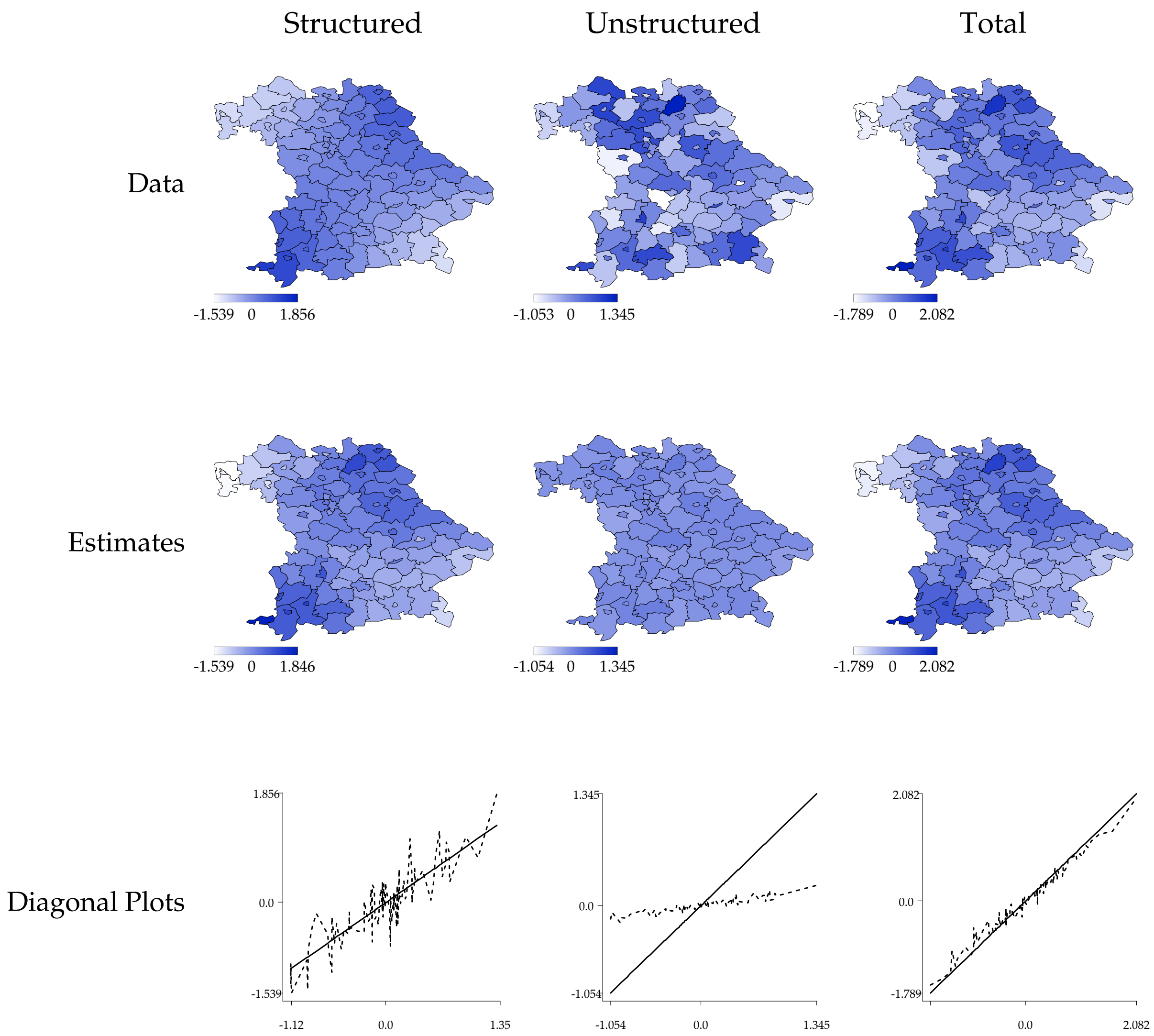

Figure 8: True and estimated structured (left row), unstructured (middle) and total (right) spatial effects toghether with diagonal plots (true versus estimated effects) obtained for the PoLN model M(0.25, 1)

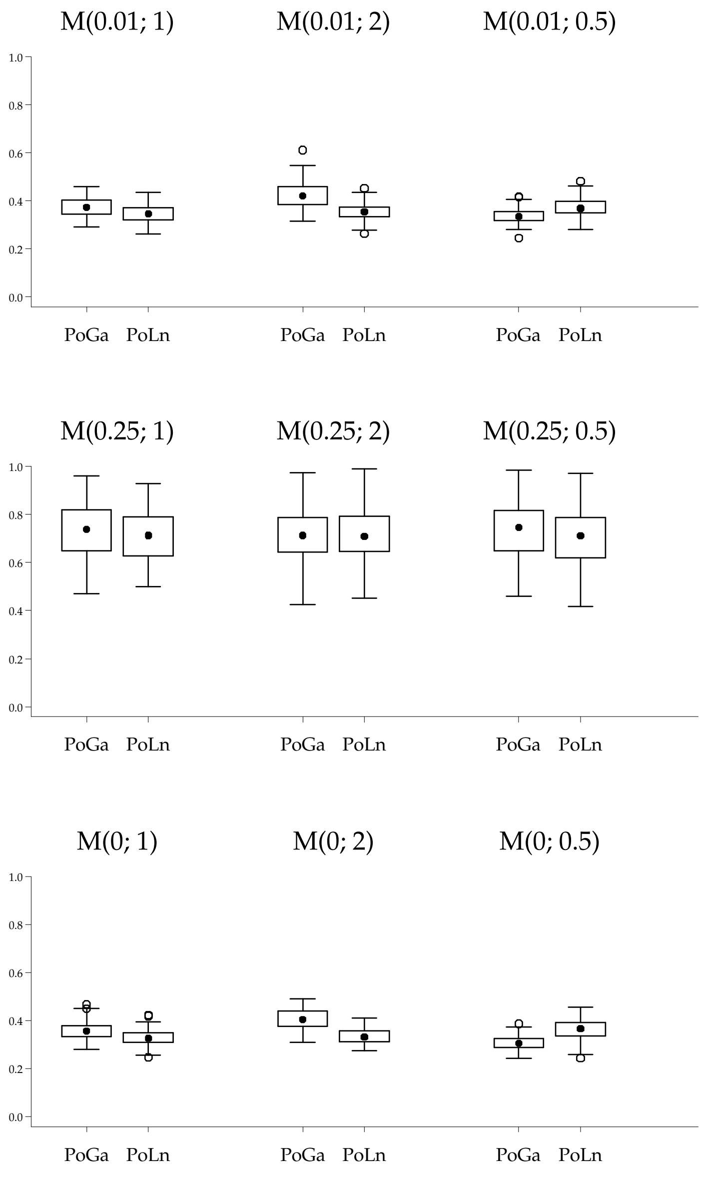

Figure 9: MSE Box Plots for posterior mean estimates of structured spatial effects

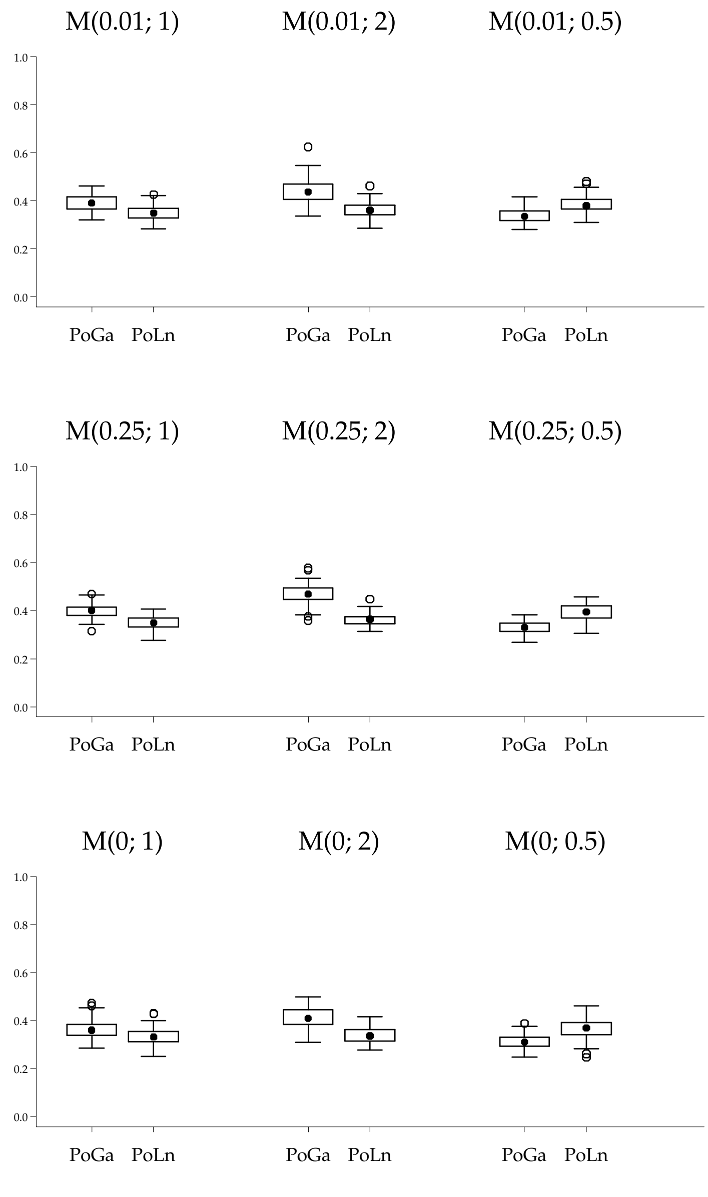

Figure 10: MSE Box Plots for posterior mean estimates of total spatial effects

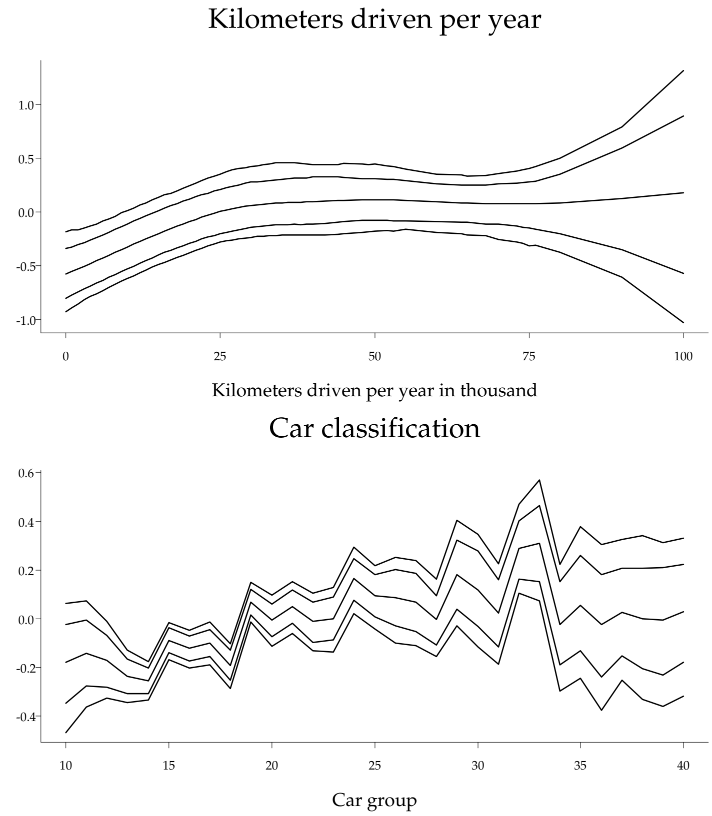

Figure 11: Estimated nonlinear functions f_1 and f_2. Shown is the posterior mean toghether with 80% and 95% pointwise credible intervals

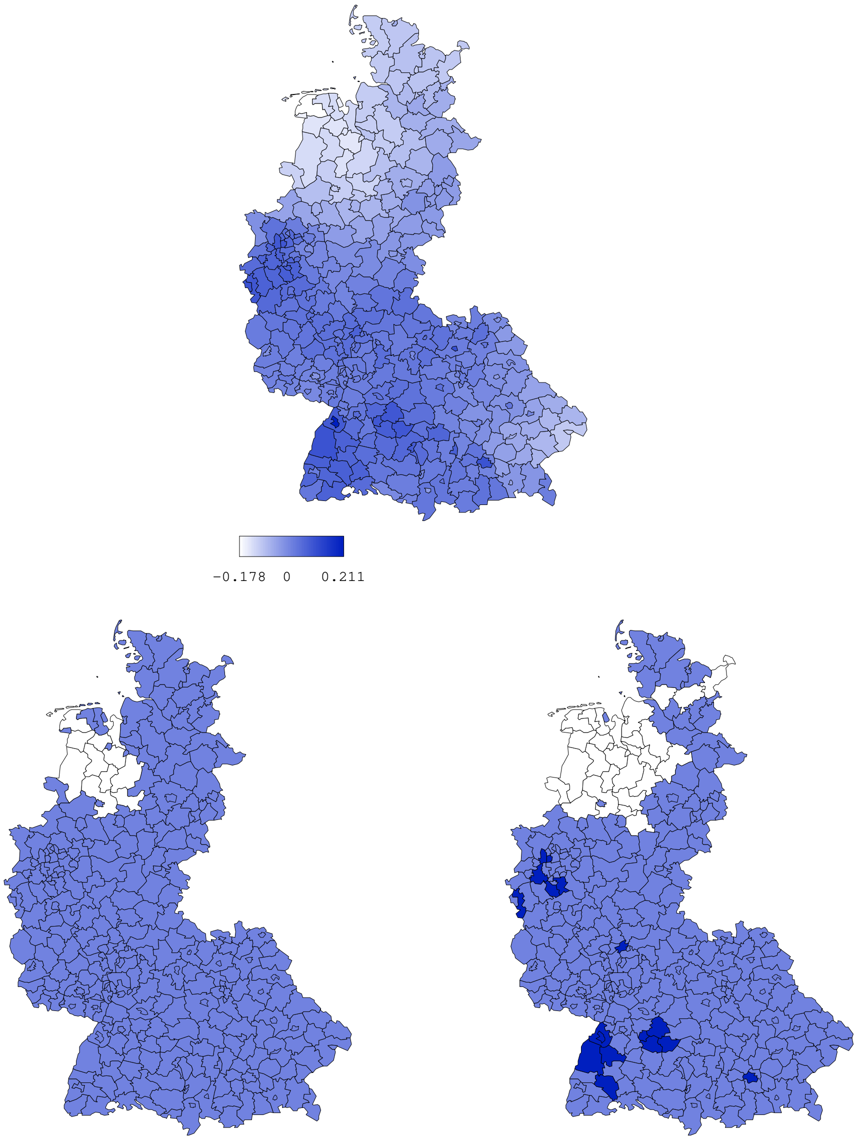

Figure 12: Structured spatial effect. The top panel shows the posterior mean, the bottom left and right panels display posterior probabilities based on nominal levels of 95% and 80%, respectively. White colored regions correspond to strictly negative credible intervals and blue colored regions to strictly positive intervals. Districs with credible intervals containing zero are colored in light blue.

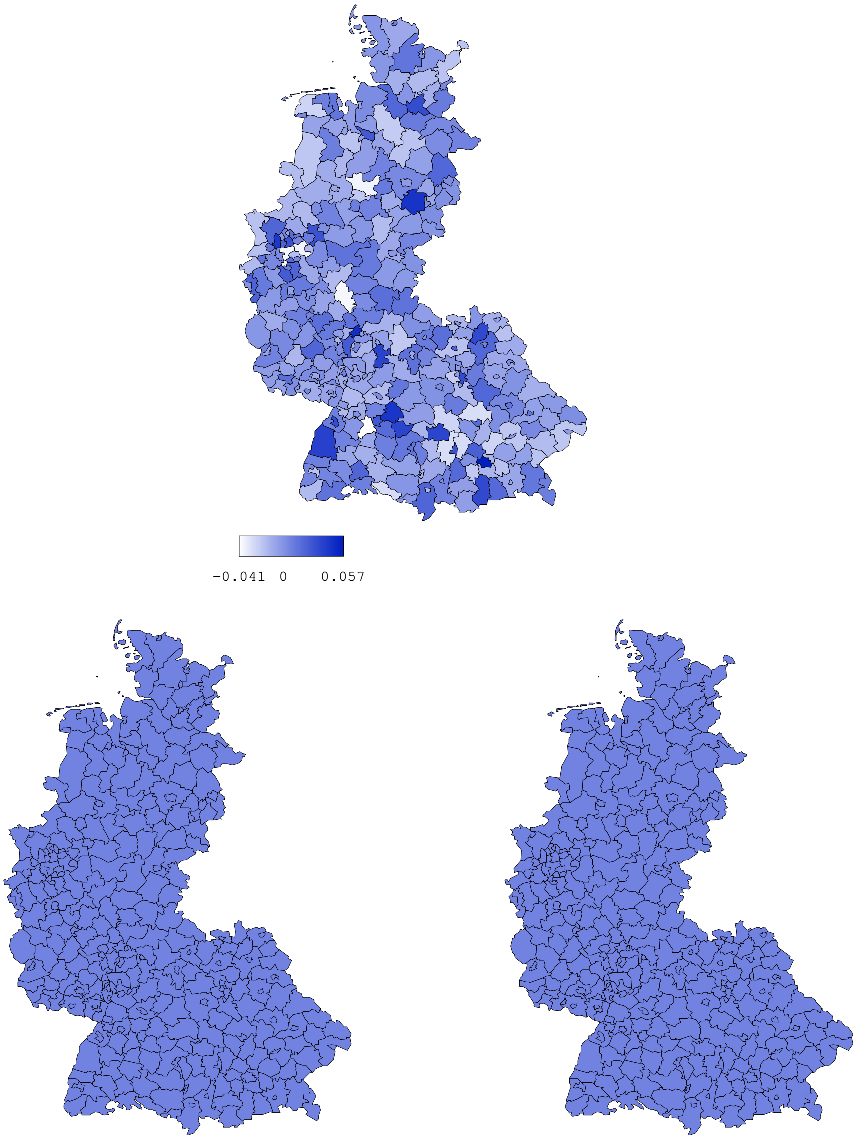

Figure 13: Unstructured spatial effect. The top panel shows the posterior mean, the bottom left and right panels display posterior probabilities based on nominal levels of 95% and 80%, respectively. White colored regions correspond to strictly negative credible intervals and blue colored regions to strictly positive intervals. Districs with credible intervals containing zero are colored in light blue.

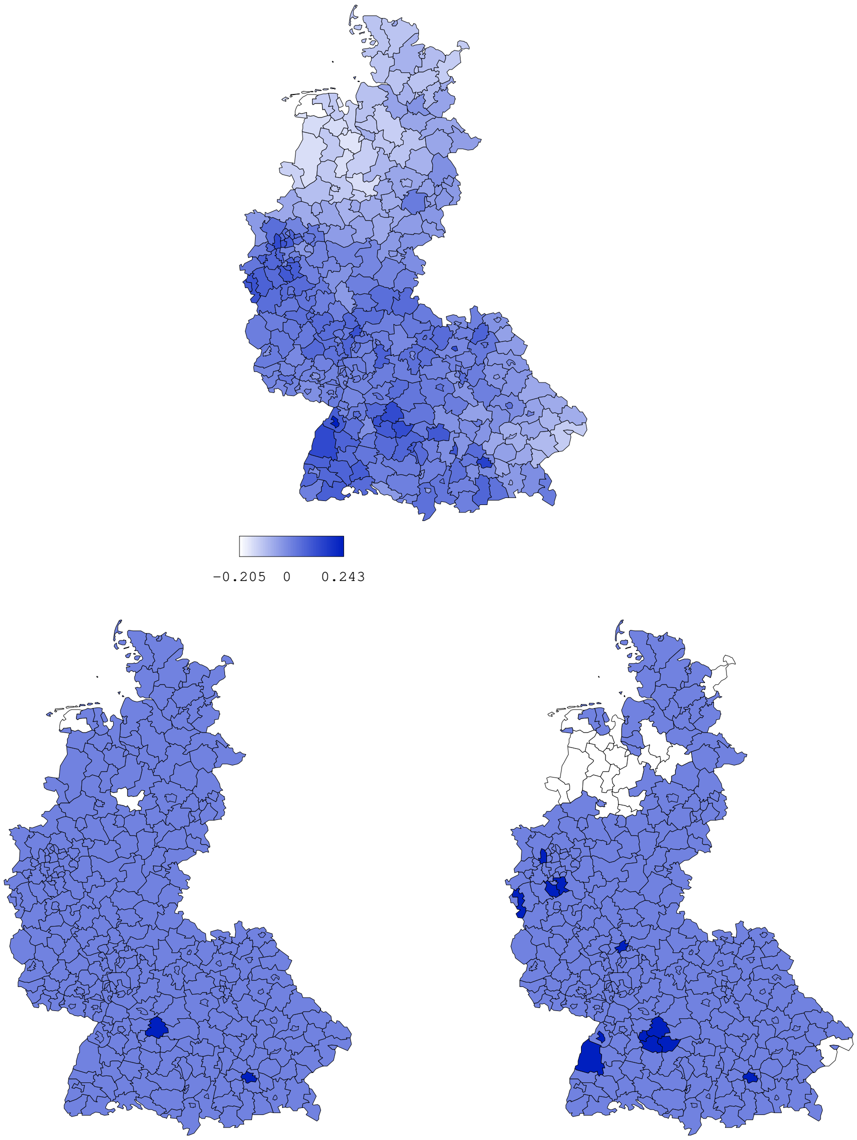

Figure 14: Sum of the structured and the unstructured spatial effect. The top panel shows the posterior mean, the bottom left and right panels display posterior probabilities based on nominal levels of 95% and 80%, respectively. White colored regions correspond to strictly negative credible intervals and blue colored regions to strictly positive intervals. Districs with credible intervals containing zero are colored in light blue.