Bayesian multivariate Poisson models for insurance ratemaking *

Lluís Bermúdez[a†]& Dimitris Karlis[b]

October 28, 2010

[a]University of Barcelona. Spain

[b]Athens University of Economics and Business. Greece

Abstract

When actuaries face the problem of pricing an insurance contract that contains different types of coverage, such as a motor insurance or homeowner's insurance policy, they usually assume that types of claim are independent. However, this assumption may not be realistic: several studies have shown that there is a positive correlation between types of claim. Here we introduce different multivariate Poisson regression models in order to relax the independence assumption, including zero-inflated models to account for excess of zeros and overdispersion. These models have been largely ignored to date, mainly because of their computational difficulties. Bayesian inference based on MCMC helps to resolve this problem (and also allows us to derive, for several quantities of interest, posterior summaries to account for uncertainty). Finally, these models are applied to an automobile insurance claims database with three different types of claims. We analyse the consequences for pure and loaded premiums when the independence assumption is relaxed by using different multivariate Poisson regression models together with their zero-inflated versions.

JEL classification: C51; *IM classification:* IM11; *IB classification:* IB40.

Keywords: Multivariate Poisson regression models, Zero-inflated models, Automobile insurance, MCMC inference, Gibbs sampling.

Acknowledgements. The first author wishes to acknowledge discussions with researchers at RFA-IREA at the University of Barcelona and the support of the Spanish Ministry of Education and FEDER grant SEJ 2007-63298.

†**Corresponding Author.** Departament de Matemàtica Econòmica, Financera i Actuarial, Universitat de Barcelona, Diagonal 690, 08034-Barcelona, Spain. Tel.:+34-93-4034854; fax: +34-93-4034892; e-mail: lbermudez@ub.edu

1 Introduction

Automobile insurance aims at covering different type of claims incurred as a result of traffic accidents. In most developed countries motor insurance is compulsory for driving a motor vehicle on public roads. The level of protection of each jurisdiction varies greatly, but essentially, the aim of compulsory motor insurance for all vehicle owners is to cover damage to third parties. This coverage is usually termed third-party liability coverage and provides financial compensation to cover any injuries caused to other people or their property.

Apart from this liability coverage, motor insurance can also cover the insured party (vehicle damage and personal injury). Property coverage or first-party coverage provides different levels of protection depending on the policy the insured purchases. Car owners may take out comprehensive coverage (damage to the vehicle caused by any unknown party, for example, damage resulting from theft, flood or fire), collision coverage (damage resulting from a collision with another vehicle or object when the policyholder is at fault), or a set of basic guarantees such as an emergency roadside assistance, legal assistance or insurance covering medical costs.

Pricing is especially complicated in the branch of motor insurance, due to the heterogeneity of the portfolios and the fact that policies cover different risks. One way to handle the problem of this heterogeneity is to segment the portfolio into homogeneous classes so that all policyholders belonging to the same class pay the same premium. To achieve this, an *a priori* ratemaking based on generalized linear models (GLM) is usually accepted. A thorough review of ratemaking systems for motor insurance, when modelling claim count data, can be found in Denuit *et al.* (2007).

With the usual ratemaking procedure, modelling the number of claims incurred using Poisson regression models, the expected number of claims (the pure premium, assuming the amount of the expected claim equals one monetary unit) is obtained for each class of guarantee as a function of different factors. Then, assuming independence between types of claims, the total motor insurance premium is obtained by the sum of the expected number of claims of each guarantee. This procedure presents at least three important limitations.

First, not all factors influencing risk can be identified, measured and introduced in the *a priori* tariff system, and hence, the tariff classes may be quite heterogeneous. To correct for this unobserved heterogeneity an *a posteriori* tariff (or bonus-malus system) can be used, by fitting

an individual premium based on the experience of claims for each insured party. There is a large amount of literature on bonus-malus systems (see Denuit *et al.*, 2007). Another way to handle unobserved heterogeneity is to introduce a random effect into the model (Cameron and Trivedi, 1998 and Boucher and Denuit, 2006).

Second, unobserved heterogeneity and serial dependence (when the data consist of repeated observations regarding the same policyholder) will often lead to overdispersion (variance greater than mean) which cannot be fully remedied by Poisson regression models. Failing to account for overdispersion may increase the number of factors considered significant by artificially increasing their level of significance. To account for overdispersion, some generalizations of the model have been considered (see e.g. zero-inflated models as in Boucher *et al.*, 2007).

Finally, it remains to be established whether the independence assumption between types of claims is realistic. This question is not widely discussed in the actuarial literature. When this assumption is relaxed, it is interesting to see how the tariff system is affected. In Frees and Valdez (2008) and Frees *et al.* (2009) a hierarchical statistical model is fitted using micro-level data. A multivariate probit model has also been suggested by Young *et al.* (2009) to account for dependencies among claim types. In Bermúdez (2009), the interpretation of a number of bivariate Poisson models was illustrated in the context of motor insurance claims and the conclusion was that using a bivariate Poisson model leads to an *a priori* ratemaking that presents larger variances and, hence, larger loadings than those obtained under the independence assumption. In that study, only two types of claim were considered: claims for third-party liability or for the rest of guarantees. Obviously, this is a limitation that other multivariate count data models can overcome: for instance, we could divide claims for third-party liability into vehicle damage and personal injury claims, or distinguish between motor collision coverage and the rest of guarantees. In the present paper we deal with this kind of extension.

Here we introduce different multivariate Poisson regression models in order to relax the independence assumption when pricing several guarantees simultaneously in automobile insurance. Creating multivariate Poisson models is not easy, as many different models can be obtained. In the present paper we use two such models and their zero-inflated variants (to account for the excess of zeros observed in automobile databases, see e.g. Boucher *et al.*, 2007 and Bermúdez, 2009). The first one, which we call the "common covariance model", has been defined in Tsionas (2001) and the second one, the "full covariance model", in Karlis and Meligkotsidou (2005). In

addition, here we extend these models with their zero-inflated variants. It is important to realize that zero inflation also introduces overdispersion in the marginal distributions. Hence, zero-inflated models can introduce improvements in several aspects of the data. Multivariate zero-inflated models are well known for claim counts data, see for example Boucher and Denuit (2008) for a credibility application. Our approach differs from this paper as we attempt to model dependence between different types of claims and not for a panel data, i.e. one type observed in different time periods. Moreover, they are focus on *a posteriori* premiums and we use these models for *a priori* ratemaking procedure.

Finally, we use a Bayesian approach for fitting the models that offers some advantages. It facilitates the estimation for such complicated models, while at the same time, allows for deriving posterior quantities of interest not as simple point estimates but together with their posterior distribution providing more insight and better understanding for correct ratemaking. To our knowledge, the derived MCMC scheme for multivariate zero-inflated Poisson models is novel.

The article is organized as follows. First, in Section 2 we introduce several multivariate Poisson regression models. In Section 3 we discuss the Bayesian methodology used to fit the statistical model to the data. In Section 4 the database from a Spanish insurance company is described. In Section 5 the results are summarized. Finally, we provide concluding remarks in Section 6.

2 Multivariate Poisson regression models

Let us consider a policyholder with N_1 the number of claims for motor third-party liability coverage, N_2 the number of claims for motor collision coverage, N_3 the number of claims for the rest of motor guarantees and $N = N_1 + N_2 + N_3$ the total number of claims during one year.

Our aim is to analyze different multivariate Poisson models as a way to relax the independence assumption between types of claims when a ratemaking procedure is developed. First, we analyze a simple multivariate Poisson model with common covariance parameter (Johnson *et al.*, 1997, Tsionas, 2001). Second, we study a multivariate Poisson model with full covariance following the model introduced by Karlis and Meligkotsidou (2005). Finally, we consider zero-inflated versions of these models to account for the excess of zero claims and the overdispersion observed typically in such datasets.

2.1 A model with common covariance

The first model is based on a simple multivariate reduction. Namely we assume that

$$\begin{aligned} N_1 &= Y_1 + Y_0 \\ N_2 &= Y_2 + Y_0 \\ N_3 &= Y_3 + Y_0 \end{aligned} \tag{1}$$

where $Y_i \sim Po(\theta_i)$, $i \in \{0,1,2,3\}, \theta_i > 0$. Then, each N_i, $i \in \{1,2,3\}$ marginally follows a Poisson distribution with parameter $\theta_i + \theta_0$. θ_0 is a common covariance parameter which measures the covariance of each pair. The covariance matrix is

$$Cov(\mathbf{N}) = \begin{bmatrix} \theta_1 + \theta_0 & \theta_0 & \theta_0 \\ \theta_0 & \theta_2 + \theta_0 & \theta_0 \\ \theta_0 & \theta_0 & \theta_3 + \theta_0 \end{bmatrix}.$$

The joint probability function of the vector $\mathbf{N}$ is given by

$$P(n_1, n_2, n_3) = \exp(-\theta) \sum_{k=0}^{s} \frac{\theta_0^k}{k!} \frac{\theta_1^{n_1-k}}{(n_1-k)!} \frac{\theta_2^{n_2-k}}{(n_2-k)!} \frac{\theta_3^{n_3-k}}{(n_3-k)!},$$

where $s = min\{n_1, n_2, n_3\}$ and $\theta = \theta_1 + \theta_2 + \theta_3 + \theta_0$. We will denote the above distribution as $MP_1(\theta_1, \theta_2, \theta_3, \theta_0)$.

Let us assume that N_{1q}, N_{2q} and N_{3q} denote respectively the random variables indicating the number of claims of each type of guarantee for the qth policyholder. We may allow for covariates by considering that $\log(\theta_{iq}) = \boldsymbol{x'_{iq}\beta_i}$, where $\boldsymbol{x_{iq}}$ is a vector of explanatory variables and $\boldsymbol{\beta_i}$ denotes the corresponding vector of regression coefficients.

Note that different covariates can be used to model each parameter θ_i, $i = 1, 2, 3$. In general we may use covariates to θ_0 as well but this would make the interpretation much more difficult. If covariates are introduced to model θ_1, θ_2 and θ_3, a multivariate Poisson regression model can be defined with the following scheme (for more details see Tsionas, 2001):

$$\begin{aligned} (N_{1q}, N_{2q}, N_{3q}) &\sim MP_1(\theta_{1q}, \theta_{2q}, \theta_{3q}, \theta_{0q}), \\ \log(\theta_{1q}) &= \boldsymbol{x'_{1q}\beta_1}, \\ \log(\theta_{2q}) &= \boldsymbol{x'_{2q}\beta_2}, \\ \log(\theta_{3q}) &= \boldsymbol{x'_{3q}\beta_3}. \end{aligned} \tag{2}$$

Limitations of this model are that it assumes a common covariance for each pair; it allows only for positive covariance (correlation); and the marginal distributions are Poisson, and so we cannot model over(under)dispersion.

There are some other models that allow for negative correlation (see van Ophem 1999, Chib and Winkelmann 2001, Berkhout and Plug 2004, Karlis and Melogkotsidou (2007), Nikoloulopoulos and Karlis, 2009), but they are much more complicated and require a special effort for parameter estimation. In the context of automobile insurance, it is not necessary to consider negative correlation for these type of claims.

However, in the next sections, we consider a more complex model to allow different covariance for each pair of variables, and zero-inflated models to deal with overdispersion which has often been observed when modelling claim counts in automobile insurance data (Dean, 1992).

2.2 A model with full covariance

In order to extend the previous model and allow for modelling the covariance structure of the data in a flexible way, we consider the case of the trivariate Poisson model with full two-way covariance structure:

$$\begin{aligned} N_1 &= Y_1 + Y_{12} + Y_{13} \\ N_2 &= Y_2 + Y_{12} + Y_{23} \\ N_3 &= Y_3 + Y_{13} + Y_{23} \end{aligned} \tag{3}$$

where $Y_i \sim Po(\mu_i)$, $i \in \{1,2,3\}$ and $Y_{ij} \sim Po(\theta_{ij})$, $i,j \in \{1,2,3\}$, $i < j$, $\mu_i, \theta_{ij} > 0$. Then, each N_i, $i \in \{1,2,3\}$ marginally follows a Poisson distribution with parameter $\mu_i + \theta_{ij} + \theta_{ik}$, $i,j,k \in \{1,2,3\}$, $i \neq j \neq k$.

Now, random variables N_1, N_2, N_3 jointly follow a trivariate Poisson distribution with parameter $\boldsymbol{\theta} = (\mu_1, \mu_2, \mu_3, \theta_{12}, \theta_{13}, \theta_{23})'$. The means of the random variables are $\mu_1 + \theta_{12} + \theta_{13}$, $\mu_2 + \theta_{12} + \theta_{23}$ and $\mu_3 + \theta_{13} + \theta_{23}$ respectively and their variance-covariance matrix is given by

$$Cov(\mathbf{N}) = \begin{bmatrix} \mu_1 + \theta_{12} + \theta_{13} & \theta_{12} & \theta_{13} \\ \theta_{12} & \mu_2 + \theta_{12} + \theta_{23} & \theta_{23} \\ \theta_{13} & \theta_{23} & \mu_3 + \theta_{13} + \theta_{23} \end{bmatrix}.$$

The parameters θ_{ij}, $i,j = 1,2,3, i \neq j$, can be interpreted straightforward as the covariances between the variables X_i and X_j and, thus, we refer to them as the covariance parameters. The

parameters μ_i, $i = 1, 2, 3$, appear only at the marginal means and variances and we refer to them as the mean parameters. It is clear that this model is more flexible for real applications than the one with common covariance. For example, if the data refer to the number of claims for different coverage of an automobile insurance, it is natural to assume that each pair of different coverage has different covariance due to the intrinsic nature of these coverages instead of assuming that all pairs have the same covariance.

Again in order to extend the applicability of the model we may assume that the parameters θ_i (including both the mean and the covariance parameters) are functions of explanatory variables. Therefore we may add covariates by assuming that $\log(\theta_{iq}) = \boldsymbol{x'_{iq}}\boldsymbol{\beta_i}$, where $\boldsymbol{x_{iq}}$ is a vector of explanatory variables and $\boldsymbol{\beta_i}$ denotes the corresponding vector of regression coefficients. To make the model easier to interpret, we consider covariates only for the mean parameters μ_i, $i = 1, 2, 3$. While covariates can also be added to the covariance parameters, this again would make the interpretation of the model very difficult and so we do not consider them here. Finally, note that the covariates associated with each parameter may be different.

The joint probability function (jpf) is given by:

$$P(n_1, n_2, n_3) = \exp(-\theta) \sum_{k_1=0}^{s_1} \sum_{k_2=0}^{s_2} \sum_{k_3=0}^{s_3} \frac{\theta_{12}^{k_1}}{k_1!} \frac{\theta_{13}^{k_2}}{k_2!} \frac{\theta_{12}^{k_3}}{k_3!} \frac{\mu_1^{(n_1-k_1-k_2)}}{(n_1-k_1-k_2)!} \frac{\mu_2^{(n_2-k_2-k_3)}}{(n_2-k_2-k_3)!} \frac{\mu_3^{(n_3-k_1-k_3)}}{(n_3-k_1-k_3)!}$$

where $s_1 = min\{n_1, n_2\}$, $s_2 = min\{n_1 - s_1, n_3\}$, $s_3 = min\{n_2 - s_1, n_3 - s_2\}$ and $\theta = \mu_1 + \mu_2 + \mu_3 + \theta_{12} + \theta_{13} + \theta_{23}$. We will denote the above distribution as $MP_2(\mu_1, \mu_2, \mu_3, \theta_{12}, \theta_{13}, \theta_{23})$.

Note that this model allows for different covariances between different pairs, making the model more realistic at the cost of having two additional parameters to estimate. The jpf is quite complicated as it involves successive summations. One may improve it by deriving a recurrence relationship between the probabilities, i.e. by calculating probabilities based on ones that have already been calculated. This reduces the computation burden by avoiding excessive summation and reducing error accumulation. On the other hand, the data augmentation offered by the multivariate reduction makes Bayesian methods appealing. More details for the model can be found in Karlis and Meligkotsidou (2005).

2.3 Zero-inflated models

The multivariate Poisson models treated above have Poisson marginal distributions and thus they cannot model overdispersion. Certain amounts of overdispersion can be introduced by

considering inflated versions of multivariate Poisson regression models, like the models described in Karlis and Ntzoufras (2003, 2005) and in Bermúdez (2009) used in the automobile insurance context for the bivariate case. In the univariate case, zero-inflated models are well understood as models to account for the excess of zeros observed in certain circumstances. In the multivariate case, inflation can occur in different patterns. A particulary interesting case in practice, is when the $(0, 0, \ldots, 0)$ cell occurs more often than the assumed model would predict. Multivariate zero-inflated models have attracted much less interest than univariate and bivariate inflated models (see, e.g. Li *et* al., 1999). For an actuarial application see Boucher and Denuit (2008).

We propose zero-inflated versions of the previous models with the following form:

$$P_{ZI}(n_1, n_2, n_3) = \begin{cases} p + (1-p)\, P(n_1, n_2, n_3) & \text{if } n_1 = n_2 = n_3 = 0 \\ (1-p)\, P(n_1, n_2, n_3) & \text{otherwise} \end{cases}$$

i.e. the model moves probability from other cells to the $(0, 0, 0)$ cell. A natural interpretation for this is that most clients never report an accident and thus the number of zeros is larger than would be expected under a Poisson model. Note that one may define more complicated models by assuming other kind of inflations. Moreover, one may add covariates to p, implying that inflation depends on external factors. We will not pursue this here.

It is important that zero inflation introduces overdispersion to the marginal distributions. One can easily see that the marginal distributions are no longer simple Poisson distributions but zero-inflated versions. It is well known (see, e.g. Bohning *et al.*, 1999) that zero-inflated Poisson models are overdispersed relative to simple Poisson models. In the bivariate (multivariate case) it has been shown that the covariance also increases (see, Wang *et al.*, 2003 and Karlis and Ntzoufras, 2005). Hence, inflated models can introduce improvements in several aspects of the data.

2.4 Moments

For the analysis presented in the following sections, some moments and covariances of the four models presented here need to be calculated. Tables 1 and 2 contain the values for the marginal expectations and variances, as well as the covariances (for ease of exposition we present the general form for N_i for the common covariance models, but for the full covariance model we present it with specific variates N_1 and N_2 in order to diminish the notational burden; of course

Common Covariance	Full Covariance
$E(N_i) = V(N_i) = \theta_i + \theta_0$	$E(N_1) = V(N_1) = \mu_1 + \theta_{12} + \theta_{13}$
$Cov(N_i, N_j) = \theta_0$	$Cov(N_i, N_j) = \theta_{ij}$
$E(N) = \sum_{i=1}^{3} \theta_i + 3\theta_0$	$E(N) = \sum_{i=1}^{3} \mu_i + 2(\theta_{12} + \theta_{13} + \theta_{23})$
$V(N) = \sum_{i=1}^{3} \theta_i + 9\theta_0$	$V(N) = \sum_{i=1}^{3} \mu_i + 4(\theta_{12} + \theta_{13} + \theta_{23})$

Table 1: Expectations and variances for CC and FC models.

Z-I Common Covariance	Z-I Full Covariance
$E(N_i) = (1-p)(\theta_i + \theta_0)$	$E(N_1) = (1-p)(\mu_1 + \theta_{12} + \theta_{13})$
$V(N_i) = (1-p)\left\{(\theta_i + \theta_0) + p(\theta_i + \theta_0)^2\right\}$	$V(N_1) = (1-p)\left\{(\mu_1 + \theta_{12} + \theta_{13}) + p(\mu_1 + \theta_{12} + +\theta_{13})^2\right\}$
$Cov(N_i, N_j) = (1-p)\left\{\theta_0 + (\theta_i + \theta_0)(\theta_j + \theta_0)\right\}$ $-\left\{(1-p)^2(\theta_i + \theta_0)(\theta_j + \theta_0)\right\}$	$Cov(N_1, N_2) = (1-p)\left\{\theta_{12} + (\mu_1 + \theta_{12} + \theta_{13})(\mu_2 + \theta_{12} + \theta_{23})\right\}$ $-\left\{(1-p)^2(\mu_1 + \theta_{12} + \theta_{13})(\mu_2 + \theta_{12} + \theta_{13})\right\}$
$E(N) = (1-p)(\sum_{i=1}^{3} \theta_i + 3\theta_0\ )$	$E(N) = (1-p)(\sum_{i=1}^{3} \mu_i + 2\theta_{12} + 2\theta_{13} + 2\theta_{23})$
$V(N) = \sum_{i=1}^{3} V(N_i) + 2 \sum_{i,j,\ i<j} Cov(N_i, N_j)$	$V(N) = \sum_{i=1}^{3} V(N_i) + 2 \sum_{i,j,\ i<j} Cov(N_i, N_j)$

Table 2: Expectations and covariances for the zero-inflated models (ZICC and ZIFC).

similar formulas are straightforward for the other variables).

Therefore, in Table 1 one can find the expressions for the multivariate Poisson model with common covariance (CC), and the multivariate Poisson model with full covariance (FC) while Table 2 presents the respective zero-inflated versions (ZICC and ZIFC). Clearly, for the zero-inflated models, the formulas are more complicated. It is also obvious that zero inflation increases both the variances and the covariance, i.e. it adds overdispersion to the data. Finally formulas for $V(N)$ are rather complicated though numerically their calculation is straightforward.

Note that our MCMC approach, described in the next section, allows to easily estimate any quantities of interest including moments ever if they are not analytically available.

3 Bayesian estimation

Bayesian approaches are widely used today. Putting aside philosophical issues for and/or against the Bayesian paradigm, the application of Bayesian methods has certain advantages. First of all, prior information can be incorporated and used in a convenient and mathematically neat way.

In actuarial applications, this prior information may be available (e.g. from past experiences) and hence it is naturally accommodated to the model. Secondly, Bayesian approaches through Markov Chain Monte Carlo (MCMC) methods allow the treatment of high dimensional problems with a large number of parameters, especially in problems where classical maximum likelihood approaches fail or are difficult to use. Moreover, the developments in the Bayesian field over last decade have meant that the methods are now widely available and can be interpreted by a wide audience. In this paper we apply a Bayesian approach through MCMC to fit and estimate the parameters of the models considered.

3.1 Full covariance model

In order to avoid repetition, we will describe only the Bayesian estimation for the model with full covariance: the model for the common covariance can be deduced in a similar manner, but we skip the details. In the next subsection, we describe the additional steps for the zero-inflated models.

The parameters to be estimated are $\Theta = (\theta_{12}, \theta_{13}, \theta_{23}, \beta_1, \beta_2, \beta_3)$ where β_i are the regression coefficients and θ_{ij} the covariance parameters.

We consider non-informative priors for the regression coefficients. Alternatively, with only minor changes in the code one may assume multivariate normal priors for β's. Prior information can be incorporated by locating the prior to specific values and allowing for smaller variance. Our non-informative approach is equivalent to assuming a prior centered at 0 and a diagonal covariance matrix, with very large variances so as to represent ignorance. For the covariance parameters we assume $Gamma(a = 0.1, b = 0.1)$ priors, i.e. gamma distributions with mean 1 and variance 100, i.e. rather diffuse priors implying ignorance. Again based on prior experience this information can be incorporated appropriately in the priors. However, if the sample size is large the impact of priors is rather small.

To run MCMC, we use the trivariate reduction technique applied to derive the multivariate model in (3). The central idea is that if we had observed all the latent variables $Y_1, Y_2, Y_3, Y_{12}, Y_{13}, Y_{23}$ then we could fully represent the data. But as we observe only the N_i's $i = 1, 2, 3$ we cannot fully recover the latent variables. The latent variables, being simple Poisson variables, have a joint likelihood which is simply the product of Poisson probability mass functions and hence is much more convenient as there is no summation. So, the simple idea behind

the Bayesian approach is to augment the observed data to the unobserved quantities Y_{12}, Y_{13}, Y_{23} (note that we really only need these, as the rest can be obtained by simple subtraction). In what follows, we denote by k_{ijm} the realization of the latent variable Y_{ij} for the m-th individual. Clearly we need to obtain the values of k_{ijm} for $i < j, \;\; i, j = 1, 2, 3, \;\; m = 1, \ldots, n$ where n is the sample size, i.e. the number of individuals.

Using $\mu_{1i} = \exp(\beta_1 x_i)$, $\mu_{2i} = \exp(\beta_2 x_i)$, $\mu_{3i} = \exp(\beta_3 x_i)$, and letting $(| \cdot)$ imply the full posterior, given all the rest parameters we can derive the full posteriors for all the quantities of interest as:

$$\begin{aligned}
k_{12i} \mid \cdot &\propto \frac{\theta_{12}^{k_{12i}}}{k_{12i}!(n_{1i} - k_{13i} - k_{12i})!(n_{2i} - k_{12i} - k_{23i})!} \left(\frac{1}{\mu_{1i}\mu_{2i}}\right)^{k_{12i}}, \\
k_{13i} \mid \cdot &\propto \frac{\theta_{13}^{k_{13i}}}{k_{13i}!(n_{1i} - k_{13i} - k_{12i})!(n_{3i} - k_{13i} - k_{23i})!} \left(\frac{1}{\mu_{1i}\mu_{3i}}\right)^{k_{13i}}, \\
k_{23i} \mid \cdot &\propto \frac{\theta_{23}^{k_{23i}}}{k_{23i}!(n_{2i} - k_{23i} - k_{12i})!(n_{3i} - k_{13i} - k_{23i})!} \left(\frac{1}{\mu_{2i}\mu_{3i}}\right)^{k_{23i}},
\end{aligned}$$

for $i = 1, \ldots, n$ with

$$\begin{aligned}
k_{12i} &= 0, \ldots, min(n_{1i} - k_{13i}, n_{2i} - k_{23i}), \\
k_{13i} &= 0, \ldots, min(n_{1i} - k_{12i}, n_{3i} - k_{23i}), \\
k_{23i} &= 0, \ldots, min(n_{2i} - k_{12i}, n_{3i} - k_{13i}).
\end{aligned}$$

For the covariance parameters we have

$$\begin{aligned}
\theta_{12} \mid \cdot &\sim Gamma(a_1 + \sum k_{12i}, b_1 + n) \\
\theta_{13} \mid \cdot &\sim Gamma(a_2 + \sum k_{13i}, b_2 + n) \\
\theta_{23} \mid \cdot &\sim Gamma(a_3 + \sum k_{23i}, b_3 + n)
\end{aligned}$$

while for the regression parameters we have that

$$\begin{aligned}
\beta_1 \mid \cdot &\propto \exp\left(-\sum_{i=1}^{n} \exp(\beta_1 x_i)\right) \exp\left(\sum_{i=1}^{n} \beta_1 x_i (n_{1i} - k_{12i} - k_{13i})\right) \\
\beta_2 \mid \cdot &\propto \exp\left(-\sum_{i=1}^{n} \exp(\beta_2 x_i)\right) \exp\left(\sum_{i=1}^{n} \beta_2 x_i (n_{2i} - k_{12i} - k_{23i})\right) \\
\beta_3 \mid \cdot &\propto \exp\left(-\sum_{i=1}^{n} \exp(\beta_3 x_i)\right) \exp\left(\sum_{i=1}^{n} \beta_3 x_i (n_{3i} - k_{13i} - k_{23i})\right)
\end{aligned}$$

All the above can be used for efficient MCMC. In each iteration we generate the latent variables using a table look up method, we generate the θ_{ij}'s by simply simulating from the gamma densities, and the β's using a Metropolis Hastings algorithm, using a multivariate normal proposal (i.e. random walk Metropolis). For the covariance of the multivariate normal proposal, in order to achieve good mixing properties we used the covariance matrix of the parameters derived from simple univariate Poisson regression models. More details can be found in Karlis and Meligkotsidou (2005).

For the common covariance model the situation is very similar in the sense that again the latent variables must be augmented to the observed data. We skip the details, as they can be found in Tsionas (2001).

3.2 Zero-inflated models

When dealing with zero-inflated models, there is one more parameter, the inflation parameter p. We assume for p a Beta(γ, δ) prior. Following standard Bayesian approaches for treating zero inflation models, we assume the existence of another latent variable Z_i, one for each individual which takes the value 1 if the observation is inflated and 0 elsewhere. Obviously for observations that are not of the form $(0, 0, 0)$ i.e. we do not observe a triplet of zeros, $Z_i = 0$. To proceed in the Bayesian way we need to generate Z_i from a Bernoulli distribution with success probability

$$p_i' = \frac{p}{p + (1-p)\exp(-\theta_i)},$$

where p is the current value and $\theta_i = \mu_{1i} + \mu_{2i} + \mu_{3i} + \theta_{12} + \theta_{13} + \theta_{23}$. Then update p by generating a beta random variable from the Beta$(\sum z_i + \gamma, n - \sum z_i + \delta)$ distribution.

The rest of the parameters are updated from the following distributions:

$$\begin{aligned}
\theta_{12} \mid \cdot &\sim Gamma(a_1 + \sum (1 - z_i) k_{12i}, b_1 + n - \sum z_i) \\
\theta_{13} \mid \cdot &\sim Gamma(a_2 + \sum (1 - z_i) k_{13i}, b_2 + n - \sum z_i) \\
\theta_{23} \mid \cdot &\sim Gamma(a_3 + \sum (1 - z_i) k_{23i}, b_3 + n - \sum z_i)
\end{aligned}$$

while for the regression parameters we have that

$$\beta_1 \mid \cdot \propto \exp\left(-\sum_{i=1}^{n} \exp(\beta_1 x_i (1 - z_i))\right) \exp\left(\sum_{i=1}^{n} \beta_1 (1 - z_i) x_i (n_{1i} - k_{12i} - k_{13i})\right)$$

$$\beta_2 \mid \cdot \quad \propto \quad \exp\left(-\sum_{i=1}^{n}\exp(\beta_2 x_i(1-z_i))\right)\exp\left(\sum_{i=1}^{n}\beta_2(1-z_i)x_i(n_{2i}-k_{12i}-k_{23i})\right)$$

$$\beta_3 \mid \cdot \quad \propto \quad \exp\left(-\sum_{i=1}^{n}\exp(\beta_3 x_i(1-z_i))\right)\exp\left(\sum_{i=1}^{n}\beta_3(1-z_i)x_i(n_{3i}-k_{13i}-k_{23i})\right)$$

One can easily see that the steps in the above algorithm are actually very similar to the previous ones. Similar ideas like the random walk Metropolis are applicable. To our knowledge, the MCMC scheme derived above for zero-inflated multivariate Poisson models is novel.

4 The database

The original database is a random sample of the automobile portfolio of a major insurance company operating in Spain in 1996. Only cars categorized as being for private use were considered. The data contains information from 20,000 policyholders. The sample is not representative of the actual portfolio as it was drawn from a larger panel of policyholders who had been customers of the company for at least seven years; however, it will be helpful for illustrative purposes.

Twelve exogenous variables were considered plus the yearly number of accidents recorded for the three types of claim. For each policy, the initial information at the beginning of the period and the total number of claims from policyholders at fault were reported within this yearly period. The exogenous variables, described in Table 3, were previously used in Pinquet *et al.* (2001), Brouhns *et al.* (2003), Bolancé *et al.* (2003, 2008), Boucher *et al.* (2007, 2009), Boucher and Denuit (2008) and Bermúdez (2009).

For this study, all customers had held a policy with the company for at least three years. Therefore, variable *v7* was rejected and variable *v8* retained its definition and its baseline was now established as a customer who had been with the company for fewer than five years.

The meaning of those variables referring to the policyholders' coverage should also be clarified. The classification here responds to the most common types of automobile insurance policies available on the Spanish market. The simplest policy only includes third-party liability (claimed and counted as N_1 type) and a set of basic guarantees such as emergency roadside assistance, legal assistance or insurance covering medical costs (claimed and counted as N_3 type). This simplest policy does not include comprehensive coverage or collision coverage (claimed and counted as N_2 type). This simplest type of policies makes up the baseline group, while variable *v10*

Variable	Definition
v1	equals 1 for women and 0 for men
v2	equals 1 when driving in urban area, 0 otherwise
v3	equals 1 when zone is medium risk (Madrid and Catalonia)
v4	equals 1 when zone is high risk (Northern Spain)
v5	equals 1 if the driving license is between 4 and 14 years old
v6	equals 1 if the driving license is 15 or more years old
v7	equals 1 if the client is in the company between 3 and 5 years
v8	equals 1 if the client is in the company for more than 5 years
v9	equals 1 of the insured is 30 years old or younger
v10	equals 1 if includes comprehensive coverage (except fire)
v11	equals 1 if includes comprehensive and collision coverages
v12	equals 1 if horsepower is greater than or equal to 5500cc

Table 3: Explanatory variables used in the models.

denotes policies which, apart from the guarantees contained in the simplest policies, also include comprehensive coverage (except fire), and variable *v11* denotes policies which also include fire and collision coverage.

5 Results

We fitted the models described in section 2 in the database described in the previous section by programming functions in R to implement MCMC algorithms.

5.1 Computational details

We ran the MCMC algorithm for each model 110,000 iterations and used the first 10,000 as a burn-in period. For the remaining 100,000 iterations we sampled every 100th value to remove autocorrelation. All values passed standard diagnostic test (we used CODA) for convergence. The autocorrelation was not significant in any lag.

In all models we used non-informative priors, by considering diffuse priors with large variance. Recall that due to the large sample size ($n = 20,000$) the effect of the prior is negligible.

5.2 Posterior summaries

The Tables 4, 5, 6 and 7 present the results for the fitted models. We also report the 95% credible interval so as to give an idea of the uncertainty around the reported mean value. A variable is considered to be relevant as a predictor of the number of claims when the zero value is not included in the 95% credible interval (significant parameters are marked in boldface in the Tables).

In general, no substantial differences regarding the coefficients were found between the four models considered here. However, there were some differences between zero-inflated models and non-inflated models, probably due to the relatively large value for the inflation parameter p.

If we focus on the claims for third-party liability (N_1), the parameters *v2* and *v8* were significant for all the models, *v9* was significant for all of them except ZICC, and finally *v6* and *v10* were almost significant in most cases. The results suggest that driving in an urban area (*v2*), driving experience (*v6*), drivers with more than 5 years with the company (*v8*) and including comprehensive coverage except fire (*v10*) caused the expected number of third-party liability claims to decrease. However, the expected number of claims was higher in young drivers (*v9*) than in older drivers.

If we focus on the number of claims for automobile collision (N_2), the parameters found significant for all the models were *v3*, *v4*, *v10* and *v11*. Moreover, parameters *v1* and *v8* were significant for the CC model and almost significant for the others. For this type of claim, we may conclude that driving in northern Spain (*v4*) and drivers with fewer than 5 years with the company (*v8*) reduced the expected number of claims, while women drivers (*v1*), drivers from Madrid and Catalonia (*v3*), and the inclusion of comprehensive coverage except fire (*v10*) or collision coverage (*v11*) increased the expected number of claims.

Finally, when looking at the number of claims related to the rest of automobile guarantees (N_3), parameters *v5*, *v8*, and parameters *v10* to *v12* were significant for all models, while parameter *v2* was significant only for FC model but almost significant for the rest of models. As in the case of N_1 and N_2 claims, drivers with more than 5 years in the company (*v8*) caused the expected number of claims to decrease. However, driving in an urban area (*v2*), drivers with intermediate experience (*v5*), the inclusion of comprehensive coverage except fire (*v10*) or collision coverage (*v11*) and vehicles with horsepower greater than or equal to 5500cc (*v12*)

	N_1			N_2			N_3		
	Coeff.	95% credible int.		Coeff.	95% credible int.		Coeff.	95% credible int.	
Intercept	**-2.098**	-2.457	-1.754	**-6.729**	-7.556	-5.890	**-4.663**	-5.263	-4.130
v1	0.004	-0.127	0.142	**0.184**	0.016	0.351	-0.089	-0.270	0.080
v2	**-0.133**	-0.232	-0.020	0.043	-0.099	0.185	0.113	-0.020	0.253
v3	0.025	-0.093	0.152	**0.363**	0.216	0.504	0.023	-0.102	0.161
v4	0.046	-0.092	0.173	**-0.330**	-0.517	-0.131	-0.094	-0.255	0.061
v5	-0.126	-0.426	0.190	0.320	-0.234	0.936	**0.570**	0.123	1.085
v6	-0.284	-0.599	0.043	0.267	-0.328	0.959	0.282	-0.171	0.821
v8	**-0.219**	-0.344	-0.096	**-0.187**	-0.383	-0.014	**-0.196**	-0.364	-0.029
v9	**0.210**	0.021	0.397	0.036	-0.255	0.310	-0.005	-0.250	0.215
v10	-0.110	-0.255	0.037	**5.060**	4.509	5.642	**1.290**	1.055	1.513
v11	0.020	-0.092	0.121	**2.497**	1.936	3.073	**1.777**	1.597	1.950
v12	0.062	-0.070	0.185	-0.043	-0.277	0.183	**0.366**	0.161	0.581
θ_{12}	0.00161	0.00109	0.00220						

Table 4: Results for the Common Covariance model.

	N_1			N_2			N_3		
	Coeff.	95% credible int.		Coeff.	95% credible int.		Coeff.	95% credible int.	
Intercept	**-2.064**	-2.442	-1.714	**-6.761**	-7.689	-5.900	**-4.963**	-5.612	-4.331
v1	0.030	-0.114	0.175	**0.177**	0.009	0.345	-0.061	-0.237	0.109
v2	**-0.128**	-0.243	-0.017	0.048	-0.099	0.190	**0.168**	0.031	0.304
v3	0.022	-0.118	0.152	**0.357**	0.204	0.514	0.017	-0.130	0.185
v4	0.023	-0.114	0.153	**-0.328**	-0.528	-0.107	-0.125	-0.312	0.061
v5	-0.210	-0.507	0.096	0.300	-0.280	0.939	0.504	-0.051	1.126
v6	**-0.359**	-0.668	-0.024	0.254	-0.330	0.911	0.218	-0.320	0.839
v8	**-0.229**	-0.358	-0.084	-0.176	-0.362	0.018	**-0.219**	-0.390	-0.056
v9	**0.226**	0.026	0.412	0.027	-0.244	0.304	0.002	-0.261	0.237
v10	-0.170	-0.371	0.002	**5.100**	4.593	5.700	**1.607**	1.355	1.855
v11	-0.054	-0.170	0.064	**2.486**	1.977	3.116	**2.002**	1.811	2.209
v12	0.041	-0.097	0.175	-0.037	-0.250	0.195	**0.385**	0.164	0.606
θ_{12}	0.00187	0.00088	0.00295						
θ_{13}	0.00749	0.00619	0.00889						
θ_{23}	0.00008	0.00000	0.00058						

Table 5: Results for the Full Covariance model.

increased the expected number of claims.

Another interesting point highlighted by the tables is the fact that the inflation parameter is rather large for both models, while some of the covariance parameters are close to zero. To explore this result further, we consider the histograms in Figure 1. The figure depicts the posterior histograms for the covariance parameters $\theta_{12}, \theta_{13}, \theta_{23}$ and p for the ZIFC model. One can easily see that while parameters θ_{12} and θ_{23} are in fact zero, θ_{13} is clearly non-zero, implying that the independent Poisson model is inappropriate. Moreover, zero inflation is evident due to the large value of p. The fact that some of the covariances are close to zero implies that more refined modelling can be made by considering different structures for the multivariate Poisson

	N_1			N_2			N_3		
	Coeff.	95% credible int.		Coeff.	95% credible int.		Coeff.	95% credible int.	
Intercept	**-0.789**	-1.165	-0.399	**-5.384**	-6.318	-4.553	**-3.305**	-3.939	-2.682
v1	-0.029	-0.161	0.116	0.111	-0.056	0.257	-0.134	-0.297	0.043
v2	**-0.136**	-0.247	-0.021	-0.045	-0.197	0.102	0.110	-0.026	0.254
v3	-0.053	-0.175	0.068	**0.264**	0.113	0.424	-0.075	-0.223	0.065
v4	0.081	-0.054	0.216	**-0.231**	-0.449	-0.047	-0.033	-0.194	0.130
v5	-0.120	-0.468	0.185	0.366	-0.164	0.979	**0.540**	0.037	1.092
v6	-0.260	-0.657	0.042	0.332	-0.257	0.964	0.290	-0.252	0.891
v8	**-0.189**	-0.328	-0.053	-0.129	-0.315	0.081	**-0.176**	-0.341	-0.008
v9	0.195	-0.018	0.378	0.051	-0.269	0.308	-0.025	-0.260	0.190
v10	-0.121	-0.280	0.027	**4.927**	4.420	5.447	**1.298**	1.080	1.511
v11	0.000	-0.111	0.117	**2.455**	1.959	3.008	**1.754**	1.588	1.955
v12	0.037	-0.100	0.183	-0.059	-0.282	0.176	**0.318**	0.114	0.543
θ_0	0.00065	0.00007	0.00185						
p	0.721	0.705	0.737						

Table 6: Results for the Z-I Common Covariance model.

	N_1			N_2			N_3		
	Coeff.	95% credible int.		Coeff.	95% credible int.		Coeff.	95% credible int.	
Intercept	**-0.837**	-1.263	-0.414	**-5.366**	-6.309	-4.528	**-3.543**	-4.282	-2.880
v1	-0.014	-0.153	0.129	0.111	-0.063	0.293	-0.119	-0.307	0.069
v2	**-0.132**	-0.248	-0.015	-0.042	-0.191	0.112	0.125	-0.027	0.253
v3	-0.051	-0.170	0.076	**0.271**	0.131	0.425	-0.063	-0.229	0.087
v4	0.087	-0.046	0.234	-0.230	-0.435	0.000	-0.022	-0.207	0.156
v5	-0.137	-0.491	0.186	0.377	-0.254	1.138	**0.548**	0.038	1.129
v6	-0.275	-0.643	0.076	0.336	-0.354	1.097	0.285	-0.260	0.907
v8	**-0.197**	-0.337	-0.042	-0.127	-0.317	0.066	-0.174	-0.353	0.003
v9	**0.209**	0.009	0.391	0.063	-0.233	0.369	-0.008	-0.244	0.225
v10	-0.107	-0.282	0.056	**4.871**	4.411	5.374	**1.433**	1.214	1.670
v11	0.010	-0.116	0.130	**2.388**	1.922	2.919	**1.891**	1.726	2.088
v12	0.040	-0.107	0.179	-0.046	-0.255	0.177	**0.361**	0.151	0.600
θ_{12}	0.00002	0.00000	0.00017						
θ_{13}	0.00788	0.00377	0.01265						
θ_{23}	0.00000	0.00000	0.00001						
p	0.715	0.698	0.728						

Table 7: Results for the Z-I Full Covariance model.

model described in section 2. We will return to this issue in the discussion.

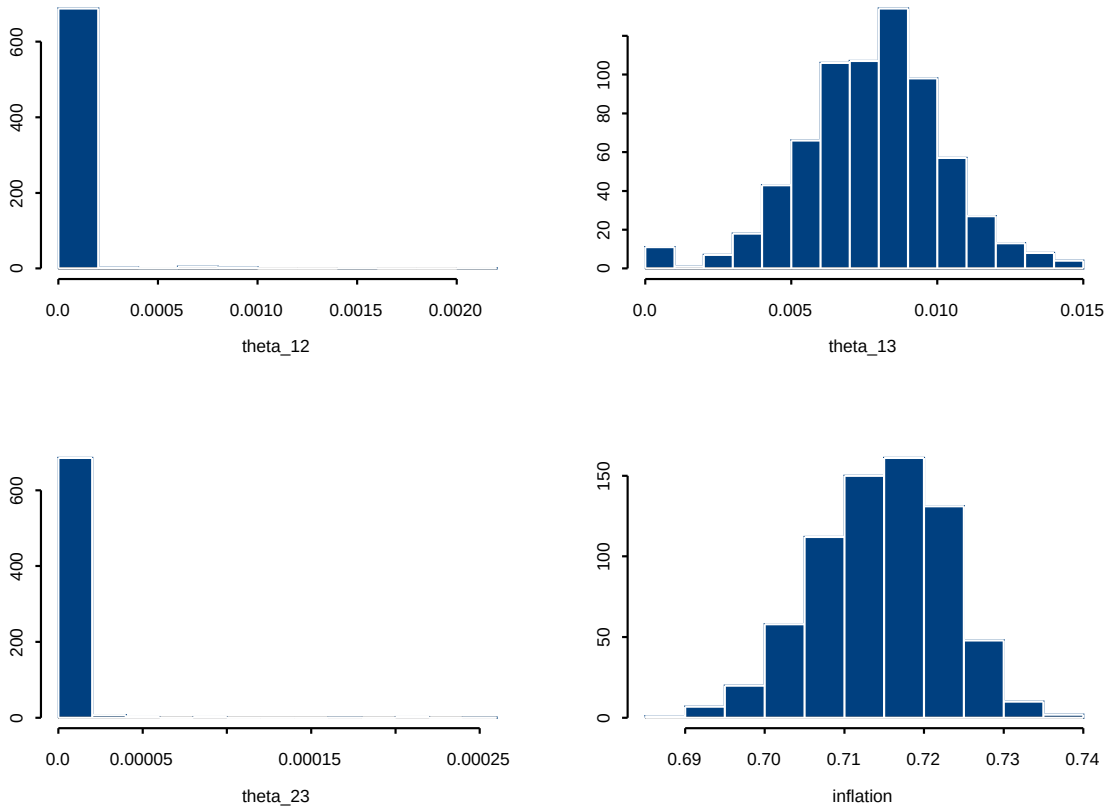

Figure 1: Histograms of the posterior values for the covariance and the zero inflation parameters based on the ZIFC model.

A final interesting point is to examine the goodness of fit of the data. As this measure we calculated the implied probability $P(0,0,0)$, i.e. the probability of no claims, for all models, and for each iteration. Note that this was the largest frequency in the data. Figure 2 shows the estimated probability of no claims for the four models. It is clear that without zero inflation the models fail to take this into account and hence the estimated probability distribution is bad. Adding a zero inflation parameter the probability is much larger, in fact very close to the true one (0.87). For other probabilities, again the zero-inflated models provide a much better fit than the simple models.

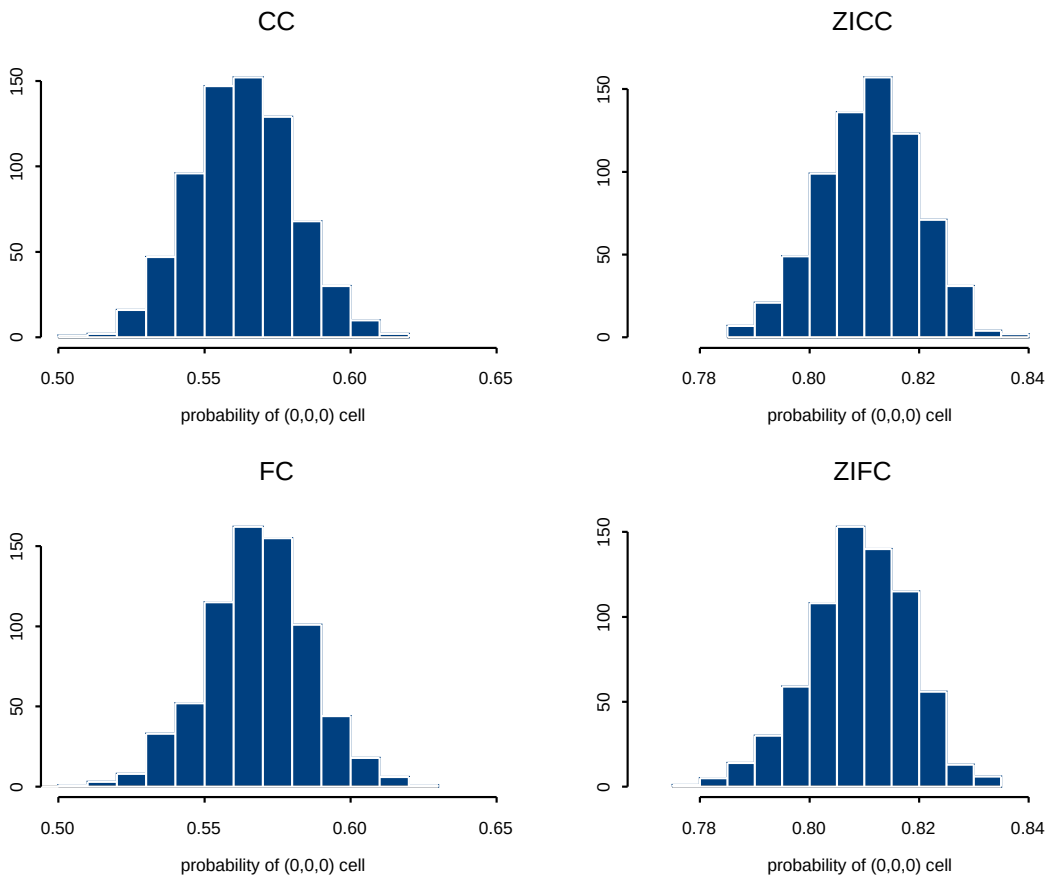

Figure 2: Boxplots of the posterior probability of no claims for each model.

5.3 Actuarial implications

At the same time we analyze the impact of using these models in *a priori* ratemaking. The differences between the models proposed in Section 2 were analyzed through the mean (*a priori* pure premium) and the variance (necessary for *a priori* loaded premium) of the number of claims per year for some profiles of the insured parties. Five different, yet representative, profiles were selected from the portfolio and classified according to their risk level. The profiles (see Table 8) are the same as in Bermúdez (2009). The first can be classified as the best profile since it presents the lowest mean score. The second was chosen from among the profiles considered as good drivers, with a lower mean value than the mean of the portfolio. The third profile was chosen with a mean score lying very close to the mean of the portfolio. Finally, a profile considered as being a bad driver (with a mean score above the mean of the portfolio) and the

Profile	v1	v2	v3	v4	v5	v6	v8	v9	v10	11	v12
Best	0	1	0	0	0	1	1	0	0	0	0
Good	0	0	1	0	0	1	0	0	0	0	1
Average	0	1	0	0	0	1	1	0	0	1	1
Bad	0	0	0	1	0	1	1	0	1	0	0
Worst	1	1	1	0	1	0	0	1	1	0	1

Table 8: Hypothetical profiles of clients to be used for illustration.

	Best		Good		Average		Bad		Worst	
Model	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
IP	0.0791	0.0791	0.1194	0.1194	0.1882	0.1882	0.2799	0.2799	0.7011	0.7011
CC	0.0832	0.0930	0.1268	0.1365	0.1896	0.1993	0.2573	0.2671	0.6868	0.6966
FC	0.0907	0.1096	0.1293	0.1482	0.1856	0.2045	0.2608	0.2797	0.6817	0.7006
ZICC	0.0856	0.3121	0.1169	0.4291	0.1918	0.7416	0.2928	1.1741	0.5381	2.3051
ZIFC	0.0845	0.2902	0.1151	0.4017	0.1920	0.7138	0.2915	1.1288	0.5542	2.3168

Table 9: Comparison of different models and different profiles (IP: independent Poisson).

worst driver profile were selected.

Table 9 shows the results for the five profiles and the five models (the four models considered here plus the independent Poisson model). There are two differences in ratemaking when using a multivariate Poisson model as opposed to the independent Poisson model. First, multivariate Poisson models produce higher means for good risks and lower means for bad risks while maintaining average risks almost equal. Second, multivariate models increase variances in most cases, meaning overdispersion. This is especially noticeable for zero-inflated models whose means are similar to non zero-inflated models, but have much higher variances. Finally, few differences were found between the models with common covariance and the ones with full covariance. Probably this is due to the fact that only covariance between N_1 and N_3 seems to be significant for the zero-inflated full covariance model.

One of the advantages of the MCMC approach is that we can easily obtain samples from the posterior distribution of any quantity of interest and then examine the variability of this quantity in a fully inferential way. A summary of the posterior distribution of the premiums for

all the models and for four different profiles (we skip over the "bad" profile for reasons of space) can be seen in Figures 3 and 4. Figure 3 refers to the $E(N)$ and Figure 4 to $V(N)$. Hence all the details about the premium are available and not just a point estimate.

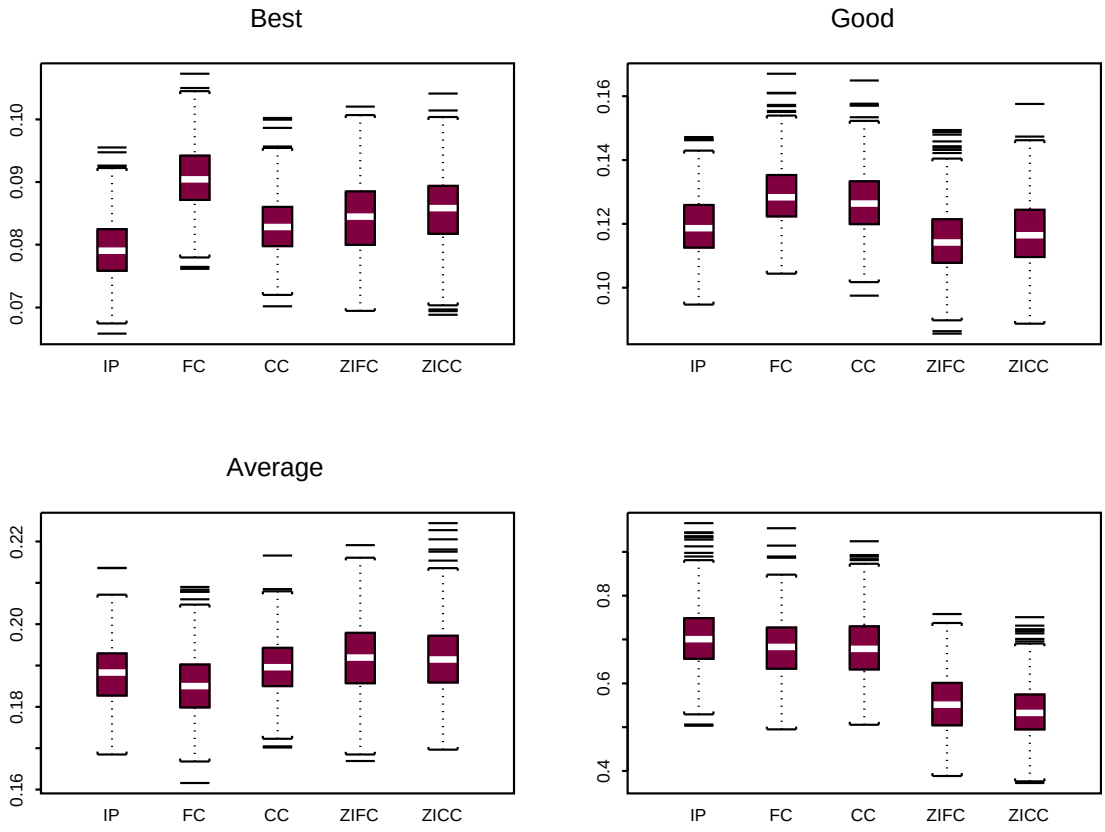

Figure 3: Boxplots of the posterior values for $E(N)$ derived from the MCMC.

When considering $E(N)$ the four models do not differ so much for the four selected profiles. For the "worst" profile the differences are greater, especially for the zero-inflated models, the reason being that a better estimate for the non-claims case exists and hence the mean is not overestimated as it is when we assume less probability at the (0,0,0) case.

It is worth to go further with this issue. In order to emphasize the importance of refined modelling for non-claims, consider the five different profiles and the model with full covariance in their two variants (FC and ZIFC). For each profile we calculated the probability of no-claims for each iteration, i.e. the posterior summary for $P(0,0,0)$. Boxplots in Figure 5 show these posterior probabilities for each profile and for the two models considered. It is clear

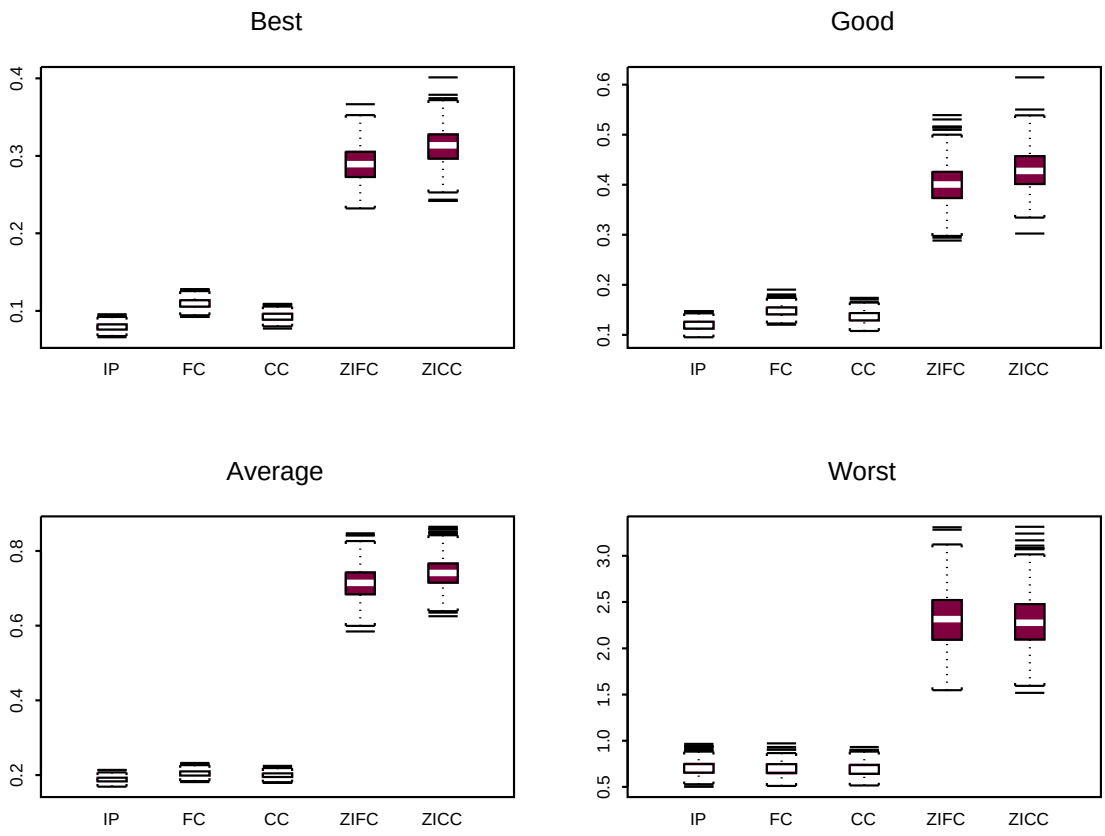

Figure 4: Boxplots of the posterior values for $V(N)$ derived from the MCMC.

that there are large differences between the different profiles. As expected, the "best" profile has by far the greatest no-claim probability, while for the "worst" profile the probability is much less. However, the probability of no-claims when using non zero-inflated model (FC) is underestimated for all profiles and especially significant for the "worst" profile. This fact would explain the greater differences observed for the pure premium in Figure 3. The best estimation of no-claims probability for the "worst" profile can be used for example in order to give bonuses to claim-free clients. Several other quantities can easily be deduced from the Bayesian output and used to obtain a better understanding and pricing of the clients.

The case of $V(N)$ is considered is much more variable than expected. For this application, zero-inflated models have larger variability. Recall that when zero-inflated models are considered, the marginal distributions are no longer Poisson and hence the assumed variance is larger. This is depicted in the estimation of $V(N)$. Note also that this is more realistic, as the Poisson

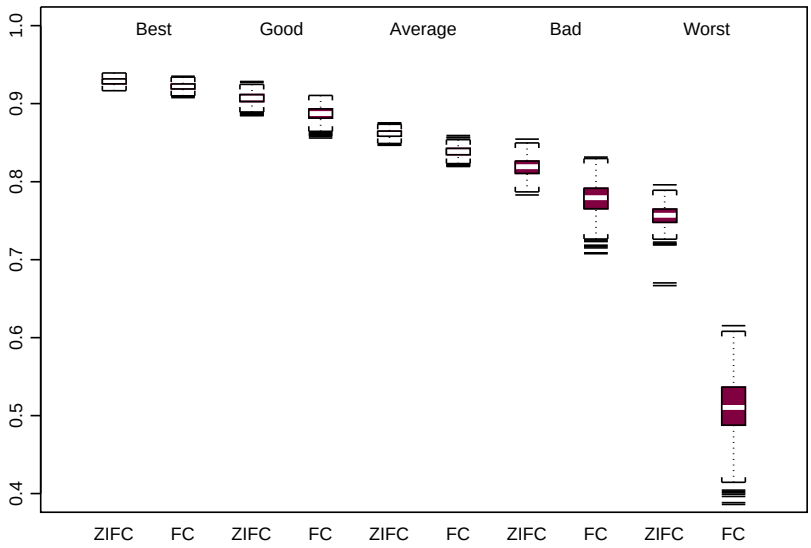

Figure 5: Boxplots of the posterior probability of no claims for each profile, based on the FC and ZIFC models.

assumption seems to be violated. Hence calculations based on this better $V(N)$ are more trustworthy.

For illustration, Table 10 presents selected percentiles of the calculated premium $E(N)$ for all the models and selected profiles. We can assume that the 75% percentile corresponds to the loaded premium, while the other two percentiles can be of interest for VaR approaches. If we focus on the calculated loaded premium for the "worst" profile, we observe that zero-inflated models present lower loaded premiums. Therefore, the observed increase in variance is compensate for the decrease in mean.

5.4 Model Selection

One interesting question that needs to be answered is to establish which model is better for describing the dataset. Recall that in practice the different models have different underlying assumptions. Table 11 presents different information criteria for model selection. Together with classical AIC and BIC that penalize with respect to the number of parameters to be estimated we also report Deviance Information Criterion (DIC) (Spiegelhalter et al, 2002) as this criterion

	Best					Good				
	IP	CC	FC	ZICC	ZIFC	IP	CC	FC	ZICC	ZIFC
75%	0.0823	0.0864	0.0942	0.0895	0.0885	0.1256	0.1333	0.1355	0.1241	0.1215
90%	0.0847	0.0897	0.0972	0.0927	0.0927	0.1313	0.1391	0.1425	0.1308	.1286
95%	0.0870	0.0917	0.0997	0.0946	0.0944	0.1368	0.1427	0.1462	0.1346	0.1321
	Average					Worst				
	IP	CC	FC	ZICC	ZIFC	IP	CC	FC	ZICC	ZIFC
75%	0.1931	0.1943	0.1905	0.1972	0.1979	0.7440	0.7323	0.7274	0.5775	0.5985
90%	0.1975	0.1990	0.1947	0.2033	0.2040	0.7991	0.7800	0.7762	0.6233	0.6423
95%	0.2009	0.2014	0.1975	0.2078	0.2074	0.8230	0.8179	0.8008	0.6475	0.6678

Table 10: Percentile points for the calculated premium for the different models and profiles.

Model	DIC	BIC	AIC
Independent Poisson	25959.42	26198.64	26073.16
Common Covariance	25863.82	26105.48	25978.78
Full Covariance	25651.56	25896.50	25768.16
Z-I Common Covariance	24256.6	24497.06	24370.96
Z-I Full Covariance	24240.46	24484.08	24356.40

Table 11: Different Information Criteria for selecting the best model for the data.

makes use of the number of "effective" parameters, which in our case is considerable due to the latent structure imposed.

All the criteria agree that the zero-inflated model with full covariance structure is the best model. It is also clear that the zero-inflated models are far better than the common covariance model without zero inflation. This can be also be seen from the values of the inflation parameter, which is very large. We do not recommend the use of a common covariance term for this application.

6 Conclusions

In the present paper we considered multivariate Poisson models and their zero-inflated extensions allowing for correlation between the different types of claims in order to improve the ratemaking

procedure. Multivariate Poisson models differ in the covariance structures that they present. We propose the analysis of a multivariate Poisson model with common covariance parameter and a multivariate Poisson model with a full covariance specification. Finally, zero-inflated versions of the previous models are considered in order to account for the excess of zeros and the overdispersion observed in automobile insurance databases.

We apply a Bayesian approach which offers advantages over standard methods; namely, estimation is feasible even for complicated models with a large number of parameters, and it allows for the use of prior information. Finally we emphasize that the Bayesian approach allows better calculation of certain quantities of interest (e.g. the loaded premium) since we obtain a distribution and not just a point estimate. This also helps to account for the uncertainty around the quantities of interest. From the modelling perspective, the derived MCMC scheme for zero-inflated multivariate Poisson regression models is novel.

The interpretation of a number of multivariate Poisson models is illustrated in the context of automobile insurance claims using a large data set. The conclusion is that even when there are small correlations between the counts, major differences in ratemaking may appear. In general, when considering the mean (*a priori* pure premium) of the number of claims per year, the expected number of claims given by the multivariate Poisson models does not differ much from the independent Poisson model; but when the variance (used for *a priori* loaded premium) is considered, larger variances are obtained with zero-inflated multivariate Poisson models and hence larger loadings in premiums must be included. Recall that when zero-inflated models are considered in order to better model the excess of (0,0,0) occurrences the marginal distributions are no longer Poisson and hence the assumed variance is larger.

From this conclusion, one can understand that the obtained loaded premiums with zero-inflated multivariate Poisson models it would be larger than those obtained with the independence assumption. However, this is not true for all policyholders, since in some cases the reduction in mean caused by account for the excess of zeros is larger than the increase in variance caused by account for the overdispersion.

All the criteria considered here agree that the zero-inflated model with full covariance structure is the best model for describing the data set and hence for use in a ratemaking procedure. This implies that claim-free clients who appear more often in the database than would be expected by a standard multivariate Poisson model. The estimated probability of claim-free clients

with this model is by far the closest to the true one. This model, as a zero-inflated model, also corrects for overdispersion relative to simple Poisson models both in the marginals and the joint distribution and hence allow us to take heterogeneity issues into account. This is especially relevant in risk measuring (Solvency II in EU).

In addition, closer examination of the fitted model reveals that one may check for the structure implied by the multivariate Poisson model considered, and derive models that better capture the structure. For example, in our case there is evidence that for some pairs the correlation is almost zero, implying that such terms can be removed. In our case going back to the model in section 2, we found that covariance between N_1 and N_3 was the only one was significant. Hence one may consider a model where only this covariance term is kept. For this data set, according to the results obtained, we may also reduce the ratemaking problem to a bivariate Poisson model considering N_2 and N_3 together, as in Bermúdez (2009). This fact does not invalidate the use of multivariate Poisson models in automobile insurance claims, since we may include other type of claims for consideration. It is worth mentioning here that it is very interesting to distinguish in third-liability claims between vehicle damage and personal injury claims, which are expected to be positively correlated. However, it was not possible to obtain the information needed for this purpose from the available data set.

A direct application for the models presented here is that the predictive distributions can be obtained to estimate financial risk measures for portfolios of policies (VaR, CTE, etc.) as in Frees et al. (2009). With the Bayesian approach to multivariate Poisson models through MCMC, we can easily derive posterior summaries for several quantities of interest to account for uncertainty. We did not pursue this problem here as it is beyond the scope of the paper.

Finally, we would like to mention some extensions for this paper. The first would be to include the cost of claims in the ratemaking procedure, in a similar way as in Frees and Valdez (2008). The second one, using a larger database panel of policyholders with information for several years, would be to model the time dependence as well. The third one refers to extending the model used in several dimensions like a) the use of covariates for the zero inflation parameter, i.e. to assume that certain covariates also affect this parameter, b) considering some model to allow for more heterogeneity like bivariate negative binomial regression models and c) by considering some variable selection approach in order to select the covariates. The latter can be based on some stepwise technique taking, however, into account that the model allows different covariates

to each dependent variable and hence the procedures should be adapted appropriately.

7 References

Berkhout, P., Plug, E., 2004. A bivariate Poisson count data model using conditional probabilities. Statistica Neerlandica 58, 349–364.

Bermúdez, L., 2009. A priori ratemaking using bivariate Poisson regression models. Insurance Mathematics & Economics 44 (1), 135–141.

Bohning D., Dietz E., Schlattmann P., Mendonca L. and Kirchner U., 1999. The Zero-inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology. Journal of the Royal Statistical Society, Series A, 162, 195–209.

Bolancé, C., Guillén, M., Pinquet, J., 2003. Time-varying credibility for frequency risk models: Estimation and tests for autoregressive specification on the random effect. Insurance: Mathematics & Economics 33 (2), 273–282.

Bolancé, C., Guillén, M., Pinquet, J., 2008. On the link between credibility and frequency premium. Insurance: Mathematics & Economics 43 (2), 209–213.

Boucher, J.-Ph., Denuit, M., 2006. Fixed versus random effects in Poisson regression models for claim counts: a case study with motor insurance. ASTIN Bulletin 36 (1), 285–301.

Boucher, J.-Ph., Denuit, M., Guillén, M., 2007. Risk classification for claims counts: a comparative analysis of various zero-inflated mixed Poisson and Hurdle models. North American Actuarial Journal 11 (4), 110–131.

Boucher, J.-Ph., Denuit, M., 2008. Credibility premiums for the zero inflated Poisson model and new hunger for bonus interpretation. Insurance: Mathematics & Economics 42 (2), 727–735.

Boucher, J.-Ph., Denuit, M., Guillén, M., 2009. Number of accidents or number of claims? An approach with zero-inflated Poisson models for panel data. Journal of Risk and Insurance 76 (4), 821–846.

Brouhns, N., Denuit, M., Guillén, M., Pinquet J., 2003. Bonus-malus scales in segmented tariffs with stochastic migration between segments. Journal of Risk and Insurance 70 (4), 577–599.

Cameron, A.C., Trivedi, P.K., 1998. Regression analysis of count data. Econometric Society Monograph 30, Cambridge University Press.

Chib, S., Winkelmann, R., 2001. Markov chain Monte Carlo analysis of correlated count data. Journal of Business and Economic Statistics 19, 428–435.

Dean, C.B., 1992. Testing for Overdispersion in Poisson and Binomial Regression Models. Journal of the American Statistical Association 87 (418), 451–457.

Denuit, M., Maréchal, X., Pitrebois, S., Walhin, J.F., 2007. Actuarial modelling of claim counts. London: John Wiley & Sons.

Frees, E.W., Valdez, E.A., 2008. Hierarchical insurance claims modelling. Journal of the American Statistical Association 103 (484), 1457–1469.

Frees, E.W., Shi, P., Valdez, E.A., 2009. Actuarial Applications of a Hierarchical Insurance Claims Model. ASTIN Bulletin 39 (1), 165–197.

Johnson, N., Kotz, S., Balakrishnan, N., 1997. Multivariate Discrete Distributions. New York: John Wiley & Sons.

Karlis, D., Ntzoufras, I., 2003. Analysis of sports data using bivariate Poisson models. Journal of the Royal Statistical Society (Statistician) 52, 381–393.

Karlis, D., Ntzoufras, I., 2005. Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R. Journal of Statistical Software 14 (10), 1–36.

Karlis, D., Meligkotsidou, L., 2005. Multivariate Poisson regression with full covariance structure. Statistics and Computing 15 (4), 255–265.

Karlis, D., Meligkotsidou, L., 2007. Finite multivariate Poisson mixtures with applications. Journal of Statistical Planning and Inference 137, 1942–1960.

Li, J.C, Park, J., Kim, K., Brinkley, P.A., Peterson, J.P., 1999. Multivariate Zero-Inflated Poisson Models and Their Applications. Technometrics 41, 29–38.

Nikoloulopoulos, A.K., Karlis, D., 2009. Finite normal mixture copulas for multivariate discrete data modelling. Journal of Statistical Planning and Inference 139, 3878–3890.

Pinquet, J., Guillén, M., Bolancé, C., 2001. Long-range contagion in automobile insurance data: estimation and implications for experience rating. ASTIN Bulletin 31 (2), 337–348.

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

Tsionas, E., 2001. Bayesian multivariate Poisson regression. Communications in Statistics-Theory and Methods 30, 243–255.

Van Ophem, H., 1999. A general method to estimate correlated discrete random variables. Econometric Theory 15, 228–237.

Wang K., Lee A., Yau K., Carrivick P., 2003. A Bivariate Zero-Inflated Poisson Regression Model to Analyze Occupational Injuries. Accident Analysis and Prevention 35, 625–629.

Young G., Valdez E.A., Kohn R., 2009. Multivariate probit models for conditional claim-types. Insurance: Mathematics & Economics 44 (2), 214–228.