Statistical Papers manuscript No.
(will be inserted by the editor)

Non nested model selection for spatial count regression models with application to health insurance

Claudia Czado · **Holger Schabenberger** ·
Vinzenz Erhardt

Received: date / Accepted: date

Abstract In this paper we consider spatial regression models for count data. We examine not only the Poisson distribution but also the generalized Poisson capable of modeling over-dispersion, the negative Binomial as well as the zero-inflated Poisson distribution which allows for excess zeros as possible response distribution. We add random spatial effects for modeling spatial dependency and develop and implement MCMC algorithms in *R* for Bayesian estimation. The corresponding R library 'spatcounts' is available on CRAN. In an application the presented models are used to analyze the number of benefits received per patient in a German private health insurance company. Since the deviance information criterion (DIC) is only appropriate for exponential family models, we use in addition the Vuong and Clarke test with a Schwarz correction to compare possibly non nested models. We illustrate how they can be used in a Bayesian context.

Keywords Spatial count regression · over-dispersion · zero-inflation · generalized Poisson · non nested comparison

---

C. Czado
Technische Universität München, Zentrum Mathematik
Lehrstuhl für Mathematische Statistik
Boltzmannstr. 3
85748 Garching
Tel.: +49-89-28917428
Fax: +49-89-28917435
E-mail: cczado@ma.tum.de

H. Schabenberger
Technische Universität München

V. Erhardt (*corresponding author*)
Technische Universität München
E-mail: erhardt@ma.tum.de

1 Introduction

We speak of count data when the data values are contained in the natural numbers. A common model for count data is the Poisson (Poi) model, which is rather restrictive since for this distribution variance and mean are equal. But often in observed count data the sample variance is considerably larger than the sample mean - a phenomenon called over-dispersion. In such cases the Poisson assumption is not appropriate for analyzing this data.

Frequently the negative Binomial (NB) distribution instead of the Poisson distribution is used to model over-dispersed data. Another possibility for modeling over-dispersion is the generalized Poisson (GP) distribution introduced by Consul and Jain (1973) which allows for a more flexible variance function than the Poisson distribution by an additional parameter (see e.g. Consul and Famoye (1992) and Famoye (1993)).

Over-dispersion may also be caused by a large proportion of zero counts in the data. Yip and Yau (2005) stress that especially claim numbers often exhibit a large number of zeros and hence traditional distributions may be insufficient. In addition to the zeros arising from the count data model, zero-inflated models (see for example Winkelmann (2008)) also allow for excess zeros. Zero-inflated models can be used in combination with any count data distribution. We consider in this paper the zero-inflated Poisson (ZIP) (see e.g. Lambert (1992)) and the zero-inflated generalized Poisson (ZIGP) model. ZIGP models have been investigated by Famoye and Singh (2003), Gupta et al. (2004), Bae et al. (2005), Joe and Zhu (2005) and Famoye and Singh (2006).

The variability in over-dispersed data can also be interpreted as unobserved heterogeneity which is not sufficiently explained by the covariates. Especially for simple models with few parameters, theoretical model predictions may not match empirical observations for higher moments. When information on the location of the individuals is known, the data is spatially indexed. For count regression models, Gschlößl and Czado (2007) include spatial random effects using a proper conditional autoregressive (CAR) model based on Pettitt et al. (2002). In other words, one assumes random effects associated with geographic areas rather than individuals and presumes that the effects in neighboring regions are similar. In contrast to Gschlößl and Czado (2007), however, we also include covariates with spatial information, e.g. measures for the degree of urbanity at a certain location. We carry out a comparison investigating whether one of these two spatial specifications or both fit our data better.

Altogether, in this paper we account for extra variability not only by addressing distributions capable of handling over-dispersion and over-dispersion caused by an excessive number of zeros, we also take extra spatial variability in the data into account.

Since in these spatial models maximum likelihood estimation and confidence interval estimation is not tractable we consider the models in a Bayesian context. Thus, for parameter estimation Markov Chain Monte Carlo (MCMC) methods are used.

Model comparison between different model classes is non standard. For nested models, i.e. when one of the two models is a super model of the other, model comparison may be carried out using tools like Akaike's information criterion or likelihood ratio tests. This condition may be violated when the distribution on which the two models are based, are different. Even within such a class of regression models, two models may be non nested when they use different link functions or when linear predictors are non hierarchical. We utilize a test proposed by Vuong (1989) and the distribution-free test proposed by Clarke (2007) for non nested model comparison and illustrate how they may be applied in a Bayesian context.

This is a novel approach since so far these two tests have only been used in classical estimation. Also, the comparison between spatial covariate and / or spatial effect specifications for count regression data has not been carried out elsewhere.

In our application we consider health insurance policies in the following context: for more than 35000 policyholders, the data contain the number of benefits received by the patients in the ambulant (i.e. outpatient) setting as well as several covariates like the total of all deductibles, age, gender, number of physicians per inhabitants, number of inhabitants per square kilometer and buying power. Further, we quantify the best fitted model according to DIC as well as Vuong and Clarke test.

This paper proceeds as follows. In Section 2 an overview on spatial count regression models as well as the modeling of spatial effects is given, where we introduce a proper Gaussian conditional autoregressive prior based on Pettitt et al. (2002). The necessary background to Bayesian inference and MCMC methods is briefly summarized in Section 3. This includes the deviance information criterion of Spiegelhalter et al. (2002) as a model selection criterion. The test proposed by Vuong (1989) and the distribution-free test utilized in a Bayesian framework are presented in Section 4. An application to private health insurance data for policyholders in Germany is presented in Section 5.

2 Spatial count regression models

2.1 Spatial effects

2.1.1 Spatial covariates

Spatial variation may sometimes be explained by covariates which vary spatially. Such covariates we call 'spatial covariates'. Examples in our data set are the number of physicians per inhabitant in a certain district, the number of inhabitants per square kilometer or the buying power per district.

2.1.2 CAR

In order to account for spatial heterogeneity we will incorporate, in addition to covariate information, spatial random effects in the regression models. Therefore we consider the Gaussian Conditional Autoregressive (CAR) formulation introduced by Pettitt et al. (2002) which permits the modeling of spatial dependence and dependence between multivariate random variables at irregularly spaced regions. Assume that J regions $\{1,\dots,J\}$ are given and let $\boldsymbol{\gamma} = (\gamma_1,\dots,\gamma_J)^t$ the vector of spatial effects for each region. Let $\boldsymbol{\gamma}$ be multivariate normal distributed with

$$\boldsymbol{\gamma} \sim N_J(0, \sigma^2 Q^{-1}) \tag{1}$$

where the precision matrix $Q = (Q_{ij})_{i,j=1,\ldots,J}$ is given by

$$Q_{ij} = \begin{cases} 1 + |\psi| \cdot N_i & i = j \\ -\psi & i \sim j \\ 0 & otherwise \end{cases} . \tag{2}$$

Here the notation $i \sim j$ indicates that the regions i and j are neighbors and N_i denotes the number of neighbors of region i. Thus the full conditional distribution of γ_i given all the other values $\boldsymbol{\gamma}_{-i}$, $i = 1, \ldots, J$ is

$$\gamma_i | \boldsymbol{\gamma}_{-i} \sim N\left(\frac{\psi}{1 + |\psi| \cdot N_i} \sum_{j \sim i} \gamma_j,\ \sigma^2 \frac{1}{1 + |\psi| \cdot N_i} \right). \tag{3}$$

Parameter ψ determines the overall degree of spatial dependence. If all regions are spatially independent, i.e. $\psi = 0$, the precision matrix Q (see (2)) reduces to the identity matrix, whereas for $\psi \to \infty$ the degree of dependence increases. The multivariate normal distribution (1) is a proper distribution since Pettitt et al. (2002) show that the precision matrix Q is symmetric and positive definite. Another convenient feature of this CAR model is that according to Pettitt et al. (2002) the determinant of Q, which is needed for the update of ψ in a MCMC algorithm, can be computed efficiently.

2.2 Count regression models

The count distributions considered in this paper will be the Poisson (Poi), the negative Binomial (NB), the generalized Poisson (GP), the zero-inflated Poisson (ZIP) and the zero-inflated generalized Poisson (ZIGP) distribution. In order to allow for a comparison between these distributions, we choose a mean parameterization for all of them. Their probability mass functions (pmf) together with means and variances are given in Table 1. Regression models for these considered distributions can be constructed similar to generalized linear models (GLM) (McCullagh and Nelder (1989)). We denote the regression model with response Y_i and (known) explanatory variables $x_i = (1, x_{i1}, \ldots, x_{ip})^t$ for the mean $i = 1, \ldots, n$. For individual observation periods, we allow exposure variables t_i, which satisfy $t_i > 0\ \forall i$ and in case without individual exposure $t_i = 1\ \forall i$.

1. **Random component:**
   $\{Y_i, 1 \leq i \leq n\}$ are independent with response distribution $Poi(\mu_i)$, $NB(\mu_i, r)$, $GP(\mu_i, \varphi)$, $ZIP(\mu_i, \omega)$ or $ZIGP(\mu_i, \varphi, \omega)$.
2 **Systematic component:**
   The linear predictor is $\eta_i^{\mu}(\boldsymbol{\beta}) = \boldsymbol{x}_i^t \boldsymbol{\beta} + \gamma_i$ which influence the response Y_i. Here, $\boldsymbol{\beta} = (\boldsymbol{\beta}^{NS}, \boldsymbol{\beta}^{S})$ are the unknown regression parameters with $\boldsymbol{\beta}^{NS} = (\beta_0, \beta_1, \ldots, \beta_r)^t$ the nonspatial explanatory factors, $\boldsymbol{\beta}^{S} = (\beta_{r+1}, \beta_{r+2}, \ldots, \beta_p)^t$ the spatial covariates and γ_i the spatial random effects (not included in our base models). The matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^t$ is called design matrix.

Table 1: Pmf's of the Poisson, NB, GP and ZIP distribution together with their means and variances in mean parameterization

	$P(Y=y)$	$E(Y)$	$Var(Y)$	Parameter restriction
$Poi(\mu)$	$\frac{\exp\{-\mu\}\mu^y}{y!}$	μ	μ	$\mu \in \mathbb{R}$
$NB(\mu,r)$	$\frac{\Gamma(y+r)}{\Gamma(r)y!}\left(\frac{r}{\mu+r}\right)^r\left(\frac{\mu}{\mu+r}\right)^y$	μ	$\mu(1+\frac{\mu}{r})$	$r>0$
$GP(\mu,\varphi)$	$\frac{\mu(\mu+(\varphi-1)y)^{y-1}}{y!}\varphi^{-y}e^{-\frac{1}{\varphi}(\mu+(\varphi-1)y)}$	μ	$\varphi^2\mu$	$\varphi>0$
$ZIP(\mu,\omega)$	$\omega\cdot\mathbb{1}_{\{y=0\}}+(1-\omega)\cdot\frac{\exp(-\mu)\mu^y}{y!}$	$(1-\omega)\mu$	$(1-\omega)\mu(1+\omega\mu)$	$\omega\in(0,1)$
$ZIGP(\mu,\varphi,\omega)$	$\omega\cdot\mathbb{1}_{\{y=0\}}+(1-\omega)\cdot$ $\cdot\frac{\mu(\mu+(\varphi-1)y)^{y-1}}{y!}\varphi^{-y}e^{-\frac{1}{\varphi}(\mu+(\varphi-1)y)}$	$(1-\omega)\mu$	$(1-\omega)\mu(\varphi^2+\omega\mu)$	$\varphi>0,\omega\in(0,1)$

3 **Parametric link component:**
To get a positive mean the linear predictor $\eta_i^{\mu}(\boldsymbol{\beta})$ is related to the parameters $\mu_i(\boldsymbol{\beta})$, $i = 1, \ldots, n$ as follows:

$$E(Y_i|\boldsymbol{\beta}) = \mu_i(\boldsymbol{\beta}) := t_i \exp\left\{\boldsymbol{x}_i^t \beta + \gamma_i\right\} = \exp\left\{\boldsymbol{x}_i^t \beta + \gamma_i + \log(t_i)\right\}$$
$$\Leftrightarrow \eta_i^{\mu}(\boldsymbol{\beta}) = \log(\mu_i(\boldsymbol{\beta})) - \log(t_i) \quad \text{(log - link)}$$

3 MCMC including model selection

In order to incorporate spatial random effects we consider the models in a Bayesian context which allows the modeling of a spatial dependency pattern. The determination of the posterior distributions require high dimensional integrations. MCMC will be used for parameter estimation, in particular we use the Metropolis Hastings sampler introduced by Metropolis et al. (1953) and Hastings (1970). For more information on Bayesian data analysis and MCMC methods see Gilks et al. (1996) and Gelman et al. (2003). Throughout this paper, an independence MH sampler using the Student's t-distribution with $\nu = 20$ degrees of freedom will be used. For details on the MCMC algorithms see Gschlößl and Czado (2008) and Schabenberger (2009b).

The DIC (Spiegelhalter et al. (2002)) is a popular information criterion which was designed to compare hierarchical models, and can easily be computed using the available MCMC output. Let $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^T$ be a sample from the posterior distribution of the model. The calculation of the DIC is based on two quantities. On one hand this is the so called *unstandardized deviance* $D(\boldsymbol{\theta}) = -2\log(p(\boldsymbol{y}|\boldsymbol{\theta}))$ where $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the observation model and on the other hand the so called effective number of parameters p_D defined by

$$p_D := \overline{D(\boldsymbol{\theta}|\boldsymbol{y})} - D(\bar{\boldsymbol{\theta}}).$$

Here $\overline{D(\boldsymbol{\theta}|\boldsymbol{y})} := \frac{1}{T}\sum_{t=1}^{T} D\left(\boldsymbol{\theta}^t\right)$ is the estimated posterior mean of the deviance and $D(\bar{\boldsymbol{\theta}})$ is the deviance of the estimated posterior means $\bar{\boldsymbol{\theta}} := \frac{1}{T}\sum_{t=1}^{T} D\left(\boldsymbol{\theta}^t\right)$. Finally the DIC determined as

$$\text{DIC} = \overline{D(\boldsymbol{\theta}|\boldsymbol{y})} + p_D = 2\overline{D(\boldsymbol{\theta}|\boldsymbol{y})} - D(\bar{\boldsymbol{\theta}}).$$

The preferred model is the one which has the smallest DIC. DIC depends on the specific values obtained in an MCMC run, thus it is difficult to assess how different DIC values have to be for different models to select among these models. For exponential family models DIC approximates the Akaike information criterion (AIC).

4 Non nested model selection

We use tests proposed by Vuong (1989) and Clarke (2003) to compare regression models which need not to be nested. These tests are based on the Kullback-Leibler information criterion (KLIC). According to Vuong (1989) the Kullback-Leibler distance is defined as

$$KLIC := E_0[\log h_0(Y_i|\boldsymbol{x}_i)] - E_0[\log f(Y_i|\boldsymbol{\upsilon}_i, \hat{\boldsymbol{\delta}})],$$

where $h_0(\cdot|\cdot)$ is the true conditional density of Y_i given $\boldsymbol{x}_i$, that is, the true but unknown model. Let E_0 denote the expectation under the true model, $\boldsymbol{\upsilon}_i$ are the covariates of the estimated model and $\hat{\boldsymbol{\delta}}$ are the pseudo-true values of $\boldsymbol{\delta}$ in model with $f(Y_i|\boldsymbol{\upsilon}_i,\boldsymbol{\delta})$, which is not the true model. Generally, the model with minimal *KLIC* is the one that is closest to the true, but unknown, specification.

4.1 Vuong test

Consider two models, $f_1 = f_1(Y_i|\boldsymbol{\upsilon}_i,\hat{\delta}^1)$ and $f_2 = f_2(Y_i|\boldsymbol{\omega}_i,\hat{\delta}^2)$ then if model 1 is closer to the true specification, we have

$$\begin{aligned} E_0[\log h_0(Y_i|\boldsymbol{x}_i)] - E_0[\log f_1(Y_i|\boldsymbol{\upsilon}_i,\hat{\boldsymbol{\delta}}^1)] &< E_0[\log h_0(Y_i|\boldsymbol{x}_i)] - E_0[\log f_2(Y_i|\boldsymbol{\omega}_i,\hat{\boldsymbol{\delta}}^2)] \\ \Leftrightarrow E_0\left[\log \frac{f_1(Y_i|\boldsymbol{\upsilon}_i,\hat{\boldsymbol{\delta}}^1)}{f_2(Y_i|\boldsymbol{\omega}_i,\hat{\boldsymbol{\delta}}^2)}\right] &> 0 \end{aligned} \tag{4}$$

Vuong defines the statistics

$$m_i := \log\left(\frac{f_1(y_i|\boldsymbol{\upsilon}_i,\hat{\boldsymbol{\delta}}^1)}{f_2(y_i|\boldsymbol{\omega}_i,\hat{\boldsymbol{\delta}}^2)}\right), \quad i = 1,\dots,n. \tag{5}$$

If h_0 is the true probability mass function, then $\mathbf{m} = (m_1,\dots,m_n)^t$ is a random vector with mean $\boldsymbol{\mu}_0^m = (\mu_1^m,\dots,\mu_n^m) := E_0(\mathbf{m})$. Hence, we can test the null hypothesis

$$H_0 : \boldsymbol{\mu}_0^m = \mathbf{0} \text{ against } H_1 : \boldsymbol{\mu}_0^m \neq \mathbf{0}.$$

The mean $\boldsymbol{\mu}_0^m$ in the above hypothesis is unknown. With convenient standardization and the central limit theorem Vuong (1989) shows that under H_0

$$\nu := \frac{\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^n m_i\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^n (m_i - \bar{m})^2}} \xrightarrow{\mathscr{D}} \mathscr{N}(0,1), \quad \text{as } n \to \infty$$

where $\bar{m} := \frac{1}{n}\sum_{i=1}^n m_i$. This allows to construct an asymptotic α-level test of $H_0 : \boldsymbol{\mu}_0^m = \mathbf{0}$ versus $H_1 :$ not H_0. It rejects H_0 if and only if $|\nu| \geq z_{1-\frac{\alpha}{2}}$, where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. The test chooses model 1 over 2, if $\nu \geq z_{1-\frac{\alpha}{2}}$. This is reasonable since according to the equivalence given in (4), significantly high values of ν indicate a higher *KLIC* of model 1 as compared to model 2. Similarly, model 2 is chosen if $\nu \leq -z_{1-\frac{\alpha}{2}}$. No model is preferred for $-z_{1-\frac{\alpha}{2}} < \nu < z_{1-\frac{\alpha}{2}}$. According to Clarke (2007, p. 349) the Vuong test must be corrected if the number of estimated coefficients in each model is different. Vuong (1989) suggests to use the Schwarz correction, which is given by

$$\left[\left(\frac{p}{2}\log n\right) - \left(\frac{q}{2}\log n\right)\right]. \tag{6}$$

Here p and q are the number of estimated coefficients in models f_1 and f_2, respectively (Clarke (2003, p. 78)). Thus the Vuong test statistic v with Schwarz correction is defined as:

$$\tilde{v} := \frac{\sqrt{n}\left(\left[\frac{1}{n}\sum_{i=1}^{n} m_i\right] - \left[\left(\frac{p}{2}\log n\right) - \left(\frac{q}{2}\log n\right)\right]/n\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2}}.$$

4.2 Clarke test

An alternative to the Vuong test is a distribution-free test (see Clarke (2007)) which applies a modified paired sign test to the differences in the individual log-likelihoods from two non nested models. The null hypothesis of the distribution-free test is

$$H_0 : P_0\left[\log \frac{f_1(Y_i|\boldsymbol{\upsilon}_i, \hat{\boldsymbol{\delta}}^1)}{f_2(Y_i|\boldsymbol{\omega}_i, \hat{\boldsymbol{\delta}}^2)} > 0\right] = 0.5. \tag{7}$$

Under the null hypothesis (7) the log-likelihood ratios should be symmetrically distributed around zero. That means that about half the log-likelihood ratios should be greater and half less than zero. Using m_i as defined in (5), Clarke considers the test statistic

$$B = \sum_{i=1}^{n} \mathbb{1}_{\{0,+\infty\}}(m_i), \tag{8}$$

where $\mathbb{1}_A$ is the indicator function which is 1 on the set A and 0 elsewhere. The quantity B is the number of positive differences and follows a Binomial distribution with parameters n and probability 0.5 under H_0. If B is, under the null hypothesis, significantly larger than its expected value, model f_1 is ”better” than model f_2. This allows to construct the following distribution-free test.

First let $m_i(Y_i)$ correspond to the random variable with value m_i, then the null hypothesis (7) is equivalent to

$$H_0^{DF} : P_0\left[m_i(Y_i) > 0\right] = 0.5 \quad \forall i = 1, \ldots, n.$$

For the test problem $H_0^{DF} : P_0\left[m_i(Y_i) > 0\right] = 0.5 \quad \forall i = 1, \ldots, n$ versus $H_{1+}^{DF} : P_0\left[m_i(Y_i) > 0\right] > 0.5,\ i = 1, \ldots, n$, the corresponding α - level upper tail test rejects H_0^{DF} versus H_{1+}^{DF} if and only if $B \geq c_{\alpha+}$, where $c_{\alpha+}$ is the smallest integer such that $\sum_{c=c_{\alpha+}}^{n} \binom{n}{c} 0.5^n \leq \alpha$. If the upper tail test rejects H_0^{DF} then we decide that model 1 is preferred over model 2. For the alternative $H_{1-}^{DF} : P_0\left[m_i(Y_i) > 0\right] < 0.5,\ i = 1, \ldots, n$, the α - level lower tail test rejects H_0^{DF} versus H_{1-}^{DF} if and only if $B \leq c_{\alpha-}$, where $c_{\alpha-}$ is the largest integer such that $\sum_{c=0}^{c_{\alpha-}} \binom{n}{c} 0.5^n \leq \alpha$ (compare to Clarke (2007, p. 349)). If H_0^{DF} versus H_{1-}^{DF} is rejected, then model 2 is preferred over model 1. If H_0^{DF} cannot be rejected, no model is preferred.

Like the Vuong test this test is sensitive to the number of estimated coefficients in each model. Once again, we need a correction for the degrees of freedom.

Since the distribution-free tests work with the individual log-likelihood ratios, we cannot apply the Schwarz correction as in the Vuong test with the ”summed” log-likelihood ratio. Clarke (2003) suggests to apply the average correction to the individual log-likelihood ratios. So we correct the individual log-likelihoods for model f_1 by a factor of $\left[\left(\frac{p}{2n}\log n\right)\right]$ and the individual log-likelihoods for model f_2 by a factor of $\left[\left(\frac{q}{2n}\log n\right)\right]$.

In the Bayesian approach we can quantify the uncertainty of the test decisions for the Vuong and Clarke test accordingly. For this we utilize the sampled parameter values from the MCMC output and determine the test decision for each sampled value. This allows to estimate the posterior percentages of how many times model 1 (model 2) was chosen over model 2 (model 1) and the percentage of no test decision.

All MCMC algorithms for model fit and the model comparison are implemented in package `spatcounts` (Schabenberger (2009a)) in *R*, which is available on CRAN.

5 Application

We now apply the models described in Section 3 to a large portfolio of a German health insurer. Before the parametric models are fitted, a basic exploratory analysis is carried out. At the end of this Section, all fitted models are compared using the DIC as well as the Vuong and the Clarke tests described in Section 4.

5.1 Data description and exploration

The data set considers 37751 insured persons of a private health insurance company in 2007. The response variable is the number of benefits received per patient for ambulant treatments. In the German private health care system, the policyholders may opt to cover a part of each invoice themselves, this amount is called deductible. Depending on the policy type and the treatment setting, deductibles can be either an annual total or a percentage of each invoice. If no bill is reimbursed throughout the whole year, the policyholder receives a refund. A variable description including the response variable and the explanatory variables is given in Table 2. Germany has 439 districts. The data includes patients from all districts.

Around 76% of the insured persons are male, which is typical for the policy line considered. To obtain a first overview of the dependent variable Y_i, a histogram of the observed count frequencies is given in Figure 1). For a better graphical illustration, outliers $Y_i > 50$ are not displayed. The histogram shows that we have a high variation in Y_i and a rather large number of zeros. In particular 43% of the response data is equal to zero. The covariates can be split up into two groups. The first group of the covariates depend on the patient like the total of all deductibles with values **DED** $\in [0, 1821]$, the age with **AGE** $\in [3, 88]$ or the gender dummy **SEX**. The second group of covariates are spatial covariates like the number of physicians per inhabitant with **PHYS.INH** $\in [0, 0.5622]$, the number of inhabitants per square kilometer with value **URBAN** $\in [39.28, 4060]$ or the average buying power **BP** $\in [12277.4, 23760.38]$ in Euros. The maps in Figure 2 show the spatial distribution of the spatial covariates. The number of physicians per inhabitant in Germany seem to be distributed very uniformly (left panel of Figure 2) whereas the most inhabitants per square kilometer

Table 2: Variable description for the analyzed health insurance data set

Variable	Type	Description
Y_i	discrete	Number of outpatient benefits received by patient i.
DED_i	continuous	Total of all deductibles of patient i.
AGE_i	discrete	Age of patient i.
SEX_i	binary	Indicator for gender of patient i. (0 = female, 1 = male)
ZIP_i	categorical	ZIP Code of the home address of patient i.
$D(i)$	categorical	Indicates the home district for patient i.
$PHYS.INH_j$	continuous with spatial information	Number of physicians per inhabitant in district j multiplied by 100.
$URBAN_j$	continuous with spatial information	Number of inhabitants per square kilometer in district j.
BP_j	continuous with spatial information	Buying power in district j.

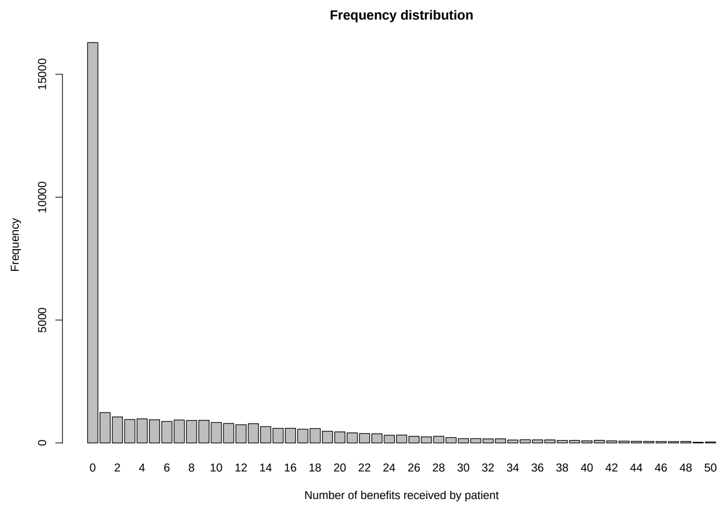

Fig. 1: Frequency distribution for the response variables ($\boldsymbol{Y} \in [0, 705]$ without outliers $Y_i > 50$).

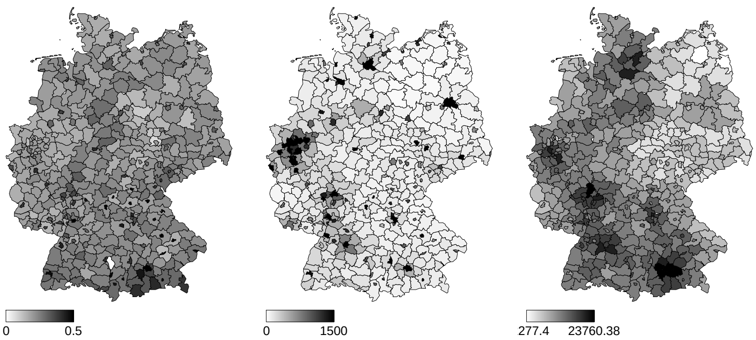

Fig. 2: Exploratory maps of the spatial covariates **PHYS.INH** (left panel), **URBAN** (middle panel) and **BP** (right panel).

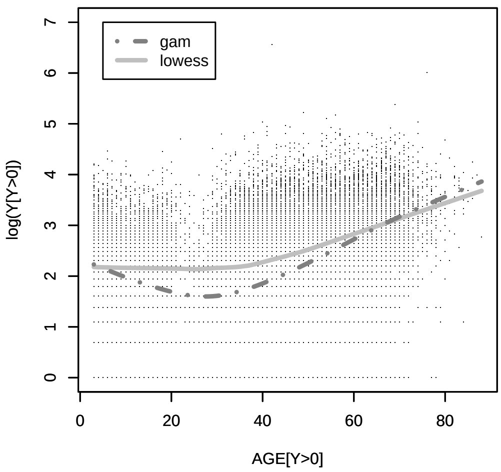

Fig. 3: Scatter plot (including gam (dashed) and lowess (solid) smoothing lines) and box plot of the number of benefits received per patient against age of the patient.

can be found in the larger cities, e.g. Berlin, Bremen, Hamburg, Munich or the Ruhr area (middle panel of Figure 2). West Germany has higher buying power with a peak around Munich compared to East Germany (former German Democratic Republic) (see right panel of Figure 2).

A natural next step is to look at scatter plots of the dependent variable $\boldsymbol{Y}$ against each of the regressors. The LOWESS (solid line) and the GAM (dashed line) smoothing curves of the scatter plot in Figure 3 indicates that the variable **AGE** has to be transformed, i.e. we allow a quadratic influence on the response. In health insurance this is not unusual since in general children and older people need more medical attendance. For numerical stability we use standardized (sometimes called autoscaled) covariates for the variables **DED**, **PHYS.INH**, **URBAN** and **BP** denoted with ".s".

5.2 Identification of base models

To establish base models we first analyze the data set in the statistical program ”R” without spatial effects. We allow for an intercept, the covariates gender (**SEX**), the standardized covariates **DED.s**, **PHYS.INH.s**, **URBAN.s** and **BP.s** as well as the orthogonal polynomial transformed covariates **AGE.p1** (polynomial transformation of degree 1), **AGE.p2** (polynomial transformation of degree 2). For maximum likelihood parameter estimates we use the function `est.zigp()` in the `R` package `ZIGP` developed by Erhardt (2009) for all models except the negative Binomial regression model which is estimated with the basic `R` library `MASS` using the function `glm.nb()`.

In a next step sequential elimination according to a Wald test with 5% α-level of significance is conducted. In Table 3 the full and reduced regression specifications are given for every model class considered in Section 2. The penalty term in the AIC statistic includes parameters which are estimated (such as φ in $GP(\mu_i,\varphi)$) and does not include them if they are fix (such as $\varphi = 1$ in $Poi(\mu_i)$). We stress that the comparison of different models based on AIC is only possible within one model class, that is when the distribution of the responses are the same and designs are hierarchical. If the models are non nested, the test decisions should be based on the Vuong test or the distribution-free test (Clarke test).

Table 4 displays for the models NB, GP, ZIP and ZIGP (defined as model (**I**) and Poi, NB, GP and ZIP (defined as model **II**) the entries of the Vuong and Clarke tests for each combination of model (**I**) and model (**II**). We choose an α-level of 5%, i.e. $z_{1-\frac{\alpha}{2}} = 1.96$. In the first line of each cell, the Vuong test statistic v is given. In the second and third line the decision of the Vuong test (V) and the Clarke test (C) is shown, i.e. if model (**I**) or (**II**) is better. The corresponding p-values for each test are given in parentheses. For example V: (I) $\left(< 2 \cdot 10^{-16}\right)$ means that the Vuong test prefers model (**I**) with p-value smaller than $2 \cdot 10^{-16}$. We now discuss the conclusions to be drawn from Table 4. Since the Poisson model is not preferred over any of the other model classes, we see evidence that the data is in fact overdispersed. Overdispersion may be explained either by a dispersion parameter as in the GP or the NB model, by excess zeros as in the ZIP model, or both. Since the GP model outperforms the NB model, we consider zero-inflation jointly with the GP distribution, i.e. we also fit a ZIGP model. In general, the tests by Vuong and Clarke are suitable for pairwise model comparison, thus they do not have to lead to an overall decision between all model classes, much less do both test necessarily decide equivalently. In our case, however, the pairwise decisions given in Table 4 are identical, and we can sort the models in a unique ranking: the GP model outperforms all other models and is followed downward by ZIGP, NB, ZIP and the Poisson model. The comparison of the $ZIGP(\mu_i,\varphi,\omega)$ model to all other model classes gives almost identical results as the comparison of the $GP(\mu_i,\varphi)$ model to these classes. The reason is that the zero-inflation parameter in the ZIGP model is estimated almost to zero (see Table 3) and therefore the ZIGP fit is almost identical to the GP fit. In the comparison between the GP and ZIGP model, the GP model by far outperforms the ZIGP model. This can be explained by the nature of the two test: even if the likelihood contributions per observation in both of these models are almost identical, there is a minimal correction toward the GP model by virtue of the larger Schwarz penalty term, which corrects for

Table 3: Model specifications and AIC for each of the models after sequential elimination of insignificant covariates according to a Wald test with $\alpha = 5\%$

Model	Model equation μ	Dispersion (SE)	Zero-inflation (SE)	$l(\hat{\boldsymbol{\theta}})$	Parameters	AIC
$Poi(\mu_i)$	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + PHYS.INH.s + URBAN.s + BP.s**	$\varphi = 1$ (not estimated)	$\omega = 0$ (not estimated)	-218 486.8	8	436 990
$NB(\mu_i, r)$ (full)	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + PHYS.INH.s + URBAN.s + BP.s**	$\hat{r} = 0.5811$ (0.0062)		-99 552.8	9	199 124
$NB(\mu_i, r)$ (reduced)	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + BP.s**	$\hat{r} = 0.5811$ (0.0062)		-99 553.3	7	199 121
$GP(\mu_i, \varphi)$ (full)	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + PHYS.INH.s + URBAN.s + BP.s**	$\hat{\varphi} = 4.6369$ (0.0397)	$\omega = 0$ (not estimated)	-96 849.1	9	193 716
$GP(\mu_i, \varphi)$ (reduced)	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + URBAN.s + BP.s**	$\hat{\varphi} = 4.6893$ (0.0410)	$\omega = 0$ (not estimated)	-96 850.5	8	193 717
$ZIP(\mu_i, \omega)$ (full)	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + PHYS.INH.s + URBAN.s + BP.s**	$\varphi = 1$ (not estimated)	$\hat{\omega} = 0.4312$ (0.0026)	-161 674.1	9	323 366
$ZIP(\mu_i, \omega)$ (reduced)	**1 + DED.s + AGE.p1 + AGE.p2 + SEX**	$\varphi = 1$ (not estimated)	$\hat{\omega} = 0.4312$ (0.0026)	-161 675.4	6	323 363
$ZIGP(\mu_i, \phi, \omega)$	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + PHYS.INH.s + URBAN.s + BP.s**	$\hat{\varphi} = 4.7010$ (0.0414)	$\hat{\omega} = 10^{-6}$ (0.0007)	-96 454.5	10	192 929

Table 4: Model comparison using the Vuong and the Distribution-Free (Clarke) test; test statistic v of the Vuong test together with decision according to Vuong (V) and Clarke (C) and their p-values, respectively.

(I) \ **(II)**	$Poi(\mu_i)$	$NB(\mu_i, r)$	$GP(\mu_i, \varphi)$	$ZIP(\mu_i, \omega)$
$NB(\mu_i, r)$	$v = 30.2$ V: (I) $\left(< 2 \cdot 10^{-16}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$			
$GP(\mu_i, \varphi)$	$v = 34.7$ V: (I) $\left(< 2 \cdot 10^{-16}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$	$v = 4.2$ V: (I) $\left(2.26 \cdot 10^{-5}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$		
$ZIP(\mu_i, \omega)$	$v = 21.4$ V: (I) $\left(< 2 \cdot 10^{-16}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$	$v = -24.7$ V: (II) $\left(< 2 \cdot 10^{-16}\right)$ C: (II) $\left(< 2 \cdot 10^{-16}\right)$	$v = -25.0$ V: (II) $\left(< 2 \cdot 10^{-16}\right)$ C: (II) $\left(< 2 \cdot 10^{-16}\right)$	
$ZIGP(\mu_i, \varphi, \omega)$	$v = 34.7$ V: (I) $\left(< 2 \cdot 10^{-16}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$	$v = 4.3$ V: (I) $\left(2.14 \cdot 10^{-5}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$	$v = -137$ V: (II) $\left(< 2 \cdot 10^{-16}\right)$ C: (II) $\left(< 2 \cdot 10^{-16}\right)$	$v = 25.0$ V: (I) $\left(< 2 \cdot 10^{-16}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$

the additional zero-inflation parameter in the ZIGP model. Notwithstanding this application, overdispersion explained by both a dispersion parameter and zero-inflation simultaneously is present in many other applications, e.g. the ZIGP model considered by Czado et al. (2007) to analyze patent filing processes.

By including a random spatial effect for each region extra heterogeneity in the data might be taken into account by assuming a finer geographic resolution. The CAR prior presented in Section 2 will be assumed for these spatial effects.

5.3 Bayesian inference using MCMC

The MCMC algorithms for the Poisson, NB, GP, ZIP and ZIGP regression models are run for 50000 iterations. The mean parameter μ_i, $i = 1, \ldots, n$ has the general form

$$\mu_i = t_i \cdot \exp\left(\boldsymbol{x}_i^t \boldsymbol{\beta} + \gamma_{D_{(i)}}\right)$$

with the observation specific exposure t_i fixed to 1. We fit models with spatial covariates only (denoted by SC), models with spatial random effects only (denoted by CAR) and models with both spatial random effects and spatial covariates (denoted by CAR+SC). Recall that we have the spatial covariates: number of physicians per inhabitants (**PHYS.INH.s**), number of inhabitants per square kilometer (**URBAN.s**) and buying power (**BP.s**).

The starting values for each parameter of the four models are taken from the regression without spatial effect. That means we use the results of the `R` functions `est.zigp()` and `glm.nb()` for all models with all covariates for SC and CAR+SC and without the spatial covariates for the CAR model. The posterior means and 80% credible intervals for the model specific parameters r, φ and ω in the different models are shown in Table 5 (the posterior means and 80% credible intervals for the regression parameter vector $\boldsymbol{\beta}$ can be found in Schabenberger (2009b, p. 59)). As in the base models in Section 5.2, the zero-inflation parameter in the ZIGP model is very close to zero for the SC, CAR and SC+CAR specifications. Note that only positive zero-inflation is allowed, therefore the credible intervals cannot contain the zero. Since the ZIGP model becomes a GP model when there is no zero-inflation present, we will no longer consider the ZIGP model for the remainder of this paper.

Estimation of the regression parameter slightly differs between the models and also changes when spatial effects are added, especially for the GP models where large spatial effects are observed. Although there are some insignificant covariates we do not reduce the models to compare whether SC, CAR or CAR+SC is preferred. Estimation of the specific parameters is rather similar in all models SC, CAR and CAR+SC. The range of the estimated spatial effects in all of the models is roughly the same in each model even though the Poisson model captures unexplained heterogeneity only by spatial effects. In the ZIP model the proportion of extras zeros ω is estimated as 43%.

In Figure 4 we present map plots of the estimated posterior means. In Figure 5 the 80% credible intervals of the spatial effects in the Poisson, negative Binomial, generalized Poisson and zero-inflated Poisson models are given. In each Figure the

Table 5: Estimated posterior means and 80% credible intervals for the model specific parameters in the considered SC, CAR, CAR+SC models

Parameter	Model	Mean	(10%, 90%)
r **in NB**	SC	0.5808	(0.5723, 0.5887)
	CAR	0.5912	(0.5831, 0.5995)
	CAR+SC	0.5910	(0.5830, 0.5993)
φ **in GP**	SC	4.6840	(4.6271, 4.7412)
	CAR	4.4492	(4.3994, 4.4999)
	CAR+SC	4.4488	(4.3985, 4.4994)
ω **in ZIP**	SC	0.4312	(0.4278, 0.4346)
	CAR	0.4310	(0.4276, 0.4345)
	CAR+SC	0.4310	(0.4276, 0.4344)
φ **in ZIGP**	SC	4.6825	$(4.6544, 4.7110)$
	CAR	4.4514	$(4.4219, 4.4805)$
	CAR+SC	4.4518	$(4.4214, 4.4792)$
ω **in ZIGP**	SC	$2.4 \cdot 10^{-4}$	$(2.0 \cdot 10^{-5}, 5.5 \cdot 10^{-4})$
	CAR	$1.6 \cdot 10^{-4}$	$(1.9 \cdot 10^{-5}, 4.0 \cdot 10^{-4})$
	CAR+SC	$1.5 \cdot 10^{-4}$	$(1.3 \cdot 10^{-5}, 3.4 \cdot 10^{-4})$

model specification SC is shown in the left panel, the CAR model specification in the middle panel and the CAR+SC model specification in the right panel. Here we see that the spatial effects of all four regression models are almost the same. The spatial covariates have nearly no influence but according to the 80% credible interval they have a negative spatial effect. According to the 80% credible intervals the CAR and the CAR+SC models have small significant spatial effect.

Unfortunately, the estimated empirical autocorrelations in some of the models decrease very slow. Therefore to compare the different models we decide to thin the 50000 MCMC output by choosing every 200th value.

In order to compare these models, the DIC, defined in Section 3, is considered. In Table 6, the DIC, the posterior mean of the deviance and the effective number of parameters are listed for each model. $E[D(\boldsymbol{\theta}|\boldsymbol{y})]$ is based only on the unscaled deviance (see Section 3) which cannot be interpreted directly as an overall goodness of fit measure of one specific model. However, $E[D(\boldsymbol{\theta}|\boldsymbol{y})]$ can be used for comparing the model fit of several models when the number of parameters is roughly the same.

For each regression model the model SC has the highest DIC value. The DIC for the CAR and CAR+SC model is roughly the same. For SC models the effective number of parameters p_D is close to the true number, which is eight for the Poisson

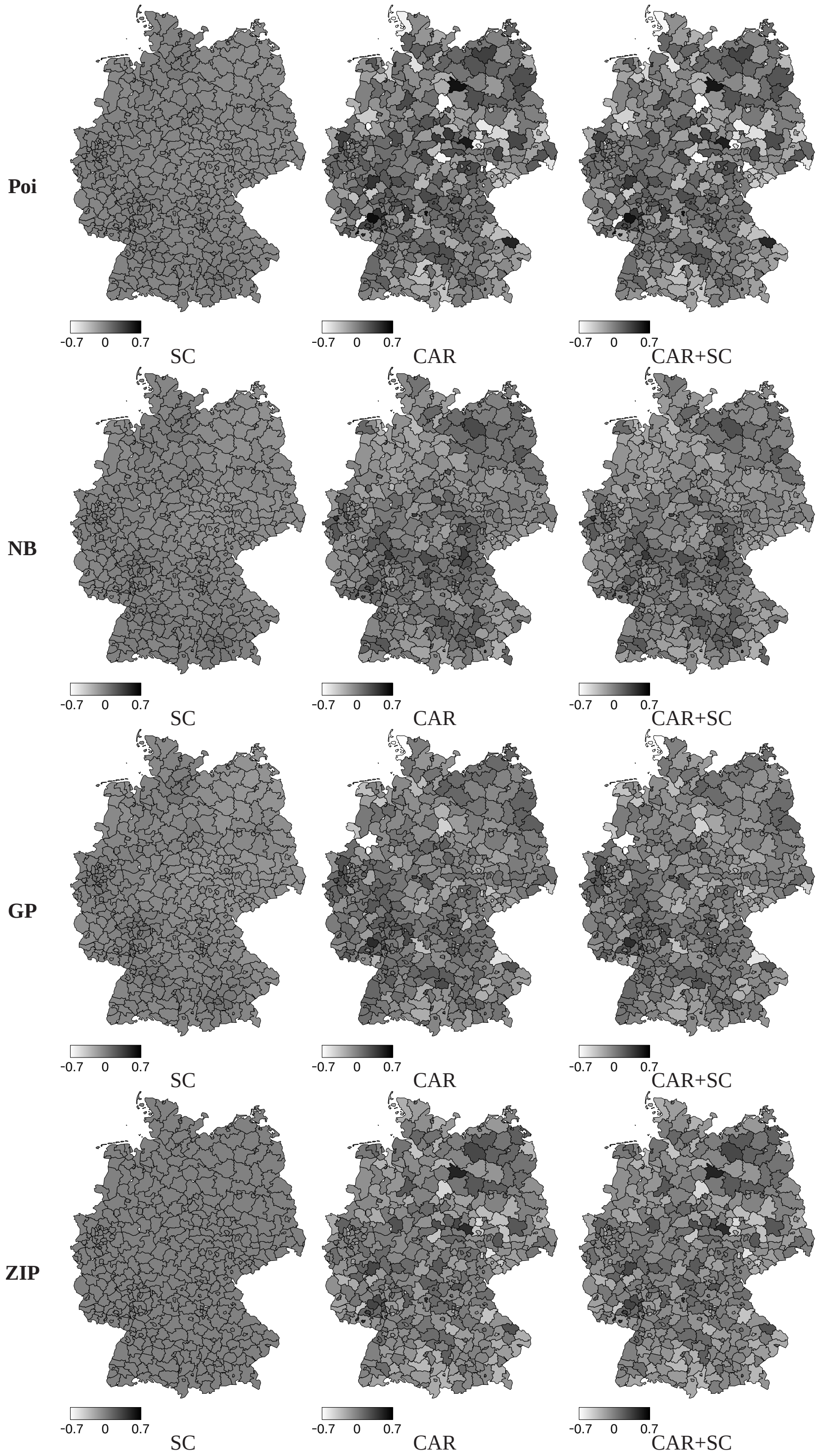

Fig. 4: Maps of the estimated posterior means (top panels) of the spatial effects in the Poi, NB, GP and ZIP regression models SC, CAR and CAR+SC

Fig. 5: Maps of the 80% credible intervals (white: strictly positive, black: strictly negative, gray: zero is contained in 80% credible interval) of the spatial effects in the Poi, NB, GP and ZIP regression models SC, CAR and CAR+SC

Table 6: DIC, $E\left[D(\boldsymbol{\theta}|\boldsymbol{y})\right]$ and effective number of parameters p_D for the different models

	Model	DIC	$E\left[D(\boldsymbol{\theta}\|\boldsymbol{y})\right]$	p_D
Poi	SC	436990.8	436982.9	7.89
	CAR	426818.5	426390.3	428.21
	CAR+SC	426817.2	426388.9	428.24
NB	SC	199124.9	199115.8	9.10
	CAR	198867.8	198651.8	216.00
	CAR+SC	198868.1	198650.8	217.29
GP	SC	192927.9	192918.5	9.33
	CAR	190764.4	190461.8	302.64
	CAR+SC	190764.7	190461.4	303.30
ZIP	SC	323367.0	323357.7	9.36
	CAR	318740.5	318364.8	375.66
	CAR+SC	318742.2	318366.2	376.03

regression model and nine for the NB, GP and ZIP regression model. This is to be expected, since these models do not include random effects. When spatial effects are added, the number of effective parameters increases rapidly. The DIC and the posterior mean of the deviance, $E\left[D(\boldsymbol{\theta}|\boldsymbol{y})\right]$, for CAR are the smallest in all regression models except for the Poisson model. Here the DIC value of CAR+SC is slightly lower than the one of CAR.

Note that the DIC must be used with care, since strictly speaking the DIC is defined for distributions of the exponential family only. Additionally, if two models have similar DIC values it is possible that the model decision varies for different MCMC runs. Therefore we make another comparison using the Vuong and the Clarke test discussed in Section 4.

5.4 Model selection

5.4.1 Selecting spatial models

First of all we compare SC, CAR and CAR+SC for each regression model $Poi(\mu_i)$, $NB(\mu_i, r)$, $GP(\mu_i, \varphi)$ and $ZIP(\mu_i, \omega)$. Table 7 shows the percentage of 250 Vuong and Clarke test decisions between model (I) and model (II). For the Vuong test we use the statistic ν and choose again an α-level of 5%, i.e. the decision border is $z_{1-\frac{\alpha}{2}} = 1.96$ For the Clarke test we report B/n. The number of parameters p and q of model **(I)**

Table 7: Decision of the Vuong and Clarke tests between model (I) and model (II) as a percentage

	Model (**I**)/ (**II**)	Test	Decision f. model (**I**)	No decision	Decision f. model (**II**)
Poi	SC/CAR+SC	Vuong	0.0%	6.0%	94.0%
		Clarke	0.0%	1.2%	98.8%
	CAR/CAR+SC	Vuong	0.0%	100.0%	0.0%
		Clarke	40.8%	19.2%	40.0%
	CAR/SC	Vuong	93.6%	6.4%	0.0%
		Clarke	98.8%	1.2%	0.0%
NB	SC/CAR+SC	Vuong	100.0%	0.0%	0.0%
		Clarke	100.0%	0.0%	0.0%
	CAR/CAR+SC	Vuong	1.2%	98.8%	0.0%
		Clarke	47.6%	4.0%	48.4%
	CAR/SC	Vuong	0.0%	0.0%	100.0%
		Clarke	0.0%	0.0%	100.0%
GP	SC/CAR+SC	Vuong	0.0%	100.0%	0.0%
		Clarke	14.4%	35.2%	50.4%
	CAR/CAR+SC	Vuong	0.4%	99.6%	0.0%
		Clarke	44.0%	18.4%	37.6%
	CAR/SC	Vuong	0.0%	100.0%	0.0%
		Clarke	55.2%	30.0%	14.8%
ZIP	SC/CAR+SC	Vuong	0.0%	100.0%	0.0%
		Clarke	100.0%	0.0%	0.0%
	CAR/CAR+SC	Vuong	0.0%	100.0%	0.0%
		Clarke	51.6%	0.0%	48.4%
	CAR/SC	Vuong	0.0%	100.0%	0.0%
		Clarke	0.0%	0.0%	100.0%

and (**II**), neccessary for the corrections, are taken from the DIC calculations, i.e. we use the effective number of parameters p_D.

The decisions of the Vuong and Clarke tests given in Table 7 are not consistent. For the Poisson regression models the SC specification performs poorly, however for the comparison between the CAR and CAR+SC specifications only the Clarke test slightly prefers CAR. Since this model has less covariates than CAR+SC, we choose this design as the preferred one within the Poisson class. For the negative Binomial model there is no distinct decision between CAR and CAR+SC, however the SC model is preferred over both of them. The same holds for the ZIP class. For the generalized Poisson regression models the test by Vuong prefers none of the models in all three comparisons. Therefore we only consider the Clarke test, which slightly decides toward the CAR model. Since this model also has the smallest DIC value (see Table 6), we choose the CAR specification within the GP class.

5.4.2 Selecting count distribution

Now we want to compare the observed preferred models Poi CAR, NB SC, GP CAR and ZIP SC to get the overall favored one. Therefore we use again the Vuong and the Clarke test like in the section before. The results are shown in Table 8. The first value in the round brackets favors model (**I**), the second one stands for no decision taken and the right one prefers model (**II**) (all in percent). For example (100%, 0%, 0%) means the test perfers model (**I**) over model (**II**) in 100% of the sampled MCMC posterior parameter values based on 250 iterations. The generalized Poisson regression model CAR seems to fit our data in terms of the Vuong test and the Clarke test the best. This model is preferred over all other models discussed (see Table 8).

Table 8: Selection of the response distributions ((I)$>$(II),(I)$=$(II),(I)$<$(II)) based on the Vuong (V) and Clarke (C) tests

(I) \ **(II)**		**Poi CAR**	**NB SC**	**GP CAR**
NB SC	V	(100%, 0%, 0%)		
	C	(100%, 0%, 0%)		
GP CAR	V	(100%, 0%, 0%)	(100%, 0%, 0%)	
	C	(100%, 0%, 0%)	(100%, 0%, 0%)	
ZIP SC	V	(100%, 0%, 0%)	(0%, 0%, 100%)	(0%, 0%, 100%)
	C	(100%, 0%, 0%)	(0%, 0%, 100%)	(0%, 0%, 100%)

6 Conclusions

For count regression data we have presented several models. In order to model overdispersion we used models with an additional parameter as in the NB and GP model or models with an extra proportion of zero observations like the zero-inflated model ZIP.

Further, in order to account for unobserved spatial heterogeneity in the data we included spatial random effects which allow for spatial correlations between observations and / or spatially varying covariates.

These models were applied to analyze the number of ambulant benefits received per patient in 2007. The DIC, the Vuong and the Clarke tests were used for model comparison. Models allowing for over-dispersion showed a significantly better fit than an ordinary non spatial Poisson regression model. For the NB and the ZIP model the inclusion of spatial effects did not improve the model fit. For the Poisson model which does not allow for over-dispersion, and the GP model, the inclusion of spatial effects led to an improved model fit. According to the considered criteria the GP regression model with spatial random CAR effects but no spatial covariates is to be preferred to all other models. However, the fitted spatial model shows no smooth surface structure. Rather it indicates isolated specific regions where the covariates provide no adequate fit.

There are several interesting avenues for further research. For instance, instead of analyzing the number of ambulant benefits received by patient for one year only, it might be interesting to include data over several years in order to examine whether the spatial pattern changes over the years. Another interesting possibility is to extend the regression models by allowing for regression on φ and ω in order to find a better model fit and to address heterogeneity on a more differentiated basis.

Acknowledgements:
C. Czado is supported by DFG (German Science Foundation) grant CZ 86/1-3. V. Erhardt is supported by a stipend.

References

Bae, S., F. Famoye, J. T. Wulu, A. A. Bartolucci, and K. P. Singh (2005). A rich family of generalized Poisson regression models. *Math. Comput. Simulation 69*(1-2), 4–11.

Clarke, K. A. (2003). Nonparametric model discrimination in international relations. *Journal of Conflict Resolution 47*, 72–93.

Clarke, K. A. (2007). A simple distribution-free test for nonnested model selection. *Political Analysis 15*(3), 347–363.

Consul, P. C. and F. Famoye (1992). Generalized poisson regression model. *Communications in Statistics - Theory and Methods 21*(1), 89–109.

Consul, P. C. and G. C. Jain (1973). A Generalization of the Poisson Distribution. *Technometrics 15*(4), 791–799.

Czado, C., V. Erhardt, A. Min, and S. Wagner (2007). Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modeling 7*(2), 125–153.

Erhardt, V. (2009). ZIGP: Zero-Inflated Generalized Poisson (ZIGP) Models. R package version 3.5.

Famoye, F. (1993). Restricted generalized poisson regression model. *Communications in Statistics - Theory and Methods 22*(5), 1335–1354.

Famoye, F. and K. P. Singh (2003). On inflated generalized Poisson regression models. *Adv. Appl. Stat. 3*(2), 145–158.

Famoye, F. and K. P. Singh (2006). Zero-inflated generalized Poisson model with an application to domestic violence data. *Journal of Data Science 4*(1), 117–130.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC.

Gilks, W., R. S., and S. D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

Gschlößl, S. and C. Czado (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal 2007*(3), 202–225.

Gschlößl, S. and C. Czado (2008). Modelling count data with overdispersion and spatial effects. *Statistical Papers 49*(3), 531–552.

Gupta, P. L., R. C. Gupta, and R. C. Tripathi (2004). Score test for zero inflated generalized Poisson regression model. *Comm. Statist. Theory Methods 33*(1), 47–64.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

Joe, H. and R. Zhu (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biom. J. 47*(2), 219–229.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics 34*(1), 1–14.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *The Journal of Chemical Physics 21*, 1087–1091.

Pettitt, A. N., I. S. Weir, and A. G. 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.

Schabenberger, H. (2009a). spatcounts: Spatial count regression. R package version 1.1.

Schabenberger, H. (2009b). Spatial count regression models with applications to health insurance data. Master's thesis, Technische Universität München (*www-m4.ma.tum.de/Diplarb/*).

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal Of The Royal Statistical Society Series B 64*(4), 583–639.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica 57*(2), 307–333.

Winkelmann, R. (2008). *Econometric Analysis of Count Data* (5th ed.). Berlin: Springer-Verlag.

Yip, K. C. and K. K. Yau (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics 36*(2), 153–163.