# Spatial modelling of claim frequency and claim size in non-life insurance

Susanne Gschlößl Claudia Czado *

April 13, 2007

**Abstract**

In this paper models for claim frequency and average claim size in non-life insurance are considered. Both covariates and spatial random effects are included allowing the modelling of a spatial dependency pattern. We assume a Poisson model for the number of claims, while claim size is modelled using a Gamma distribution. However, in contrast to the usual compound Poisson model, we allow for dependencies between claim size and claim frequency. A fully Bayesian approach is followed, parameters are estimated using Markov Chain Monte Carlo (MCMC). The issue of model comparison is thoroughly addressed. Besides the deviance information criterion and the predictive model choice criterion, we suggest the use of proper scoring rules based on the posterior predictive distribution for comparing models. We give an application to a comprehensive data set from a German car insurance company. The inclusion of spatial effects significantly improves the models for both claim frequency and claim size and also leads to more accurate predictions of the total claim sizes. Further we detect significant dependencies between the number of claims and claim size. Both spatial and number of claims effects are interpreted and quantified from an actuarial point of view.

Key words: compound Poisson model, non-life insurance, proper scoring rules, spatial regression models, Bayesian inference

*Both at Center of Mathematical Sciences, Munich University of Technology, Boltzmannstr.3, D-85747 Garching, Germany, email: cczado@ma.tum.de, susanne@ma.tum.de, http://www.ma.tum.de/m4/

# 1 Introduction

In this paper statistical models for the number of claims and the average claim size in non-life insurance are discussed in a Bayesian context. Based on these models the total claim size can be simulated which is fundamental for premium calculation. In particular we consider regression models for spatially indexed data and allow for an underlying spatial dependency pattern by the inclusion of correlated spatial random effects. One important contribution of this paper is that we further allow for dependencies between the number of claims and claim size. This is in contrast to the classical compound Poisson model going back to Lundberg (1903), where independence between claim frequency and claim size is assumed. Further the issue of model comparison is discussed in detail. In particular, we aim to present that proper scoring rules (Gneiting and Raftery (2007)) can be applied for comparing models with regard to their posterior predictive distribution. We apply our approach to a large data set from a German car insurance and quantify the impact of spatial and number of claims effects from an actuarial perspective.

In the classical Poisson-Gamma model the number of claims is assumed to follow a Poisson distribution and to be independent of the claim sizes which are modelled by a Gamma distribution. The use of generalised linear models (GLMs) in actuarial science has been discussed by Haberman and Renshaw (1996) who give several applications, including premium rating in non-life insurance based on models for claim frequency and average claim size. A more detailed study of GLMs for claim frequency and average claim sizes taking covariate information into account is given in Renshaw (1994). Taylor (1989) and Boskov and Verrall (1994) analyse household contents insurance data incorporating geographic information. Whereas Taylor (1989) uses spline functions, Boskov and Verrall (1994) assume a spatial Bayesian model based on Besag et al. (1991). In both papers adjusted loss ratios are fitted, although Taylor (1989) states that separate models for claim frequency and claim size are preferable.

Another approach, which also does not include a separate analysis of claim size and frequency is given by Jørgensen and de Souza (1994) and Smyth and Jørgensen (2002). They use a compound Poisson model, which they call Tweedie's compound Poisson model due to its association to exponential dispersion models. Based on the joint distribution of the number of claims and the total claim sizes, they model the claim rate, defined by total costs per exposure unit, directly. Separate models for claim frequency and claim size have been used by Dimakos and Frigessi (2002) for determining premiums. Based on a spatial Poisson regression model and a spatial Gamma regression model for the average claim size, they estimate premiums by the product of the expected claim frequency and the expected claim size. This approach relies on the independence assumption of claim frequency and claim size. Here the spatial structure is modelled using an improper Markov Random Field following Besag et al. (1991).

This paper extends the approach by Dimakos and Frigessi (2002). We also prefer a separate analysis of claim frequency and claim size and assume a spatial Poisson regression model for claim frequency and a Gamma model for the average claim size per policyholder. In particu-

lar, separate risk factors for claim frequency and claim size models can be included. However, in contrast to Dimakos and Frigessi (2002), we allow for dependencies between the number of claims and claim size. In particular, claim size is modelled conditionally on the number of claims which allows us to include the observed number of claims as covariate.

We follow a fully Bayesian approach, since random parameters can be used to adjust for parameter uncertainty. Panjer and Willmot (1983) state in this context: "The operational actuarial interpretation is that the risk is first selected from the whole set of risks in accordance with the risk distribution, and the performance of the selected risk is then monitored. The statistical interpretation is essentially Bayesian." Markov Chain Monte Carlo (MCMC) is used for parameter estimation and thus facilitates the desired Bayesian inference.

In this paper spatial dependencies are modelled using a Gaussian conditional autoregressive (CAR) prior introduced by Pettitt et al. (2002) for the spatial effects. CAR models are based on the assumption that the effects of adjacent sites are similar, leading to a spatially smoothed dependency pattern. In contrast to the often used intrinsic CAR model introduced by Besag and Kooperberg (1995) the spatial prior considered here leads to a proper joint distribution of the spatial effects. Other proper modifications of the intrinsic CAR model have been proposed by Sun et al. (1999) and Czado and Prokopenko (2004).

Based on the MCMC output of the models for claim frequency and the average claim size, we approximate the posterior predictive distribution of the total claim sizes using simulation. We would like to emphasise again, that independence of claim size and claim frequency is not necessary here.

We analyse a large data set from a German car insurance company using the above models. In particular, we consider policyholders with full comprehensive car insurance and claims caused by traffic accidents. One of our main interests is to investigate whether, after adjusting for covariate information, models are improved by adding spatial random effects. Further, we study the impact of the observed number of claims as additional covariate for the claim size models and quantify the effects on the expected claim sizes.

Models are compared using several criteria. Next to the well known deviance information criterion (DIC) suggested by Spiegelhalter et al. (2002), the predictive model choice criterion (PMCC), see for example Gelfand and Ghosh (1998), is used for comparing model fit and complexity of the considered models. A novel contribution of this paper consists in the investigation of proper scoring rules (Gneiting and Raftery (2007)) based on the posterior predictive distribution in the context of model comparison. In particular, we compare models under out of sample conditions. Up to now proper scoring rules have only been used for assessing the quality of probabilistic forecasts and determining parameter estimates based on the highest score.

The inclusion of spatial effects leads to a significantly improved model fit both for claim frequency and claim size and more accurate predictions of the total claim sizes are obtained. When spatial effects are neglected the posterior predictive means of the total claim sizes in some regions with particular high (low) observed total claims are estimated considerably lower (higher) than based on the spatial models. Further, effects for the number of claims are significant in the

claim size models. For an increasing number of claims, the average claim sizes tend to decrease. The paper is organized as follows. In Section 2 models for claim frequency and claim size are discussed, information on the assumed prior distributions and the developed MCMC algorithms is given. The criteria used for model comparison are presented in Section 3. In Section 4 we develop and compare models for German car insurance data. Finally, we summarize our results and draw conclusions.

# 2 Spatial regression models for claim frequency and claim size

In insurance premiums are based on the expected total claim size which is determined both by the number of claims and the average claim size. In the following we consider spatial regression models for these quantities. Note that we discuss in this paper models for individual, policyholder specific data and not data aggregated for specific groups of policyholders.

## 2.1 Models for claim frequency

For claim frequency we choose a Poisson regression model with spatial effects. In particular, we assume for the number of claims $N_i, i = 1, .., n$, observed at $J$ regions

$$N_i \sim Poisson(\mu_i^N),$$

with mean $\mu_i^N$ given by

$$\mu_i^N = t_i \exp(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_{j(i)}).$$

Here, $t_i$ denotes the exposure time of policyholder $i$. The covariate vector for the $i$-th observation including an intercept is given by $\mathbf{x}_i = (1, x_{i1}, .., x_{ip})'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, .., \beta_p)'$ denotes the vector of unknown regression parameters. Spatial dependencies are modelled by introducing a random effect $\gamma_j, j = 1, .., J$ for each region. The index $j(i)$ denotes the region where the $i$-th policyholder is residing. All spatial effects are combined in the random vector $\boldsymbol{\gamma} = (\gamma_1, .., \gamma_J)'$. The following prior assumptions complete the model. Since we have little prior information we consider a normal prior with large standard deviation, in particular we assume

$$\boldsymbol{\beta} \sim N_{p+1}(0, \sigma_\beta^2 I_{p+1}) \quad \text{with} \quad \sigma_\beta^2 = 100.$$

Here $I_{p+1}$ denotes the $p+1$-dimensional identity matrix. The proper conditional autoregressive prior with hyperparameters $\sigma^2$ and $\psi$

$$\boldsymbol{\gamma}|\sigma^2, \psi \sim N(\mathbf{0}, \sigma^2 Q^{-1}) \tag{2.1}$$

based on Pettitt et al. (2002) is chosen for the spatial effects. The $(g, h)$-th element of the spatial precision matrix $Q$ is specified by

$$Q_{gh} = \begin{cases} 1 + |\psi| \cdot m_g & g = h \\ -\psi & g \neq h, g \sim h \\ 0 & \text{otherwise} \end{cases} \qquad \forall g, h = 1, .., J,$$

where $g \sim h$ denotes adjacent regions and $m_g$ gives the number of neighbours of region $g$. We define regions to be neighbours if they share a common border. The parameter $\psi$ determines the degree of spatial dependence. For $\psi = 0$ the spatial effects are independent, while for increasing values of $\psi$ an increasing spatial dependency is obtained. Proper priors are assumed for the spatial hyperparameters $\sigma^2$ and $\psi$. In particular, we choose for $\sigma^2$ the noninformative prior

$$\sigma^2 \sim IGamma(a, b) \quad \text{with} \quad a = 1 \quad \text{and} \quad b = 0.005$$

which is a common parameter choice for vague Gamma priors. For $\psi$ the prior with density function $\frac{1}{(1+\psi)^2}$ is utilized which takes high values for small $\psi$. We restrict $\psi \geq 0$, since we expect positive conditional partial correlations between regions. Only the hyperparameter $\frac{1}{\sigma^2}$ can be sampled directly from a Gamma distribution, for the remaining parameters a single component Metropolis Hastings algorithm is implemented. We choose independence proposals for the spatial effects, in particular a t-distribution with 20 degrees of freedom having the same mode and inverse curvature at the mode as the target distribution. As investigated in Gschlößl and Czado (2005) this proposal distribution leads to very good mixing with low Monte Carlo standard errors. For the regression parameters $\boldsymbol{\beta}$ and the spatial hyperparameter $\psi$ symmetric random walk proposal distributions are taken. See for example Gilks et al. (1996) for an introduction to MCMC, details on the Metropolis Hastings algorithm and the choice of proposal distributions.

### 2.2 Modelling the average claim size

It is natural for the analysis of claim size to take only observations with positive claims into account. We will consider models for the average claim size of a policyholder. Gschlößl (2006) also investigates models for the individual claim sizes, but concludes that this leads to very similar results for the data set considered in this paper. Since we aim to allow for the modelling of dependencies between the number and the size of the claims, we consider models for the claim size conditionally on the number of claims. Prior specifications are given at the end of this section. For policyholder $i = 1, .., n$ let $S_{ik}, k = 1, .., N_i$, denote the individual claim sizes for the $N_i$ observed claims. In this paper we are interested in models for claim sizes resulting from traffic accidents in car insurance, but not including IBNR (incurred but not reported) losses. The latter type of data are typically skewed and do not contain extremely high claims, which would require the use of heavy tailed distributions like the Pareto distribution (see for example Mikosch (2004)). Therefore a Gamma model is sufficient. In particular, we assume that individual claim sizes conditionally on $N_i$ are independently Gamma distributed, i.e.

$$S_{ik}|N_i \sim Gamma(\mu_i^S, v),\ k = 1, .., N_i, i = 1, .., n \tag{2.2}$$

with mean and variance given by $E(S_{ik}|N_i) = \mu_i^S$ and $Var(S_{ik}|N_i) = \frac{(\mu_i^S)^2}{v}$. We use the following parameterisation of the Gamma distribution: $f(s_{ik}|\mu_i^S, v) = \frac{v}{\mu_i^S \Gamma(v)} \left(\frac{v s_{ik}}{\mu_i^S}\right)^{v-1} \exp\left(-\frac{v s_{ik}}{\mu_i^S}\right)$. The average claim size $S_i$ for policyholder $i$ is given by

$$S_i := \sum_{k=1}^{N_i} \frac{S_{ik}}{N_i}.$$

Since we assume $S_{ik}|N_i, k = 1, .., N_i$ to be independent and identically distributed, the average claim size $S_i$ given the observed number of claims $N_i$ is again Gamma distributed with mean $E(S_i|N_i) = \mu_i^S$ and variance $Var(S_i|N_i) = \frac{(\mu_i^S)^2}{N_i v}$, i.e.

$$S_i|N_i \sim Gamma(\mu_i^S, N_i v). \tag{2.3}$$

We perform a regression on the mean $\mu_i^S$ including covariates $\mathbf{w}_i$ and spatial effects $\boldsymbol{\zeta}$ for the $J$ geographical regions. By choosing a log link we obtain the following mean specification:

$$\mu_i^S = \exp(\mathbf{w}_i'\boldsymbol{\alpha} + \zeta_{j(i)}).$$

Here $\mathbf{w}_i = (1, w_{i1}, .., w_{iq})'$ denotes the vector of covariates for the $i$-th observation with intercept, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, .., \alpha_q)'$ the vector of unknown regression parameters and $\boldsymbol{\zeta} = (\zeta_1, .., \zeta_J)$ the vector of spatial effects, which are modelled by a similar CAR prior as in the Poisson model for claim frequency. Since we consider a model for average claim sizes conditionally on the number of claims, the observed number of claims $N_i$ may be introduced as a covariate as well. The number of claims per policyholder observed in car insurance data is typically very low, therefore we include $N_i$ as a factor covariate with reference level $N_i = 1$, denoted by $\alpha_{N_i=k}, k = 2, .., \text{argmax}_i N_i$. The mean $\mu_i^S$ is therefore given by

$$\mu_i^S \quad = \quad \exp(\mathbf{w}_i'\boldsymbol{\alpha} + \zeta_{j(i)}) = \exp(\sum_{l=0}^{q} w_{il}\alpha_l + \sum_{k=2}^{\text{argmax}_i N_i} D_{ki}\alpha_{N_i=k} + \zeta_{j(i)}), \tag{2.4}$$

where $D_{ki} = \begin{cases} 1, & N_i = k \\ 0, & \text{otherwise} \end{cases}$. Similar to the Poisson model for the number of claims we have little prior knowledge on the regression parameters $\boldsymbol{\alpha}$ in the model for the average claim size. Therefore we assume a normal prior with large standard deviation, in particular,

$$\boldsymbol{\alpha} \sim N_{q+\text{argmax}_i N_i}(0, \sigma_\alpha^2 I_{q+\max_i N_i})$$

with $\sigma_\alpha^2 = 100$ which is a rather uninformed prior. For the scale parameter $v$ the gamma prior $v|a, b \sim Gamma(a, b)$, i.e. $\pi(v|a, b) = \frac{b^a}{\Gamma(a)} v^{a-1} \exp(-vb)$ with $a = 1$ is assumed, the conditional mean and variance are given by $E(v|a = 1, b) = \frac{1}{b}$ and $Var(v|a = 1, b) = \frac{1}{b^2}$, respectively. Following a fully Bayesian approach we also assign a noninformative gamma prior to the hyperparameter $b$, in particular $b|c, d \sim Gamma(c, d)$ with $c = 1$ and $d = 0.005$, yielding $E(b|c, d) = 200$ and $Var(b|c, d) = 40000$. However, the models turn out to be very robust with respect to the prior on $b$, a very similar estimated posterior mean of $v$ is obtained when $b$ is fixed to 0.005, which is a popular choice for a flat gamma prior.
The spatial effects are modelled using the conditional autoregressive prior (2.1), i.e.

$$\boldsymbol{\zeta}|\sigma^2, \psi \sim N_J(\mathbf{0}, \sigma^2 Q^{-1}),$$

assuming the same prior distributions for the hyperparameters $\sigma^2$ and $\psi$ as in Section 2.1. The hyperparameter $b$ can be sampled directly from a gamma distribution, for the regression parameters, the spatial effects and spatial hyperparameters a single component Metropolis Hastings

algorithm with the same proposal distributions as in the Poisson model discussed in Section 2.1 is used. For the scale parameter $v$ a symmetric random walk proposal distribution is taken. MCMC algorithms are implemented in Matlab, but OpenBugs could have been used as well. However we prefer Matlab since we had more control over the implementation.

# 3 Model comparison

For complex hierarchical models like those considered in this paper, the computation of Bayes factors (see for example Kass and Raftery (1995)) requires substantial efforts (compare to Han and Carlin (2001)). Therefore, we consider here model choice criteria and scoring rules which can be easily computed using the available MCMC output. In this paper only nested models are compared, however, the criteria presented in this section also can be used for comparing non nested models.

## 3.1 Deviance Information Criterion (DIC) and Predictive Model Choice Criterion (PMCC)

The deviance information criterion (DIC), suggested by Spiegelhalter et al. (2002), for a probability model $p(\mathbf{y}|\boldsymbol{\theta})$ with observed data $\mathbf{y} = (y_1, .., y_n)$ and unknown parameters $\boldsymbol{\theta}$ is defined by

$$DIC := E[D(\boldsymbol{\theta}|\mathbf{y})] + p_D.$$

It considers both model fit as well as model complexity. The goodness-of-fit is measured by the posterior mean of the Bayesian deviance $D(\boldsymbol{\theta})$ defined as

$$D(\boldsymbol{\theta}) = -2 \log p(\mathbf{y}|\boldsymbol{\theta}) + 2 \log f(\mathbf{y})$$

where $f(\mathbf{y})$ is some fully specified standardising term. Model complexity is measured by the effective number of parameters $p_D$ defined by

$$p_D \quad := \quad E[D(\boldsymbol{\theta}|\mathbf{y})] - D(E[\boldsymbol{\theta}|\mathbf{y}]).$$

According to this criterion the model with the smallest DIC is to be preferred. Using the available MCMC output both $p_D$ and DIC are easily computed by taking the posterior mean of the deviance $E[D(\boldsymbol{\theta}|\mathbf{y})]$ and the plug-in estimate of the deviance $D(E[\boldsymbol{\theta}|\mathbf{y}])$. We will compute the DIC with the standardising term $f(\mathbf{y})$ set to zero. An information theoretic discussion of the DIC as criterion for posterior predictive model comparison is given in van der Linde (2005).
A related model comparison approach is given by the predictive model choice criterion (PMCC) considered by Laud and Ibrahim (1995) and Gelfand and Ghosh (1998). It is based on the posterior predictive distribution given by $p(\mathbf{y}_{rep}|\mathbf{y}) = \int p(\mathbf{y}_{rep}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$ where $\mathbf{y}_{rep} = (y_{rep,1}, .., y_{rep,n})$ denotes a new, replicated data set. Here $\mathbf{y}_{rep}$ and $\mathbf{y}$ are assumed to be independent given $\boldsymbol{\theta}$.

The posterior predictive distribution can be estimated by $\hat{p}(\mathbf{y}_{rep}|\mathbf{y}) := \frac{1}{R}\sum_{r=1}^{R} p(\mathbf{y}|\hat{\boldsymbol{\theta}}^r)$ where $\hat{\boldsymbol{\theta}}^r, r = 1, .., R$ denotes the $r$-th MCMC iterate of $\boldsymbol{\theta}$ after burnin. The PMCC is defined by

$$PMCC := \sum_{i=1}^{n}(\mu_i - y_i)^2 + \sum_{i=1}^{n}\sigma_i^2, \tag{3.5}$$

where $\mu_i := E(y_{rep,i}|\mathbf{y})$ and $\sigma_i^2 := Var(y_{rep,i}|\mathbf{y})$ denote the expected value and the variance of a replicate $y_{rep,i}$ of the posterior predictive distribution. Similar to DIC, models with a smaller PMCC value are preferred. While the first term $\sum_{i=1}^{n}(\mu_i - y_i)^2$ gives a goodness-of-fit measure which will decrease with increasing model complexity, the second term $\sum_{i=1}^{n}\sigma_i^2$ can be considered as penalty term which will tend to be large both for poor and overfitted models (see Gelfand and Ghosh (1998)). The quantities $\mu_i$ and $\sigma_i^2$ can be estimated based on the MCMC output $\hat{\boldsymbol{\theta}}^r, r = 1, .., R$ by $\hat{\mu}_i := \frac{1}{R}\sum_{r=1}^{R}\mu_i(\hat{\boldsymbol{\theta}}^r)$ and $\hat{\sigma}_i^2 := \frac{1}{R}\sum_{r=1}^{R}\sigma_i^2(\hat{\boldsymbol{\theta}}^r)$, where $\mu_i(\boldsymbol{\theta})$ and $\sigma_i^2(\boldsymbol{\theta})$ denote the mean and the variance of the underlying model $p(\mathbf{y}|\boldsymbol{\theta})$ depending on the parameters $\boldsymbol{\theta}$. When the mean $\mu_i(\boldsymbol{\theta})$ and the variance $\sigma_i^2(\boldsymbol{\theta})$ of the model are not explicitly available, the PMCC can be alternatively evaluated using simulation. For every MCMC iteration $r = 1, .., R$ after burnin, a replicated data set $\mathbf{y}_{rep}^r = (y_{rep,1}^r, .., y_{rep,n}^r)$ can be simulated from $p(\mathbf{y}|\hat{\boldsymbol{\theta}}^r)$. The mean $\mu_i$ and the variance $\sigma_i^2$ can then be estimated by the empirical counterparts $\hat{\mu}_i := \frac{1}{R}\sum_{r=1}^{R} y_{rep,i}^r$ and $\hat{\sigma}_i^S := \frac{1}{R-1}\sum_{r=1}^{R}(y_{rep}^r - \hat{\mu}_i)^2$. In the application given in Section 4 we compare models for average and total claim sizes using PMCC. Since the mean $\mu_i(\boldsymbol{\theta})$ and the variance $\sigma_i^2(\boldsymbol{\theta})$ are explicitly given in the models for the number of claims and for the average claim sizes, we will compute the PMCC directly using the MCMC output for these models. The distribution of the total claim sizes however, is not available in an analytically closed form, therefore here, the PMCC will be evaluated using simulation as described above.

### 3.2 Scoring rules for continuous variables

Gneiting and Raftery (2007) consider scoring rules for assessing the quality of probabilistic forecasts. A scoring rule assigns a numerical score based on the forecast of the predictive distribution for a specific model and the value that was observed. It can be used for comparing the predictive distribution of several models. Ideally, both calibration and sharpness of the predictive distribution are taken into account. Gneiting and Raftery (2007) also use scoring rules in estimation problems for assessing the optimal score estimator for the unknown model parameters. Assume a parametric model $P_\theta := p(\mathbf{y}|\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$ based on the sample $\mathbf{y} = (y_1, .., y_n)$. Then, the mean score

$$S_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} S(P_\theta, y_i)$$

can be taken as a goodness-of-fit measure, where $S$ is a strictly proper scoring rule, i.e. the highest score is obtained for the true model. Since for the true parameter vector $\boldsymbol{\theta}_0$ (see Gneiting and Raftery (2007))

$$\text{argmax}_\theta S_n(\boldsymbol{\theta}) \to \boldsymbol{\theta}_0, n \to \infty,$$

the optimum score estimator based on scoring rule $S$ is given by $\hat{\boldsymbol{\theta}}_n = \text{argmax}_\theta S_n(\boldsymbol{\theta})$. We will use scoring rules in a Bayesian context as measures for comparing models based on their posterior predictive distribution. Gneiting and Raftery (2007) provide and discuss several scoring rules, we will present some of the proper scoring rules for continuous variables here. In particular, we consider the logarithmic score (LS), the continuous ranked probability score (CRPS), the interval score (IS) and a score for quantiles which we will denote as quantile score (QS). All these scores are positively oriented, i.e. the model with the highest mean score $S_n(\boldsymbol{\theta})$ is favoured.
The logarithmic score LS is given by

$$LS(p(\mathbf{y}_{rep}|\mathbf{y}), y_i) := \log p(y_{rep} = y_i|\mathbf{y}),$$

where $p(y_{rep} = y_i|\mathbf{y})$ denotes the posterior predictive density at $y_{rep} = y_i$ of the model under consideration. When a sample of MCMC iterates $\hat{\boldsymbol{\theta}}^r, r = 1, .., R$ after burnin is available, an estimate of $\log p(y_{rep} = y_i|\mathbf{y})$ for the $i$-th observation is straightforward, i.e.

$$\hat{\log} p(y_{rep} = y_i|\mathbf{y}) := \log\Big(\frac{1}{R}\sum_{r=1}^{R} p(y_i|\hat{\boldsymbol{\theta}}^r)\Big),$$

where $p(y|\hat{\boldsymbol{\theta}}^r)$ denotes the density at the observed value $y$ based on the $r$-th MCMC iterates. In contrast to the logarithmic score which only considers the posterior predictive density evaluated at the observed value, the following scoring rules take both calibration and sharpness into account.
The continuous ranked probability score CRPS for a parametric model $P_\theta$ with posterior predictive cumulative density function (cdf) $F(x) := \int_{-\infty}^{x} p(\tilde{y}|\mathbf{y})d\tilde{y}$ is defined by

$$CRPS(F, y_i) \;=\; -\int_{-\infty}^{\infty} (F(x) - 1_{\{x \geq y_i\}})^2 dx,$$

where $1_{\{x \geq y\}}$ takes the value 1 if $x \geq y$ and 0 otherwise. Hence, the CRPS can be interpreted as the integrated squared difference between the predictive and the empirical cdf based on the single observation $y_i$. A graphical illustration of the CRPS is presented in Figure 1 when $P_\theta$ is a normal distribution. Here the pdf of a normal distribution with mean 0 and standard deviation 1 (left panel in first row) and 4 (left panel in second row) respectively is plotted. The difference between the corresponding cdf and the empirical cdf for two observations $y = 0$ and $y = 2$ is indicated in the middle and right plot of each row as dashed regions. These plots show that the CRPS rewards sharp distributions, but also takes into account if the observation $y$ is close to the center or rather in the tails of the distribution. According to Székely (2003) the CRPS can be expressed as

$$CRPS(P_\theta, y_i) \;=\; \frac{1}{2}E|y_{rep,i} - \tilde{y}_{rep,i}| - E|y_{rep,i} - y_i|. \tag{3.6}$$

Here $y_{rep,i}, \tilde{y}_{rep,i}$ are independent replicates from the posterior predictive distribution $p(\cdot|\mathbf{y})$ and the expectation is taken with respect to $p(\cdot|\mathbf{y})$. Estimation of the CRPS is again straightforward using the available MCMC output: for $r = 1, .., R$ simulate independently two replicated data sets

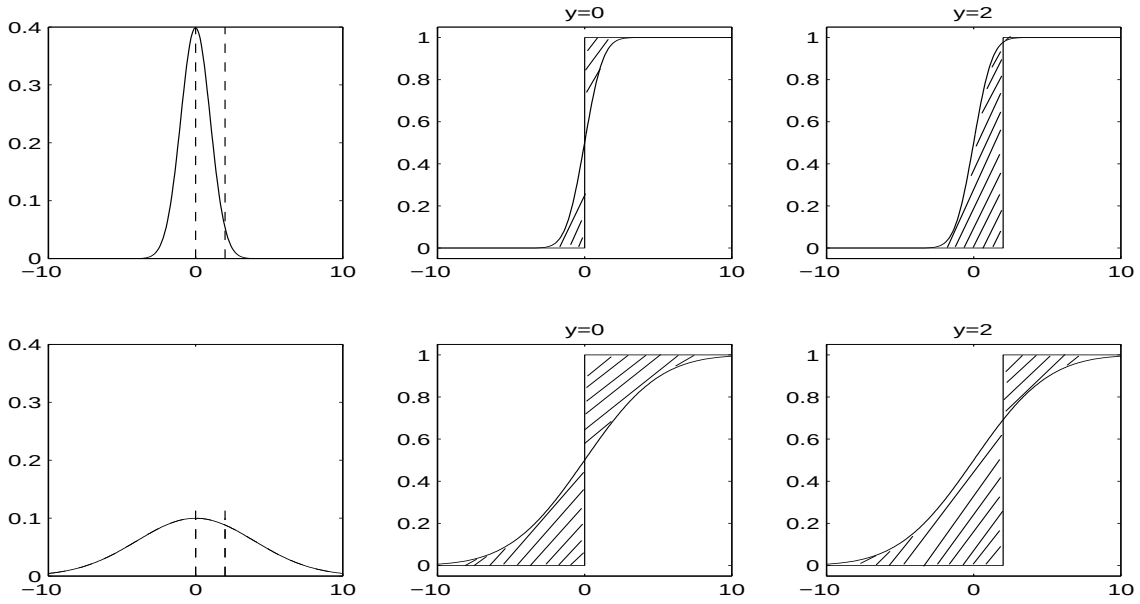


Figure 1: Pdf (left column with $y = 0, 2$ indicated as dashed lines) and cdf of a normal distribution with mean 0 and standard deviation 1 (first row) and 4 (second row), respectively. The differences between the cdf and the empirical cdf for two observations $y = 0$ (middle) and $y = 2$ (right) are indicated as dashed regions.

$\mathbf{y}^r_{rep} = (y^r_{rep,1}, .., y^r_{rep,n}), \tilde{\mathbf{y}}^r_{rep} = (\tilde{y}^r_{rep,1}, .., \tilde{y}^r_{rep,n})$ based on the distribution $p(\mathbf{y}|\hat{\boldsymbol{\theta}}^r)$ and estimate (3.6) by $\hat{E}|y_{rep,i} - \tilde{y}_{,rep,i}| := \frac{1}{R}\sum_{r=1}^{R} |y^r_{rep,i} - \tilde{y}^r_{rep,i}|$ and $\hat{E}|y_{rep,i} - y_i| := \frac{1}{R}\sum_{r=1}^{R} |y^r_{rep,i} - y_i|$.
The interval score $IS_\alpha$ is based on the $(1-\alpha)$ 100 % posterior prediction interval defined by $I = [l_i, u_i]$ where $l_i$ and $u_i$ denote the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantile of the posterior predictive distribution for the $i$-th observation. It rewards narrow prediction intervals and assigns a penalty for observations which are not covered by the interval. The interval score is defined by

$$IS_\alpha(l_i, u_i, y_i) = \begin{cases} -(u_i - l_i) - \frac{2}{\alpha}(l_i - y_i) & \text{if} \quad y_i < l_i \\ -(u_i - l_i) & \text{if} \quad l_i \leq y_i \leq u_i \\ -(u_i - l_i) - \frac{2}{\alpha}(y_i - u_i) & \text{if} \quad y_i > u_i \end{cases} .$$

Using the available MCMC output, replicated data sets $\mathbf{y}^r_{rep} = (y^r_{rep,1}, .., y^r_{rep,n}), r = 1, .., R$, can be simulated from which $l_i$ and $u_i, i = 1, .., n$ can be estimated. In order to compare models based on prediction intervals with both moderate and large coverage, we will use $\alpha = 0.1$ and $\alpha = 0.5$, respectively.
As will be seen in the application, the posterior predictive distribution of the total claim size in car insurance typically has most of its mass at zero. In particular, zero will in general be included in the posterior prediction intervals and the interval score will not be appropriate for model comparison. Here one-sided scores might be more interesting to investigate. Gneiting and Raftery (2007) propose a proper scoring rule based on the quantiles $Q_{\alpha,i}$ at level $\alpha \in (0, 1)$ of

the predictive distribution for the $i$-th observation given by

$$S(Q_{\alpha,i}, y_i) = \alpha s(Q_{\alpha,i}) + (s(y_i) - s(Q_{\alpha,i}))1_{\{y_i \leq Q_{\alpha,i}\}} + h(y_i)$$

for a nondecreasing function $s$ and $h$ arbitrary. We will use this score with the special choice $s(x) = x$ and $h(x) = -\alpha x$ and refer to the resulting scoring rule as quantile score $QS_\alpha$ which is given by

$$QS_\alpha(Q_{\alpha,i}, y_i) = (y_i - Q_{\alpha,i})[1_{\{y_i \leq Q_{\alpha,i}\}} - \alpha].$$

Similar to the interval score, the $\alpha$-quantile $Q_{\alpha,i}$ of the posterior predictive distribution can be computed based on the MCMC output and evaluation of the quantile score is straightforward. The predictive model choice criterion PMCC discussed in the previous section, can be expressed as a scoring rule as well. The corresponding positively oriented score function is defined by

$$S(P_\theta, y_i) = -(E(y_{rep,i}|\mathbf{y}) - y_i)^2 - Var(y_{rep,i}|\mathbf{y}).$$

However, this is not a proper scoring rule (see Gneiting and Raftery (2007)) and should be used with care.

# 4 Application

The models considered in Section 2 will now be used to analyse a data set from a German car insurance company. Our main questions of interest for this application are the following: Does the inclusion of spatial effects, after having adjusted for covariate information, improve the model fit and can we observe a spatial pattern for the expected number of claims and the expected claim sizes? Does average claim size of a policyholder depend on the number of observed claims, i.e. are there significant number of claims effects in the models for claim size? Based on the models for the number of claims and claim size, we finally approximate the posterior predictive distribution of the total claim sizes. Here again, we are interested to what extent the inclusion of spatial and claim number effects influences the total claim sizes.

## 4.1 Data description

The data set contains information on policyholders in Germany with full comprehensive car insurance within the year 2000. Not all policyholders were insured during the whole year, however the exposure time $t_i$ of each policyholder is known. Several covariates like age and gender of the policyholder, kilometers driven per year, type of car and age of car are given in the data. The deductible which differs between policyholders will also be included as covariate. Germany is divided into 440 regions, for each policyholder the region he/she is living in is known. We analyse a subset of these data, in particular we only consider traffic accident data for policyholders with three types of midsized cars. The resulting data set contains about 350000 observations. Table 1 shows that there is a very large amount of observations with no claim in the data set and the maximum number of observed claims is only 4.

| number of claims | percentage of observations | number of observations |
|---|---|---|
| 0 | 0.960 | 338330 |
| 1 | 0.039 | 13816 |
| 2 | $6.9 \cdot 10^{-4}$ | 243 |
| 3 | $1.7 \cdot 10^{-5}$ | 6 |
| 4 | $2.8 \cdot 10^{-6}$ | 1 |

Table 1: Summary of the observed claim frequencies in the data.

The histogram of the observed positive individual claim sizes in DM given in Figure 2 reveals that the distribution of the claim sizes is highly skewed. The average individual claim size is given by DM 5371.0, the largest observed claim size takes the value DM 49339.1 which is less than 0.01 % of the sum of all individual claim sizes. Therefore, the use of a Gamma model seems to be justified.

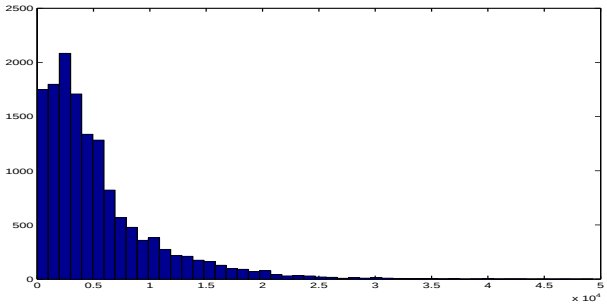


Figure 2: Histogram of the observed positive individual claim sizes.

For an increasing number of observed claims Table 2 shows that the average individual claim sizes decrease, indicating a negative correlation between claim size and the number of claims.

| | all observations | $N_i = 1$ | $N_i = 2$ | $N_i = 3$ | $N_i = 4$ |
|---|---|---|---|---|---|
| mean | 5371.0 | 5389.9 | 4403.8 | 3204.2 | 330.5 |

Table 2: Mean of the observed individual claim sizes taken over all observations and over observations with $N_i = k, k = 1, 2, 3, 4$ observed claims separately.

### 4.2 Modelling claim frequency

We first consider models for claim frequency. In order to identify significant covariates and interactions, the data set is first analysed in Splus using a Poisson model without spatial effects. The obtained covariates specification is then used as a starting specification for our MCMC algorithms. An intercept, seventeen covariates like age, gender of the policyholders or mileage and interactions were found to be significant for explaining claim frequency. However, for rea-

sons of confidentiality no details about these covariates and their effects will be reported. In order to obtain low correlations between covariates, we use centered and standardized covariates throughout the whole application. Since Germany is divided into 440 irregular spaced regions, 440 spatial effects are introduced for the MCMC analysis. We are interested in spatial effects after adjusting for population effects, therefore the population density in each region is included as covariate as well. In particular, the population density is considered on a logarithmic scale which turned out to give the best fit in the initial Splus analysis.

The MCMC algorithm for the spatial Poisson regression model introduced in Section 2.1 is run for 10000 iterations with starting values determined by the corresponding GLM without spatial effects using an iterative weighted least squares algorithm. A burnin of 1000 iterations is found to be sufficient after investigation of the MCMC trace plots. The trace plots for the regression parameters are presented in Figure 3. The estimated empirical autocorrelations for the

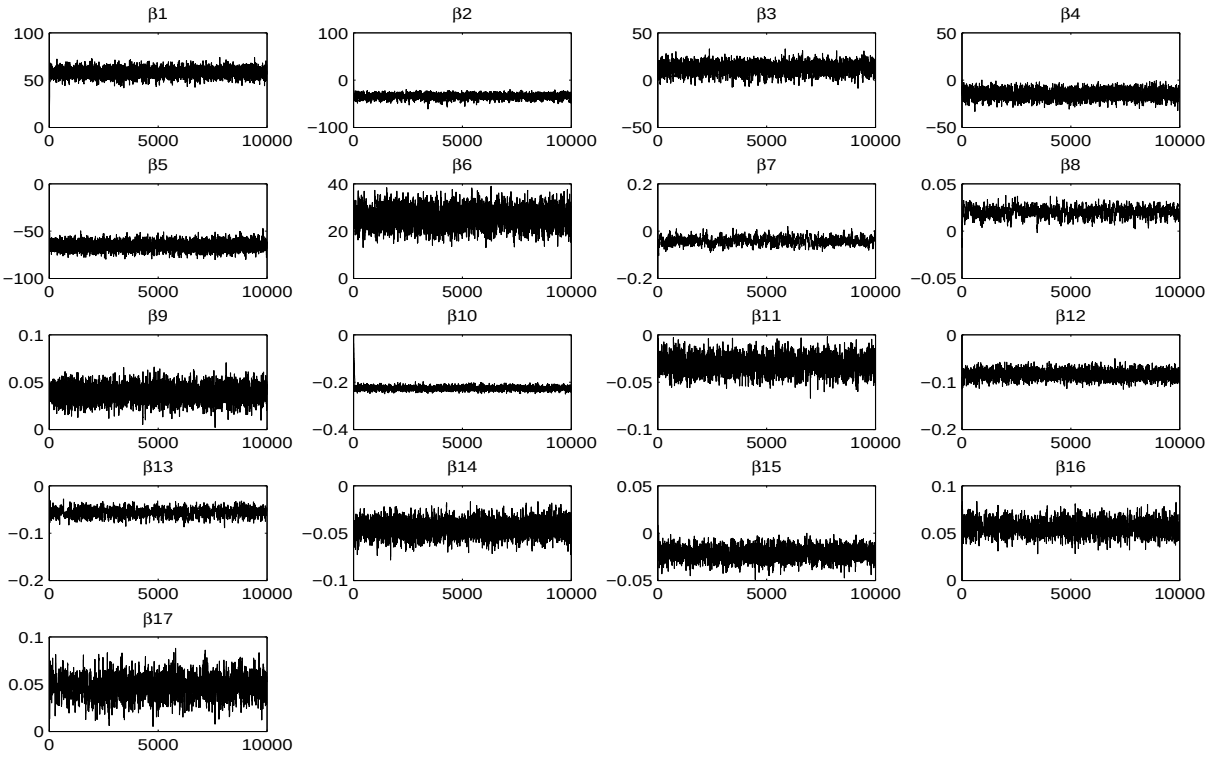


Figure 3: MCMC trace plots for the regression parameters $\beta_1, .., \beta_{17}$.

regression parameters and for 50 randomly chosen spatial effects centered around the intercept, plotted in Figure 4, decrease reasonably fast. We assume both a model including and without spatial effects, both containing the same covariates, and compare them using DIC and PMCC. Although the effective number of parameters $p_D$ increases from 16.0 to 98.2, the improvement in the goodness-of-fit expressed by the posterior mean of the deviance $E[D(\boldsymbol{\theta}|\mathbf{y})]$, leads to a lower value of the DIC when spatial effects are included (see Table 3). This shows that, after taking the information given by the covariates into account, there is still some unexplained spatial heterogeneity present in the data which is captured by the spatial effects. Although the inclusion of

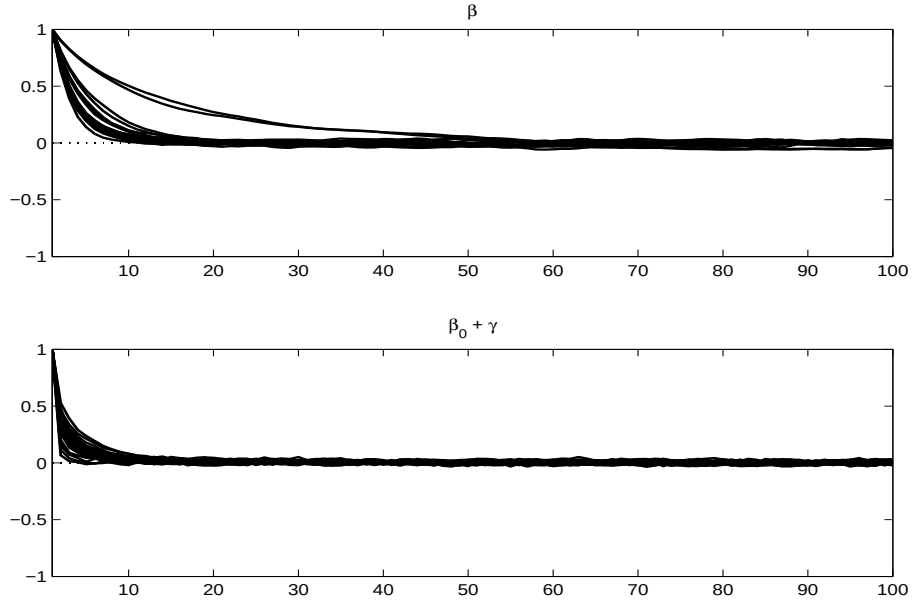


Figure 4: Estimated empirical autocorrelations for the regression parameters $\beta_1, .., \beta_{17}$ and the spatial effects centered around the intercept $\beta_0 + \gamma_i$ for 50 randomly chosen indices $i$.

the population density in each region allows for geographic differences already, spatial random effects still have a significant influence on explaining the expected number of claims. The PMCC given in Table 3 as well confirms these results.

| $\boldsymbol{\gamma}$ | DIC | $E[D(\boldsymbol{\theta}\|\mathbf{y})]$ | $p_D$ | PMCC | $\sum_{i=1}^{n}(\mu_i - y_i)^2$ | $\sum_{i=1}^{n}\sigma_i^2$ |
|---|---|---|---|---|---|---|
| no | 122372 | 122356 | 16.0 | 28624 | 14297 | 14328 |
| yes | 122143 | 122045 | 98.2 | 28613 | 14280 | 14334 |

Table 3: DIC, posterior mean of the deviance $D(\boldsymbol{\theta}|\mathbf{y})$, effective number of parameters $p_D$ and PMCC, split in its two components, for the Poisson models for claim frequency with and without spatial effects $\boldsymbol{\gamma}$.

The left panel in the top row in Figure 5 shows a map of the estimated posterior means of the spatial effects, ranging from -0.441 to 0.285. The corresponding posterior means of the risk factors $\exp(\gamma_i)$ for the minimum and maximum spatial effects are given by 0.65 and 1.34, respectively. A trend from the east to the west of Germany is visible for the spatial effects, the risk for claims tends to be lower in the east and increases towards the south western regions. A map of 80 % credible intervals for the spatial effects is given in the right panel. For the eastern and south western regions significant spatial effects are present, whereas the spatial effects for the regions in the middle of Germany do not significantly differ from zero.

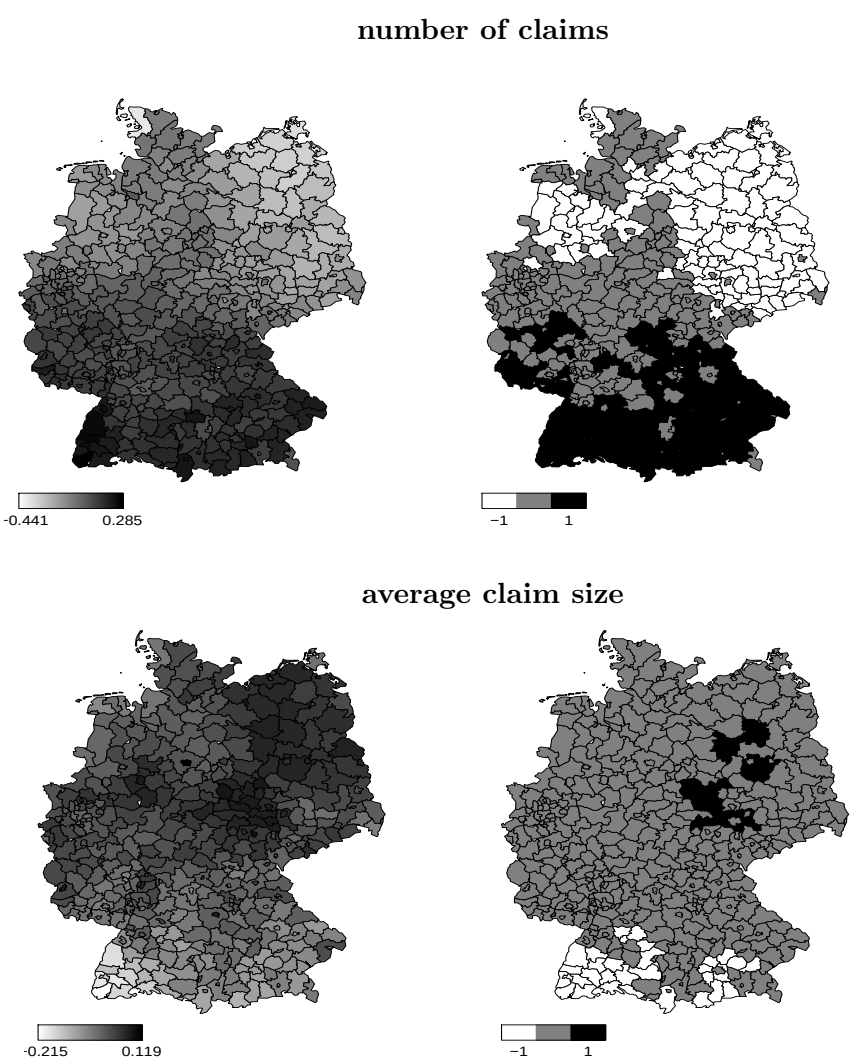


Figure 5: Map of the estimated posterior means (left) together with map of the 80 % credible intervals (right) for the spatial effects in the Poisson (top row) and average (bottom row) claim size regression model. For grey regions, zero is included in the credible interval, black regions indicate strictly positive, white regions strictly negative credible intervals.

### 4.3 Modelling the average claim size

In this section the average claim sizes $S_i := \sum_{k=1}^{N_i} S_{ik}$ are analysed using the spatial Gamma regression Model (2.3), i.e. $S_i|N_i \sim Gamma(\mu_i^S, N_i v)$ with mean specification (2.4). Considering only observations with a positive number of claims, altogether 14066 observations are obtained. Again, significant covariates and interactions are identified by analysing the data in Splus first, assuming a Gamma model without spatial effects. An intercept and fourteen covariates including gender, type and age of car as well as the population density in each region, modelled as polynomial of order four, have been found to have significant influence. Further, the observed number of claims $N_i$ is included as covariate. Since the highest number of claims is four, the

number of claims is treated as a factor with three levels where $N_i = 1$ is taken as reference level. These covariates will be taken into account when analysing the data set using MCMC. Therefore, including spatial and number of claims effects the mean $\mu_i^S$ is specified by

$$\mu_i^S = \exp(\mathbf{w}_i'\boldsymbol{\alpha} + \zeta_{j(i)}) = \exp(\sum_{l=0}^{14} w_{il}\alpha_l + \sum_{k=2}^{4} D_{ki}\alpha_{N_i=k} + \zeta_{j(i)})$$

where $D_{ki} = \begin{cases} 1, & N_i = k \\ 0, & \text{otherwise} \end{cases}$. In the following we consider both models with and without number of claims effects $\alpha_{N_i=k}, k = 2, 3, 4$ to quantify the influence of these effects. The MCMC sampler for Model (2.3) including and without spatial effects and the observed number of claims is run for 10000 iterations. Again a burnin of 1000 is found to be sufficiently large, plots of the trace plots and the empirical autocorrelations are omitted for brevity reasons. Models are compared using DIC, PMCC and some of the scoring rules given in Section 3. The lowest value of the DIC is obtained for the model both including $N_i$ and spatial effects (see Table 4). Although the increase of the estimated effective number of parameters is very small when

| model with $\alpha_{N_i=k}$ | model with $\boldsymbol{\zeta}$ | DIC | $E[D(\boldsymbol{\theta}\|\mathbf{y})]$ | $p_D$ |
|---|---|---|---|---|
| yes | yes | 269092 | 269020 | 72.7 |
| yes | no | 269136 | 269119 | 16.5 |
| no | yes | 269122 | 269048 | 73.9 |
| no | no | 269175 | 269159 | 15.9 |

| $\alpha_{N_i=k}$ | $\boldsymbol{\zeta}$ | LS | $IS_{\alpha=0.5}$ | $IS_{\alpha=0.1}$ | CRPS | PMCC | $\sum_{i=1}^n(\mu_i - y_i)^2$ | $\sum_{i=1}^n \sigma_i^2$ |
|---|---|---|---|---|---|---|---|---|
| yes | yes | -9.5642 | -55760 | -20412 | -2471.9 | $7.332 \cdot 10^{11}$ | $3.540 \cdot 10^{11}$ | $3.792 \cdot 10^{11}$ |
| yes | no | -9.5699 | -56125 | -20526 | -2481.8 | $7.363 \cdot 10^{11}$ | $3.575 \cdot 10^{11}$ | $3.788 \cdot 10^{11}$ |
| no | yes | -9.5669 | -55805 | -20434 | -2474.2 | $7.290 \cdot 10^{11}$ | $3.538 \cdot 10^{11}$ | $3.752 \cdot 10^{11}$ |
| no | no | -9.5734 | -56170 | -20575 | -2484.3 | $7.321 \cdot 10^{11}$ | $3.580 \cdot 10^{11}$ | $3.741 \cdot 10^{11}$ |

Table 4: DIC, posterior mean of the deviance $D(\boldsymbol{\theta}|\mathbf{y})$, effective number of parameters $p_D$ and mean score $S_n(\boldsymbol{\theta})$ for scoring rules LS, $IS_\alpha$ ($\alpha = 0.5, 0.1$), CRPS and PMCC, split in its two components, for the average claim size models including and without spatial and claim number effects.

the number of claims is included as covariate, the value of the posterior mean of the deviance decreases by 28 and 40 in the model with and without spatial effects respectively, indicating a significant improvement. The results for the scoring rules and the PMCC, divided into its two components are reported in Table 4 as well. The computation of DIC, PMCC and the scores is based on 5000 iterations of the MCMC output, the first 5000 iterations are neglected. Note, that the computation of the interval score $IS_\alpha$ and the continuous ranked probability score CRPS is based on simulated data, whereas the logarithmic score LS and the PMCC are calculated directly

using the MCMC output. For the logarithmic score LS, the interval score $IS_\alpha$ and the CRPS the highest mean score $S_n(\boldsymbol{\theta})$ is obtained for the model including spatial effects and number of claims effects. This confirms the significance of spatial and number of claims effects. According to the negatively oriented PMCC the spatial models also are to be preferred to the non-spatial ones. However, lower values of PMCC are obtained for the models without number of claims effects which is mainly caused by the second term of PMCC. Here it should be kept in mind, that PMCC is not a proper scoring rule as noted in Section 3.2.
A map of the posterior means and the 80 % credible intervals of the spatial effects for the model both including spatial effects and $N_i$ is given in the bottom row in Figure 5. Similar results are obtained for the model without $N_i$. The estimated posterior means of the risk factors $\exp(\gamma_i)$ for the minimum and maximum spatial effects range from 0.81 to 1.13. Contrary to the estimated spatial effects for claim frequency, the average claim size tends to be higher for some regions in the east of Germany, whereas for regions in the south western part lower claim sizes are to be expected. Again, according to the 80 % credible intervals, the spatial effects are only significant for some regions in the east and the south west of Germany.

| parameter | posterior mean of $\alpha_{N_i=k}$ | posterior mean of $\exp(\alpha_{N_i=k})$ |
|---|---|---|
| $\alpha_{N(i)=2}$ | -0.295 | 0.745 |
| | $(-0.382, -0.203)$ | $(0.683, 0.817)$ |
| $\alpha_{N(i)=3}$ | -1.951 | 0.146 |
| | $(-2.376, -1.462)$ | $(0.093, 0.232)$ |
| $\alpha_{N(i)=4}$ | -2.642 | 0.082 |
| | $(-3.473, -1.544)$ | $(0.031, 0.214)$ |

Table 5: Estimated posterior means of the number of claims effects and the risk factors $\exp(\alpha_{N(i)=k})$ in the Gamma model for average claim sizes including spatial effects, with the 95 % credible intervals given in brackets.

The estimated posterior means together with 95 % credible intervals of the number of claims effects $\alpha_{N_i=k}, k = 2, 3, 4$ are reported in Table 5. For an easier interpretation of the results we also give the estimated posterior means of the factors $\exp(\alpha_{N_i=k}), k = 2, 3, 4$, which quantify the relative risk in contrast to observations with the same covariates but only one observed claim. Compared to a policyholder with one observed claim, the expected average claim size for an observation with two observed claims decreases by about 25 %. If three or four claims have been reported, the expected average claim size even decreases by about 75 % and 92 %, respectively.

### 4.4 Posterior predictive distribution of the total claim size

The distribution of the total claim size is not available analytically, but can be determined numerically using recursion formulas going back to Panjer (1981) when independence of claim size and claim frequency is assumed and no regression is present. In contrast in our approach, the independence assumption is violated and spatial regression models are considered. However,

based on the MCMC output of the models for the number of claims and the average claim size the posterior predictive distribution of the total claim size can be approximated. For this independence of claim size and the number of claims is not required. In the following we describe how the total claim size $S_i^T = \sum_{k=1}^{N_i} S_{ik}$ for policyholder $i = 1, .., n$ can be simulated based on the MCMC output. Let $\hat{\boldsymbol{\beta}}^r$, $\hat{\boldsymbol{\gamma}}^r$,$\hat{\boldsymbol{\alpha}}^r$, $\hat{\boldsymbol{\zeta}}^r, r = 1, .., R$ denote the MCMC draws after burnin for the regression parameters and spatial effects of the claim frequency and average claim size model, respectively. The quantities $\hat{v}^r$ denote the MCMC draws of $v$ in the Gamma model for average claim sizes. Then, for $r = 1, .., R$, proceed as follows.

- simulate $N_i^r \sim Poisson(\hat{\mu_i}^{Nr})$ where $\hat{\mu_i}^{Nr} := t_i \exp(\mathbf{x}_i' \hat{\boldsymbol{\beta}}^r + \hat{\gamma}_{j(i)}^r)$
- if $N_i^r = 0$ set $S_i^{Tr} = 0$
- otherwise simulate:
  $S_i^r \sim Gamma(\hat{\mu}_i^{Sr}, \hat{v}^r N_i^r)$ where $\hat{\mu_i}^{Sr} := \exp(\mathbf{w}_i' \hat{\boldsymbol{\alpha}}^r + \hat{\zeta}_{j(i)}^r)$ and set $S_i^{Tr} := N_i^r \cdot S_i^r$

Thus, a sample $S_i^{Tr}, r = 1, .., R$ of the total claim size $S_i^T$ is obtained for which the posterior predictive distribution of $S_i^T$ can be approximated. Note that the Bayesian inference only provides information on number of claims effects $\alpha_{N_i=k}$ for number of claims up to $k = 4$, while the simulated number of claims $N_i^r$ are not restricted to take values less or equal to 4. However, since in our simulations only very rarely, if at all, a number of claims greater than 4 was simulated, this effect can be neglected. In order to compare the simulated total claim sizes $S_i^T$ based on the different models for claim size and claim frequency, we compute the continuous ranked probability score CRPS and the predictive model choice criterion PMCC. DIC and the logarithmic score cannot be computed here, since they require the explicit form of the total claim size distribution which is not available. The interval score will also be omitted out of the following reasons: Due to the large amount of observations with zero claims in our data set, the percentage of simulations with total claim size equal to zero is also very high. Zero will be included in the $(1-\alpha)$ 100 % posterior predictive intervals of the total claim sizes for $\alpha = 0.5, 0.1$ for almost all observations. Therefore only observations above the upper quantiles of the prediction intervals would be considered as outliers and be penalized. Hence, the use of the interval score will not be appropriate any more. Instead we consider one-sided quantities here like the quantiles $Q_\alpha$ at level $\alpha = 0.95, 0.99$ and the number of observations falling above these quantiles and compute the quantile score $QS_\alpha$ described in Section 3.2 for $\alpha = 0.95, 0.99$. Both the scores as well as the PMCC are computed using 5000 simulations of the total claim sizes $S_i^T$. The results are reported in Table 6.

The PMCC favours the simulations based on the models including spatial effects for the number of claims only, further better results are achieved when number of claims effects are taken into account. This is caused especially by the second term of the PMCC, representing the model variances, which are considerably lower when the number of claims is included as covariate in the claim size models.

The mean scores for the CRPS and the quantile scores $QS_\alpha, \alpha = 0.95, 0.99$, are very close for all

| freq $\gamma$ | size $\zeta$ | PMCC | $\sum_{i=1}^n (\mu_i - y_i)^2$ | $\sum_{i=1}^n \sigma_i^2$ |
|---|---|---|---|---|
| with $\alpha_{N_i=k}$ | | | | |
| yes | yes | $1.5757 \cdot 10^{12}$ | $7.8032 \cdot 10^{11}$ | $7.9540 \cdot 10^{11}$ |
| no | no | $1.5735 \cdot 10^{12}$ | $7.8069 \cdot 10^{11}$ | $7.9279 \cdot 10^{11}$ |
| yes | no | $1.5710 \cdot 10^{12}$ | $7.8046 \cdot 10^{11}$ | $7.9089 \cdot 10^{11}$ |
| no | yes | $1.5856 \cdot 10^{12}$ | $7.8088 \cdot 10^{11}$ | $8.0477 \cdot 10^{11}$ |
| without $\alpha_{N_i=k}$ | | | | |
| yes | yes | $1.5960 \cdot 10^{12}$ | $7.8055 \cdot 10^{11}$ | $8.1541 \cdot 10^{11}$ |
| no | no | $1.5909 \cdot 10^{12}$ | $7.8088 \cdot 10^{11}$ | $8.1000 \cdot 10^{11}$ |
| yes | no | $1.5894 \cdot 10^{12}$ | $7.8072 \cdot 10^{11}$ | $8.0871 \cdot 10^{11}$ |
| no | yes | $1.6052 \cdot 10^{12}$ | $7.8108 \cdot 10^{11}$ | $8.2414 \cdot 10^{11}$ |

| freq | size | | 95 % | | | 99 % | | |
|---|---|---|---|---|---|---|---|---|
| $\gamma$ | $\zeta$ | CRPS | quantile | outliers | $QS_{0.95}$ | quantile | outliers | $QS_{0.99}$ |
| with $\alpha_{N_i=k}$ | | | | | | | | |
| yes | yes | -212.26 | 476.3 | 3.60 % | -205.2 | 6400.6 | 1.15 % | -134.9 |
| no | no | -212.35 | 456.0 | 3.65 % | -205.7 | 6426.9 | 1.16 % | -135.5 |
| yes | no | -212.27 | 480.7 | 3.60 % | -205.3 | 6393.5 | 1.17 % | -135.4 |
| no | yes | -212.36 | 460.0 | 3.65 % | -205.8 | 6454.0 | 1.16 % | -135.5 |
| without $\alpha_{N_i=k}$ | | | | | | | | |
| yes | yes | -212.30 | 473.6 | 3.60 % | -205.2 | 6433.6 | 1.16 % | -135.3 |
| no | no | -212.30 | 453.8 | 3.65 % | -205.7 | 6455.1 | 1.16 % | -135.5 |
| yes | no | -212.28 | 477.8 | 3.60 % | -205.3 | 6422.6 | 1.17 % | -135.7 |
| no | yes | -212.34 | 456.2 | 3.65 % | -205.9 | 6482.1 | 1.16 % | -135.7 |

Table 6: In the upper table the PMCC, split in its two components is given for several models for the simulated total claim sizes $S_i^T$. In the lower table the mean score $S_n(\boldsymbol{\theta})$ for the CRPS, the average 95 % and 99 % quantiles given by $\frac{1}{n}\sum_{i=1}^n Q_{\alpha,i}$, the percentage of observations lying above these quantiles and the corresponding quantile mean scores $QS_\alpha, \alpha = 0.95, 0.99$, are given.

models, in general slightly higher scores are obtained for simulations based on a spatial Poisson model for the number of claims. Further, the simulations based on a spatial model for both claim frequency and claim size and including number of claims effects tend to achieve the highest score. The average size of the quantiles seems to be mainly determined by the inclusion or neglect of spatial effects in the Poisson model for the number of claims. The quantiles at level $\alpha = 0.95$ are higher when spatial effects are included for the number of claims, reflecting a higher model complexity. The percentage of observations above the 95 % quantile ranges from 3.60 % to 3.65 %, lying below the expected 5 %. This might be caused by the above noted fact. Since for some observations even the 95 % quantile will be zero, a zero observation will not be regarded as an outlier. This might be overcome by randomizing zero observations, i.e. considering zero

observations as outliers with a certain probability when the 95 % quantile takes the value zero. The 99% quantiles in contrast, are slightly higher when no spatial Poisson models are assumed, the percentage of outliers is close to the expected 1 %.

In Figure 6 map plots of the observed total claim sizes and the posterior predictive means $\frac{1}{R}\sum_{r=1}^{R} S_i^{Tr}$ of the simulated total claim sizes, averaged over each region, are given. Since we only consider the posterior predictive mean of the simulated total claim sizes, it is natural that the map displaying the true total claim sizes shows more extreme values. Hence, for a better visual comparison of the maps, we have built six classes for the total claim size in these plots, assuming equal length for the four middle classes, but summarizing extremely small or high values in broader classes. The simulations are based on the models for average claim sizes with the number of claims included as covariate. When spatial effects are included in the Poisson model (middle row), an increasing trend from the east to the west is observable for the simulated total claim sizes. The additional inclusion of spatial effects for the average claim size leads to small changes in the very eastern and south western parts of Germany. The rough spatial structure of the observed total claim sizes (top) is represented reasonable well for these two models. However, if spatial effects are only included for the average claim sizes, the regions with high observed total claim sizes in the middle and south western parts of Germany are not detected. The same holds for the simulations based on the models without any spatial effects. For regions in the east of Germany with rather low true total claim sizes for example, the mean of the total claim sizes is estimated up to 1.27 times as high when no spatial effects at all are taken into account compared to a spatial modelling of claim frequency and claim size. For one south western region with large observed total claims in contrast, the posterior mean of the simulated total claim size based on non spatial models is only estimated 0.69 times as large compared to the simulations based on spatial models for claim frequency and claim size.

The estimated probabilities for the total claim sizes being equal to zero as well as density estimates of the positive simulated total claim sizes of the policyholders in the two regions Hannover and Lörrbach are given in Figure 7. For Hannover which is located in the northern middle part of Germany the largest posterior mean of the spatial effect in the average claim size model was estimated ($\hat{\zeta} = 0.12$), while in Lörrbach which is situated in the south west of Germany the smallest effect $\hat{\zeta} = -0.22$ was observed. The observed total claim size, averaged over all policyholders in the region, is given by DM 335.0 in Hannover and DM 220.0 in Lörrbach, respectively. The estimated posterior means of the spatial effects in the Poisson model for the number of claims are given by $\hat{\gamma} = -0.10$ in Hannover and $\hat{\gamma} = 0.29$ in Lörrbach. Figure 7 shows that the estimated probability for zero total claim sizes and the density estimates of the positive total claim sizes notedly change when spatial effects are included in the models for claim frequency and average claim size. In Hannover, the inclusion of spatial effects in the models for claim frequency and the average claim size leads to a higher estimated probability of zero total claim sizes and heavier tails for the estimated density of the positive total claim sizes. The posterior predictive mean of the total claim sizes, averaged over all policyholders in Hannover, takes the value 254.3 when spatial effects are included which is closer to the observed

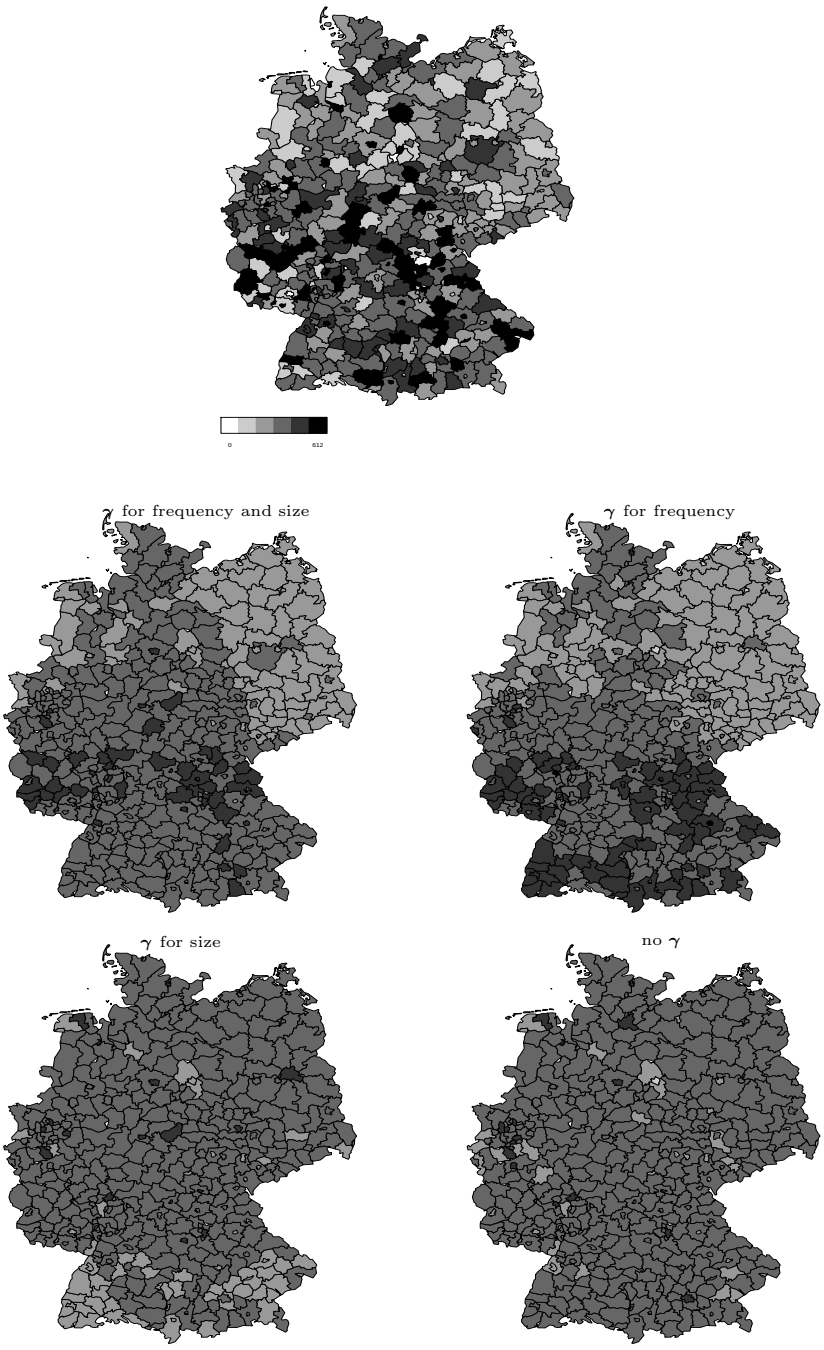


Figure 6: Observed total claim sizes (top) and posterior predictive means of the simulated total claim sizes based on Poisson and Gamma models for average claim sizes with and without spatial effects. Grey level classification: $[0, 100), [100, 150), [150, 200), [200, 250), [200, 300), [300, \infty)$

total claim size than without spatial effects where the posterior predictive mean takes the value 252.3. In Lörrbach in contrast, the estimated probability for zero total claim sizes decreases and more mass is given to small positive total claim sizes when spatial effects are added. Here again the posterior predictive mean of the total claim sizes, averaged over all policyholders in Lörrbach, is closer to the observed total claim size when spatial effects are taken into account (218.4) compared to simulations based on the non spatial models where the posterior predictive mean is estimated as 201.9.

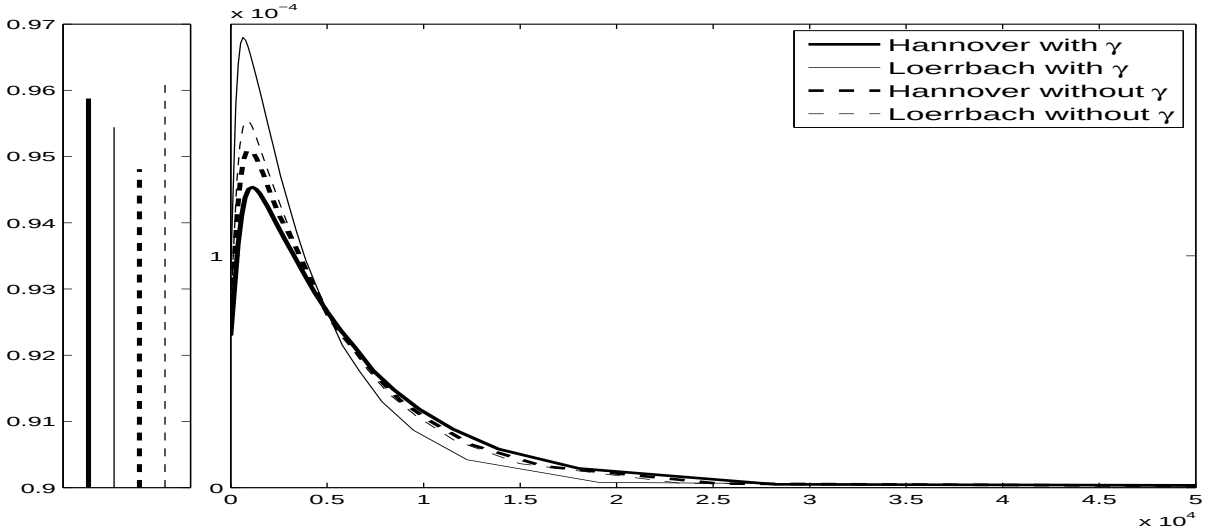


Figure 7: Estimated probabilities for zero total claim sizes (left panel) and density estimates of the positive total claim sizes (right panel) of the policyholders in the regions Hannover and Lörrbach based on spatial (solid lines) and non spatial (dashed lines) models for both claim frequency and average claim size.

Ideally, when the predictive quality of models is of interest, the data should not be used twice, i.e. parameter estimation should be based on part of the data only and predictions should be done for the remaining data. Since in this section model comparison rather than prediction was the focus, all data were used for parameter estimation and simulation of the total claim sizes. However, for the sake of completeness, we also fitted the Poisson models for claim frequency and the Gamma models for the average claim size based on 75 % of the data only. Then we simulated from the predictive distributions of the total claim sizes for the remaining 25 % of the data and assessed performance by comparing to the observed total claim sizes. The in Table 7 reported results for PMCC, CRPS and the quantile scores are qualitatively the same as observed before. The mean scores are very close for all models, the quantile scores give a slight preference to simulations based on a spatial Poisson model. Note, that the mean scores take lower values now compared to the simulations based on all data. Further, about 9 % of the observations exceed the 95 % quantile, about 2.9 % are above the 99 % quantile. This shows, that prediction of the true total claim sizes is worse here. However, this is to be expected, since the information given in these 25 % of the data has not been accounted for in parameter estimation.

| freq $\gamma$ | size $\zeta$ | PMCC | $\sum_{i=1}^{n}(\mu_i - y_i)^2$ | $\sum_{i=1}^{n}\sigma_i^2$ |
|---|---|---|---|---|
| with $\alpha_{N_i=k}$ | | | | |
| yes | yes | $4.0026 \cdot 10^{11}$ | $2.0460 \cdot 10^{11}$ | $1.9567 \cdot 10^{11}$ |
| no | no | $4.0013 \cdot 10^{11}$ | $2.0465 \cdot 10^{11}$ | $1.9548 \cdot 10^{11}$ |
| yes | no | $3.9932 \cdot 10^{11}$ | $2.0460 \cdot 10^{11}$ | $1.9470 \cdot 10^{11}$ |
| no | yes | $4.0193 \cdot 10^{11}$ | $2.0475 \cdot 10^{11}$ | $1.9718 \cdot 10^{11}$ |
| without $\alpha_{N_i=k}$ | | | | |
| yes | yes | $4.0472 \cdot 10^{11}$ | $2.0467 \cdot 10^{11}$ | $2.0005 \cdot 10^{11}$ |
| no | no | $4.0447 \cdot 10^{11}$ | $2.0468 \cdot 10^{11}$ | $1.9979 \cdot 10^{11}$ |
| yes | no | $4.0384 \cdot 10^{11}$ | $2.0467 \cdot 10^{11}$ | $1.9917 \cdot 10^{11}$ |
| no | yes | $4.0627 \cdot 10^{11}$ | $2.0482 \cdot 10^{11}$ | $2.0145 \cdot 10^{11}$ |

| freq | size | | 95 % | | | 99 % | | |
|---|---|---|---|---|---|---|---|---|
| $\gamma$ | $\zeta$ | CRPS | quantile | outliers | $QS_{0.95}$ | quantile | outliers | $QS_{0.99}$ |
| with $\alpha_{N_i=k}$ | | | | | | | | |
| yes | yes | -213.73 | 470.3 | 8.9 % | -207.2 | 6376.2 | 2.9 % | -139.4 |
| no | no | -213.75 | 451.4 | 9.0 % | -207.5 | 6411.0 | 2.9 % | -139.3 |
| yes | no | -213.76 | 472.3 | 8.9 % | -207.3 | 6370.3 | 2.9 % | -139.3 |
| no | yes | -213.82 | 452.0 | 9.0 % | -207.7 | 6247.5 | 2.9 % | -139.8 |
| without $\alpha_{N_i=k}$ | | | | | | | | |
| yes | yes | -213.77 | 467.2 | 8.9 % | -207.3 | 6412.0 | 3.0 % | -139.3 |
| no | no | -213.80 | 448.3 | 9.0 % | -207.4 | 6444.1 | 2.9 % | -139.4 |
| yes | no | -213.72 | 469.3 | 8.9 % | -207.4 | 6403.0 | 2.9 % | -139.5 |
| no | yes | -213.75 | 448.6 | 9.0 % | -207.8 | 6459.9 | 2.9 % | -140.6 |

Table 7: PMCC, split in its two components, mean score $S_n(\boldsymbol{\theta})$ for the CRPS, the average 95 % and 99 % quantiles given by $\frac{1}{n}\sum_{i=1}^{n} Q_{\alpha,i}$, the percentage of observations lying above these quantiles and the corresponding quantile mean scores $QS_\alpha, \alpha = 0.95, 0.99$, for several models for the simulated total claim sizes $S_i^T$. Parameter estimation is based on 75 % of the data, for remaining 25 % of the data total claim sizes are simulated from the posterior predictive distribution.

# 5 Summary and conclusions

We have presented a Bayesian approach for modelling claim frequency and average claim size taking both covariates and spatial effects into account. In contrast to the common approach where independence of the number of claims and claim size is assumed, we do not need this assumption. Instead, we have shown, that by including the observed number of claims as covariate for claim size, the models for claim size are significantly improved. If for example a policyholder caused two or three claims, the expected average claim size decreases by about 25 % and 75 %, respectively, compared to a policyholder with only one claim. Based on the models for claim

frequency and average claim sizes, we finally approximated the posterior predictive distribution of the total claim sizes using simulation.
For model selection we suggest the use of proper scoring rules for continuous variables based on the posterior predictive distribution. In particular, some of the presented scoring rules can also be estimated for the total claim size, although the distribution of the total claim size is not available analytically. According to the scoring rules and the PMCC, especially the inclusion of spatial effects in the model for claim frequency leads to improved predictions of the total claim sizes in our data set. However, the inclusion of number of claims effects in the claim size models for our data set hardly affects the total claim sizes according to the scoring rules. This can be explained by the fact, that very rarely more than one claim is simulated in accordance with the observed data and therefore number of claims effects have almost no impact. It is to be expected that for other data sets which are based for example on longer time horizons, thus having more observations with more than one claim, this modelling framework which allows for dependencies between number of claims and claim size will become very important for total claim size models. The number of claims might also be modelled using more flexible models allowing for overdispersion like for example the negative binomial distribution, the generalised Poisson distribution introduced by Consul and Jain (1973) or zero inflated models, see Gschlößl and Czado (2006) for more details. However, for the data set analysed in this paper, the Poisson distribution was sufficient.

## Acknowledgement

We would like to thank Tilmann Gneiting for fruitful discussions and helpful suggestions on the use of proper scoring rules. The first author is supported by a doctoral fellowship within the Graduiertenkolleg *Angewandte Algorithmische Mathematik*, while the second author is supported by Sonderforschungsbereich 386 *Statistische Analyse Diskreter Strukturen*, both sponsored by the *Deutsche Forschungsgemeinschaft*.

## References

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

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

Boskov, M. and R. Verrall (1994). Premium rating by geographic area using spatial models. *ASTIN Bullletin 24 (1)*, 131–143.

Consul, P. and G. Jain (1973). A generalization of the Poisson distribution. *Technometrics 15*, 791–799.

Czado, C. and S. Prokopenko (2004). Modeling transport mode decisions using hierarchical binary spatial regression models with cluster effects. *Discussion paper 406, SFB 386 Statistische Analyse diskreter Strukturen*. http://www.stat.uni-muenchen.de/sfb386/.

Dimakos, X. and A. Frigessi (2002). Bayesian premium rating with latent structure. *Scandinavian Actuarial Journal* (3), 162–184.

Gelfand, A. and S. Ghosh (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika 85 (1)*, 1–11.

Gilks, W., S. Richardson, and D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. Boca Raton: Chapman & Hall/CRC.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, to appear.

Gschlößl, S. (2006). *Hierarchical Bayesian spatial regression models with applications to non-life insurance*. Ph. D. thesis. Munich University of Technology.

Gschlößl, S. and C. Czado (2005). Does a Gibbs sampler approach to spatial Poisson regression models outperform a single site MH sampler? *submitted*.

Gschlößl, S. and C. Czado (2006). Modelling count data with overdispersion and spatial effects. *Statistical Papers*, Online first.

Haberman, S. and A. Renshaw (1996). Generalized linear models and actuarial science. *The Statistician 45*, 407–436.

Han, C. and B. Carlin (2001). Markov chain Monte Carlo methods for computing bayes factors: A comparative review. *Journal of the American Statistical Association 96*, 1122–1132.

Jørgensen, B. and M. C. P. de Souza (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal 1*, 69–93.

Kass, R. and A. Raftery (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association 90*, 773–795.

Laud, P. and J. Ibrahim (1995). Predictive model selection. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 57*(1), 247–262.

Lundberg, F. (1903). Approximerad framställning av sannolikhetsfunktionen. aterförsäkring av kollektivrisker. Akad. afhandling. almqvist och wicksell, uppsala, Almqvist och Wicksell, Uppsala.

Mikosch, T. (2004). *Non-Life Insurance Mathematics. An Introduction with Stochastic Processes*. New York: Springer.

Panjer, H. (1981). Recursive evaluation of a family of compound distributions. *ASTIN Bulletin 11*, 22–26.

Panjer, H. and G. Willmot (1983). Compound Poisson models in actuarial risk theory. *Journal of Econometrics 23*, 63–76.

Pettitt, A., I. Weir, and A. Hart (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing 12 (4)*, 353–367.

Renshaw, A. (1994). Modelling the claims process in the presence of covariates. *ASTIN Bulletin 24*, 265–285.

Smyth, G. K. and B. Jørgensen (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin 32 (1)*, 143–157.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *J. R. Statist. Soc. B 64 (4)*, 583–640.

Sun, D., R. K. Tsutakawa, and P.L.Speckman (1999). Posterior distribution of hierarchical models using CAR(1) distributions. *Biometrika 86*, 341–350.

Székely, G. (2003). $\epsilon$-statistics: The energy of statistical samples. Technical report no. 2003-16, Department of Mathematics and Statistics, Bowling Green State University, Ohio.

Taylor, G. (1989). Use of spline functions for premium rating by geographic area. *ASTIN Bullletin 19 (1)*, 89–122.

van der Linde, A. (2005). DIC in variable selection. *Statistica Neerlandica 59 (1)*, 45–56.

Address for correspondence:
Claudia Czado, Center of Mathematical Sciences, Munich University of Technology, Boltzmannstr.3, D-85747 Garching, Germany.
E-mail: cczado@ma.tum.de