Allowing for time and cross dependence assumptions between claim counts in ratemaking models

Lluís Bermúdez[a,*], Montserrat Guillén[b], Dimitris Karlis[c]

[a] *Universitat de Barcelona, Riskcenter-IREA, Departament de Matemàtica Econòmica, Financera i Actuarial, Spain*
[b] *Universitat de Barcelona, Riskcenter-IREA, Departament d'Estadística, Econometria i Economia Aplicada, Spain*
[c] *Athens University of Economics and Business, Department of Statistics, Greece*

Abstract

For purposes of ratemaking, time dependence and cross dependence have been treated as separate entities in the actuarial literature. Indeed, to date, little attention has been paid to the possibility of considering the two together. To discuss the effect of the simultaneous inclusion of different dependence assumptions in ratemaking models, a bivariate INAR(1) regression model is adapted to the ratemaking problem of pricing an automobile insurance contract with two types of coverage, taking into account both the correlation between claims from different coverage types and the serial correlation between the observations of the same policyholder observed over time. A numerical application using an automobile insurance claims database is conducted and the main finding is that the improvement obtained with a BINAR(1) regression model, compared to the outcomes of the simplest models, is marked, implying that we need to consider both time and cross correlations to fit the data at hand. In addition, the BINAR(1) specification shows a third source of dependence to be significant, namely, cross-time dependence.

Keywords: Multivariate longitudinal data, Time dependence, Cross dependence, Automobile insurance, BINAR(1) model

[*] **Corresponding Author.** Departament de Matemàtica Econòmica, Financera i Actuarial, Universitat de Barcelona, Diagonal 690, 08034-Barcelona, Spain. Tel.:+34-93-4034853; fax: +34-93-4034892; e-mail: `lbermudez@ub.edu`

1. Introduction

Insurance ratemaking is one of the main tasks that actuaries perform. To calculate a premium, actuaries will typically obtain the conditional expectation of the number of claims, given a set of observable risk characteristics, and then combine this with the corresponding conditional expectation of claim amounts. As such, modelling insurance claim count data represents an essential part of their task. Indeed, the actuarial science literature contains many studies examining count data models that seek to take into consideration specific features of their data, i.e., unobserved heterogeneity (including overdispersion and excess of zeros) and, more recently, dependence between claim counts.

In this paper, we examine the effect of simultaneously including different dependence assumptions in ratemaking models. Specifically, we focus on three sources of dependence: first, *time dependence*, or the serial dependence between observations of the same policyholder at different points in time; second, *cross dependence*, or the dependence between observations of the same policyholder for different types of claim or coverage; and third, a source of dependence that combines these first two sources, defined as *cross-time dependence*, or the dependence between observations from different types of claim made by the same policyholder at different points in time.

In the context of automobile insurance, the behaviour of a driver is likely to change after they have made a claim and, therefore, some kind of time dependence should be found in a panel count dataset. At the same time, an automobile insurance contract includes different types of guarantees. A third-party liability guarantee is often combined with a set of other guarantees related to driving such as, for example, damage resulting from a collision with another vehicle/object when the policyholder is at fault. In this case, when policyholders make a third-party liability claim, it is common for them also to file a claim on their collision coverage. Hence, the multi-guarantee nature of the insurance contract gives rise to a source of cross dependence. But, at the same time, it is possible that a collision claim reported in the past will influence the number of third-party liability claims reported in the future, giving rise to cross-time dependence.

Time and cross dependence have been widely addressed in the ratemaking literature as separate entities. Traditionally, ratemaking has been tackled in two steps: *a priori* ratemaking and *a posteriori* ratemaking. The first step uses count regression analysis to identify risk factors and to predict

the expected frequency of claims given the observable characteristics of the policyholders. However, not all the factors influencing a risk can be identified, measured and introduced in the a priori tariff. In *a posteriori* ratemaking, actuaries consider the past claims record of each policyholder in order to update their *a priori* premiums, assuming that the number of claims reported by policyholders reveals unobservable risk characteristics, such as driving ability or driver aggression. An exhaustive review of ratemaking systems in automobile insurance using cross-section data can be found in Denuit *et al.* (2007).

However, in recent years, insurers have been able to accumulate longitudinal information on their policyholders. In parallel, a growing body of literature has developed panel count data models applied to the field of insurance. By using these models, actuaries can use repeated observations of each policyholder over time, thus allowing for time dependence. As Boucher and Inoussa (2014) stress, the advantage of using this information when modelling the number of claims is that it becomes possible to estimate premiums that depend simultaneously on risk characteristics and on claim experience and, so, actuaries can avoid the classical two-step approach which is devoid of all coherence in a panel data setting. Following Molenberghs and Verbeke (2005), models for discrete panel data can be classified into three categories: conditional models (e.g. autoregressive and integer-valued autoregressive models), marginal models (e.g. multivariate models with serial correlation) and subject-specific models (e.g. random effects models). An exhaustive overview of such models applied to the actuarial sciences is provided by Boucher *et al.* (2008).

When actuaries are faced with the problem of pricing an insurance contract containing different types of coverage, they usually assume that claim types are independent. However, such an assumption may not be realistic. Indeed, Bermúdez (2009), Bermúdez and Karlis (2011, 2012) and Shi and Valdez (2014) have reported a positive correlation between types of claim and introduced different bivariate (or multivariate) regression models to relax the independence assumption between claims counts arising from the same policy in *a priori* ratemaking and cross-section data settings. They concluded that using a bivariate (or multivariate) regression model provides a better fit, resulting in an *a priori* ratemaking that presents larger variances and, hence, larger loadings than those obtained under the independence assumption.

In short, on the one hand, Boucher *et al.* (2008) and Gourieroux and Jasiak (2004) showed that integer-valued autoregressive (INAR) models are

an acceptable alternative for modelling univariate claim count data when a panel data structure is available for ratemaking, allowing for time dependence (autocorrelation or time series correlation). On the other hand, Bermúdez (2009), Bermúdez and Karlis (2011, 2012) and Shi and Valdez (2014) showed that, when the ratemaking consists of pricing different types of coverage, bivariate (or multivariate) regression models for cross-section data provide a better fit than when using regression models assuming independence, allowing for cross dependence (cross correlation).

The present paper combines these two approaches and extends INAR models for panel claim count data to the bivariate case. More specifically, the bivariate INAR process of order 1, BINAR(1), as introduced by Pedeli and Karlis (2011, 2013), is adapted to the ratemaking problem of pricing an automobile insurance contract with two types of coverage (third-party liability guarantee and other guarantees), taking into account both the cross-correlation between claims from different types of coverage and the serial correlation between the observations of a given policyholder over time. To date, little attention has been given to multivariate longitudinal data analysis for actuarial applications (Shi, 2012); however, Boudreault and Charpentier (2011) apply BINAR(1) to model earthquake counts, but unlike the ratemaking problem, no covariates were included.

In the section that follows, the BINAR(1) regression model is defined. In Section 3, a numerical application using an automobile insurance claims database is presented. Finally, some concluding remarks are given in Section 4.

2. BINAR(1) regression models

2.1. The Models

Let N_1 be the number of claims for third-party liability and N_2 the number of claims for all the other guarantees contained in an automobile insurance. Assuming that for each individual we have data for different time points, we denote as N_{jit} the number of claims for claim type j and i-th individual at time point t, where $j = 1, 2$, $i = 1, \ldots, n$, and $t = 1, \ldots, T_i$ (i.e. we may have different numbers of observations for each client).

The bivariate integer-valued autoregressive process of order 1, BINAR(1), is a generalisation of the simple INAR model introduced by Al-Osh and Alzaid (1987) based on thinning and is described in detail in Pedeli and

Karlis (2011, 2013). It can be defined as

$$\mathbf{N}_t = \mathbf{A} \circ \mathbf{N}_{t-1} + \mathbf{R}_t$$

where $\mathbf{N}$ and $\mathbf{R}$ are non-negative integer-valued random 2-vectors and $\mathbf{A}$ is a 2×2 matrix with independent elements $\{\alpha_{jk}\}_{j,k=1,2}$. It holds that $0 \leq \alpha_{jk} < 1, \quad j,k = 1,2$. The operator '$\circ$', known as the binomial thinning operator, is defined as $\alpha \circ N = \sum_{s=1}^{N} Z_s = Z$ where Z_s are independently and identically distributed Bernoulli random variables with $P(Z_s = 1) = 1 - P(Z_s = 0) = \alpha$ and $\alpha \in [0,1]$. This operator, developed by Steutel and van Harn (1979), mimics the scalar multiplication used for normal time series models so as to ensure that only integer values occur. The elements $\mathbf{R}_t$ that entered the system in the interval $(t-1, t]$ are usually referred to as innovations.

In Pedeli and Karlis (2011) the case with a diagonal matrix $\mathbf{A}$ was examined. For this simpler structure, hereinafter called the Basic BINAR(1) model,

$$\mathbf{A} = \begin{bmatrix} \alpha_{11} & 0 \\ 0 & \alpha_{22} \end{bmatrix},$$

where each series is represented as

$$N_{1it} = \alpha_{11} \circ N_{1i,t-1} + R_{1it}$$
$$N_{2it} = \alpha_{22} \circ N_{2i,t-1} + R_{2it}.$$

As in every INAR-type process, each series N_t is composed of two parts. The first consists of the survivors of the elements of the process at the preceding point in time $t-1$, denoted by N_{t-1}. The autocorrelation derives from this part. The second part consists of the innovations R_t that are assumed to be correlated. The cross correlation derives from the joint distribution assumed for the R_{jt}.

The case with a non-diagonal matrix $\mathbf{A}$ was considered in Boudreault and Charpentier (2011) and in Pedeli and Karlis (2013). This case allows for a more complicated structure and, hence, for a new source of dependence. The case of non-diagonal matrix $\mathbf{A}$, hereinafter called the Full BINAR(1) model, with

$$\mathbf{A} = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix}$$

where each series is now represented as follows:

$$\begin{aligned} N_{1it} &= \alpha_{11} \circ N_{1i,t-1} + \alpha_{12} \circ N_{2i,t-1} + R_{1it} \\ N_{2it} &= \alpha_{22} \circ N_{2i,t-1} + \alpha_{21} \circ N_{1i,t-1} + R_{2it}. \end{aligned}$$

The assumption of the diagonality of matrix $\mathbf{A}$ implies that the correlation between innovations is the only source of dependence between the two series. On relaxing the diagonality assumption, the value of each univariate series at time t is directly associated not only to its own survivors but also to the survivors of the elements of the other series at the preceding point in time $t-1$. Hence, this association forms a second source of dependence, referred to cross autocorrelation in Boudreault and Charpentier (2011).

This more complicated structure allows for the three different sources of dependence defined in the introduction. The number of claims for third-party liability at time t (N_{1t}) is thus correlated with the number of claims for third-party liability at time $t-1$ ($N_{1,t-1}$), the number of claims for all the other guarantees at time t (N_{2t}), and the number of claims for all the other guarantees at time $t-1$ ($N_{2,t-1}$) - that is, autocorrelation, cross correlation and cross autocorrelation, respectively.

In this paper, we assume that both innovation terms (R_{1it}, R_{2it}) jointly follow a bivariate Poisson distribution with parameters λ_{1it}, λ_{2it} and ϕ_{it} with the joint probability mass function being given by:

$$P(R_{1it}=x,\, R_{2it}=y) = e^{-(\lambda_{1it}+\lambda_{2it}+\phi_{it})}\,\frac{(\lambda_{1it}-\phi_i t)^x}{x!}\,\frac{(\lambda_{2it}-\phi_{it})^y}{y!} \times \sum_{s=0}^{\min(x,y)} \binom{x}{s}\binom{y}{s}\, s!\, \left(\frac{\phi_{it}}{(\lambda_{1it}-\phi_{it})(\lambda_{2it}-\phi_{it})}\right)^s.$$

The bivariate Poisson distribution defined above allows for positive dependence between the random variables N_1 and N_2, which is what we expect for claims of this type. In the case of negatively correlated claims (a case not considered here) a more general specification would be necessary. Moreover, ϕ_{it} is a measure of this dependence at time t. Obviously, if $\phi_{it}=0$ the two random variables are independent and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions. For a comprehensive treatment of the bivariate Poisson distribution, the reader is referrred to Kocherlakota and Kocherlakota (1992). In Pedeli and Karlis (2011), the model with bivariate negative binomial innovations used in modelling overdispersed bivariate time series, and some additional specifications for time series data with negative correlation, are also considered.

Despite using a bivariate Poisson distribution, the Full BINAR(1) model allows for overdispersion since its marginals are no longer Poisson distributions. This is not the case for the Basic BINAR(1) model.

Finally note that parameters α, ϕ and λ's are defined as client dependent and, so, can be related to some covariate information. For ratemaking purposes, we seek to introduce covariates and, thus, we further assume that

$$\begin{aligned} \log \lambda_{1it} &= \mathbf{x}_{it}\beta_1 \\ \log \lambda_{2it} &= \mathbf{z}_{it}\beta_2 \end{aligned}$$

where $\mathbf{x_{it}}$ and $\mathbf{z_{it}}$ are time dependent covariates for the i-th individual, not necessarily the same, and β's are the relevant regression coefficients. To keep the model parsimonious, we do not assume covariates for α's and ϕ and we also assume that they keep constant across time. We also denote as $\alpha = (\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22})$.

2.2. Fitting the models

As in Pedeli and Karlis (2013) the conditional likelihood for a BINAR(1) model is the convolution of binomials and a bivariate Poisson and, hence, the contribution of the i-th policyholder is

$$L_i = \prod_{t=1}^{T_i} f(n_{1it}, n_{2it} | n_{1i,t-1}, n_{2i,t-1}, \alpha, \lambda_{1it}, \lambda_{2it}, \phi)$$

In this formula, $f(n_{1it}, n_{2it}|\cdot)$ represents the conditional distribution at time t given the previous time point. In the general case of non-diagonal matrix $\mathbf{A}$, this bivariate distribution is represented as a convolution of four binomial random variables and a bivariate Poisson, and, hence, it involves a quadruple summation

$$\sum_{k=0}^{n_{1it}} \sum_{s=0}^{n_{2it}} \sum_{m=0}^{n_{1it}-k} \sum_{\ell=0}^{n_{2it}-s} f_1(n_{1it}-k) f_2(n_{2it}-s) f_3(n_{1it}-k-m) f_4(n_{2it}-s-\ell) f_5(k,s)$$

with f_j, $j = 1, \ldots, 4$ being the probability function of a binomial distribution, while $f_5(\cdot, \cdot)$ is the probability function of a bivariate Poisson distribution. Based on all of these, the full likelihood is simply the product of the likelihood for each client, i.e.

$$\ell(\mathbf{\Theta}) = \prod_{i=1}^{n} L_i$$

Maximization is possible via standard numerical optimization. Further details can be consulted in Pedeli and Karlis (2013). We have used `R` code to fit the model to the data. Obviously, for reduced models, e.g. the diagonal case, the calculation is much easier as fewer summations are needed.

2.3. Predictions

Bearing in mind that INAR models are constructed for use as predictive distributions, the following results are used in the numerical application for premium calculations. For further details, see Pedeli and Karlis (2011, 2013). The conditional expectations (given the past value) are

$$E(N_{1i,t}|N_{1i,t-1}=x)=\alpha_{11}x+\alpha_{12}y\lambda_{1it}+\phi_{it}$$

$$E(N_{2i,t}|N_{2i,t-1}=y)=\alpha_{22}y+\alpha_{21}x\lambda_{2it}+\phi_{it}$$

while for the variances we obtain that

$$Var(N_{1i,t}|N_{1i,t-1}=x)=\alpha_{11}(1-\alpha_{11})x+\alpha_{12}(1-\alpha_{12})y+\lambda_{1it}+\phi_{it}$$

$$Var(N_{2i,t}|N_{2i,t-1}=y)=\alpha_{22}(1-\alpha_{22})y+\alpha_{21}(1-\alpha_{21})x+\lambda_{2it}+\phi_{it}.$$

Finally, for $N_t=N_{1t}+N_{2t}$ we obtain that

$$\begin{aligned}
E(N_{it}|N_{1i,t-1}=x,N_{2i,t-1}=y) &= (\alpha_{11}+\alpha_{21})x+(\alpha_{22}+\alpha_{12})y\\
&+ \lambda_{1it}+\lambda_{2it}+2\phi_{it}\\
Var(N_{it}|N_{1i,t-1}=x,N_{2i,t-1}=y) &= ((1-\alpha_{11})\alpha_{11}+(1-\alpha_{21})\alpha_{21})x\\
&+ ((1-\alpha_{22})\alpha_{22}+(1-\alpha_{12})\alpha_{12})y\\
&+ \lambda_{1it}+\lambda_{2it}+4\phi_{it}.
\end{aligned}$$

3. Ratemaking application

3.1. Data

The data used in this section are drawn from an automobile portfolio belonging to a major insurance company operating in Spain. The data have been used previously in Bermúdez (2009) and Bermúdez and Karlis (2011, 2012, 2017). Only cars categorized as being for private use were considered. The data contain information for 14,386 policyholders with full coverage, policies that include third-party liability (claimed and counted as N_1 type), a set of basic guarantees that include emergency roadside assistance or legal and medical assistance (claimed and counted as N_2 type) and, finally, comprehensive coverage (damage to the policyholder's vehicle caused by any unknown party, including damage resulting from theft, flood or fire) and collision coverage (damage resulting from a collision with another vehicle or

object when the policyholder is at fault), also claimed and counted as N_2 type.

We use seven years of data for each policyholder. This means that for each individual we have seven observations made at successive time points for the two types of claim considered here (i.e. third-party liability and all other guarantees). For each individual, we also dispose of a set of covariates, some of which vary across time. For illustrative purposes, we have only opted to employ several of the more usual covariates for pricing an automobile insurance contract. In Table 1 these exogenous variables are described. Figure 1 presents the observed proportions for each variable across the 7 years of study. We see that for some of the covariates, as expected, the proportion changes during the years. Also Figure 2 shows the joint frequencies for the two types of claims across years. We can see that this remains relatively stable across years with minor changes.

GEN	Equals 1 for women and 0 for men
ZON	Equals 1 when zone is deemed high risk (northern Spain)
LOY	Equals 1 if the client has been with the company for more than five years
AGE	Equals 1 if the insured is 30 years old or younger
POW	Equals 1 if the vehicle's horsepower is equal to or greater than 5500 cc

Table 1: Explanatory variables used in the application

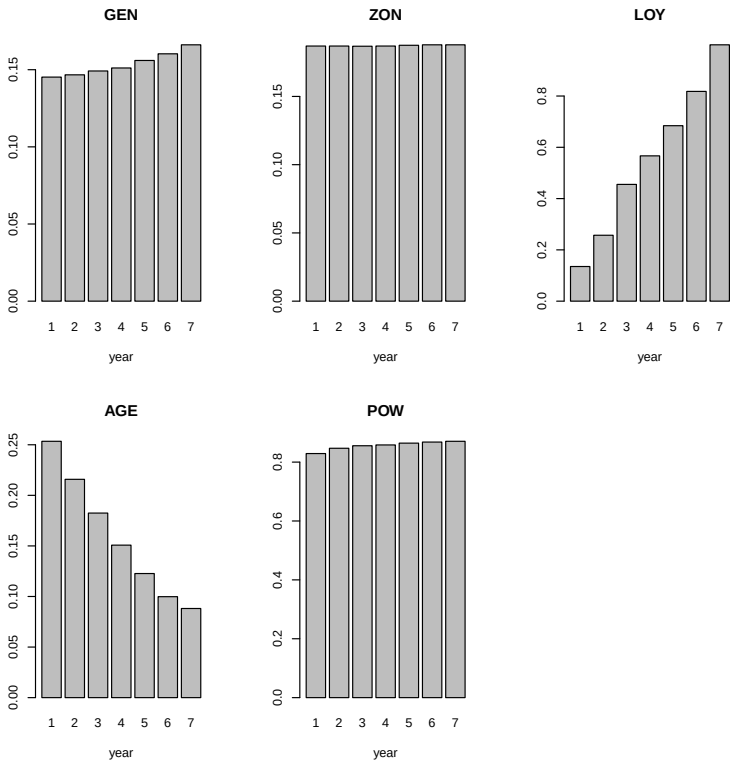

Figure 1: Observed proportions for the variables in Table 1 across years of study

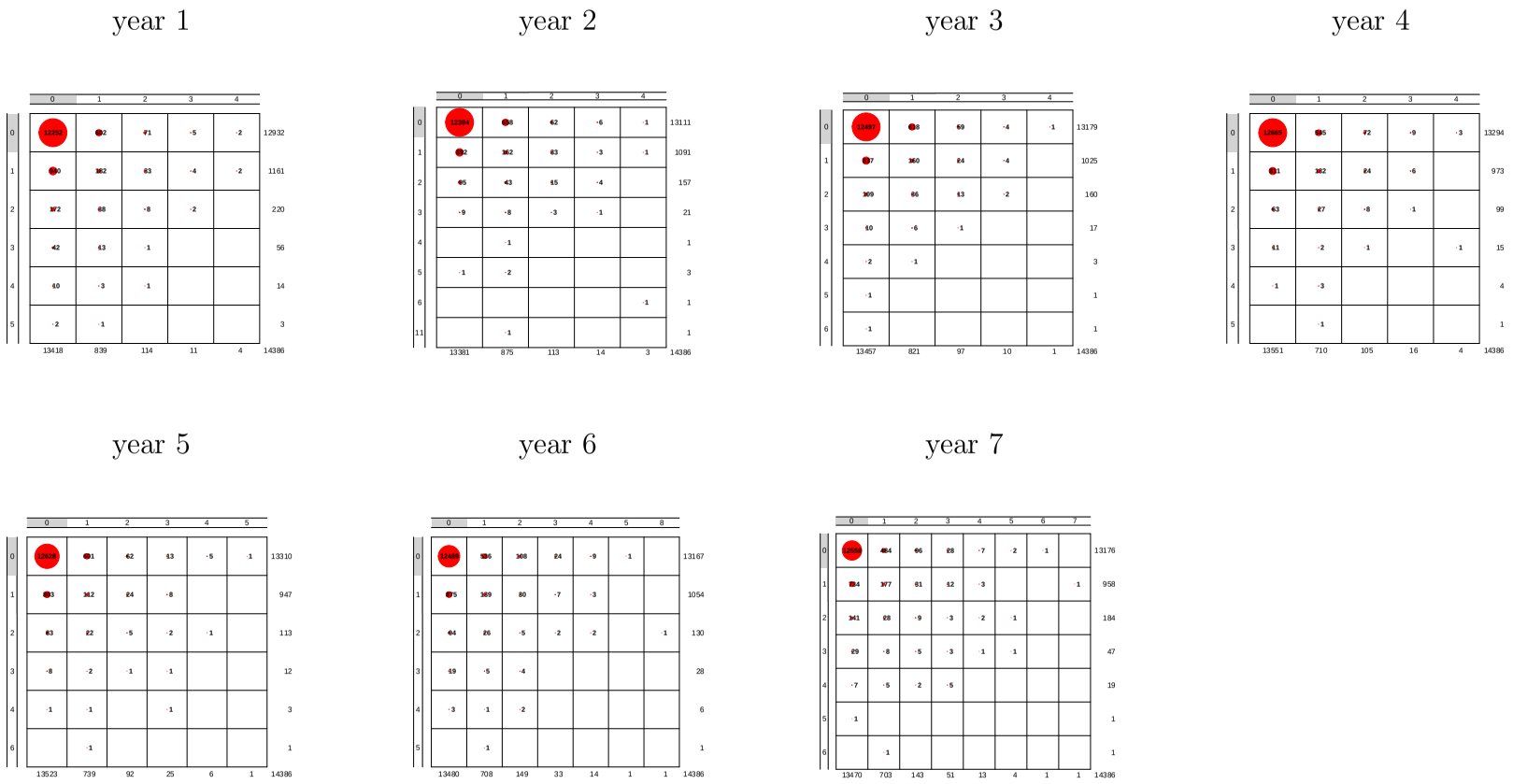

Figure 2: Observed frequencies for the two type of claims across years of study

For comparative purposes, three different, yet representative, profiles were selected from the portfolio (see Table 2). The first was chosen from among the profiles considered to be good drivers, with a lower mean value than that of the average for the portfolio. A profile with a mean lying very close to this average was chosen for the second profile. Finally, a profile considered to represent a bad driver (with a mean above the average) was selected.

	GEN	ZON	LOY	AGE	POW
Good	0	1	1	0	0
Medium	0	0	1	0	1
Bad	0	0	0	1	1

Table 2: Three different profiles for comparison

3.2. Results

Leaving the covariates to one side in order to demonstrate the convenience of using models with full dependence assumptions, we first fitted the BINAR(1) models considered above together with three models that present more restrictions as regards their dependence assumptions. These included a model that does not assume any dependence assumptions; a model that assumes no time series correlation, and hence considers the data as bivariate Poisson observations, as in Bermúdez (2009); and a model where no cross dependence is considered, and hence it fits two independent INAR(1) models, as in Boucher *et al.* (2008). The estimated parameters and the log-likelihood of the fitted models are presented in Table 3.

Model	$\hat{\alpha}_{11}$	$\hat{\alpha}_{22}$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\phi}$	Log-likelihood
No time nor cross dependence			0.0766	0.0969		-53465.48
No cross dependence	0.0391	0.0667	0.0736	0.0902		-53171.69
No time dependence			0.0625	0.0828	0.0141	-52420.80
Basic BINAR(1)	0.0349	0.0627	0.0601	0.0768	0.0138	-52149.43
Full BINAR(1)	$\hat{A}=\begin{bmatrix} 0.033 & 0.058 \\ 0.075 & 0.063 \end{bmatrix}$		0.0583	0.0756	0.0137	-52100.40

Table 3: Fitting different models, without covariates, to the data

Table 3 shows the marked improvement achieved when using the BINAR models compared to the results obtained with the simplest models. This

means that we need to consider both time and cross correlations to fit the available data. Once the effectiveness of the BINAR(1) models had been assessed, covariates to model λ_1 and λ_2 were included. Table 4 shows the log-likelihoods, together with AIC and BIC, of the regression models fitted.

Model	Restrictions	Log-likelihood	AIC	BIC
No time nor cross dependence	$\{\alpha_{jk}\}_{j,k=1,2} = \phi = 0$	-53244.29	106512.58	106626.81
No cross dependence	$\phi = 0$	-52982.14	105990.28	106114.78
No time dependence	$\{\alpha_{jk}\}_{j,k=1,2} = 0$	-52228.24	104482.48	104606.91
Basic BINAR(1)	$\alpha_{12} = \alpha_{21} = 0$	-51999.60	104029.20	104172.57
Full BINAR(1)	No restriction	-51968.70	103971.40	104133.81

Table 4: Fitting comparison for different regression models

The same conclusions as those obtained in the case with no covariates can be derived from the respective regression models. Again, the improvement achieved with the BINAR(1) regression models, compared to the simplest models, is apparent. If we compare the simplest regression models, it seems that cross correlation is more significant than autocorrelation for these data, since the improvement in log-likelihood is larger for the model with only cross correlation. Finally, if we compare the BINAR(1) regression models, we see that a more complicated structure, allowing for cross autocorrelation, is needed for these data.

Tables 5 and 6 show the results of fitting the Basic BINAR(1) and the Full BINAR(1) regression models, respectively. In the case of dependence parameters, α's and ϕ are significant in all cases, implying that time dependence and cross dependence must be considered to fit the data at hand. Moreover, Table 6 validates the presence of cross autocorrelation because of the significance of α_{12} and α_{21}. The largest dependence effect is that provided by α_{21}, which measures the influence of past third-party liability claims on the number of claims against all the other guarantees.

If we focus on the covariates, most are significant and present similar effects in both models. However, note that GEN and LOY are only significant with respect to the claims for all other guarantees. In this case, women are more likely to report claims of this type, while policyholders with more than 5 years in the company are less likely to do so. Some covariates, i.e. ZON and POW, present opposite effects on the number of claims depending on the type of coverage. Northern Spain is really a higher risk zone with respect to third-party liability coverage; however, driving in that zone would decrease

the expected number of claims for all other guarantees. A similar pattern is found when the horsepower of the car is equal to or greater than 5500cc, reducing the expected number of claims for the third-party liability guarantee and increasing it for the rest of guarantees. Finally, AGE is significant in both models and being a younger driver caused the expected number of claims to increase for both types of claim.

	N_1			N_2		
	Estimate	s. err	z-value	Estimate	s. err	z-value
(Intercept)	-2.8122	0.0447	-62.913*	-2.4884	0.0408	-60.990*
GEN	0.0476	0.0390	1.221	0.0860	0.0339	2.537*
ZON	0.2241	0.0346	6.477*	-0.3052	0.0361	-8.454*
LOY	0.0135	0.0305	0.443	-0.2586	0.0261	-9.908*
AGE	0,3001	0.0388	7.735*	0.2837	0.0330	8.597*
POW	-0.1115	0.0397	-2.809*	0.1124	0.0377	2.981*
α_{11}	0.0322	0.0012	26.817*			
α_{22}	0.0490	0.0014	36.275*			
ϕ	0.0129	0.0008	16.177*			

Table 5: Results from fitting the Basic BINAR(1) regression model

	N_1			N_2		
	Estimate	s. err	z-value	Estimate	s. err	z-value
(Intercept)	-2.8435	0.0460	-61.617*	-2.7352	0.0446	-61.322*
GEN	0.0611	0.0403	1.517	0.1034	0.0351	2.949*
ZON	0.1827	0.0363	5.033*	-0.2715	0.0369	-7.355*
LOY	-0.0151	0.0316	-0.477	-0.1403	0.0273	-5.144*
AGE	0.2987	0.0401	7.448*	0.2934	0.0345	8.494*
POW	-0.1058	0.0411	-2.573*	0.2404	0.0410	5.856*
α_{11}	0.0397	0.0016	25.341*			
α_{22}	0.0389	0.0019	20.129*			
α_{12}	0.0178	0.0010	18.653*			
α_{21}	0.0582	0.0017	34.121*			
ϕ	0.0133	0.0004	34.676*			

Table 6: Results from fitting the Full BINAR(1) regression model

3.3. Ratemaking

An analysis of the impact of using these models for ratemaking was also conducted, as the differences between the models proposed in Section 2 were

analysed through the mean (pure premium) and the variance (necessary for loaded premium) of the number of total claims $(N_1 + N_2)$ per year for certain profiles of the insured parties. First, Tables 7 and 8 compare the models' means and variances, respectively, of the three profiles for the different claims reported in the last year. Secondly, Tables 9 and 10 compare the BINAR(1) regression models by focusing on the Medium profile and expanding the number of claims reported in the last year.

Tables 7 and 8 show that including different dependence assumptions in the models may lead to very different premiums. The first two columns (models without autocorrelation) present constant means and variances since they are independent of the number of claims in the last year. The model in the second column, which includes cross correlation, differs from that in the first column in that the variances are larger (due to cross correlation). The model in the third column (with only autocorrelation) presents means that depend on the claims reported in the last year and that are larger than those in the two previous models when a claim was reported. In the last two columns, i.e. BINAR (1) models that simultaneously include autocorrelation and cross correlation, the means also depend on the claims reported in the last year. While the means and variances of the Basic BINAR(1) model are close to those for the model with only autocorrelation, the Full BINAR(1) model presents larger means and variances when a claim was reported in the last year, especially a third-party liability claim. Otherwise, lower means and variances are obtained when no claims were reported. Therefore, allowing for cross autocorrelation leads to a significant change in premiums. In particular, their range of variation is larger. It is worth recalling that this model allows for overdispersion.

Following the above discussion concerning the inclusion of cross autocorrelation, and hence allowing for overdispersion, Tables 9 and 10 show that the range of premiums for the Medium profile is much wider in the case of the Full BINAR(1) model than it is in that of the Basic BINAR(1) model. The former moves from a premium of 0.152 when no claim of any type was reported to a premium of 0.616 when three claims were reported for each type of claim. In the case of the Basic BINAR(1) model, this range moves from 0.161 to 0.405. A closer inspection shows that the presence of third-party liability claims in the last year has a crucial role in this pattern. When no third-party liability claims were reported, premiums for the Full BINAR(1) model were lower than they were for the Basic BINAR(1) model. This reduction in premiums is offset by the larger premiums obtained when

Profile	Last year (x, y)	$\alpha_{11}=\alpha_{22}$ $=\phi=0$	$\alpha_{11}=\alpha_{22}$ $=0$	$\phi = 0$	Basic BINAR(1)	Full BINAR(1)
Good	(0,0)	0.1446	0.1255	0.1386	0.1493	0.1384
	(0,1)	0.1446	0.1255	0.2010	0.1983	0.1951
	(1,0)	0.1446	0.1255	0.1767	0.1815	0.2363
	(1,1)	0.1446	0.1255	0.2391	0.2305	0.2930
Medium	(0,0)	0.1776	0.1477	0.1673	0.1616	0.1524
	(0,1)	0.1776	0.1477	0.2297	0.2106	0.2091
	(1,0)	0.1776	0.1477	0.2054	0.1938	0.2502
	(1,1)	0.1776	0.1477	0.2678	0.2427	0.3069
Bad	(0,0)	0.2420	0.2118	0.2275	0.2218	0.2078
	(0,1)	0.2420	0.2118	0.2899	0.2708	0.2645
	(1,0)	0.2420	0.2118	0.2656	0.2540	0.3057
	(1,1)	0.2420	0.2118	0.3280	0.3030	0.3624

Table 7: Premium Calculations from different models: Means

a third-party liability claim was reported.

3.4. Predictive ability

In order to assess the predictive ability of the Full BINAR(1) model we ran the following experiment. We randomly selected 11,500 clients (that is, 80% of our data set) and used them as a training set, while the remaining 2,886 (20%) clients were used for out of sample prediction. For this set, we predicted the value at $t = 7$, based on previous experience and available covariate information.

Table 11 presents the prediction sum of squared error (PSSE) for the different models together with the observed frequency of some basic cells, namely (0,0), (0,1), (1,0) and (1,1). As can be seen from the PSSE, all the models behave almost the same, although the Full BINAR(1) model behaves slightly better. This is no surprise since all the models can capture the mean effect. The interesting contribution of BINAR(1) models lies in the effect of time dependence, cross dependence and overdispersion. In this sense, models with cross correlation predict the cells much better. The Full BINAR(1) model also allows for overdispersion and, hence, is better for prediction purposes. Note that this model presents the best predictions among the fitted models, but still not good enough for pairs (0,1) and (1,0).

To examine in greater detail the predictions from the Full BINAR(1)

Profile	Last year (x, y)	$\alpha_{11}=\alpha_{22}$ $=\phi=0$	$\alpha_{11}=\alpha_{22}$ $=0$	$\phi = 0$	Basic BINAR(1)	Full BINAR(1)
Good	(0,0)	0.1446	0.2214	0.1386	0.1752	0.1650
	(0,1)	0.1446	0.2214	0.1971	0.2218	0.2199
	(1,0)	0.1446	0.2214	0.1753	0.2063	0.2579
	(1,1)	0.1446	0.2214	0.2338	0.2529	0.3128
Medium	(0,0)	0.1776	0.2436	0.1673	0.1875	0.1789
	(0,1)	0.1776	0.2436	0.2258	0.2340	0.2338
	(1,0)	0.1776	0.2436	0.2039	0.2186	0.2719
	(1,1)	0.1776	0.2436	0.2624	0.2652	0.3267
Bad	(0,0)	0.2420	0.3077	0.2275	0.2477	0.2344
	(0,1)	0.2420	0.3077	0.2860	0.2943	0.2893
	(1,0)	0.2420	0.3077	0.2641	0.2788	0.3274
	(1,1)	0.2420	0.3077	0.3226	0.3254	0.3822

Table 8: Premium Calculations from different models: Variances

model for each observation in the test set, we calculated the joint pmf conditional on the past. By summing all the observations in the test set, we created the expected frequencies for all pairs. Figure 3 presents barplots for the two different types of claim, i.e. we report only the marginal frequencies. A comparison with the observed frequencies shows that the model makes sufficiently good predictions of the expected number of claims for the last period conditional on the information from the previous time points. Thus, the use of the model for premium calculations is well supported.

4. Conclusions

In this paper, we have discussed the effect of simultaneously including different dependence assumptions in ratemaking models. Specifically, we have focused our attention on three sources of dependence: *cross dependence*, *time dependence* and *cross-time dependence*. BINAR(1) regression models are presented as an instrument that can account for the underlying correlation between two types of claim arising from the same policy, the serial correlation between the observations of the same policyholder and the correlation resulting from a combination of these two previous correlations.

Using an automobile insurance database recording the claim frequency history of policyholders (seven years) with two types of coverage (third-party

	N_1							
	0		1		2		3	
N_2	Mean	Var	Mean	Var	Mean	Var	Mean	Var
0	0.161	0.187	0.193	0.218	0.225	0.249	0.258	0.280
1	0.210	0.234	0.242	0.265	0.274	0.296	0.307	0.327
2	0.259	0.280	0.291	0.311	0.323	0.342	0.356	0.374
3	0.308	0.327	0.340	0.358	0.372	0.389	0.405	0.420

Table 9: Premiums based on the Basic BINAR(1) model and past history for Medium policyholder profile

	N_1							
	0		1		2		3	
N_2	Mean	Var	Mean	Var	Mean	Var	Mean	Var
0	0.152	0.179	0.250	0.272	0.348	0.364	0.446	0.458
1	0.209	0.233	0.306	0.326	0.404	0.419	0.502	0.512
2	0.265	0.288	0.363	0.381	0.461	0.474	0.559	0.567
3	0.322	0.343	0.420	0.436	0.518	0.529	0.616	0.622

Table 10: Premiums based on the Full BINAR(1) model and past history for Medium policyholder profile

liability guarantee and all other guarantees), we fitted the BINAR(1) regression models presented above together with a number of other models that included various restrictions with regards to the dependence assumptions. The implications for the ratemaking problem of pricing an automobile insurance contract, including the fitting of the models and a predictive analysis, have been considered.

The best fit, in terms of AIC and BIC, was obtained for the Full BINAR(1) regression model, implying that the three sources of dependence must be taken into account simultaneously. A comparison of the simplest regression models shows that cross correlation appears to be more significant than autocorrelation for these data, since the improvement in log-likelihood is greater for the model with only cross correlation.

Different dependence assumptions in the models can lead to very different premiums. In fact, the dependence assumptions may reveal different aspects of the data. We know, for example, that the assumption of time dependence

Model	PSSE for N_1	PSSE for N_2	PSSE for N_1+N_2	Frequency of (0,0)	Frequency of (0,1)	Frequency of (1,0)	Frequency of (1,1)
No time nor cross dependence	415.7	466.2	881.9	2451	217	182	16
No cross dependence	416.8	464.9	881.6	2450	217	183	17
No time dependence	416.1	466.3	882.3	2480	187	152	46
Basic BINAR(1)	415.5	464.5	880.1	2482	188	148	44
Full BINAR(1)	415.2	464.4	879.6	2488	185	137	42
Observed				2518	144	108	36

Table 11: Out of sample prediction summary

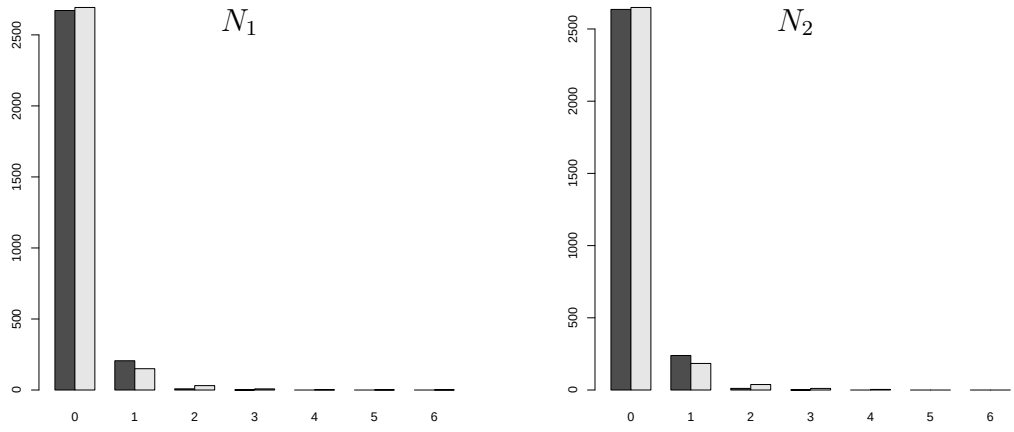

Figure 3: Observed (dark grey) and predicted (light grey) frequencies for the 2,886 clients in the test set. Full BINAR(1) model

enables us to account for the unobserved heterogeneity captured by the past claims experience of policyholders. Likewise, the assumption of cross dependence enables us to account for the positive correlation derived from the fact that the same accident can lead to a claim of each type and, hence, we take this extra variability into consideration. Finally, when cross-time dependence is assumed, as in the Full BINAR(1) with a bivariate Poisson innovation distribution, overdispersion is taken into account, since the marginals are no longer a Poisson distribution.

To test the model's predictive ability, an out of sample study was conducted, from which we conclude that the Full BINAR(1) regression model presents the best prediction sum of squared error and the best estimations for the most common cells, which supports its use for premium calculations.

Finally, the study reported here might be extended in two directions. First, although overdispersion is sufficiently captured here by the covariates and by the model structure of the Full BINAR(1) model, the assumption of bivariate Poisson innovation might be replaced with an overdispersed bivariate pdf at the added cost of a more complicated, yet less well-known model. Second, the time dependence assumption is overly limited by the number of claims reported in the last insurance period. This means, for example, that no time dependence is implied for an insured with no claims in his last contract. Models based more firmly on time structure could be used like BINAR(p) models.

Acknowledgments. Research for this paper was initiated while the third author was visiting the Riskcenter Research Group at the University of Barcelona. The authors wish to acknowledge the support of the Spanish Ministry for grant ECO2015-66314-R and ECO2016-76203-C2-2-P.

References

Al-Osh, M. and Alzaid, A. (1987). First-Order Integer-Valued Autoregressive Process. *Journal of Time Series Analysis*, **8(3)**, 261-275.

Bermúdez, L. (2009). A priori ratemaking using bivariate Poisson regression models. *Insurance: Mathematics and Economics*, **44(1)**, 135-141.

Bermúdez, L. and Karlis, D. (2011). Bayesian multivariate Poisson models for insurance ratemaking. *Insurance: Mathematics and Economics*, **48(2)**, 226-236.

Bermúdez, L. and Karlis, D. (2012). A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking. *Computational Statistics & Data Analysis*, **56**, 3988-3999.

Bermúdez, L. and Karlis, D. (2017). A posteriori ratemaking using bivariate Poisson models. *Scandinavian Actuarial Journal*, **2**, 148-158.

Boucher, J.-P., Denuit, M. and Guillén, M. (2008). Models of Insurance Claim Counts with Time Dependence Based on Generalisation of Poisson and Negative Binomial Distributions, *Variance*, **2(1)**, 135-162.

Boucher, J.-P. and Inoussa, R. (2014). A posteriori ratemaking with panel data. *ASTIN Bulletin*, **44**, 587-612.

Boudreault, M. and Charpentier, A. (2011). Multivariate integer-valued autoregressive models applied to earthquake counts. http://arxiv.org/abs/1112.0929.

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

Gourieroux, C. and Jasiak J. (2004). Heterogeneous INAR(1) Model with Application to Car Insurance. *Insurance: Mathematics and Economics*, **34**, 177-192.

Kocherlakota, S. and Kocherlakota, K. (1992). Bivariate Discrete Distributions. In: Statistics: Textbooks and Monographs, vol. 132. New York: Markel Dekker.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Pedeli, X. and Karlis, D. (2011). A Bivariate INAR(1) Process with Application. *Statistical Modelling*, **11(4)**, 327-351.

Pedeli, X. and Karlis, D. (2013). Some properties of multivariate INAR(1) processes. *Computational Statistics and Data Analysis*, **67**, 213-225.

Shi, P. (2012). Multivariate longitudinal modelling of insurance company expenses. *Insurance: Mathematics and Economics*, **51(1)**, 204-215.

Shi, P. and Valdez, E.A. (2014). Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics*, **55(1)**, 18-29.

Steutel, F. and van Harn, K. (1979). Discrete Analogues of Self-Decomposability and Stability. *The Annals of Probability*, **7**, 893-899.