North American Actuarial Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uaaj20

Multivariate Zero-Inflated INAR(1) Model with an Application in Automobile Insurance

Pengcheng Zhang, Zezhun Chen, George Tzougas, Enrique Calderín–Ojeda, Angelos Dassios & Xueyuan Wu

To cite this article: Pengcheng Zhang, Zezhun Chen, George Tzougas, Enrique Calderín–Ojeda, Angelos Dassios & Xueyuan Wu (19 Sep 2024): Multivariate Zero-Inflated INAR(1) Model with an Application in Automobile Insurance, North American Actuarial Journal, DOI: 10.1080/10920277.2024.2381726

To link to this article: https://doi.org/10.1080/10920277.2024.2381726

Published online: 19 Sep 2024.

North American Actuarial Journal, 0(0), 1–19, 2024

ISSN: 1092-0277 print / 2325-0453 online
DOI: 10.1080/10920277.2024.2381726

Multivariate Zero-Inflated INAR(1) Model with an Application in Automobile Insurance

Pengcheng Zhang,[1] Zezhun Chen,[2] George Tzougas,[3] Enrique Calderín–Ojeda,[4] Angelos Dassios,[2] and Xueyuan Wu[4]

[1]*School of Insurance, Shandong University of Finance and Economics, Jinan, China.*
[2]*Department of Statistics, London School of Economics and Political Science, London, UK.*
[3]*Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, UK.*
[4]*Department of Economics, University of Melbourne, Melbourne, Victoria, Australia.*

The objective of this article is to propose a comprehensive solution for analyzing multidimensional non-life claim count data that exhibits time and cross-dependence, as well as zero inflation. To achieve this, we introduce a multivariate INAR(1) model, with the innovation term characterized by either a multivariate zero-inflated Poisson distribution or a multivariate zero-inflated hurdle Poisson distribution. Additionally, our modeling framework accounts for the impact of individual and coverage-specific covariates on the mean parameters of each model, thereby facilitating the computation of customized insurance premiums based on varying risk profiles. To estimate the model parameters, we employ a novel expectation-maximization (EM) algorithm. Our model demonstrates satisfactory performance in the analysis of European motor third-party liability claim count data.

1. INTRODUCTION

Claim count modeling is an essential part in the calculation of premiums. Due to the inclusion of multiple types of coverage in insurance policies, there is a need for multivariate count models to effectively capture the dependence structures between different count responses. Many attempts have been made in actuarial literature to develop appropriate multivariate count models. One common approach is to leverage common shock variables. This method proves useful in the multivariate Poisson model because the sum of independent Poisson random variables still follows a Poisson distribution. The application of the multivariate Poisson model in an actuarial setting can be found in Bermúdez and Karlis (2011). Another method involves the use of copulas, which offers the advantage of treating marginals and dependence structures separately. Shi and Valdez (2014b) demonstrated the application of copulas directly on negative binomial marginals, and Zhang et al. (2023) constructed a copula based on the mixing parameters of mixed Poisson distributions. Sarmanov distributions serve as an additional approach. Bolancé and Vernic (2019) considered three trivariate Sarmanov distributions combined with generalized linear models for marginals, and they fit these distributions to car insurance data. Mixture count models have also been explored to describe correlations in insurance. Fung, Badescu, and Lin (2019) and Tzougas and Pignatelli di Cerchiara (2021) investigated the application of mixture count models in this context.

As evidenced by empirical data, non-life insurance claim count data often exhibit a high prevalence of zeros. Zhang, Pitt, and Wu (2022) noted that the zero inflation phenomenon in the multivariate context is more intricate compared to the univariate case, emphasizing the need to employ multivariate zero-inflated models to explore cross-dependence. The existing methods have primarily focused on scenarios where margins follow Poisson distributions. Li et al. (1999) proposed a multivariate zero-inflated Poisson model comprising $m + 2$ components of m-dimensional discrete distributions. However, the complexity of

Address correspondence to Zezhun Chen, Department of Statistics, London School of Economics and Political Science, London WC2A 2AE, UK. E-mail: czz0328@outlook.com

this model makes maximum likelihood estimation challenging for large m. Recently, Liu and Tian (2015) introduced a new multivariate zero-inflated Poisson model that addresses computational issues associated with dimensionality. Nonetheless, the assumption of a standard Poisson distribution for each margin limits the model's adaptability to diverse features. To tackle this limitation, Zhang, Pitt, and Wu (2022) proposed a multivariate zero-inflated hurdle model, assuming zero-modified distributions for margins. This model offers greater flexibility in handling various features across margins.

The insurance company often keeps track of their policyholders for several years, enabling repeated observations that yield valuable insights into individuals' risk profiles. Consequently, an expanding research field in insurance has emerged to study panel count data models. One of the most commonly employed models is the multiplicative random effect Poisson model, where the random effect is assumed to follow a Gamma distribution. For a comprehensive description of this random effect model and its generalization, refer to Boucher, Denuit, and Guillen (2009). Copulas have been explored as an alternative method to model panel counts. In Shi and Valdez (2014a), the utility of copulas in capturing relationships within panel count data was investigated, and a jittering method was employed for estimation purposes. Furthermore, the application of time series models to analyze longitudinal insurance data is available in the literature. For instance, Gourieroux and Jasiak (2004) utilized the integer-valued autoregressive (INAR) model of order one based on the binomial thinning operator to update premiums in car insurance. An exhaustive overview of the aforementioned models as applied to actuarial science was provided in Boucher, Denuit, and Guillén (2008).

However, it is worth noting that though there is existing literature on models that solely considered cross-dependence or time dependence, the integration of these two types of dependence into a single model remains a relatively unexplored area of research. One natural approach is to generalize the univariate INAR model to the multivariate case. In this approach, the innovation term in the model captures cross-dependence, and the lag term accounts for time dependence. Within this framework, Bermúdez, Guillén, and Karlis (2018) utilized a bivariate Poisson distribution as the innovation term to allow for cross-correlations. Bermúdez and Karlis (2021) considered multivariate discrete distributions defined using the Sarmanov family as the innovation term.

In this article, we introduce two multivariate INAR(1) models by modeling the innovation terms based on the multivariate zero-inflated Poisson distribution (INAR-MZIP) and the multivariate zero-inflated hurdle Poisson (INAR-MZIHP) distribution. Our contribution to the existing literature is threefold. Firstly, we propose a flexible framework that simultaneously addresses various features observed in the data, such as an excess of common zeros, cross-dependence, and time dependence. Unlike previous studies, our models can be applied in a general multivariate case, not limited to just a bivariate setting. This distinguishes our work from that of Bermúdez, Guillén, and Karlis (2018), which focused only on a two-dimensional scenario. Furthermore, we consider multivariate zero inflation, which is commonly observed in automobile insurance data, whereas Bermúdez, Guillén, and Karlis (2018) and Bermúdez and Karlis (2021) failed to incorporate this phenomenon. Thus, our models can be regarded as important additions to the existing literature. Secondly, we develop a novel expectation-maximization (EM) algorithm to estimate the parameters in our proposed models. The log-likelihood function in these models has a complex form, making direct maximization challenging. However, by employing the EM algorithm, we can significantly simplify the inference procedure. To the best of our knowledge, the EM algorithm we establish for the multivariate INAR model is innovative and has not been previously proposed. Finally, we evaluate the suitability of the multivariate zero-inflated INAR(1) models from multiple perspectives. Through extensive analysis, we demonstrate that our proposed models outperform other potential alternatives in terms of model fitting and predictive performance.

The rest of this article is organized as follows. In Section 2 we provide a brief review of multivariate zero-inflated distribution, focusing on two specific cases: the multivariate zero-inflated Poisson (MZIP) distribution and the multivariate zero-inflated hurdle Poisson distribution (MZIHP). Section 3 details the formulation of two types of multivariate INAR(1) models with MZIP and MZIHP as innovations. The corresponding EM algorithms for parameter estimation are presented in each case. Section 4 presents a simulation study aimed at illustrating the efficacy of our proposed EM algorithms. In Section 5, we delve into a practical application to illustrate the utility of our proposed models. The last section concludes the article.

2. MULTIVARIATE ZERO-INFLATED DISTRIBUTION

The multivariate zero-inflated distribution can be defined as follows. Let $\boldsymbol{Y} = (Y_1, ..., Y_m)^\top$ denote a discrete random vector where Y_j, $j = 1, ..., m$, are independent of each other and defined on $\mathbb{N}$. Then $\boldsymbol{N} = (N_1, ..., N_m)^\top$ is said to follow the multivariate zero-inflated distribution if

$$\boldsymbol{N} = U_0 \boldsymbol{Y} = \begin{cases} \boldsymbol{0}_m, & U_0 = 0 \\ \boldsymbol{Y}, & U_0 = 1, \end{cases} \tag{2.1}$$

where $U_0 \sim \text{Bernoulli}(\pi_0),\ 0 < \pi_0 < 1,$ and U_0 is independent of $\boldsymbol{Y}$. The probability mass function (pmf) of $\boldsymbol{N}$ can be derived as

$$\Pr(\boldsymbol{N} = \boldsymbol{n}) = \left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} \Pr(Y_j = 0)\right]^{v} \left[\pi_0 \prod_{j=1}^{m} \Pr(Y_j = n_j)\right]^{1-v}, \tag{2.2}$$

where $\boldsymbol{n} = (n_1, ..., n_m)^\top$ is a vector of observed values, $v = \mathbb{I}(\boldsymbol{n} = \boldsymbol{0}_m),$ and $\mathbb{I}(\cdot)$ is an indicator function.

2.1. Multivariate Zero-Inflated Poisson Distribution

Let $Y_j \sim \text{Poisson}(\lambda_j),$ for $j = 1, ..., m$. Then $\boldsymbol{N}$ is said to follow the multivariate zero-inflated Poisson distribution with the parameter vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_m)^\top$ and a zero inflation parameter $\pi_0,$ denoted by $\boldsymbol{N} \sim \text{MZIP}(\boldsymbol{\lambda}, \pi_0)$. The pmf of $\boldsymbol{N}$ is

$$\Pr(\boldsymbol{N} = \boldsymbol{n}) = \left(1 - \pi_0 + \pi_0 e^{-\sum_{j=1}^{m} \lambda_j}\right)^{v} \left(\pi_0 \prod_{j=1}^{m} \frac{\lambda_j^{n_j} e^{-\lambda_j}}{n_j!}\right)^{1-v}. \tag{2.3}$$

2.2. Multivariate Zero-Inflated Hurdle Poisson Distribution

We shall assume that each $Y_j,\ j = 1, ..., m,$ follows a zero-modified Poisson distribution, which can be characterized as follows:

$$Y_j = U_j W_j = \begin{cases} 0, & U_j = 0, \\ W_j, & U_j = 1, \end{cases} \tag{2.4}$$

where W_j follows a unit-shifted Poisson distribution with the following pmf:

$$\Pr(W_j = n_j) = \frac{\lambda_j^{n_j - 1} e^{-\lambda_j}}{(n_j - 1)!}, \quad n_j > 0. \tag{2.5}$$

$U_j \sim \text{Bernoulli}(\pi_j),\ 0 < \pi_j < 1,$ and U_j is independent of W_j. Then $\boldsymbol{N}$ is said to follow the multivariate zero-inflated hurdle Poisson distribution with parameter vectors $\boldsymbol{\pi} = (\pi_1, ..., \pi_m)^\top,\ \boldsymbol{\lambda} = (\lambda_1, ..., \lambda_m)^\top$ and a zero inflation parameter $\pi_0,$ denoted by $\boldsymbol{N} \sim \text{MZIHP}(\boldsymbol{\pi}, \boldsymbol{\lambda}, \pi_0)$. The pmf of $\boldsymbol{N}$ is

$$\begin{aligned} \Pr(\boldsymbol{N} = \boldsymbol{n}) \quad &= \left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} (1 - \pi_j)\right]^{v} \\ &\times \left[\pi_0 \prod_{j:n_j=0} (1 - \pi_j) \prod_{j:n_j \neq 0} \pi_j \frac{\lambda_j^{n_j - 1} e^{-\lambda_j}}{(n_j - 1)!}\right]^{1-v}. \end{aligned} \tag{2.6}$$

3. MULTIVARIATE INAR(1) MODEL

3.1. The Model

The multivariate INAR model of order 1 is defined as follows:

$$\boldsymbol{N_t} = \boldsymbol{P} \circ \boldsymbol{N_{t-1}} + \boldsymbol{R_t}, \tag{3.1}$$

where $\boldsymbol{N_t} = (N_{1t}, ..., N_{mt})^\top$ and $\boldsymbol{N_{t-1}} = (N_{1,t-1}, ..., N_{m,t-1})^\top$. $\boldsymbol{R}_t = (R_{1t}, ..., R_{mt})^\top$ is referred to as innovations. $\boldsymbol{P}$ is assumed to be a diagonal matrix in our model, which is written as

$$\boldsymbol{P} = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_m \end{pmatrix}, \tag{3.2}$$

$\boldsymbol{P}\circ$ acts as the usual matrix multiplication, while reserving the properties of the binomial thinning operation. Each series then can be written as

$$\begin{aligned} N_{1t} &= p_1 \circ N_{1,t-1} + R_{1t}, \\ N_{2t} &= p_2 \circ N_{2,t-1} + R_{2t}, \\ &\vdots \\ N_{mt} &= p_m \circ N_{m,t-1} + R_{mt}. \end{aligned} \tag{3.3}$$

Each operation is defined as $p \circ N = \sum_{i=1}^{N} Z_i,$ where $Z_i \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(p)$.

The pmf of $\boldsymbol{N}_t$ conditional on the last state $\boldsymbol{N}_{t-1}$ is given by

$$f(\boldsymbol{n}_t|\boldsymbol{n}_{t-1}) = \sum_{y_{1t}=0}^{s_{1t}} \cdots \sum_{y_{mt}=0}^{s_{mt}} \left(\prod_{j=1}^{m} g_j(y_{jt}; n_{j,t-1}, p_j) \right) f_{\boldsymbol{R}}(\boldsymbol{n}_t - \boldsymbol{y}_t), \tag{3.4}$$

where $\boldsymbol{n_t} = (n_{1t}, ..., n_{mt})^\top$ and $\boldsymbol{n_{t-1}} = (n_{1,t-1}, ..., n_{m,t-1})^\top$ are observed values. $s_{jt} = \min(n_{j,t-1}, n_{jt})$ and $\boldsymbol{y}_t = (y_{1t}, ..., y_{mt})^\top$. We denote $g_j(y; n, p)$ as the pmf of a binomial variable with parameters n and p and $f_{\boldsymbol{R}}$ as the joint pmf of the random vector $\boldsymbol{R}_t$. For our purpose, we assume that $\boldsymbol{R}_t$ follows a multivariate zero-inflated distribution.

3.2. Multivariate INAR(1) Model with MZIP as Innovations

We denote N_{ijt} as the number of claims for the ith individual and for claim type j at time point t, where $i = 1, ..., n$, $j = 1, ..., m,\ t = 1, ..., T_i,$ and n_{ijt} as the observed value. Now we introduce some covariates $\boldsymbol{x}_{it},$ where $\boldsymbol{x}_{it} = (1, x_{it1}, ..., x_{itp})^\top$. Here we use the same set of covariates for each claim type. However, a specific policyholder's information may change over time. The location parameter λ_{ijt} can then be modeled as

$$\lambda_{ijt} = \exp(\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_j), \quad i = 1, ..., n, \quad j = 1, ..., m, \quad t = 1, ..., T_i, \tag{3.5}$$

where $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, ..., \beta_{jp})^\top$ is the parameter vector to estimate. For the purpose of easy interpretation, we do not inject covariates in π_0 and p_j.

Suppose now we observe the values $\boldsymbol{y}_t$ from the latent random vector $\boldsymbol{Y}_t = (p_1 \circ N_{1,t-1}, ..., p_m \circ N_{m,t-1})^\top$. Then the joint pmf can be written as (we omit i for simplicity)

$$f(\boldsymbol{n}_t|\boldsymbol{n}_{t-1}, \boldsymbol{y}_t) = \left(\prod_{j=1}^{m} g_j(y_{jt}; n_{j,t-1}, p_j) \right) f_{\boldsymbol{R}}(\boldsymbol{n}_t - \boldsymbol{y}_t). \tag{3.6}$$

Furthermore, suppose we also observe latent variables v_t and $u_t,$ where $v_t = \mathbb{I}(\boldsymbol{n}_t - \boldsymbol{y}_t = \boldsymbol{0}_m)$ and $u_t = 1$ indicates that the common zeros come from the zero inflation part. The joint pmf can be further decomposed as

$$\begin{aligned} f(\boldsymbol{n}_t|\boldsymbol{n}_{t-1}, \boldsymbol{y}_t, u_t, v_t) &= \left(\prod_{j=1}^{m} g_j(y_{jt}; n_{j,t-1}, p_j) \right) \left[(1-\pi_0)^{u_t} \left(\pi_0 \prod_{j=1}^{m} h_j(0) \right)^{1-u_t} \right]^{v_t} \\ &\quad \times \left[\pi_0 \prod_{j=1}^{m} h_j(n_{jt} - y_{jt}) \right]^{1-v_t}, \end{aligned} \tag{3.7}$$

where h_j denotes the pmf of the Poisson distribution.

The complete log-likelihood function is then given by

$$\begin{aligned}\ell_c(\Theta) \quad \propto & \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T_i}\left[y_{ijt}\log p_j + (n_{ij,t-1}-y_{ijt})\log(1-p_j)\right] \\ & + \sum_{i=1}^{n}\sum_{t=1}^{T_i}[u_{it}v_{it}\log(1-\pi_0) + (1-u_{it}v_{it})\log\pi_0] \\ & + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T_i}\left[(n_{ijt}-y_{ijt})\log\lambda_{ijt} - (1-u_{it}v_{it})\lambda_{ijt}\right],\end{aligned} \tag{3.8}$$

where $\Theta = (\boldsymbol{p}, \pi_0, \boldsymbol{\beta})$, $\boldsymbol{p} = (p_1, ..., p_m)^\top$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m)$.

The Q function at the rth iteration is given by

$$\begin{aligned}Q(\Theta;\Theta^{(r)}) \quad = & \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T_i}\left[y_{ijt}^{(r)}\log p_j + \left(n_{ij,t-1}-y_{ijt}^{(r)}\right)\log(1-p_j)\right] \\ & + \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left[w_{it}^{(r)}\log(1-\pi_0) + \left(1-w_{it}^{(r)}\right)\log\pi_0\right] \\ & + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T_i}\left[\left(n_{ijt}-y_{ijt}^{(r)}\right)\log\lambda_{ijt} - \left(1-w_{it}^{(r)}\right)\lambda_{ijt}\right],\end{aligned} \tag{3.9}$$

where $w_{it} = u_{it}v_{it}$.

- **E-step:**

  1. The conditional expectation $y_{ijt}^{(r)}$ is given by

$$\begin{aligned}y_{ijt}^{(r)} \quad & = \mathbb{E}(y_{ijt}|\Theta^{(r)}, \boldsymbol{n}_{it}, \boldsymbol{n}_{i,t-1}) \\ & = \begin{cases} \dfrac{p_j^{(r)} n_{ij,t-1} f(\boldsymbol{n}_{it}-\mathbf{1}_j|\boldsymbol{n}_{i,t-1}-\mathbf{1}_j;\Theta^{(r)})}{f(\boldsymbol{n}_{it}|\boldsymbol{n}_{i,t-1};\Theta^{(r)})}, & n_{ijt}>0 \text{ and } n_{ij,t-1}>0, \\ 0, & \text{otherwise,} \end{cases}\end{aligned} \tag{3.10}$$

  where $\boldsymbol{n}_{it} = (n_{i1t}, ..., n_{imt})^\top$ and $\boldsymbol{n}_{i,t-1} = (n_{i1,t-1}, ..., n_{im,t-1})^\top$ are observed values, and $\mathbf{1}_j = (0, ..., 1, ..., 0)^\top$ is a unit vector with the jth element equal to one.

  2. The conditional expectation $w_{it}^{(r)}$ is given by

$$\begin{aligned}w_{it}^{(r)} \quad & = \mathbb{E}(u_{it}v_{it}|\Theta^{(r)}, \boldsymbol{n}_{it}, \boldsymbol{n}_{i,t-1}) = \Pr(u_{it}=v_{it}=1|\Theta^{(r)}, \boldsymbol{n}_{it}, \boldsymbol{n}_{i,t-1}) \\ & = \begin{cases} \dfrac{\prod_{j=1}^{m} g_j(n_{ijt}; n_{ij,t-1}, p_j^{(r)})(1-\pi_0^{(r)})}{f(\boldsymbol{n}_{it}|\boldsymbol{n}_{i,t-1};\Theta^{(r)})}, & n_{ijt}\le n_{ij,t-1} \text{ for } j=1,...,m, \\ 0, & \text{otherwise.} \end{cases}\end{aligned} \tag{3.11}$$

- **M-step**: Update the parameter set Θ to make $Q(\Theta^{(r+1)};\Theta^{(r)}) > Q(\Theta^{(r)};\Theta^{(r)})$.

  1. Update the parameter p_j:

$$p_j^{(r+1)} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i} y_{ijt}^{(r)}}{\sum_{i=1}^{n}\sum_{t=1}^{T_i} n_{ij,t-1}}, \quad j=1,...,m. \tag{3.12}$$

2. Update the parameter π_0:

$$\pi_0^{(r+1)} = 1 - \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i} w_{it}^{(r)}}{\sum_{i=1}^{n} T_i}. \tag{3.13}$$

3. Update the parameter vector $\boldsymbol{\beta}_j$ by implementing the Newton-Raphson method for one cycle:

$$\boldsymbol{\beta}_j^{(r+1)} = \boldsymbol{\beta}_j^{(r)} - \left[H(\boldsymbol{\beta}_j^{(r)})\right]^{-1} s(\boldsymbol{\beta}_j^{(r)}), \quad j = 1, ..., m, \tag{3.14}$$

where the score equation and Hessian matrix are given as follows:

$$\begin{aligned} s(\boldsymbol{\beta}_j^{(r)}) &= \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left[n_{ijt} - y_{ijt}^{(r)} - (1 - w_{it}^{(r)})\lambda_{ijt}^{(r)}\right]\boldsymbol{x}_{it}, \\ H(\boldsymbol{\beta}_j^{(r)}) &= -\sum_{i=1}^{n}\sum_{t=1}^{T_i}(1 - w_{it}^{(r)})\lambda_{ijt}^{(r)}\boldsymbol{x}_{it}\boldsymbol{x}_{it}^{\top}. \end{aligned} \tag{3.15}$$

3.3. Multivariate INAR(1) Model with MZIHP as Innovations

In this model, we introduce covariates in both π_j and λ_j. The parameter π_{ijt} can be modeled as

$$\pi_{ijt} = \frac{\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\alpha}_j)}{1 + \exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\alpha}_j)} \quad i = 1, ..., n, \quad j = 1, ..., m, \quad t = 1, ..., T_i, \tag{3.16}$$

where $\boldsymbol{\alpha}_j = (\alpha_{j0}, \alpha_{j1}, ..., \alpha_{jp})^{\top}$ is the parameter vector to estimate. The parameter λ_{ijt} can be modeled as

$$\lambda_{ijt} = \exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_j), \quad i = 1, ..., n, \quad j = 1, ..., m, \quad t = 1, ..., T_i, \tag{3.17}$$

where $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, ..., \beta_{jp})^{\top}$ is the parameter vector to estimate.

Suppose now we observe the values $\boldsymbol{y}_t$ from the random vector $\boldsymbol{Y}_t = (p_1 \circ N_{1,t-1}, ..., p_m \circ N_{m,t-1})^{\top}$. Then the joint pmf can be written as (we omit i for simplicity)

$$f(\boldsymbol{n}_t|\boldsymbol{n}_{t-1}, \boldsymbol{y}_t) = \left(\prod_{j=1}^{m} g_j(y_{jt}; n_{j,t-1}, p_j)\right) f_R(\boldsymbol{n}_t - \boldsymbol{y}_t). \tag{3.18}$$

Furthermore, suppose we also observe latent variables v_t and u_t, where $v_t = \mathbb{I}(\boldsymbol{n}_t - \boldsymbol{y}_t = \boldsymbol{0}_m)$ and $u_t = 1$ indicates that the common zeros come from the zero inflation part. The joint pmf can be further decomposed as

$$\begin{aligned} f(\boldsymbol{n}_t|\boldsymbol{n}_{t-1}, \boldsymbol{y}_t, u_t, v_t) &= \left(\prod_{j=1}^{m} g_j(y_{jt}; n_{j,t-1}, p_j)\right)\left[(1 - \pi_0)^{u_t}\left(\pi_0 \prod_{j=1}^{m}(1 - \pi_{jt})\right)^{1-u_t}\right]^{v_t} \\ &\quad \times \left[\pi_0 \prod_{j:y_{jt}=n_{jt}} (1 - \pi_{jt}) \prod_{j:y_{jt}<n_{jt}} \pi_{jt} f_{W_j}(n_{jt} - y_{jt})\right]^{1-v_t}, \end{aligned} \tag{3.19}$$

where f_{W_j} denotes the pmf of the unit-shifted Poisson distribution.

The complete log-likelihood function is then given by

$$\begin{aligned}\ell_c(\Theta) \quad \propto & \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T_i}\left[y_{ijt}\log p_j + (n_{ij,t-1}-y_{ijt})\log(1-p_j)\right] \\ & + \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left[u_{it}v_{it}\log(1-\pi_0) + (1-u_{it}v_{it})\log\pi_0\right] \\ & + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T_i}\left[\tau_{ijt}\log\pi_{ijt} + (1-u_{it}v_{it}-\tau_{ijt})\log(1-\pi_{ijt})\right] \\ & + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T_i}\left[(n_{ijt}-y_{ijt}-\tau_{ijt})\log\lambda_{ijt} - \tau_{ijt}\lambda_{ijt}\right],\end{aligned} \tag{3.20}$$

where $\tau_{ijt} = \mathbb{I}(y_{ijt} < n_{ijt})$, $\Theta = (\boldsymbol{p}, \pi_0, \boldsymbol{\alpha}, \boldsymbol{\beta})$, $\boldsymbol{p} = (p_1, ..., p_m)^\top$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_m)$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m)$.

The Q function at the rth iteration is given by

$$\begin{aligned}Q(\Theta;\Theta^{(r)}) \quad = & \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T_i}\left[y_{ijt}^{(r)}\log p_j + \left(n_{ij,t-1}-y_{ijt}^{(r)}\right)\log(1-p_j)\right] \\ & + \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left[w_{it}^{(r)}\log(1-\pi_0) + \left(1-w_{it}^{(r)}\right)\log\pi_0\right] \\ & + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T_i}\left[\tau_{ijt}^{(r)}\log\pi_{ijt} + (1-w_{it}^{(r)}-\tau_{ijt}^{(r)})\log(1-\pi_{ijt})\right] \\ & + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T_i}\left[\left(n_{ijt}-y_{ijt}^{(r)}-\tau_{ijt}^{(r)}\right)\log\lambda_{ijt} - \tau_{ijt}^{(r)}\lambda_{ijt}\right],\end{aligned} \tag{3.21}$$

where $w_{it} = u_{it}v_{it}$.

- **E-step:**

1. The conditional expectation $y_{ijt}^{(r)}$ is given by

$$\begin{aligned}y_{ijt}^{(r)} \quad & = \mathbb{E}(y_{ijt}|\Theta^{(r)}, \boldsymbol{n}_{it}, \boldsymbol{n}_{i,t-1}) \\ & = \begin{cases} \dfrac{p_j^{(r)} n_{ij,t-1} f(\boldsymbol{n}_{it}-\mathbf{1}_j|\boldsymbol{n}_{i,t-1}-\mathbf{1}_j;\Theta^{(r)})}{f(\boldsymbol{n}_{it}|\boldsymbol{n}_{i,t-1};\Theta^{(r)})}, & n_{ijt}>0 \text{ and } n_{ij,t-1}>0, \\ 0, & \text{otherwise,} \end{cases}\end{aligned} \tag{3.22}$$

where $\boldsymbol{n}_{it} = (n_{i1t}, ..., n_{imt})^\top$ and $\boldsymbol{n}_{i,t-1} = (n_{i1,t-1}, ..., n_{im,t-1})^\top$ are observed values, and $\mathbf{1}_j = (0, ..., 1, ..., 0)^\top$ is a unit vector with the jth element equal to one.

2. The conditional expectation $w_{it}^{(r)}$ is given by

$$\begin{aligned}w_{it}^{(r)} \quad & = \mathbb{E}(u_{it}v_{it}|\Theta^{(r)}, \boldsymbol{n}_{it}, \boldsymbol{n}_{i,t-1}) = \Pr(u_{it}=v_{it}=1|\Theta^{(r)}, \boldsymbol{n}_{it}, \boldsymbol{n}_{i,t-1}) \\ & = \begin{cases} \dfrac{\prod_{j=1}^{m} g_j(n_{ijt}; n_{ij,t-1}, p_j^{(r)})(1-\pi_0^{(r)})}{f(\boldsymbol{n}_{it}|\boldsymbol{n}_{i,t-1};\Theta^{(r)})}, & n_{ijt} \le n_{ij,t-1} \text{ for } j=1,...,m, \\ 0, & \text{otherwise.} \end{cases}\end{aligned} \tag{3.23}$$

3. The conditional expectation $\tau_{ijt}^{(r)}$ is given by

$$\begin{aligned}\tau_{ijt}^{(r)} &= \mathbb{E}(\tau_{ijt}|\Theta^{(r)}, \boldsymbol{n}_{it}, \boldsymbol{n}_{i,t-1}) = \Pr(\tau_{ijt}=1|\Theta^{(r)}, \boldsymbol{n}_{it}, \boldsymbol{n}_{i,t-1}) \\ &= \begin{cases} 1 - \dfrac{f(\boldsymbol{n}_{it}|\boldsymbol{n}_{i,t-1}, y_{ijt}=n_{ijt}; \Theta^{(r)})}{f(\boldsymbol{n}_{it}|\boldsymbol{n}_{i,t-1}; \Theta^{(r)})}, & n_{ijt} \le n_{ij,t-1}, \\ 1, & n_{ijt} > n_{ij,t-1}. \end{cases}\end{aligned} \tag{3.24}$$

- **M-step**: Update the parameter set Θ to make $Q(\Theta^{(r+1)}; \Theta^{(r)}) > Q(\Theta^{(r)}; \Theta^{(r)})$.
  1. Update the parameter p_j:

$$p_j^{(r+1)} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i} y_{ijt}^{(r)}}{\sum_{i=1}^{n}\sum_{t=1}^{T_i} n_{ij,t-1}}, \quad j = 1, ..., m. \tag{3.25}$$

  2. Update the parameter π_0:

$$\pi_0^{(r+1)} = 1 - \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i} w_{it}^{(r)}}{\sum_{i=1}^{n} T_i}. \tag{3.26}$$

  3. Update the parameter vector $\boldsymbol{\alpha}_j$ by implementing the Newton-Raphson method for one cycle:

$$\boldsymbol{\alpha}_j^{(r+1)} = \boldsymbol{\alpha}_j^{(r)} - \left[H(\boldsymbol{\alpha}_j^{(r)})\right]^{-1} s(\boldsymbol{\alpha}_j^{(r)}), \quad j = 1, ..., m, \tag{3.27}$$

  where the score equation and Hessian matrix are given as follows:

$$\begin{aligned} s(\boldsymbol{\alpha}_j^{(r)}) &= \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left[\tau_{ijt}^{(r)} - (1-w_{it}^{(r)})\pi_{ijt}^{(r)}\right]\boldsymbol{x}_{it}, \\ H(\boldsymbol{\alpha}_j^{(r)}) &= -\sum_{i=1}^{n}\sum_{t=1}^{T_i}(1-w_{it}^{(r)})\pi_{ijt}^{(r)}(1-\pi_{ijt}^{(r)})\boldsymbol{x}_{it}\boldsymbol{x}_{it}^{\top}. \end{aligned} \tag{3.28}$$

  4. Update the parameter vector $\boldsymbol{\alpha}_j$ by implementing the Newton-Raphson method for one cycle:

$$\boldsymbol{\beta}_j^{(r+1)} = \boldsymbol{\beta}_j^{(r)} - \left[H(\boldsymbol{\beta}_j^{(r)})\right]^{-1} s(\boldsymbol{\beta}_j^{(r)}), \quad j = 1, ..., m, \tag{3.29}$$

  where the score equation and Hessian matrix are given as follows:

$$\begin{aligned} s(\boldsymbol{\beta}_j^{(r)}) &= \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left[n_{ijt} - y_{ijt}^{(r)} - \tau_{ijt}^{(r)} - \tau_{ijt}^{(r)}\lambda_{ijt}^{(r)}\right]\boldsymbol{x}_{it}, \\ H(\boldsymbol{\beta}_j^{(r)}) &= -\sum_{i=1}^{n}\sum_{t=1}^{T_i}\tau_{ijt}^{(r)}\lambda_{ijt}^{(r)}\boldsymbol{x}_{it}\boldsymbol{x}_{it}^{\top}. \end{aligned} \tag{3.30}$$

4. SIMULATION STUDY

In this section, a simulation study is carried out. Firstly, we seek to validate the effectiveness of our proposed EM algorithms for the specialized models, namely, INAR-MZIP and INAR-MZIHP. Secondly, we aim to showcase the versatility of our models by applying them in a broader multivariate context, beyond just a bivariate scenario. In our study, the programming is implemented using the R language. The R codes for the implementation of the two models can be found at https://github.com/qingdaozpc/multivariate-zero-inflated-INAR-model.git.

4.1. Study Setup

We simulate a portfolio comprising $n = 2000$ policyholders and $m = 3$ claim types, each with an insured period $T_i = 5$. Two predictors are independently generated in the simulation: x_1 from a standard normal distribution and x_2 from a Bernoulli distribution with $p = 0.5$, and we assume they do not vary with time. The true values for the underlying parameters in the two models are presented in Table 4.1. For simplicity, covariates are not included in π_j, $j = 1, 2, 3$, when data are simulated from the INAR-MZIHP model.

4.2. Results

The estimation results are summarized in Table 4.1. In both cases, the results are derived from 100 replications. We provide the average estimates along with their standard errors. As anticipated, the mean estimates are very close to the true values, all of which fall within the 95% confidence intervals. This verifies the effectiveness of our proposed EM algorithms.

TABLE 1
Mean Estimates with Standard Errors from Simulations from the INAR-MZIP Model and the INAR-MZIHP Model

Parameter	Estimate	SE ($\times 10^{-3}$)
$p_1 = 0.1$	0.10	1.25
$p_2 = 0.2$	0.20	1.44
$p_3 = 0.3$	0.30	1.08
$\pi_0 = 0.5$	0.50	1.45
$\beta_{11} = -3$	−3.02	8.85
$\beta_{12} = -1$	−1.00	3.84
$\beta_{13} = 1$	1.01	9.11
$\beta_{21} = -2$	−2.00	6.42
$\beta_{22} = -1$	−1.00	3.74
$\beta_{23} = -1$	−1.00	7.75
$\beta_{31} = -1$	−0.99	4.94
$\beta_{32} = 1$	1.00	2.44
$\beta_{33} = -1$	−1.00	6.21

Parameter	Estimate	SE ($\times 10^{-3}$)
$p_1 = 0.1$	0.10	1.21
$p_2 = 0.2$	0.20	1.27
$p_3 = 0.3$	0.30	1.23
$\pi_0 = 0.5$	0.50	1.99
$\pi_1 = 0.3$	0.30	1.35
$\pi_2 = 0.2$	0.20	0.97
$\pi_3 = 0.1$	0.10	0.60
$\beta_{11} = -3$	−2.98	14.27
$\beta_{12} = -1$	−1.00	6.42
$\beta_{13} = 1$	0.98	13.62
$\beta_{21} = -2$	−2.00	13.03
$\beta_{22} = -1$	−1.00	7.68
$\beta_{23} = -1$	−1.02	19.48
$\beta_{31} = -1$	−1.00	10.41
$\beta_{32} = 1$	0.99	6.97
$\beta_{33} = -1$	−1.00	19.06

(a) INAR-MZIP model
(b) INAR-MZIHP model

5. APPLICATION

5.1. Data Description

The study is based on a dataset comprising automobile insurance policies from a major European insurance company during the underwriting years 2014 to 2019. This dataset includes bodily injury (BI) and property damage (PD) claims, denoted by N_1 and N_2, respectively, along with risk factors that impact both N_1 and N_2. An exploratory analysis was conducted to select the subset of covariates with the highest predictive power for N_1 and N_2. The description and empirical distribution of the selected explanatory variables are presented in Tables 5.2 and 5.3, respectively.

For our study, we randomly take 10,000 policyholders from the portfolio. The records in 2015 to 2018 are regarded as training data, and the records in 2019 are treated as test data. Therefore, we have 40,000 observations as training data to develop the models and 10,000 observations as a hold-out sample to evaluate model performance. The empirical joint distributions for claim numbers N_1 and N_2 across the 4 years (2015–2018) are displayed in Table 5.4. It is worth noting the presence of a multivariate zero inflation feature in the dataset, which confirms the validity of our proposed models.

5.2. Model Fitting

We begin by considering the scenario where no covariates are included. It is important to note that for the first type of claim, the claim number is either 0 or 1. Therefore, when utilizing relevant hurdle models such as MZIHP, INAR-MZIHP, and INAR-HP-copula, there is no need to model the positive part. In addition to our proposed INAR-MZIP and INAR-MZIHP models, we apply several typical alternatives associated with Poisson distributions as benchmark models for comparison

TABLE 2
Description of Explanatory Variables

Variable	Description
Car horsepower:	
$v1 = v2 = v3 = 0$	C1: 0–1299 cc
$v1 = 1, v2 = v3 = 0$	C2: 1300–1399 cc
$v2 = 1, v1 = v3 = 0$	C3: 1400–1599 cc
$v3 = 1, v1 = v2 = 0$	C4: $\geq$ 1600 cc
Policy type:	
$v4 = v5 = 0$	C1: Economic type, which includes only MTPL coverage.
$v4 = 1, v5 = 0$	C2: Middle type, which includes other types of coverage like legal protection etc.
$v5 = 1, v4 = 0$	C3: Expensive type—own coverage
Region:	
$v6 = v7 = 0$	C1: Capital city
$v6 = 1, v7 = 0$	C2: Provincial cities of mainland
$v7 = 1, v6 = 0$	C3: Provincial cities of island area
Vehicle age:	
$v8 = v9 = 0$	C1: 0–10 years
$v8 = 1, v9 = 0$	C2: 10–20 years
$v9 = 1, v8 = 0$	C3: >20 years

TABLE 3
Empirical Distribution of Explanatory Variables over the Years 2015–2018

	Car horsepower	Policy type	Region	Vehicle age
C1	9,648	2,024	19,139	14,199
C2	13,777	8,248	16,451	22,932
C3	10,323	29,728	4,410	2,869
C4	6,252	—	—	—

TABLE 4
Empirical Joint Distribution of N_1 and N_2 in Each Year from 2015 to 2018

	N_2			
N_1	0	1	2	3
2015				
0	9521	429	18	1
1	4	24	3	0
2016				
0	9484	470	23	0
1	10	13	0	0
2017				
0	9448	508	22	2
1	4	16	0	0
2018				
0	9467	483	26	2
1	2	19	1	0

purposes. The description of each model is outlined below. Further details on these benchmark models can be found in Appendix A.

- IP: Independent Poisson model.
- MZIP: Multivariate zero-inflated Poisson model; see (2.3).
- MZIHP: Multivariate zero-inflated hurdle Poisson model; see (2.6).
- Poi-copula: Gaussian copula to connect two Poisson marginals.
- INAR-Poi-copula: Gaussian copula to connect two INAR marginals, where the innovation parts are both modeled as Poisson distributions.
- INAR-ZIP-copula: Gaussian copula to connect two INAR marginals, where the innovation parts are both modeled as zero-inflated Poisson distributions.
- INAR-HP-copula: Gaussian copula to connect two INAR marginals, where the innovation parts are both modeled as hurdle Poisson distributions.
- INAR-BP: Bivariate INAR model with bivariate Poisson proposed in Bermúdez, Guillén, and Karlis (2018).
- INAR-PS: Bivariate INAR model with bivariate Poisson-Sarmanov proposed in Bermúdez and Karlis (2021).

The comparison results are presented in Table 5.5. It is evident that the IP model performs the worst among all models, indicating the need to introduce correlated effects between the two types of claims, in the form of either copula or multivariate random variables. Furthermore, the INAR models generally outperform those without autocorrelation counterparts (INAR-Poi-copula vs. Poi-copula, INAR-MZIP vs. MZIP, INAR-MZIHP vs. MZIHP), suggesting the importance of considering time correlation to fit the claim data. Overall, all of the information criteria support our proposed INAR-MZIHP model as the best performer.

We next turn to parameter estimation when covariates are introduced in our INAR-MZIHP model. The estimation results are displayed in Table 5.6. The 95% confidence interval of π_0 is $(0.063, 0.083)$, with the upper bound significantly below the boundary of 1. This indicates the presence of a multivariate zero inflation feature within the dataset. For the claims of N_1 type, p_1=0, suggesting that the number of claims in the previous year does not influence the number of claims in the current year. However, for the N_2 type, p_2=0.034, with the lower boundary of the 95% confidence interval exceeding 0, confirming the significance of the lag term. Moving on to the influence of covariates on π_1, π_2, and λ_2, it is observed that no predictor significantly affects the occurrence of N_1 type claims. As for the occurrence of the claims of N_2 type, $v6$, $v7$, and $v9$ are all statistically significant, indicating that driving in provincial cities on the mainland or island ($v6$ and $v7$) and vehicle age greater than 20 ($v9$) are all associated with decreased chances of a claim in this category. Additionally, conditional on the occurrence

TABLE 5
Information Criteria of Several Relevant Fitted Models

Model	Parameters	Log-likelihood	Akaike information criterion	Bayesian information criterion
IP	2	−9,221.82	18,447.64	18,464.84
MZIP	3	−9,141.52	18,289.03	18,314.82
MZIHP	4	−9,027.68	18,063.36	18,097.74
Poi-copula	3	−9,086.45	18,178.90	18,204.69
INAR-Poi-copula	5	−9,065.27	18,140.54	18,183.52
INAR-ZIP-copula	7	−9,055.38	18,124.76	18,184.93
INAR-HP-copula	6	−9,054.59	18,121.17	18,172.75
INAR-BP	5	−9,040.77	18,091.53	18,134.51
INAR-PS	5	−9,031.91	18,073.82	18,116.80
INAR-MZIP	5	−9,121.99	18,253.98	18,296.96
INAR-MZIHP	6	−9,006.56	18,025.13	18,076.71

TABLE 6
Estimates of the Full INAR-MZIHP Model

	π_1		π_2		λ_2	
	Estimate	t Ratio	Estimate	t Ratio	Estimate	t Ratio
Intercept	−2.885	−2.038*	1.383	2.248*	−3.269	−3.819***
$v1$	0.259	0.815	0.245	1.113	0.645	2.036*
$v2$	−0.186	−0.553	0.163	0.685	0.847	2.840**
$v3$	−0.224	−0.381	0.178	0.455	0.487	1.364
$v4$	0.038	0.026	0.132	0.278	−0.133	−0.128
$v5$	−0.286	−0.202	0.184	0.439	−0.627	−0.593
$v6$	−0.334	−1.312	−1.390	−5.015***	−0.002	−0.007
$v7$	−0.646	−1.410	−1.336	−4.503***	−0.588	−0.828
$v8$	−0.169	−0.644	−0.113	−0.679	0.289	1.027
$v9$	−0.985	−1.483	−0.818	−2.962**	−0.290	−0.471
	Estimate	95% Confidence interval				
π_0	0.073	(0.063, 0.083)				
p_1	0.000	(0.000, 0.000)				
p_2	0.034	(0.021, 0.047)				

Note: *$p < .05$. **$p < .01$. ***$p < .001$.

of N_2 type claims, $v1$ and $v2$ are positively associated with the positive number of claims. This suggests that higher horsepower correlates with more claims of this type.

5.3. Predictive Performance

In insurance claims modeling, we are concerned about the overall distribution of the portfolio, which can be used for premium calculation, risk management, and so forth. To evaluate the predictive performance, we then calculate the predicted claim frequencies (expected frequencies) by summing individual probabilities of joint events $(N_1, N_2) \in \{(0,0),(0,1),(0,2),(1,0),(1,1),(1,2)\}$ based on estimated parameters. These are then compared to observed frequencies from the test sample. In addition to our proposed INAR-MZIHP model, we try the INAR-BP and INAR-PS models as the benchmark for comparison, because these two models generally exhibit superior Akaike information criterion and Bayesian information criterion compared to other candidates, as seen from Table 5.5. To mitigate overfitting, a stepwise variable selection process is conducted, and only relevant covariates are retained in the model. The corresponding estimation results are

TABLE 7
Estimates of the Reduced INAR-MZIHP Model

	π_1		π_2		λ_2	
	Estimate	t Ratio	Estimate	t Ratio	Estimate	t Ratio
Intercept	−3.383	−27.379***	1.583	5.371***	−3.393	−17.054***
$v1$					0.390	1.472
$v2$					0.615	2.289*
$v6$			−1.333	−6.523***		
$v7$			−1.285	−5.031***		
$v9$			−0.716	−3.020**		
	Estimate	95% Confidence interval				
π_0	0.073	(0.065, 0.081)				
p_1	0.000	(0.000, 0.000)				
p_2	0.036	(0.023, 0.048)				

Note: $^*p < .05$. $p < ^{**}.01$. $^{***}p < .001$.

TABLE 8
Estimates of the Reduced INAR-BP Model

	π_1		π_2		λ_2	
	Estimate	t Ratio	Estimate	t Ratio	Estimate	t Ratio
Intercept	−7.542	−33.772***	−2.756	−92.034***	−6.282	−53.883***
$v6$			−0.453	−9.191***		
$v7$			−0.457	−5.604***		
$v9$			−0.344	−3.376***		
	Estimate	95% Confidence interval				
p_1	0.000	(0.000, 0.000)				
p_2	0.038	(0.025, 0.050)				

Note: $^*p < .05$. $p < ^{**}.01$. $^{***}p < .001$.

TABLE 9
Estimates of the Reduced INAR-PS Model

	λ_1		λ_2	
	Estimate	t Ratio	Estimate	t Ratio
Intercept	−6.032	−59.099***	−2.719	−92.530***
$v6$			−0.452	−9.337***
$v7$			−0.453	−5.663***
$v9$			−0.360	−3.570***
	Estimate	95% Confidence interval		
p_1	0.000	(0.000, 0.000)		
p_2	0.038	(0.025, 0.051)		
ω	24.541	(19.762, 29.320)		

Note: $^*p < .05$. $p < ^{**}.01$. $^{***}p < .001$.

presented in Tables 5.7, 5.8, and 5.9. These reduced models are utilized for prediction and ratemaking purposes. Furthermore, we incorporate results for the MZIHP model, given its second-best performance in fitting the training data, as evidenced by Table 5.5.

The predictive joint frequencies (N_1, N_2) generated by these models are summarized in Table 5.10. It is evident that both the MZIHP and INAR-MZIHP models exhibit superior predictive performance based on χ^2 statistics. Apart from the frequency comparison, the overall predictive ability of the models can be assessed by examining the log-likelihood values on the test data. As shown in Table 5.10, once again, the superiority of our proposed INAR-MZIHP model is apparent. This finding consistently aligns with the conclusions drawn from the model fitting part.

5.4. Application to Ratemaking

In this subsection, we analyze several fitted models for ratemaking. We have chosen three representative risk profiles under different models, labeled as Good, Average, and Bad. The three risk profiles under the four models are presented in Table 5.11.

We then calculate the means and variances of $N_{1t} + N_{2t}$ for three representative risk profiles under each model. The formulas for these computations are detailed in Appendix B. Tables 5.12 and 5.13 compare the means and variances of three profiles across the four models, given the claim counts from the previous year. Because $p_1 = 0$ in the INAR models, the means and variances in these models are unaffected by the number of claims of the first type. For the MZIHP model, claim history is irrelevant to the prediction of future claims.

TABLE 10
Predicted Frequencies and Log-Likelihood Values on Test Data under Each Model

(N_1, N_2)	Observed	INAR-MZIHP	MZIHP	INAR-BP	INAR-PS
(0, 0)	9420	9482.00	9484.73	9475.11	9469.14
(0, 1)	524	471.46	467.41	487.35	492.78
(0, 2)	30	21.97	23.23	13.32	13.82
(1, 0)	4	7.86	7.69	5.02	11.54
(1, 1)	21	15.38	15.52	17.96	11.94
(1, 2)	1	0.74	0.77	0.92	0.46
χ^2		13.24	13.04	24.69	33.61
Log-likelihood		−2, 440.023	−2, 446.275	−2, 442.309	−2, 447.954

TABLE 11
Three Different Risk Profiles under the Four Models

	$v1$	$v2$	$v6$	$v7$	$v9$
Good	0	0	1	0	1
Average	1	0	0	1	0
Bad	0	1	0	0	0

	$v6$	$v7$	$v9$
Good	0	1	1
Average	1	0	0
Bad	0	0	0

(a) INAR-MZIP and MZIHP models
(b) INAR-BP and INAR-PS models

TABLE 12
Premium Calculations from Different Models: Means

Profile	$(n_{1,t-1}, n_{2,t-1})$	INAR-MZIHP	MZIHP	INAR-BP	INAR-PS
Good	(0,0)	0.0315	0.0353	0.0329	0.0317
	(0,1)	0.0671	0.0353	0.0705	0.0693
	(1,0)	0.0315	0.0353	0.0329	0.0317
	(1,1)	0.0671	0.0353	0.0705	0.0694
Average	(0,0)	0.0464	0.0471	0.0447	0.0444
	(0,1)	0.0820	0.0471	0.0823	0.0820
	(1,0)	0.0464	0.0471	0.0447	0.0444
	(1,1)	0.0820	0.0471	0.0823	0.0821
Bad	(0,0)	0.0668	0.0686	0.0678	0.0683
	(0,1)	0.1024	0.0686	0.1054	0.1060
	(1,0)	0.0668	0.0686	0.0678	0.0683
	(1,1)	0.1024	0.0686	0.1054	0.1060

TABLE 13
Premium Calculations from Different Models: Variances

Profile	$(n_{1,t-1}, n_{2,t-1})$	INAR-MZIHP	MZIHP	INAR-BP	INAR-PS
Good	(0,0)	0.0335	0.0375	0.0366	0.0330
	(0,1)	0.0678	0.0375	0.0728	0.0693
	(1,0)	0.0335	0.0375	0.0366	0.0330
	(1,1)	0.0678	0.0375	0.0728	0.0693
Average	(0,0)	0.0501	0.0513	0.0484	0.0463
	(0,1)	0.0844	0.0513	0.0846	0.0825
	(1,0)	0.0501	0.0513	0.0484	0.0463
	(1,1)	0.0844	0.0513	0.0846	0.0826
Bad	(0,0)	0.0724	0.0744	0.0715	0.0713
	(0,1)	0.1067	0.0744	0.1077	0.1076
	(1,0)	0.0724	0.0744	0.0716	0.0713
	(1,1)	0.1067	0.0744	0.1077	0.1076

6. CONCLUDING REMARKS

This article introduces a flexible framework aimed at addressing time dependence and cross-dependence in multivariate insurance claim counts. We utilized the recently proposed multivariate zero-inflated Poisson and multivariate zero-inflated hurdle Poisson distributions to model the innovation term of the multivariate INAR(1) model. The resulting models effectively capture overdispersion and the high prevalence of zeros observed in real insurance data. The article distinguishes itself from previous works by employing the EM algorithm for parameter estimation.

In our numerical analysis, we utilized a dataset comprising automobile insurance policies from the underwriting years 2014 to 2019. Here, observations from 2015 to 2018 were designated as training data, and those from 2019 were used as test data to evaluate model performance. In addition to the proposed INAR-MZIP and INAR-MZIHP models, we applied several alternatives related to Poisson distributions as benchmark models for comparison. Generally, the INAR(1) models outperformed their counterparts without autocorrelation, highlighting the significance of considering time correlation when modeling claim counts data. Overall, all information criteria favored the superior performance of the INAR-MZIHP model. These models were further employed in a ratemaking scenario for pricing an automobile insurance contract.

In this study, our methodology effectively addresses the time dependence structure by assuming a diagonal matrix $\boldsymbol{P}$ in the model, thereby substantially reducing the correlation structure. However, exploring more intricate arrangements, such as a non-diagonal matrix $\boldsymbol{P}$, could provide insights into additional sources of dependence. Furthermore, our current time dependence

assumption is limited to claims counts reported in the preceding period. To mitigate this limitation, exploring higher-order models, such as MINAR(p) with $p > 1$, warrants consideration in future research endeavors.

ACKNOWLEDGMENTS

The authors express their gratitude to the editor and the two reviewers for their valuable comments and suggestions, which have significantly contributed to the improvement of the article.

DISCLOSURE STATEMENT

The authors report there are no competing interests to declare.

FUNDING

The research of Pengcheng Zhang was supported by the National Natural Science Foundation of China (No. 12301607), the Shandong Provincial Natural Science Foundation (No. ZR2023QA091, No. ZR2022MG027, No. ZR2022MG057), the Shandong Provincial Social Science Project Planning Research Project (No. 22CGLJ21, No. 22CJJJ29), and the Taishan Scholars Program of Shandong Province (No. tsqnz20230620).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

Bermúdez, L., M. Guillén, and D. Karlis. 2018. Allowing for time and cross dependence assumptions between claim counts in ratemaking models. *Insurance: Mathematics and Economics* 83:161–69. 10.1016/j.insmatheco.2018.06.003

Bermúdez, L., and D. Karlis. 2011. Bayesian multivariate Poisson models for insurance ratemaking. *Insurance: Mathematics and Economics* 48 (2):226–36. 10.1016/j.insmatheco.2010.11.001

Bermúdez, L., and D. Karlis. 2021. Multivariate INAR(1) regression models based on the Sarmanov distribution. *Mathematics* 9 (5):505. 10.3390/math9050505

Bolancé, C., and R. Vernic. 2019. Multivariate count data generalized linear models: Three approaches based on the Sarmanov distribution. *Insurance: Mathematics and Economics* 85:89–103. 10.1016/j.insmatheco.2019.01.001

Boucher, J.-P., M. Denuit, and M. Guillén. 2008. Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions. *Variance* 2 (1):135–62.

Boucher, J.-P., M. Denuit, and M. Guillen. 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):821–46. 10.1111/j.1539-6975.2009.01321.x

Fung, T. C., A. L. Badescu, and X. S. Lin. 2019. A class of mixture of experts models for general insurance: Application to correlated claim frequencies. *ASTIN Bulletin* 49 (3):647–88. 10.1017/asb.2019.25

Gourieroux, C., and J. Jasiak. 2004. Heterogeneous INAR (1) model with application to car insurance. *Insurance: Mathematics and Economics* 34 (2):177–92. 10.1016/j.insmatheco.2003.11.005

Li, C.-S., J.-C. Lu, J. Park, K. Kim, P. A. Brinkley, and J. P. Peterson. 1999. Multivariate zero-inflated Poisson models and their applications. *Technometrics* 41 (1):29–38. 10.1080/00401706.1999.10485593

Liu, Y., and G.-L. Tian. 2015. Type I multivariate zero-inflated Poisson distribution with applications. *Computational Statistics & Data Analysis* 83:200–222. 10.1016/j.csda.2014.10.010

Shi, P., and E. A. Valdez. 2014a. Longitudinal modeling of insurance claim counts using jitters. *Scandinavian Actuarial Journal* 2014 (2):159–79. 10.1080/03461238.2012.670611

Shi, P., and E. A. Valdez. 2014b. Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics* 55:18–29. 10.1016/j.insmatheco.2013.11.011

Tzougas, G., and A. Pignatelli di Cerchiara. 2021. The multivariate mixed negative binomial regression model with an application to insurance a posteriori ratemaking. *Insurance: Mathematics and Economics* 101:602–25. 10.1016/j.insmatheco.2021.10.001

Zhang, P., E. Calderín-Ojeda, S. Li, and X. Wu. 2023. Bayesian multivariate mixed Poisson models with copula-based mixture. *North American Actuarial Journal* 27 (3):560–78. 10.1080/10920277.2022.2112233

Zhang, P., D. Pitt, and X. Wu. 2022. A new multivariate zero-inflated hurdle model with applications in automobile insurance. *ASTIN Bulletin* 52 (2):393–416. 10.1017/asb.2021.39

APPENDIX A. DESCRIPTION FOR SOME BENCHMARK MODELS

A.1. Copula Model

We aim to compare our proposed INAR-MZIP and INAR-MZIHP models with some copula models. For our purpose, we use the Gaussian copula as an example. Instead of pairing the bivariate sequence (N_{1t}, N_{2t}) by the joint inflated probability π_0, we introduce correlation using the Gaussian copula to connect the two discrete marginals. The bivariate Gaussian Copula is defined as

$$C^{Gauss}(u_1, u_2) = \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2)),$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal distribution, $\Phi^{-1}(\cdot)$ is its quantile function, and $\Phi_\rho(\cdot,\cdot)$ is the cdf of a bivariate normal distribution with density function $\phi_\rho(\cdot,\cdot)$ defined as follows:

$$\phi_\rho(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{1-\rho^2}\right\}, \tag{A.1}$$

where ρ is the correlation parameter. Here, u_1 and u_2 are the cdf of discrete marginals N_{1t} and N_{2t}, respectively. We consider the following choices for marginals:

- Poison model (Poi-copula).
- INAR(1) model: N_{1t} and N_{2t} are both modeled as univariate INAR(1). The model is defined as

$$N_t = p \circ N_{t-1} + R_t, \tag{A.2}$$

  where R_t denotes the innovation part. The pmf of N_t is given by

$$f(n_t | n_{t-1}) = \sum_{k=0}^{\min(n_t, n_{t-1})} \binom{n_{t-1}}{k} p^k (1-p)^k f_R(n_t - k). \tag{A.3}$$

For the innovation part, three choices are considered:

- Poisson (INAR-Poi-copula)
- Zero-inflated Poisson (INAR-ZIP-copula). The pmf is given by

$$f_R(r) = \begin{cases} 1 - \pi_0 + \pi_0 e^{-\lambda}, & r = 0, \\ \pi_0 \dfrac{e^{-\lambda}\lambda^r}{r!}, & r > 0. \end{cases} \tag{A.4}$$

- Hurdle Poisson (INAR-HP-copula). The pmf is given by

$$f_R(r) = \begin{cases} 1 - \pi_0, & r = 0, \\ \pi_0 \dfrac{e^{-\lambda}\lambda^{r-1}}{(r-1)!}, & r > 0. \end{cases} \tag{A.5}$$

A.2. Bivariate INAR(1) Model

In addition to our proposed INAR-MZIP and INAR-MZIHP models, we try other bivariate INAR models with different innovation terms. We consider the following two choices for the innovation term:

- Bivariate Poisson (INAR-BP). This distribution, characterized by three parameters $(\lambda_0, \lambda_1, \lambda_2)$, is formulated as follows:

$$R_1 = Z_1 + Z_0, \quad R_2 = Z_2 + Z_0, \tag{A.6}$$

where each $Z_j,\ j = 0, 1, 2,$ independently follows a simple Poisson distribution with parameter λ_j. Its pmf is given by

$$f_{\boldsymbol{R}}(\boldsymbol{r}) = e^{-\lambda_0-\lambda_1-\lambda_2} \sum\nolimits_{j=0}^{\min(r_1, r_2)} \frac{\lambda_0^j}{j!} \frac{\lambda_1^{r_1-j}}{(r_1-j)!} \frac{\lambda_2^{r_2-j}}{(r_2-j)!}. \tag{A.7}$$

- Poisson-Sarmanov (INAR-PS). The pmf of this distribution, characterized by three parameters $(\omega, \lambda_1, \lambda_2)$, is given as follows:

$$f_{\boldsymbol{R}}(\boldsymbol{r}) = \frac{e^{-\lambda_1}\lambda_1^{r_1}}{r_1!} \frac{e^{-\lambda_2}\lambda_2^{r_2}}{r_2!} \left[1 + \omega(e^{-r_1} - e^{-\lambda_1 c})(e^{-r_2} - e^{-\lambda_2 c})\right], \tag{A.8}$$

where $c = 1 - e^{-1}$. To ensure the nonnegativity of this pmf, the following restrictions apply to ω :

$$-\min\left(\frac{1}{L_1 L_2}, \frac{1}{(L_1-1)(L_2-1)}\right) \le \omega \le \min\left(\frac{1}{(1-L_1)L_2}, \frac{1}{(1-L_2)L_1}\right), \tag{A.9}$$

where $L_j = e^{-\lambda_j c},\ j = 1, 2,$ is the value of the Laplace transform function of Poisson distribution evaluated at 1.

APPENDIX B. DISTRIBUTIONAL PROPERTIES FOR SOME INAR(1) MODELS

B.1. Two Multivariate Zero-Inflated INAR(1) Models

The conditional mean and variance of two multivariate zero-inflated INAR(1) models (INAR-MZIP and INAR-MZIHP) are

$$\begin{aligned} &\mathbb{E}(N_{jt}|N_{j,t-1} = n_{j,t-1}) = p_j n_{j,t-1} + \mathbb{E}(R_{jt}), \\ &\mathrm{Var}(N_{jt}|N_{j,t-1} = n_{j,t-1}) = p_j(1-p_j)n_{j,t-1} + \mathrm{Var}(R_{jt}), \end{aligned} \tag{B.1}$$

where $\mathbb{E}(R_{jt})$ is given by

$$\mathbb{E}(R_{jt}) = \begin{cases} \pi_0\lambda_j, & \text{MZIP}, \\ \pi_0\pi_j(\lambda_j + 1), & \text{MZIHP}, \end{cases} \tag{B.2}$$

and $\mathrm{Var}(R_{jt})$ is given by

$$\mathrm{Var}(R_{jt}) = \begin{cases} \pi_0\lambda_j + \pi_0(1-\pi_0)\lambda_j^2, & \text{MZIP}, \\ \pi_0\pi_j\lambda_j + \pi_0\pi_j(1-\pi_0\pi_j)(\lambda_j+1)^2, & \text{MZIHP}. \end{cases} \tag{B.3}$$

The covariance between N_{jt} and $N_{j't},\ j \neq j'$ is given by

$$\mathrm{Cov}(N_{jt}, N_{j't}) = \begin{cases} \pi_0(1-\pi_0)\lambda_j\lambda_{j'}, & \text{INAR-MZIP}, \\ \pi_0(1-\pi_0)\pi_j\pi_{j'}(\lambda_j+1)(\lambda_{j'}+1), & \text{INAR-MZIHP}. \end{cases} \tag{B.4}$$

B.2. Two Bivariate INAR(1) Models

The conditional mean and variance of two bivariate INAR(1) models (INAR-BP and INAR-PS) are

$$\begin{aligned} &\mathbb{E}(N_{jt}|N_{j,t-1} = n_{j,t-1}) = p_j n_{j,t-1} + \mathbb{E}(R_{jt}), \\ &\mathrm{Var}(N_{jt}|N_{j,t-1} = n_{j,t-1}) = p_j(1-p_j)n_{j,t-1} + \mathrm{Var}(R_{jt}), \end{aligned} \tag{B.5}$$

where $\mathbb{E}(R_{jt})$ is given by

$$\mathbb{E}(R_{jt}) = \begin{cases} \lambda_j + \lambda_0, & \text{BP}, \\ \lambda_j, & \text{PS}, \end{cases} \tag{B.6}$$

and $\text{Var}(R_{jt})$ is given by

$$\text{Var}(R_{jt}) = \begin{cases} \lambda_j + \lambda_0, & \text{BP}, \\ \lambda_j, & \text{PS}. \end{cases} \tag{B.7}$$

The covariance between N_{1t} and N_{2t} is given by

$$\text{Cov}(N_{1t}, N_{2t}) = \begin{cases} \lambda_0, & \text{BP}, \\ \omega\lambda_1\lambda_2 c^2 e^{-(\lambda_1+\lambda_2)c}, & \text{PS}. \end{cases} \tag{B.8}$$