Insurance Mathematics and Economics 125 (2025) 103136

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Bayesian CART models for aggregate claim modeling

Yaojun Zhang [a], Lanpeng Ji [b,*], Georgios Aivaliotis [b], Charles C. Taylor [b]

[a] School of Statistics, Southwestern University of Finance and Economics, China
[b] Department of Statistics, University of Leeds, United Kingdom of Great Britain and Northern Ireland

A R T I C L E I N F O

JEL classification:
C11
C14
C51
G22

Keywords:
Average severity
Dependence
DIC
Information sharing
Zero-inflated compound Poisson gamma distribution

A B S T R A C T

This paper proposes three types of Bayesian CART (or BCART) models for aggregate claim amount, namely, frequency-severity models, sequential models and joint models. We propose a general framework for BCART models applicable to data with multivariate responses, which is particularly useful for the joint BCART models with a bivariate response: the number of claims and the aggregate claim amount. To facilitate frequency-severity modeling, we investigate BCART models for the right-skewed and heavy-tailed claim severity data using various distributions. We discover that the Weibull distribution is superior to gamma and lognormal distributions, due to its ability to capture different tail characteristics in tree models. Additionally, we find that sequential BCART models and joint BCART models, which can incorporate more complex dependence between the number of claims and severity, are beneficial and thus preferable to the frequency-severity BCART models in which independence is commonly assumed. The effectiveness of these models' performance is illustrated by carefully designed simulations and real insurance data.

1. Introduction

Aggregate claim amount estimation in non-life insurance is an important task of actuaries for the calculation of premiums, pricing of insurance contracts and management of risk. A common model for aggregate claim amount is the so-called *collective risk model* defined as

$$S = \begin{cases} \sum_{i=1}^{N} Y_i, & \text{if } N > 0 \\ 0, & \text{if } N = 0 \end{cases}, \tag{1}$$

where N is a nonnegative integer valued random variable for the *number of claims* and $\{Y_i\}_{i=1}^{\infty}$ are positive random variables for the *individual severity*. Note that the above aggregate claim amount can also be expressed as

$$S = N\bar{S}, \quad \text{with} \quad \bar{S} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} Y_i, & \text{if } N > 0 \\ 0, & \text{if } N = 0 \end{cases}, \tag{2}$$

where $\bar{S}$ is the so-called *average severity*. Following the convention in Oh et al. (2021), we call $(N, Y_1, \dots, Y_N)$ micro-level data and $(N, \bar{S})$ (or (N, S)) summarized data.

The classical theory in actuarial science is based on the full independence assumption between the number of claims and the individual severities, i.e., N is independent of $\{Y_i\}_{i=1}^{\infty}$ which is a sequence of independent and identically distributed (IID) random variables. In a regression setting given risk factors (or covariates), the *frequency-severity models* treat these two components separately by using generalized linear models (GLMs), assuming distributions from the exponential family; see e.g., Ohlsson and Johansson (2010). The frequency study focuses on the occurrences of claims, and the severity study — provided that a claim has occurred — investigates the claim amount. In the current literature, the severity component has been discussed from two perspectives: by modeling individual severity Y_i or by modeling the average severity $\bar{S}$, depending on the format of data available (micro-level or summarized). Under the full independence assumptions, the expected aggregate claim (or *pure premium*) can be determined by multiplying the estimated frequency with the estimated severity or conditional average severity.

In recent years, a growing body of literature emphasizes the importance of understanding the interrelated nature of claim occurrences and their associated claim amounts to improve model applicability. Since the scope of the current paper is on the regression setting, hereafter we review some of the relevant literature. We refer to e.g., Cossette et al. (2019); Blier-Wong et al. (2024) for recent developments on this topic under a non-regression setting where different aspects of the distribu-

* Corresponding author.
E-mail addresses: zyaojun@swufe.edu.cn (Y. Zhang), l.ji@leeds.ac.uk (L. Ji), g.aivaliotis@leeds.ac.uk (G. Aivaliotis), c.c.taylor@leeds.ac.uk (C.C. Taylor).

https://doi.org/10.1016/j.insmatheco.2025.103136
Received 3 September 2024; Received in revised form 14 April 2025; Accepted 23 July 2025

Available online 30 July 2025

tion of the aggregate claim S are investigated. Research has shown that ignoring the dependence between number of claims and claim severities may lead to serious bias in inference and thus the evaluation of risks; see Oh et al. (2021); Shi and Zhao (2020) and references therein. There are mainly three strategies to address the dependence between (i) the number of claims and the individual severities (including dependence among individual severities) and (ii) the number of claims and the average severity; namely, using a coupla, using shared random effects and using a two-part model where the severity component directly depends on the frequency.

Copula models for both micro-level data and summarized data have been discussed; see, e.g., Ahn et al. (2021); Shi et al. (2015); Czado et al. (2012); Krämer et al. (2013); Shi and Zhao (2020); Oh et al. (2021). Estimation of parameters in these models is usually obtained using a likelihood-based method which requires a tractable formulation of the likelihood. Working with micro-level data, (Oh et al., 2021) models the full dependence of number of claims and individual severities using multivariate Gaussian and t-copula functions which is also extended to include a vine copula. In Shi and Zhao (2020), the authors assume conditional independence of the individual severities given the number of claims so that they only require a single bivariate copula to construct the likelihood function, by doing so a general class of copulas can be adopted and incomplete data due to censoring or truncation can also be accounted for. Working with summarized data, Czado et al. (2012); Krämer et al. (2013); Shi et al. (2015) analyze different copula models for the number of claims and average severity, where the challenge persists in selecting the appropriate copula family. As noticed in Shi et al. (2015); Shi and Zhao (2020), for regression copula models, the association between the frequency and severity is introduced by both the covariates and the copula used; the dependence introduced by copula (interpreted as *residual dependence*) should be seen as an extra layer of association in addition to that introduced by the covariates. It has been noted that in all these copula models the dependence parameter in the copula is assumed to be constant and does not vary across covariates (i.e., *portfolio-level dependence*). To overcome this drawback, a regression approach using shared random effects is introduced in Baumgartner et al. (2015) where the dependence between the number of claims and average severity is induced by shared random effects. By analyzing a German car insurance portfolio, they show that the proposed shared random effects model can reflect the varying dependence characteristics across different geographical regions.

The two-part models enable the severity component to depend on the frequency component more explicitly. Specifically, the number of claims is introduced as a covariate in the (average) severity modeling; see, e.g., Frees et al. (2011); Garrido et al. (2016); Gschlößl and Czado (2007); Shi et al. (2015). It is noted that when dealing with the average severity, Garrido et al. (2016); Gschlößl and Czado (2007) include the number of claims as weights as well as covariates assuming the individual severities are conditionally IID and from the exponential family. Whereas, Shi et al. (2015) imposes a generalized gamma distribution for the average severity where the number of claims is only included as a covariate. Under the classical Poisson-gamma risk model with full independence, there is an intrinsic dependence between the number of claims and the average severity, which can be difficultly modeled by any copula as illustrated in Oh et al. (2021) using a simulation example. This difficulty may suggest a preference for simple two-part models for summarized data. It is also noticed that in these two-part models the dependence introduced by the coefficient of the number of claims for the severity component can only interpret a portfolio-level dependence which has the same issue as pointed out above for the current copula models.

As a comparable alternative, the aggregate claim amount can be directly modeled using Tweedie's distribution which assumes a Poisson sum of gamma variables for the aggregate claim amount. This modeling approach simplifies the analysis by accommodating discrete claim numbers and continuous claim amounts in one distribution; see, e.g., Jørgensen and Paes De Souza (1994). Concurrently, discussions regarding the suitability of GLMs for aggregate claim amount analysis have focused on the trade-off between model complexity and predictive performance, emphasizing the benefits and contexts where Tweedie's model excels and where alternative methodologies may be better; see, e.g., Quijano Xacur and Garrido (2015); Delong et al. (2021). Particularly, Delong et al. (2021) has investigated two parametrization approaches and provided theoretical evidence supporting the industry preference for the Poisson-gamma parametrization over the Tweedie's compound Poisson parametrization. We refer to Zhou et al. (2022); Gao (2024); Lian et al. (2023) for recent developments of Tweedie's model.

The above literature review seems to indicate that almost all the models rely on likelihood-based approaches for parameter estimation, assuming a linear additive structure when incorporating covariates. This makes it difficult for these models to achieve variable selection and interaction detection automatically. Moreover, the above discussions show that there is no simple answer to the question on the association/dependence between number of claims and severities, because it can be introduced by the covariates or model dependence or a combination of them.

More recently, machine learning methods have been introduced in the context of insurance by adopting actuarial loss distributions to capture the characteristics of insurance claims. We refer to Blier-Wong et al. (2020); Denuit and Trufin (2019); Wuthrich and Merz (2022); Wuthrich and Buser (2022) for recent discussions. Insurance pricing models are heavily regulated and must meet specific requirements before being deployed in practice, which poses challenges for most machine learning methods. As discussed in Henckaerts et al. (2021); Zhang et al. (2024), tree-based models are considered appropriate for insurance rate-making due to their transparent nature. Furthermore, tree-based models are known to be able to capture nonlinearities and complex higher order interactions among covariates, and automatically implement variable selection. In our previous work (Zhang et al., 2024), we have demonstrated the superiority of Bayesian classification and regression tree (BCART) models in claim frequency analysis. In this sequel, working with summarized data $(N, \bar{S})$ (or (N, S)), we construct some novel insurance pricing models using BCART for both average severity and aggregate claim amount.

Specifically, inspired by the claim loss models discussed in the literature, we introduce and investigate three types of BCART models. We first discuss a benchmark frequency-severity BCART model, where the number of claims and average severity are modeled separately using BCART. Furthermore, we propose two other types of BCART models, with an aim to better incorporate the underlying complex association between the number of claims and average severity. These are sequential BCART models (motivated by Frees et al. (2011); Garrido et al. (2016); Gschlößl and Czado (2007)) and joint BCART models (motivated by Jørgensen and Paes De Souza (1994); Smyth and Jørgensen (2002); Delong et al. (2021)). In contrast to the frequency-severity and sequential BCART models which result in two separate trees for the number of claims and average severity, the joint BCART models generate one joint tree for the bivariate response (N, S) which suggests a joint effect of covariates to both parts. In the real data analysis below, we observe that different cells of the resulting tree partition can exhibit varying (positive or negative) conditional dependencies between the number of claims and average severity, overcoming the drawback of most existing models which are only able to capture a unique portfolio-level dependence; see Remark 6.

The main contributions of this paper are as follows:

- We implement BCART models for average severity including gamma, lognormal and Weibull distributions and for aggregate claim amount including compound Poisson gamma (CPG) and zero-inflated compound Poisson gamma (ZICPG) distributions. These are not currently available in any R package.
- To explore the complex association between the number of claims and average severity, we propose novel sequential BCART models

that treat the number of claims (or its estimate) as a covariate in average severity modeling. For these models, the aggregate claim amount cannot be obtained analytically due to the assumed dependence, for which a Monte Carlo method will be used. The effectiveness is illustrated using simulated and real insurance data.

- We present a general framework for the BCART models applicable for multivariate responses, extending the MCMC algorithms discussed in Zhang et al. (2024). To the best of our knowledge, there have been very few discussions on Bayesian tree models with multivariate responses in the current literature, with the only exception (Um et al., 2023). As a particular application, we propose novel joint BCART models with a bivariate response to simultaneously model the number of claims and aggregate claim amount. In doing so, we employ the commonly used distributions such as CPG and ZICPG. The potential advantages of information sharing using one joint tree compared with two separate trees are also illustrated by simulated and real insurance data.
- For the comparison of one joint tree (generated from the joint BCART models) with two separate trees (generated from the frequency-severity or sequential BCART models), we propose some evaluation metrics which involve a combination of trees using an idea of Rocková et al. (2020). We also propose an application of the adjusted Rand Index (ARI) in assessing the similarity between trees. Although ARI is widely used in cluster analysis, its application to tree comparisons seems to be a novel idea. The use of ARI enhances the understanding of the necessity of information sharing, an aspect not covered in relevant literature; see, e.g., Linero et al. (2020).

Outline of the rest of the paper: In Section 2, we briefly review the BCART framework, introducing a more general MCMC algorithm for BCART models with multivariate responses. Section 3 introduces the notation for insurance claim data and investigates three types of BCART models for the aggregate claim amount. Section 4 develops a performance assessment of the proposed aggregate claim models using simulation examples. In Section 5, we present a detailed analysis of real insurance data using the proposed models. Section 6 concludes the paper. Some of the technical details are included in an online Supplementary Material.

2. Bayesian CART: a general framework

The BCART models, as introduced in the seminal papers (Chipman et al., 1998; Denison et al., 1998), provide a Bayesian perspective on CART models. In this section, we give a brief review of the BCART model using a more general framework that applies to multivariate response data; see Zhang et al. (2024) for the univariate case.

2.1. Data, model and training algorithm

Consider a matrix-form dataset $(\boldsymbol{X},\boldsymbol{Y}) = \left((\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2),\dots,(\boldsymbol{x}_n,\boldsymbol{y}_n)\right)^\top$ with n independent observations. For the i-th observation, $\boldsymbol{x}_i=(x_{i1},x_{i2},\dots,x_{ip})$ is a vector of p explanatory variables (or covariates) sampled from a space $\mathcal{X}$, while $\boldsymbol{y}_i=(y_{i1},y_{i2},\dots,y_{iq})$ is a vector of q response variables sampled from a space $\mathcal{Y}$. For the severity (or frequency) modeling, $\mathcal{Y}$ is a space of real positive (or integer) values. For aggregate claim modeling, $\mathcal{Y}$ is a space of 2-dimensional vectors with two components: an integer number of claims and a real valued aggregate claim amount.

A CART has two main components: a binary tree $\mathcal{T}$ with b terminal nodes which induces a partition of the covariate space $\mathcal{X}$, denoted by $\{\mathcal{A}_1,\dots,\mathcal{A}_b\}$, and a parameter $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\dots,\boldsymbol{\theta}_b)$ which associates the parameter value $\boldsymbol{\theta}_t$ with the t-th terminal node. Note that here we do not specify the dimension and range of the parameter $\boldsymbol{\theta}_t$ which should be clear from the context. If $\boldsymbol{x}_i$ is located in the t-th terminal node (i.e., $\boldsymbol{x}_i\in\mathcal{A}_t$), then $\boldsymbol{y}_i$ has a (joint) distribution $f\left(\boldsymbol{y}_i\mid\boldsymbol{\theta}_t\right)$, where f represents a parametric family indexed by $\boldsymbol{\theta}_t$. By associating observations with the b terminal nodes in the tree $\mathcal{T}$, we can re-order the n observations such that

$$(\boldsymbol{X},\boldsymbol{Y}) = \left((\boldsymbol{X}_1,\boldsymbol{Y}_1),(\boldsymbol{X}_2,\boldsymbol{Y}_2)\dots,(\boldsymbol{X}_b,\boldsymbol{Y}_b)\right)^\top,$$

where $\boldsymbol{Y}_t=\left(\boldsymbol{y}_{t1},\dots\boldsymbol{y}_{tn_t}\right)^\top$ is an $n_t\times q$ matrix with n_t denoting the number of observations and $\boldsymbol{y}_{ti}$ denoting the i-th observed response in the t-th terminal node, and $\boldsymbol{X}_t$ is an analogously defined $n_t\times p$ design matrix. We make the typical assumption that conditionally on $(\boldsymbol{\theta},\mathcal{T})$, response variables are independent and identically distributed (IID). The CART model likelihood is then

$$p(\boldsymbol{Y}\mid\boldsymbol{X},\boldsymbol{\theta},\mathcal{T})=\prod_{t=1}^{b}f\left(\boldsymbol{Y}_t\mid\boldsymbol{\theta}_t\right)=\prod_{t=1}^{b}\prod_{i=1}^{n_t}f\left(\boldsymbol{y}_{ti}\mid\boldsymbol{\theta}_t\right). \tag{3}$$

Given $(\boldsymbol{\theta},\mathcal{T})$, a Bayesian analysis involves specifying a prior distribution $p(\boldsymbol{\theta},\mathcal{T})$, and inference about $\boldsymbol{\theta}$ and $\mathcal{T}$ is based on the joint posterior $p(\boldsymbol{\theta},\mathcal{T}|\boldsymbol{Y},\boldsymbol{X})$ using a suitable MCMC algorithm. Since $\boldsymbol{\theta}$ indexes the parametric model whose dimension depends on the number of terminal nodes of the tree, it is usually convenient to apply the relationship $p(\boldsymbol{\theta},\mathcal{T})=p(\boldsymbol{\theta}\mid\mathcal{T})p(\mathcal{T})$, and specify the tree prior distribution $p(\mathcal{T})$ and the terminal node parameter prior distribution $p(\boldsymbol{\theta}\mid\mathcal{T})$, respectively.

The prior distribution $p(\mathcal{T})$ has two components: a tree topology and a decision rule for each of the internal nodes. We follow (Chipman et al., 1998), in which a draw of the tree is obtained by generating, for each node at depth d (with $d=0$ for the root node), two child nodes with probability

$$p(d)=\gamma(1+d)^{-\rho}, \tag{4}$$

where $\gamma\in(0,1],\rho\geq 0$ are parameters controlling the structure and size of the tree. This process iterates for $d=0,1,\dots$ until we reach a depth at which all the nodes cease growing. After the tree topology is generated, each internal node is associated with a decision rule which will be drawn uniformly among all the possible decision rules for that node. We refer to Zhang et al. (2024) for detailed discussion on the choice of the prior distribution $p(\mathcal{T})$. It is important to choose the form $p(\boldsymbol{\theta}\mid\mathcal{T})$ for which it is possible to analytically margin out $\boldsymbol{\theta}$ to obtain the integrated likelihood

$$\begin{aligned}p(\boldsymbol{Y}\mid\boldsymbol{X},\mathcal{T})&=\int p(\boldsymbol{Y}\mid\boldsymbol{X},\boldsymbol{\theta},\mathcal{T})p(\boldsymbol{\theta}\mid\mathcal{T})d\boldsymbol{\theta}=\prod_{t=1}^{b}\int f\left(\boldsymbol{Y}_t\mid\boldsymbol{\theta}_t\right)p(\boldsymbol{\theta}_t)d\boldsymbol{\theta}_t\\&=\prod_{t=1}^{b}\int\prod_{i=1}^{n_t}f\left(\boldsymbol{y}_{ti}\mid\boldsymbol{\theta}_t\right)p(\boldsymbol{\theta}_t)d\boldsymbol{\theta}_t,\end{aligned} \tag{5}$$

where in the second equality we assume that conditional on the tree $\mathcal{T}$ with b terminal nodes as above, the parameters $\boldsymbol{\theta}_t, t=1,2,\dots,b$, have IID priors $p(\boldsymbol{\theta}_t)$, which is a common assumption. Examples where this integration has a closed-form expression can be found in, e.g., Chipman et al. (1998); Linero (2017).

When there is no obvious prior distribution $p(\boldsymbol{\theta}_t)$ such that the integration in (5) is of closed-form, particularly, for non-Gaussian distributed data $\boldsymbol{Y}$, a data augmentation method is usually utilized in the literature, e.g., Murray (2021); Zhang et al. (2024). Here, we present a general framework in which apart from including a data augmentation, some components of $\boldsymbol{\theta}_t$ are assumed to be known a priori, but some others are assumed to be unknown. More precisely, we assume $\boldsymbol{\theta}_t=(\boldsymbol{\theta}_{t,M},\boldsymbol{\theta}_{t,B})$, where $\boldsymbol{\theta}_{t,M}$ are the parameters that are treated as known and computed using Method of Moments Estimation (MME), or Maximum Likelihood Estimation (MLE), and $\boldsymbol{\theta}_{t,B}$ are the unknown parameters that need to be estimated in the Bayesian framework. This newly proposed framework aims to reduce the overall computational time of the algorithm and overcome the difficulty of finding an appropriate prior for some parameters even with the data augmentation (that is why $\boldsymbol{\theta}_{t,M}$ is assumed known a priori). Under this framework, we augment the data $\boldsymbol{Y}$ by introducing a latent variable $\boldsymbol{Z}=(\boldsymbol{z}_1,\boldsymbol{z}_2,\dots,\boldsymbol{z}_n)^\top$ so that the integration in (6) is computable for the augmented data $(\boldsymbol{Y}, \boldsymbol{Z})$. The integrated likelihood is given as

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}_M, \mathcal{T}) = \int p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}_M, \mathcal{T}) d\boldsymbol{Z},$$

where

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}_M, \mathcal{T}) &= \int p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}_M, \boldsymbol{\theta}_B, \mathcal{T}) p(\boldsymbol{\theta}_B \mid \mathcal{T}) d\boldsymbol{\theta}_B \\ &= \prod_{t=1}^{b} \int \prod_{i=1}^{n_t} f\left(\boldsymbol{y}_{ti}, \boldsymbol{z}_{ti} \mid \boldsymbol{\theta}_{t,M}, \boldsymbol{\theta}_{t,B}\right) p(\boldsymbol{\theta}_{t,B}) d\boldsymbol{\theta}_{t,B}, \end{aligned} \tag{6}$$

with $\boldsymbol{Z}_t = (\boldsymbol{z}_{t1}, \boldsymbol{z}_{t2}, \ldots, \boldsymbol{z}_{tn_t})^\top$ defined according to the partition of $\mathcal{X}$.

Combining the augmented integrated likelihood $p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}_M, \mathcal{T})$ with tree prior $p(\mathcal{T})$, allows us to calculate the posterior of $\mathcal{T}$

$$p(\mathcal{T} \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}_M, \boldsymbol{Z}) \propto p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}_M, \mathcal{T}) p(\mathcal{T}). \tag{7}$$

When using MCMC to conduct Bayesian inference, $\mathcal{T}$ can be updated using a Metropolis-Hastings (MH) algorithm with the right-hand side of (7) used to compute the acceptance ratio. Starting from the root node, the MCMC algorithm for simulating a Markov chain sequence of pairs $(\boldsymbol{\theta}^{(1)}, \mathcal{T}^{(1)}), (\boldsymbol{\theta}^{(2)}, \mathcal{T}^{(2)}), \ldots$, using the posterior given in (7), is given in Algorithm 1 in which commonly used proposals (or transitions) for $q(\cdot,\cdot)$ include grow, prune, change and swap (see Chipman et al. (1998)). See Zhang et al. (2024) for further details.

Algorithm 1 One step of the MCMC algorithm for the BCART models parameterized by $(\boldsymbol{\theta}_M, \boldsymbol{\theta}_B, \mathcal{T})$ using data augmentation with both known and unknown parameters.

Input: Data $(\boldsymbol{X}, \boldsymbol{Y})$ and current values $(\hat{\boldsymbol{\theta}}_M^{(m)}, \boldsymbol{\theta}_B^{(m)}, \boldsymbol{Z}^{(m)}, \mathcal{T}^{(m)})$

1: Generate a candidate value $\mathcal{T}^*$ with probability distribution $q(\mathcal{T}^{(m)}, \mathcal{T}^*)$

2: Estimate $\hat{\boldsymbol{\theta}}_M^{(m+1)}$, using MME (or MLE)

3: Sample $\boldsymbol{Z}^{(m+1)} \sim p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{Y}, \hat{\boldsymbol{\theta}}_M^{(m+1)}, \boldsymbol{\theta}_B^{(m)}, \mathcal{T}^{(m)})$

4: Set the acceptance ratio

$$\alpha(\mathcal{T}^{(m)}, \mathcal{T}^*) = \min\left\{ \frac{q(\mathcal{T}^*, \mathcal{T}^{(m)}) p(\boldsymbol{Y}, \boldsymbol{Z}^{(m+1)} \mid \boldsymbol{X}, \hat{\boldsymbol{\theta}}_M^{(m+1)}, \mathcal{T}^*) p(\mathcal{T}^*)}{q(\mathcal{T}^{(m)}, \mathcal{T}^*) p(\boldsymbol{Y}, \boldsymbol{Z}^{(m)} \mid \boldsymbol{X}, \hat{\boldsymbol{\theta}}_M^{(m)}, \mathcal{T}^{(m)}) p(\mathcal{T}^{(m)})}, 1 \right\}$$

5: Update $\mathcal{T}^{(m+1)} = \mathcal{T}^*$ with probability $\alpha(\mathcal{T}^{(m)}, \mathcal{T}^*)$, otherwise, set $\mathcal{T}^{(m+1)} = \mathcal{T}^{(m)}$

6: Sample $\boldsymbol{\theta}_B^{(m+1)} \sim p(\boldsymbol{\theta}_B \mid \boldsymbol{X}, \boldsymbol{Y}, \hat{\boldsymbol{\theta}}_M^{(m+1)}, \boldsymbol{Z}^{(m+1)}, \mathcal{T}^{(m+1)})$

Output: New values $(\hat{\boldsymbol{\theta}}_M^{(m+1)}, \boldsymbol{\theta}_B^{(m+1)}, \boldsymbol{Z}^{(m+1)}, \mathcal{T}^{(m+1)})$

Remark 1.

(a). In Algorithm 1, the sampling steps should be done only as required. For example, in Step 2, $\hat{\boldsymbol{\theta}}_M^{(m+1)}$ needs to be estimated only for those nodes that were involved in the proposed move from $\mathcal{T}^{(m)}$ to $\mathcal{T}^*$.

(b). Algorithm 1 is a general algorithm from which we can retrieve all the algorithms discussed in Zhang et al. (2024), e.g., the algorithm for the zero-inflated Poisson model therein can be retrieved by assuming there is no component $\boldsymbol{\theta}_M$.

(c). The proposed parameter separation framework $\boldsymbol{\theta}_t = (\boldsymbol{\theta}_{t,M}, \boldsymbol{\theta}_{t,B})$ is designed for simplicity. While general non-conjugate priors could offer greater flexibility in modeling complex data, they would require computationally intensive methods in approximating the posteriors, which significantly increases computational cost. Exploring non-conjugate priors, using, e.g., Laplace approximation as in Chipman et al. (2003), may be a direction for future research.

2.2. Model selection and prediction

The MCMC algorithm described in Algorithm 1 can be used to search for desirable trees, and we use the three-step approach proposed in Zhang et al. (2024) based on deviance information criterion (DIC) to select an "optimal" tree among those visited trees; see Table 1. Note that in the following sections, we introduce the DIC for different models based on the idea that DIC="goodness of fit"+"complexity". See Spiegelhalter et al. (2002); Celeux et al. (2006) for discussion on DIC in a general Bayesian framework.

Suppose $\mathcal{T}$, with b terminal nodes and parameter $\overline{\boldsymbol{\theta}}$, is the selected tree from the above approach. For new $\boldsymbol{x}$ the predicted $\hat{\boldsymbol{y}}$ is defined as

$$\hat{\boldsymbol{y}} \mid \boldsymbol{x} = \sum_{t=1}^{b} E(\boldsymbol{y} \mid \bar{\boldsymbol{\theta}}_t) I_{(\boldsymbol{x} \in \mathcal{A}_t)}, \tag{8}$$

where $I_{(\cdot)}$ denotes the indicator function and $\{\mathcal{A}_t\}_{t=1}^b$ is the partition of $\mathcal{X}$ by $\mathcal{T}$.

3. Aggregate claim amount modeling with Bayesian CART

This section introduces the BCART models for aggregate claim amount by specifying the response distribution within the framework outlined in Section 2. We begin by introducing the type of insurance claim data that will be discussed in this paper. A claim dataset with n policyholders can be described by $(\boldsymbol{X}, \boldsymbol{v}, \boldsymbol{N}, \boldsymbol{S}) = \left((\boldsymbol{x}_1, v_1, N_1, S_1), \ldots, (\boldsymbol{x}_n, v_n, N_n, S_n)\right)^\top$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip}) \in \mathcal{X}$ represents rating variables (e.g., driver age, age of the car and car brand in car insurance); $v_i \in (0, 1]$ is the exposure in yearly units, quantifying the duration the policyholder i is exposed to risk; N_i is the number of claims reported during exposure time of the policyholder, and S_i is the aggregate (total) claim amount. Following the convention in Oh et al. (2021), we have a summarized dataset as the individual severities are not accessible.

Before describing our BCART models, we briefly recall some basics on the aggregate claim amount (1)-(2); see, e.g., Wuthrich (2022); Garrido et al. (2016); Frees et al. (2016) for discussions. Consider a given (generic) policyholder, and assume unit exposure (i.e., $v = 1$), for simplicity. We are primarily interested in estimating the pure premium defined as $\mathbb{E}(S)$ (we remark that generally the pure premium should be defined as $\mathbb{E}(S)/v$, i.e., the expected claim amount per year). Under the classical collective risk model with full independence between N and $\{Y_i\}_{i\geq 1}$, we have

$$\mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(\bar{S}|N > 0). \tag{9}$$

When a vector $\boldsymbol{x} = (x_1, \ldots, x_p)$ of covariates for this policyholder is available, it can be incorporated into separate GLMs for frequency N and (conditional) average severity $\bar{S}|N > 0$, this is the classical frequency-severity model. In particular, a distribution from exponential distribution family (EDF) with N as a weight is used for the average severity $\bar{S}|N > 0$; see e.g., Garrido et al. (2016); Gschlößl and Czado (2007). More specifically, assuming $Y_j \sim \text{EDF}(\mu, \phi)$, with mean μ and dispersion ϕ, in (1), we have, $\bar{S}|N = n \sim \text{EDF}(\mu, \phi/n)$ due to the convolution property, which means that modeling individual severity is equivalent to modeling the average severity where N is included as a weight. A common distribution for Y_j is the gamma distribution. This approach introduces an intrinsic functional dependence of the distribution of average severity on N and can be used for summarized data. The pure premium can then be calculated by multiplying the estimations for the two parts by (9). Another approach when dealing with summarized data $(N, \bar{S})$ is to use only the first formulation of (2), imposing a distribution (independent of N) directly for the average severity $\bar{S}|N > 0$; see e.g., Czado et al. (2012); Baumgartner et al. (2015); Shi et al. (2015); Krämer et al. (2013). In Baumgartner et al. (2015) a shared random effects model is used to induce a conditional independence between the number of claims and average severity. In Shi et al. (2015); Czado et al. (2012) copulas are used to model the dependence between the two parts, and the performance of these models is compared with the independence case, that is, technically speaking, N independent of $\bar{S}$ given $N > 0$. It is clear that for this independence case the pure premium can still be estimated based on the product form (9). However, for the general dependence case, the product form as in (9) is not valid any more, and instead we apply a Monte Carlo method for pure premium estimation.

Based on these discussions, we shall introduce the frequency-severity BCART models as a benchmark, where the frequency part and severity part are dealt with independently; see Subsection 3.1. In order to explore methods to capture the association between $\bar{S}$ and N more comprehensively, we propose two other types of BCART models. First, following the idea of Frees et al. (2011); Garrido et al. (2016) we introduce the *sequential BCART models* by including N (or its estimate $\hat{N}$) as a covariate when modeling the average severity. Second, motivated by Jørgensen and Paes De Souza (1994); Smyth and Jørgensen (2002); Delong et al. (2021), we introduce *joint BCART models* by considering (N, S) as a bivariate response. In this framework, association between the number of claims and average severity induced by potentially shared information (through covariates) is naturally incorporated within a selected single tree structure. These BCART models will be discussed in detail in Subsections 3.2 and 3.3, respectively.

3.1. Frequency-severity BCART models

Recall that the BCART models for the frequency component $\mathbb{E}(N)$ of (9) have been discussed in Zhang et al. (2024). Here we shall focus on the BCART modeling of the average severity component $\mathbb{E}(\bar{S}|N>0)$ of (9). More precisely, we will discuss a gamma distribution (as an example in the EDF) with N included as a weight, and three other distributions to directly model the average severity without including N as a weight, namely, gamma, lognormal, and Weibull. For this purpose, we will only consider a data subset with $N_i > 0$, and denote by $\bar{n}$ $(\leq n)$ the size of this subset. The subset of average severity data will be denoted by $(\boldsymbol{X}, \boldsymbol{N}, \bar{\boldsymbol{S}}) = \left((\boldsymbol{x}_1, N_1, \bar{S}_1), \ldots, (\boldsymbol{x}_{\bar{n}}, N_{\bar{n}}, \bar{S}_{\bar{n}})\right)^\top$.

3.1.1. Average severity modeling using gamma distribution with N as a weight

Assume the generic average severity $\bar{S}|N>0$ follows a gamma distribution with parameters being multipliers of N, i.e., $\bar{S}|N>0 \sim \text{Gamma}\,(N\alpha, N\beta)$, with $\alpha, \beta > 0$. Note that this is equivalent to assuming that the individual severity Y_j follows Gamma(α,β) distribution, due to the convolution property. Recall that the probability density function (pdf) of the Gamma(α,β) distribution and its mean and variance are given as

$$f_G(x) = \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x},\ x>0, \quad \mu_G = \frac{\alpha}{\beta},\ \ \sigma_G^2 = \frac{\alpha}{\beta^2}, \tag{10}$$

where $\Gamma(\cdot)$ is the gamma function. It is known that gamma distribution is right-skewed and relatively light-tailed.

According to the general BCART framework in Section 2, considering a tree $\mathcal{T}$ with b terminal nodes and $\boldsymbol{\theta}_t = (\alpha_t, \beta_t)$ the two-dimensional parameter for the t-th terminal node, we assume $\bar{S}_i|\boldsymbol{x}_i, N_i \sim \text{Gamma}\,\left(N_i\alpha(\boldsymbol{x}_i), N_i\beta(\boldsymbol{x}_i)\right)$ for the i-th observation, where $\alpha(\boldsymbol{x}_i) = \sum_{t=1}^b \alpha_t I_{(\boldsymbol{x}_i\in\mathcal{A}_t)}$ and $\beta(\boldsymbol{x}_i) = \sum_{t=1}^b \beta_t I_{(\boldsymbol{x}_i\in\mathcal{A}_t)}$, with $\{\mathcal{A}_t\}_{t=1}^b$ being the corresponding partition of $\mathcal{X}$. Specifically, for i-th observation such that $\boldsymbol{x}_i \in \mathcal{A}_t$, we have (with N_i compressed in f_G)

$$f_{\mathrm{G}}\left(\bar{S}_i|\alpha_t,\beta_t\right) = \frac{(N_i\beta_t)^{N_i\alpha_t}\bar{S}_i^{N_i\alpha_t-1}}{\Gamma(N_i\alpha_t)} e^{-N_i\beta_t\bar{S}_i}.$$

The mean and variance of $\bar{S}_i$ are thus given by α_t/β_t and $\alpha_t/(N_i\beta_t^2)$, respectively.

For each terminal node t, we treat α_t as known and β_t as unknown and shall not apply any data augmentation. According to the notation used in Section 2 this means $\boldsymbol{\theta}_{t,M} = \alpha_t$ and $\boldsymbol{\theta}_{t,B} = \beta_t$. Here α_t will be estimated using MME, i.e.,

$$\hat{\alpha}_t = \frac{(\bar{S})_t^2}{\mathrm{Var}(\bar{S})_t \bar{N}_t}, \tag{11}$$

where $(\bar{S})_t$ and $\mathrm{Var}(\bar{S})_t$ are the empirical mean and variance of the average severity, respectively, and $\bar{N}_t$ is the average claim number of the data in the t-th terminal node. We treat β_t as uncertain and use a conjugate gamma prior with hyper-parameters $\alpha_\pi, \beta_\pi > 0$. Denote the associated data in terminal node t as $\left(\boldsymbol{X}_t, \boldsymbol{N}_t, \bar{\boldsymbol{S}}_t\right) = ((X_{t1}, N_{t1}, \bar{S}_{t1}), \ldots, (X_{t\bar{n}_t}, N_{t\bar{n}_t}, \bar{S}_{t\bar{n}_t}))^\top$. The integrated likelihood for the terminal node t can then be obtained as

$$\begin{aligned}
&p_{\mathrm{G}}(\bar{\boldsymbol{S}}_t \mid \boldsymbol{X}_t, \boldsymbol{N}_t, \hat{\alpha}_t) \\
&= \int_0^\infty f_{\mathrm{G}}(\bar{\boldsymbol{S}}_t \mid \boldsymbol{N}_t, \hat{\alpha}_t, \beta_t) p(\beta_t) d\beta_t \\
&= \int_0^\infty \prod_{i=1}^{\bar{n}_t} \left(\frac{(N_{ti}\beta_t)^{N_{ti}\hat{\alpha}_t} \bar{S}_{ti}^{N_{ti}\hat{\alpha}_t-1} e^{-N_{ti}\beta_t\bar{S}_{ti}}}{\Gamma(N_{ti}\hat{\alpha}_t)} \right) \frac{\beta_\pi^{\alpha_\pi}\beta_t^{\alpha_\pi-1}e^{-\beta_\pi\beta_t}}{\Gamma(\alpha_\pi)} d\beta_t \qquad (12)\\
&= \frac{\beta_\pi^{\alpha_\pi} \prod_{i=1}^{\bar{n}_t}\left(N_{ti}^{N_{ti}\hat{\alpha}_t}\bar{S}_{ti}^{N_{ti}\hat{\alpha}_t-1}\right)}{\Gamma(\alpha_\pi)\prod_{i=1}^{\bar{n}_t}\Gamma(N_{ti}\hat{\alpha}_t)} \frac{\Gamma(\sum_{i=1}^{\bar{n}_t} N_{ti}\hat{\alpha}_t + \alpha_\pi)}{(\sum_{i=1}^{\bar{n}_t} N_{ti}\bar{S}_{ti} + \beta_\pi)^{\sum_{i=1}^{\bar{n}_t} N_{ti}\hat{\alpha}_t + \alpha_\pi}}.
\end{aligned}$$

Clearly, from (12), we see that the posterior distribution of β_t conditional on data $(\boldsymbol{N}_t,\ \bar{\boldsymbol{S}}_t)$ and the estimated parameter $\hat{\alpha}_t$, is given by

$$\beta_t \mid \boldsymbol{N}_t, \bar{\boldsymbol{S}}_t, \hat{\alpha}_t \ \sim\ \text{Gamma}\left(\sum_{i=1}^{\bar{n}_t} N_{ti}\hat{\alpha}_t + \alpha_\pi, \sum_{i=1}^{\bar{n}_t} N_{ti}\bar{S}_{ti} + \beta_\pi\right).$$

The integrated likelihood for the tree $\mathcal{T}$ is thus given by

$$p_{\mathrm{G}}\left(\bar{\boldsymbol{S}} \mid \boldsymbol{X}, \boldsymbol{N}, \hat{\boldsymbol{\alpha}}, \mathcal{T}\right) = \prod_{t=1}^b p_{\mathrm{G}}\left(\bar{\boldsymbol{S}}_t \mid \boldsymbol{X}_t, \boldsymbol{N}_t, \hat{\alpha}_t\right).$$

Next, we discuss the DIC for this tree. Following (Zhang et al., 2024), a DIC$_t$ for terminal node t can be defined as $\mathrm{DIC}_t = D(\bar{\beta}_t) + 2p_{Dt}$, where the posterior mean of β_t is given by

$$\bar{\beta}_t = \frac{\sum_{i=1}^{\bar{n}_t} N_{ti}\hat{\alpha}_t + \alpha_\pi}{\sum_{i=1}^{\bar{n}_t} N_{ti}\bar{S}_{ti} + \beta_\pi}; \tag{13}$$

the goodness-of-fit is given as

$$\begin{aligned}
D(\bar{\beta}_t) &= -2\sum_{i=1}^{\bar{n}_t} \log f_{\mathrm{G}}(\bar{S}_{ti} \mid N_i, \hat{\alpha}_t, \bar{\beta}_t) \\
&= -2\sum_{i=1}^{\bar{n}_t} \left[N_{ti}\hat{\alpha}_t \log(N_{ti}\bar{\beta}_t) + (N_{ti}\hat{\alpha}_t - 1)\log(\bar{S}_{ti}) \right.\\
&\qquad \left. - \bar{\beta}_t N_{ti}\bar{S}_{ti} - \log\left(\Gamma(N_{ti}\hat{\alpha}_t)\right)\right],
\end{aligned}$$

and the effective number of parameters p_{Dt} is defined by

$$\begin{aligned}
p_{Dt} &= 1 + \overline{D(\beta_t)} - D(\bar{\beta}_t) \\
&= 1 + 2\sum_{i=1}^{\bar{n}_t}\left\{\log\left(f_{\mathrm{G}}(\bar{S}_{ti} \mid \hat{\alpha}_t, \bar{\beta}_t)\right) - \mathbb{E}_{\text{post}}\left(\log(f_{\mathrm{G}}\left(\bar{S}_{ti} \mid \hat{\alpha}_t, \beta_t\right))\right)\right\},
\end{aligned} \tag{14}$$

where 1 is added for the parameter α_t which was estimated upfront, and the difference of the last two terms on the right-hand side of the first line is the effective number for the unknown parameter β_t. Some direct calculations yield that

$$p_{Dt} = 1 + 2\left(\log\left(\sum_{i=1}^{\bar{n}_t} N_{ti}\hat{\alpha}_t + \alpha_\pi\right) - \psi\left(\sum_{i=1}^{\bar{n}_t} N_{ti}\hat{\alpha}_t + \alpha_\pi\right)\right)\sum_{i=1}^{\bar{n}_t} N_{ti}\hat{\alpha}_t,$$

with $\psi(x) = \Gamma'(x)/\Gamma(x)$ being the digamma function, and thus DIC$_t$ can be derived. Consequently, the DIC of the whole tree $\mathcal{T}$ is obtained as

Table 1
Estimations for average severity in terminal node t. Here $(\bar{S})_t$ and $\mathrm{Var}(\bar{S})_t$ denote the empirical mean and variance of the average severity in the t-th node, respectively. See Supplementary Material SM.A for details.

Distribution	$\mathrm{Gamma}(\alpha_t, \beta_t)$	$\mathrm{LN}(\mu_t, \sigma_t)$	$\mathrm{Weib}(\alpha_t, \beta_t)$
Prediction $\hat{\bar{S}}_t$	$\hat{\alpha}_t/\bar{\beta}_t$	$\exp(\bar{\mu}_t + \hat{\sigma}_t^2/2)$	$\bar{\beta}_t^{1/\hat{\alpha}_t}\Gamma(1+1/\hat{\alpha}_t)$
Parameter	$\hat{\alpha}_t = \frac{(\bar{S})_t^2}{\mathrm{Var}(\bar{S})_t}$	$\hat{\sigma}_t$ obtained using MME	$\hat{\alpha}_t$ obtained using MME
estimation	$\bar{\beta}_t = \frac{\bar{n}_t\hat{\alpha}_t+\alpha_\pi}{\sum_{i=1}^{\bar{n}_t}\bar{S}_{ti}+\beta_\pi}$	$\bar{\mu}_t = \frac{\hat{\sigma}_t^2\sigma_\pi^2}{\bar{n}_t\sigma_\pi^2+\hat{\sigma}_t^2}\left(\frac{\mu_\pi}{\sigma_\pi^2} + \frac{\sum_{i=1}^{\bar{n}_t}\log(\bar{S}_{ti})}{\hat{\sigma}_t^2}\right)$	$\bar{\beta}_t = \frac{\sum_{i=1}^{\bar{n}_t}\bar{S}_{ti}^{\hat{\alpha}_t}+\beta_\pi}{\bar{n}_t+\alpha_\pi-1}$

$$\mathrm{DIC} = \sum_{t=1}^{b} \mathrm{DIC}_t. \tag{15}$$

With the above formulas derived for the gamma case, we can use the approach presented in Table 1 of Zhang et al. (2024), together with Algorithm 1, to search for an optimal tree which can then be used to predict new data such that the estimated average severity $\hat{\alpha}_t/\bar{\beta}_t$ in each terminal node t can be determined using (11) and (13).

3.1.2. Average severity modeling using distributions without N as a weight

Three distributions (gamma, lognormal and Weibull) will be used to model the average severity $\bar{S}|N > 0$. Selecting among these three distributions for certain data may pose a considerable challenge, and scholars have extensively explored this topic; see, e.g., Siswadi and Quesenberry (1982). In average severity modeling, insurers want to gain more insights into the right tail. The gamma distribution would be a suitable model for losses that are not catastrophic, such as auto insurance. The lognormal distribution is more suitable for fire insurance, which may exhibit more extreme values than auto insurance. Moreover, the Weibull distribution has the ability to handle different scenarios by tuning the shape parameter to adapt to different tail characteristics.

We demonstrate how to apply these distributions in BCART models for the average severity data. The idea, as in the previous section, is to specify the distributions/parameters in the general BCART framework of Section 2. We only give some key information below, and defer some detailed calculations to Section SM.A of the Supplementary Material.

Consider a tree $\mathcal{T}$ with b terminal nodes for the average severity data. In gamma and Weibull models, we respectively assume $\bar{S}_i \mid \boldsymbol{x}_i \sim \mathrm{Gamma}\left(\alpha(\boldsymbol{x}_i), \beta(\boldsymbol{x}_i)\right)$, and $\bar{S}_i \mid \boldsymbol{x}_i \sim \mathrm{Weib}\left(\alpha(\boldsymbol{x}_i), \beta(\boldsymbol{x}_i)\right)$, where $\alpha(\boldsymbol{x}_i) = \sum_{t=1}^{b}\alpha_t I_{(\boldsymbol{x}_i\in\mathcal{A}_t)}$, $\beta(\boldsymbol{x}_i) = \sum_{t=1}^{b}\beta_t I_{(\boldsymbol{x}_i\in\mathcal{A}_t)}$. In a lognormal model, we assume that $\bar{S}_i \mid \boldsymbol{x}_i \sim \mathrm{LN}\left(\mu(\boldsymbol{x}_i), \sigma(\boldsymbol{x}_i)\right)$, where $\mu(\boldsymbol{x}_i) = \sum_{t=1}^{b}\mu_t I_{(\boldsymbol{x}_i\in\mathcal{A}_t)}$, and $\sigma(\boldsymbol{x}_i) = \sum_{t=1}^{b}\sigma_t I_{(\boldsymbol{x}_i\in\mathcal{A}_t)}$.

For each terminal node t, we treat one parameter as known and the other as unknown, that is, according to the notation in Section 2, $\boldsymbol{\theta}_{t,M}$ is α_t for the gamma and Weibull models and is σ_t for the lognormal model, and $\boldsymbol{\theta}_{t,B}$ is β_t for the gamma and Weibull models and is μ_t for the lognormal model. Furthermore, we apply a conjugate prior for β_t and μ_t, namely, a $\mathrm{Gamma}(\alpha_\pi, \beta_\pi)$ prior for the β_t in the gamma model, a $\mathrm{Normal}(\mu_\pi, \sigma_\pi)$ prior for the μ_t in the lognormal model, and an inverse-$\mathrm{Gamma}(\alpha_\pi, \beta_\pi)$ prior for the β_t in the Weibull model, i.e.,

$$p(\beta_t) = \frac{\beta_\pi^{\alpha_\pi}}{\Gamma(\alpha_\pi)}\beta_t^{-\alpha_\pi-1}\exp(-\beta_\pi/\beta_t), \tag{16}$$

with $\alpha_\pi, \beta_\pi > 0$. Estimates for the unknown parameters, calculations of the integrated likelihood and DIC_t for these three models are given in Section SM.A of the Supplementary Material. We can then use the above procedure leading to the predictions obtained using (8) from different models, as displayed in Table 1.

Remark 2.

(a). There are different ways to parameterize the Weibull distribution, either with two or three parameters; see, e.g., Rinne (2008). For simplicity, we adopt the common parameterization with two parameters; see, e.g., Fink (1997).

(b). In the above BCART models for average severity we have assumed that one parameter of the distribution is treated as known and the other is treated as unknown which is given a conjugate prior. We note that this is not the only way to implement the BCART algorithms. There are other ways to treat the parameters. For example, for the gamma distribution, the following two alternative approaches can be considered:
- Treat the parameter β_t as known and use a prior for α_t, i.e., $p(\alpha_t) \propto a_0^{\alpha_t-1}\beta_t^{\alpha_t c_0}/\Gamma(\alpha_t)^{b_0}$ where a_0, b_0, c_0 are prior hyper-parameters.
- Treat both α_t and β_t as unknown and use a joint prior for them, i.e., $p(\alpha_t, \beta_t) \propto 1/(\Gamma(\alpha_t)^{c_0}\beta_t^{-\alpha_t d_0})\ {a_0}^{\alpha_t-1}e^{-\beta_t b_0}$ where a_0, b_0, c_0, d_0 are prior hyper-parameters; see, e.g., Fink (1997).

Although the joint prior can be used to obtain estimators for α_t and β_t simultaneously in the Bayesian framework, it is not formulated as an exact distribution, leading to less accurate estimators. The first way also has this shortcoming. For the lognormal distribution, a normal and inverse-gamma joint prior can be used for the parameters μ and σ^2; see, e.g., Fink (1997). These more complicated cases are not considered in our current implementation.

(c). Many other distributions can also be used to model average severity, such as Pareto, generalized gamma, generalized Pareto distributions, and so on. However, they either have too many parameters or are challenging to make explicit calculations in the Bayesian framework. We believe further research into the selection of these distributions is worth exploring; see, e.g., Shi et al. (2015); Mehmet and Saykan (2005); Farkas et al. (2021) for some insights on the application of these distributions to insurance pricing.

In the frequency-severity BCART models, we obtain two trees for frequency and average severity respectively. The pure premium can be calculated using the predictions from these two trees together with the pricing formula (9). There can be many different combinations of predictions for the frequency-severity models, i.e., any model discussed in Zhang et al. (2024) for frequency and any model introduced above for average severity can be adopted.

One benefit of modeling frequency and average severity separately using two trees is that the important risk factors associated with each component can be discovered separately. However, it can be challenging to interpret two trees as a whole, since several policyholders may be classified in one cell by the frequency tree but in a different cell by the average severity tree. In the next section, we discuss the combination of two trees for prediction and interpretation.

3.1.3. Evaluation metrics for frequency-severity BCART models

In this section, we begin by exploring some performance evaluation metrics for average severity BCART models, then we introduce the idea of combining two trees to derive evaluation metrics for the frequency-severity BCART models. Application of these evaluation metrics will be discussed in Sections 4 and 5.

Evaluation metrics for average severity trees

We use the same performance measures that were introduced in Zhang et al. (2024). Suppose we have obtained a tree with b terminal nodes and the corresponding predictions $\hat{\bar{S}}_t$ $(t = 1, \ldots, b)$ given in Table 1. Consider a test dataset with $\bar{m}$ observations. Denote the number

Table 2
Variance ($\hat{V}_t$) of the average severity distribution in terminal node t using the BCART estimations. Below GammaN means the gamma model with N as a weight, while Gamma means the gamma model without N as a weight.

Dist.	GammaN	Gamma	Lognormal	Weibull
$\hat{V}_t$	$\hat{\alpha}_t/(\bar{N}_t\hat{\beta}_t^2)$	$\hat{\alpha}_t/\hat{\beta}_t^2$	$\left(e^{\hat{\sigma}_t^2}-1\right)e^{2\hat{\mu}_t+\hat{\sigma}_t^2}$	$\hat{\beta}_t^{2/\hat{\alpha}_t}\left[\Gamma\left(1+2/\hat{\alpha}_t\right)-\left(\Gamma\left(1+1/\hat{\alpha}_t\right)\right)^2\right]$

of test data in terminal node t by $\bar{m}_t$, and denote the associated data in terminal node t as $\left(\boldsymbol{X}_t, \boldsymbol{N}_t, \bar{\boldsymbol{S}}_t\right) = ((\boldsymbol{x}_{t1}, N_{t1}, \bar{S}_{t1}), \ldots, (\boldsymbol{x}_{t\bar{m}_t}, N_{t\bar{m}_t}, \bar{S}_{t\bar{m}_t})\Big)^\top$ $(t = 1, \ldots, b)$. The evaluation metrics are listed below.

M1: The residual sum of squares (RSS) is given by $\text{RSS}(\bar{\boldsymbol{S}}) = \sum_{t=1}^b \sum_{i=1}^{\bar{m}_t} (\bar{S}_{ti} - \hat{\bar{S}}_{ti})^2$.

M2: The squared error (SE), based on a sub-portfolio (i.e., those instances in the same terminal node) level, is defined by $\text{SE}(\bar{\boldsymbol{S}}) = \sum_{t=1}^b \left(\sum_{i=1}^{\bar{m}_t} S_{ti} / \sum_{i=1}^{\bar{m}_t} N_{ti} - \hat{\bar{S}}_t\right)^2$.

M3: Discrepancy statistic (DS) is defined as a weighted version of SE, given by $\text{DS}(\bar{\boldsymbol{S}}) = \sum_{t=1}^b \left(\sum_{i=1}^{\bar{m}_t} S_{ti} / \sum_{i=1}^{\bar{m}_t} N_{ti} - \hat{\bar{S}}_t\right)^2 / \hat{V}_t$, where $\hat{V}_t$ for different models are given in Table 2.

M4: Model Lift indicates the ability to differentiate between cells of policyholders with low and high risks (average severity here), and is defined by using the data and their predicted values in the most and least risky cells. We use a similar approach as in Zhang et al. (2024) to calculate Lift for the average severity tree models; more details on these calculations can be found in Zhang (2024).

Evaluation metrics for two trees from the frequency-severity model

The frequency-severity BCART model yields two trees. We now explain how to combine these two trees to evaluate model performance based on the aggregate claim amount (or pure premium) prediction $\hat{S}$ for a test dataset with m observations. The idea is natural – individual tree partitions are superimposed to form a joint partition of the covariate space $\mathcal{X}$. This process evolves by merging all the splitting rules from both trees. The splits of each tree contribute to a refined segmentation of the covariate space, resulting in a joint partition that represents the collective behavior of the original two tree partitions; see Rocková et al. (2020).

Suppose we have obtained a joint partition with c cells. The corresponding prediction $\hat{S}_t$ for cell t $(t = 1, \ldots, c)$ is obtained by (9),

$$\hat{S}_t = \hat{N}_{s(t)}\hat{\bar{S}}_{l(t)}, \tag{17}$$

where $s(t)$ is the corresponding node index of t in the frequency tree and $l(t)$ is the corresponding one in the severity tree, and $\hat{N}_{s(t)}$ and $\hat{\bar{S}}_{l(t)}$ are the corresponding estimates from these two individual trees. We further denote m_t as the number of observations in cell t. Using the notation above we have

M1': The residual sum of squares $\text{RSS}(\boldsymbol{S}) = \sum_{t=1}^c \sum_{i=1}^{m_t} (S_{ti} - \hat{S}_{ti})^2$.

M2': The squared error $\text{SE}(\boldsymbol{S}) = \sum_{t=1}^c \left(\sum_{i=1}^{m_t} S_{ti} / \sum_{i=1}^{m_t} v_{ti} - \hat{S}_t\right)^2$.

M3': The discrepancy statistic $\text{DS}(\boldsymbol{S}) = \sum_{t=1}^c \left(\sum_{i=1}^{m_t} S_{ti} / \sum_{i=1}^{m_t} v_{ti} - \hat{S}_t\right)^2 / \hat{V}_t$, where $\hat{V}_t$ is the estimated model variance of S in the t-th cell which is derived using the model specific assumptions and its parameter estimates. More specifically, if the average severity model is as in Subsection 3.1.1, we can rewrite $S = \sum_{j=1}^N Y_j$ with Y_j following independent gamma distribution with parameters α, β as in (10). Thus, we use $\text{Var}(S) = \mathbb{E}(N)\text{Var}(Y) + (\mathbb{E}(Y))^2 \text{Var}(N)$ to derive an estimate of $\hat{V}_t$ for the t-th cell, together with corresponding estimated parameters for α, β given in Subsection 3.1.1 and for different frequency models in Zhang et al. (2024). Further, if the average severity model is as in Subsection 3.1.2, assuming N and $\bar{S}|N > 0$ are independent, we use $\text{Var}(S) = \left(\text{Var}(N) + (\mathbb{E}(N))^2\right) \text{Var}(\bar{S}|N > 0) + \text{Var}(N)(\mathbb{E}(\bar{S}|N > 0))^2$ to derive an estimate for $\hat{V}_t$ for the t-th cell, together with corresponding estimated parameters for different average severity models in Tables 1-2 and for different frequency models in Zhang et al. (2024).

M4': Model Lift - similarly defined as M4 and Zhang et al. (2024).

Remark 3. We remark that for each of the BCART models, we apply the three-step approach in Table 1 of Zhang et al. (2024) to select a tree model. The above evaluation metrics are then used to evaluate the performance of these tree models on test data, based on which we can select a best tree model among different types of BCART models. As observed in Zhang et al. (2024) when discussing different types of frequency BCART models, all four metrics yield the same type of tree model choice based on their performance on test data. It is worth pointing out that within a certain type of BCART model, the test data performance using SE and DS aligns with the tree model selected using the three-step approach and thus they are deemed to be preferred metrics for comparison. See also Sections 4 and 5 for further discussion.

3.2. Sequential BCART models

In this section, we introduce the sequential model to better capture the potential association between the number of claims and average severity. One popular approach in the literature, e.g., Frees et al. (2011); Garrido et al. (2016); Gschlößl and Czado (2007), is to treat the number of claims N as a covariate for the average severity modeling. Following this idea, our sequential BCART model consists of two steps: 1) model the frequency component of (9) using the BCART models developed in Zhang et al. (2024); 2) treat the number of claims N as a covariate (also treated as a model weight in the GammaN model) for the average severity component in (9) using the BCART models introduced in Subsection 3.1.1 or 3.1.2.

When modeling average severity with N as a covariate, there are usually two ways to treat N, namely, either use N as a numeric covariate (see Garrido et al. (2016)) or treat N as a factor (see Gschlößl and Czado (2007)). In this paper, we propose including the information of claim count for the average severity modeling, using the estimation of claim count $\hat{N}$ from the frequency BCART model as a numeric covariate. The underlying ideas for this proposal are as follows. First, for a new observation we do not know N but can only obtain its estimate $\hat{N}$ through frequency model. Second, the frequency tree will classify the policyholders with similar risk (in terms of claim frequency) into the same cell and assign similar estimations $\hat{N}$ (the value of them depends also on their exposure). If the claim count information is highly correlated to the average severity, then the estimated value $\hat{N}$ will be chosen as the splitting covariate and the policyholders in the same frequency cell will be more likely (than using N) to be classified into the same cell by the average severity tree. In doing so, we expect the sequential model would be able to better capture the potential dependence between the number of claims and average severity. This is demonstrated to be true by our simulated and real data below.

The sequential BCART models will also result in two trees, one for frequency and the other for average severity. If the resulting severity tree does not include the number of claims N (or $\hat{N}$) as its splitting variable, then the evaluation metrics and estimates introduced in Subsection 3.1.3 can still be employed. Otherwise, the aggregate claim amount $\hat{S}$ cannot be estimated directly by multiplying the estimates from these two trees due to lack of independence, for which we will apply a Monte Carlo method following the idea in Gschlößl and Czado (2007). In this case, there are also challenges involved in combining two trees to form a joint partition of the covariate space $\mathcal{X}$, since now the average severity tree does not only involve splitting variables $\boldsymbol{x}$ but also involve N (or $\hat{N}$). To overcome this, we propose to construct a joint partition using only the splitting rules introduced by $\boldsymbol{x}$ in these two trees, ignoring any ones based on N (or $\hat{N}$). As a result, some changes for M1'-M4' are necessary, which are as follows. Assume the joint partition has c cells.

- M1': we will use $\mathrm{RSS}(\boldsymbol{S}) = \sum_{i=1}^{m}(S_i - \hat{S}_i)^2$, where the prediction $\hat{S}_i$ is obtained using Monte Carlo simulation for the i-th test data given $(\boldsymbol{x}_i, v_i)$. The simulation algorithm is included in Section SM.B of the Supplementary Material; see Algorithm SM.1.
- M2'-M4': we need estimates for $\hat{S}_t, \hat{V}_t$ $(t = 1, \dots, c)$ which can be obtained using similar Monte Carlo method as for M1'. Specifically, we consider a generic data in node t, denoted by $\boldsymbol{x}$ with an exposure $v = 1$. Inputting this data $(\boldsymbol{x}, v = 1)$ into the Algorithm SM.1, we obtain a sequence of realizations $S_t^1, S_t^2, \dots, S_t^R$, from which we can obtain $\hat{S}_t$ and $\hat{V}_t$ using their empirical mean and variance.

Remark 4. In Quan et al. (2023), a hybrid approach is used for insurance pricing where the terminal nodes of a frequency tree serve as the partition of $\mathcal{X}$ and a regression model is used in each terminal node for claim severity. In contrast, our approach for partition incorporates information from both frequency tree and severity tree. We have validated the effectiveness and superiority of our method (over a partition using only frequency tree) in both simulated and real data analysis.

3.3. Joint BCART models

Different from the previous two types of BCART models where separate tree models are used for the frequency and average severity, in this section we introduce the third type of BCART models, called joint BCART models, where we consider (N, S) as a bivariate response; see Jørgensen and Paes De Souza (1994); Smyth and Jørgensen (2002) for a similar treatment in GLMs. We discuss two commonly used distributions for aggregate claim amount S, namely, CPG and ZICPG distributions. The presence of a discrete mass at zero makes them suitable for modeling aggregate claim amount; see, e.g., Quijano Xacur and Garrido (2015); Yang et al. (2018); Denuit et al. (2021). As in Zhang et al. (2024), for the ZICPG models we need to employ a data augmentation technique. We also explore different ways to embed exposure. The advantage of modeling frequency and (average) severity components separately has been recognized in the literature; see, e.g., Quijano Xacur and Garrido (2015); Frees et al. (2016). In particular, this separate treatment can reflect the situation when the covariates that affect the frequency and severity are very different. However, one disadvantage is that it takes more effort to combine the two resulting tree models, as we have already seen in Subsection 3.1.3. Compared to the use of two separate tree models, the advantage of joint modeling is that the resulting single tree is easier to interpret as it simultaneously gives estimates for frequency, pure premium and thus average severity. Additionally, for the situation where frequency and average severity are linked through shared covariates, using a parsimonious joint tree model might be advantageous; this will be illustrated in the examples in Section 4.

Since the ZICPG model is an extension of the CPG model and accounts for the high proportion of zero claims in practice, we focus primarily on the ZICPG model and refer to Section SM.C of the Supplementary Material for details of the CPG model.

3.3.1. Zero-inflated compound Poisson gamma model

We consider a response (N, S) with exposure $v = 1$, where N is zero-inflated Poisson distributed with parameters μ and λ, and $S = \sum_{j=1}^{N} Y_j$ with individual severity Y_j following independent gamma distribution with parameters $\alpha, \beta > 0$. In the following, S is called a zero-inflated compound Poisson gamma random variable, denoted by ZICPG$(\mu, \lambda, \alpha, \beta)$. According to the general BCART framework in Section 2, considering a tree $\mathcal{T}$ with b terminal nodes and $\boldsymbol{\theta}_t = (\mu_t, \lambda_t, \alpha_t, \beta_t)$ $(t = 1, \dots, b)$, we assume $N_i|\boldsymbol{x}_i \sim \mathrm{ZIP}\left(\mu(\boldsymbol{x}_i), \lambda(\boldsymbol{x}_i)\right)$, and $S_i|\boldsymbol{x}_i, N_i > 0 \sim \mathrm{Gamma}\left(N_i\alpha(\boldsymbol{x}_i), \beta(\boldsymbol{x}_i)\right)$ for the i-th observation, where $\mu(x_i) = \sum_{t=1}^{b} \mu_t I_{(x_i \in \mathcal{A}_t)}$, $\lambda(\boldsymbol{x}_i) = \sum_{t=1}^{b} \lambda_t I_{(x_i \in \mathcal{A}_t)}$, $\alpha(\boldsymbol{x}_i) = \sum_{t=1}^{b} \alpha_t I_{(x_i \in \mathcal{A}_t)}$ and $\beta(\boldsymbol{x}_i) = \sum_{t=1}^{b} \beta_t I_{(x_i \in \mathcal{A}_t)}$.

For ZICPG models, we need to introduce a data augmentation strategy as in Zhang et al. (2024) to obtain a closed form expression for the integrated likelihood; see (20) below. Motivated by the discussion on the ZIP-BCART models in Zhang et al. (2024), we construct three ZICPG models according to how the exposure is embedded into the modeling. We try to cover all three ZICPG models in a general set-up, which requires some general notation for exposure. Specifically, for i-th observation such that $\boldsymbol{x}_i \in \mathcal{A}_t$, we have the joint distribution (recalling that the associated data for terminal node t is denoted by $(\boldsymbol{X}_t, \boldsymbol{v}_t, \boldsymbol{N}_t, \boldsymbol{S}_t)$)

$$
\begin{aligned}
&f_{\mathrm{ZICPG}}(N_{ti}, S_{ti} \mid \mu_t, \lambda_t, \alpha_t, \beta_t) = f_{\mathrm{ZIP}}(N_{ti} \mid \mu_t, \lambda_t) f_{\mathrm{G}}(S_{ti} \mid N_{ti}, \alpha_t, \beta_t) \\
&= \begin{cases} \frac{1}{1+\mu_t w_{ti}} + \frac{\mu_t w_{ti}}{1+\mu_t w_{ti}} e^{-\lambda_t u_{ti}} & (N_{ti}, S_{ti}) = (0,0), \\ \frac{\mu_t w_{ti}}{1+\mu_t w_{ti}} \frac{(\lambda_t u_{ti})^{N_{ti}} e^{-\lambda_t u_{ti}}}{N_{ti}!} \frac{\beta_t^{N_{ti}\alpha_t} S_{ti}^{N_{ti}\alpha_t - 1} e^{-\beta_t S_{ti}}}{\Gamma(N_{ti}\alpha_t)} & (N_{ti}, S_{ti}) \in (\mathbb{N} \times \mathbb{R}^+), \end{cases}
\end{aligned}
\tag{18}
$$

where we use w_{ti} to denote the "exposure" for the zero mass part and u_{ti} to denote the "exposure" for the Poisson part. The above general formulation can cover three different models as special cases. Namely, 1) setting $w_{ti} = 1$ and $u_{ti} = v_{ti}$, then the exposure is only embedded in the Poisson part, yielding the ZICPG1 model; 2) setting $w_{ti} = v_{ti}$ and $u_{ti} = 1$ then the exposure is only embedded in the zero mass part, yielding the ZICPG2 model; 3) setting $w_{ti} = u_{ti} = v_{ti}$ means the exposure is embedded in both parts, yielding the ZICPG3 model. Note that $1/(1 + \mu_t w_{ti}) \in (0, 1)$ is the probability that zero is due to the point mass component.

For computational convenience, a data augmentation scheme is used. To this end, we introduce two latent variables $\boldsymbol{\phi}_t = (\phi_{t1}, \phi_{t2}, \dots, \phi_{tn_t}) \in (0, \infty)^{n_t}$ and $\boldsymbol{\delta}_t = (\delta_{t1}, \delta_{t2}, \dots, \delta_{tn_t}) \in \{0, 1\}^{n_t}$, and define the data augmented likelihood for the i-th data instance in terminal node t as

$$
\begin{aligned}
&f_{\mathrm{ZICPG}}(N_{ti}, S_{ti}, \delta_{ti}, \phi_{ti} \mid \mu_t, \lambda_t, \alpha_t, \beta_t) \\
&= e^{-\phi_{ti}(1+\mu_t w_{ti})} \left(\frac{\mu_t w_{ti} \left(\lambda_t u_{ti}\right)^{N_{ti}}}{N_{ti}!} e^{-\lambda_t u_{ti}} \right)^{\delta_{ti}} \\
&\quad \times \left(\left(\frac{\beta_t^{N_{ti}\alpha_t} S_{ti}^{N_{ti}\alpha_t - 1} e^{-\beta_t S_{ti}}}{\Gamma(N_{ti}\alpha_t)} - 1 \right) I_{(N_{ti}>0)} + 1 \right),
\end{aligned}
\tag{19}
$$

where the support of the function f_{ZICPG} is $(\{0\} \times \{0\} \times \{0, 1\} \times (0, \infty)) \cup \left(\mathbb{N} \times \mathbb{R}^+ \times \{1\} \times (0, \infty)\right)$. It can be shown that (18) is the marginal distribution of the above augmented distribution; see Section SM.D of the Supplementary Material for more details.

By conditional arguments, we can also check that δ_{ti}, given data $N_{ti} = S_{ti} = 0$ and parameters (μ_t and λ_t), has a Bernoulli distribution, i.e., $\delta_{ti} \mid N_{ti} = 0, \mu_t, \lambda_t \sim \mathrm{Bern}\left(\frac{\mu_t w_{ti} e^{-\lambda_t u_{ti}}}{1+\mu_t w_{ti} e^{-\lambda_t u_{ti}}}\right)$, and $\delta_{ti} = 1$ if $N_{ti} > 0$. Furthermore, $\phi_{ti} \mid \mu_t \sim \mathrm{Exp}\left(1 + \mu_t w_{ti}\right)$.

For each terminal node t, we treat α_t as known, μ_t, λ_t and β_t as unknown and apply the above data augmentation. According to the notation used in Section 2 this means $\boldsymbol{\theta}_{t,M} = \alpha_t$ and $\boldsymbol{\theta}_{t,B} = (\mu_t, \lambda_t, \beta_t)$. Here α_t will be estimated as in (11) using a subset of data with $N > 0$. We treat μ_t, λ_t and β_t as uncertain and use independent conjugate gamma priors, i.e., $\mu_t \sim \mathrm{Gamma}(\alpha^{(\mu)}, \beta^{(\mu)})$, $\lambda_t \sim \mathrm{Gamma}(\alpha^{(\lambda)}, \beta^{(\lambda)})$, $\beta_t \sim \mathrm{Gamma}(\alpha^{(\beta)}, \beta^{(\beta)})$, where the superscript (μ) (or (λ) and (β)) indicates this hyper-parameter is assigned for the parameter μ_t (or λ_t and β_t). Then, given the estimated parameter $\hat{\alpha}_t$, the integrated augmented likelihood for terminal node t can be obtained as

$$
p_{\mathrm{ZICPG}}\left(\boldsymbol{N}_t, \boldsymbol{S}_t, \boldsymbol{\delta}_t, \boldsymbol{\phi}_t \mid \boldsymbol{X}_t, \hat{\alpha}_t\right)
$$

$$= \int_0^\infty \int_0^\infty \int_0^\infty f_{\text{ZICPG}}\left(\boldsymbol{N}_t, \boldsymbol{S}_t, \boldsymbol{\delta}_t, \boldsymbol{\phi}_t \mid \mu_t, \lambda_t, \hat{\alpha}_t, \beta_t\right) p(\mu_t) p(\lambda_t) p(\beta_t) d\mu_t d\lambda_t d\beta_t$$

$$= \iiint \prod_{i=1}^{n_t} \left(e^{-\phi_{ti}(1+\mu_t w_{ti})} \left(\frac{\mu_t w_{ti} \left(\lambda_t u_{ti}\right)^{N_{ti}}}{N_{ti}!} e^{-\lambda_t u_{ti}} \right)^{\delta_{ti}} \right)$$

$$\times \prod_{i:N_{ti}>0} \frac{\beta_t^{N_{ti}\hat{\alpha}_t} S_{ti}^{N_{ti}\hat{\alpha}_t - 1} e^{-\beta_t S_{ti}}}{\Gamma(N_{ti}\hat{\alpha}_t)}$$

$$\times \frac{\beta^{(\mu)\alpha^{(\mu)}} \mu_t^{\alpha^{(\mu)}-1} e^{-\beta^{(\mu)}\mu_t}}{\Gamma\left(\alpha^{(\mu)}\right)} \frac{\beta^{(\lambda)\alpha^{(\lambda)}} \lambda_t^{\alpha^{(\lambda)}-1} e^{-\beta^{(\lambda)}\lambda_t}}{\Gamma\left(\alpha^{(\lambda)}\right)}$$

$$\times \frac{\beta^{(\beta)\alpha^{(\beta)}} \beta_t^{\alpha^{(\beta)}-1} e^{-\beta^{(\beta)}\beta_t}}{\Gamma\left(\alpha^{(\beta)}\right)} d\mu_t d\lambda_t d\beta_t$$

$$= \frac{\beta^{(\mu)\alpha^{(\mu)}}}{\Gamma\left(\alpha^{(\mu)}\right)} \frac{\beta^{(\lambda)\alpha^{(\lambda)}}}{\Gamma\left(\alpha^{(\lambda)}\right)} \frac{\beta^{(\beta)\alpha^{(\beta)}}}{\Gamma\left(\alpha^{(\beta)}\right)} \prod_{i=1}^{n_t} \left(e^{-\phi_{ti}} w_{ti}^{\delta_{ti}} u_{ti}^{\delta_{ti} N_{ti}} \left(N_{ti}!\right)^{-\delta_{ti}} \right)$$

$$\times \prod_{i:N_{ti}>0} \frac{S_{ti}^{N_{ti}\hat{\alpha}_t - 1}}{\Gamma(N_{ti}\hat{\alpha}_t)}$$

$$\times \frac{\Gamma\left(\sum_{i=1}^{n_t} \delta_{ti} + \alpha^{(\mu)}\right)}{\left(\sum_{i=1}^{n_t} \phi_{ti} w_{ti} + \beta^{(\mu)}\right)^{\sum_{i=1}^{n_t} \delta_{ti} + \alpha^{(\mu)}}} \frac{\Gamma\left(\sum_{i=1}^{n_t} \delta_{ti} N_{ti} + \alpha^{(\lambda)}\right)}{\left(\sum_{i=1}^{n_t} \delta_{ti} u_{ti} + \beta^{(\lambda)}\right)^{\sum_{i=1}^{n_t} \delta_{ti} N_{ti} + \alpha^{(\lambda)}}}$$

$$\times \frac{\Gamma(\sum_{i:N_{ti}>0} N_{ti}\hat{\alpha}_t + \alpha^{(\beta)})}{(\sum_{i:N_{ti}>0} S_{ti} + \beta^{(\beta)})^{\sum_{i:N_{ti}>0} N_{ti}\hat{\alpha}_t + \alpha^{(\beta)}}}. \tag{20}$$

The integrated augmented likelihood for the tree $\mathcal{T}$ is thus given by

$$p_{\text{ZICPG}}\left(\boldsymbol{N}, \boldsymbol{S}, \boldsymbol{\delta}, \boldsymbol{\phi} \mid \boldsymbol{X}, \hat{\boldsymbol{\alpha}}, \mathcal{T}\right) = \prod_{t=1}^{b} p_{\text{ZICPG}}\left(\boldsymbol{N}_t, \boldsymbol{S}_t, \boldsymbol{\delta}_t, \boldsymbol{\phi}_t \mid \boldsymbol{X}_t, \hat{\alpha}_t\right).$$

We now discuss DIC which can be derived similarly as in Subsection 3.1.1 with a three-dimensional unknown parameter $(\mu_t, \lambda_t, \beta_t)$. We first focus on DIC_t of terminal node t. It follows that

$$D\left(\bar{\mu}_t, \bar{\lambda}_t, \bar{\beta}_t\right)$$
$$= -2 \log f_{\text{ZICPG}}\left(\boldsymbol{N}_t, \boldsymbol{S}_t \mid \bar{\mu}_t, \bar{\lambda}_t, \hat{\alpha}_t, \bar{\beta}_t\right)$$
$$= -2 \sum_{i=1}^{n_t} \log \left(\frac{1}{1+\bar{\mu}_t w_{ti}} I_{(N_{ti}=0)} + \frac{\bar{\mu}_t w_{ti}}{1+\bar{\mu}_t w_{ti}} \frac{\left(\bar{\lambda}_t u_{ti}\right)^{N_{ti}} e^{-\bar{\lambda}_t u_{ti}}}{N_{ti}!} \right.$$
$$\left. \times \left(\left(\frac{\bar{\beta}_t^{N_{ti}\hat{\alpha}_t} S_{ti}^{N_{ti}\hat{\alpha}_t - 1} e^{-\bar{\beta}_t S_{ti}}}{\Gamma\left(N_{ti}\hat{\alpha}_t\right)} - 1 \right) I_{(N_{ti}>0)} + 1 \right) \right), \tag{21}$$

where

$$\bar{\mu}_t = \frac{\sum_{i=1}^{n_t} \delta_{ti} + \alpha^{(\mu)}}{\sum_{i=1}^{n_t} \phi_{ti} w_{ti} + \beta^{(\mu)}}, \qquad \bar{\lambda}_t = \frac{\sum_{i=1}^{n_t} \delta_{ti} N_{ti} + \alpha^{(\lambda)}}{\sum_{i=1}^{n_t} \delta_{ti} u_{ti} + \beta^{(\lambda)}},$$
$$\bar{\beta}_t = \frac{\sum_{i:N_{ti}>0} N_{ti}\hat{\alpha}_t + \alpha^{(\beta)}}{\sum_{i:N_{ti}>0} S_{ti} + \beta^{(\beta)}}. \tag{22}$$

Furthermore, direct calculations yield the effective number of parameters for terminal node t given by

$$p_{Dt} = 1 - 2\mathbb{E}_{\text{post}}\left(\log f_{\text{ZICPG}}(\boldsymbol{N}_t, \boldsymbol{S}_t, \boldsymbol{\delta}_t, \boldsymbol{\phi}_t \mid \mu_t, \lambda_t, \hat{\alpha}_t, \beta_t)\right)$$
$$+ 2 \log f_{\text{ZICPG}}(\boldsymbol{N}_t, \boldsymbol{S}_t, \boldsymbol{\delta}_t, \boldsymbol{\phi}_t \mid \bar{\mu}_t, \bar{\lambda}_t, \hat{\alpha}_t, \bar{\beta}_t)$$
$$= 1 + 2\left(\log\left(\sum_{i=1}^{n_t} \delta_{ti} + \alpha^{(\mu)} \right) - \psi\left(\sum_{i=1}^{n_t} \delta_{ti} + \alpha^{(\mu)} \right) \right) \sum_{i=1}^{n_t} \delta_{ti}$$
$$+ 2\left(\log\left(\sum_{i=1}^{n_t} \delta_{ti} N_{ti} + \alpha^{(\lambda)} \right) - \psi\left(\sum_{i=1}^{n_t} \delta_{ti} N_{ti} + \alpha^{(\lambda)} \right) \right) \sum_{i=1}^{n_t} \delta_{ti} N_{ti}$$
$$+ 2\left(\log\left(\sum_{i:N_{ti}>0} N_{ti}\hat{\alpha}_t + \alpha^{(\beta)} \right) - \psi\left(\sum_{i:N_{ti}>0} N_{ti}\hat{\alpha}_t + \alpha^{(\beta)} \right) \right)$$
$$\times \sum_{i:N_{ti}>0} N_{ti}\hat{\alpha}_t,$$

and thus $\text{DIC} = \sum_{t=1}^{b} \text{DIC}_t = \sum_{t=1}^{b} \left(D(\bar{\mu}_t, \bar{\lambda}_t, \bar{\beta}_t) + 2p_{Dt} \right)$.

Using these formulas for ZICPG, we can follow the approach presented in Table 1 of Zhang et al. (2024), together with Algorithm 1 (here $\boldsymbol{z}_t = (\boldsymbol{\delta}_t, \boldsymbol{\phi}_t)$), to search for a tree which can then be used for prediction with (8). Given a tree, the estimated pure premium per year in terminal node t is given as

$$\hat{S}_t = \frac{\bar{\mu}_t \bar{\lambda}_t \hat{\alpha}_t}{\bar{\beta}_t (1+\bar{\mu}_t)}, \tag{23}$$

which can be determined using (11) and (22).

Remark 5. We observe that the effective number of parameters does not depend on exposures w_{ti} and u_{ti}, illustrating that the way to embed the exposure does not affect the effective number of parameters. This is intuitively reasonable and is in line with the observations for NB and ZIP models in Zhang et al. (2024).

3.3.2. *Evaluation metrics for joint models*

Note that the ultimate goal in insurance rate-making is to set the pure premium based on the estimate of the aggregate claim amount S. Thus, for joint models, we focus on evaluation metrics defined via the second component S in the bivariate response (N, S). We follow the definitions of M1'–M4' in Subsection 3.1.3; here the number of cells c is the number of terminal nodes b.

Suppose we have obtained a tree with b terminal nodes and corresponding predictions $\hat{S}_t$ $(t = 1, \ldots, b)$ given in (23). Consider a test dataset with m observations. Denote the test data in terminal node t by $\left(\boldsymbol{X}_t, \boldsymbol{v}_t, \boldsymbol{N}_t, \boldsymbol{S}_t\right) = \left((\boldsymbol{x}_{t1}, v_{t1}, N_{t1}, S_{t1}), \ldots, (\boldsymbol{x}_{tm_t}, v_{tm_t}, N_{tm_t}, S_{tm_t})\right)^\top$. The RSS, SE and Lift are defined by M1', M2' and M4' respectively, with c replaced by b. The DS is also similarly defined by M3', but with $\hat{V}_t$ being equal to $\bar{\lambda}_t\hat{\alpha}_t(1+\hat{\alpha}_t)/\bar{\beta}_t^2$ for the CPG model, and $\bar{\mu}_t\bar{\lambda}_t\hat{\alpha}_t(1+\hat{\alpha}_t+\bar{\mu}_t+\hat{\alpha}_t\bar{\mu}_t+\hat{\alpha}_t\bar{\lambda}_t)/\left((1+\bar{\mu}_t)\bar{\beta}_t^2\right)$ for the ZICPG model.

3.4. *Two separate trees versus one joint tree: adjusted rand index*

In this section, we extend our focus to examine the similarity between the BCART generated optimal trees. This exploration will give us confidence and valuable insights into whether information sharing through one joint tree is essential for model accuracy and effectiveness, compared to separate trees.

Measuring the similarity of two trees is generally challenging, particularly when there are variations in the number of terminal nodes or the structure (balanced/unbalanced) of the two trees; see Nye et al. (2006) and the references therein. We propose to explore one simple index commonly employed in cluster analysis comparison, namely, the adjusted Rand Index (ARI) which is a widely recognized metric for assessing the similarity of different clusterings; see, e.g., Rand (1971); Hubert and Arabie (1985); Gates and Ahn (2017). We extend its application to evaluate the similarity of two trees. This is a natural application since a tree generates a partition of the covariate space which automatically induces clusters (i.e. observations belonging to the same leaf) of policyholders in the insurance context.

The ARI measures the similarity between two data partitions by comparing the number of pairwise agreements and disagreements, adjusting for the possibility of random clustering to ensure that the index values are corrected for chance. This results in a score ranging from −1 to 1, where 1 indicates perfect agreement, 0 suggests a similarity no better than random chance, and negative values imply less agreement than expected by chance. The ARI is particularly valued for its ability to account for different cluster sizes and number of clusters, making it a robust metric. Our results use the `adj.rand.index` function in the R package `fossil` (see more details in Vavrek (2020)).

4. Simulation examples

In this section, we investigate the performance of the BCART models introduced in Section 3 by using simulated data. In Scenario 1, the effectiveness of sequential BCART models in capturing the dependence between the number of claims and average severity is examined, along with their performance when using claims count N treated as a numeric variable (or its estimate $\hat{N}$). Full details can be found in Section SM.E of the Supplementary Material; below we only present the simulation framework and conclusion for the sake of brevity. Scenario 2 focuses on the influence of shared information between the number of claims and average severity.

In the sequel, we use the abbreviation Gamma-CART to denote CART for the Gamma model, and other abbreviations can be similarly understood (e.g., ZICPG1-BCART denotes the BCART for ZICPG1 model).

4.1. Scenario 1: varying dependence between the number of claims and average severity

We simulate $\left\{(\boldsymbol{x}_i, v_i, N_i, \bar{S}_i)\right\}_{i=1}^n$ with $n = 5,000$ independent observations. Here $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}\right)$, with independent components $x_{ik} \sim N(0,1)$ for $k = 1,2$. We assume exposure $v_i \equiv 1$ for simplicity. Moreover, $N_i \sim \mathrm{Poi}(\lambda(x_{i1}, x_{i2}))$, where

$$\lambda\left(x_1, x_2\right) = \begin{cases} 1 & \text{if } x_1 x_2 \leq 0, \\ 7 & \text{if } x_1 x_2 > 0. \end{cases} \tag{24}$$

We obtained $N_i = 0$ for 901 occurrences, for which we set $\bar{S}_i = 0$. For the remaining 4099 cases, $\bar{S}_i$ is generated from a gamma distribution with a pre-specified and varying dependence parameter ζ, i.e., $\bar{S}_i \mid N_i \sim \mathrm{Gamma}\left(1, 0.001 + \zeta N_i\right)$, in which the shape (fixed as 1 for simplicity) and rate parameters are chosen to maintain the empirical average claim amount $\bar{S}_i$ to be around 500, aligning with real-world scenarios. The data is split into two subsets: a training set with $n - m = 4,000$ observations and a test set with $m = 1,000$ observations. In this case, our goal is to investigate how the dependence modulated by ζ influences the performance of both frequency-severity models and sequential models, and the performance of incorporating N (or $\hat{N}$) into the sequential models. If N (or $\hat{N}$) is selected as a splitting covariate, it would indicate that the claim count plays an important role in average severity modeling, and thus sequential models should be preferred. As ζ changes, the conditional correlation between the number of claims and average severity varies. Stronger dependence (e.g., $\zeta = 0.001$) is expected to favor sequential models, while weaker dependence (e.g., $\zeta = 0.00001$) makes both model types perform similarly, as N (or $\hat{N}$) is less likely to be used in the sequential model tree.

The observed relatively large DIC differences in training data between the Gamma model without claim count (or its estimate) as a covariate (i.e., Gamma-BCART) and those with it (i.e., Gamma1-BCART or Gamma2-BCART) suggests that incorporating claim count (or its estimate) as a covariate improves performance in the presence of stronger inherent dependence in the data, where Gamma2-BCART (with DIC = 2618) performs best. On test data, Gamma2-BCART also outperforms other models across all evaluation metrics. Additional experiments, repeated with new training and test datasets, confirm the robustness of these results.

Further simulations with different pairs of values for ζ and λ in (24) reveal that: 1) Weak dependence between the number of claims and average severity results in similar performance between frequency-severity and sequential models, in this case the former is preferred for computational efficiency; 2) Strong dependence allows sequential models to perform better by incorporating the number of claims (or its estimate) as a covariate; 3) Gamma2-BCART consistently outperforms Gamma1-BCART, validating our discussion in Section 3.2.

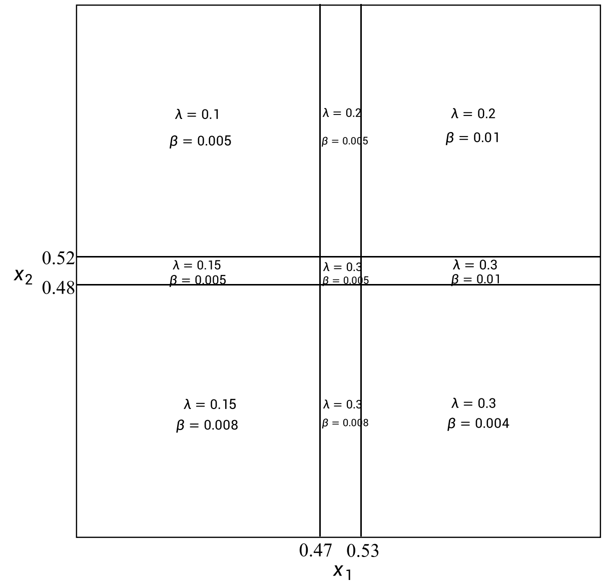

Fig. 1. Covariate space partition for a CPG-distributed simulation. The values of parameters λ and β are provided in each region.

4.2. Scenario 2: covariates sharing between the number of claims and average severity

In this scenario, we consider two simulations where common covariates are used for parameters representing the number of claims and average severity. The objective is to assess the effectiveness of frequency-severity BCART models and joint BCART models, that is, whether it is preferred to share information using one joint tree. To this end, we consider CPG distribution in joint BCART models, and correspondingly, Poisson distribution and gamma distribution involving N as a model weight in the frequency-severity BCART models to keep consistency for comparison. We first explain these two simulations and then present some findings and suggestions.

<u>**Simulation 2.1**</u>: We simulate a dataset $\left\{(\boldsymbol{x}_i, v_i, N_i, \bar{S}_i)\right\}_{i=1}^n$ with $n = 5,000$ independent observations. Here $\boldsymbol{x}_i = \left(x_{i1}, \ldots, x_{i5}\right)$, with independent components $x_{i1} \sim N(0,1)$, $x_{i2} \sim U(-1,1)$, $x_{i3} \sim U(-5,5)$, $x_{i4} \sim N(0,5)$, $x_{i5} \sim U\{1,2,3,4\}$ and $v_i \sim U(0,1)$, where $U(\cdot,\cdot)$ (or $U\{\cdot,\cdot\}$) stands for continuous (or discrete-type) uniform distribution. Moreover, $N_i \sim \mathrm{Poi}(\lambda(x_{i1}, x_{i2}) v_i)$, where

$$\lambda\left(x_1, x_2\right) = \begin{cases} 0.1 & \text{if } x_1 \leq 0.47,\ x_2 > 0.52, \\ 0.2 & \text{if } x_1 > 0.47,\ x_2 > 0.52, \\ 0.3 & \text{if } x_1 > 0.47,\ x_2 \leq 0.52, \\ 0.15 & \text{if } x_1 \leq 0.47,\ x_2 \leq 0.52. \end{cases}$$

If $N_i = 0$ then $\bar{S}_i = 0$, otherwise $\bar{S}_i \sim \mathrm{Gamma}\left(N_i \alpha, N_i \beta\left(x_{i1}, x_{i2}\right)\right)$, where

$$\beta\left(x_1, x_2\right) = \begin{cases} 0.005 & \text{if } x_1 \leq 0.53,\ x_2 > 0.48, \\ 0.01 & \text{if } x_1 > 0.53,\ x_2 > 0.48, \\ 0.004 & \text{if } x_1 > 0.53,\ x_2 \leq 0.48, \\ 0.008 & \text{if } x_1 \leq 0.53,\ x_2 \leq 0.48. \end{cases}$$

For simplicity, we assume $\alpha = 1$, and the values of β are selected such that the average claim amount $\bar{S}_i$ is around 200, which is close to the situation in real-world scenarios. See Fig. 1 for an illustration of the true covariate space partition and corresponding values of parameters.

<u>**Simulation 2.2**</u>: We keep most simulation settings as in Simulation 2.1, except the partition of the covariate space; see Fig. 2. Specifically,

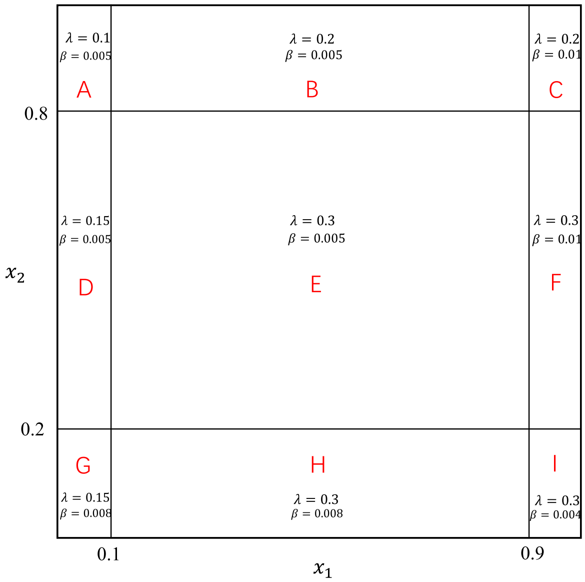

Fig. 2. Covariate space partition for a CPG-distributed simulation. The values of parameters λ and β are provided in each region, which has been labeled with names.

Table 3
Hyper-parameters, p_D and DIC on training data for Simulation 2.1. The number in brackets after the abbreviation of the model indicates the number of terminal nodes for that tree. Bold font indicates DIC selected model.

Model	γ	ρ	p_D	DIC
Poisson-BCART (3)	0.95	15	2.98	3697
Poisson-BCART (4)	0.99	13	3.98	**3572**
Poisson-BCART (5)	0.99	10	4.97	3616
Gamma-BCART (3)	0.95	10	5.97	30586
Gamma-BCART (4)	0.99	10	7.97	**30319**
Gamma-BCART (5)	0.99	8	9.96	30414
CPG-BCART (3)	0.99	5	8.92	34017
CPG-BCART (4)	0.99	4	11.90	**33582**
CPG-BCART (5)	0.99	3	14.89	33711

$$\lambda\left(x_1,x_2\right)=\begin{cases}0.1 & \text{if } x_1\le 0.1,\ x_2>0.8,\\ 0.2 & \text{if } x_1>0.1,\ x_2>0.8,\\ 0.3 & \text{if } x_1>0.1,\ x_2\le 0.8,\\ 0.15 & \text{if } x_1\le 0.1,\ x_2\le 0.8,\end{cases}$$

and for non-zero N_i, generate $\bar{S}_i \sim \text{Gamma}\left(N_i, N_i\beta(x_{i1},x_{i2})\right)$, where

$$\beta\left(x_1,x_2\right)=\begin{cases}0.005 & \text{if } x_1\le 0.9,\ x_2>0.2,\\ 0.01 & \text{if } x_1>0.9,\ x_2>0.2,\\ 0.004 & \text{if } x_1>0.9,\ x_2\le 0.2,\\ 0.008 & \text{if } x_1\le 0.9,\ x_2\le 0.2.\end{cases}$$

The specific design here is that both components of the response variable $(N,\bar{S})$ are affected by the same covariates x_1 and x_2. In Simulation 2.1 they share similar split points, while for Simulation 2.2 they have quite different split points. The variables $x_k, k=3,4,5$ are all noise variables. We aim to compare separate BCART trees versus one joint BCART tree.

Each simulation dataset is split into a training set with $n-m=4{,}000$ observations and a test set with $m=1{,}000$ observations. The outputs from training data and test data for the BCART models are presented in Tables 3-4 for Simulation 2.1 and Tables 5-6 for Simulation 2.2.

We start by looking at the DICs on training data in Tables 3 and 5. For Simulations 2.1 and 2.2, both Poisson-BCART and Gamma-BCART can find the optimal tree with the correct 4 terminal nodes. However, the selected joint CPG-BCART tree for Simulation 2.1 has only 4 terminal nodes which is different from its simulation scheme that should result in a covariate space partition with 9 cells. In contrast, the selected joint CPG-BCART tree for Simulation 2.2 has 9 terminal nodes which is consistent with its simulation scheme. This result for the frequency-severity BCART models is expected, based on our earlier discussion of the separated frequency and average severity models. Now we look into the details of the selected joint tree to explore the reason. For Simulation 2.1, we see that both x_1 and x_2 are used in the tree and the split points for them are close to 0.5 which is the mean of 0.47 and 0.53 for x_1, and also the mean of 0.52 and 0.48 for x_2. Since these split values are very close, it is reasonable for the joint BCART model to select a split value around their mean, resulting in a selected joint tree with 4 terminal nodes. For Simulation 2.2 the selected joint tree includes 9 terminal nodes, which is also reasonable because the split values for both variables are far apart.

Table 4
Model performance on test data with bold entries determined by DIC (see Table 3). F_PS_G denotes the frequency-severity models by using Poisson and gamma distributions separately. The number in brackets after the abbreviation of the model indicates the number of terminal nodes for those trees.

Model	RSS(S) $\times 10^{-8}$	SE	DS	Lift
F_PS_G-BCART (3/3)	3.03	0.1324	0.0833	2.13
F_PS_G-BCART (4/4)	2.89	**0.1245**	**0.0791**	2.21
F_PS_G-BCART (5/5)	2.81	0.1273	0.0812	2.23
CPG-BCART (3)	3.01	0.1319	0.0820	2.15
CPG-BCART (4)	2.84	**0.1211**	**0.0769**	2.27
CPG-BCART (5)	2.78	0.1254	0.0795	2.29

Table 5
Hyper-parameters, p_D and DIC on training data for Simulation 2.2. The number in brackets after the abbreviation of the model indicates the number of terminal nodes for that tree. Bold font indicates DIC selected model.

Model	γ	ρ	p_D	DIC
Poisson-BCART (3)	0.95	15	2.97	3875
Poisson-BCART (4)	0.99	12	3.97	**3669**
Poisson-BCART (5)	0.99	10	4.96	3724
Gamma-BCART (3)	0.95	10	5.97	32156
Gamma-BCART (4)	0.99	10	7.96	**31798**
Gamma-BCART (5)	0.99	8	9.96	31904
CPG-BCART (8)	0.99	5	23.85	36174
CPG-BCART (9)	0.99	3	26.81	**35622**
CPG-BCART (10)	0.99	2	29.79	35781

Table 6
Model performance on test data with bold entries determined by DIC (see Table 5). F_PS_G denotes the frequency-severity models by using Poisson and gamma distributions separately. The number in brackets after the abbreviation of the model indicates the number of terminal nodes for those trees.

Model	RSS(S)$\times 10^{-8}$	SE	DS	Lift
F_PS_G-BCART (3/3)	3.21	0.152	0.091	2.18
F_PS_G-BCART (4/4)	3.04	**0.140**	**0.073**	2.30
F_PS_G-BCART (5/5)	2.95	0.141	0.079	2.32
CPG-BCART (8)	3.23	0.160	0.097	2.15
CPG-BCART (9)	3.08	**0.142**	**0.088**	2.24
CPG-BCART (10)	3.01	0.146	0.090	2.25

The results for test data are shown in Tables 4 and 6. For Simulation 2.1 we see that the joint model performs better than the frequency-severity model, while for Simulation 2.2 the opposite is observed.

In the above, we focused on using evaluation metrics to assess model performance. We now calculate the ARI for these three trees, using the test data. First, for Simulation 2.1 we have

$$\text{ARI(Poisson-BCART (4), Gamma-BCART (4))} = 0.87,$$

Table 7
Description of variables (*dataCar*).

Variable	Description	Type
numclaims (N)	number of claims	numeric
exposure (v)	in yearly units, between 0 and 1	numeric
claimscst0 (S)	total claim amount for each policyholder	numeric
veh_value	vehicle value, in $10,000s	numeric
veh_age	vehicle age category, 1 (youngest), 2, 3, 4	numeric
agecat	driver age category, 1 (youngest), 2, 3, 4, 5, 6	numeric
veh_body	vehicle body, one of: HBACK, UTE, STNWG, HDTOP, PANVN, SEDAN, TRUCK, COUPE, MIBUS, MCARA, BUS, CONVT, RDSTR	character
gender	Female or Male	character
area	coded as A B C D E F	character

Table 8
Numerical summary of the average severity in *dataCar*.

Statistics	Min	Mean	Max	Standard Deviation	Skewness	Kurtosis
Average severity	200	1916	55922	3461	5	48

ARI(Poisson-BCART (4), CPG-BCART (4)) $= 0.94$,

ARI(Gamma-BCART (4), CPG-BCART (4)) $= 0.92$.

This confirms the preference of joint models in Simulation 2.1, as the ARI values indicate strong similarity. This suggests that information sharing can avoid redundant use of similar information, which is evident in the similarities between the two trees. Next, for Simulation 2.2 we have

ARI(Poisson-BCART (4), Gamma-BCART (4)) $= 0.28$,

ARI(Poisson-BCART (4), CPG-BCART (9)) $= 0.77$,

ARI(Gamma-BCART (4), CPG-BCART (9)) $= 0.73$.

This indicates that the frequency-severity model should be preferred in Simulation 2.2 as the ARI values are smaller, especially the first. It is not obvious how to determine a specific ARI threshold that indicates when sharing information becomes worthwhile. This requires further research.

Building on the above findings in Simulation 2.2, our investigation shows that both the frequency-severity model and joint model identify the optimal trees as expected, indicating that information sharing may not be necessary. Further exploration reveals that the parameter estimates of the joint tree are not as accurate as those for the frequency-severity trees. This discrepancy arises because the joint tree uses less data for the estimates in some of the 9 terminal notes, compared to the separate two trees, each of which only has 4 terminal nodes. We suspect that, with fewer observations the differences will increase, and vice versa. To investigate this intuition, we use the same data generation scheme with different sample sizes (ranging from 1,000 to 50,000) to conduct 10 repeated experiments for each size. We find that as the amount of data increases, the differences between the two model estimates become smaller (see Supplementary Material SM.F for details). This investigation suggests if less data is available the frequency-severity models may be preferred as they produce more accurate parameter estimates than the joint models, while if more data is available the joint models may be preferred to save computation time.

We also run several other simulation examples which are not shown here. From our results, we conclude that: when two trees have similar splitting rules (high ARI), one joint tree is more effective through information sharing. Conversely, if all covariates affecting claim frequency and average severity are different (ARI is close to -1), two trees outperform one joint tree. This conclusion aligns with our intuition and can be generalized to a wider field; see also Linero et al. (2020).

5. Real data analysis

We illustrate our methodology with the insurance dataset *dataCar*, available from the library `insuranceData` in R; see Wolny–Dominiak and Trzesiok (2014) for details. This dataset is based on one-year vehicle insurance policies taken out in 2004 or 2005. There are 67,856 policies of which 93.19% made no claims. A description of the variables is given in Table 7. We split this dataset into training (80%) and test (20%) datasets such that the proportion of non-zero claims remains the same in both training and test datasets.

5.1. Average severity modeling

For average severity modeling, we consider a subset of the data with $N > 0$. Among all 67,856 policies, 4,624 policies satisfy this requirement (3,699 in the training data, and 925 in the test data). We calculate the average severity by dividing the total claim amount by the number of claims for each policyholder. A numerical summary of the average severity data is displayed in Table 8, indicating that the average severity data exhibit right-skewness and heavy tails. We start with some exploratory analysis, fitting gamma, lognormal and Weibull distributions to the whole data. From the histogram and Q-Q plots (see Supplementary Material SM.G), we find that all distributions can capture the right-skewed feature, however, none of them correctly captures the heavy right tail of the distribution. It appears that the lognormal distribution fits slightly better when the data are treated as IID. However, as we will see below the lognormal distribution is not the best choice in the BCART models for this data.

For comparison, we first run some benchmark CART models. We have tried to fit the ANOVA-CART using both the original and log-transformed training data. Neither of them gives us any reasonable result since no split is identified, resulting in only a root node tree. We use the R package `distRforest` (cf. Henckaerts (2020)) to fit Gamma-CART and LN-CART and both of the trees, after cost-complexity pruning, have 5 terminal nodes. As far as we are aware there is no R package with the Weibull distribution implemented for regression trees.

We then apply the proposed Gamma-BCART, LN-BCART, and Weib-BCART to the same data. The DICs in Table 9 indicate that all these BCART models choose a tree with 4 terminal nodes.

We also examine the splitting rules used in each tree. Gamma-CART uses both "*agecat*" and "*veh_value*" twice, with the first one being "*agecat*". In contrast, LN-CART uses three different variables, "*veh_value*" first, followed by "*veh_body*" and "*area*". All trees from BCART models, i.e., Gamma-BCART, LN-BCART, and Weib-BCART, have the same tree structure and splitting variables ("*agecat*", "*veh_value*", and "*area*"), while the split values/categories are slightly different. Weib-BCART, in particular, can identify a more risky cell (i.e., the one with estimated average severity equal to 2743.41); see Fig. 3. This may be because, as discussed in Subsection 3.1.2, Weib-BCART can flexibly control the shape parameter to adapt to data with different tail characteristics, allowing it to handle cases where some cells (terminal nodes) have lighter tails, and others have heavier tails. We generate Q-Q plots of the opti-

Table 9
Hyper-parameters, p_D and DIC on training data (*dataCar*), and model performance on test data for average severity models with bold entries determined by DIC. The number in brackets after the abbreviation of the model indicates the number of terminal nodes for this tree.

	Training data				Test data			
Model	γ	ρ	p_D	DIC	RSS($\boldsymbol{S}$) $\times 10^{-10}$	SE	DS	Lift
Gamma-GLM	-	-	-	-	1.4335	-	-	-
Gamma-CART (5)	-	-	-	-	1.4173	464	0.00171	1.625
LN-CART (5)	-	-	-	-	1.4168	458	0.00168	1.629
Gamma-BCART (3)	0.99	4	5.97	78061	1.4201	486	0.00181	1.567
Gamma-BCART (4)	0.99	3.5	7.97	**77779**	1.4176	**457**	**0.00154**	1.615
Gamma-BCART (5)	0.99	2	9.95	77982	1.4158	472	0.00167	1.643
LN-BCART (3)	0.99	5	5.97	78022	1.4193	483	0.00178	1.570
LN-BCART (4)	0.99	4	7.97	**77741**	1.4171	**449**	**0.00149**	1.628
LN-BCART (5)	0.99	3	9.96	77893	1.4153	456	0.00161	1.649
Weib-BCART (3)	0.99	7	5.98	77932	1.4177	473	0.00164	1.604
Weib-BCART (4)	0.99	5	7.98	**77646**	1.4154	**433**	**0.00131**	1.661
Weib-BCART (5)	0.99	4	9.98	77821	1.4136	446	0.00144	1.693

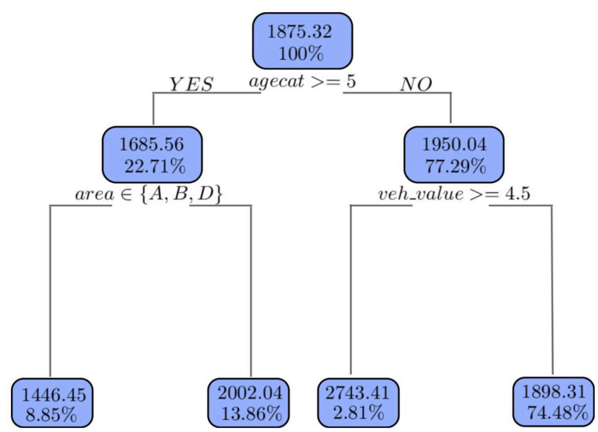

Fig. 3. Optimal tree from Weib-BCART. Numbers at each node give the estimated average severity and the percentage of observations.

mal Weib-BCART tree for average severity data in each terminal node (see Supplementary Material SM.G). The plots reveal that although all shape parameters are smaller than one, indicating heavy tails for average severity data within each terminal node, the selected Weib-BCART tree shows improved data fitting compared to the initial model where covariates are not taken into account. Similar improvements are obtained from Gamma-BCART and LN-BCART. We also use a standard Gamma-GLM for this dataset. We find that only the variable "*gender*" is significant, and thus no interaction is considered in the Gamma-GLM. Interestingly, "*gender*" does not appear in any of the CART and BCART models. In summary, though the variables used for different models may differ, there seems to be a consensus that "*agecat*" is still significantly important for average severity modeling, as Gamma-CART and all BCART models use it in the first split, and "*veh_value*" is another relatively important variable. This observation aligns, to some extent, with our initial analysis of the relationship between covariates and average severity; see Table 12 below. Particularly, in comparison to CARTs, BCART models reveal another important variable, "*area*".

The performance on the test data of the selected tree models above is given in Table 9. It is evident that the Gamma-GLM is not as good as the tree models, as reflected in RSS($\boldsymbol{S}$). The study yields the following model ranking: Weib-BCART, LN-BCART, Gamma-BCART, LN-CART, Gamma-CART. This ranking is consistent with our expectations. First, it is common that average severity data is heavy-tailed. Second, the Weibull distribution is advantageous, because it can effectively handle varying tail characteristics in different tree nodes.

5.2. Aggregate claim modeling

5.2.1. Model fitting and comparison

We now fit the three BCART models with aggregate claim data, namely, frequency-severity models, sequential models, and joint models. For frequency-severity models, numerous combinations of claim frequency and average severity models are possible; see Zhang et al. (2024) for the frequency models and Section 5.1 for average severity models. Here, we choose ZIP2-BCART and Weib-BCART as the optimal tree models for frequency-severity models (see Zhang et al. (2024) and Section 5.1). Note that although ZIP2-BCART and Weib-BCART are identified as the best for claim frequency and average severity separately, it is unclear whether they remain optimal when combined. This will be examined and discussed below. For joint models, we discuss the CPG-BCART model and three types of ZICPG-BART models. Because of these choices, we also include the frequency-severity BCART model with Poisson and gamma distributions for comparison. For sequential BCART models, we consider Poisson-BCART (or called P-BCART) and ZIP2-BCART for claim frequency. Subsequently, we treat the claim count N (or $\hat{N}$) as a covariate in the corresponding Gamma-BCART and Weib-BCART for average severity. The resulting models are called Gamma1-BCART (or Gamma2-BCART, with $\hat{N}$ from P-BCART) and Weib1-BCART (or Weib2-BCART, with $\hat{N}$ from ZIP2-BCART).

Table 10 presents the DICs for the average severity part in the sequential models and for the joint models. We see that in the average severity modeling with N (or $\hat{N}$) as a covariate, all of them choose an optimal tree with 4 terminal nodes. Upon inspecting the tree structure, N (or $\hat{N}$) is indeed used in the first step in all those optimal trees. All of them replace the previously used variable "*agecat*" by the covariate N (or $\hat{N}$). We suspect this may be due to a strong relationship between the covariates N and "*agecat*", as verified in the claim frequency analysis (see Zhang et al. (2024) and also Table 12 below). Furthermore, by comparing the DICs of all Gamma-BCART and Weib-BCART models in Tables 9 and 10 (with/without N or $\hat{N}$ as a covariate), we find that the model performance improves when considering N (or $\hat{N}$) as a covariate, especially when using $\hat{N}$.

For joint models, i.e., CPG-BCART and three ZICPG-BCART models, all of them choose optimal trees with 5 terminal nodes. Among them, ZICPG3-BCART, with the smallest DIC (= 102120), is deemed to be the best. We again examine the splitting rules used in the selected trees among joint models. All trees use the same splitting variables ("*agecat*", "*veh_value*", "*veh_body*", and "*area*"), but the order of use and the tree structures vary. Notably, "*agecat*" is consistently the first variable. Among them, ZICPG3-BCART demonstrates the ability to identify a riskier cell (i.e., the one with an estimated pure premium equal to 657.45; see Fig. 4), possibly due to the same reason as discussed in Zhang et al. (2024) for the outstanding performance of ZIP2-BCART for claim

Table 10
Hyper-parameters, p_D and DIC on training data (*dataCar*) for aggregate claim models. The number in brackets after the abbreviation of the model indicates the number of terminal nodes for this tree. The Gamma1/Weib1 and Gamma2/Weib2 models treat the claim count N and $\hat{N}$ as a covariate respectively, where $\hat{N}$ for Gamma2 comes from P-BCART and that for Weib2 comes from ZIP2-BCART. Bold font indicates DIC selected model.

Model	γ	ρ	p_D	DIC
Gamma1-BCART (3)	0.99	4	5.97	78032
Gamma1-BCART (4)	0.99	3.5	7.97	**77750**
Gamma1-BCART (5)	0.99	2	9.96	77854
Gamma2-BCART (3)	0.99	4	5.98	78024
Gamma2-BCART (4)	0.99	3.5	7.97	**77743**
Gamma2-BCART (5)	0.99	2	9.97	77849
Weib1-BCART (3)	0.99	7	5.98	77911
Weib1-BCART (4)	0.99	5	7.98	**77619**
Weib1-BCART (5)	0.99	4	9.98	77804
Weib2-BCART (3)	0.99	7	5.98	77893
Weib2-BCART (4)	0.99	5	7.98	**77608**
Weib2-BCART (5)	0.99	4	9.98	77787
CPG-BCART (4)	0.99	10	11.96	105710
CPG-BCART (5)	0.99	8	14.93	**105626**
CPG-BCART (6)	0.99	7	17.92	105643
ZICPG1-BCART (4)	0.99	11	15.97	102314
ZICPG1-BCART (5)	0.99	10	19.95	**102198**
ZICPG1-BCART (6)	0.99	7.5	23.92	102225
ZICPG2-BCART (4)	0.99	12	15.95	102265
ZICPG2-BCART (5)	0.99	11	19.94	**102134**
ZICPG2-BCART (6)	0.99	8	23.92	102167
ZICPG3-BCART (4)	0.99	14	15.94	102247
ZICPG3-BCART (5)	0.99	12	19.93	**102120**
ZICPG3-BCART (6)	0.99	9	23.90	102158

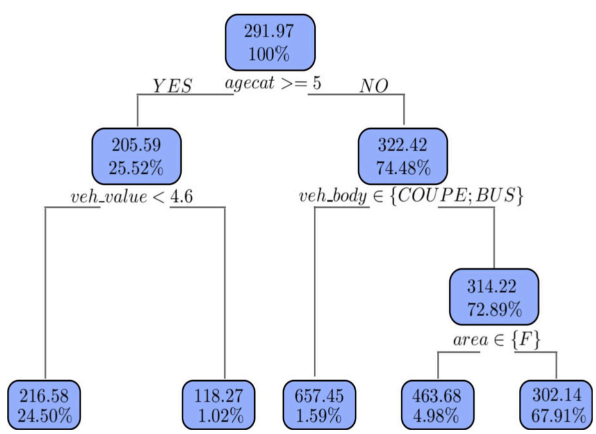

Fig. 4. Optimal tree from ZICPG3-BCART. Numbers at each node give the estimated premium and the percentage of observations.

frequency. Besides, we observe that the tree structure of ZICPG3-BCART is quite similar to ZIP2-BCART. However, ZICPG3-BCART identifies another important variable "*area*", which was recognized as important for average severity before (see Section 5.1). We also fit a CPG-GLM to the data. We find that only the variable "*agecat*" is significant, aligning with its consistent selection as the first splitting variable in almost all the BCART models. It is also worth mentioning that CART is not included in this analysis due to the absence of R packages that can directly use CPG (or ZICPG) to process the data.

The performance of the selected trees for the test data is given in Table 11. As before, GLM exhibits poorer performance compared to tree models, as evidenced by RSS($\boldsymbol{S}$). Below we discuss the three types of

Table 11
Model performance on test data (*dataCar*) for aggregate claim models with bold entries determined by DIC (see Table 10). $F_P S_G$ denotes frequency-severity models using Poisson and gamma distributions separately; other abbreviations can be explained similarly referring to Table 10. The number in brackets after the abbreviation of the model indicates the number of terminal nodes for those trees.

Model	RSS ($\boldsymbol{S}$)$\times 10^{-10}$	SE	DS $\times 10^4$	Lift
CPG-GLM	1.5187	-	-	-
$F_P S_G$-BCART (5/4)	1.4874	242.12	8.32	2.532
$F_P S_{G1}$-BCART (5/4)	1.4796	237.51	8.17	2.540
$F_P S_{G2}$-BCART (5/4)	1.4792	237.44	8.16	2.541
$F_{ZIP2} S_{Weib}$-BCART (5/4)	1.4844	240.89	8.27	2.534
$F_{ZIP2} S_{Weib1}$-BCART (5/4)	1.4783	236.03	8.01	2.545
$F_{ZIP2} S_{Weib2}$-BCART (5/4)	1.4770	235.83	7.92	2.549
CPG-BCART (4)	1.4791	237.42	8.15	2.542
CPG-BCART (5)	1.4781	**235.98**	**7.93**	2.547
CPG-BCART (6)	1.4778	236.29	8.02	2.549
ZICPG1-BCART (4)	1.4670	232.87	7.85	2.560
ZICPG1-BCART (5)	1.4497	**229.73**	**7.56**	2.584
ZICPG1-BCART (6)	1.4478	231.35	7.79	2.587
ZICPG2-BCART (4)	1.4612	232.15	7.81	2.563
ZICPG2-BCART (5)	1.4434	**229.41**	**7.52**	2.595
ZICPG2-BCART (6)	1.4417	231.10	7.77	2.597
ZICPG3-BCART (4)	1.4598	231.24	7.79	2.570
ZICPG3-BCART (5)	1.4415	**228.88**	**7.45**	2.601
ZICPG3-BCART (6)	1.4409	229.53	7.69	2.604

aggregate claim models from various perspectives. The meaning of the abbreviations can be found in the captions of Tables 10-11.

1. A comparison of our frequency-severity models suggests using the combination of two best models for claim frequency and average severity respectively based on all evaluation metrics, i.e., $F_{ZIP2}S_{Weib}$-BCART > $F_P S_G$-BCART. In the sequential models, the same conclusion as in Section 4.1 is reached: using the estimate of the claim count $\hat{N}$ is superior to using N itself when treating them as a covariate in the average severity tree. Regarding joint models, ZICPG models outperform the CPG model, with ZICPG3-BCART being the best.
2. When comparing frequency-severity models and sequential models, it is evident that adding N (or $\hat{N}$) as a covariate improves performance, as shown by all evaluation metrics, i.e., $F_P S_{G2}$-BCART > $F_P S_{G1}$-BCART > $F_P S_G$-BCART. The same ranking is observed for another combination, i.e., $F_{ZIP2}S_{Weib2}$-BCART > $F_{ZIP2}S_{Weib1}$-BCART > $F_{ZIP2}S_{Weib}$-BCART. This is reasonable, as real data often exhibit some correlation between the number of claims and average severity, favoring sequential models that consider this feature directly over frequency-severity models assuming independence.
3. In comparing frequency-severity models and joint models, all evaluation metrics indicate that the optimal CPG-BCART (or ZICPG-BCART) chosen by DIC consistently outperforms frequency-severity models, suggesting that sharing information is beneficial for this dataset, i.e., one joint tree exhibits better performance. Exploration of the reasons is provided below.
4. As for sequential models and joint models, they address dependence in different ways. The former uses two trees, treating the number of claims (or its estimate) as a covariate in average severity modeling to address the dependence. In contrast, the latter uses one joint tree, potentially hiding some dependence in the common variables used to split the nodes and incorporating the number of claims as a model weight in the aggregate claim amount distribution. Joint models employing ZICPG distributions perform better than all the sequential models as demonstrated by all evaluation metrics, possibly due to a small negative conditional correlation between the

Table 12
Correlation coefficients between covariates (numerical ones and transformed categorical ones) and claim frequency and average severity. Bold font indicates the strongest correlation (though generally small) in each row.

	veh_value	veh_age	agecat	veh_body	gender	area
Claim frequency	−0.0047	0.0013	**−0.0131**	0.0022	0.0008	−0.0021
Average severity	0.0135	−0.0059	**−0.0274**	−0.0035	0.0003	−0.0034

Table 13
Values of ARI between different trees. The number in brackets after the abbreviation of the model indicates the number of terminal nodes.

	Poisson-BCART (5)	ZIP2-BCART (5)	Gamma-BCART (4)	Weib-BCART (4)	CPG-BCART (5)	ZICPG3-BCART (5)
Poisson-BCART (5)	1	0.7396	0.5398	0.5163	0.8585	0.7823
ZIP2-BCART (5)	-	1	0.5599	0.5179	0.8587	0.8152
Gamma-BCART (4)	-	-	1	0.7430	0.6921	0.6423
Weib-BCART (4)	-	-	-	1	0.6491	0.6022
CPG-BCART (5)	-	-	-	-	1	0.6351
ZICPG3-BCART (5)	-	-	-	-	-	1

number of claims and average severity (−0.0336) and the dataset involving a high proportion of zeros (93.19%).

Remark 6. Our primary motivation for the joint models was to incorporate the association between the number of claims and average severity through splitting variables. To evaluate its effectiveness, we examine the conditional dependence within each terminal node of the selected joint tree from ZICPG3-BCART. We identify two cells of positive conditional correlation (out of a partition consisting of 5 cells), occurring when the estimated premium is at its minimum and maximum levels, respectively. This finding reveals the nuanced dependence between the number of claims and average severity, emphasizing the role of covariates, and provides valuable implications for insurance pricing and risk management.

5.2.2. *Information sharing - why is it beneficial?*

From the above data analysis, we have seen the advantages of being able to share information between claim frequency and severity (through common covariates in the tree) in the joint modeling of this dataset. In this section, we will further investigate the superior performance of the joint models by looking at the correlation between covariates and claim frequency and average severity.

The correlation coefficients are displayed in Table 12. Note that some of the covariates are categorical, for which we use some transformations to replace them with numerical values for the correlation calculation; see Supplementary Material SM.H for further details. For claim frequency, variables with the strongest correlation coefficients include "*agecat*" and "*veh_value*" which are used in the tree selected by the frequency BCART models in Zhang et al. (2024). For average severity, we observe the same two variables exhibit the strongest correlation. Furthermore,"*veh_age*" has the third strongest correlation, but the selected average severity tree (cf. Fig. 3) does not include it. We suspect this is due to a strong relationship between covariates "*veh_age*" and "*veh_value*" (see Supplementary Material SM.H) and thus one of them is dropped to avoid multicollinearity. Additionally, both claim frequency and average severity show the strongest correlation with *"agecat"* which is the first splitting variable in the selected optimal frequency, average severity and joint trees, illustrating the effectiveness of BCART models for variable selection. The above discussion suggests that some common covariates exhibit strong relationships with both claim frequency and average severity, it is thus beneficial to share this information using joint modeling. This validates the conclusion of Subsection 5.2.1.

Next, we consider ARI (see Section 3.4) to assess the similarity between different trees and determine whether information sharing is necessary; see Table 13. As discussed in Section 4.2, although a specific threshold of ARI for making a direct judgment about the necessity for information sharing is unknown, it is evident that ARI values between all claim frequency and average severity trees are greater than 0.5. This suggests that significant intrinsic similarities of these models cannot be ignored. This also validates the preference of adopting a joint model from Subsection 5.2.1.

After conducting a thorough analysis of this dataset, we suggest that insurers need to pay more attention to policyholders who are younger and have vehicles with higher values since they are more likely to have higher risks.

Remark 7. We have also applied the proposed BCART models to other datasets (*dataOhlsson* included in the library `insuranceData` in R, *freMTPL2freq* and *freMTPL2sev* included in the library `CASdatasets`; see more details in Charpentier (2014)). Due to the similarity in analysis methods and the consistency of conclusions, we omit the details.

6. Summary

This work develops BCART models for insurance pricing, building upon the foundation of previous claim frequency analysis (see Zhang et al. (2024)). In particular, for average severity, we incorporated BCART models with gamma, lognormal, and Weibull distributions, which have different abilities to handle data with varying tail characteristics. We found that the Weib-BCART performs better than Gamma-BCART or LN-BCART since it can deal with cases where some cells have lighter tails, while others have heavier tails. Besides, in the comparison between Gamma-BCART and LN-BCART, the former is preferable for data with a lighter tail and the latter is more suitable for data with a heavier tail. This finding provides us with a practical strategy for choosing models. Concerning aggregate claim modeling, we proposed three types of models. First, we found that the sequential models treating the number of claims (or its estimate) as a covariate in the average severity modeling perform better than the standard frequency-severity models when the underlying true dependence between the number of claims and average severity is stronger. Second, we explored the choice between using two trees or one joint tree. In particular, when there are indeed common covariates affecting claim frequency and average severity and there is a relatively large amount of data available, it may be beneficial to use one joint tree, which supports the conclusion in Linero et al. (2020) by illuminating the benefits of information sharing. Third, we provided details of evaluation metrics in the case of two trees and proposed the use of ARI to quantify the similarity between two trees, which can assist in explaining the necessity of information sharing. Finally, in the analysis of various joint models, especially the three ZICPG models employing different ways to embed exposure, we found that ZICPG3-BCART, which embeds exposure in both the zero mass component and Poisson component, delivers the most favorable results in real insurance data.

We address their similarities to the analysis of ZIP models discussed in Zhang et al. (2024). Furthermore, we introduced a more general MCMC algorithm for BCART models. These enhancements extend the applicability of BCART models to a broader range of applications. Although this paper does not fully resolve the challenge of dependence modeling between the number of claims and average severity, it highlights the importance and a need for further research, with copula-based approaches emerging as a promising direction.

CRediT authorship contribution statement

Yaojun Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Lanpeng Ji:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Georgios Aivaliotis:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Charles C. Taylor:** Supervision, Methodology, Investigation, Formal analysis, Conceptualization, Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the writing process

There is no use of generative AI and AI-assisted technologies in the writing process.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We are thankful to the anonymous referees for their constructive suggestions which have led to a significant improvement of the manuscript.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.insmatheco.2025.103136.

Data availability

Data used are included in software R packages.

References

Ahn, J.Y., Fuchs, S., Oh, R., 2021. A copula transformation in multivariate mixed discrete-continuous models. Fuzzy Sets Syst. 415, 54–75.

Baumgartner, C., Gruber, L.F., Czado, C., 2015. Bayesian total loss estimation using shared random effects. Insur. Math. Econ. 62, 194–201.

Blier-Wong, C., Cossette, H., Lamontagne, L., Marceau, E., 2020. Machine learning in P&C insurance: a review for pricing and reserving. Risks 9, 4.

Blier-Wong, C., Cossette, H., Marceau, E., 2024. Collective risk models with FGM dependence. Scand. Actuar. J., 1–29.

Celeux, G., Forbes, F., Robert, C.P., Titterington, D.M., 2006. Deviance information criteria for missing data models. Bayesian Anal. 1, 651–673.

Charpentier, A., 2014. Computational Actuarial Science with R. CRC Press, Boca Raton, FL.

Chipman, H., George, E., McCulloch, R., 2003. Bayesian treed generalized linear models. Bayesian Stat. 7, 323–349.

Chipman, H.A., George, E.I., McCulloch, R.E., 1998. Bayesian CART model search. J. Am. Stat. Assoc. 93, 935–948.

Cossette, H., Marceau, E., Mtalai, I., 2019. Collective risk models with dependence. Insur. Math. Econ. 87, 153–168.

Czado, C., Kastenmeier, R., Brechmann, E.C., Min, A., 2012. A mixed copula model for insurance claims and claim sizes. Scand. Actuar. J. 2012, 278–305.

Delong, Ł., Lindholm, M., Wüthrich, M.V., 2021. Making Tweedie's compound Poisson model more accessible. Eur. Actuar. J. 11, 185–226.

Denison, D.G., Mallick, B.K., Smith, A.F., 1998. A Bayesian CART algorithm. Biometrika 85, 363–377.

Denuit, M., Trufin, J., 2019. Effective Statistical Learning Methods for Actuaries. Springer.

Denuit, M., Charpentier, A., Trufin, J., 2021. Autocalibration and Tweedie-dominance for insurance pricing with machine learning. Insur. Math. Econ. 101, 485–497.

Farkas, S., Lopez, O., Thomas, M., 2021. Cyber claim analysis using generalized Pareto regression trees with applications to insurance. Insur. Math. Econ. 98, 92–105.

Fink, D., 1997. A compendium of conjugate priors. [Online]. https://www.johndcook.com/CompendiumOfConjugatePriors.pdf. (Accessed August 2024).

Frees, E.W., Gao, J., Rosenberg, M.A., 2011. Predicting the frequency and amount of health care expenditures. N. Am. Actuar. J. 15, 377–392.

Frees, E.W., Lee, G., Yang, L., 2016. Multivariate frequency-severity regression models in insurance. Risks 4, 4.

Gao, G., 2024. Fitting Tweedie's compound Poisson model to pure premium with the EM algorithm. Insur. Math. Econ. 114, 29–42.

Garrido, J., Genest, C., Schulz, J., 2016. Generalized linear models for dependent frequency and severity of insurance claims. Insur. Math. Econ. 70, 205–215.

Gates, A.J., Ahn, Y.-Y., 2017. The impact of random models on clustering similarity. arXiv preprint arXiv:1701.06508.

Gschlößl, S., Czado, C., 2007. Spatial modelling of claim frequency and claim size in non-life insurance. Scand. Actuar. J. 2007, 202–225.

Henckaerts, R., 2020. DistRforest: random forests with distribution-based loss functions. [Online]. Available: https://henckr.github.io/distRforest/.

Henckaerts, R., Côté, M.-P., Antonio, K., Verbelen, R., 2021. Boosting insights in insurance tariff plans with tree-based machine learning methods. N. Am. Actuar. J. 25, 255–285.

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193–218.

Jørgensen, B., Paes De Souza, M.C., 1994. Fitting Tweedie's compound Poisson model to insurance claims data. Scand. Actuar. J. 1994, 69–93.

Krämer, N., Brechmann, E.C., Silvestrini, D., Czado, C., 2013. Total loss estimation using copula-based regression models. Insur. Math. Econ. 53 (3), 829–839.

Lian, Y., Yang, A.Y., Wang, B., Shi, P., Platt, R.W., 2023. A Tweedie compound Poisson model in reproducing kernel Hilbert space. Technometrics 65, 281–295.

Linero, A.R., 2017. A review of tree-based Bayesian methods. Commun. Stat. Appl. Methods 24, 543–559.

Linero, A.R., Sinha, D., Lipsitz, S.R., 2020. Semiparametric mixed-scale models using shared Bayesian forests. Biometrics 76, 131–144.

Mehmet, M., Saykan, Y., 2005. On a bonus-malus system where the claim frequency distribution is geometric and the claim severity distribution is Pareto. Hacet. J. Math. Stat. 34, 75–81.

Murray, J.S., 2021. Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. J. Am. Stat. Assoc. 116, 756–769.

Nye, T.M., Lio, P., Gilks, W.R., 2006. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics 22, 117–119.

Oh, R., Ahn, J.Y., Lee, W., 2021. On copula-based collective risk models: from elliptical copulas to vine copulas. Scand. Actuar. J. 2021 (1), 1–33.

Ohlsson, E., Johansson, B., 2010. Non-life Insurance Pricing with Generalized Linear Models. Springer.

Quan, Z., Wang, Z., Gan, G., Valdez, E.A., 2023. On hybrid tree-based methods for short-term insurance claims. Probab. Eng. Inf. Sci. 37 (2), 597–620.

Quijano Xacur, O.A., Garrido, J., 2015. Generalised linear models for aggregate claims: to Tweedie or not? Eur. Actuar. J. 5, 181–202.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. 66, 846–850.

Rinne, H., 2008. The Weibull Distribution: a Handbook. CRC Press.

Rocková, V., Van der Pas, S., et al., 2020. Posterior concentration for Bayesian regression trees and forests. Ann. Stat. 48, 2108–2131.

Shi, P., Zhao, Z., 2020. Regression for copula-linked compound distributions with applications in modeling aggregate insurance claims. Ann. Appl. Stat. 14, 357–380.

Shi, P., Feng, X., Ivantsova, A., 2015. Dependent frequency-severity modeling of insurance claims. Insur. Math. Econ. 64, 417–428.

Siswadi, Quesenberry, C., 1982. Selecting among weibull, lognormal and gamma distributions using complete and censored smaples. Nav. Res. Logist. Q. 29, 557–569.

Smyth, G.K., Jørgensen, B., 2002. Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. ASTIN Bull. 32, 143–157.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc., Ser. B Stat. Methodol. 64, 583–639.

Um, S., Linero, A.R., Sinha, D., Bandyopadhyay, D., 2023. Bayesian additive regression trees for multivariate responses. Stat. Med. 42, 246–263. https://doi.org/10.1002/sim.9613. [Online]. Available: https://par.nsf.gov/biblio/10403199.

Vavrek, M.J., 2020. Fossil: palaeoecological and palaeogeographical analysis tools, R package version 0.4.0. [Online]. Available: https://CRAN.R-project.org/package=fossil.

Wolny–Dominiak, A., Trzesiok, M., 2014. InsuranceData: a collection of insurance datasets useful in risk classification in non-life insurance, R package version 1.0. [Online]. Available: https://CRAN.R-project.org/package=insuranceData.

Wuthrich, M., Merz, M., 2022. Statistical Foundations of Actuarial Learning and Its Applications. Springer Actuarial. https://link.springer.com/book/10.1007/978-3-031-12409-9.

Wuthrich, M.V., 2022. Non-life Insurance: Mathematics & Statistics. Available at SSRN 2319328.

Wuthrich, M.V., Buser, C., 2022. Data analytics for non-life insurance pricing (January 9, 2023). Available at SSRN: https://ssrn.com/abstract=2870308.

Yang, Y., Qian, W., Zou, H., 2018. Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. J. Bus. Econ. Stat. 36, 456–470.

Zhang, Y., 2024. Insurance Pricing using Bayesian Tree Models. PhD thesis. University of Leeds.

Zhang, Y., Ji, L., Aivaliotis, G., Taylor, C., 2024. Bayesian CART models for insurance claims frequency. Insur. Math. Econ. 114, 108–131.

Zhou, H., Qian, W., Yang, Y., 2022. Tweedie gradient boosting for extremely unbalanced zero-inflated data. Commun. Stat., Simul. Comput. 51, 5507–5529.