Klein, Nadja; Denuit, Michel; Lang, Stefan; Kneib, Thomas

Working Paper

Nonlife ratemaking and risk management with Bayesian additive models for location, scale and shape

Working Papers in Economics and Statistics, No. 2013-24

Provided in Cooperation with:
Institute of Public Finance, University of Innsbruck

Suggested Citation: Klein, Nadja; Denuit, Michel; Lang, Stefan; Kneib, Thomas (2013) : Nonlife ratemaking and risk management with Bayesian additive models for location, scale and shape, Working Papers in Economics and Statistics, No. 2013-24, University of Innsbruck, Research Platform Empirical and Experimental Economics (eeecon), Innsbruck

This Version is available at:
https://hdl.handle.net/10419/101078

University of Innsbruck

Nonlife ratemaking and risk management with bayesian additive models for location, scale and shape

**Nadja Klein, Michel Denuit,
Stefan Lang, Thomas Kneib**

Working Papers in Economics and Statistics

2013-24

Research Platform
eeecon
Empirical and Experimental Economics
University of Innsbruck
http://eeecon.uibk.ac.at/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact Address:
University of Innsbruck
Department of Public Finance
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 7171
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
http://eeecon.uibk.ac.at/wopec/

For a list of recent papers see the backpages of this paper.

Nonlife Ratemaking and Risk Management with Bayesian Additive Models for Location, Scale and Shape

Nadja Klein
University of Göttingen

Michel Denuit
Université Catholique de Louvain

Stefan Lang
University of Innsbruck

Thomas Kneib
University of Göttingen

Abstract

Generalized additive models for location, scale and shape define a flexible, semi-parametric class of regression models for analyzing insurance data in which the exponential family assumption for the response is relaxed. This approach allows the actuary to include risk factors not only in the mean but also in other parameters governing the claiming behavior, like the degree of residual heterogeneity or the no-claim probability. In this broader setting, the Negative Binomial regression with cell-specific heterogeneity and the zero-inflated Poisson regression with cell-specific additional probability mass at zero are applied to model claim frequencies. Models for claim severities that can be applied either per claim or aggregated per year are also presented. Bayesian inference is based on efficient Markov chain Monte Carlo simulation techniques and allows for the simultaneous estimation of possible nonlinear effects, spatial variations and interactions between risk factors within the data set. To illustrate the relevance of this approach, a detailed case study is proposed based on the Belgian motor insurance portfolio studied in Denuit and Lang (2004).

Key words: overdispersed count data; mixed Poisson regression; zero-inflated Poisson; Negative Binomial; zero-adjusted models; MCMC; probabilistic forecasts.

1 Introduction

Calculations of motor insurance premiums are based on detailed statistical analyses of large data bases maintained by insurance companies, recording individual claim experience. The actuarial evaluation relies on a statistical model incorporating all the available information about the risk. Premiums then often vary by the territory in which the vehicle is garaged, the use of the vehicle (driving to and from work or business use) and individual characteristics (such as age, gender, occupation and marital status of the main driver of the vehicle, for instance). If the policyholders misrepresent any of these classification variables in their declaration, they are subject to loss of coverage when they are involved in a claim. There is thus a strong incentive for accurate reporting of risk characteristics making insurance data reliable.

It is now common practice to achieve a priori risk classification with the help of Generalized Linear Models (GLMs), see, e.g., Denuit et al. (2007) for an introduction in relation with motor insurance. They are so called because they generalize the classical linear model based on the Normal distribution to similar regression models for Poisson, Binomial, Gamma or Inverse-Gaussian responses, for instance. The main drawback of GLMs is that covariate effects are modeled in the form of a linear predictor. This is not a problem for categorical explanatory variables coded by means of binary variables, but a strong restriction for continuous explanatory variables which may have a nonlinear effect on the score. It has been common practice in insurance companies to model possibly nonlinear effects by means of polynomials. However, it is now well documented that low-degree polynomials are often not flexible enough to capture the variability in the data and that increasing their degree produces unstable estimates, especially for extreme values of the covariates. Although banding results in a loss of information, a model employing a banded version of a continuous covariate is sometimes considered more practical than one which employs the (untransformed) continuous variable. However, there is no general rule to determine the optimal choice of cutoffs so that banding may bias risk evaluation.

Among continuous covariates, geographic area plays a particular role. It can either be seen as a function of two coordinates if exact locations are available or a function of an administrative areal variable if spatial information is aggregated for confidentiality

reasons. In any case, actuaries wish to estimate the spatial variation in risk premium and to price accordingly. Spatial zip code methods for insurance rating attempt to extract information which is in addition to that contained in standard factors (like age or gender for instance). With the regression models discussed in the present paper, the effect of continuous and spatial covariates is modeled on the score scale by means of smooth, unspecified functions estimated from the data.

Generalized additive models (GAMs) as developed in Hastie and Tibshirani (1990) and popularized by Wood (2006) provide a convenient framework to overcome the linearity assumptions inherent to GLMs when smooth effects of continuous covariates need to be included in an additive predictor. Inference can be realized by cross validation as in Wood (2004), by mixed model representations as in Ruppert et al. (2003), Fahrmeir et al. (2004) and Wood (2008) or by Markov chain Monte Carlo (MCMC) simulations as in Brezger and Lang (2006), Julion and Lambert (2007) and Lang et al. (2013).

The framework of generalized additive models for location, scale and shape (GAMLSS) introduced by Rigby and Stasinopoulos (2005) allows to extend GAMs to more complex response distributions where not only the expectation but multiple parameters are related to structured additive predictors with the help of suitable link functions. Structured additive regression relies on a unified representation of different model terms like parametric linear effects, smooth nonlinear effects of continuous covariates, interaction terms based on varying coefficients and spatial effects (Fahrmeir et al., 2013, Brezger and Lang, 2006). In particular, zero-inflated, skewed and zero-adjusted distributions can be embedded in this framework as special cases where all occurring parameters are related to regression predictors and may depend on a complex covariate structure. All these model terms rely on a unifying representation based on non-standard basis function specifications in combination with quadratic penalties in a frequentist formulation or Gaussian priors in a Bayesian approach.

In this broader setting, the Poisson assumption for claim frequencies made in Denuit and Lang (2004) is replaced with a mixed Poisson one, with Gamma distributed random effect (yielding the Negative Binomial distribution with cell-specific heterogeneity) or Bernoulli distributed random effect (yielding the zero-inflated Poisson distribution with cell-specific additional probability mass at zero).

In addition to claim frequencies, we also consider regression models for claim severities. Much attention has been paid in the actuarial literature to find suitable distributions to model claim sizes; see for example Klugman et al. (2004). Whereas Denuit and Lang (2004) studied claim frequencies and claim severities separately, we consider in this paper the so-called zero-adjusted models that allow to account for zeros in the analysis of the amount of loss directly without resorting to models for claim frequencies. Zero-adjusted distributions combine a continuous distribution on the positive real line and a point mass at zero, such that the probabilities for a claim and quantiles of the claim size distribution can be estimated in one model. Zero-adjusted models are in the line of Jørgensen and Paes de Souza (1994) and Smyth and Jørgensen (2002) where the zero claims are included using the Tweedie distribution. However, this model has the disadvantage that the probability at zero cannot depend on covariates whereas here, this key actuarial indicator is allowed to vary according to risk characteristics.

To select an appropriate response distribution and to specify several predictors that correspond for instance to variances, skewness or overdispersion of the distribution, we rely mainly on the deviance information criterion (DIC) of Spiegelhalter et al. (2002) whose performance in Bayesian count data regression within the framework of GAMLSS has been tested in Klein et al. (2013a). The choice of the distribution will be supported by normalized quantile residuals (Dunn and Smyth, 1996) and proper scoring rules (Gneiting and Raftery, 2007).

We highlight the advantages of complex Bayesian count data, skewed and zero-adjusted regression models for insurance claims data with a detailed analysis of a Belgian data set with more than 160,000 policies. Specifically,

- we consider the Poisson, zero-inflated Poisson and Negative Binomial regression models for claim frequencies, where suitable predictors are specified for the expected number of claims as well as for the probability of the structural zeros in zero-inflated Poisson and for the scale parameter of the Negative Binomial distribution.

- for claim severities, we extend the continuous models to zero-adjusted versions of the Gamma, Inverse-Gaussian and LogNormal distributions, we estimate the corresponding location and scale or shape parameters as well as the probability

of a claim in terms of relevant covariates in an additive fashion.

- inference in all model formulations is based on iteratively weighted least squares approximations to the full conditionals in Markov chain Monte Carlo (MCMC) simulation techniques as suggested in Gamerman (1997) or Brezger and Lang (2006) and extended to the general framework of GAMLSS by Klein et al. (2013b).

- we benefit from a numerically efficient implementation in the free open software BayesX also available via the R add-on package R2BayesX.

- compared to frequentist GAMLSS approaches, the approach adopted here directly includes the choice of smoothing parameters in the estimation run and provides valid confidence intervals which are difficult to obtain from asymptotic maximum likelihood theory.

Our approach to zero-inflated, skewed and zero-adjusted models has therefore the full flexibility in the parametric distribution assumption. The structured additive modeling of all parameters allows to focus on specific aspects of the data that go beyond the mean. In particular,

- the separate modeling of the probability mass at zero as a function of the observable characteristics allows for an accurate analysis of this key actuarial indicator.

- cell-specific residual heterogeneity in the Negative Binomial model allows for more accurate risk predictions when deriving the predictive distributions of future claims.

- zero-augmented models for the annual claim amounts are in accordance with the individual model of risk theory so that the actuarial analysis benefits from the numerous tools developed in that setting.

The claim frequencies models considered in the present paper have already been applied to insurance data. See, e.g., Yip and Yau (2005) or Boucher et al. (2006). However, previous applications of Negative Binomial or zero-inflated Poisson regression models to insurance data only allowed for linear effects of the covariates or applied

preliminary banding techniques to transform continuous covariates into categorical ones. Zero-adjusted Gamma and Inverse-Gaussian models have been proposed by Heller et al. (2006), Bortoluzzo et al. (2011) and Resti et al. (2013) but their analysis only allowed for linear effects of the covariates, too. See also Heller et al. (2007) for a related model extending the Tweedie construction beyond the Poisson-Gamma setup. The present paper innovates in that nonlinear effects are allowed using the efficient inference techniques developed by Klein et al. (2013a). The effect of continuous covariates on the score are quantified by means of unknown smooth functions that do not need to be specified a priori under parametric form but are estimated directly from the data.

Let us now briefly present the data used to illustrate the techniques described in this paper. The data set is the one analyzed in Denuit and Lang (2004) by means of Poisson and LogNormal regression techniques. It relates to a Belgian motor third-party liability insurance portfolio observed during the year 1997, comprising more than 160,000 policies. The following information is available on an individual basis. As far as policyholders' characteristics are concerned, we know gender, age, place of residence and use of the car. Concerning the insured vehicle, we have its ancientness, the type of fuel, its power and whether the vehicle belongs to a fleet. About the contract, we know the type of coverage (compulsory motor third party liability only, or motor third party liability together with some optional coverages) and the level occupied in the former Belgian bonus-malus scale. In addition to these covariates, the number of claims filed by each policyholder during 1997, the exposure-to-risk from which these claims originated, as well as the resulting total claim amount are given. See Table 1 for a list of available explanatory variables with some descriptive statistics.

Notice that the -1/+1 coding has been used for the binary covariates in Table 1 whereas actuaries usually resort to a 0/1 coding, with 0 for the most populated class taken as reference (see, e.g., Denuit and Lang, 2004). This is because the -1/1 coding often has a positive mixing behavior in MCMC. Of course, the actuary can easily revert to the standard 0/1 coding, if needed, by an appropriate linear transformation of the regression parameters. The extent of coverage has been coded in a similar way

	Continuous covariates			
variable	description	mean	std dev.	min/max
ageph	age of policyholder	47	14.83	18/78
agec	age of vehicle	7.35	4.12	0/30
power	engine power	56.01	19.02	10/243
	Binary covariates			
variable	description	levels	proportions in %	
fuel	fuel oils	gas=1/diesel=-1	69.1/30.9	
use	use of vehicle	work=1/private=-1	4.8/95.2	
fleet	belongings to a fleet	yes=1/no=-1	3.2/96.8	
sex	gender of policyholder	male=1/female=-1	73.5/26.5	
	Categorical covariates			
variable	description	levels	proportions in %	
coverage	guarantees subscribed	TPL only (1)	58.2	
		TPL+limited material damage and theft (2)	28.3	
		TPL+comprehensive damage (3)	13.6	
bm	bonus-malus level	0,...,22		
district	spatial information	1,...,589		

Table 1: Covariates available in the car insurance data set.

by means of two auxiliary covariates cov1 and cov2 defined as follows:

$$\begin{aligned} cov1 &= I[coverage = 1] - I[coverage = 3] \\ cov2 &= I[coverage = 2] - I[coverage = 3] \end{aligned}$$

where $I[\cdot]$ denotes the indicator function.

Even if it is common to include the gender of the main driver in the actuarial ratemaking, some states have banned the commercial use of this rating factor (as in EU member countries, for instance). Here, we assume that we deal with the technical price list and we do not discuss the various adaptations made to produce the commercial price list.

The remaining sections are organized as follows: In Section 2 we introduce the specification of Bayesian models for analyses of claim frequencies as well as severities. We describe the underlying inference and we give guidelines for model choice. Sections 3 and 4 deal with claim frequency and claim severity models, respectively. The applicability of the proposed models is demonstrated on the Belgian data set presented before. The analysis we conduct shows major improvements compared to existing approaches and reveal new features of insurance data. Conclusions are given in the last Section 5.

2 Regression Models

2.1 Generalized additive models for location scale and shape (GAMLSS)

Suppose we have observations $(y_i, \boldsymbol{x}_i)$, $i = 1, 2, \ldots, n$ on a response variable Y and a vector of covariates $\boldsymbol{x}$. Here, Y may be either the claim size or the claim frequency. The vector $\boldsymbol{x}$ is split up into a subset of continuous covariates, another subset of categorical covariates, and geographical information. We assume that there are p continuous covariates $x_1, \ldots, x_p$, spatial region s, and that the remaining categorical explanatory variables are coded by means of a vector $\boldsymbol{z}_0$ of covariates.

GAMLSS require a parametric distribution assumption for the response variable, involving several parameters represented as functions of explanatory variables. In

GAMLSS the exponential family distribution assumption for the response Y is relaxed so that the actuarial analysis is no more restricted to the distributions used in the classical GLM setting. The systematic part of the model is expanded compared to GLMs to allow modeling not only the mean (or location) but other parameters of the distribution of the response as linear and/or nonlinear parametric and/or additive non-parametric functions of explanatory variables.

More precisely, the model is built as follows: The form of the distribution assumed for the response variable can be very general. The only restriction is that the individual contribution to the log-likelihood and its first two derivatives with respect to each of the parameters must be computable. Known monotonic link functions relate the distribution parameters to explanatory variables. Specifically, distribution parameters $\vartheta_1, \vartheta_2, \ldots$ are decomposed into $h_k(\vartheta_k) = \eta_k$ for some known link function h_k where the score η_k is assumed to be of a semi-parametric additive form

$$\eta_k = \boldsymbol{z}_0' \boldsymbol{\beta}_{0k} + \sum_{j=1}^{p} f_{kj}(x_j) + f_{k,\text{spat}}(s)$$

where the functions f_{kj} express the effect of continuous covariates on the score scale, $f_{k,\text{spat}}$ accounts for spatial variations in the risk distribution, $\boldsymbol{z}_0' \boldsymbol{\beta}_{0k}$ contains parametric, linear effects of covariates.

All the distribution parameters can be decomposed on the score scale as linear, nonlinear parametric, non-parametric (smooth) functions or spatial variations of the explanatory variables. The parameters are estimated within the GAMLSS framework by maximizing a penalized likelihood function. More details on how the penalized log likelihood is maximized are given in Rigby and Stasinopoulos (2005). The available distributions include all those commonly used in actuarial analyses. See Table 1 in Stasinopoulos and Rigby (2007) for an exhaustive list.

2.2 Bayesian structured additive regression

Hereafter, we give a brief overview of the statistical concepts used in the GAMLSS model terms. A more tutorial style introduction is given in Denuit and Lang (2004) and in the textbook by Fahrmeir et al. (2013). In structured additive regression each vector $\boldsymbol{f}$ containing the evaluations of a function f at the observed covariates is approximated in terms of appropriate basis functions which allows the unified

representation

$$\boldsymbol{f} = \boldsymbol{Z}\boldsymbol{\beta}$$

where $\boldsymbol{Z}$ is a design matrix arising from evaluations of the basis functions and $\boldsymbol{\beta}$ is the vector of regression coefficients.

Each predictor from the previous section can therefore be expressed in terms of

$$\boldsymbol{\eta} = \boldsymbol{Z}_0\boldsymbol{\beta}_0 + \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \ldots + \boldsymbol{Z}_p\boldsymbol{\beta}_p + \boldsymbol{Z}_{\text{spat}}\boldsymbol{\beta}_{\text{spat}}$$

where for simplicity we drop the parameter name.

In a Bayesian framework a standard smoothness prior is a (possibly improper) Gaussian prior of the form

$$p(\boldsymbol{\beta}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{\text{rank}(\boldsymbol{K})/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}'\boldsymbol{K}\boldsymbol{\beta}\right) \cdot I[\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}] \tag{1}$$

where $\boldsymbol{\beta}$ may be any of the vectors $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p, \boldsymbol{\beta}_{\text{spat}}$. The key components of the prior are the penalty matrix $\boldsymbol{K}$, the variance parameter τ^2 and the constraint $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$.

The structure of the penalty or prior precision matrix $\boldsymbol{K}$ depends on the covariate type and on prior assumptions about smoothness of f, see below. Usually, the penalty matrix in our examples is rank deficient resulting in a partially improper prior. The amount of smoothness is governed by the variance parameter τ^2: The smaller the variance, the smoother the function estimates and the other way around. A conjugate inverse Gamma prior is employed for τ^2 with small values for the hyperparameters resulting in an uninformative prior on the log scale.

For instance, for modeling the nonlinear effect $f(ageph)$ we apply Bayesian P(enalized)-splines as proposed in Lang and Brezger (2004). P-splines assume that the unknown functions can be approximated by a polynomial spline which can be written in terms of a linear combination of B-spline basis functions. Hence, the columns of the design matrix $\boldsymbol{Z}$ are given by the B-spline basis functions evaluated at the observations. Lang and Brezger (2004) propose to use first or second order random walks as smoothness priors for the regression coefficients, i.e.

$$\beta_l = \beta_{l-1} + u_l, \quad \text{or} \quad \beta_l = 2\beta_{l-1} - \beta_{l-2} + u_l, \tag{2}$$

with Gaussian centered errors u_l with common variance τ_j^2 and diffuse priors $p(\beta_1) \propto const$, or $p(\beta_1)$ and $p(\beta_2) \propto const$, for initial values. This prior is of the form (1) with

penalty matrix given by $\boldsymbol{K} = \boldsymbol{D}'\boldsymbol{D}$, where $\boldsymbol{D}$ is a first or second order difference matrix.

A common way to deal with the spatial covariate *distr* is to define a neighborhood structure ∂_s on the set $\{1, \ldots, 589\}$ of districts in Belgium and to assume that neighboring regions are similar. In our case neighborhoods are simply given by common borders. The design matrix is an indicator matrix connecting individual observations with corresponding regions, i.e. the entry (i, s) of the matrix $\boldsymbol{Z}$ is one if observation i belongs to region s and zero otherwise. To implement spatial smoothness, $\boldsymbol{K}$ is chosen as an adjacency matrix indicating which regions are neighbors of each others, see Rue and Held (2005) for details. The simplest smoothness prior is called Markov random field. In this case, given β_r, $r \neq s$ and hyperparameter τ^2, β_s is Normally distributed with mean $\sum_{r \in \partial_s} \frac{1}{\mathrm{N}_s} \beta_r$ and variance $\frac{\tau^2}{\mathrm{N}_s}$, where N_s is the number of neighbors of region s. In consequence, the conditional mean of β_s given all other coefficients is the average of the neighborhood regions. Further background about Markov random fields can be found in Fahrmeir et al. (2013).

Especially in the application on claim sizes, it is useful to split up the effect f_{spat} into a spatially structured (smooth) effect f_{str} and a spatially unstructured effect f_{unstr}, i.e. $f_{\text{spat}} = f_{\text{str}} + f_{\text{unstr}}$. The unstructured part is modeled by independent and identically distributed Gaussian random effects, i.e. $\boldsymbol{Z}$ is again an indicator matrix as in the Markov random field and $\boldsymbol{\beta}$ is multivariate Normal with zero mean and diagonal covariance matrix.

The type of modeling discussed so far can be embedded in a computationally very powerful multilevel version of structured additive regression models where regression coefficients may themselves depend on covariates and can be modeled by a structured additive predictor. We refer the reader to Lang et al. (2013) for details. Bayesian inference is based on highly efficient Markov chain simulation based on iteratively weighted least squares proposals as suggested by Gamerman (1997) or Brezger and Lang (2006) and for the models at hand in a recent paper by Klein et al. (2013b).

2.3 Model choice and predictive ability

Since we are dealing with many (possibly nonlinear) covariate effects and usually several additive predictors, adequate model choice plays a crucial role when apply-

ing GAMLSS. Unfortunately, many diagnostic tools in Bayesian inference based on MCMC are only useful for restricted classes of simple models with just a few parameters. Klein et al. (2013b) proposed a practical procedure for model selection in Bayesian GAMLSS to find a suitable response distribution as well as parsimonious specifications for all model predictors. We mainly follow their proposals since they do not only focus on improving the fit to the data but also the quality of predictions for new observations.

A first measure for the model fit that takes the estimated model complexity into account is the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) that has become quite popular in Bayesian statistics as it can easily be computed from the MCMC output. Let $\boldsymbol{\vartheta}$ denote the vector of model parameters. If $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(T)}$ is an MCMC sample from the posterior distribution of model parameters, then the DIC is based on the following two quantities: The deviance $D(\boldsymbol{\vartheta}) = -2\log(p(\boldsymbol{y}|\boldsymbol{\vartheta}))$ of the model that reflects the fit of the data and model complexity through the effective number of parameters in the model p_D. As derived by Spiegelhalter et al. (2002), the latter is given by the difference of the posterior mean of the deviance and the deviance of the posterior means, i.e. $p_D = \overline{D(\boldsymbol{\vartheta})} - D(\overline{\boldsymbol{\vartheta}})$, where

$$\overline{D(\boldsymbol{\vartheta})} = \frac{1}{T}\sum_{t=1}^{T} D\left(\boldsymbol{\vartheta}^{(t)}\right) \text{ and } \overline{\boldsymbol{\vartheta}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\vartheta}^{(t)}.$$

The DIC is then defined as

$$\text{DIC} = \overline{D(\boldsymbol{\vartheta})} + 2p_D = 2\overline{D(\boldsymbol{\vartheta})} - D(\bar{\boldsymbol{\vartheta}})$$

indicating a close relationship to the frequentist Akaike information criterion that provides a similar compromise between fidelity to the data and model complexity.

A rough rule of thumb says that DIC differences of 10 and more between two competing models indicate the model with the lower DIC to be superior. For fixed distributional assumptions for the response, we use the DIC for variable selection in all predictors, as already described in the previous section. If for one distribution with different parameter specifications several models have similar DIC with differences smaller than 10, we usually decide for the sparser model in the sense that additional non-significant effects are excluded from the predictors. Our variable selection typically starts with the examination of the location parameter of the distributions. After

having a reasonable predictor for this main parameter, we continue in determining the remaining predictors by a stepwise search. Note that the performance of the DIC in count data models has been evaluated in the supplement of Klein et al. (2013a) and rated to provide valid guidance.

For discriminating between competing response distributions, we rely on normalized quantile residuals and proper scoring rules. Stasinopoulos et al. (2008) suggested normalized quantile residuals as a graphical device. For continuous response data, the quantile residual relies on $u_i = F_i(y_i|\hat{\boldsymbol{\vartheta}})$, where F_i is the cumulative distribution function estimated for the ith individual, $\hat{\boldsymbol{\vartheta}}$ contains all estimated model parameters and y_i is the corresponding observation. If the estimated F_i is close to the true distribution of y_i then u_i approximately follows a uniform distribution. To use standard quantile-quantile plots for the graphical investigation of the model fit, the residual is finally defined as $\hat{r}_i = \Phi^{-1}(u_i)$ where Φ^{-1} is the inverse cumulative distribution function of the standard Normal distribution. Hence, r_i is approximately standard Normal if the estimated model is close to the true one. For discrete response distributions, the definition has to be slightly extended and u_i is defined as a random number from the uniform distribution on the interval $[F_i(y_i - 1|\hat{\boldsymbol{\vartheta}}), F_i(y_i|\hat{\boldsymbol{\vartheta}})]$. In any case, the model fit can be evaluated by means of usual quantile-quantile plots: The closer the residuals to the bisecting line, the better the fit to the data.

Gneiting and Raftery (2007) propose proper scoring rules as summary measures for the evaluation of probabilistic forecasts based on the predictive distribution and the observed realizations. In our analysis, we apply a selection of these scores that have been assigned to the general framework of regression in GAMLSS by Klein et al. (2013b). There are several candidates for the score function S, depending on whether the response variable is discrete, continuous or mixed discrete-continuous. In all three cases, we use the logarithmic, quadratic and spherical score which are all proper scoring rules meaning that if G_i is the true distribution of the response y_i then $S(G_i, G_i) \geq S(G_i, F_i)$ holds for all $F_i \neq G_i$. We refer the reader to Gneiting and Raftery (2007) for further details on these scores. In practice, the predictive distributions F_i for observations y_i are obtained by ten-fold cross validation, that is, 10% of the data set is left out and the parameters of the distributions are predicted based on the estimations of the remaining 90%. The scores are then the sum of individual con-

tributions, where higher scores deliver better probabilistic forecasts when comparing different models.

Since the total score is only a summary measure for the complete predictive distribution, it does not allow to assess which parts of the true response distribution are reflected well in the model and which aspects differ from the truth. For example, there may be a distribution that fits the central part of the data well but fails to fit the tails. In terms of a summarized score, it may then be difficult to distinguish this behavior from another candidate distribution that differs from the true distribution slightly over the whole domain. In such a case, it can be helpful to follow Gneiting and Ranjan (2011) who suggest a quantile decomposition of the quantile score given by

$$-2 \int_0^1 \left(I\left[y_i \leq F_i^{-1}(\alpha)\right] - \alpha \right) \left(F_i^{-1}(\alpha) - y_i \right) \mathrm{d}\alpha.$$

In this way, the performance of the distributions can be compared with respect to the ability to fit specific quantiles of the true distribution. We apply this score later on to compare distributions for analyzing the claim sizes and to determine claim sizes that cause the largest deteriorations in the score. The whole integral over the quantile decomposition leads to the continuous ranked probability score (CRPS) which is also a proper scoring rule directly formulated in terms of the cumulative distribution functions while the logarithmic, quadratic and spherical scores are computed from the densities.

3 Modelling Claim Frequencies: Count Data Regression

3.1 Negative Binomial regression with cell-specific heterogeneity

Insurance data often exhibits overdispersion because several important risk attitudes cannot be observed (swiftness of reflexes, aggressiveness behind the wheel, consumption of drugs, etc.). This can be modeled by the inclusion of a random heterogeneity factor. Specifically, given $\Theta_i = \theta$, the number of claims N_i reported by policyholder i conforms to the Poisson distribution with mean $\mu_i\theta$. Here, we consider that

$\Theta_1, \Theta_2, \Theta_3, \ldots$ are independent and Gamma distributed random variables, with unit mean and scale parameter δ_i depending on covariates (so that the heterogeneity may be specific to risk classes). Therefore we obtain a Negative Binomial (NB) distribution for N_i, that is,

$$\Pr[N_i = k] = \frac{\Gamma(k+\delta_i)}{\Gamma(k+1)\Gamma(\delta_i)} \Big(\frac{\mu_i}{\delta_i+\mu_i}\Big)^k \Big(\frac{\delta_i}{\delta_i+\mu_i}\Big)^{\delta_i}, \quad k = 0, 1, 2, \ldots \tag{3}$$

where $\Gamma(a) = \int_0^\infty \exp(-t)t^{a-1}dt$ for $a > 0$.

To compare with the results obtained by Denuit and Lang (2004), we also consider the Poisson regression model for claim counts where N_i obeys the Poisson distribution with mean λ_i. Both μ and λ are expressed in terms of frequency scores η_{freq} using log link, i.e. $\lambda_{\text{freq}} = \exp(\eta^{\lambda}_{\text{freq}})$ and $\mu_{\text{freq}} = \exp(\eta^{\mu}_{\text{freq}})$. In both cases, we end up with a frequency score of the form

$$f_1^{\mu}(ageph) + sex\, f_2^{\mu}(ageph) + f_3^{\mu}(agec) + f_4^{\mu}(bm) + f_5^{\mu}(power) + f_{\text{spat}}^{\mu}(distr) + (\boldsymbol{z}^{\mu})'\boldsymbol{\beta}.$$

where μ is replaced by λ to obtain the frequency score of the Poisson model. Here, $\boldsymbol{z}$ contains the linear effects *fuel*, *coverage* and *fleet*, an intercept and the logarithm of the contract periods in days. The other categorical covariates do not significantly influence the claim frequencies. Recall that in a Bayesian context an effect is seen as not significant if the zero line is totally contained in the pointwise confidence interval. The nonlinear functions $f_1, \ldots, f_5$ are smooth estimates of continuous covariates. As gender and age often interact, in the sense that the effect of age on the average claim frequency is different for males than for females (typically, young male drivers are more dangerous than young female drivers), we allow for such an effect in our study. Formally, f_2 is an interaction between the gender of the policyholder and age. With the coding used in this paper, this means that the effect of age on the score scale is $f_1(ageph) + f_2(ageph)$ for males and $f_1(ageph) - f_2(ageph)$ for females. Then, f_{spat} captures the spatial variation of the claim frequencies. Notice that we did not include the logarithmic contract period in days as an offset but included it as a linear covariate. If the assumption of an offset is justified then the corresponding estimated coefficient should be close to one.

The scale parameter δ_{freq} is expressed in terms of a score $\eta^{\delta}_{\text{freq}}$ by means of a log link, i.e. $\delta_{\text{freq}} = \exp(\eta^{\delta}_{\text{freq}})$. We start from a model with the intercept, only. Based on the deviance information criterion (DIC) and significances in effects, we included step by

step further covariates and estimated in this way several models with different score structures. This finally gives an optimal score of the form

$$\eta^{\delta}_{\text{freq}} = f^{\delta}_{\text{spat}}(distr) + \left(\boldsymbol{z}^{\delta}\right)' \boldsymbol{\beta}^{\delta},$$

with $\boldsymbol{z}^{\delta}$ consisting of *fuel*, *coverage*, an intercept and the logarithm of the contract periods in days.

Estimates of linear effects are given in Table 2 for the Poisson model and in Table 3 for the Negative Binomial regression model. Considering the values displayed in these two tables, we see that the Negative Binomial and Poisson regression models produce almost identical estimations for the linear part of the scores $\eta^{\lambda}_{\text{freq}}$ and η^{μ}_{freq}. We see from the estimations of β^{λ}_1 and β^{μ}_1 that gasoline vehicles appear to be less risky compared to diesel ones. This effect prevails in Belgium and is often explained by the higher annual mileage for diesel vehicles which are usually driven over longer distances to compensate the higher buying cost of the car by the cheaper price of fuel oils. The estimated values of β^{λ}_2 and β^{μ}_2 suggest that belonging to a fleet decreases the expected claim number. The estimated values of $(\beta^{\lambda}_3, \beta^{\lambda}_4)$ and $(\beta^{\mu}_3, \beta^{\mu}_4)$ reveal that buying only TPL increases the frequency score by β^{λ}_3 or β^{μ}_3 (and thus the expected claim number) whereas extending the coverage lowers it (by β^{λ}_4 or β^{μ}_4 for TPL+limited material damage and theft and by $-\beta^{\lambda}_3 - \beta^{\lambda}_4$ or $-\beta^{\mu}_3 - \beta^{\mu}_4$ for TPL+comprehensive coverage). This confirms the tendency for drivers subscribing more guarantees to report less third-party liability claims. Considering the estimated β^{λ}_5 and β^{μ}_5, we see that the exposure-to-risk is multiplied by a coefficient strictly less than 1. This contradicts the widely used offset construction which tends to grant too much importance to this effect and questions the use of calendar time as the appropriate measure of risk exposure, instead of the distance travelled (which is of course difficult to measure accurately and subject to misreporting, but this may well change if pay-as-you-drive systems become more popular).

In Figures 1 and 2, the estimated nonlinear effects on the mean λ in the Poisson regression case and on the mean μ in the Negative Binomial regression model are plotted together with 80% and 95% pointwise credible intervals. Vertical stripes indicate the relative amount of data of the corresponding covariate values. Again, we see that the estimated nonlinear effects are almost identical in the Poisson and Negative Binomial regression models. Let us summarize the main findings as follows:

The effect of age on the score scale is depicted in Figure 1. The estimated functions f_1^λ, f_2^λ, f_1^μ and f_2^μ are displayed there. The interaction age-gender is significant and reveals the higher claim frequencies for young, unexperienced drivers. This effect is much more pronounced for young male drivers. The total estimated age effect $f_1^\lambda + f_2^\lambda$ or $f_1^\mu + f_2^\mu$ for males and $f_1^\lambda - f_2^\lambda$ or $f_1^\mu - f_2^\mu$ for females is also described in Figure 1. Ages 35-45 correspond to the reference level for males whereas females are subject to a peak around age 45. This phenomenon is usually attributed to accidents caused by children learning to drive behind the wheel of their mother's car. Also, it was not uncommon in Belgium during the late 1990s to ask older relatives, often the mother, to subscribe the policy in order to avoid premium surcharges imposed to young drivers. Male drivers around 70 reveal a very low risk level. This beneficial effect disappears at older ages (after age 80, the analysis suggests an increasing risk level but no firm conclusion can be drawn as the right part of the estimated $f_1 + f_2$ does not significantly differ from 0). The risk seems to stay constant after age 60 for female drivers. The other estimated nonlinear effects are displayed in Figure 2. Considering the estimation of f_3, we see that vehicles up to 3 years seem to be riskier, which can be explained by the first annual mechanical check up imposed by the Belgian state at that time. Drivers with high annual mileage tend to sell their cars after 3 to 4 years, which explains the shape in f_3. Occupying a higher level in the bonus-malus scale results from a worse claim frequency experience so that f_4 is increasing, as expected. Also, f_5 appears to be increasing so that more powerful cars tend to cause more accidents. There is a clearly visible kink around power 50, with a close-to-linear behavior before and after that value.

Figure 3 depicts the estimated spatial effects on λ and μ. The estimated spatial variation in λ and μ is very similar for the Poisson and Negative Binomial regression models. The structured spatial effect clearly indicates the higher risk associated with the main cities: Brussels area in the center, Antwerp in the North, Liège in the East and Charleroi in the South. On the contrary, living in the countryside reduces the expected claim number, like in the Ardenne, South-East. These effects clearly dominate as shown by the probability plots which isolate the significant effects: The districts corresponding to significantly negative effects are colored in black, those with significantly positive effects are colored in white whereas those colored in grey are not

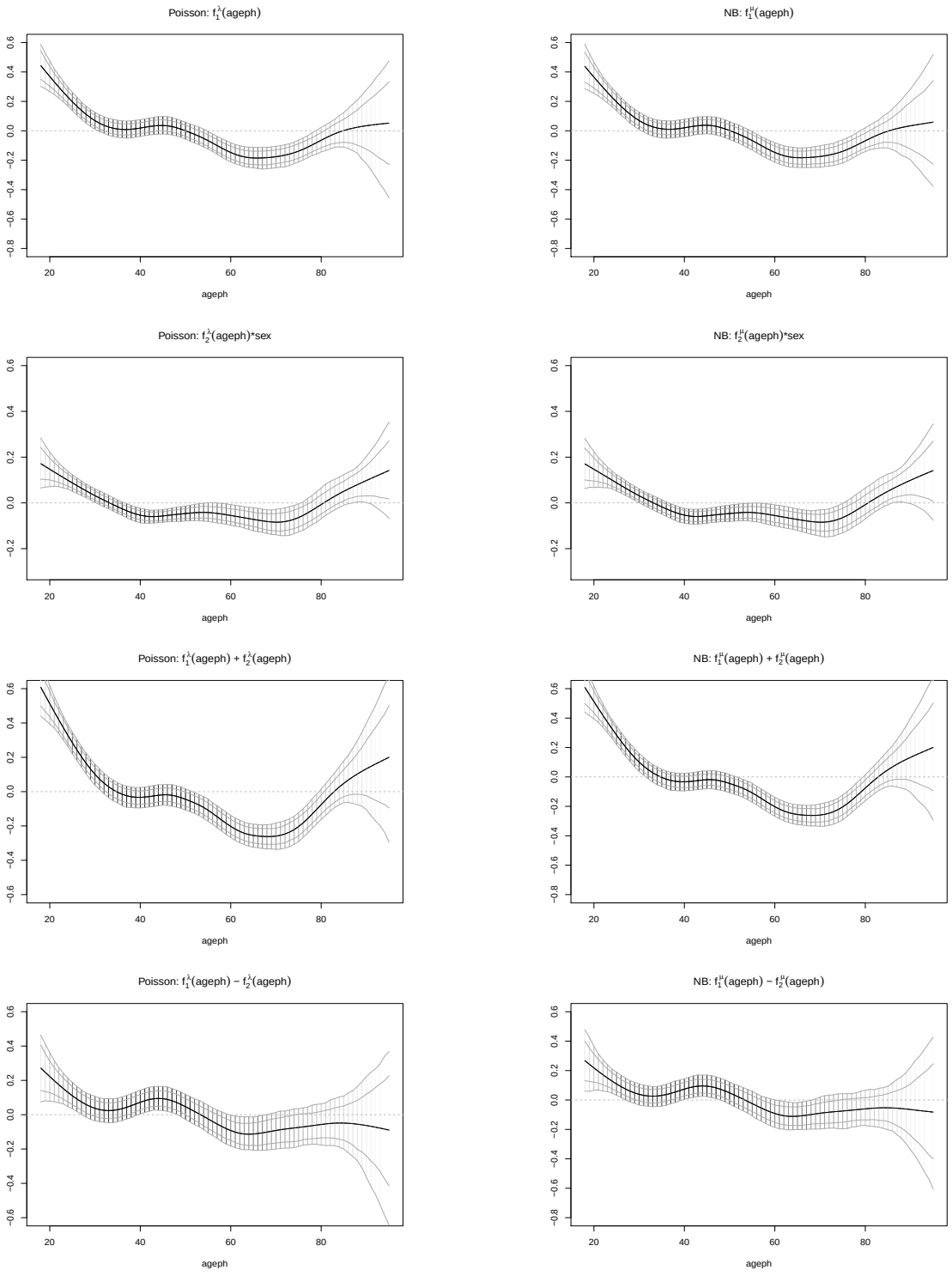

Figure 1: Posterior mean estimates of nonlinear age effects (centered around zero) on the expectation parameter together with pointwise 80% and 95% confidence intervals in the Poisson (left panels) and Negative Binomial models (right panels).

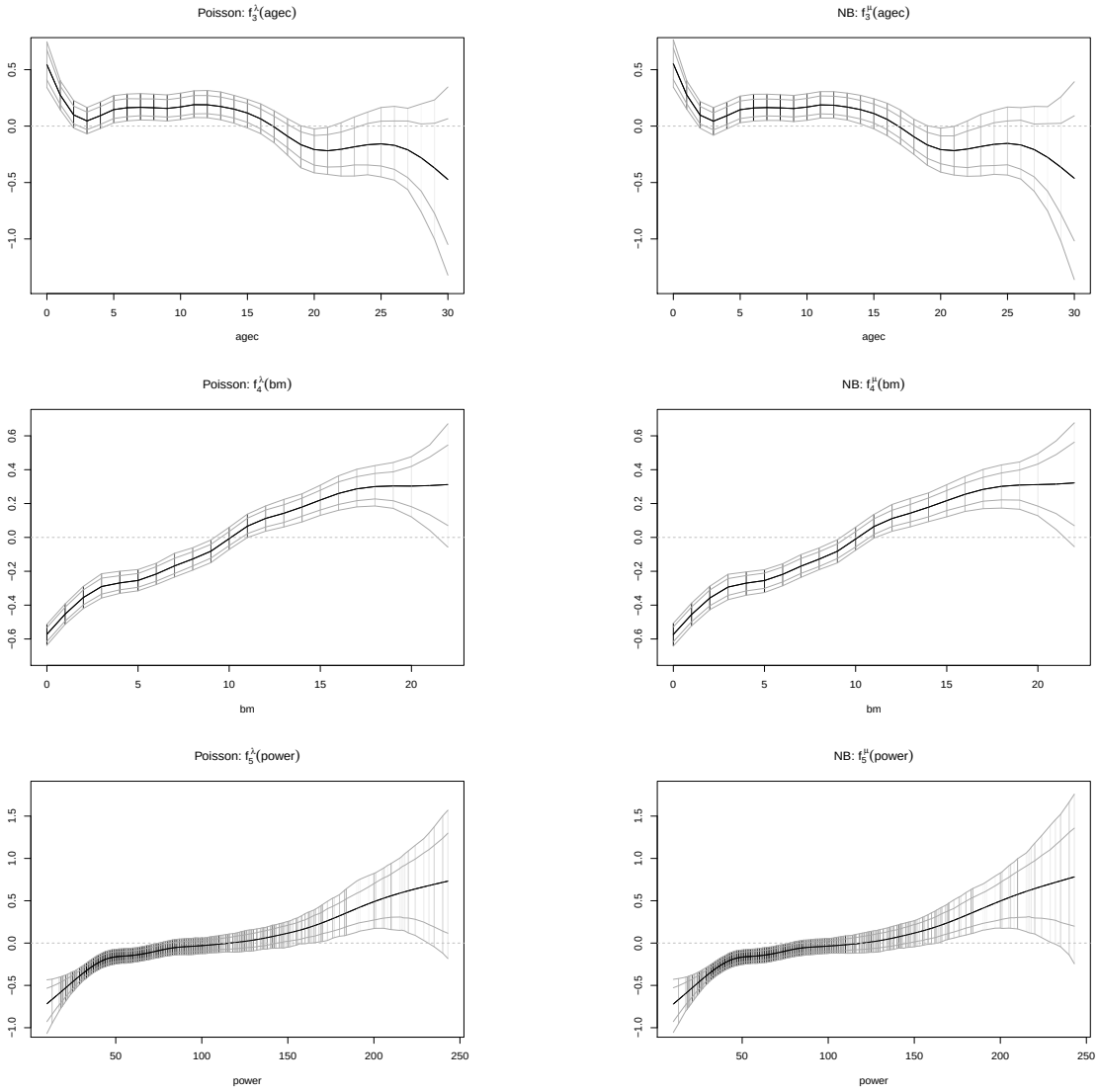

Figure 2: Posterior mean estimates of nonlinear effects (centered around zero) on the expectation parameter together with pointwise 80% and 95% confidence intervals in the Poisson (left panels) and Negative Binomial models (right panels).

significant (i.e. their 80% confidence interval contains 0).

Considering the numerical results discussed so far, the Poisson and Negative Binomial distributions provide very similar answers. The Negative Binomial regression model nevertheless comprises and additional scale parameter δ controlling the degree of residual heterogeneity within risk classes. The estimated score can easily be interpreted from the variance

$$\text{Var}[\Theta_i] = \frac{1}{\delta_i} = \exp(-\eta^{\delta}_{\text{freq}})$$

which measures the cell-specific residual heterogeneity. We see from Table 3 that the residual heterogeneity tends to be smaller for diesel vehicles compared to gasoline ones and higher for the vehicles covered in TPL only or in TPL+limited material damage and theft compared to vehicles with comprehensive coverage. These categories have lower expected claim frequencies but tend to be less homogeneous. Considering the spatial variation in δ displayed in Figure 4, we see that the residual heterogeneity seems to be higher in the North-West part of the country, especially along the North Sea coast. However, most districts have unsignificant spatial effects as shown by the accompanying probability plot. These new findings are particularly interesting as the a posteriori corrections must be more severe for higher residual heterogeneity, i.e. the past claim experience must play a more important role for predicting future claims when $\text{Var}[\Theta_i]$ gets larger. We come back to this issue in Section 3.4.2.

Parameter	mean	standard error	2.5% quantile	median	97.5% quantile
β^{λ}_0 $(const)$	-5.14	0.14	-5.42	-5.14	-4.87
β^{λ}_1 $(fuel)$	-0.09	0.01	-0.11	-0.09	-0.08
β^{λ}_2 $(fleet)$	-0.06	0.02	-0.10	-0.06	-0.02
β^{λ}_3 $(cov1)$	0.07	0.01	0.04	0.07	0.09
β^{λ}_4 $(cov2)$	-0.04	0.01	-0.06	-0.04	-0.02
β^{λ}_5 $(risk)$	0.57	0.02	0.53	0.57	0.61

Table 2: Summary of estimated linear effects for λ in the Poisson model.

3.2 Zero-inflated Poisson regression model

Often, the number of observed zeros in insurance data sets is much larger than under the Poisson assumption. This may be explained by the reluctance of some insured

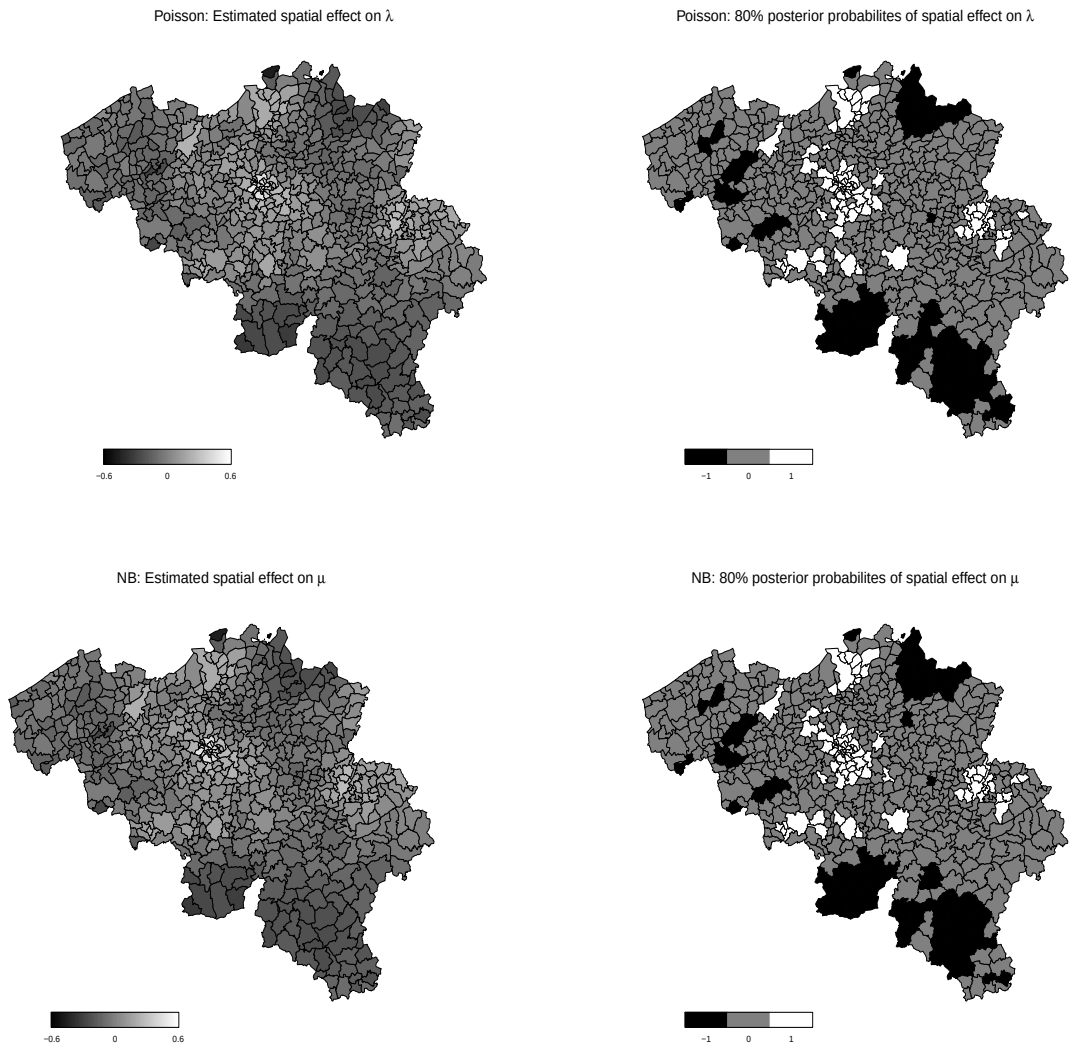

Figure 3: Estimated posterior mean spatial effects f_{spat} (centered around zero) on the mean value of the Poisson (top panels) and Negative Binomial (bottom panels) models together with corresponding probability plots.

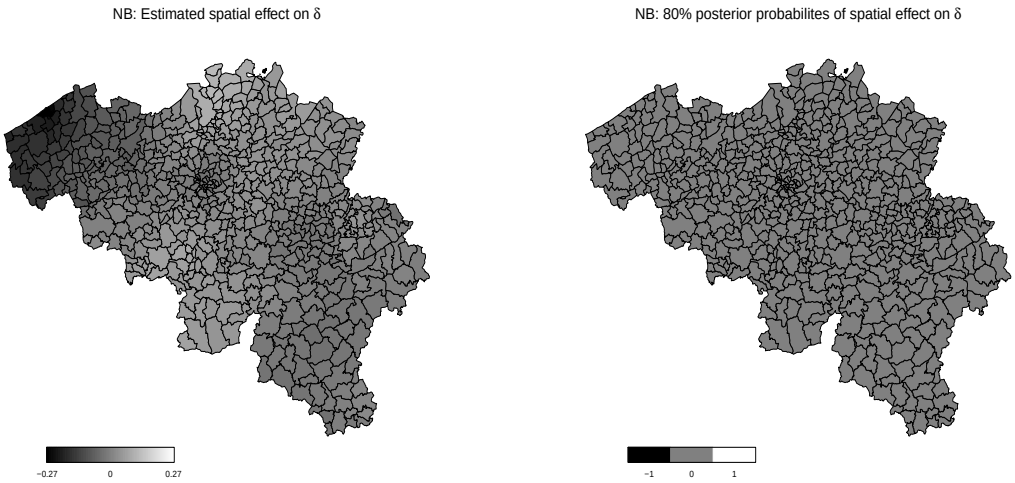

Figure 4: Estimated posterior mean spatial effect f_{spat} (centered around zero) in the scale parameter of the Negative Binomial model, together with the corresponding probability plot.

Parameter	mean	standard error	2.5% quantile	median	97.5% quantile
β_0^μ $(const)$	-5.16	0.15	-5.44	-5.16	-4.86
β_1^μ $(fuel)$	-0.09	0.01	-0.11	-0.09	-0.08
β_2^μ $(fleet)$	-0.06	0.02	-0.11	-0.06	-0.02
β_3^μ $(cov1)$	0.07	0.01	0.04	0.07	0.09
β_4^μ $(cov2)$	-0.04	0.01	-0.06	-0.04	-0.01
β_5^μ $(risk)$	0.58	0.02	0.53	0.58	0.61
β_0^δ $(const)$	-1.63	0.95	-3.46	-1.63	0.37
β_1^δ $(fuel)$	-0.24	0.10	-0.44	-0.25	-0.06
β_3^δ $(cov1)$	-0.39	0.26	-0.91	-0.34	-0.03
β_4^δ $(cov2)$	-0.50	0.29	-1.17	-0.44	-0.06
β_5^δ $(risk)$	0.52	0.16	0.16	0.54	0.81

Table 3: Summary of estimated linear effects for μ and δ in the Negative Binomial model.

drivers to report their accident: Due to bonus-malus mechanisms, some claims are not filed to the company because policyholders think it is cheaper for them to defray the third party (or to pay for their own costs in first party coverages) to avoid premium surcharges. Deductibles also increase the proportion of zeros, since small claims are not reported by insured drivers. For more details, see Denuit et al. (2007).

The Negative Binomial regression model indirectly accounts for this phenomenon as it inflates the probability mass at zero compared to the Poisson law with the same mean. Another, more direct approach consists in using a mixture of two distributions: A degenerated distribution for the zero case combined with a Poisson distribution, giving the zero-inflated Poisson (ZIP) distribution with probability mass function

$$\Pr[N_i = k] = \begin{cases} \pi_i + (1-\pi_i)\exp(-\lambda_i) & \text{for } k = 0 \\ (1-\pi_i)\exp(-\lambda_i)\frac{\lambda_i^k}{k!} & \text{for } k = 1, 2, ... \end{cases} \tag{4}$$

where λ_i is the Poisson parameter and π_i is the additional probability mass at zero. The two first moments of the ZIP distribution are

$$\begin{aligned} \mathrm{E}[N_i] &= (1-\pi_i)\lambda_i \\ \mathrm{Var}[N_i] &= \mathrm{E}[N_i] + \mathrm{E}[N_i](\lambda_i - \mathrm{E}[N_i]) = (1-\pi_i)\lambda_i + \pi_i(1-\pi_i)\lambda_i^2. \end{aligned}$$

We see that both π_i and λ_i enter the expected claim number, which differs from the Poisson and Negative Binomial cases examined before (where the mean was one of the parameters linked to covariates) and makes the interpretation of the results more cumbersome. The variance clearly exceeds the mean so that the ZIP model accounts for the overdispersion generally present in insurance data. Note that the ZIP model can also be seen as a special case of a mixed Poisson distribution obtained with Θ_i equal to 0 or 1 (with respective probabilities π_i and $1-\pi_i$) and conditional mean $\lambda_i\Theta_i$.

Both parameters λ and π are related to structured additive regression predictors constructed from covariates via suitable link functions. The specification for $\lambda_{\text{freq}} = \exp(\eta_{\text{freq}})$ is the same as before with Poisson and Negative Binomial distributions and we follow Baetschmann and Winkelmann (2012) who suggest the complementary log log link for π when the exposure time varies, that is, we specify

$$\pi_{\text{freq}} = \exp(-\exp(\eta^{\pi}_{\text{freq}})). \tag{5}$$

We started from very simple models where π_{freq} is only estimated by a constant. Based on the deviance information criterion (DIC) and significances in effects, further covariates are included step by step. In this way, several models are estimated with different predictor structures, ending up with the predictor structure

$$\eta^{\pi}_{\text{freq}} = f^{\pi}_1(ageph) + f^{\pi}_{\text{spat}}(distr) + (\boldsymbol{z}^{\pi})' \boldsymbol{\beta}^{\pi},$$

for π_{freq}, where $\boldsymbol{z}^{\pi}$ consists of *fuel*, an intercept and the logarithm of the contract period in days.

Estimates of linear effects are given in Table 4. We see that the gasoline vehicles have a lower λ and a higher π so that both effects agree and tend to decrease the expected number of claims. Belonging to a fleet reduces λ but does not impact π. Subscribing to TPL only increases λ compared to when optional coverages are included in the policy. Note also that risk exposure enters both λ and π, increasing λ but decreasing π which conforms with intuition.

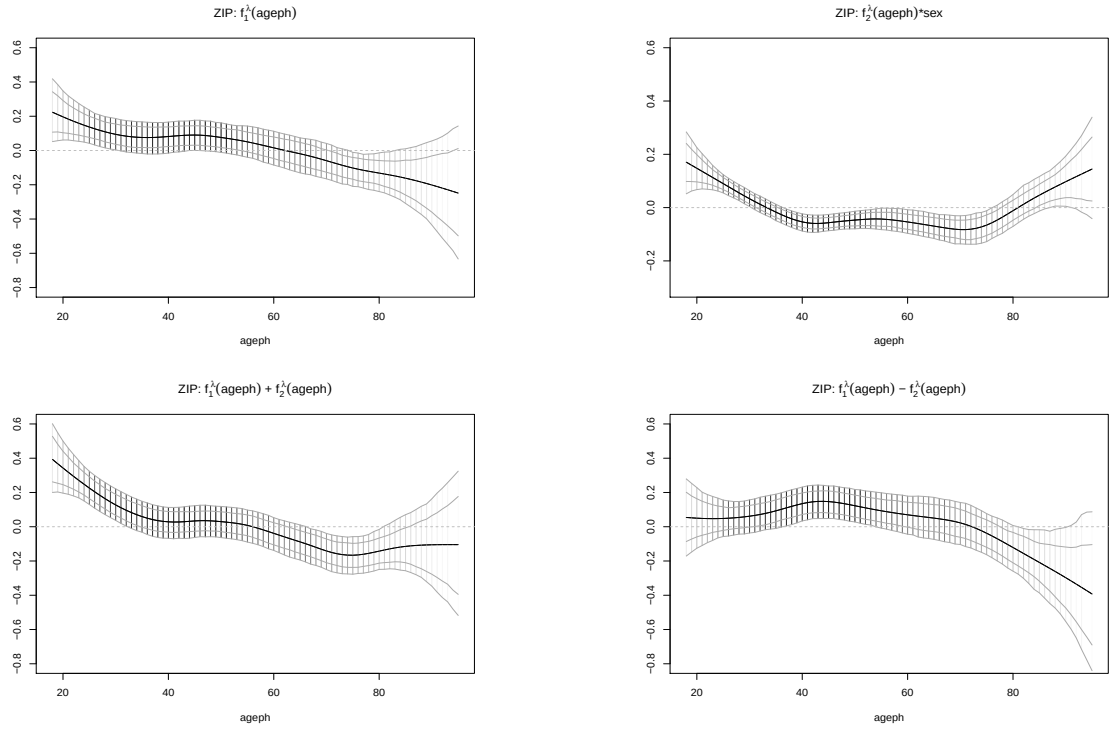

Figure 5: Posterior mean estimates of nonlinear age effects (centered around zero) together with pointwise 80% and 95% confidence intervals in the ZIP model.

In Figures 5 and 6, the estimated nonlinear effects on λ and π are plotted together with 80% and 95% pointwise confidence intervals. As before, vertical stripes indicate

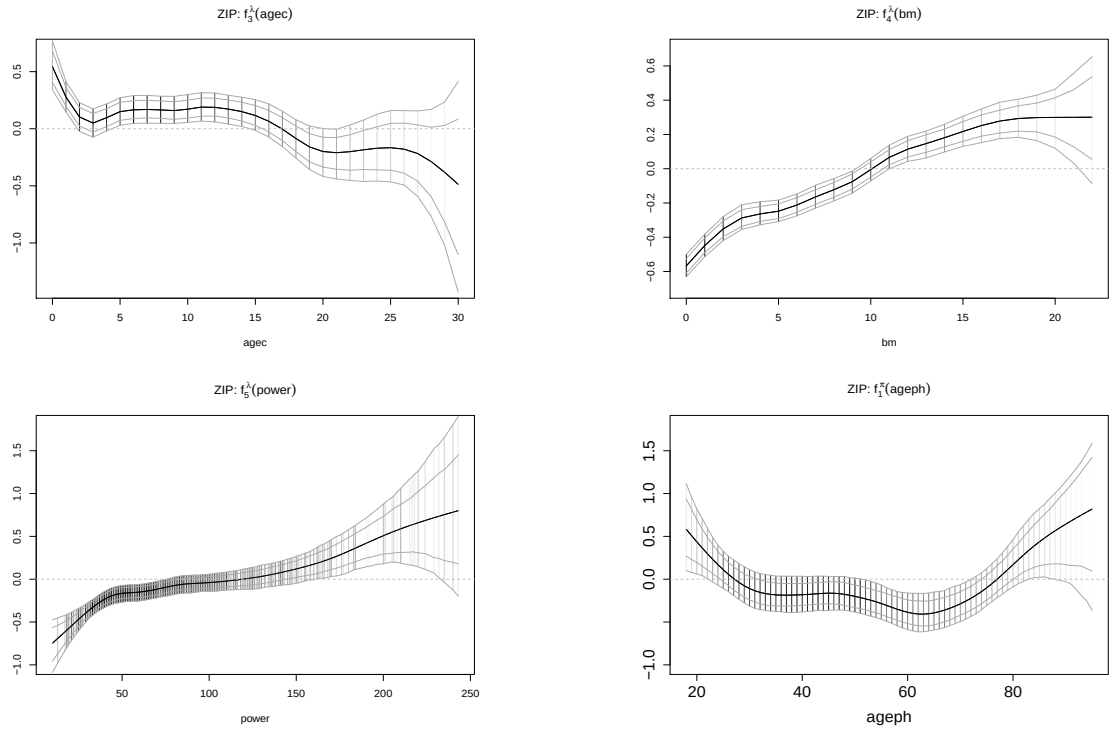

Figure 6: Posterior mean estimates of nonlinear effects (centered around zero) together with pointwise 80% and 95% confidence intervals in the ZIP model.

Parameter	mean	standard error	2.5% quantile	median	97.5% quantile
β_0^λ $(const)$	-2.38	0.39	-3.13	-2.37	-1.62
β_1^λ $(fuel)$	-0.04	0.02	-0.07	-0.04	0.004
β_2^λ $(fleet)$	-0.06	0.02	-0.11	-0.06	- 0.02
β_3^λ $(cov1)$	0.07	0.01	0.04	0.07	0.09
β_4^λ $(cov2)$	-0.04	0.01	-0.06	-0.04	-0.01
β_5^λ $(risk)$	0.14	0.06	0.02	0.14	0.27
β_0^π $(const)$	-3.53	0.50	-4.48	-3.53	-2.53
β_1^π $(fuel)$	-0.11	0.04	-0.20	-0.11	-0.03
β_5^π $(risk)$	0.69	0.09	0.51	0.69	0.86

Table 4: Summary of estimated linear effects for λ and π in the ZIP model.

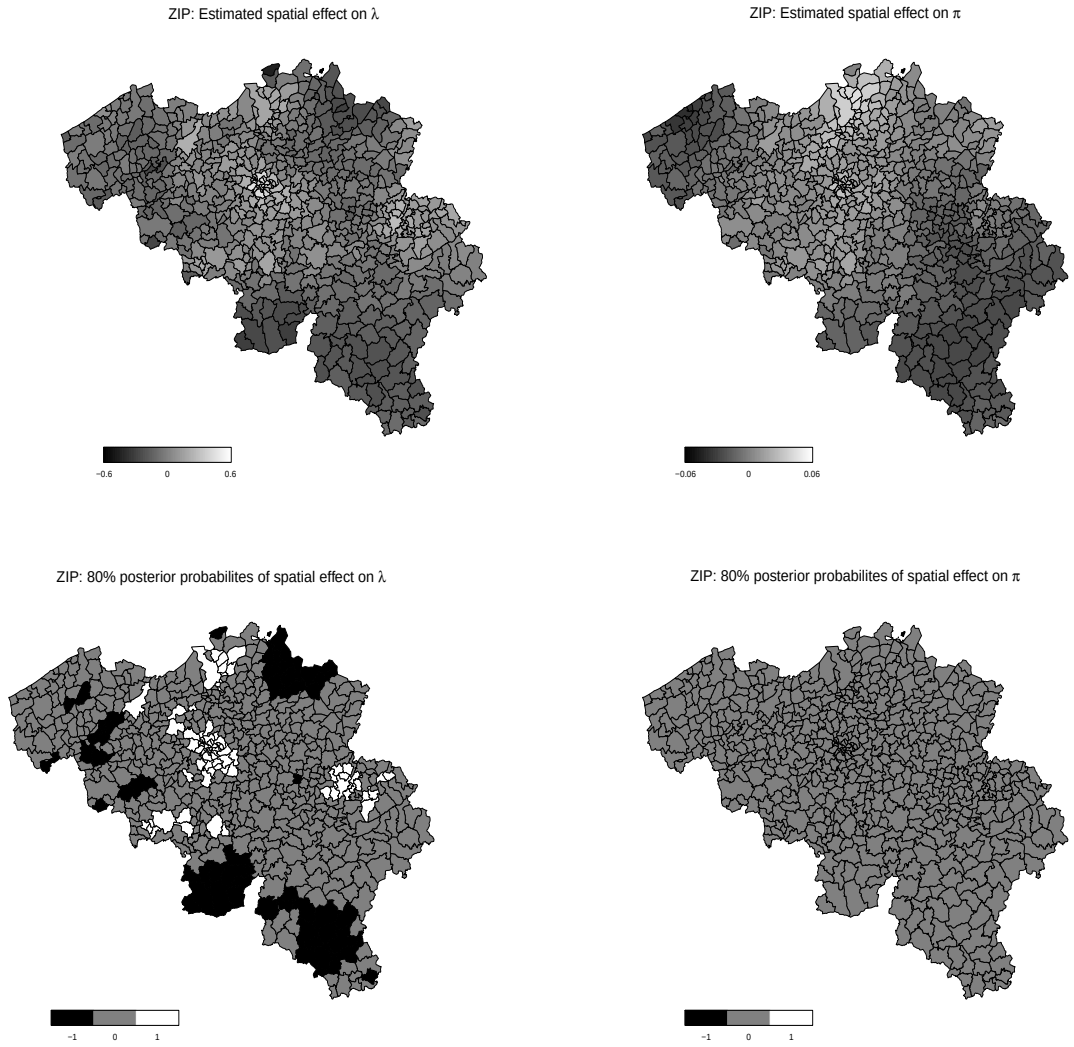

Figure 7: Estimated posterior mean spatial effects in the ZIP model (centered around zero) together with corresponding probability plots.

the relative amount of data of the corresponding covariate values. The estimated nonlinear effects and the spatial effect in λ are in line with those obtained before in the Poisson and Negative Binomial models. The interaction between age and gender is significant and indicates that males younger than 35 cause more accidents than females of the same ages. In contrast, in the ages in between the behavior is converse which can be explained as before. Cars seem to be more dangerous in the first three years, as long as the annual check up organized by the state is not obligatory. The effects of *power* of the car and the policyholder's bonus-malus score are not far from being linear with a kink at 50 for power. As expected, claim frequencies seem to increase with the power of the vehicle and with the level occupied in the bonus-malus scale.

Figure 7 depicts the estimated spatial effects on λ and π. Figure 7 clearly indicates higher claim frequencies for urban areas like Brussels, Antwerp and Liège.

In the zero inflation parameter, we see that young drivers tend to have a lower π whereas drivers around 60 have a significantly higher π. The spatial effect depicted in Figure 7 is much weaker than in λ and basically suggests an inverse relation compared to the findings in the parameter λ indicating that the expected excess of zeros is smaller in urban areas and higher in rural areas. However, most districts have no significant effects on π, as revealed by the accompanying probability plot.

3.3 Comparison of claim frequency models

For specifying all relevant effects in the different predictors, we estimated several models and compared them based on DIC as explained before. Figure 8 shows the normalized (random) quantiles of the final models for the three distributions Poisson, NB and ZIP. Table 5 summarizes the corresponding calculated scores.

Model	Brier Score	Logarithmic Score	Spherical Score	DIC
Poisson	-32,220.61	-61,946.27	145,416.6	123,455
ZIP	**-32,213.16**	-61,803.56	**145,420.2**	**123,226**
NB	-32,215.88	**-61,803.40**	145,419.7	123,230

Table 5: Summarized score contributions from a ten-fold cross validation and DIC from estimations of the whole data set, optimal values appearing printed in bold.

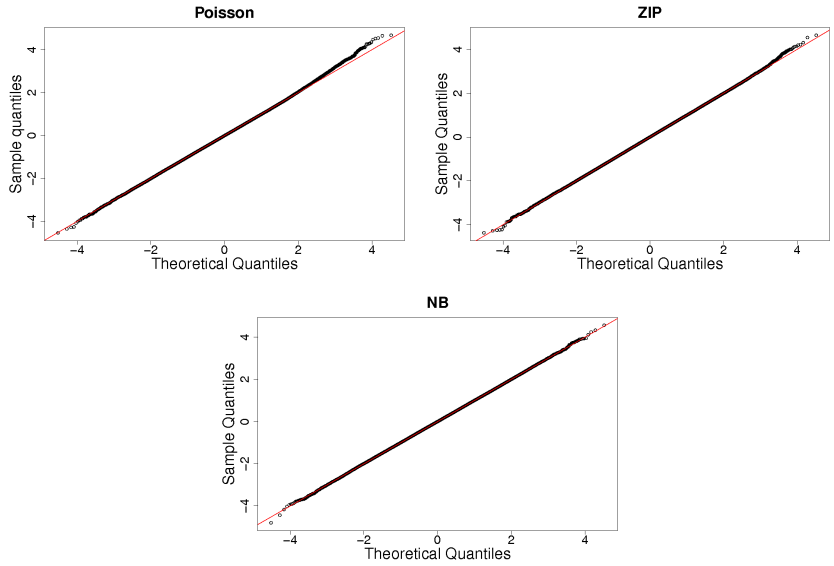

Figure 8: Comparison of quantile residuals in the Poisson, ZIP and NB models.

The residuals of NB and ZIP are very close to the diagonal and indicate a very good fit to the distribution of the claim frequencies while the residuals of the Poisson regression indicate a lack of fit with a clear departure from linearity in the right end. The differences in the computed scores are relatively small but provide further evidence for zero-inflation or overdispersion since the Poisson model always yields the smallest scores. The Brier and spherical score are greatest for the ZIP model while the logarithmic score gives some evidence in favor of the Negative Binomial distribution. In conclusion, NB and ZIP model both give reasonable fits and are clearly preferable to the classical Poisson approach of Denuit and Lang (2004).

3.4 Actuarial applications

3.4.1 No-claim probability

The no-claim probability is a key actuarial indicator for underwriting and annual policy renewal. This probability is given by

$$\Pr[N_i = 0] = \begin{cases} \exp(-\lambda_i) \text{ in the Poisson case} \\ \left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i} \text{ in the NB case} \\ \pi_i + (1-\pi_i)\exp(-\lambda_i) \text{ in the ZIP case.} \end{cases}$$

Not surprisingly, given the high percentage of zero observations in the data basis, the scores in the NB scale parameter δ_i and in the ZIP parameter π_i essentially target

the no-claim probability.

Table 6 gives the no-claim probability predicted by the Poisson, NB and ZIP models for the following three risk profiles:

- high risk: unexperienced male driver aged 18 living in Brussels, driving a brand new, gasoline, powerful car (power 200), occupying the initial level 11 in the bonus-malus scale, subscribing TPL only.
- medium risk: middle-aged male driver aged 40 living in Liège driving a 4-year diesel car with power 100, occupying level 2 in the bonus malus scale, subscribing TPL+limited material damage and theft.
- low risk: retired male driver aged 70 living in the countryside, driving a 15-year gasoline car with power 30, occupying level 0 in the bonus malus scale, subscribing TPL+comprehensive coverage.

No vehicle belongs to a fleet and all policies are assumed to be in force for the whole year.

As we can see from Table 6, the three models provide similar values for the expected number of claims. This is also the case for no-claim probabilities, except for the high risk profile where the Poisson model tends to underestimate this actuarial quantity, making this profile less attractive at renewal compared to NB and ZIP models.

3.4.2 Credibility updates

In the Negative Binomial case, the probability density function of Θ_i given $N_i = k$ is

$$\frac{\exp\left(-\theta_i\left(\delta_i+\mu_i\right)\right)\theta_i^{\delta_i+k-1}}{\int_0^{+\infty}\exp\left(-\xi\left(\delta_i+\mu_i\right)\right)\xi^{\delta_i+k-1}d\xi} = \exp\left(-\theta_i\left(\delta_i+\mu_i\right)\right)\theta_i^{\delta_i+k-1}\frac{\left(\delta_i+\mu_i\right)^{\delta_i+k}}{\Gamma\left(\delta_i+k\right)},$$

so that Θ_i given $N_i = k$ obeys the Gamma distribution with updated parameter values $\delta_i + k$ and $\delta_i + \mu_i$. Therefore, the expected relative risk level revised on the basis of past experience is given by

$$\mathrm{E}[\Theta_i|N_i = k] = \frac{\delta_i + k}{\delta_i + \mu_i}.$$

If the risk profile remains unchanged for the next year, the predicted expected claim number μ_i is replaced with its update $\mu_i\,\mathrm{E}[\Theta_i|N_i = k]$ based on the information

	Low risk	Medium risk	High risk
Poisson, E$[N_i]$	0.0433	0.1664	1.8913
Confidence Interval 95%	(0.0327,0.056)	(0.1366,0.1982)	(1.0796,3.0574)
Confidence Interval 80%	(0.0359,0.0512)	(0.1457,0.1871)	(1.2860,2.6021)
NB, E$[N_i]$	0.0438	0.1667	1.8738
Confidence Interval 95%	(0.0327,0.0581)	(0.1386,0.1993)	(1.0335,3.0189)
Confidence Interval 80%	(0.0362,0.0519)	(0.1478,0.1869)	(1.2965,2.539)
ZIP, E$[N_i]$	0.0435	0.1648	1.8239
Confidence Interval 95%	(0.0327,0.0578)	(0.1357,0.1974)	(0.9892,2.9565)
Confidence Interval 80%	(0.0359,0.0516)	(0.1441,0.1856)	(1.2269,2.4853)
Poisson, Pr$[N_i = 0]$	0.9576	0.8468	0.1699
Confidence Interval 95%	(0.9456,0.9678)	(0.8202,0.8723)	(0.047,0.3397)
Confidence Interval 80%	(0.9501,0.9647)	(0.8294,0.8644)	(0.0741,0.2764)
NB, Pr$[N_i = 0]$	0.9574	0.8501	0.2800
Confidence Interval 95%	(0.9438,0.968)	(0.8241,0.8734)	(0.1509,0.4317)
Confidence Interval 80%	(0.9497,0.9645)	(0.8339,0.8657)	(0.1909,0.3773)
ZIP, Pr$[N_i = 0]$	0.9579	0.8510	0.2207
Confidence Interval 95%	(0.9449,0.9683)	(0.8245,0.8748)	(0.0898,0.3956)
Confidence Interval 80%	(0.9502,0.9652)	(0.8343,0.8678)	(0.1244,0.3200)

Table 6: Predictions of expected claim numbers and no-claim probabilities for different risk profiles according to the Poisson, NB and ZIP models.

contained in the number k of claims filed by policyholder i. The theoretical bonus-malus coefficients $\mathrm{E}[\Theta_i|N_i = k]$ exhibit some well-known features (see, e.g., Denuit et al., 2007):

- the a posteriori corrections become more severe when the residual heterogeneity, measured by $\mathrm{Var}[\Theta_i]$, increases.
- considering two policyholders (numbered i_1 and i_2) such that i_1 is a priori a better driver than i_2, that is, $\mu_{i_1} < \mu_{i_2}$,
  - if these policyholders do not report any claim (i.e., $N_{i_1} = N_{i_2} = 0$) then the a priori worse driver receives more discount provided $\delta_{i_1} = \delta_{i_2}$.
  - if these policyholders report $k \geq 1$ claims (i.e., $N_{i_1} = N_{i_2} = k$) then the penalty for the a priori bad driver is less severe than for the good one provided $\delta_{i_1} = \delta_{i_2}$.

Compared to this classical setting, we have here a new effect coming from the respective values of δ_{i_1} and δ_{i_2} which may interfere with the preceding discussion.

Table 7 gives the theoretical bonus malus correction $\mathrm{E}[\Theta_i|N_i = k]$ for the three risk profiles according to the value of k and Table 8 displays the corresponding revised expected claim frequencies. Considering the three risk profiles defined previously, we know from Table 3 that δ tends to be larger for diesel vehicles compared to gasoline ones and smaller for the vehicles covered in TPL only or in TPL+limited material damage and theft compared to vehicles with comprehensive coverage. The spatial variation in δ displayed in Figure 4 has no significant impact on the three risk profiles we consider here. Including the intercept and the exposure-to-risk gives estimated values of δ equal to $\exp(2.09)$ for the low risk profile, $\exp(1.18)$ for the medium risk profile, and $\exp(0.81)$ for the high risk profile. Hence, the past claim experience plays a more important role as the quality of the risk deteriorates, which conforms with actuarial intuition.

If the low risk profile does not report any accident during the first year, we see from Table 7 that the revision consists in multiplying the expected claim frequency by 99.27%. This modest decrease is to be compared with the corresponding values 95.09% for the medium risk profile and 52.97% for the high risk profile. The discounts

Low risk			
k	$\mathrm{E}[\Theta_i \mid N_i = k]$	Confidence Interval 95%	Confidence Interval 80%
0	0.9927	(0.9815,0.9993)	(0.9866,0.9984)
1	1.1617	(1.0172,1.4034)	(1.0349,1.293)
2	1.3307	(1.0351,1.8257)	(1.0715,1.6001)
3	1.4997	(1.0531,2.2479)	(1.1081,1.905)
4	1.6688	(1.0711,2.6702)	(1.1448,2.21)
5	1.8378	(1.089,3.0925)	(1.1814,2.5165)
Medium risk			
k	$\mathrm{E}[\Theta_i \mid N_i = k]$	Confidence Interval 95%	Confidence Interval 80%
0	0.9509	(0.923,0.9733)	(0.9348,0.9669)
1	1.2459	(1.1387,1.3773)	(1.1717,1.327)
2	1.5409	(1.305,1.8292)	(1.3752,1.7167)
3	1.8359	(1.4703,2.2851)	(1.5786,2.105)
4	2.131	(1.6356,2.7356)	(1.7831,2.4988)
5	2.426	(1.8009,3.1861)	(1.9874,2.8907)
High risk			
k	$\mathrm{E}[\Theta_i \mid N_i = k]$	Confidence Interval 95%	Confidence Interval 80%
0	0.5297	(0.338,0.6975)	(0.4079,0.6455)
1	0.7889	(0.5824,0.991)	(0.6566,0.9174)
2	1.048	(0.796,1.3009)	(0.8813,1.2135)
3	1.3072	(0.9963,1.6561)	(1.0974,1.5229)
4	1.5664	(1.1922,2.0188)	(1.3166,1.8275)
5	1.8256	(1.3854,2.3749)	(1.5235,2.1403)

Table 7: Theoretical bonus malus correction $\mathrm{E}[\Theta_i \mid N_i = k]$ for different risk profiles.

Low risk			
k	$\mu_i \, \mathrm{E}[\Theta_i \vert N_i = k]$	Confidence Interval 95%	Confidence Interval 80%
0	0.0434	(0.0324,0.0577)	(0.036,0.0514)
1	0.0508	(0.0367,0.0694)	(0.0411,0.0614)
2	0.0581	(0.0394,0.0867)	(0.0448,0.0733)
3	0.0654	(0.0407,0.1052)	(0.0473,0.0866)
4	0.0728	(0.0415,0.1229)	(0.0491,0.0993)
5	0.0801	(0.0427,0.1392)	(0.0513,0.1129)
Medium risk			
k	$\mu_i \, \mathrm{E}[\Theta_i \vert N_i = k]$	Confidence Interval 95%	Confidence Interval 80%
0	0.1584	(0.132,0.188)	(0.1408,0.1769)
1	0.2075	(0.1701,0.254)	(0.1816,0.2352)
2	0.2567	(0.2003,0.3267)	(0.217,0.2977)
3	0.3058	(0.2277,0.4036)	(0.251,0.3612)
4	0.3549	(0.2575,0.4781)	(0.286,0.4261)
5	0.4041	(0.2868,0.558)	(0.3192,0.4905)
High risk			
k	$\mu_i \, \mathrm{E}[\Theta_i \vert N_i = k]$	Confidence Interval 95%	Confidence Interval 80%
0	0.958	(0.6408,1.3239)	(0.7384,1.1807)
1	1.4283	(1.0242,1.8487)	(1.1769,1.6907)
2	1.8986	(1.3382,2.3944)	(1.563,2.2394)
3	2.3689	(1.6642,3.0078)	(1.922,2.8068)
4	2.8392	(1.9959,3.6243)	(2.2817,3.3782)
5	3.3095	(2.2739,4.2618)	(2.6371,3.9529)

Table 8: Resulting revised expected claim frequency $\mu_i \, \mathrm{E}[\Theta_i \vert N_i = k]$ for different risk profiles.

awarded to policyholders who do not report any accident to the insurance company are thus increasing with the a priori annual expected claim frequency: The more claims are expected by the insurance company on the basis of observable characteristics, the higher the discount in case no claims are reported. This classical effect is amplified here by the values of δ which put more weight on past experience for higher risk profiles. Note however from Table 8 that the revised expected claim frequency is still larger the higher the risk profile.

Now, considering the penalty in case one claim is reported, we see that when the low risk profile reports one claim, the expected claim frequency is multiplied by 116.17%. The corresponding values for the medium risk profile and for the high risk profile are 124.59% and 78.89%, respectively. The penalties in case an accident is reported to the company are thus decreasing with the a priori annual expected claim frequencies. Compared to the classical analyses where reporting a claim typically entails a penalty comprised between 50 and 75%, we see that the present system based on accurate risk classification appears to be less severe, and thus easier to implement in practice. Moreover, the high risk profile still receives a discount in case only one claim is reported, the penalty appearing when 2 or more claims are filed. Similar comments apply when more claims are reported, i.e. for higher values of k. The corresponding revised expected claim frequencies are displayed in Table 8.

Let us now consider the ZIP model. In this case, Θ_i is Bernoulli distributed with mean $1-\pi_i$ and

$$\Pr[\Theta_i = 0 | N_i = k_i] = \begin{cases} \frac{\Pr[\Theta_i=0]}{\Pr[N_i=0]} = \frac{\pi_i}{\pi_i+(1-\pi_i)\exp(-\lambda_i)} \text{ if } k_i = 0 \\ 0 \text{ if } k_i \geq 1. \end{cases}$$

If $N_i = 0$ then the number of claims for next year is still ZIP with increased probability mass at zero given by $\Pr[\Theta_i = 0|N_i = 0]$. If $N_i \geq 1$ then the number of claims for next year becomes Poisson distributed with mean λ_i and no further re-evaluation based on claim experience is needed. As these re-evaluations are rather crude, the ZIP model does not provide accurate enough credibility updates for claim frequencies and the NB model is preferable.

4 Modelling Claim Sizes: Zero-Adjusted Regression

4.1 From individual to aggregate claim sizes

Large claims generally affect liability coverages. These major claim sizes require a separate analysis as no simple standard parametric model seems to emerge as providing an acceptable fit to both small and large claims. Let us nevertheless mention the composite models considered, e.g., in Pigeon and Denuit (2011) which can accomodate a mix of small and large claims but for which no regression analysis is available, yet. Here, we restrict the analysis to observations with total claim size less than exp(15) which corresponds to a threshold approximately equal to EUR 80,000 (or 3.2 millions Belgian francs) in the spirit of Cebrian et al. (2003).

With the noticeable exception of Jørgensen and Paes de Souza (1994), the vast majority of actuarial analyses of the pure premium so far have examined frequencies and severities separately. The Tweedie GLM used by these authors is however quite restrictive as the no-claim probability is not allowed to depend on covariates. Heller et al. (2007) extended this approach to Poisson, ZIP or NB claim frequencies combined with Gamma or Inverse-Gaussian severities. The Tweedie distribution appears as a special case corresponding to the Poisson-Gamma choice. In this section, we target the total annual claim amount using zero-adjusted regression models. This provides an alternative to these mixed Poisson compound models.

Before the Poisson regression became popular among actuaries, claims data were often analyzed using logistic regression; see, e.g., Beirlant et al. (1991). Zero-adjusted models are closely related to this approach but avoid the two-step analysis. As the likelihood factors in two parts, one with the probability of reporting no claim and another one with positive claim amounts, the two strategies lead to very similar results. But analyzing the total claim amount by means of a single zero-adjusted regression model allows the actuary to get accurate confidence intervals for several key risk indicators, such as the expected claim cost, for instance. Confidence intervals are more difficult to derive in the two-step approach because estimation errors from the zero part and from the continuous part need to be combined. Moreover, it is not

that easy to obtain global DIC values when performing separate analyzes.

4.2 Zero-adjusted models

Zero-adjusted (ZA) models are discrete-continuous distributions with a probability mass at zero and a continuous component which can be any parametric distribution as long as first and second derivative of the log-likelihood can be computed. Therefore, these models additionally allow to estimate the probability for positive claim sizes. The idea of using such models to describe insurance data goes back to Heller et al. (2006) who applied the zero-adjusted Inverse-Gaussian distribution as a model for claim sizes, including zero claims. Whereas these authors confined to linear effects of the covariates (a logit-linear model for the occurrence of a claim and log-linear models for the expected claim size and for its dispersion when at least one claim is reported), we allow here for nonlinear effects of the explanatory variables, including spatial effects.

Consider the total claim sizes Y_i and covariate vectors as in the previous section. The distribution of Y_i is defined by $\Pr[Y_i = 0] = 1 - \pi_i$ and, given $Y_i > 0$ (which happens with probability π_i) Y_i has probability density function $g(\cdot)$ corresponding to some continuous distribution with support in $\mathbb{R}^+$. Here, $1 - \pi_i$ is the no-claim probability, not to be confused with the parameter π_i in the ZIP model.

This representation corresponds to the individual model of risk theory. See, e.g., Chapter 2 in Kaas et al. (2008) for an introduction. In this model, the total claim cost is decomposed into the product of an indicator for the event "the policy produces at least one claim during the reference period" and a positive random variable representing the total claim amount produced by the policy when at least one claim has been filed. This exactly corresponds to the construction of the ZA models. Numerous powerful actuarial techniques have been developed for the individual model, which are thus directly applicable to the ZA modeling output and facilitates the actuarial analysis.

4.3 Numerical illustration

To model the total claim size, we used the LogNormal (LN), Gamma (GA) and Inverse-Gaussian (IG) distribution supplemented with a probability mass at zero. This allows for the simultaneous estimation of characteristics of the claim size distribution and the no-claim probability. Explanatory variables are included in the three parameters, the probability mass at zero $1 - \pi$ as well as both parameters of these three continuous distributions. We use log link functions for positive parameters and the identity link for the real-valued parameter of the LogNormal distribution. For the claim probability π, the complementary log log link has been used. As for claim frequencies, all models are compared graphically applying normalized quantile residuals and by proper scoring rules. Predictor specifications were chosen by DIC together with significances of the effects as described before. The spatial effect is modeled in the spirit of Lang et al. (2013) consisting of a Markov random field and an additional independent and identically distributed random effect.

Figure 9 shows the normalized quantile residuals computed from the subset of observations with positive claims. The residuals indicate that LN and IG are better assumptions than the GA distribution since the residuals of the former two are very close to the diagonal while the sample quantiles of the GA distribution greater than 3 are higher than the theoretical quantiles. This indicates that the GA model may only be appropriate for modeling small claim sizes.

In Table 9, the computed scores and the DIC underline the preferences for applying either the ZALN model (which would be chosen by DIC as well as the logarithmic, quadratic and spherical score) or the ZAIG (with highest CRPS).

Model	Brier Score	Logarithmic Score	Spherical Score	CRPS	DIC
ZALN	**105,610.7**	**-171,614.4**	**125,911.6**	-134,845.7	**90,202**
ZAIG	103,211.7	-172,983.5	124,908.4	**-127,940.1**	90,323
ZAGA	104,931.5	-174,303.7	125,533.1	-148,244.3	94,064

Table 9: Summarized scores from a ten-fold cross validation and DIC obtained from estimations of the whole data set, optimal values appearing printed in bold.

For further illustration of this score, we performed a quantile decomposition and plotted the scores against the quantiles in Figure 10. It gets obvious that for quantiles

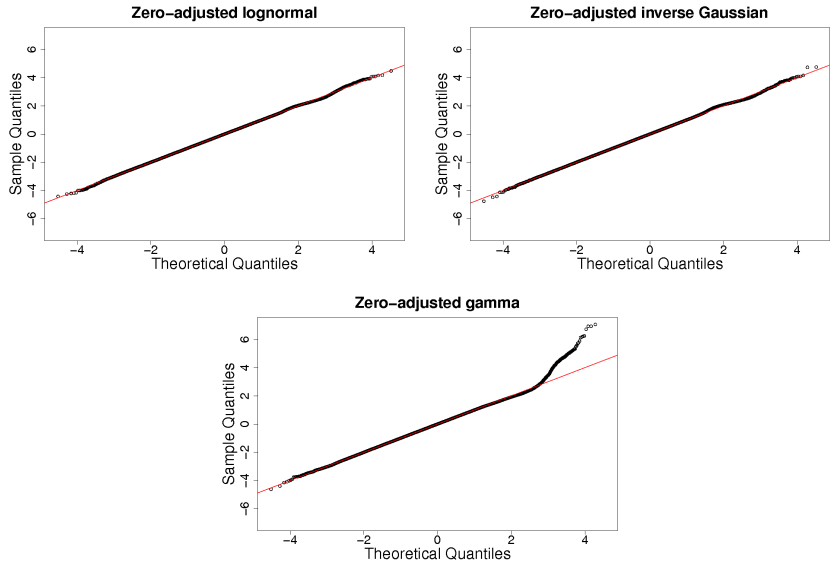

Figure 9: Comparison of quantile residuals in the LogNormal, Inverse Gaussian and Gamma model.

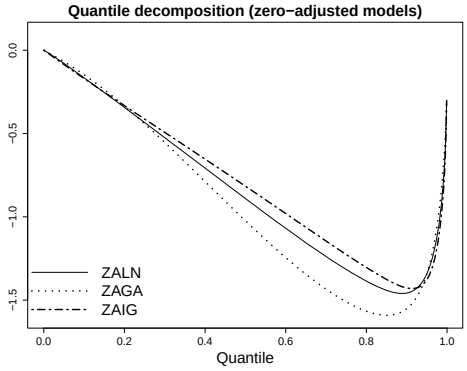

Figure 10: Quantile decomposition of the CRPS.

smaller than 0.2 and greater than 0.9 the scores of ZALN and ZAIG are very similar, while in between the ZAIG model delivers smaller losses. In a nutshell it can be concluded that both the ZALN and ZAIG distributions seem to be very promising candidates for the aggregate claim size distributions. For illustration and interpretation of effects we choose the ZALN model in the remainder of this section.

Given that the total claim size Y_i is positive, we get

$$\begin{aligned} \mathrm{E}[Y_i|Y_i>0] &= \exp\left(\mu_i+\frac{\sigma_i^2}{2}\right) \\ \mathrm{Var}[Y_i|Y_i>0] &= \exp\left(2\mu_i+\sigma_i^2\right)\left(\exp(\sigma_i^2)-1\right). \end{aligned}$$

Here, $\exp(\mu_i)$ corresponds to the median claim cost when at least one claim has been reported.

Based on the DIC we specified the following predictor structures for the location parameter $\mu_{\text{costs}} = \eta^{\mu}_{\text{costs}} \in \mathbb{R}$ and scale parameter $\sigma^2_{\text{costs}} = \exp\left(\eta^{\sigma^2}_{\text{costs}}\right) > 0$. Here,

$$\eta^{\mu}_{\text{costs}} = f^{\mu}_1(ageph) + sex\, f^{\mu}_2(ageph) + f^{\mu}_3(agec) + f^{\mu}_4(bm) + f^{\mu}_{\text{spat}}(distr) + (\boldsymbol{z}^{\mu})'\boldsymbol{\beta}^{\mu},$$

where $\boldsymbol{z}^{\mu}$ consists of *fleet*, *coverage* and an overall constant. For σ^2 we get

$$\eta^{\sigma^2}_{\text{costs}} = f^{\sigma^2}_1(ageph) + f^{\sigma^2}_{\text{spat}}(distr) + \left(\boldsymbol{z}^{\sigma^2}\right)'\boldsymbol{\beta}^{\sigma^2},$$

where $\boldsymbol{z}^{\sigma^2}$ consists of *fleet*, *coverage* and an intercept.

For π_{costs}, we follow again Baetschmann and Winkelmann (2012) since there are several policyholders with a contract shorter than one year and we choose

$$\pi_{\text{costs}} = 1 - \exp(-\exp(\eta^{\pi}_{\text{costs}})). \tag{6}$$

It is instructive to compare the specifications (5) and (6): In the ZIP case π models the probability of observing structural zeros (compare Section 3.2) while in the ZALN model π stands for the probability of observing a positive claim size, i.e. $\pi_i = \Pr[Y_i > 0]$. Therefore, equations (5) and (6) are reasonable since an increasing contract period, included in the predictor, reduces π in the ZIP model and raises the probability of observing a positive claim connected with a positive amount. Based on this specification, we found as best predictor specification

$$\eta^{\pi}_{\text{costs}} = f^{\pi}_1(ageph) + sex\, f^{\pi}_2(ageph) + f^{\pi}_3(agec) + f^{\pi}_4(power) + f^{\pi}_5(bm) + f^{\pi}_{\text{spat}}(distr) + (\boldsymbol{z}^{\pi})'\boldsymbol{\beta}^{\pi},$$

where $\boldsymbol{z}^{\pi}$ consists of *fuel*, *fleet*, *coverage* an intercept and the logarithm of the contract period in days.

Figure 14 depicts the estimated spatial effects on μ, σ^2 and π. Estimates of nonlinear effects are shown in Figures 11, 12 and 13 while Table 10 summarizes all linear effects. Considering the effects of the categorical covariates on the claim probability π, we see from Table 10 that driving a diesel vehicle decreases the no-claim probability whereas belonging to a fleet increases this probability. Also, extending the coverage increases the no-claim probability. As expected, all estimated nonlinear effects on π are quite similar to the ones we estimated in the ZIP model, so that the same comments apply here, mutatis mutandis. This can be explained by the large number of policyholders who either report no or only one claim. For the spatial effects, we can summarize

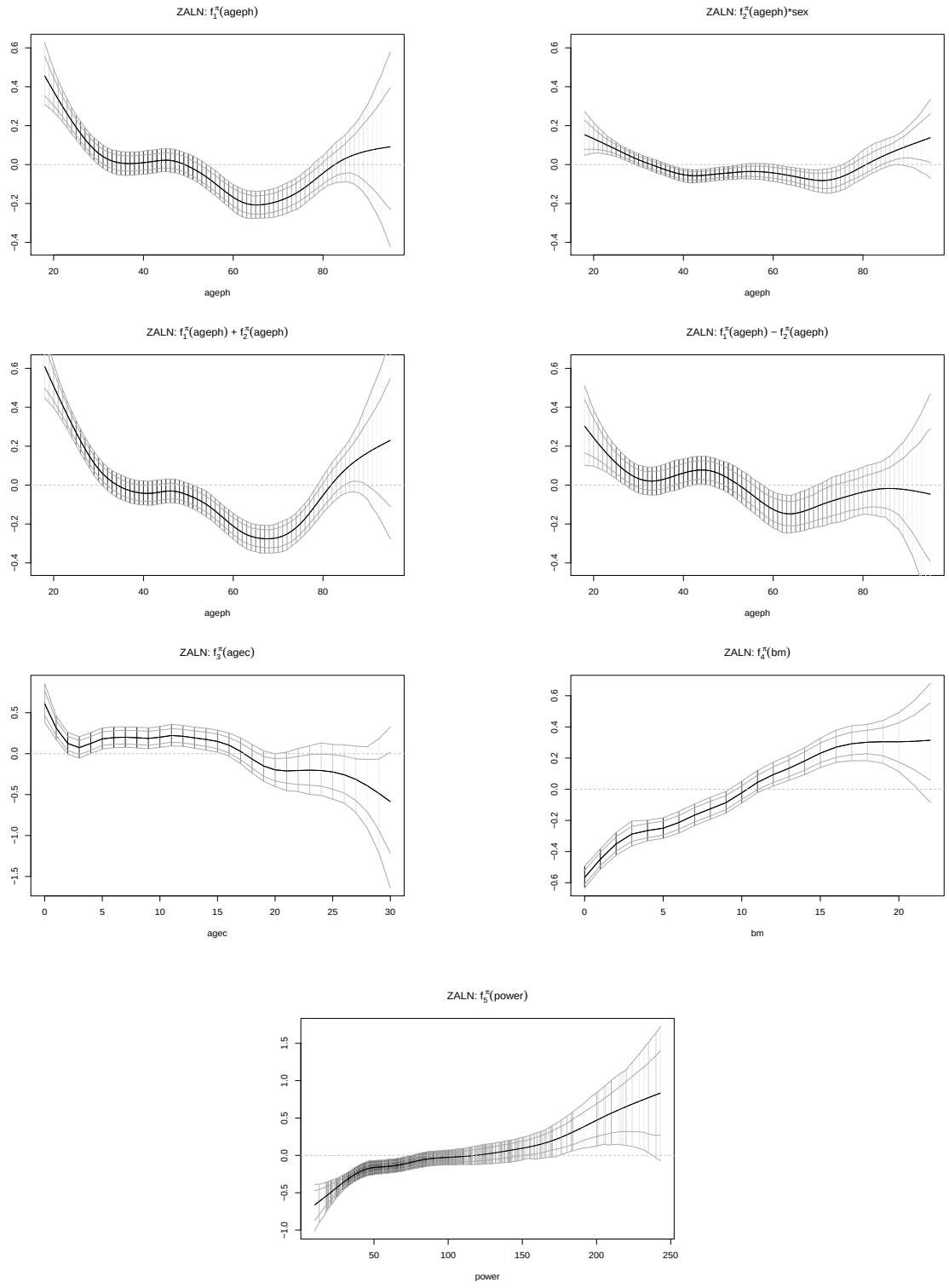

Figure 11: Posterior mean estimates of nonlinear effects on π (centered around zero) together with pointwise 80% and 95% confidence intervals in the ZALN model.

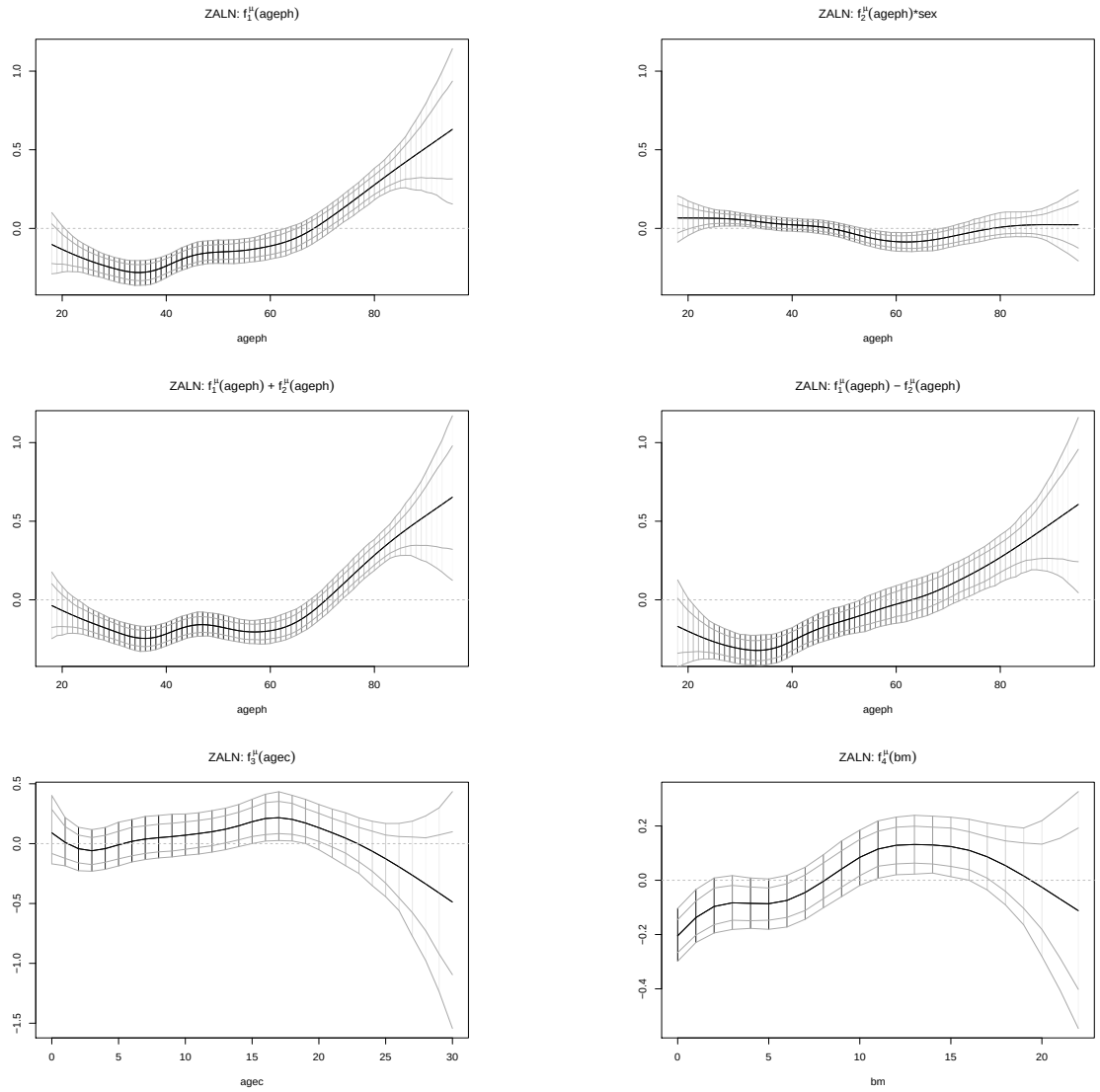

Figure 12: Posterior mean estimates of nonlinear effects on μ (centered around zero) together with pointwise 80% and 95% confidence intervals in the ZALN model

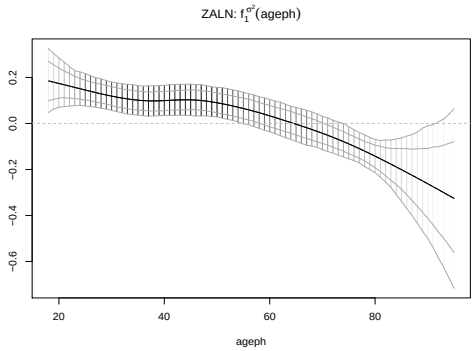

Figure 13: Posterior mean estimates of nonlinear effects on σ^2 (centered around zero) together with pointwise 80% and 95% confidence intervals in the ZALN model.

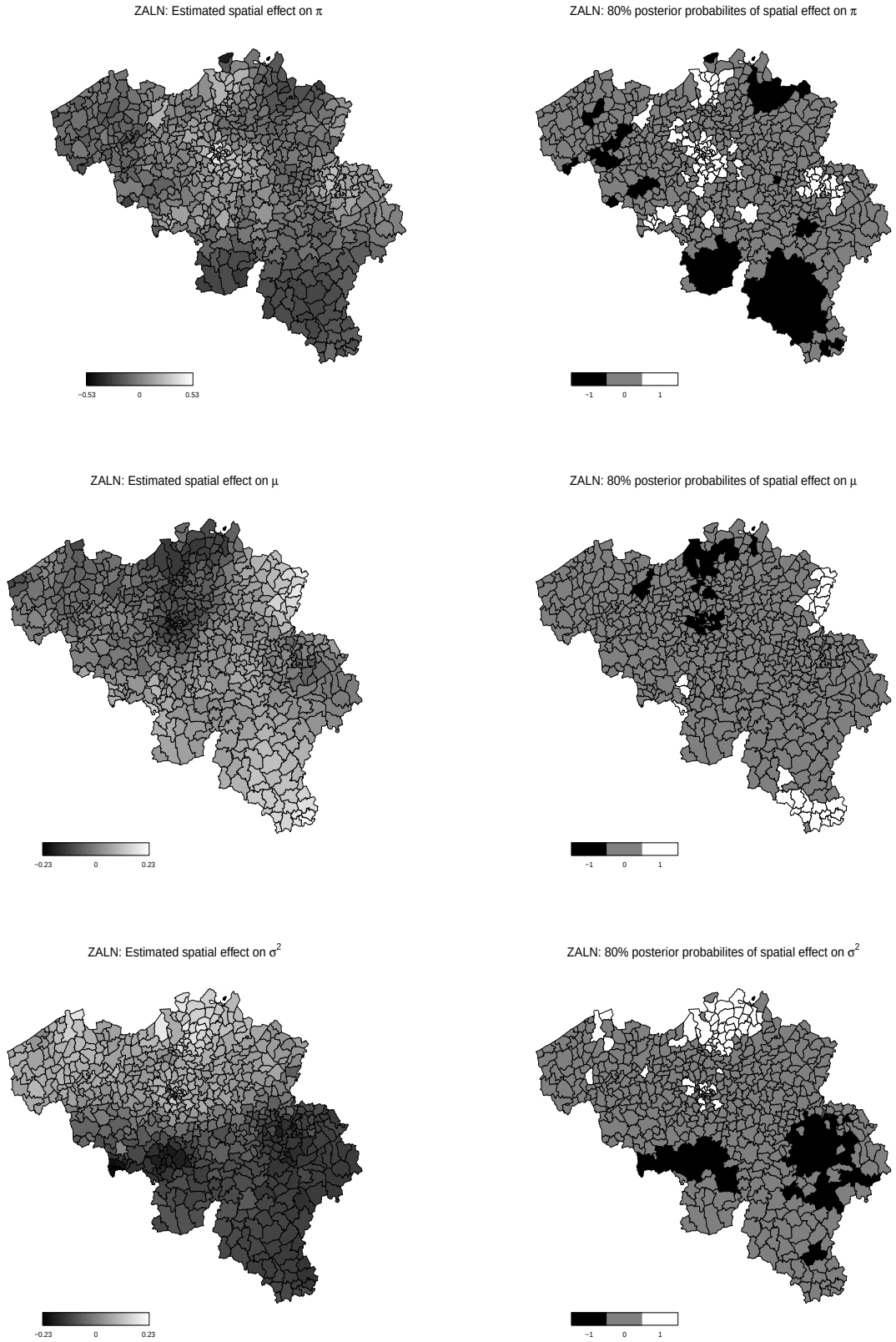

Figure 14: Estimated posterior mean complete spatial effects in the ZALN model (centered around zero) together with corresponding probability plots.

Parameter	mean	standard error	2.5% quantile	median	97.5% quantile
β_0^π $(const)$	-5.25	0.15	-5.55	-5.25	-4.94
β_1^π $(fuel)$	-0.10	0.01	-0.12	-0.10	-0.08
β_2^π $(fleet)$	-0.06	0.02	-0.10	-0.06	-0.02
β_3^π $(coverage1)$	0.06	0.01	0.03	0.06	0.08
β_4^π $(coverage2)$	-0.04	0.01	-0.07	-0.04	-0.02
β_5^π $(risk)$	0.58	0.02	0.54	0.58	0.63
β_0^μ $(const)$	10.20	9.31	9.99	10.20	10.40
β_1^μ $(fleet)$	0.07	0.03	0.02	0.07	0.13
β_2^μ $(coverage1)$	-0.02	0.02	-0.05	-0.01	0.02
β_3^μ $(coverage2)$	-0.15	0.02	-0.18	-0.15	-0.11
$\beta_0^{\sigma^2}$ $(const)$	0.61	0.04	0.52	0.61	0.69
$\beta_1^{\sigma^2}$ $(fleet)$	-0.14	0.03	-0.20	-0.14	-0.07
$\beta_2^{\sigma^2}$ $(coverage1)$	-0.05	0.01	-0.08	-0.05	-0.02
$\beta_3^{\sigma^2}$ $(coverage2)$	-0.12	0.02	-0.15	-0.12	-0.08

Table 10: Summary of estimated linear effects for π, μ and σ^2 in the ZALN model.

that the effect in π is also similar to the effect on λ with higher chances of reporting claims in urban areas. Considering the linear effects on μ and σ^2, we see from Table 10 that belonging to a fleet increases μ but decreases σ^2, increasing the median claim severity but leaving the expected claim cost almost unchanged. Subscribing TPL only or limited material damage decreases μ and σ^2 compared to comprehensive coverage (remember that we only deal here with the TPL guarantee). Looking at the estimated nonlinear effects on μ, we see that when claims are reported, the median cost is higher for older drivers. Taking into account the monotonically decreasing estimated effect of *ageph* in σ^2 sowewhat tempers this effect when expected claim size is considered but the overall effect is still that older drivers report more expensive annual claims. The effect of the ancientness of the car is much less significant. Also, the level occupied in the bonus-malus scale does not seem to bring a lot of information about μ, which can be easily understood as movements in this scale only result from the number of claims and not their size. It is interesting to discuss the spatial effects on μ and σ^2 as they reveal new characteristics of claim sizes. The effect on μ relates to the smaller expected claim cost in Brussels and Antwerp areas (major cities with low average speed because of traffic congestion) compared to rural areas in Limburg and Ardenne

(with weaker traffic intensity and higher average speed resulting in more expensive accidents when they occur). This main effect is refined by the spatial structure in σ^2, somewhat increasing the expected claim severities in some urban districts and decreasing the average claim cost in Wallonia compared to Flanders and Brussels. This might be explained by the higher living standards in Flanders and Brussels where more expensive cars may be driven, increasing the average claim size.

	Low risk	Medium risk	High risk
ZALN, $\mathrm{E}[Y_i]$	3965.87	5699.8	88031.46
Confidence Interval 95%	(2620.23,5833)	(4246.14,7566.32)	(49504.56,150846.18)
Confidence Interval 80%	(2969.85,5044.94)	(4674.02,6758.25)	(58426.16,121233.93)
ZALN, $\pi_i = \Pr[Y_i > 0]$	0.0425	0.1522	0.7801
Confidence Interval 95%	(0.032,0.0548)	(0.1274,0.1798)	(0.5867,0.9383)
Confidence Interval 80%	(0.0351,0.0501)	(0.1353,0.1706)	(0.6617,0.8896)
ZALN, $\mu_i = \mathrm{E}[\log(Y_i)\|Y_i > 0]$	10.3627	9.6232	10.1496
Confidence Interval 95%	(10.1217,10.5796)	(9.4573,9.7911)	(9.8048,10.5242)
Confidence Interval 80%	(10.2165,10.5127)	(9.5106,9.7353)	(9.923,10.3756)
ZALN, $\sigma_i^2 = \mathrm{Var}[\log(Y_i)\|Y_i > 0]$	2.1375	1.802	2.8976
Confidence Interval 95%	(1.6951,2.6315)	(1.5231,2.0831)	(2.2863,3.6579)
Confidence Interval 80%	(1.8561,2.4481)	(1.6246,1.9902)	(2.4674,3.3468)

Table 11: Predictions of expected claim amounts in Belgian francs (1EUR being approximately equal to 40 Belgian francs) and of parameters π, μ and σ^2 for different risk profiles in the ZALN model.

Table 11 displays the predictions obtained from the ZALN model for the three risk profiles defined earlier in the paper. Again, we discover there new features of the insurance data set as the three risk profiles greatly differ in their respective no-claim probabilities but not that much on the expected claim size when claims are reported. We see from Table 11 that $\mathrm{E}[Y_i|Y_i > 0]$ is similar for the low and high risk profiles, but smaller for the medium risk profile. This conforms with actuarial intuition: Individual drivers' characteristics are more efficient at explaining the number of reported TPL claims than their corresponding costs which depend to a large extent on the third party involved. Considering the variations in the no-claim probabilities, we nevertheless end up with an expected claim cost $\mathrm{E}[Y_i]$ which increases with the risk profile considered.

5 Discussion

The present paper extends the study conducted by Denuit and Lang (2004) to generalized additive models for location, scale and shape (GAMLSS). This flexible, semiparametric class of regression models turns out to be particularly suitable for analyzing insurance data. The restrictive exponential family assumption for the response is relaxed compared to standard GLM-based analysis which opens the door to numerous response distributions in line with the specificities of insurance data. While ordinary regression analyzes only the effects of covariates on the mean of a response, the approach proposed in the present paper allows the actuary to include risk factors not only in the mean but also in other parameters governing the claiming behavior, like the degree of residual heterogeneity or the no-claim probability. In this broader setting, the Poisson assumption made in Denuit and Lang (2004) is replaced with a mixed Poisson one, either the Negative Binomial distribution with cell-specific heterogeneity or the zero-inflated Poisson distribution with cell-specific probability mass at zero. Bayesian inference is based on efficient Markov chain Monte Carlo simulation techniques with appropriate proposal densities based on iteratively weighted least square approximations to the full conditionals. In this way, simultaneous estimation of possible nonlinear effects, spatial variations, random effects and interactions between risk factors within the data set are possible.

In addition to these models for claim frequencies, new models for claim severities are also presented, either per claim or aggregated per year. In the former case, location and scale parameters in the LogNormal, Gamma or Inverse Gaussian distributions may depend on covariates. Total claim amounts distributions are then obtained by combining estimated frequencies with estimated claim sizes by means of Panjer algorithm, for instance. In the latter case, an additional probability mass is placed at zero, corresponding to the policies without claims, allowing the actuary to account for zeros in the analysis of the amount of loss directly without resorting to models for claim frequencies. This yields zero-adjusted LogNormal, Gamma or Inverse Gaussian distributions able to deal with the large number of zero observations within the insurance data sets. Such zero-adjusted models based on distributions with a probability mass at zero and a continuous component thus allow for analyzing claim sizes in due

consideration of claim-free policyholders. In all cases, linear and nonlinear effects of the covariates can be accounted for, with or without parametric formulation in the latter case. Interactions between risk factors are allowed and spatial correlations as well as unobserved heterogeneity can be captured.

To illustrate the relevance of this approach, we performed a detailed illustrative analysis based on the Belgian motor insurance portfolio studied in Denuit and Lang (2004). These highly flexible regression models reveal new features of insurance data and have important consequences for a priori risk management and for credibility updating mechanisms. Guidelines for model choice with respect to the response distributions and for further model specifications are also provided. For the comparison of models with respect to the distribution, we consider quantile residuals as an effective graphical device and scoring rules that allow to quantify the predictive ability of the models. The deviance information criterion is used for further model specification.

To end with, let us mention that GAMLSS models have numerous additional potential applications to insurance. For instance, the GAMLSS family includes the Beta distribution inflated at 0, at 1 or at 0 and 1. This framework is thus appropriate to model material damage claims which are expressed as a percentage of the sum insured (the value of the insured vehicle, in motor insurance). Applications to life and health insurance are also promising, in the spirit of Gschlossl et al. (2011).

References

Baetschmann, G., Winkelmann, R., 2012. Modelling zero-inflated count data when exposure varies with an application to sick leave. Technical Report.

Beirlant, J., Derveaux, V., De Meyer, A.M., Goovaerts, M.J., Labies, E., Maenhoudt, B., 1991. Statistical risk evaluation applied to (Belgian) car insurance. Insurance: Mathematics and Economics 10, 289-302.

Bortoluzzo, A.B., Claro, D.P., Caetano, M.A.L., Artes, R., 2011. Estimating total claim size in the auto insurance industry: A comparion between Tweedie and zero-adjusted Inverse Gaussian distribution. Brazilian Administration Review 8, 37-47.

Boucher, J.-Ph., Denuit, M., Guillen, M., 2007. Risk classification for claim counts: A comparative analysis of various zero-inflated mixed Poisson and hurdle models. North American Actuarial Journal 11, 110-131.

Brezger, A., Lang, S., 2004. Bayesian P-splines. Journal of Computational and Graphical Statistics

13, 183–212.

Brezger, A., Lang, S., 2006. Generalized structured additive regression based on Bayesian P-splines Computational Statistics & Data Analysis 50, 967–991.

Cebrian, A., Denuit, M., Lambert, Ph., 2003. Generalized Pareto fit to the society of Actuaries large claims database. North American Actuarial Journal 7, 18-36.

Denuit, M., Lang, S., 2004. Nonlife ratemaking with Bayesian GAM's. Insurance: Mathematics and Economics 35, 627-647.

Denuit, M., Marechal, X., Pitrebois, S., Walhin, J.-F., 2007. Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems. Wiley, New York.

Dunn, P.K., Smyth, G.K., 1996. Randomized quantile residuals. Computational and Graphical Statistics 5, 236–245.

Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). Statistical Science 11, 89-121.

Fahrmeir, L., Kneib, T., Lang, S., 2004. Penalized additive regression for space-time data: A Bayesian perspective. Statistica Sinica 14, 713-743.

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

Fahrmeir, L., Lang, S., 2001b. Bayesian semiparametric regression analysis of multicategorical time-space data. Annals of the Institute of Statistical Mathematics 53, 10-30.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer-Verlag, New York.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression - Models, Methods and Applications. Springer.

Friedman, J.H., 1991. Multivariate adaptive regression splines (with discussion). Annals of Statistics 19, 1-141.

Gamerman, D., 1997. Sampling from the posterior distribution in Generalized Linear Mixed Models. Statistics and Computing 7, 57–68.

George, A., Liu, J.W., 1981. Computer Solution of Large Sparse Positive Definite Systems. Prentice-Hall.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378.

Gneiting, T., Ranjan, R., 2011. Comparing density forecasts using threshold and quantile-weighted scoring rules. Journal of Business & Economic Statistics 29, 411–421.

Green, P.J., 2001. A primer in Markov Chain Monte Carlo. In: Barndorff-Nielsen, O.E., Cox, D.R., Klüppelberg, C. (Eds.), Complex Stochastic Systems. Chapmann and Hall, London, pp. 1-62.

Gschlossl, S., Schoenmaekers, P., Denuit, M., 2011. Risk classification in life insurance: Methodology and case study. European Actuarial Journal 1, 23-41.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.

Hastie, T., Tibshirani, R., 1993. Varying-coefficient models. Journal of the Royal Statistical Society – Series B 55, 757-796.

Heller, G., Stasinopoulos, D. M., Rigby R. A., 2006. The zero-adjusted Inverse Gaussian distribution as a model for insurance data. Proceedings of the 21th International Workshop on Statistical Modelling, J. Hinde, J. Einbeck, J. Newell Editors, pp. 226-233.

Heller, G., Stasinopoulos, D. M., Rigby R. A., de Jong, P., 2007. Mean and dispersion modeling for policy claim costs. Scandinavian Actuarial Journal, 281-292.

Jørgensen, B., Paes de Souza, M.C., 1994. Fitting Tweedie's compound Poisson model to insurance claims data. Scandinavian Actuarial Journal 69-93.

Jullion, A., Lambert, P., 2007. Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. Computational Statistics & Data Analysis 51, 2542–2558.

Kaas, R., Goovaerts, M.J., Dhaene, J., Denuit, M., 2008. Modern Actuarial Risk Theory Using R. Springer, New York.

Klein, N., Kneib, T., Lang, S., 2013a. Bayesian Generalized Additive Models for Location, Scale and Shape for Zero-Inflated and Overdispersed Count Data. Technical Report, http://econpapers.repec.org/paper/innwpaper/2013-12.htm.

Klein, N., Kneib, T., Lang, S., 2013b. Bayesian Structured Additive Distributional Regression. Technical Report.

Klugman, S., Panjer, H., Willmot, G., 2004. Loss Models: From Data to Decisions. Wiley, New York.

Lang, S., Brezger, A., 2004. Bayesian P-splines. Journal of Computational and Graphical Statistics 13, 183-212.

Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K., Kneib, T., 2013. Multilevel Structured Additive Regression. Statistics and Computing 23, in press.

Lin, X., Zhang, D., 1999. Inference in Generalized Additive Mixed Models by using Smoothing Splines. Journal of the Royal Statistical Society – Series B 61, 381-400.

Ngo, L.,Wand, M.P., 2003. Smoothing with mixed model software. Journal of Statistical Software 9. Also available at http://www.maths.unsw.edu.au/Ëœwand/.

Pigeon, M., Denuit, M., 2011. Composite Lognormal-Pareto model with random threshold. Scandinavian Actuarial Journal, 177-192.

Resti, Y., Ismail, N., Jamaan, S.H., 2013. Estimation of claim cost data using zero-adjusted Gamma and Inverse Gaussian regression models. Journal of Mathematics and Statistics 9, 186-192.

Rigby, R.A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape (with discussion). Applied Statistics 54, 507–554.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields. Chapman & Hall / CRC.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Linde, A., 2002. Bayesian measures of model complexity and fits (with discussion). Journal of the Royal Statistical Society – Series B 64, 583-639.

Stasinopoulos, D.M., Rigby, R.A., 2007 Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software 23, 1-46.

Stasinopoulos, D.M., Rigby, B., Akantziliotou, C., 2008. Instructions on how to use the gamlss package in R, Second Edition.

Smyth, G.K., Jørgensen, B., 2002. Fitting Tweedie's compund Poisson model to insurance claims data: Dispersion modeling. Astin Bulletin 32, 143–157.

Wand, M.P., 2003. Smoothing and mixed models. Computational Statistics 18, 223-249.

Wood, S.N., 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society – Series B 62, 413-428.

Wood, S.N., 2001. mgcv: GAMs and generalized ridge regression for R. R News 1, 20-25.

Wood, S.N., 2003. Thin plate regression splines. Journal of the Royal Statistical Society – Series B 65, 95-114.

Wood, S. N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99, 673–686.

Wood, S.N., 2006. Generalized Additive Models : An Introduction with R. Chapmann & Hall.

Wood, S. N., 2008. Fast stable direct fitting and smoothness selection for Generalized Additive Models. Journal of the Royal Statistical Society – Series B 70, 495–518.

Yip, K.C.H., Yau, K.K.W., 2005. On modeling claim frequency data in general insurance with extra zeros. Insurance: Mathematics and Economics 36, 153-163.

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

http://eeecon.uibk.ac.at/wopec/

2013-24 **Nadja Klein, Michel Denuit, Stefan Lang, Thomas Kneib:** Nonlife ratemaking and risk management with bayesian additive models for location, scale and shape

2013-23 **Nadja Klein, Thomas Kneib, Stefan Lang:** Bayesian structured additive distributional regression

2013-22 **David Plavcan, Georg J. Mayr, Achim Zeileis:** Automatic and probabilistic foehn diagnosis with a statistical mixture model

2013-21 **Jakob W. Messner, Georg J. Mayr, Achim Zeileis, Daniel S. Wilks:** Extending extended logistic regression to effectively utilize the ensemble spread

2013-20 **Michael Greinecker, Konrad Podczeck:** Liapounoff's vector measure theorem in Banach spaces *forthcoming in Economic Theory Bulletin*

2013-19 **Florian Lindner:** Decision time and steps of reasoning in a competitive market entry game

2013-18 **Michael Greinecker, Konrad Podczeck:** Purification and independence

2013-17 **Loukas Balafoutas, Rudolf Kerschbamer, Martin Kocher, Matthias Sutter:** Revealed distributional preferences: Individuals vs. teams

2013-16 **Simone Gobien, Björn Vollan:** Playing with the social network: Social cohesion in resettled and non-resettled communities in Cambodia

2013-15 **Björn Vollan, Sebastian Prediger, Markus Frölich:** Co-managing common pool resources: Do formal rules have to be adapted to traditional ecological norms?

2013-14 **Björn Vollan, Yexin Zhou, Andreas Landmann, Biliang Hu, Carsten Herrmann-Pillath:** Cooperation under democracy and authoritarian norms

2013-13 **Florian Lindner, Matthias Sutter:** Level-k reasoning and time pressure in the 11-20 money request game *forthcoming in Economics Letters*

2013-12 **Nadja Klein, Thomas Kneib, Stefan Lang:** Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data

2013-11 **Thomas Stöckl:** Price efficiency and trading behavior in limit order markets with competing insiders *forthcoming in Experimental Economics*

2013-10 **Sebastian Prediger, Björn Vollan, Benedikt Herrmann:** Resource scarcity, spite and cooperation

2013-09 **Andreas Exenberger, Simon Hartmann:** How does institutional change coincide with changes in the quality of life? An exemplary case study

2013-08 **E. Glenn Dutcher, Loukas Balafoutas, Florian Lindner, Dmitry Ryvkin, Matthias Sutter:** Strive to be first or avoid being last: An experiment on relative performance incentives.

2013-07 **Daniela Glätzle-Rützler, Matthias Sutter, Achim Zeileis:** No myopic loss aversion in adolescents? An experimental note

2013-06 **Conrad Kobel, Engelbert Theurl:** Hospital specialisation within a DRG-Framework: The Austrian case

2013-05 **Martin Halla, Mario Lackner, Johann Scharler:** Does the welfare state destroy the family? Evidence from OECD member countries

2013-04 **Thomas Stöckl, Jürgen Huber, Michael Kirchler, Florian Lindner:** Hot hand belief and gambler's fallacy in teams: Evidence from investment experiments

2013-03 **Wolfgang Luhan, Johann Scharler:** Monetary policy, inflation illusion and the Taylor principle: An experimental study

2013-02 **Esther Blanco, Maria Claudia Lopez, James M. Walker:** Tensions between the resource damage and the private benefits of appropriation in the commons

2013-01 **Jakob W. Messner, Achim Zeileis, Jochen Broecker, Georg J. Mayr:** Improved probabilistic wind power forecasts with an inverse power curve transformation and censored regression

2012-27 **Achim Zeileis, Nikolaus Umlauf, Friedrich Leisch:** Flexible generation of e-learning exams in R: Moodle quizzes, OLAT assessments, and beyond

2012-26 **Francisco Campos-Ortiz, Louis Putterman, T.K. Ahn, Loukas Balafoutas, Mongoljin Batsaikhan, Matthias Sutter:** Security of property as a public good: Institutions, socio-political environment and experimental behavior in five countries

2012-25 **Esther Blanco, Maria Claudia Lopez, James M. Walker:** Appropriation in the commons: variations in the opportunity costs of conservation

2012-24 **Edgar C. Merkle, Jinyan Fan, Achim Zeileis:** Testing for measurement invariance with respect to an ordinal variable *forthcoming in Psychometrika*

2012-23 **Lukas Schrott, Martin Gächter, Engelbert Theurl:** Regional development in advanced countries: A within-country application of the Human Development Index for Austria

2012-22 **Glenn Dutcher, Krista Jabs Saral:** Does team telecommuting affect productivity? An experiment

2012-21 **Thomas Windberger, Jesus Crespo Cuaresma, Janette Walde:** Dirty floating and monetary independence in Central and Eastern Europe - The role of structural breaks

2012-20 **Martin Wagner, Achim Zeileis:** Heterogeneity of regional growth in the European Union

2012-19 **Natalia Montinari, Antonio Nicolo, Regine Oexl:** Mediocrity and induced reciprocity

2012-18 **Esther Blanco, Javier Lozano:** Evolutionary success and failure of wildlife conservancy programs

2012-17 **Ronald Peeters, Marc Vorsatz, Markus Walzl:** Beliefs and truth-telling: A laboratory experiment

2012-16 **Alexander Sebald, Markus Walzl:** Optimal contracts based on subjective evaluations and reciprocity

2012-15 **Alexander Sebald, Markus Walzl:** Subjective performance evaluations and reciprocity in principal-agent relations

2012-14 **Elisabeth Christen:** Time zones matter: The impact of distance and time zones on services trade

2012-13 **Elisabeth Christen, Joseph Francois, Bernard Hoekman:** CGE modeling of market access in services

2012-12 **Loukas Balafoutas, Nikos Nikiforakis:** Norm enforcement in the city: A natural field experiment *forthcoming in European Economic Review*

2012-11 **Dominik Erharter:** Credence goods markets, distributional preferences and the role of institutions

2012-10 **Nikolaus Umlauf, Daniel Adler, Thomas Kneib, Stefan Lang, Achim Zeileis:** Structured additive regression models: An R interface to BayesX

2012-09 **Achim Zeileis, Christoph Leitner, Kurt Hornik:** History repeating: Spain beats Germany in the EURO 2012 Final

2012-08 **Loukas Balafoutas, Glenn Dutcher, Florian Lindner, Dmitry Ryvkin:** The optimal allocation of prizes in tournaments of heterogeneous agents

2012-07 **Stefan Lang, Nikolaus Umlauf, Peter Wechselberger, Kenneth Harttgen, Thomas Kneib:** Multilevel structured additive regression

2012-06 **Elisabeth Waldmann, Thomas Kneib, Yu Ryan Yu, Stefan Lang:** Bayesian semiparametric additive quantile regression

2012-05 **Eric Mayer, Sebastian Rueth, Johann Scharler:** Government debt, inflation dynamics and the transmission of fiscal policy shocks *forthcoming in Economic Modelling*

2012-04 **Markus Leibrecht, Johann Scharler:** Government size and business cycle volatility; How important are credit constraints? *forthcoming in Economica*

2012-03 **Uwe Dulleck, David Johnston, Rudolf Kerschbamer, Matthias Sutter:** The good, the bad and the naive: Do fair prices signal good types or do they induce good behaviour?

2012-02 **Martin G. Kocher, Wolfgang J. Luhan, Matthias Sutter:** Testing a forgotten aspect of Akerlof's gift exchange hypothesis: Relational contracts with individual and uniform wages

2012-01 **Loukas Balafoutas, Florian Lindner, Matthias Sutter:** Sabotage in tournaments: Evidence from a natural experiment *published in Kyklos*

University of Innsbruck

Working Papers in Economics and Statistics

2013-24

Nadja Klein, Michel Denuit, Stefan Lang, Thomas Kneib

Nonlife ratemaking and risk management with bayesian additive models for location, scale and shape

Abstract
Generalized additive models for location, scale and shape define a flexible, semiparametric class of regression models for analyzing insurance data in which the exponential family assumption for the response is relaxed. This approach allows the actuary to include risk factors not only in the mean but also in other parameters governing the claiming behavior, like the degree of residual heterogeneity or the no-claim probability. In this broader setting, the Negative Binomial regression with cell-specific heterogeneity and the zero-inflated Poisson regression with cell-specific additional probability mass at zero are applied to model claim frequencies. Models for claim severities that can be applied either per claim or aggregated per year are also presented. Bayesian inference is based on efficient Markov chain Monte Carlo simulation techniques and allows for the simultaneous estimation of possible nonlinear effects, spatial variations and interactions between risk factors within the data set.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)