Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 8(12): 28710–28727.
DOI: 10.3934/math.20231469
Received: 29 July 2023
Revised: 12 October 2023
Accepted: 17 October 2023
Published: 23 October 2023

Research article

Bayesian premium of a credibility model based on a heterogeneous SETINAR(2,1) process

Shuo Zhang[1] **and Jianhua Cheng**[2,*]

[1] School of Mathematics and Information Science, Shandong Technology and Business University, Yantai 264005, China

[2] School of Mathematics, Jilin University, Changchun 130012, China

* **Correspondence:** Email: chengjh@jlu.edu.cn.

Abstract: In this paper, we propose a new credibility model based on heterogeneous integer-valued self-exciting threshold autoregressive time series, in which the SETINAR(2,1) process is used to fit the claim numbers of policyholders for consecutive periods, and the unobservable heterogeneity is assumed to follow Gamma distribution. We obtain the Bayesian pricing formula for the proposed model and present some numerical examples to illustrate how the claim history affects the future premiums. We also apply the proposed model to a real panel dataset from the Wisconsin Local Government Property Insurance Fund. By comparing with some existing models, we find that our model can exploit the past information more efficiently and has better predictive performance.

Keywords: automobile insurance; credibility model; Bayesian premium; SETINAR(2,1) process
Mathematics Subject Classification: 62P05, 91B30, 97M30

1. Introduction

Credibility theory is one of the most important and applicable methods of property and casualty ratemaking, particularly in the context of automobile insurance. It is called a "cornerstone" in the field of actuarial science, and has now been widely used in many insurance companies. Numerous authors devote themselves to developing tools to evaluate the existing credibility models, and then designing improved Bonus-Malus Systems (BMS) or No-Claim Discount (NCD). Contributions to the related literature include, among others, references [1–4].

As we all know, an insurance company insures many kinds of risks, and it is one of the main tasks of actuaries to analyze individual risks quantitatively, predict future claims and calculate equitable premium for each policyholder. As a rule, ratemaking process in non-life actuarial science always consists of two separate steps. In the first step, the policyholders are grouped into several classes on the basis of their observable characteristics, and those people in the same class are collected the same premium because they are considered to have similar expected losses. Examples of such risk characteristics are the age, gender and occupation of the policyholders, the type and use of their vehicles, the place where they reside, etc.

However, many important factors, such as respectfulness towards the law, reaction speed, aggressiveness behind the wheel and knowledge of the highway code, cannot be taken into account at this stage because they are usually impossible to measure and incorporate in a price list. Therefore, some heterogeneity still remains within the risk classes. Credibility models are used to distribute premiums fairly among a heterogeneous group of policyholders during the second step. The unobserved heterogeneity is taken into account by the introduction of a latent random variable Θ, with Gamma, LogNormal or Inverse Gaussian distribution. Making use of Bayesian methodology, the base premium determined in the first step will be adjusted by discounts or surcharges depending on past claims of the policyholder. As result of this experience rating process, BMS or NCD is established to motivate people to drive more safely by rewarding good drivers and penalizing bad drivers.

Traditionally, actuaries assume that the random effect Θ in the credibility model does not vary with time. One of major problems in this so-called time-independent (or static) heterogeneous model is that the age of claims has not been taken into account. In other words, the classical models conflict with the intuition that the predictability based on a period of the policyholder's history should decrease with time. To overcome this defect, more and more actuaries have been paying attention to the evolutionary (or dynamic) credibility models, in which the unknown characteristics are represented by time-varying random effects $\{\Theta_t, t \geq 1\}$, see for instance [5–10] and the references therein. However, one drawback of these models is that the closed-form expressions for the predictive distribution and predictive premium are difficult to derive.

On the other hand, the dynamics on insurance data can also be obtained from endogenous approaches. As a typical representation among these methods, Gourieroux and Jasiak [11] has applied a heterogeneous integer-valued time series to describe the serial dependence in claim counts process of the policyholders. The authors show that their model is more flexible in computing predictive premiums by deriving the closed-form expressions of the Bayesian premium. They also find that the premium depends on not only the number of claims but also the timing of claim arrivals.

In the model of Gourieroux and Jasiak [11], the famous integer-valued autoregressive (INAR) process plays a central role. The classical INAR(1) process proposed by Al-Osh and Alzaid [12] can be written as

$$N_t = \phi \circ N_{t-1} + \epsilon_t, \ t = 2, 3, \ldots, \tag{1.1}$$

where $\phi \in [0, 1)$, and the so-called thinning operator "$\circ$" is defined by

$$\phi \circ N_{t-1} = \sum_{k=1}^{N_{t-1}} B_{t,k},$$

in which $\{B_{t,k}, t = 2, 3, \ldots, k = 1, 2, \ldots\}$ is an array of independent and identically distributed (i.i.d.) Bernoulli random variables with mean ϕ, and independent of the sequence $\{\epsilon_2, \epsilon_3, \ldots\}$.

Model (1.1) implies that the number of claims in period t consists of two parts. The first part is a sum of Bernoulli variables indexed by the claims occurred in the preceding period. This term captures claim trends of the policyholders by providing a causality relationship and temporal dependence among the claim numbers of different periods. It can be seen that each of the claims occurred in the previous period could contribute one claim to the claim count of the next period, with probability ϕ. This could be explained by the incentives to careful driving in automobile insurance, i.e., car accidents may modify a driver's perception of danger behind the wheel, such that a policyholder who reports claims in previous period would like to lower his risk level during next period, in consideration of safety and BMS. Furthermore, the other part ϵ_t is a number of claims without relationship with those occurred in the past, representing an innovation or fluctuations.

To generalize the application of model (1.1), Bermúdez et al. [13] adopts the bivariate INAR(1) process to price an automobile insurance contract with two types of coverage, in which the dependence structures arising from different sources of the claim numbers are taken into account. Afterwards, Bermúdez and Karlis [14] extends the results in the aforementioned paper to a multivariate setting using the multivariate INAR(1) process, and Zhang et al. [15] provides a flexible framework to fit the claim count data sets that contain a large number of zeros by modelling the innovations of multivariate INAR(1) process with zero-inflated distributions. Recently, Hu and Yao [16] proposes a modification of a combined INAR(p) process to address higher-order dependence among the number of claims in an insurance portfolio.

In practice, the real-world data sets sometimes exhibit nonlinear structure, such as the so-called piecewise phenomenon, which means that the autoregressive parameter ϕ in model (1.1) may be affected by some factors, saying for example, the value of N_{t-1}. To capture this attribute, Monteiro et al. [17] proposes the integer-valued self-exciting threshold autoregressive process of order one with two regimes (referred to as SETINAR(2,1) for short). In this paper, it is our goal to extend the classical SETINAR(2,1) process to accommodate the unobserved heterogeneity and then apply the modified SETINAR(2,1) process to credibility model for ratemaking problem. We will show that our approach provides a superior pattern to the predictive premiums calculation when compared to the outcomes of several alternative models.

The layout of this paper is organized as follows. In Sections 2, we introduce the proposed model and derive the Bayesian premium formula. In Section 3, some numerical examples are conducted to illustrate our methods and results. In section 4, we present an empirical analysis of insurance claims count data. Section 5 concludes this paper.

2. Modelling and results

For automobile insurance, retemaking should be calculated based on both claim frequency and severity. However, it is believed that the number of claims can better reveal the inherent risk of a driver. In this paper, we follow most of the actuarial literature and focus only on the frequency part.

Let us consider one policyholder from an insurance portfolio, suppose T to be the number of periods elapsed since this policy has been issued, and $N_t, t = 1, 2, \ldots, T$, denotes the number of claims reported by this policyholder during period t of insurance. Our task is to draw prediction on the premium for the subsequent period $T + 1$ of this policyholder. To this end, we assume that:

A1. The count variables $N_1, \ldots, N_T, N_{T+1}$ follow the SETINAR(2,1) process:

$$N_t = I_{1,t} \cdot (\phi_1 \circ N_{t-1}) + I_{2,t} \cdot (\phi_2 \circ N_{t-1}) + \epsilon_t, \quad t = 2, 3, \ldots, \tag{2.1}$$

where

- $I_{1,t} = I\{N_{t-1} \leq r\}$, $I_{2,t} = 1 - I_{1,t} = I\{N_{t-1} > r\}$, in which r is the so-called threshold variable;
- the thinning operator "$\circ$" is defined as

$$\phi_j \circ N_{t-1} = \sum_{k=1}^{N_{t-1}} B_{t,k}^{(j)}, \quad j = 1, 2,$$

in which

$$\{B_{t,k}^{(j)}, t = 2, 3, \ldots, k = 1, 2, \ldots\}, \quad j = 1, 2,$$

are arrays of i.i.d. Bernoulli random variables with mean $\phi_j \in (0, 1)$, respectively. Furthermore, $\{B_{t,k}^{(1)}, k = 1, 2, \ldots\}$ and $\{B_{t,k}^{(2)}, k = 1, 2, \ldots\}$ are supposed to be independent of ϵ_t and $\{N_{t-1}, \ldots, N_1\}$ for any fixed t.

A2. Given the unobservable heterogeneity $\Theta = \theta$, N_1 follows the Poisson distribution with mean $\lambda\theta$, i.e.,

$$P(N_1 = n|\Theta = \theta) = \frac{(\lambda\theta)^n}{n!} \exp\{-\lambda\theta\}, \ n = 0, 1, \ldots, \tag{2.2}$$

where $\lambda = e^{\boldsymbol{\beta}\boldsymbol{x}_1'}$, in which $\boldsymbol{x}_1 = (1, x_{1,1}, x_{1,2}, \ldots, x_{1,q})$ is the observable risk characteristics of the policyholder in the first period, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)$ collects the regression coefficients and $\boldsymbol{x}_1'$ denotes the transpose of $\boldsymbol{x}_1$.

A3. Given the unobservable heterogeneity $\Theta = \theta$, $\{\epsilon_t, t = 2, 3, \ldots\}$ is a sequence of independent non-negative integer-valued random variables confirming to the Poisson distribution with mean $\eta_t\theta$, where $\eta_t = e^{\boldsymbol{\omega}\boldsymbol{x}_t'}$ with $\boldsymbol{x}_t = (1, x_{t,1}, x_{t,2}, \ldots, x_{t,q})$ representing the characteristics of the policyholder in the period t and $\boldsymbol{\omega} = (\omega_0, \omega_1, \ldots, \omega_q)$ denoting the regression coefficients. Furthermore, ϵ_t and $N_1, \ldots, N_{t-1}$ are independent conditional on Θ and $\boldsymbol{x}_t, \ldots, \boldsymbol{x}_1$ for any fixed t.

A4. The unobservable heterogeneity Θ follows Gamma distribution whose density function is

$$f_\Theta(\theta) = \begin{cases} \dfrac{\alpha^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\alpha\theta}, & \theta > 0, \\ 0, & \theta \leq 0, \end{cases} \tag{2.3}$$

in which the shape parameter and scale parameter are identical so that the premium for this policyholder in the first period results from a priori rating system and equals to $\lambda = e^{\boldsymbol{\beta}\boldsymbol{x}_1'}$.

Remark 2.1. *The model defined by (2.1) implies that the autoregressive parameter is a function of N_{t-1}, i.e., the effect of N_{t-1} on N_t varies with different values of N_{t-1}. In the insurance practice, this could be explained that policyholders may adjust their efforts to prevent losses according to their experience with past claims. Taking the NCD system of automobile insurance in China for example, if a policyholder has reported more than one claim, his premium will be raised up to 2 times of the base premium. Therefore, this policyholder might become more risk-averse and may have a greater of initiative to reduce the claim numbers in the future, so that it is reasonable for us to believe the constraint condition of $\phi_1 > \phi_2$. By contrast, more claims mean higher risks, i.e., the claims of period $t-1$ could produce another claim in period t with a higher probability, making us could acknowledge that $\phi_1 < \phi_2$. In brief, it is very necessary to consider the SETINAR(2,1) process with two different autoregressive coefficients for actuarial applications from a practical view.*

Remark 2.2. *The distributional assumptions A2–A4 on top of SETINAR(2,1) process are made in accordance with the Poisson-Gamma model that have been widely applied for modeling claim counts when the portfolio is heterogeneous to construct the optimal experience rating system. Under these assumptions, we are able to derive the closed-form expression for Bayesian pricing formula and illustrate more effectively our main results regarding the threshold structure of SETINAR(2,1) process. However, the limitations of these assumptions and the potential impacts they may have on the usefulness and applicability of our proposed model are also very obvious. For better and more flexible application in practice, our proposed model can be generalized to some alternatives to the Poisson-Gamma assumption, the computations become however more complicated, because numerical methods to perform integration are usually required. Perhaps credibility premium, the linear estimator (based on past claims data) that best approximates the Bayesian premium, can be considered to overcome these difficulties in these cases.*

From Denuit et al. [3], we know that given the observations of claim counts for T periods $N_1 = n_1, \ldots, N_T = n_T$, the Bayesian premium of the policyholder for period $T+1$ with the quadratic loss function can be obtained as

$$
\begin{aligned}
P_{T+1} =& E(N_{T+1}|N_1 = n_1, \ldots, N_T = n_T) \\
=& E[I_{1,T+1} \cdot (\phi_1 \circ N_T) + I_{2,T+1} \cdot (\phi_2 \circ N_T) + \epsilon_{T+1}|N_1 = n_1, \ldots, N_T = n_T] \\
=& E[I_{1,T+1} \cdot (\phi_1 \circ N_T)|N_1 = n_1, \ldots, N_T = n_T] \\
&+ E[I_{2,T+1} \cdot (\phi_2 \circ N_T)|N_1 = n_1, \ldots, N_T = n_T] \\
&+ E[E(\epsilon_{T+1}|\Theta, N_1 = n_1, \ldots, N_T = n_T)|N_1 = n_1, \ldots, N_T = n_T] \\
=& \phi_1 n_T I\{n_T \le r\} + \phi_2 n_T I\{n_T > r\} + \eta_{T+1} E[\Theta|N_1 = n_1, \ldots, N_T = n_T].
\end{aligned} \tag{2.4}
$$

As shown in (2.4), the pure premium depends on the conditional distribution of the heterogeneity given the claim history. Therefore, we obtain the analytical expression of the conditional distribution of the heterogeneity in the following theorem.

Theorem 2.1. *Under the assumptions A1–A4, we have that:*

(1) For $T = 1$, given $N_1 = n_1$, the heterogeneity Θ follows Gamma distribution with conditional density function

$$
f_\Theta(\theta|N_1 = n_1) = \begin{cases} \dfrac{(\alpha+\lambda)^{\alpha+n_1}}{\Gamma(\alpha+n_1)} \theta^{\alpha+n_1-1} e^{-(\alpha+\lambda)\theta}, & \theta > 0, \\ 0, & \theta \le 0. \end{cases} \tag{2.5}
$$

(2) For any $T \ge 2$, given $N_1 = n_1, \ldots, N_T = n_T$, the conditional density function of Θ is given by

$$
f_\Theta(\theta|N_1 = n_1, \ldots, N_T = n_T) \tag{2.6}
$$

$$
= \frac{\sum_{z_T=0}^{\min\{n_{T-1}, n_T\}} \cdots \sum_{z_2=0}^{\min\{n_1, n_2\}} \pi(z_T, \ldots, z_2, n_T, \ldots, n_1) \gamma(\theta; \tilde{\alpha}_1(z_T, \ldots, z_2, n_T, \ldots, n_1), \tilde{\alpha}_2)}{\sum_{z_T=0}^{\min\{n_{T-1}, n_T\}} \cdots \sum_{z_2=0}^{\min\{n_1, n_2\}} \pi(z_T, \ldots, z_2, n_T, \ldots, n_1)}, \tag{2.7}
$$

in which

$$\tilde{\alpha}_1(z_T,\dots,z_2,n_T,\dots,n_1) = \alpha + n_1 + \cdots + n_T - z_2 - \cdots - z_T, \tag{2.8}$$

$$\tilde{\alpha}_2 = \alpha + \eta_2 + \cdots + \eta_T + \lambda, \tag{2.9}$$

$$\begin{aligned}\pi(z_T,\dots,z_2,n_T,\dots,n_1) =& \frac{1}{\eta_2^{z_2}}\cdots\frac{1}{\eta_T^{z_T}} \times \binom{n_{T-1}}{z_T}\cdots\binom{n_1}{z_2} \times \frac{1}{(n_T-z_T)!}\cdots\frac{1}{(n_2-z_2)!} \\ &\times \left[\phi_1^{z_T}(1-\phi_1)^{n_{T-1}-z_T} I\{n_{T-1}\le r\} + \phi_2^{z_T}(1-\phi_2)^{n_{T-1}-z_T} I\{n_{T-1}> r\}\right] \\ &\times\cdots\times \left[\phi_1^{z_2}(1-\phi_1)^{n_1-z_2} I\{n_1\le r\} + \phi_2^{z_2}(1-\phi_2)^{n_1-z_2} I\{n_1> r\}\right] \\ &\times \frac{\Gamma(\tilde{\alpha}_1(z_T,\dots,z_2,n_T,\dots,n_1))}{\tilde{\alpha}_2^{\tilde{\alpha}_1(z_T,\dots,z_2,n_T,\dots,n_1)}},\end{aligned} \tag{2.10}$$

and $\gamma(\theta;\tilde{\alpha}_1,\tilde{\alpha}_2)$ *is the density function of Gamma distribution with shape parameter* $\tilde{\alpha}_1$ *and scale parameter* $\tilde{\alpha}_2$*.*

Proof. For $T = 1$, because $N_1|\Theta = \theta \sim \mathrm{P}(\lambda\theta)$, we know that the joint distribution of N_1 and Θ is

$$\begin{aligned} f(n_1,\theta) &= P(N_1 = n_1|\Theta=\theta)\times f_\Theta(\theta) \\ &= \frac{(\lambda\theta)^{n_1}}{n_1!}e^{-\lambda\theta}\times\left(\frac{\alpha^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\alpha\theta}\right) \\ &= \frac{\alpha^\alpha}{\Gamma(\alpha)}\times\frac{(\lambda)^{n_1}}{n_1!}\times\theta^{\alpha+n_1-1}e^{-(\alpha+\lambda)\theta}, \end{aligned} \tag{2.11}$$

then, we have

$$\begin{aligned} f_\Theta(\theta|N_1=n_1) &= \frac{f(n_1,\theta)}{\int_0^\infty f(n_1,\theta)\mathrm{d}\theta} = \frac{\theta^{\alpha+n_1-1}e^{-(\alpha+\lambda)\theta}}{\int_0^\infty \theta^{\alpha+n_1-1}e^{-(\alpha+\lambda)\theta}\mathrm{d}\theta} \\ &= \frac{(\alpha+\lambda)^{\alpha+n_1}}{\Gamma(\alpha+n_1)}\theta^{\alpha+n_1-1}e^{-(\alpha+\lambda)\theta}, \end{aligned}$$

which implies that (2.5) holds.

Similarly, for any $T \ge 2$, we have

$$f_\Theta(\theta|N_1=n_1,\dots,N_T=n_T) = \frac{f(n_1,\dots,n_T,\theta)}{\int_0^\infty f(n_1,\dots,n_T,\theta)\mathrm{d}\theta}, \tag{2.12}$$

in which

$$\begin{aligned} f(n_1,\dots,n_T,\theta) =& P(N_T=n_T|N_1=n_1,\dots,N_{T-1}=n_{T-1},\Theta=\theta) \\ &\times P(N_{T-1}=n_{T-1}|N_1=n_1,\dots,N_{T-2}=n_{T-2},\Theta=\theta) \\ &\times\cdots\times P(N_1=n_1|\Theta=\theta)\times f_\Theta(\theta). \end{aligned} \tag{2.13}$$

Noting that

$$P(N_j=n_j|N_1=n_1,\dots,N_{j-1}=n_{j-1},\Theta=\theta),\quad j=2,\dots,T,$$

is the probability mass function of the sum of two discrete random variables, we can rewrite it as the following convolutions:

$$P(N_j = n_j | N_1 = n_1, \ldots, N_{j-1} = n_{j-1}, \Theta = \theta) \tag{2.14}$$
$$= \sum_{z_j=0}^{\min\{n_{j-1}, n_j\}} \binom{n_{j-1}}{z_j} \Big[\phi_1^{z_j} (1-\phi_1)^{n_{j-1}-z_j} I\{n_{j-1} \le r\}$$
$$+ \phi_2^{z_j} (1-\phi_2)^{n_{j-1}-z_j} I\{n_{j-1} > r\} \Big] \times \frac{(\eta_j \theta)^{n_j - z_j}}{(n_j - z_j)!} e^{-\eta_j \theta}. \tag{2.15}$$

By applying (2.11), (2.13) and (2.14) into (2.12), we can obtain (2.6). □

One can see that the conditional distribution of the heterogeneity Θ is a mixture of Gamma distributions with parameters that depend on the claim history. From (2.4) and the above theorem, it is easy to derive the analytical formulas of the pure premium for period $T+1$ as follows.

Theorem 2.2. *If the assumptions A1–A4 hold, then we have:*

(1) When $T = 0$, no claim history is available at the beginning of the contract, and the base premium determined in the first step of ratemaking process is collected at the time of signing the insurance contract, i.e.,

$$P_1 = E(N_1) = \lambda.$$

(2) When $T = 1$, the Bayesian premium of the policyholder for period $T = 2$ is

$$P_2 = \phi_1 n_1 I\{n_1 \le r\} + \phi_2 n_1 I\{n_1 > r\} + \eta_2 \widehat{\Theta}_1,$$

where the predicted heterogeneity $\hat{\Theta}_1$ is given by

$$\widehat{\Theta}_1 = \frac{\alpha + n_1}{\alpha + \lambda}. \tag{2.16}$$

(3) When $T \ge 2$, the Bayesian premium of the policyholder for period $T+1$ is

$$P_{T+1} = \phi_1 n_T I\{n_T \le r\} + \phi_2 n_T I\{n_T > r\} + \eta_{T+1} \hat{\Theta}_T,$$

where the predicted heterogeneity $\hat{\Theta}_T$ is given by

$$\widehat{\Theta}_T = \frac{\displaystyle\sum_{z_T=0}^{\min\{n_{T-1}, n_T\}} \cdots \sum_{z_2=0}^{\min\{n_1, n_2\}} \pi(z_T, \ldots, z_2, n_T, \ldots, n_1) \frac{\tilde{\alpha}_1(z_T, \ldots, z_2, n_T, \ldots, n_1)}{\tilde{\alpha}_2}}{\displaystyle\sum_{z_T=0}^{\min\{n_{T-1}, n_T\}} \cdots \sum_{z_2=0}^{\min\{n_1, n_2\}} \pi(z_T, \ldots, z_2, n_T, \ldots, n_1)}, \tag{2.17}$$

in which $\tilde{\alpha}_1(z_T, \ldots, z_2, n_T, \ldots, n_1)$, $\tilde{\alpha}_2$ and $\pi(z_T, \ldots, z_2, n_T, \ldots, n_1)$ are accordingly defined by the Eqs (2.8)–(2.10).

3. Numerical illustration

In this section, we present some simulation results to illustrate the premium updating for one given policyholder. To this end, we fix all the values of the parameters, compute and compare the Bayesian premiums for different models.

We mainly focus on the following three different models for the number of claims.

Model 1: Credibility model based on SETINAR(2,1) process with Gamma heterogeneity, abbreviated as CM-SETINAR(2,1). In this model. we assume that $\lambda = 0.4286$, $\alpha = 9$ and $\eta_t = \eta = 0.3$ for $t = 2, 3, \ldots$. From Remark 2.1, we know that the policyholders who report more than one claim would be penalized by some premium surcharges. Therefore, we take $r = 1$ to illustrate the results. Furthermore, in order to compare different situations, we consider the following two scenarios for ϕ_1 and ϕ_2:

(a) $\phi_1 = 0.3$, $\phi_2 = 0.2$;

(b) $\phi_1 = 0.3$, $\phi_2 = 0.4$.

Model 2: Credibility model based on INAR(1) process with Gamma heterogeneity, abbreviated as CM-INAR(1). When $\phi_1 = \phi_2 = \phi$, CM-SETINAR(2,1) reduces to the case studied in Gourieroux and Jasiak [11]. For the other parameters, we assume that $\lambda = 0.4286$, $\phi = 0.3$, $\alpha = 9$ and $\eta_t = \eta = 0.3$ for $t = 2, 3, \ldots$.

Model 3: Credibility model based on mixed Poisson distribution with Gamma heterogeneity, abbreviated as CM-MP. When $\phi_1 = \phi_2 = 0$, Model 1 becomes the standard case, i.e., negative binomial distribution. Furthermore, let $\lambda = 0.4286$ and $\alpha = 9$.

Remark 3.1. *In Gourieroux and Jasiak [11], the number of claims is fitted by the following model:*

$$N_t = \phi \circ N_{t-1} + \epsilon_t, \quad t = 2, 3, \ldots,$$

where conditional on the unobserved heterogeneity $\Theta = \theta$*, the variables* $\{\epsilon_t, t = 2, 3, \ldots\}$ *are independent with distribution* $P(\eta\theta)$*, so that* $\{N_t, t = 1, 2, \ldots\}$ *has Poisson marginal that is distributed by* $P(\frac{\eta}{1-\phi}\theta)$*, to make the process itself satisfy the stationarity. Therefore, we also set* $\lambda = \frac{\eta}{1-\phi}$ *in our model for the convenience of comparison. Additionally, the mixed Poisson distribution with Gamma heterogeneity implies*

$$\begin{cases} N_t|\Theta = \theta \sim P(\lambda\theta), t = 1, 2, \ldots, \\ \Theta \sim Ga(\alpha, \alpha), \end{cases}$$

which results in the standard negative binomial model.

Remark 3.2. *Note that we fix* $\eta_t = \eta$ *for* $t = 2, 3, \ldots$ *in CM-SETINAR(2,1) and CM-INAR(1), so that we can use the same setting of the parameters with Gourieroux and Jasiak [11] to calculate the premiums, and show the effect of the threshold on the results. In practice,* η_t *could vary because the risk characteristics (e.g., age, coverage, etc.) of the policyholder may be different for every period. In this case, the claims history* $\{N_t, t = 1, 2, \ldots\}$ *may be not stationary any more. As stated in Asmussen [18], the transient distributions in a Bonus-Malus system are quite far from the stationary ones usually, i.e., the rate of convergence to stationarity may be slow in comparison to the typical sojourn time of a customer in the portfolio. We accept this view in this paper and relax the condition of stationarity. In* *the next section (Empirical Study), we will illustrate that our proposed model is able to fit the real data better than some existing models.*

As an illustration, we consider several different claim histories, and provide a summary of the evolution of the corresponding premiums over the first 4 periods in Table 1.

From Table 1, it is easy to see that the predicted premiums of CM-SETINAR(2,1) and CM-INAR(1) are determined by the explicit claim frequency history, and not just by the total number of claims as that in the case of CM-MP. For example, we compare the predicted premiums based on claim histories $(0, 2)$ and $(2, 0)$, respectively. For CM-MP, P_3 has the same value 0.4783 because the total numbers of claims in the first 2 periods are both 2. However, for CM-SETINAR(2,1) and CM-INAR(1), the two claim histories are distinguishable and the claim arrival (i.e., in the first or in the second year) matters. Therefore, the predicted premiums are different. Furthermore, P_3 for claim history $(0, 2)$ is higher than P_3 for claim history $(2, 0)$ because the former is the premium after a claim-free period and a second period with 2 claims, while the later is the premium after a first period with 2 claims and a free-claim period. This result shows that the most recent claims are more predictive than older ones.

Let us now consider the claim history $(1, 2, 0)$ to present the difference of predicted premiums between CM-SETINAR(2,1) (including Scenario (a) and (b)) and CM-INAR(1).

I. Initially, all the models have the same premium 0.4286, because all of them calculate P_1 according to λ.

II. If there is I claim filed during the first period, P_2 for CM-SETINAR(2,1) (a), CM-SETINAR(2,1) (b) and CM-INAR(1) also have the same value 0.6182, because N_1 does not exceed the threshold value $r = 1$ at this moment, implying that both CM-SETINAR(2,1) (a) and CM-SETINAR(2,1) (b) predict N_2 according to $N_2 = \phi_1 \circ N_1 + \epsilon_2$, which is identical to that of CM-INAR(1).

III. As time passes by, when II claims are reported in the next period, and then a claim-free period comes, the predicted premiums for CM-SETINAR(2,1) (a), CM-SETINAR(2,1) (b) and CM-INAR(1) become quite different.

(1) The values of P_3 for CM-SETINAR(2,1) (a) and CM-INAR(1) are 0.7479 and 0.9479, respectively. Moreover, it can be calculated that the surcharges for these two models accordingly are $(0.7479 - 0.6182)/0.6182 = 20.98\%$ and $(0.9479 - 0.6182)/0.6182 = 53.33\%$, implying that the policyholder for CM-SETINAR(2,1) (a) will get milder penalty than the individual for CM-INAR(1), which is in line with the situation $\phi_1 > \phi_2$ (see Remark 2.1). Correspondingly, after a claim-free period, the policyholder for CM-SETINAR(2,1) (a) will receive lower reward than the driver for CM-INAR(1), since the discounts for these two models are $(0.7479 - 0.3374)/0.7479 = 54.89\%$ and $(0.9479 - 0.3374)/0.9479 = 64.41\%$, respectively.

(2) By contrast, for CM-SETINAR(2,1) (b) with $\phi_1 < \phi_2$, the policyholder will get more severe penalty or receives higher reward than driver that is described in CM-INAR(1).

To sum up, compared to CM-INAR(1), CM-SETINAR(2,1) (a) brings less variation to the premiums, while CM-SETINAR(2,1) (b) causes more variation to the premiums. We can also get these trends visually from Figure 1. The same conclusion can also be made from the results of the other claim histories, revealing that the threshold structure has significant implications on the dynamics of the predictive premiums.

Table 1. Premiums comparison for different models.

Claim history	Premium	CM-SETINAR(2,1) (a)	CM-SETINAR(2,1) (b)	CM-INAR(1)	CM-MP
(0, 1, 2)	P_1	0.4286	0.4286	0.4286	0.4286
	P_2	0.2864	0.2864	0.2864	0.4091
	P_3	0.6084	0.6084	0.6084	0.4348
	P_4	0.7513	1.1513	0.9513	0.5000
(1, 0, 2)	P_1	0.4286	0.4286	0.4286	0.4286
	P_2	0.6182	0.6182	0.6182	0.4546
	P_3	0.3084	0.3084	0.3084	0.4348
	P_4	0.7590	1.1590	0.9590	0.5000
(1, 1, 1)	P_1	0.4286	0.4286	0.4286	0.4286
	P_2	0.6182	0.6182	0.6182	0.4546
	P_3	0.6213	0.6213	0.6213	0.4783
	P_4	0.6243	0.6243	0.6243	0.5000
(0, 2, 1)	P_1	0.4286	0.4286	0.4286	0.4286
	P_2	0.2864	0.2864	0.2864	0.4091
	P_3	0.7392	1.1392	0.9392	0.4783
	P_4	0.6409	0.6350	0.6374	0.5000
(2, 0, 1)	P_1	0.4286	0.4286	0.4286	0.4286
	P_2	0.7500	1.1500	0.9500	0.5000
	P_3	0.3392	0.3392	0.3392	0.4783
	P_4	0.6590	0.6590	0.6590	0.5000
(2, 1, 0)	P_1	0.4286	0.4286	0.4286	0.4286
	P_2	0.7500	1.1500	0.9500	0.5000
	P_3	0.6517	0.6455	0.6479	0.4783
	P_4	0.3409	0.3350	0.3374	0.5000
(1, 2, 0)	P_1	0.4286	0.4286	0.4286	0.4286
	P_2	0.6182	0.6182	0.6182	0.5000
	P_3	0.7479	1.1479	0.9479	0.4783
	P_4	0.3374	0.3374	0.3374	0.5000

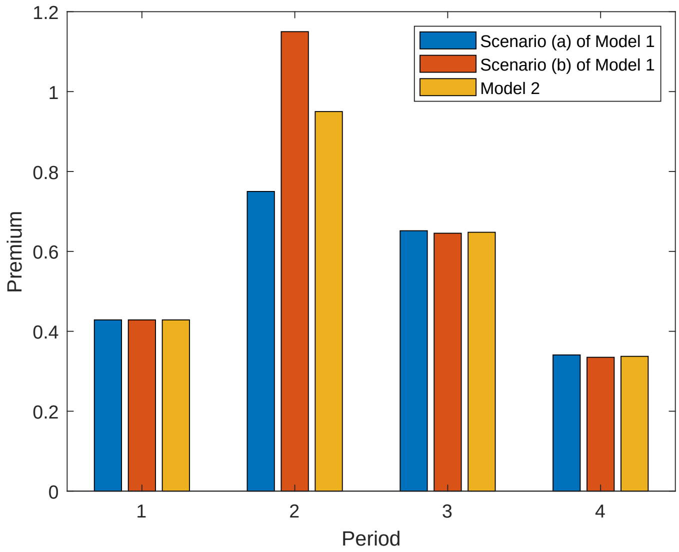

Figure 1. Predicted premiums for (1,2,0).

4. Empirical study

In Section 3, we fix the values of the parameters for different models in a special way to numerically illustrate the calculation of the Bayesian premiums for one policyholder. In this section, we implement all the models presented in this paper with a sample of insurance data, obtain the statistical estimates of the parameters and compare the three models in a more scientific way.

Suppose that the insurance portfolio consists of I policies, denote the unobservable heterogeneity of the ith policyholder by Θ_i, $i = 1, \ldots, I$ and define $N_{i,t}$ to be the claim count of the ith policyholder in period t, $t = 1, \ldots, T_i$. Then, for each policyholder in the portfolio, the related variables can be expressed as $(\Theta_i, N_{i,1}, \ldots, N_{i,T_i})$, in which the time length T_i may differ among the policyholders in the case when the data are unbalanced. Moreover, the sequences $\{(\Theta_i, N_{i,1}, \ldots, N_{i,T_i}), i = 1, 2, \ldots, I\}$ are supposed to be independent at the portfolio level.

Rewrite the parameters to be estimated as $\mathbf{\Phi} = (\boldsymbol{\beta}, \boldsymbol{\omega}, \alpha, \phi_1, \phi_2)$ in our proposed model, and let $\boldsymbol{n}_i = (n_{i,1}, \ldots, n_{i,T_i})$ denote the observations of $\boldsymbol{N}_i = (N_{i,1}, \ldots, N_{i,T_i})$, then the likelihood for the ith policyholder could be shown as

$$L_i(\mathbf{\Phi}; \boldsymbol{n}_i) = \frac{\lambda_i^{n_{i,1}}}{n_{i,1}!} \times \frac{\alpha}{\Gamma(\alpha)} \times \eta_2^{n_{i,2}} \cdots \eta_{T_i}^{n_{i,T_i}} \times \left[\sum_{z_{i,T_i}=0}^{\min\{n_{i,T_i-1}, n_{i,T_i}\}} \cdots \sum_{z_{i,2}=0}^{\min\{n_{i,1}, n_{i,2}\}} \pi(z_{i,T_i}, \ldots, z_{i,2}, n_{i,T_i}, \ldots, n_{i,1}) \right], \quad (4.1)$$

in which $\pi(z_{i,T_i}, \ldots, z_{i,2}, n_{i,T_i}, \ldots, n_{i,1})$ is defined by (2.10). Then, we can obtain the estimators $\hat{\mathbf{\Phi}} = (\hat{\boldsymbol{\beta}}, \hat{\omega}, \hat{\alpha}, \hat{\phi}_1, \hat{\phi}_2)$ by maximizing the following total log-likelihood function of the observed data derived by summing up the logarithm of (4.1) over all policyholders:

$$l(\mathbf{\Phi}) = \sum_{i=1}^{I} \log L_i(\mathbf{\Phi}; \boldsymbol{n}_i). \quad (4.2)$$

In the following part, we apply our proposed model to a panel data set for demonstrating our methods and results. The data are drawn from Wisconsin Local Government Property Insurance Fund (LGPIF), and have been used previously in Frees et al. [19], Quan and Valdez [20], Oh et al. [21, 22] and Chen et al. [23]. The LGPIF provides property insurance such as buildings insurance, vehicles insurance and equipment insurance, for various governmental entities, mostly including counties, cities, towns, villages, school districts, fire departments and other miscellaneous entities. Detailed information on the LGPIF data can be found in Frees et al. [19]. There are six types of insurance coverage in the data, and here we will consider only the collision coverage for new vehicles (i.e., vehicles of current model year, or 1∼2 years prior to the current model year), which provides coverage for the impact of a vehicle with an object, impact of vehicle with an attached vehicle, or the overturn of a vehicle. We first remove the policyholders who have zero coverage and then take the policyholders that have complete claim histories over the policy years from 2006 to 2010 as the training dataset for simplicity. Therefore, a longitudinal data with 199 governmental entities is used for our analysis to calibrate the model. We describe some summary statistics for the numbers of claims and the risk characteristics in Tables 2 and 3, respectively. As can be seen from the results, there is a clear increasing trend in the frequency, implying that the sequence of claim numbers for a policyholder could be nonstationary during these years.

Table 4 displays the effects of risk characteristics on the numbers of the claims, showing that theses variables are significant. For one thing, it seems that the county entities have the highest accident rate, while the town entities have the lowest risk. For another, we separate all the policyholders into two classes according to the median of the coverage, so that CoverageCN1 means the level with coverage $\in (0, 0.2051]$, and CoverageCN2 indicates the level with coverage $\in (0.2051, +\infty)$. As we can see, the drivers with higher coverage have more accidents, probably because of the moral hazard, i.e., the coverage of insurance provides incentives for reckless driving behavior, and make the policyholders have less risk aversion towards the loss.

In Frees et al. [19], the authors fit several commonly employed count models which include Poisson, zero-inflated Poisson, zero-one-inflated Poisson, negative binomial, zero-inflated negative binomial and zero-one-inflated negative binomial to the LGPIF dataset, and use Chi-square goodness-of-fit statistics to compare these different models. Their results show that negative binomial is significantly better than other models apart from zero-one-inflated negative binomial. The authors prefer and recommend the simpler model, so negative binomial is selected for further investigation. Follow this related study, it is reasonable for us to take CM-MP as the benchmark, then compare the fitting effect and predictive performance of CM-SETINAR(2,1) and CM-INAR(1) with CM-MP, to illustrate our main discoveries.

We now apply the three models aforementioned, i.e., CM-SETINAR(2,1), CM-INAR(1) and CM-MP, to the LGPIF data. The parameters are estimated via the MLE (maximum likelihood estimation) method based on the log-likelihood function defined by (4.2), and the estimates obtained using *optim* function in R are presented in Table 5. In the meantime, we also exhibit two goodness-of-fit measures, including the log-likelihood function calculated at the estimated parameters (LogLik) and the Akaike information criterion (AIC), to evaluate different models which we focus on. It can be seen that both of the criteria support our proposed model, because it has the largest value of LogLik and the smallest value of AIC.

Table 2. Numbers of claims over policy years.

Year	Frequency	Count													
		0	1	2	3	4	5	6	7	8	9	11	13	14	15
2006	0.4874	141	35	14	5	2	1	1	0	0	0	0	0	0	0
2007	0.5980	137	33	19	2	2	4	0	2	0	0	0	0	0	0
2008	0.6532	142	34	8	4	6	0	1	0	2	0	2	0	0	0
2009	0.6783	138	39	9	5	2	2	0	0	2	0	0	0	1	1
2010	0.6934	137	34	14	5	3	2	1	0	0	1	1	1	0	0

Table 3. Summary statistics for the risk characteristics.

No.	Variable name	Type	Description	Proportion/Mean
1	TypeCity	Categorical	Indicator for city entity	0.1457
2	TypeCounty	Categorical	Indicator for county entity	0.2261
3	TypeSchool	Categorical	Indicator for school entity	0.3316
4	TypeTown	Categorical	Indicator for town entity	0.1005
5	TypeVillage	Categorical	Indicator for village entity	0.1658
6	CoverageCN	Continuous	Log coverage amount of CN (in millions of dollars)	0.4624

Table 4. Percentages of claims by risk characteristics.

Variables	Percentage by count								
	0	1	2	3	4	5	6	7	8+
TypeCity	0.6414	0.2207	0.0966	0.0138	0.0069	0.0138	0.0000	0.0000	0.0069
TypeCounty	0.3244	0.2667	0.1733	0.0800	0.0578	0.0311	0.0133	0.0089	0.0442
TypeSchool	0.8697	0.1182	0.0091	0.0030	0.0000	0.0000	0.0000	0.0000	0.0069
TypeTown	0.9300	0.0700	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
TypeVillage	0.7394	0.2061	0.0485	0.0000	0.0060	0.0000	0.0000	0.0000	0.0000
CoverageCN1	0.8554	0.1406	0.0020	0.0020	0.0000	0.0000	0.0000	0.0000	0.0000
CoverageCN2	0.5412	0.2112	0.1268	0.0402	0.0302	0.0181	0.0060	0.0040	0.0220

It should be noted that we have to determine the value of the threshold parameter r in our proposed CM-SETINAR(2,1) before estimating the other parameters. To this end, we can search the integer that maximizes the LogLik criterion function, in an appropriate subset of positive integers $[1, 2 \ldots, R]$. For the LGPIF data, we may set $R = 14$ because a maximum of 15 claims has been filed by the policyholders. Therefore, the algorithm to get the unknown threshold r can be summarized in the following two steps:

Step 1. For each fixed $r \in [1, 2, \ldots, 14]$, find $\hat{\mathbf{\Phi}}^{(r)}$ such that

$$\hat{\mathbf{\Phi}}^{(r)} = \arg\max_{\mathbf{\Phi}} l(\mathbf{\Phi}),$$

Table 5. Parameter estimates and goodness of fit for different models.

CM-SETINAR(2,1)				CM-INAR(1)				CM-MP	
Parameter	Estimate	Parameter	Estimate	Parameter	Estimate	Parameter	Estimate	Parameter	Estimate
β_0	-2.2771	ω_0	-2.6284	β_0	-4.1113	ω_0	-2.6509	β_0	-1.6564
β_1	1.2622	ω_1	1.7255	β_1	3.2982	ω_1	1.8678	β_1	1.0227
β_2	2.2116	ω_2	2.7621	β_2	4.3132	ω_2	2.9645	β_2	2.1359
β_3	0.1427	ω_3	0.5719	β_3	1.7646	ω_3	0.5707	β_3	-0.2981
β_4	-0.7662	ω_4	-0.2850	β_4	0.2627	ω_4	-0.1644	β_4	-1.1208
β_5	1.3018	ω_5	0.9861	β_5	3.2895	ω_5	1.1964	β_5	0.4774
β_6	0.3275	ω_6	0.3008	β_6	0.0712	ω_6	0.1649	β_6	0.1059
α	1.9141			α	1.4418			α	1.1863
ϕ_1	0.1105			ϕ	0.1491				
ϕ_2	0.5297								
LogLik		-789.7122			-796.9141			-807.5219	
AIC		1613.424			1625.828			1631.044	

in which $l(\mathbf{\Phi})$ is defined by (4.2).

Step 2. The threshold parameter is estimated by searching over all candidates, i.e.,

$$\hat{r} = \arg \max_{r \in [1,2,\ldots,14]} l(\hat{\mathbf{\Phi}}^{(r)}).$$

In practice, we also can search the threshold parameter by minimizing the values of AIC or optimize other criterion functions. Figures 2 and 3 show different r and the corresponding values of LogLik and AIC, respectively. It is found that $r = 6$ is the best choice. Accordingly, we can then obtain the estimates of the other parameters, as shown in Table 5.

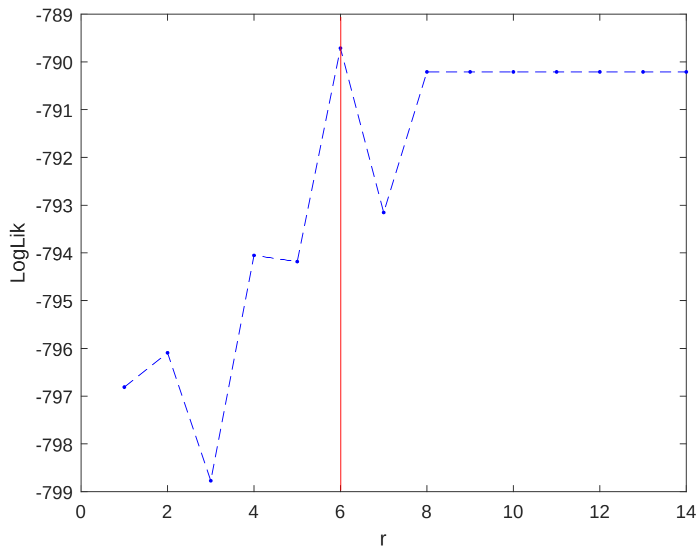

Figure 2. Different r and LogLik.

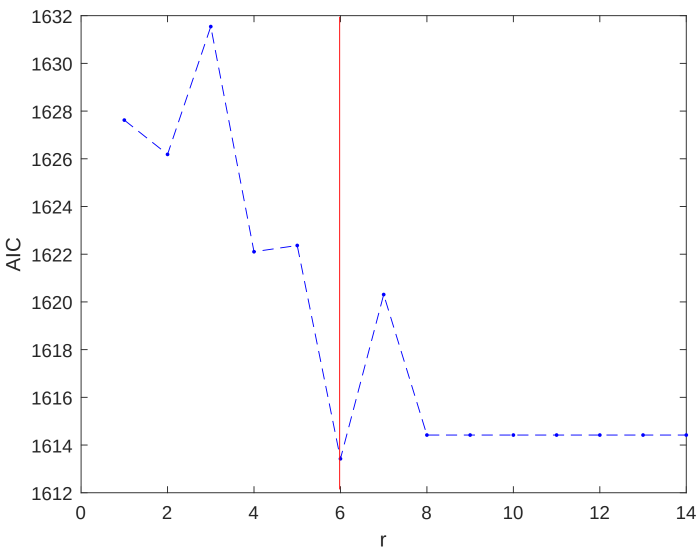

Figure 3. Different r and AIC.

At last, we compare the predictive performance of CM-SETINAR(2,1), CM-INAR(1) and CM-MP by performing an out-of-sample experiment. To this end, the observations in year 2011 with 1098 policyholders are considered. Of these samples, only 174 policyholders are identical to the governmental entities in the training dataset. Therefore, a longitudinal data with the 174 samples is reserved as the testing dataset for model validation. Since the dataset used is balanced now, we denote $N_{j,T+1}$ to be the claim number in the testing period of the jth policyholder, $j = 1, \ldots, J$, where J is the size of the testing dataset and equals to 174 in our analysis. Correspondingly, the observation of $N_{j,T+1}$ is denoted by $n_{j,T+1}$

In insurance claims modeling, it is more important to check the overall distribution for all policyholders rather than prediction of the claim frequency for each policyholder, which could be more useful for premium calculation, risk management and so forth. As a result, we adopt three measures to assess the predictive performance for each model, i.e., the log-likelihood function (LogLik), the mean squared error of prediction (MSPE) and the mean absolute error of prediction (MAPE), which are accordingly defined by

$$\text{LogLik} = \sum_{j=1}^{J} \log\left(P(N_{j,T+1} = n_{j,T+1} | N_{j,1} = n_{j,1}, \ldots, N_{j,T} = n_{j,T})\right),$$

$$\text{MSPE} = \sum_{j=1}^{J} \log\left[n_{j,T+1} - E(N_{j,T+1} | N_{j,1} = n_{j,1}, \ldots, N_{j,T} = n_{j,T})\right]^2,$$

$$\text{MAPE} = \sum_{j=1}^{J} \log\left|n_{j,T+1} - E(N_{j,T+1} | N_{j,1} = n_{j,1}, \ldots, N_{j,T} = n_{j,T})\right|.$$

From the results displayed in Table 6, it is clear that our proposed model, CM-SETINAR(2,1), has the best predictive performance with the largest LogLik among other two models. Furthermore, the smallest MSPE and MAPE also show that CM-SETINAR(2,1) outperforms other two models, which is consistent with the fitting result in Table 5.

Table 6. Predictive performance of different models.

	CM-SETINAR(2,1)	CM-INAR(1)	CM-MP
LogLik	-121.65672	-125.4629	-128.2358
MSPE	0.5903074	0.6320958	0.9499252
MAPE	0.4722917	0.4904699	0.5391067

5. Conclusions

In this paper, we apply the integer-valued self-exciting threshold autoregressive time series with Gamma distribution for the heterogeneity to the credibility theory and propose a credibility model based on heterogeneous SETINAR(2,1) process. We derive the explicit Bayesian premium formula, and consider the ratemaking problem of pricing the automobile insurance contract. Numerical examples show that our proposed model can account for the claim history and the behavior of the policyholders more flexibly. A real data analysis demonstrates that our approach provides a superior pattern to the Bayesian premium calculation when compared to the outcomes of two existing models, in view of goodness of fit and predictive ability.

Finally, the study of this paper might be extended in several directions. First, some other distributions, such as inverse Gaussian, LogNormal, mixed Gamma, etc., could be taken into account for the unobserved heterogeneity. Besides, the dynamic heterogeneity is also a very interesting topic for further research. Second, higher-order processes could be adopted to make the credibility models more practical. Third, multivariate integer-valued time series models could be considered to deal with the contracts with different types of coverage.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors would like to thank the editor and the referees for their constructive and pertinent comments and suggestions.

Conflict of interest

The authors declare that they have no competing interests.

References

1. J. Lemaire, *Bonus-malus systems in automobile insurance*, Dordrecht: Kluwer Academic Publisher, 1995. https://doi.org/10.1007/978-94-011-0631-3
2. H. Bühlmann, G. Alois, *A course in credibility theory and its applications*, Berlin, Heidelberg: Springer, 2005. https://doi.org/10.1007/3-540-29273-X
3. M. Denuit, X. Maréchal, S. Pitrebois, J. F. Walhin, *Actuarial modelling of claim counts: risk classification, credibility and bonus-malus scales*, New York: Wiley, 2007. https://doi.org/10.1002/9780470517420
4. J. Pinquet, Experience rating in non-life insurance, In: G. Dionne, *Handbook of insurance*, New York: Springer, 2013, 471–485. https://doi.org/10.1007/978-1-4614-0155-1_17
5. J. Pinquet, M. Guillén, C. Bolancé, Allowance for age of claims in bonus-malus systems, *ASTIN Bull.: J. IAA*, **31** (2001), 337–348. https://doi.org/10.2143/AST.31.2.1009
6. N. Brouhns, M. Guillén, M. Denuit, J. Pinquet, Bonus-malus scales in segmented tariffs with stochastic migration between segments, *J. Risk Insur.*, **70** (2003), 577–599. https://doi.org/10.1046/j.0022-4367.2003.00066.x
7. O. Purcaru, M. Guillén, M. Denuit, Linear credibility models based on time series for claim counts, *Belg. Actuarial Bull.*, **4** (2004), 62–74.
8. C. Bolancé, M. Denuit, M. Guillén, P. Lambert, Greatest accuracy credibility with dynamic heterogeneity: the Harvey-Fernandes model, *Belg. Actuarial Bull.*, **7** (2007), 14–18.

9. P. Shi, E. A. Valdez, Longitudinal modeling of insurance claim counts using jitters, *Scand. Actuar. J.*, **2** (2014), 159–179. https://doi.org/10.1080/03461238.2012.670611
10. A. Abdallah, J. P. Boucher, H. Cossette, Sarmanov family of multivariate distributions for bivariate dynamic claim counts model, *Insur. Math. Econ.*, **68** (2016), 120–133. https://doi.org/10.1016/j.insmatheco.2016.01.003
11. C. Gourieroux, J. Jasiak, Heterogeneous INAR(1) model with application to car insurance, *Insur. Math. Econ.*, **34** (2004), 177–192. https://doi.org/10.1016/j.insmatheco.2003.11.005
12. M. A. Al-Osh, A. A. Alzaid, First-order integer-valued autoregressive INAR(1) process, *J. Time Ser. Anal.*, **8** (1987), 261–275. https://doi.org/10.1111/j.1467-9892.1987.tb00438.x
13. L. Bermúdez, M. Guillén, D. Karlis, Allowing for time and cross dependence assumptions between claim counts in ratemaking models, *Insur. Math. Econ.*, **83** (2018), 161–169. https://doi.org/10.1016/j.insmatheco.2018.06.003
14. L. Bermúdez, D. Karlis, Multivariate INAR(1) regression models based on the Sarmanov distribution, *Mathematics*, **9** (2021), 505. https://doi.org/10.3390/math9050505
15. P. C. Zhang, Z. Z. Chen, G. Tzougas, E. Calderín-Ojeda, A. Dassios, X. Y. Wu, Multivariate zero-inflated INAR(1) model with an application in automobile insurance, *SSRN*, 2022. https://doi.org/10.2139/ssrn.4170555
16. X. Hu, J. Yao, A combined integer-valued autoregressive process with actuarial applications, unpublished paper, 2023. https://doi.org/10.21203/rs.3.rs-2734214/v1
17. M. Monteiro, M. G. Scotto, I. Pereira, Integer-valued self-exciting threshold autoregressive processes, *Commun. Stat.-Theory M.*, **41** (2012), 2717–2737. https://doi.org/10.1080/03610926.2011.556292
18. S. Asmussen, Modeling and performance of bonus-malus systems: stationarity versus age-correction, *Risks*, **2** (2014), 49–73. https://doi.org/10.3390/risks2010049
19. E. W. Frees, G. Lee, L. Yang, Multivariate frequency-severity regression models in insurance, *Risks*, **4** (2016), 4. https://doi.org/10.3390/risks4010004
20. Z. Y. Quan, E. A. Valdez, Predictive analytics of insurance claims using multivariate decision trees, *Depend. Model.*, **6** (2018), 377–407.
21. R. Oh, S. Peng, J. Y. Ahn, Bonus-malus premiums under the dependent frequency-severity modeling, *Scand. Actuar. J.*, **2020** (2020), 172–195. https://doi.org/10.1080/03461238.2019.1655477
22. R. Oh, J. H. T. Kim, J. Y. Ahn, Designing a bonus-malus system reflecting the claim size under the dependent frequency-severity model, *Probab. Eng. Informa. Sci.*, **36** (2022), 963–987. https://doi.org/10.1017/S0269964821000188
23. Z. Z. Chen, A. Dassios, G. Tzougas, EM estimation for bivariate mixed poisson INAR(1) claim count regression models with correlated random effects, *Eur. Actuar. J.*, 2023. https://doi.org/10.1007/s13385-023-00351-7