USING OPEN FILES FOR INDIVIDUAL LOSS RESERVING IN PROPERTY AND CASUALTY INSURANCE

A PREPRINT

Mathieu Pigeon
Département de mathémathiques
Université du Québec à Montréal (UQAM)
Montreal, QC, Canada
pigeon.mathieu.2@uqam.ca

Hélène Cossette
École d'actuariat
Université Laval
Quebec City, QC, Canada
helene.cossette@act.ulaval.ca

October 25, 2021

ABSTRACT

Machine learning techniques have recently shown to being an interesting and effective tool to be used in the establishment of loss reserves on an individual basis. These new approaches to tackle the reserving task pose some challenges. After a brief literature review on the recent research works on the use of machine learning strategies in a loss reserving context, we propose in this project to investigate in detail the tree-based methodologies of Lopez *et al.* (2016, 2019) and Duval and Pigeon (2019) to include open files within the loss reserving process. We compare their performance with some classical aggregate approaches (Mack's Chain-Ladder, generalized linear models, etc.) through a simulation study. We also begin to study the possible influence of dependency structures within and between business lines.

1 Introduction

Non-life insurers face high volatility due to the nature of losses they must provide coverage for. Regulation thus requires them to maintain funds under solvency constraints to ensure that up to a certain risk level, there will be no solvency issues. There exist very specific guidelines to determine the capital requirement to be maintained by an insurer varying from one country and even one state/province to another. For Property and Casualty insurance companies, United States regulation as defined by the National Association of Insurance Commissioners uses risk-based capital requirements (see Feldblum (1996)), while the Canadian requirement is set forth as the conditional tail expectation (CTE) at a 99 % level for insurance risk (OSFI , 2018).

There are many classical (or aggregate) methods to evaluate reserves, see Wüthrich and Merz (2008) and Friedland (2010) for an extensive discussion of existing methods. Recently, several approaches based on statistical learning techniques have been proposed for individual (or granular, or micro-level) loss reserving. These generally presuppose the availability of many closed files, i.e. files for which the entire development of the claim – from the occurrence until the final closure of the file – is known. In practice, this assumption is never verified, and the actuary must include open files in the modeling process. In this work, these open files will be considered as censored data, i.e. files for which the development is not fully known. We plan to analyze some challenges that this situation poses as well as various strategies to overcome them. More specifically, the main objective of our project is **to analyze how open claims should be integrated into an individual reserve valuation process when statistical learning-based approaches are used**. We present in detail and investigate the tree-based strategies proposed in Lopez *et al.* (2016, 2019) and Duval and Pigeon (2019) to include open files within the loss reserving process. With the use of sampled data sets, we compare their performance to classical aggregate loss reserving methods and study the possible impacts of copula-based dependence structures within and between lines of business.

In Section 2, we present a short review of the actuarial literature on individual loss reserving focusing on the use of statistical learning tools. In Section 3, we define both, individual and collective frameworks for loss reserving. In Section 4, we present various strategies to include open files in the modeling process. More specifically, we detail approaches proposed in Lopez *et al.* (2016) and Duval and Pigeon (2019). We perform many simulation studies in Section 5, and finally, we conclude and present some remarks in Section 6.

2 A Short Literature Review on Individual Loss Reserving using Statistical Learning Approaches

Individual loss reserving can be traced back to the 1980s with the development of a mathematical framework in continuous time by Arjas (1989) and Norberg (1986). However, it is in 2007 that the subject really took off with the availability of detailed data, and the development of computing resources. Many approaches have been proposed, e.g. Larsen (2007), Zhao *et al.* (2009), Pigeon *et al.* (2013), and Antonio and Plat (2014). On the one hand, statistical learning techniques are widely used in the field of data analytic. On the other hand, only few approaches based on these techniques have been developed in loss reserving using micro-level information. The tools used in these works usually fall in two main categories: tree-based machine learning methods (see Subsection 2.1) and neural networks (see Subsection 2.2).

2.1 Tree-based Approaches

Tree-based machine learning methods rely on decision trees to predict an outcome. To our knowledge, the first one to have used such an approach in an individual loss reserving framework is Wüthrich (2018a) who uses the popular and well-known Classification And Regression Tree (CART) algorithm, introduced by Breiman *et al.* (1984). In this paper, regression trees are considered in a discrete context to predict number of payments only. First, numbers of payments for reported but not settled claims are predicted with feature components on an individual basis. The regression tree model allows to obtain individual predictions for each claim given its feature components at a given time. Second, incurred but not reported claims are considered. For such claims, individual claim-specific information is unknown hence no individual predictions can be obtained. Wüthrich (2018a) assumes that claim occurrences and reporting process can be described by a homogeneous marked Poisson point process enabling him to apply the Chain-Ladder method to obtain the predictions. Predictions for closed claims, RBNS claims and IBNR claims are then aggregated to obtain a prediction of all payments for all accident years considered. Finally, a prediction for the final reserve amount can be calculated based on these predictions.

The work of Wüthrich (2018a) is the foundation of the work of De Felice and Moriconi (2019) which also uses classification and regression tree (CART) within their prediction model. Contrarily to Wüthrich (2018a), claim amounts paid are considered within a frequency-severity model. CARTs are applied in both the frequency (classification trees) and severity (regression trees) predictions. An important addition in this work is an assumption of multiple payment types meaning that different regimes are used to handle incurred claims. This double-claim regime allowing two different types of compensation for the same claim is shown to be suitable in an application to Italian Motor Third Party Liability data given that incurred claims here can be handled under two regimes: direct compensation and indirect compensation.

Lopez *et al.* (2016, 2019) propose an adaptation of the CART algorithm to censored data (open claims) and implement the procedure to obtain ultimate individual reserves for RBNS claims. This extension of the CART algorithm introduces a weighting scheme based on a Kaplan-Meier estimator to compensate for the censoring of the data in the sample. More precisely, a weighted quadratic loss is used as splitting criterion rather than the quadratic loss of the classical CART algorithm (we describe this approach in Section 4). In Lopez (2019), a construction based on copulas is introduced in a model similar to the one proposed in Lopez *et al.* (2016) based on survival analysis to account for a possible dependence between the length of time from the occurrence to the closure of a claim and the amount of the claim. Tree-based approaches are flexible techniques commended notably for being able to detect interaction effects within the data and to handle structured and unstructured information. Their ease of interpretation is another one of their strong points. The sensibility to changes in the data of CARTs may however lead to different optimal trees lowering their level of predictive accuracy given this increase in variability. Some ensemble methods which combine multiple decision trees leading to a variance reduction and better performance for the predictions are interesting alternatives.

In this vein, Baudry and Robert (2019) proposed a general recursive approach based on Extremely randomized trees (ExtraTrees) to assess outstanding liabilities based on all available information since the reporting of the claim. Applications are done for specific recursive one-period ahead predictions as in the framework proposed by Wüthrich (2018a). Duval and Pigeon (2019) propose an individual loss reserving model based on an application of the gradient boosting algorithm, more precisely the XGBoost algorithm. On the basis of the prediction distribution of the RBNS claims, they compare this non-parametric approach using a machine learning algorithm with more classical reserving techniques such as a bootstrapped version of Mack's collective model (see England and Verrall (2002)), a collective generalized linear model (GLM) (see Wüthrich and Merz (2008)) and an individual GLM loss reserving model. Four XGBoost models are considered with different approaches to deal with censored data due to opened claims at evaluation date. We detail this approach in Subsection 4.2. The main challenges associated with tree-based approaches are summarized in Table 1.

2.2 Neural Networks

Another machine learning methodology which has received much attention lately in a loss reserving setting is neural networks. Taylor (2019) presents briefly the first work which dates back to the application of neural networks with a single hidden layer in Mulquiney (2006). Harej *et al.* (2017) investigate the use of neural networks in a cascade fashion in a presupposed Chain-Ladder structure but on an individual claim basis.

Wüthrich (2018b) aims at improving predictions obtained under the homogeneity assumption within an extended version of Mack's Chain-Ladder method by using neural networks. This neural network extended Chain-Ladder model includes individual claims feature information and captures heterogeneity within the data. Reserves are not evaluated for individual claims but rather a Chain-Ladder method is used to obtain reserves for different homogeneous sub-portfolios of claims. Multiplicative Chain-Ladder factors based on claim features are used for each sub-portfolio. Used to generate data for applications within many following works, Gabrielli and Wüthrich (2018) propose a stochastic scenario generator based on neural networks to simulate a non-life insurance portfolio of individual claims. Gabrielli *et al.* (2019) propose the embedding in a neural network framework of the over-dispersed Poisson (ODP) reserving model. The weight initialization of the neural network is such that it corresponds to the ODP reserving model. During training, the network learns meaningful interactions (if any) between covariates and delivers what should be the optimal form of the regression function within the ODP reserving model. Gabrielli (2019) pursues his work on the ODP model with an extension taking into account both claim counts and claim amounts with the embedding in a neural network of two separate ODP models. Again in a run-off triangle setting, Kuo (2019) uses a deep neural network architecture to forecast paid losses.

Very recent works include Kuo (2020) who proposes the use of a neural networks architecture in an individual claim reserving setting. More precisely, Bayesian mixture density networks are used to forecast individual claims and the application relies on data from a synthetic insurance portfolio of individual claims histories generated with the simulator of Gabrielli and Wüthrich (2018). As for Delong and Wüthrich (2020), they investigate the joint development process of individual payments and incurred claims through the estimation of regression models with neural networks. Finally, Delong *et al.* (2020) pursue in the same vein as Delong and Wüthrich (2020) by simplifying and enhancing the performance of the previous models. Estimation of incurred but not reported claims is an important added element to the previous work done in Delong and Wüthrich (2020). The proposed global architecture rests on six neural networks with each one having a different modeling purpose related to the estimation of frequency and severity elements of RBNS and IBNR claims.

From the recent papers, one see that researchers have shown growing interest in the use of neural networks in an individual loss reserving context. Much more work is ahead in the understanding and the implementation of this promising technique. A future call for proposal based solely on this type of methodology to tackle the evaluation of granular reserves would greatly benefit the actuarial practice. In the present project, we have chosen to concentrate our efforts to tree-based methods.

The main challenges associated with neural networks are summarized in Table 1.

2.3 Others Approaches

Other works related to the use of machine learning techniques in an individual loss reserving setting must also be mentioned, notably Jamal *et al.* (2018) which is a report from the ASTIN Working Party of the International Actuarial Association and the paper Taylor (2019). Jamal *et al.* (2018) present a well rounded detailed study in which traditional methods (Mack's Chain-Ladder and GLM) and several machine learning techniques (random forests, gradient boosting machine, neural networks, TDBoost) are clearly described and tested on a Swiss Re data set. Pros and cons of the different methods are discussed. Taylor (2019) presents the evolution of loss reserving methods with emphasis put on a survey of the latest works on individual loss reserving methodologies including machine learning approaches. Comparisons of the techniques are made highlighting their strong points and guiding the choice to be made between possible strategies.

Table 1: Main challenges associated with tree-based models and neural networks

Challenge	Tree-based approaches	Neural networks
Robustness	Regression trees are not very robust toward changes in observations. It is therefore necessary to rely on more complex and more computationally expensive algorithms, (e.g. random forests or boosting) to use them.	Has not yet really been studied in the actuarial literature.
Interpretability	Rather simple and generally satisfying.	Difficult and needs to be improved. Models are often highly complex and understanding them requires a considerable investment of time.
Dependence	Modeling of dependence (e.g. between lines of business) needs to be developed.	Modeling of dependence needs to be developed.
Schedule of individual payments	Tree-based models do not provide a detailed schedule of individual payments made for each claim. Some models allow a modeling of the schedule of individual payments.	Some models presented in the literature provide a schedule of individual payments. Research on this topic is progressing promptly.
Covariates	Methods to include dynamic covariates, i.e. covariates that can change between the reporting date and the closure date, still need to be developed.	Same comment.
Others	-	Establish guidelines for the selection of hyperparameters involved in models.

3 Individual Loss Reserving Framework

In Property and Casualty insurance, a claim starts with an accident happening at the *occurrence point* (see Figure 2). For some situations, e.g. for bodily injury liability coverage, a *reporting delay* is observed between the occurrence date and the reporting to the insurance company at the *reporting point*. At this moment, the insurer could observe details about the accident, as well as some information about the insured. From this moment, until the *closing date* of the file, a series of random payments is therefore triggered. We illustrate in Figure 1 the development of 5 individual claims.

The *valuation date* t_{val} is the moment on which the insurance company wants to evaluate its solvency and calculate reserves. At this date, we may classify each claim under one of those categories according to the stage reached by its development:

- if the valuation date is between the occurrence and the reporting date, the claim is classified as *Incurred But Not Reported*, or IBNR;
- if the valuation date is between the reporting date and the moment of the closure of the file, the claim is classified as *Reported But Not Settled*, or RBNS; and
- otherwise, the claim is classified as *Closed*.

In this report, we mainly focus on RBNS claims, i.e. claims for which the accident has been reported to the insurer but the file is still not settled.

In Subsection 3.1, we define the notation for the censorship of a random variable that we will use throughout this report. Thereafter, we will specify the loss reserving problem in Subsection 3.2.

3.1 Notation

Let

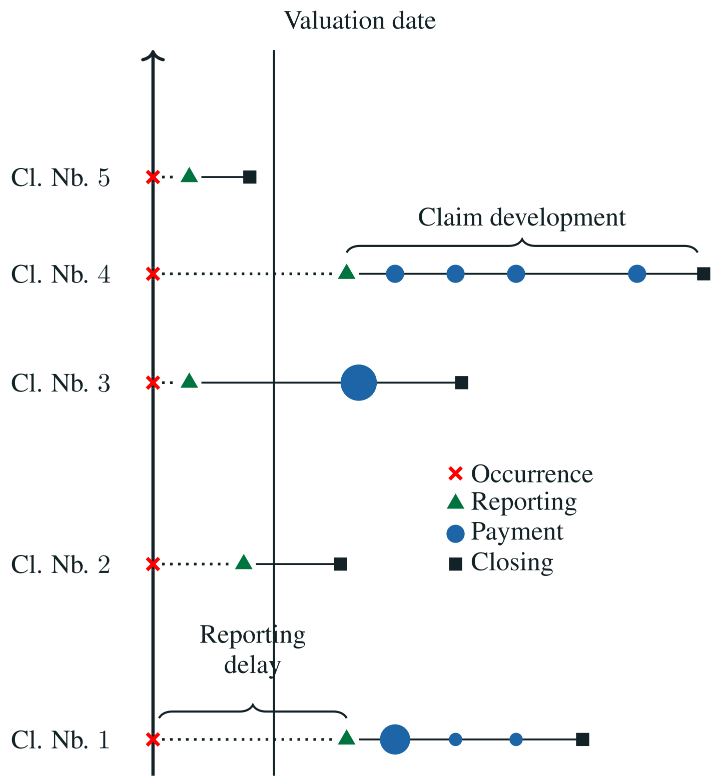

Figure 1: Typical development of 5 individual claims. At the valuation date, claims 1 and 4 are classified as IBNR and claims 2 and 3 are classified as RBNS. Claim 5 is closed.

- $\{Y_1, \dots, Y_n\}$ be a random sample[1] of duration random variables from an unknown cumulative distribution function (cdf) $F : \mathbb{R}^+ \to [0,1]$. In the context of loss reserving, Y_i is the time elapsed between the occurrence date and the closure date for claim i.
- $\{M_1, \dots, M_n\}$ be a set of random variables $M_i \in \mathbb{R}$, $i = 1, \dots, n$. In this report, M_i is a real-valued random variable representing the total paid amount for the i^{th} claim.
- $\{C_1, \dots, C_n\}$ be a random sample from an unknown censoring cdf G. The censoring variable C_i is the delay between the occurrence date and the valuation date. Consequently, open and closed claims are considered as censored and uncensored observations, respectively.
- $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ be a set of covariates, $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^p$, $i = 1, \dots, n$.

We define

$$Z_i = \min(Y_i, C_i), \quad \delta_i = \mathbb{I}\left(Y_i \le C_i\right), \text{ and } N_i = \delta_i M_i.$$

Thus, Z_i and N_i represent, for claim i, the duration and severity observed in the database at the valuation date. Without loss of generality, we assume that $Z_1 < Z_2 < \dots < Z_n$, with δ_i and N_i, $i = 1, \dots, n$, constructed accordingly. In this general framework, (M_i, Y_i) may not be observed due to the censoring effect of C_i, but $\boldsymbol{x}_i$ are always observed. Thus, in a dataset, we have $\{N_i, Z_i, \delta_i, \boldsymbol{x}_i\}_{i=1,\dots,n}$. Finally, we assume that C_i is independent of (Y_i, M_i), and

$$\Pr(Y_i \le C_i | M_i, Y_i, \boldsymbol{x}_i) = \Pr(Y_i \le C_i | Y_i)\,,$$

for $1 \le i \le n$.

Based on that, the main objective is to construct an estimator for

$$\pi_0 = \text{argmin}_{\pi \in \mathcal{P}} \mathbb{E}[\phi\left(M, \pi(Y, \boldsymbol{x})\right)]\,, \tag{1}$$

where $\mathcal{P}$ is an appropriate subset of a functional space and ϕ is a loss function. Informally, this means that we are looking for the function π which minimizes a loss function ϕ calculated between M on one side and $(Y, \boldsymbol{x})$ on the other

[1] A set of independent and identically distributed (iid) random variables.

side. Using the quadratic loss function and $\mathcal{P} = L^2(\mathbb{R}^p)$ or $L^2(\mathbb{R}^{p+1})$, we obtain the classical mean regression model where

$$\pi_0 = \mathbb{E}[M|\boldsymbol{x}] \text{ or } \pi_0 = \mathbb{E}[M|Y, \boldsymbol{x}] \, .$$

3.2 Loss Reserving Model

Often in the actuarial literature, censored variables are discussed when a contract has a limit and/or deductible. It is important to note that in this report we are only interested in the censorship present in the duration of a file, and censored data corresponds to an open claim at the valuation date.

Individual framework. Figure 2 illustrates the structure of the development of an individual claim with

$$M_i = W_{t_{i3}} + W_{t_{i4}} \text{ and } N_i = 0.$$

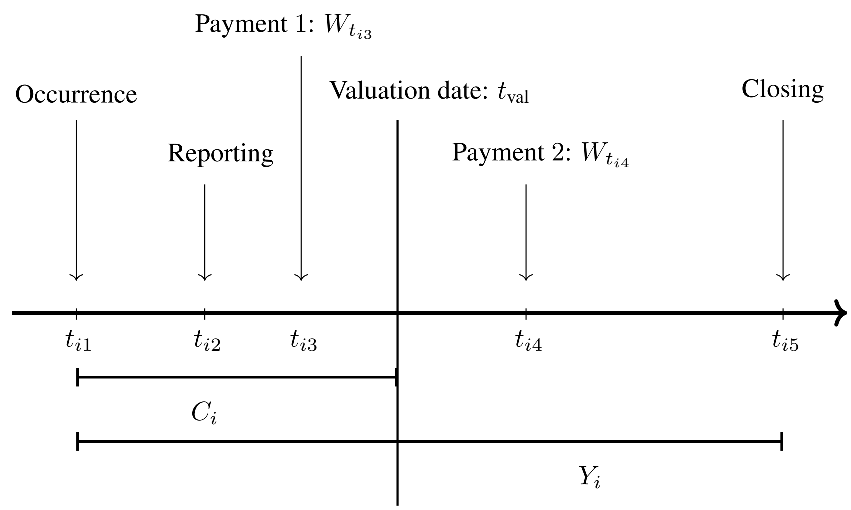

Figure 2: Development of an individual claim

In the loss reserving framework, our main objective is to construct an estimator $\widehat{M_i}$ for

$$\mathbb{E}[M_i|N_i, Z_i, \delta_i, \boldsymbol{x}_i] \, ,$$

which is the best L^2-predictor of the total paid amount M_i. We call this approach the *purely individual framework (IF)*. A prediction of the RBNS reserve amount is given by

$$\widehat{R}^{\text{RBNS}} = \sum_{i=1}^{n} \left(\widehat{M_i} - m_i^* \right) ,$$

where m_i^* is the observed total paid amount at the valuation date for claim i. It should be noted that for closed files, we then have $\widehat{M_i} = m_i^*$. Table 6 illustrates an example of a dataset in this framework.

Collective framework. Traditionally, insurance companies aggregate information by accident year and by development year. Claims with accident year a, $a = 1, \ldots, J$, are all claims that occurred in the ath year after τ, an *ad hoc* starting point common to all claims. For a claim i, a payment made in development year j, $j = 1, \ldots, J$ is a payment made in the jth year after the occurrence t_{i1}, namely a payment $W_{t_{im}}$ for which $j - 1 < t_{im} - t_{i1} < j$. For development years $j = 1, \ldots, J$, we define

$$W_j^{(i)} = \sum_{m \in \mathcal{S}_j^{(i)}} W_{t_{im}},$$

Table 2: Run-off triangle based on Figure 3

Occurrence	Development period		
period	1	2	3
1	W_{11}	W_{12}	W_{13}
2	W_{21}	W_{22}	
3	W_{31}		

where $\mathcal{S}_j^{(i)} = \{m : j-1 < t_{im} - t_{i1} < j\}$, as the total paid amount for claim i during year j and we define the corresponding cumulative paid amount as

$$C_j^{(i)} = \sum_{s=1}^{j} W_s^{(i)}.$$

A collective approach groups every claim in the same accident year to form the aggregate incremental payment

$$W_{aj} = \sum_{i \in \mathcal{K}_a} W_j^{(i)}, \qquad a,j = 1, \ldots, J,$$

where $\mathcal{K}_a$ is the set of all claims with accident year a. For portfolio-level models, a prediction of the reserve is obtained by

$$\widehat{R}^{\text{RBNS+IBNR}} = \sum_{a=2}^{J} \sum_{j=J+2-a}^{J} \widehat{W}_{aj}, \tag{2}$$

where the $\widehat{W}_{aj}$ are usually predicted using only the accident year and the development year. This is the *collective framework* (CF). It is worth noting that this framework does not allow to distinguish the RBNS reserve from the IBNR reserve. In Figure 3, we illustrate the 3 frameworks used in this report. In Table 2, we present the structure of a dataset in the collective framework, also called *run-off triangle*, and in Table 8, we present an example.

Partially individual framework. In the collective framework, each cell contains a series of payments, information about the claims and some information about policyholders. These payments can also be modeled within an individual framework. Hence, a prediction of the total reserve amount is given by

$$\widehat{R}^{\text{RBNS+IBNR}} = \underbrace{\sum_{a=2}^{J} \sum_{j=J+2-a}^{J} \sum_{i \in \mathcal{K}_a} \widehat{W}_j^{(i)}}_{\text{RBNS reserve}} + \underbrace{\sum_{a=2}^{J} \sum_{j=J+2-a}^{J} \sum_{i \in \mathcal{K}_a^{\text{unobs.}}} \widehat{W}_j^{(i)}}_{\text{IBNR reserve}}, \tag{3}$$

where $\mathcal{K}_a^{\text{unobs.}}$ is the set of IBNR claims with occurrence year a. It should be noted that in Equations (2) and (3), we assume that there will be no future payments on claims in the earliest occurrence period ($a = 1$). We call this approach the *partially individual framework* (PIF) because a partial aggregation of the information is made (by development period), but for the remaining part the information has been preserved. In this work, we are mainly interested in the reserve associated with the claims present in the database: the RBNS reserve, which is the first part on the right-hand side of Equation (3). Table 7 illustrates an example of a dataset in this framework.

4 Including Open Files

Assume we have a portfolio $\mathcal{S}$ on which we want to train a model for loss reserving. This portfolio contains both, open and closed claims. At this point, we have some options: (1) considering only non-censored claims in the process, (2) correcting the selection bias using an inverse probability of censoring weighting strategy (see Subsection 4.1), and (3) developing censored claims using a classical model before applying a statistical learning-based model (see Subsection 4.2). Obviously, the first strategy, using only closed files in the training set, leads to building the model using a too high proportion of "simple cases" and underestimating the risk associated with the portfolio. This was clearly shown in Duval and Pigeon (2019). In order to illustrate the other two strategies, we consider a tree-based model, similar to the one proposed in Lopez *et al.* (2016).

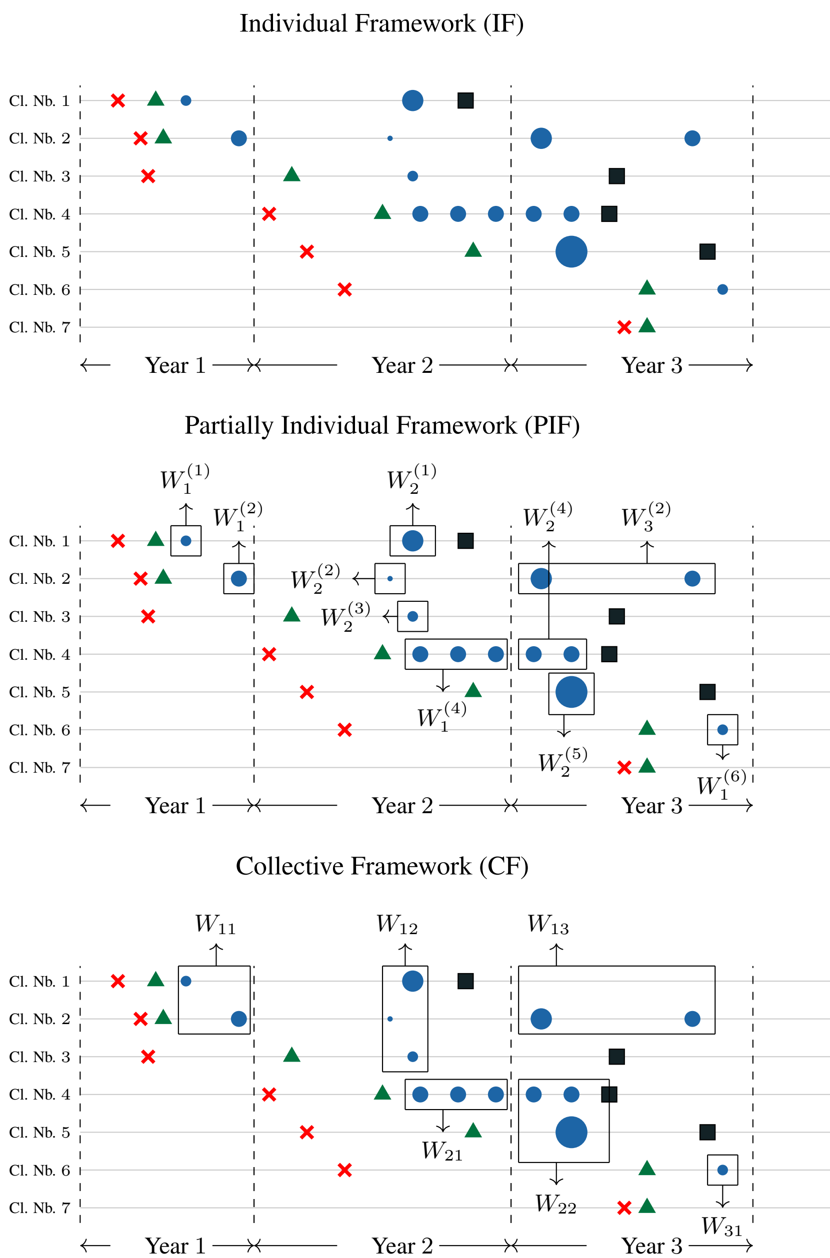

Figure 3: Small artificial portfolio illustrating the 3 frameworks: individual (top), partially individual (center) and collective (bottom)

Before explaining how the model is constructed, we assume that at each step $s \in \{1, 2, \ldots\}$, the tree contains $L^{(s)}$ leaves $\{\mathcal{T}_j^{(s)}\}_{j=1,\ldots,L^{(s)}}$ which are a partition[2] of the space $\mathcal{T} = \mathbb{R}^+ \times \mathcal{X}$. An observation $\widetilde{\boldsymbol{X}}_i = (Y_i, \boldsymbol{X}_i)$ belongs to the leaf ℓ if $\widetilde{\boldsymbol{X}}_i \in \mathcal{T}_\ell^{(s)}$.

- **Step 1: Construction of the maximal tree.** At the beginning of the algorithm ($s = 1$), there is only one leaf in the tree corresponding to the set of all uncensored observations. At each subsequent step, a new tree ($s + 1$) is created by dividing one of the existing leaves. For the leaf ℓ, this split is made based on an optimization: (1) for each covariate $x^{(j)}$ (the j^{th} component of $\widetilde{\boldsymbol{x}}$), one determines the threshold $x_\ell^{(j)}$ that minimizes the function $\mathcal{L}_\ell(j, x_\ell^{(j)})$ defined by

$$\begin{aligned}\mathcal{L}_\ell(j,&x_\ell^{(j)}) = \\ &\min_{(\pi,\pi')\in\Gamma^2} \int \phi(m,\pi)\mathbb{I}\left(\widetilde{\boldsymbol{x}} \in \mathcal{T}_\ell^{(s)}\right)\mathbb{I}\left(x^{(j)} \leq x_\ell^{(j)}\right) d\widehat{F}_n(m,\widetilde{\boldsymbol{x}}) \\ &+ \int \phi(m,\pi')\mathbb{I}\left(\widetilde{\boldsymbol{x}} \in \mathcal{T}_\ell^{(s)}\right)\mathbb{I}\left(x^{(j)} > x_\ell^{(j)}\right) d\widehat{F}_n(m,\widetilde{\boldsymbol{x}}),\end{aligned}$$

where ϕ is a loss function, and $\Gamma \subset \mathbb{R}$; (2) determines

$$j_0 = \text{argmin}_{j=1,\ldots,p+1}\left(\mathcal{L}_\ell(j, x_\ell^{(j)})\right).$$

Finally, two new leaves are created by applying the splitting rule: $x_i^{(j_0)} \leq x_\ell^{(j_0)}$ and $x_i^{(j_0)} > x_\ell^{(j_0)}$. In the absence of censored data, the empirical distribution function $\widehat{F}_n$ can be easily calculated. However, in the presence of censorship, this distribution is unavailable. The procedure ends when there is only one uncensored observation left in each of the leaves, or when all the uncensored observations in the same leaf are identical. This entire step can be performed using the rpart function available in the rpart package.

Example 4.1. *To illustrate the general procedure, we consider the* freaggnumber *dataset available in the* CASdatasets *package. The dataset consists of* 12,513 *classes for which we have the driver age, the age of driving licence, the vehicle age, the exposure and the claim number. In this toy example, the response variable is*

$$Y_i = \frac{\textit{Frequency}_i}{\textit{Exposure}_i}, \qquad i = 1, \ldots, 12{,}513$$

and we use only the vehicle age as a covariate. We randomly split the dataset into a training dataset ($\pm 70\ \%$*) and a validation dataset (*$\pm 30\ \%$*), and we grow the maximal tree. We start with a tree with only one leaf (see Figure 4, left), we try all splitting rules*

$$\mathbb{I}\left(\textit{vehicle age} \leq s\right), \qquad s = 1, 2, \ldots.$$

and then we select the optimal one (see Figure 5, right). We repeat the procedure and, finally, we obtain the maximal tree illustrated in Figure 5.

- **Step 2: Pruning the tree.** Let $K \leq n$ be the number of leaves in the maximal tree. The final tree is a sub-tree S, with $K_S \leq K$ leaves, selected from the set $\mathcal{S}$ of all sub-trees of the maximal tree. The pruning strategy is based on the following optimization problem:

$$S(\alpha) = \text{argmin}_{S\in\mathcal{S}}\left(\int \phi(m, \widehat{\pi}^S)\, d\widehat{F}_n(m, \widetilde{\boldsymbol{x}}) + \frac{\alpha K_S}{n}\right),$$

where

$$\begin{aligned}\widehat{\pi}^S &= \sum_{\ell=1}^{K_S} \widehat{\gamma}_\ell R_\ell(\widetilde{\boldsymbol{x}}), \\ \widehat{\gamma}_\ell &= \text{argmin}_{\pi\in\Gamma} \int \phi(m,\pi) R_\ell(\widetilde{\boldsymbol{x}})\, d\widehat{F}_n(m,\widetilde{\boldsymbol{x}}),\end{aligned}$$

and

$$R_\ell(\widetilde{\boldsymbol{x}}) = \mathbb{I}\left((\widetilde{\boldsymbol{x}}) \in \mathcal{T}_\ell\right).$$

In order to determine the optimal value of α, α^*, a cross-validation procedure is applied. Again, this procedure can be implemented quite directly using the xval argument of the rpart function.

[2]The partition is such that $\mathcal{T}_j^{(s)} \cap \mathcal{T}_{j'}^{(s)} = \emptyset$ for $j \neq j'$, and $\bigcup_{j=1}^{L^{(s)}} \mathcal{T}_j^{(s)} = \mathcal{T}$.

Figure 4: Tree with only one leaf (left) and tree with the first 2 leaves (right) for Example 4.1. The value shown in each of the leafs is the empirical mean of the group.

Example 4.1 (part 2). *Based on the previously obtained maximal tree, we calculate the optimal value of the complexity parameter (α in our equations and* `cp` *in the* R *code). First, we calculate the mean squared error on the validation dataset (vMSE) using a grid of values for the complexity parameter. Then, we find the optimal one (α^*) and we prune the tree using* `cp` $= \alpha^*$*. We present the vMSE against several values for the complexity parameter and the resulting optimal tree in Figures 6 and 7, respectively.*

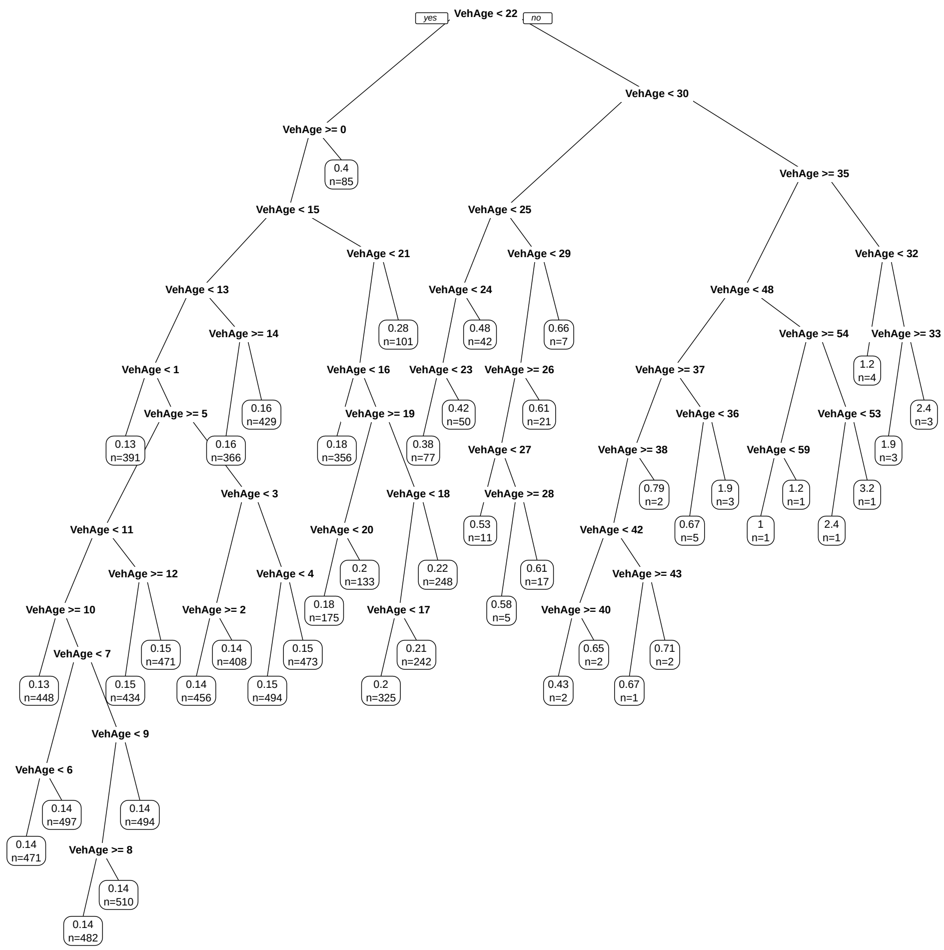

Figure 5: Maximal tree for Example 4.1

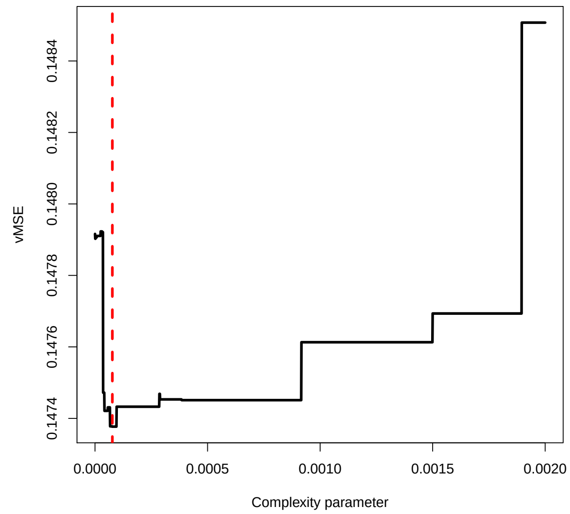

Figure 6: Mean squared error calculated on the validation set for several values of the complexity parameter

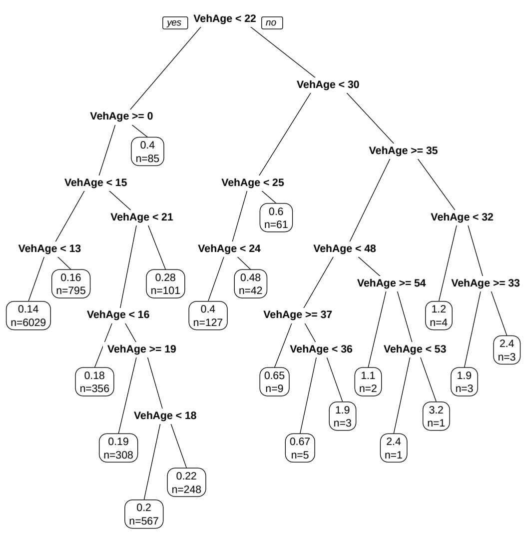

Figure 7: Optimal tree for Example 4.1

Finally, the estimator of π_0 defined by Equation (1) is given by

$$\widehat{M} = \widehat{\pi}^{S(\alpha^*)} = \sum_{\ell=1}^{K_{S(\alpha^*)}} \widehat{\gamma}_\ell R_\ell(\widetilde{\boldsymbol{x}}).$$

As mentioned, in a context of loss reserving, the challenge comes from the unavailability of $\widehat{F}_n(m, \widetilde{\boldsymbol{x}}) = \widehat{F}_n(m, y, \boldsymbol{x})$ in the presence of censored data.

4.1 Strategies Based on Survival Analysis

In this section, we introduce the main ideas of the approach proposed in Lopez *et al.* (2016). The authors propose a weighted regression tree procedure for censored data. This approach is based on Kaplan-Meier (KM) weights given by (see Appendix A for the details)

$$w_k = \left(\frac{\delta_k}{n-k+1}\right) \prod_{i=1}^{k-1} \left(\frac{n-i}{n-i+1}\right)^{\delta_i}, \tag{4}$$

where $w_1 = \delta_1/n$, and

$$w_n = \prod_{i=1}^{n-1} \left(\frac{n-i}{n-i+1}\right)^{\delta_i}.$$

In Lopez *et al.* (2016), the authors explain in detail the theoretical bases of the proposed approach and demonstrate the consistency of the estimator obtained. In the presence of censorship, they suggest replacing $\widehat{F}_n(m, y, \boldsymbol{x})$ (in **Step 1** and **Step 2**) by

$$\tilde{F}(m, y, \boldsymbol{x}) = \frac{1}{n} \sum_{j=1}^{n} \frac{\delta_j \mathbb{I}(N_j \leq m, Z_j \leq y, \boldsymbol{x}_j \leq \boldsymbol{x})}{(1 - \widehat{G}(Z_j^-))} = \sum_{j=1}^{n} w_j \mathbb{I}(N_j \leq m, Z_j \leq y, \boldsymbol{x}_j \leq \boldsymbol{x}),$$

where $\widehat{G}$ is given by

$$\widehat{G}(Z_k^-) = 1 - \prod_{i=1}^{k-1} \left(\frac{n-i}{n-i+1}\right)^{1-\delta_i}.$$

Moreover, in order to determine the optimal value of α (**Step 2**), they propose a cross-validation procedure minimizing

$$\sum_{j=1}^{n^*} \frac{\delta_j \phi\left(N_j, \widehat{\pi}^{S(\alpha)}\right)}{1 - \widehat{G}(Z_j^-)}.$$

For closed claims ($\delta_i = 1$), we simply have $\widehat{M_i} = m_i^*$, the observed total paid amount. For $\delta_i = 0$, several estimators are possible. Here we focus on two of the main ones. The first one is based on

$$\begin{aligned} M_i^{(1)} = M_i^{(1)}(m_i^*, z_i, \boldsymbol{x}_i) = \mathbb{E}[M_i | M_i > m_i^*, Y_i > z_i, \boldsymbol{x}_i] &= \frac{\mathbb{E}[M_i \mathbb{I}(M_i > m_i^*, Y_i > z_i) | \boldsymbol{x}_i]}{\Pr(M_i > m_i^*, Y_i > z_i | \boldsymbol{x}_i)} \\ &= \frac{\mathbb{E}[\psi_2(m_i^*, z_i) | \boldsymbol{x}_i]}{\mathbb{E}[\psi_1(m_i^*, z_i) | \boldsymbol{x}_i]} \\ &= \frac{\pi_2(\boldsymbol{x}_i)}{\pi_1(\boldsymbol{x}_i)}, \end{aligned}$$

where $\psi_1(m, z) = \mathbb{I}(M > m, Y > z)$ and $\psi_2(m, z) = M\psi_1(m, z)$. In Lopez *et al.* (2019), the authors propose to define

$$\widehat{M}_i^{(1)} = \frac{\widehat{\pi}_2(\boldsymbol{x}_i)}{\widehat{\pi}_1(\boldsymbol{x}_i)}, \tag{5}$$

where both estimators are constructed using the regression tree procedure introduced previously with Kaplan-Meier weights. These weights are equal to 0 for open (censored) claims; otherwise, the larger the delay between the occurrence date and the evaluation date the higher the weight. This compensates for the fact that only few claims with large observed development are observed in a dataset.

A second strategy is using one single tree directly estimating

$$M_i^{(2)} = M_i^{(2)}\left(N_i, Y_i, \boldsymbol{x}_i\right) = \pi_5(\boldsymbol{x}_i, N_i, Y_i) = \mathbb{E}[M_i|N_i, Y_i, \boldsymbol{x}_i]\,.$$

Because Y_i is unknown for open claims (censored), we need, as a preliminary step, to obtain a predicted value $\widehat{y}_i$. Thus, the steps are

(a) construct a model for $(Y_i|Y_i \geq z_i, \boldsymbol{x}_i)$:

$$\begin{aligned} Y_i = \mathbb{E}[Y_i|Y_i > z_i, \boldsymbol{x}_i] &= \frac{\mathbb{E}[Y_i\mathbb{I}(Y_i > z_i)|\boldsymbol{x}_i]}{\Pr(Y_i > z_i|\boldsymbol{x}_i)} \\ &= \frac{\mathbb{E}[\psi_4(z_i)|\boldsymbol{x}_i]}{\mathbb{E}[\psi_3(z_i)|\boldsymbol{x}_i]} \\ &= \frac{\pi_4(\boldsymbol{x}_i)}{\pi_3(\boldsymbol{x}_i)}, \end{aligned}$$

where $\psi_4(z) = Y\psi_3(z)$, and $\psi_3(z) = \mathbb{I}(Y > z)$;

(b) for an open claim, obtain a prediction for the duration

$$\widehat{y}_i = \frac{\widehat{\pi}_4(\boldsymbol{x}_i)}{\widehat{\pi}_3(\boldsymbol{x}_i)},$$

using regression trees with Kaplan-Meier weights; and

(c) for an open claim, obtain a prediction

$$\widehat{M}_i^{(2)} = M_i^{(2)}\left(N_i, \widehat{y}_i, \boldsymbol{x}_i\right). \tag{6}$$

4.2 Strategies Based on Imputation of Missing Data

In this section, we introduce the main ideas of the approach proposed in Duval and Pigeon (2019). In the initial paper, the model is built using a gradient boosting algorithm but can be directly modified to be used with a tree-based model. In order to be able to better study the impact of the strategy used to include open cases, we replace the gradient boosting algorithm with a simple tree model such as the one described at the beginning of Section 4, but with equal weights replacing weights based on Kaplan-Meier.

The main idea is as follows: artificially generating values, or *pseudo-responses*, for all open files in order to "complete" the portfolio. Then, it becomes possible to calculate $\widehat{F}_n(m, y, \boldsymbol{x})$.

In the collective framework, we assume that incremental aggregate payments W_{aj} are independent, and $W_{aj} \sim$ Exp. family with expected value given by $g\left(\mathbb{E}[W_{aj}]\right) = g(\mu_{aj}) = \beta_0 + \kappa_a + \beta_j + \nu_{aj}$, where $g()$ is the link function, κ_a, $a = 2, 3, \ldots, J$ is the accident period effect, β_j, $j = 2, 3, \ldots, J$ is the development period effect, β_0 is the intercept, and ν_{aj} is an offset term for the volume of payments in cell (a, j). Moreover, we have $\mathbb{V}\text{ar}[W_{aj}] = \varphi\mathcal{V}\left(\mathbb{E}[W_{aj}]\right)$, where $\mathcal{V}()$ is the variance function and φ is the dispersion parameter (see Wüthrich and Merz (2008)). The predicted expected value is given by

$$\widehat{\mu}_{aj} = g^{-1}(\widehat{\beta}_0 + \widehat{\kappa}_a + \widehat{\beta}_j + \nu_{aj}).$$

Back to the partially individual framework, we have, for an open claim with accident period a_i,

$$\widehat{\mu}_J^{(i)} = \underbrace{\sum_{j=1}^{J+1-a_i} w_j^{(i)}}_{\text{observed part}} + \sum_{j=J+2-a_i}^{J} \widehat{\mu}_{aj},$$

and

$$\widetilde{M}_i^{(3)} = \widehat{F}^{-1}_{C_J^{(i)}}\left(q\right),$$

Table 3: Portfolio for Example 4.2

Claim id	Acc. year	Dev. year 1	Dev. year 2	Dev. year 3	Status (val. date)
1	2000	200	400	100	Closed
2	2000	300	400	150	Closed
3	2001	250	450	–	Open
4	2001	300	500	–	Open
5	2001	350	600	–	Closed
6	2002	400	–	–	Open
7	2002	200	–	–	Open

Table 4: Portfolio for a strategy based on survival analysis

Claim id	Total paid amount	Duration (Z)	Status (val. date)	$w^{\text{class.}}$	w^{KM}
1	700	2.9961	Closed	$1/7$	0.4
2	850	2.9992	Closed	$1/7$	0.4
3	700	2.0052	Open	$1/7$	0
4	800	2.0057	Open	$1/7$	0
5	950	2.0024	Closed	$1/7$	0.2
6	400	0.9913	Open	$1/7$	0
7	200	1.0011	Open	$1/7$	0

which is the level q quantile of the distribution of $C_J^{(i)}$ with expected value $\widehat{\mu}_J^{(i)}$. This quantile can be obtained using simulations, a bootstrap procedure, etc. As suggested in Duval and Pigeon (2019), we estimate the level q by using cross-validation. For closed claims, we simply set $\widetilde{M}_i^{(3)} = M_i$. We can now fit the tree model described at the beginning of Section 4 using this artificially completed database:

$$\widehat{M}_i^{(3)} = \mathbb{E}\left[\widetilde{M}_i^{(3)} | \boldsymbol{x}_i\right] = M_i^{(3)}\left(\boldsymbol{x}_i\right). \tag{7}$$

It is also possible to replace the generalized linear model by a classic collective model such as Mack's model (see Duval and Pigeon (2019)).

Example 4.2. *In order to illustrate the two strategies, we use a small artificial portfolio made up of* 7 *claims and presented in Table 3. In Table 4, we present all information for the strategy based on survival analysis and Kaplan-Meier weights. The "duration" column corresponds to the development period (in years) slightly modified by a random number (the* jitter *function in* R*) in order to avoid the difficulties linked to equal values in the calculation of weights. We have*

$$w_i^{KM} = \frac{\delta_i}{1 - \widehat{G}(Z_i^-)},$$

where δ_i is the status of claim i ($\delta_i = 0$ for an open claim and $\delta_i = 1$ for a closed claim), and Z_i is defined in Section 3.1. *Thus, using a strategy based on survival analysis, the model is adjusted using the weights presented in column w^{KM} rather than those presented in column $w^{class.}$.*

Using a generalized linear model with the over-dispersed Poisson distribution and a logarithmic link function, we obtain the following estimated values: $\widehat{\beta}_0 = 5.4917$, $\widehat{\kappa}_{2001} = 0.2283$, $\widehat{\kappa}_{2002} = 0.2121$, $\widehat{\beta}_2 = 0.5179$ and $\widehat{\beta}_3 = -0.6634$. Then, we use these estimated parameters to complete all open claims in the portfolio:

$$\mu_3^{(1)} = 200 + 400 + 100 = 700 \qquad \textit{(already completed)}$$
$$\mu_3^{(2)} = 300 + 400 + 150 = 850 \qquad \textit{(already completed)}$$
$$\widehat{\mu}_3^{(3)} = 250 + 450 + \exp\left(5.4917 + 0.2283 - 0.6634\right) = 857.0556$$
$$\widehat{\mu}_3^{(4)} = 300 + 500 + \exp\left(5.4917 + 0.2283 - 0.6634\right) = 957.0556$$
$$\widehat{\mu}_3^{(5)} = 350 + 600 = 950 \qquad \textit{(already completed)}$$
$$\widehat{\mu}_3^{(6)} = 400 + \exp\left(5.4917 + 0.2121 + 0.5179\right) + \exp\left(5.4917 + 0.2121 - 0.6634\right) = 1{,}058.09$$
$$\widehat{\mu}_3^{(7)} = 200 + \exp\left(5.4917 + 0.2121 + 0.5179\right) + \exp\left(5.4917 + 0.2121 - 0.6634\right) = 858.0904.$$

Finally, for open claims, we obtain pseudo-responses using a quantile q of the over-dispersed Poisson distribution $\widetilde{M}_i^{(3)} = \widehat{F}^{-1}_{C_3^{(i)}}(q)$ with expected value $\widehat{\mu}_3^{(i)}$, $i = 3, 4, 6, 7$. Thus, in the algorithm detailed at the beginning of Section 4, we can now estimate the empirical cumulative distribution function using 7 (artificially) closed claims.

We can adapt this model to the partially individual framework (PIF), which will make it possible to include individual covariates such as the status of the files (open or closed), information on the accident, etc. The implementation of the model is quite similar (see Duval and Pigeon (2019) and Charpentier and Pigeon (2016) for the details). Finally, in the PIF, we assume that covariates remain identical after the valuation date, which is not exactly accurate in the presence of dynamic variables.

For an open claim with accident period a_i, we have

$$\widehat{\mu}_j^{(i)} = g^{-1}(\widehat{\beta}_0 + \widehat{\beta}_j + \boldsymbol{\lambda}\boldsymbol{x}_i),$$
$$\widehat{C}_J^{(i)} = \sum_{j=1}^{J+1-a_i} W_j^{(i)} + \sum_{j=J+2-a_i}^{J} \widehat{\mu}_j^{(i)},$$

and

$$\widetilde{M}_i^{(4)} = F^{-1}_{\widehat{C}_J^{(i)}}(q),$$

where $\boldsymbol{\lambda}$ is a vector of parameters. Finally, using a tree model, we have

$$\widehat{M}_i^{(4)} = \mathbb{E}\left[\widetilde{M}_i^{(4)} | \boldsymbol{x}_i\right] = M_i^{(4)}(\boldsymbol{x}_i). \tag{8}$$

5 Simulation Studies

5.1 Comparison of the weighted and the classical regression tree procedure

As a first step, we want to illustrate the utility and the performance of the algorithm explained in Subsection 4.1. Informally, we want to determine to what extent the use of the KM weights defined by Equation 4 are useful compared to the classical approach which uses identical weights for all observations ($w_k = 1/n$, $\forall k$). We propose a small simulation study based on the parameters indicated in Section 4 of Lopez *et al.* (2016). For sake of simplicity, we consider the case where $Y_i = M_i$ and we focus on estimating $\mathbb{E}[Y_i|\boldsymbol{x}_i]$. Informally, it means we assume that each additional day the claim is opened costs the insurer an additional \$1.

We consider various levels of censorship (10%, 30% and 50%), various sample sizes, and we apply both, the classical regression tree procedure and the regression tree procedure with KM weights. We present, in Table 5, the means over 5,000 simulations of the sums of weighted squared errors given by

$$WSE = \sum_{i=1}^{n} w_i \left(\tilde{Z}_i - \mathbb{E}[Z_i]\right)^2,$$

where $\tilde{Z}_i$ and $\mathbb{E}[Z_i]$ stand for the prediction and the expected value for observation i, respectively[3].

Table 5: Sums of Weighted Squared Errors (means over 5,000 simulations)

% Cens.	Approach	$n = 100$	$n = 500$	$n = 1{,}000$	$n = 5{,}000$	$n = 10{,}000$
10%	KM	57.29	23.44	14.06	3.18	1.62
	class.	47.10	18.66	12.68	5.75	4.91
30%	KM	72.33	47.50	34.99	12.81	7.32
	class.	68.80	45.45	41.43	33.33	31.49
50%	KM	71.82	71.76	70.92	42.93	31.34
	class.	93.38	84.79	81.02	76.11	71.47

As expected, we note that when the sample size (n) and/or the % of censorship increases, the performance of the KM estimator tends to exceed that of the classical estimator.

[3]Recall that all the codes used are available in the R Markdown document.

5.2 Structure of a dataset

In Table 6, we present a generic database in the individual framework from which it will be possible to perform various simulation studies. In this table, "Status" is a dichotomous variable indicating the status of the claim (0 for closed files and 1 for open files) at the payment date. $X_1, \ldots, X_p$ represent the time independent explanatory variables available in the database. In this specific example, X_1 and X_p represent the age and gender of the insured, respectively. Tables 7 and 8 present the same dataset in the partially individual framework and the collective framework, respectively.

Table 6: Generic Database in the individual framework (IF)

Claim_id	Acc_date	Rep_date	Pmt_date	Paid	Status	X_1	$\cdots$	X_p
001	01/01/1995	01/04/1995	03/01/1995	3,234	1	34	$\cdots$	M
001	01/01/1995	01/04/1995	05/01/1995	450	1	34	$\cdots$	M
001	01/01/1995	01/04/1995	10/01/1996	1,200	0	34	$\cdots$	M
002	01/02/1995	01/05/1995	04/01/1995	3,000	1	47	$\cdots$	M
002	01/02/1995	01/05/1995	06/01/1995	200	0	47	$\cdots$	M
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
303	04/01/2005	05/01/2005	05/07/2005	2,000	1	52	$\cdots$	M
303	04/01/2005	05/01/2005	07/01/2005	450	1	52	$\cdots$	M
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 7: Generic Database in the partially individual framework (PIF)

Claim_id	Acc_date	Rep_date	Pmt_year	Paid	Status	X_1	$\cdots$	X_p
001	01/01/1995	01/04/1995	1995	3,684	1	34	$\cdots$	M
001	01/01/1995	01/04/1995	1996	1,200	0	34	$\cdots$	M
002	01/02/1995	01/05/1995	1995	3,200	1	47	$\cdots$	M
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
303	04/01/2005	05/01/2005	2005	2,450	1	52	$\cdots$	M
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 8: Generic Database in the collective framework (CF)

Acc_year	Pmt_year	Paid
1995	1995	6,884
1995	1996	1,200
$\vdots$	$\vdots$	$\vdots$
2005	2005	2,450
$\vdots$	$\vdots$	$\vdots$

To respect a replicability criteria, we use simulated data by the *Individual Claims History Simulation Machine*, or ICHSM, described in Gabrielli and Wüthrich (2018) in our analyzes. The aim of the ICHSM project was to develop a stochastic simulation machine that generates individual claims histories of non-life insurance claims. The simulation machine is based on neural networks calibrated on real, but unknown to us and to the public, non-life insurance data. This database contains four unidentified lines of business, and the available covariates suggest that these are bodily injury coverages. Thus, we have access to the following covariates: line of business (LoB), labor sector of the injured (cc), age of the injured (age), part of the body injured (inj_part) and reporting delay (RepDel). The ICHSM did not allow us to include adjuster-set case reserves in our analyzes. However, if the latter are available, and they are consistent over time, they could be used as a covariate in the model (see Antonio and Plat (2014) for example). Moreover, the simulated individual data are aggregated on an annual basis: we thus have 12 annual photographs of each claim from the accident date. Finally, in order to simplify the analyzes, we assume that there is no possible reopening nor reimbursement. Appendices A and C in the paper Gabrielli and Wüthrich (2018) provide more details regarding the actual database used to calibrate the ICHSM.

5.3 Comparison of Strategies

In this subsection, we compare, based on simulated databases, the performance of several approaches:

- Mack's model with bootstrap (Gamma distribution);
- collective over-dispersed Poisson model for reserves (see Wüthrich and Merz (2008));
- tree-based model using strategies based on survival analysis (estimators $\widehat{M}_i^{(1)}$ and $\widehat{M}_i^{(2)}$); and
- tree-based model using strategies based on imputation (estimators $\widehat{M}_i^{(3)}$ and $\widehat{M}_i^{(4)}$).

All approaches are applied to three scenarios: (1) one line of business without inflation, (2) two lines of business without inflation, and (3) two lines of business with inflation in the frequency.

Scenario I: 1 line of business without inflation. We construct a validation dataset containing $1{,}060$ claims, $1{,}060 \times 12 = 12{,}720$ annual photographs and accident years between 1994 and 2005. This dataset assumes only one line of business and no inflation for frequency. We present some descriptive statistics in Table 19 and Figure 15 in Appendix B, as well as in Table 10. In this dataset, we have access to the complete development of all claims. Therefore, we can choose various valuation dates t_{val} and split the validation dataset into an *available dataset* (everything before t_{val}) and an *outstanding dataset* (everything after t_{val} about claims with occurrence data before t_{val}). Then, we use information in the available dataset to make a prediction of the final total cost, we calculate the reserve, and finally we compare this value with the true value observed in the outstanding dataset. In Table 9, we summarize the dataset by occurrence and development year.

Table 9: Validation incremental run-off triangle (in \$1 000)

	1	2	3	4	5	6	7	8	9	10	11	12
1994	39.58	11.03	2.54	1.93	1.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1995	23.53	12.50	2.70	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1996	32.95	44.53	13.47	8.83	6.42	4.86	8.22	5.64	3.42	4.75	3.20	3.25
1997	43.25	14.04	1.62	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1998	32.17	14.91	3.45	0.00	0.49	0.97	0.00	0.00	0.00	0.00	0.00	0.00
1999	94.75	64.04	22.39	10.29	7.89	7.11	6.09	5.33	4.03	0.00	0.00	0.00
2000	30.56	10.96	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2001	43.88	16.10	6.84	3.40	1.80	1.51	1.44	1.33	0.00	1.22	1.00	0.00
2002	32.66	13.26	3.28	0.87	0.67	0.00	1.46	0.00	0.00	0.00	0.00	0.00
2003	55.36	168.10	84.14	42.41	30.60	23.92	20.34	18.17	15.14	15.47	13.71	12.48
2004	61.33	25.52	3.90	2.27	1.98	1.81	1.68	0.00	1.53	1.17	0.00	0.00
2005	72.83	95.80	25.61	3.59	15.02	10.24	8.55	5.87	5.48	4.66	4.28	3.72

Table 10: Validation dataset (in \$1,000) for Scenario I

Valuation date	% of censored data	RBNS amount	IBNR amount
01/01/2005	11.9	350	4
01/01/2006	11.7	406	8
01/01/2007	7.7	260	1
01/01/2008	6.6	192	1
01/01/2009	5.4	162	0
01/01/2010	4.2	124	0
01/01/2011	3.7	93	0
01/01/2012	2.6	68	0

In order to build our estimators we generate training databases using *Individual Claims History Simulation Machine* again. As a preliminary step, for the estimators defined by Equations (7) and (8), we must first determine the level q to be used in the completion of the databases. To do this, we generate databases of size 2,000 and calculate the mean absolute error of prediction (MAE) for a grid of values of q. For the two estimators, the results are presented in Figure 8 for valuation date 01/01/2011. Graphs for valuation dates 01/01/2006 and 01/01/2010 are similar and are not presented here. Selected values are $\widehat{q}^{(2006,3)} = 0.85$, $\widehat{q}^{(2006,4)} = 0.85$, $\widehat{q}^{(2010,3)} = 0.8$, $\widehat{q}^{(2010,4)} = 0.7$ $\widehat{q}^{(2012,3)} = 0.6$ and $\widehat{q}^{(2012,4)} = 0.4$, where $\widehat{q}^{(i,j)}$ is the selected quantile for estimator j and valuation year i. In Table 11, we present covariates used in all models. It is important to note that the limited number of covariates available in the simulated databases is not the best scenario for tree-based models. Unfortunately, the *Individual Claims History Simulation Machine* used does not provide access to more covariates. When it comes to individual approaches, the availability of a detailed dataset is key, and there is not, at present and to our knowledge, this kind of data openly available in the scientific community. However, we believe that the limited number of covariates does not have a major impact on the validity of the analyzes made in this report. However, in an application on a real portfolio, it would be necessary to make sure to have a larger number of covariates.

Table 11: Covariates used in models

Model	Component	Covariates
Mack	-	acc. year, dev. year
GLM-ODP	-	acc. year, dev. year
$\widehat{M}_i^{(1)}$	$\pi_1(\boldsymbol{x}_i)$ $\pi_2(\boldsymbol{x}_i)$	age, RepDel, cc, inj_part age, RepDel, cc, inj_part
$\widehat{M}_i^{(2)}$	$\pi_3(\boldsymbol{x}_i)$ $\pi_4(\boldsymbol{x}_i)$ $\pi_5(\boldsymbol{x}_i, N_i, Y_i)$	age, RepDel, cc, inj_part age, RepDel, cc, inj_part age, RepDel, cc, inj_part, Y
$\widehat{M}_i^{(3)}$	imputation tree	acc. year, dev. year acc. year, Last Paid, RepDel, cc, inj_part
$\widehat{M}_i^{(4)}$	imputation tree	acc. year, dev. year, cc, age, inj_part, RepDel, Status acc. year, Last Paid, RepDel, cc, inj_part

Then, we evaluate the four estimators defined using several values for the size of the training database (n_{train}). We present some of the results in Table 12. Based on these results we conclude that using training databases of size $2{,}000 - 5{,}000$ seems sufficient to obtain relatively stable results in a reasonable time.

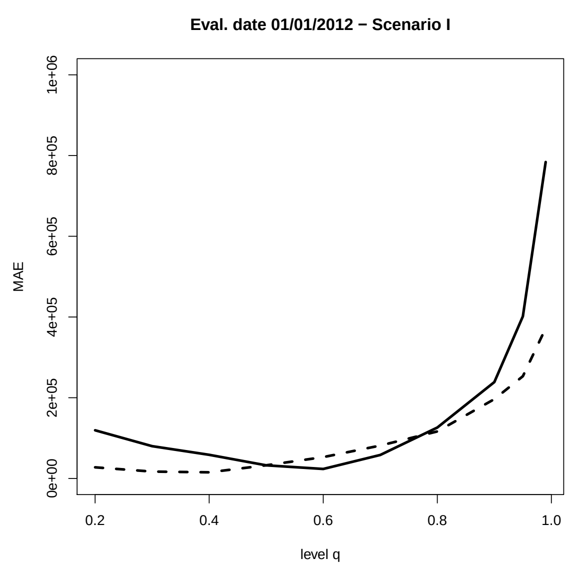

Figure 8: Mean absolute error of prediction as a function of the level q for estimator $\widehat{M}_i^{(3)}$ (solid line) and estimator $\widehat{M}_i^{(4)}$ (broken line)

Table 12, Table 14 and Table 16 present the expected values of the reserves using Tree $\widehat{M}_i^{(1)}$, Tree $\widehat{M}_i^{(2)}$, Tree $\widehat{M}_i^{(3)}$ and Tree $\widehat{M}_i^{(4)}$ models for 3 levels of portfolio maturity: 01/01/2006, 01/01/2010 and 01/01/2012. It is much more informative to look at the predictive distributions of the reserves which are illustrated in Figure 9. Remember that

Table 12: Predicted Reserve (in \$1,000) for Scenario I

Val. year (% cens.) (obs. value)	$\mathbb{E}\left[\widehat{M}_i^{(1)}\right]$	$\mathbb{E}\left[\widehat{M}_i^{(2)}\right]$	$\mathbb{E}\left[\widehat{M}_i^{(3)}\right]$	$\mathbb{E}\left[\widehat{M}_i^{(4)}\right]$	Mack (Gamma)	GLM (ODP)
2012 (2.6%) (68)	787	27	58	57	13	29
2010 (4.2%) (124)	282	189	215	206	69	64
2006 (11.7%) (414)	677	209	480	650	217	234

the most recent occurrence year is 2005 and therefore, for the subsequent valuation dates, there is no new claim after December 31, 2005. Figure 9 presents the predictive distributions for the total reserve, IBNR and RBNS, because with the collective models, it is not possible to separate the two types of reserve. For the valuation dates 01/01/2010 and 01/01/2012, this is not a problem since there is no longer an IBNR claim in the database. For the valuation date 01/01/2006, we added the true observed value of the IBNR reserve to the simulated values of the RBNS reserve for Tree $\widehat{M}_i^{(1)}$, Tree $\widehat{M}_i^{(2)}$, Tree $\widehat{M}_i^{(3)}$ and Tree $\widehat{M}_i^{(4)}$ models (similar to what is done in Duval and Pigeon (2019)). Because the IBNR reserve (\$8,000) is a very small part of the total reserve (\$414,000), the impact on the analysis is negligible. Of course, if IBNR represents a significant part of the total amount of the reserve, the comparison of the results based on individual approaches with the results based on collective approaches will then be strongly biased. This bias can be corrected in an analysis, for example, by subtracting the amount actually paid for IBNR claims from the total amount of the reserve obtained from a collective approach (see Duval and Pigeon (2019) for an example). Or one can complete the individual approach with a simple model for the frequency and the severity of IBNR claims (e.g. see Wüthrich (2018a) and Baudry and Robert (2019)).

In order to determine if the integration of open claims improve the results of the methods tested, we present in Figure 10 predictive distributions of the reserve amount using all claims and only closed claims in the calibration process. We note that, in practically all cases, the fact of not considering the open files in the calibration process leads to an underestimation of the risk. This underestimation is particularly important for estimators based on Tree $\widehat{M}_i^{(3)}$ and Tree $\widehat{M}_i^{(4)}$. In addition, this conclusion is similar to that obtained following the analysis made in Duval and Pigeon (2019). Therefore, it seems clear to us that a simplistic strategy in which open files would be removed from the calibration process is not advisable.

Scenario II: 2 lines of business with no inflation. We construct a validation dataset containing $1{,}063$ claims, $1{,}063 \times 12 = 12{,}756$ annual photographs and accident years between 1994 and 2005. This dataset assumes two lines of business, LoB 1 and 2, and no inflation for frequency. We present some descriptive statistics in Table 19 and Figure 15 in Appendix B, as well as in Table 13. Selected values are $\widehat{q}^{(2006,3)} = 0.80$, $\widehat{q}^{(2006,4)} = 0.30$, $\widehat{q}^{(2010,3)} = 0.95$, $\widehat{q}^{(2010,4)} = 0.95$, $\widehat{q}^{(2012,3)} = 0.95$ and $\widehat{q}^{(2012,4)} = 0.9$.

Table 13: Validation dataset (in \$1,000) for Scenario II

Valuation date	% of censored data	RBNS amount	IBNR amount
01/01/2005	8.4	90	19
01/01/2006	8.1	54	9
01/01/2007	4.1	10	6
01/01/2008	3.3	4	6
01/01/2009	2.7	7	0
01/01/2010	2.2	7	0
01/01/2011	1.7	3	0
01/01/2012	1.0	3	0

We present results in Table 14. Figure 11 presents the predictive distribution of the reserve amount using all models for the same 3 levels of portfolio maturity.

Scenario III: 2 lines of business with inflation (frequency). We construct a validation dataset containing $1{,}060$ claims, $1{,}060 \times 12 = 12{,}720$ annual photographs and accident years between 1994 and 2005. This dataset

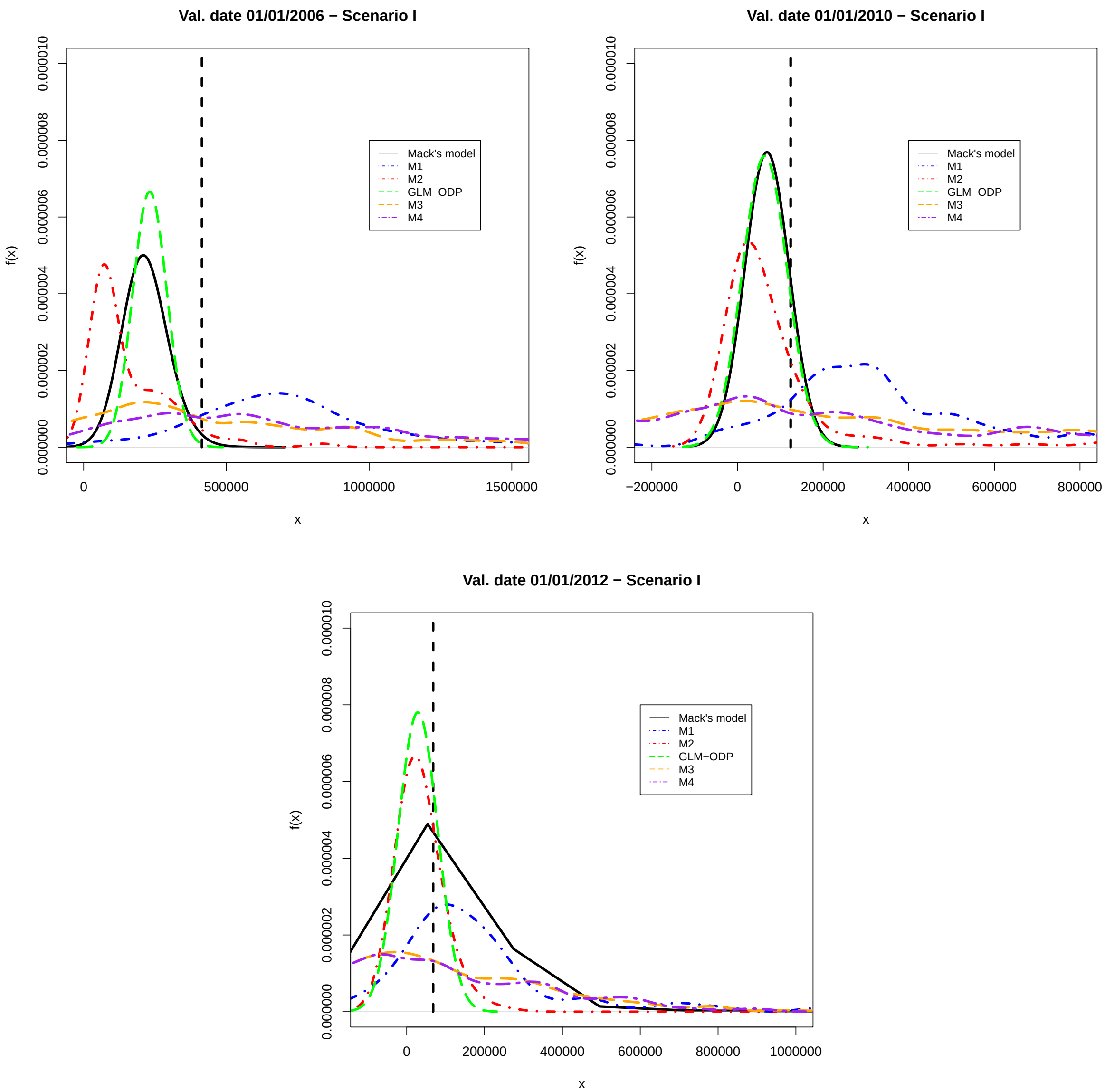

Figure 9: Predictive distribution of the reserve amount. The observed value is $414,000 for 2006 (top, left), $124,000 for 2010 (top, right) and $68,000 for 2012 (bottom)

Table 14: Predicted Reserve (in $1,000) for Scenario II

Val. year (% cens.) (obs. value)	$\mathbb{E}\left[\widehat{M}_i^{(1)}\right]$	$\mathbb{E}\left[\widehat{M}_i^{(2)}\right]$	$\mathbb{E}\left[\widehat{M}_i^{(3)}\right]$	$\mathbb{E}\left[\widehat{M}_i^{(4)}\right]$	Mack (Gamma)	GLM (ODP)
2012 (1.0%) (3)	152	98	-1	-3	1	1
2010 (2.2%) (7)	176	93	13	11	5	9
2006 (8.1%) (63)	1,135	256	178	237	90	110

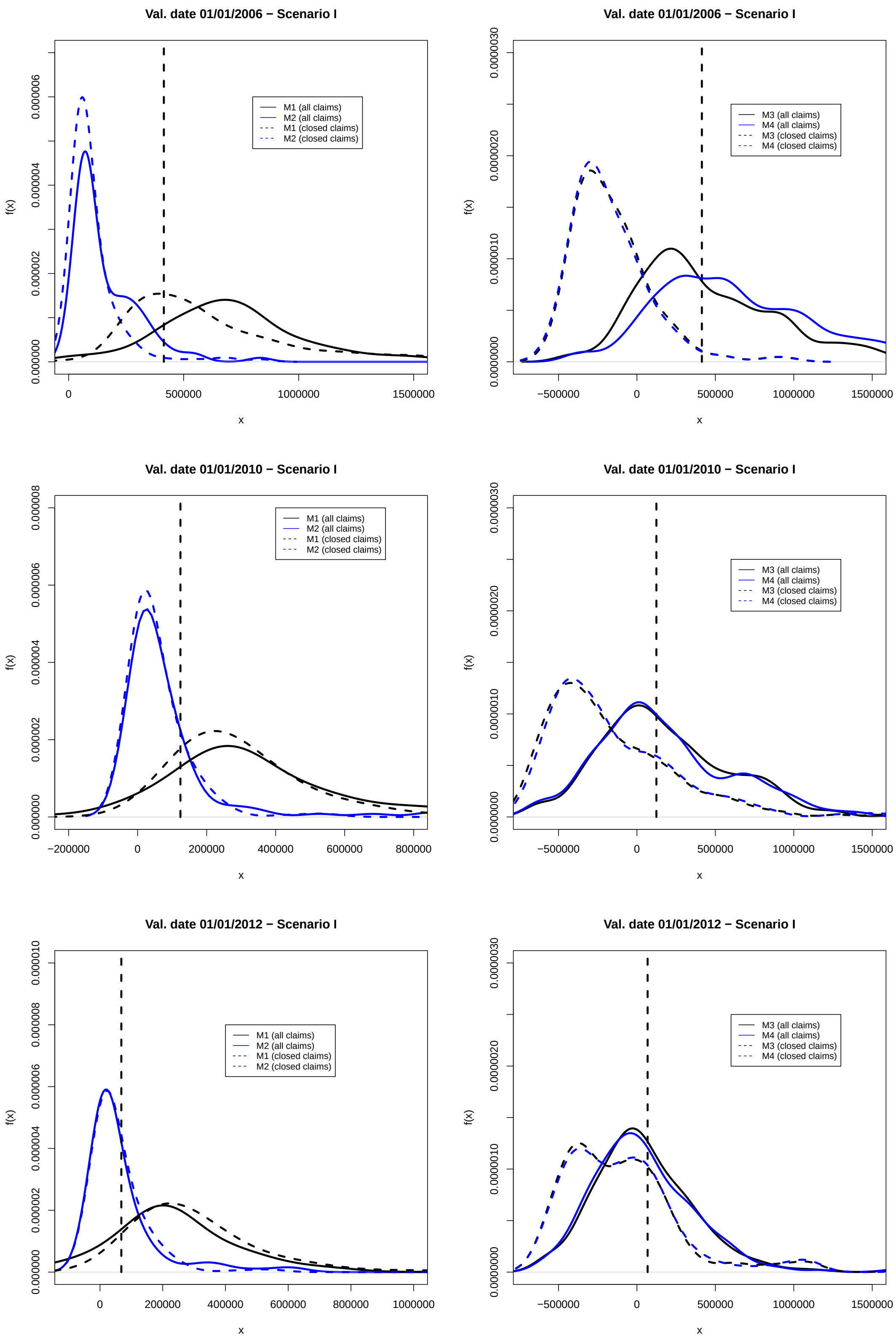

Figure 10: Predictive distribution of the reserve amount using all claims (solid lines) and only closed claims (broken lines) in the calibration process

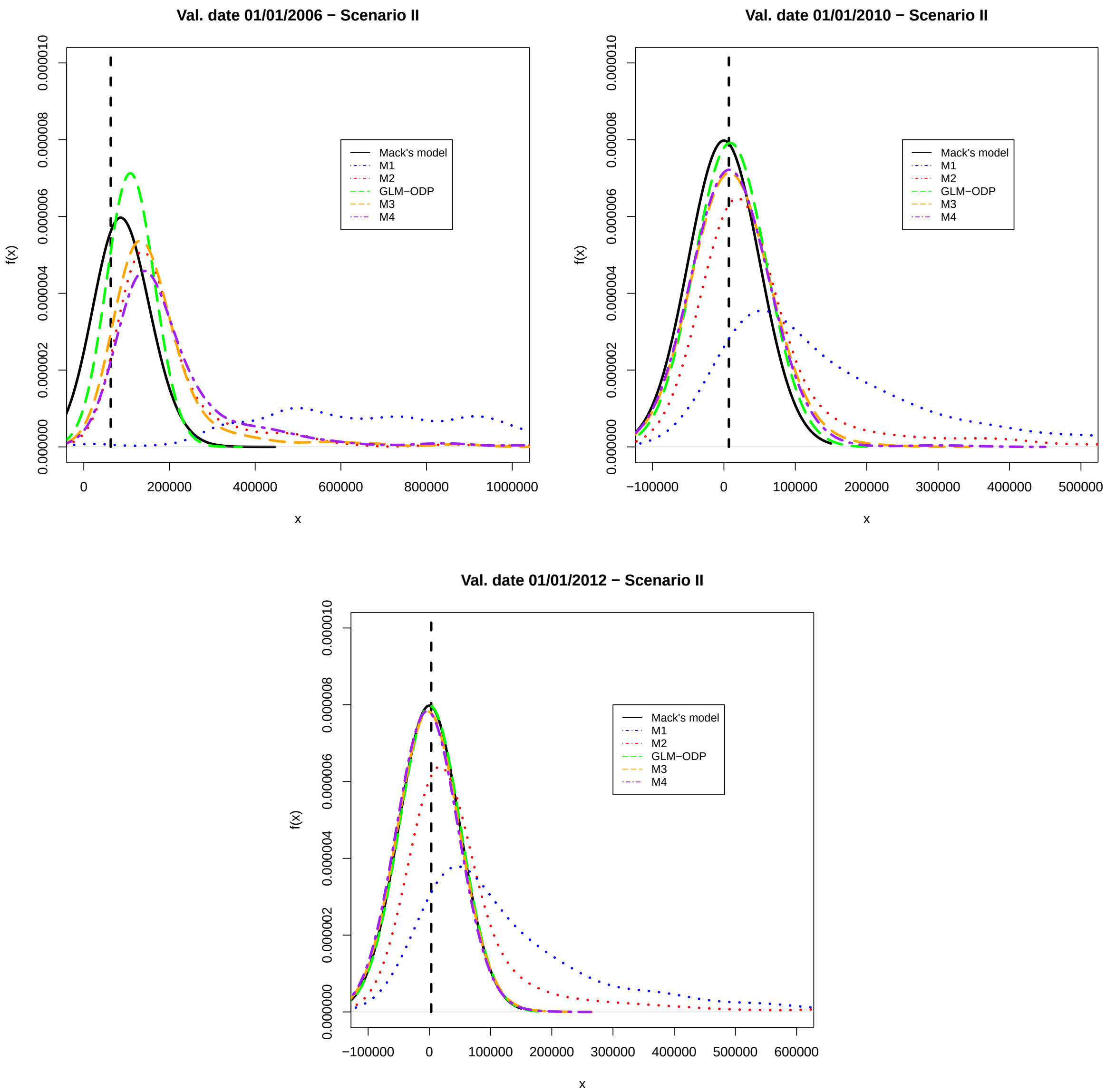

Figure 11: Predictive distribution of the reserve amount. The observed value is \$63,000 for 2006 (top, left), \$7,000 for 2010 (top, right) and \$3,000 for 2012 (bottom)

assumes two lines of business, LoB 1 and 2, and an inflation rate of 5%/year for frequency. We present some descriptive statistics in Table 19 and Figure 15 in Appendix B, and in Table 15. Selected values are $\widehat{q}^{(2006,3)} = 0.8$, $\widehat{q}^{(2006,4)} = 0.4$, $\widehat{q}^{(2010,3)} = 0.9$, $\widehat{q}^{(2010,4)} = 0.9$, $\widehat{q}^{(2012,3)} = 0.95$ and $\widehat{q}^{(2012,4)} = 0.95$.

Table 15: Validation dataset (in $1,000) for Scenario III

Valuation date	% of censored data	RBNS amount	IBNR amount
01/01/2005	9.5	46	19
01/01/2006	10.5	52	8
01/01/2007	4.5	6	6
01/01/2008	3.4	5	6
01/01/2009	2.8	9	0
01/01/2010	2.1	8	0
01/01/2011	1.9	6	0
01/01/2012	1.6	6	0

We present results in Table 16. Figure 12 presents the predictive distribution of the reserve amount using all models for the same 3 levels of portfolio maturity.

Table 16: Predicted Reserve (in $1,000) for Scenario III

Val. year (% cens.) (obs. value)	$\mathbb{E}\left[\widehat{M}_i^{(1)}\right]$	$\mathbb{E}\left[\widehat{M}_i^{(2)}\right]$	$\mathbb{E}\left[\widehat{M}_i^{(3)}\right]$	$\mathbb{E}\left[\widehat{M}_i^{(4)}\right]$	Mack	GLM-ODP
2012 (1.6%) (6)	152	233	23	19	1	1
2010 (2.1%) (8)	164	308	26	23	2	5
2006 (10.5%) (60)	269	1,493	249	420	115	123

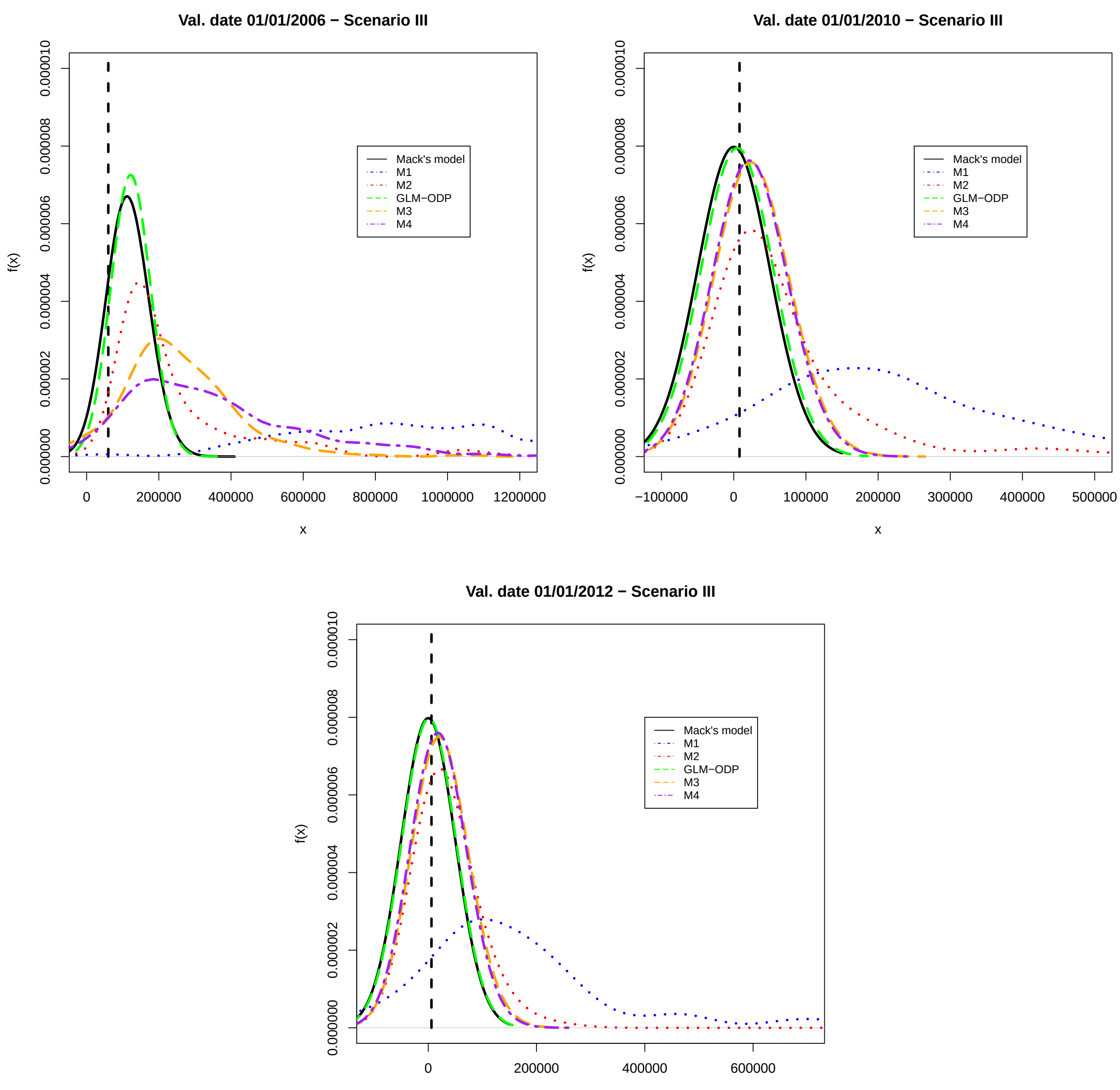

Figure 12: Predictive distribution of the reserve amount. The observed value is \$60,000 for 2006 (top, left), \$8,000 for 2010 (top, right) and \$6,000 for 2012 (bottom)

For all scenarios, we note that the Tree $\widehat{M}_i^{(1)}$ model (blue line) produces very variable reserves resulting in very high expected values and very flattened predictive distributions. This instability is explained by the structure of the estimator which suffers from the lack of data related to the estimation of $\mathbb{I}\left(M > m, Y > z\right)$. This effect is less pronounced for a more mature portfolio because there are much fewer open claims. The Tree $\widehat{M}_i^{(2)}$ model (red line) is much more stable, which is mainly due to the fact that there is more data to estimate $\mathbb{I}\left(Y > z\right)$ than $\mathbb{I}\left(M > m, Y > z\right)$, and that Y is a variable generally less dispersed than M. This conclusion confirms that made in a study (Lopez and Milhaud, 2020) which, among these two estimators, suggests "... *we recommend to use the strategy (B)a)* [Tree $\widehat{M}_i^{(2)}$ model] *to make the reserve predictions, as it outperforms all other methods and shows stable results in terms of prediction error...*".

In all situations, estimators $\widehat{M}_i^{(3)}$ and $\widehat{M}_i^{(4)}$ offer similar performance, which seems to indicate that the use of individual explanatory variables when imputing missing values does not significantly improve the performance of the model. We still add a caveat to this remark due to the small number of micro-level covariates in the database. Furthermore, estimators $\widehat{M}_i^{(3)}$ and $\widehat{M}_i^{(4)}$ require much shorter computation times than estimators $\widehat{M}_i^{(1)}$ and $\widehat{M}_i^{(2)}$. For scenarios II and III, although Tree $\widehat{M}_i^{(2)}$, Tree $\widehat{M}_i^{(3)}$ and Tree $\widehat{M}_i^{(4)}$ models seem appropriate (see Figures 11 and 12), we would recommend the use of Tree $\widehat{M}_i^{(3)}$ model for its simplicity and its saving in computation time.

As a concluding remark, we note that for some scenarios and some estimators, the expected values for the reserve are sometimes quite far from the observed values. However, for all three scenarios, the observed value is always within the range of plausible values.n Moreover, we notice in several cases (for example for scenario III) a skewed predictive distribution, which results in an empirical median always lower than the empirical mean. Thus, the latter is strongly impacted by the slightly more extreme cases observed in the right tail of the distribution.

5.4 Comparison of several future developments

Each of the validation databases contains only one realization of the future development of claims. To enrich the analysis, we now use the *Loss Simulator 2.0* available on the Casualty Actuarial Society website (`https://www.casact.org/research/lsmwp/index.cfm?fa=software`). This tool will allow us to generate several future file developments based on various hypotheses. The *Loss Simulator 2.0* provides claim level simulation functions and is mainly based on Parodi (2013). We focus on two types of dependence: within line of business dependence between reporting delay, settlement delay and severity; and the dependence between claim amounts for two lines of business.

For each scenario, we simulated the exact dates of occurrence, declaration and settlement for each file based on the information available in the database. To do this, we used appropriate uniformity assumptions. Then, the claimFitting function available in the package cascsim tries to fit four distributions for each line of business: (monthly) frequency, total severity, report lag, and settlement lag. Finally, empirical probability of zero-payment claim by development year is also calculated.

Within line of business dependence. We use the copulaFit function to estimate the dependence there may be between severity, reporting lag and settlement lag. We present the results obtained in Table 17. Then, we use the claimSimulation function to simulate future developments of our three validation databases. In order to obtain realizations of a conditional distribution for the settlement lag and for the severity, we slightly modified the code of two functions of the cascsim package, as illustrated in Figure 16.

In Figure 13, we present the predictive distribution of the reserve amount using results in Table 17 and various dependence structures: independent, Frank copula, Gumbel copula, Student copula, and Joe copula.

Table 17: Fitting results for the three scenarios

Component	Scenario I	Scenario II	Scenario III
Severity (LoB A)	$\mathrm{Lognormal}(5.6877, 1.1422)$	$\mathrm{Lognormal}(5.5955, 1.0612)$	$\mathrm{Lognormal}(5.5981, 1.1431)$
(LoB B)	-	$\mathrm{Lognormal}(7.2629, 1.132)$	$\mathrm{Lognormal}(7.2537, 1.2654)$
Report Lag (LoB A)	$\mathrm{Weibull}(1.3921, 223.62)$	$\mathrm{Weibull}(1.3479, 226.6305)$	$\mathrm{Weibull}(1.3822, 218.8912)$
(LoB B)	-	$\mathrm{Weibull}(1.4635, 210.5343)$	$\mathrm{Weibull}(1.362, 214.5786)$
Settlement Lag (LoB A)	$\mathrm{Weibull}(1.5587, 796.87)$	$\mathrm{Weibull}(1.6453, 778.0198)$	$\mathrm{Weibull}(1.4144, 768.9874)$
(LoB B)	-	$\mathrm{Weibull}(1.6382, 760.9602)$	$\mathrm{Weibull}(1.722, 745.8955)$
within LoB dependence			
t (LoB A)	$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1 & 0.04 & -0.03 \\ 0.04 & 1 & 0.06 \\ -0.03 & 0.06 & 1 \end{bmatrix}$	$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1 & 0.14 & 0.05 \\ 0.14 & 1 & 0.09 \\ 0.05 & 0.09 & 1 \end{bmatrix}$	$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1 & 0.01 & 0.09 \\ 0.01 & 1 & 0.05 \\ 0.09 & 0.05 & 1 \end{bmatrix}$
	$\mathrm{df} = 33$	$\mathrm{df} = 142$	$\mathrm{df} = 60$
(LoB B)	-	$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1 & 0.10 & -0.06 \\ 0.10 & 1 & 0.09 \\ -0.06 & 0.09 & 1 \end{bmatrix}$	$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1 & 0.07 & 0.01 \\ 0.07 & 1 & -0.01 \\ 0.01 & -0.01 & 1 \end{bmatrix}$
	-	$\mathrm{df} = 228$	$\mathrm{df} = 22$
Frank (LoB A)	-	$\widehat{\alpha} = 0.433$	$\widehat{\alpha} = 0.1557$
(LoB B)	-	$\widehat{\alpha} = 0.2112$	$\widehat{\alpha} = 0.0482$
Gumbel (LoB A)	$\widehat{\alpha} = 1.0065$	$\widehat{\alpha} = 1.0544$	$\widehat{\alpha} = 1.0264$
(LoB B)	-	$\widehat{\alpha} = 1.0222$	$\widehat{\alpha} = 1.0000$
Joe (LoB A)	$\widehat{\alpha} = 1.0113$	$\widehat{\alpha} = 1.0905$	$\widehat{\alpha} = 1.0471$
(LoB B)	-	$\widehat{\alpha} = 1.0397$	$\widehat{\alpha} = 1.0000$

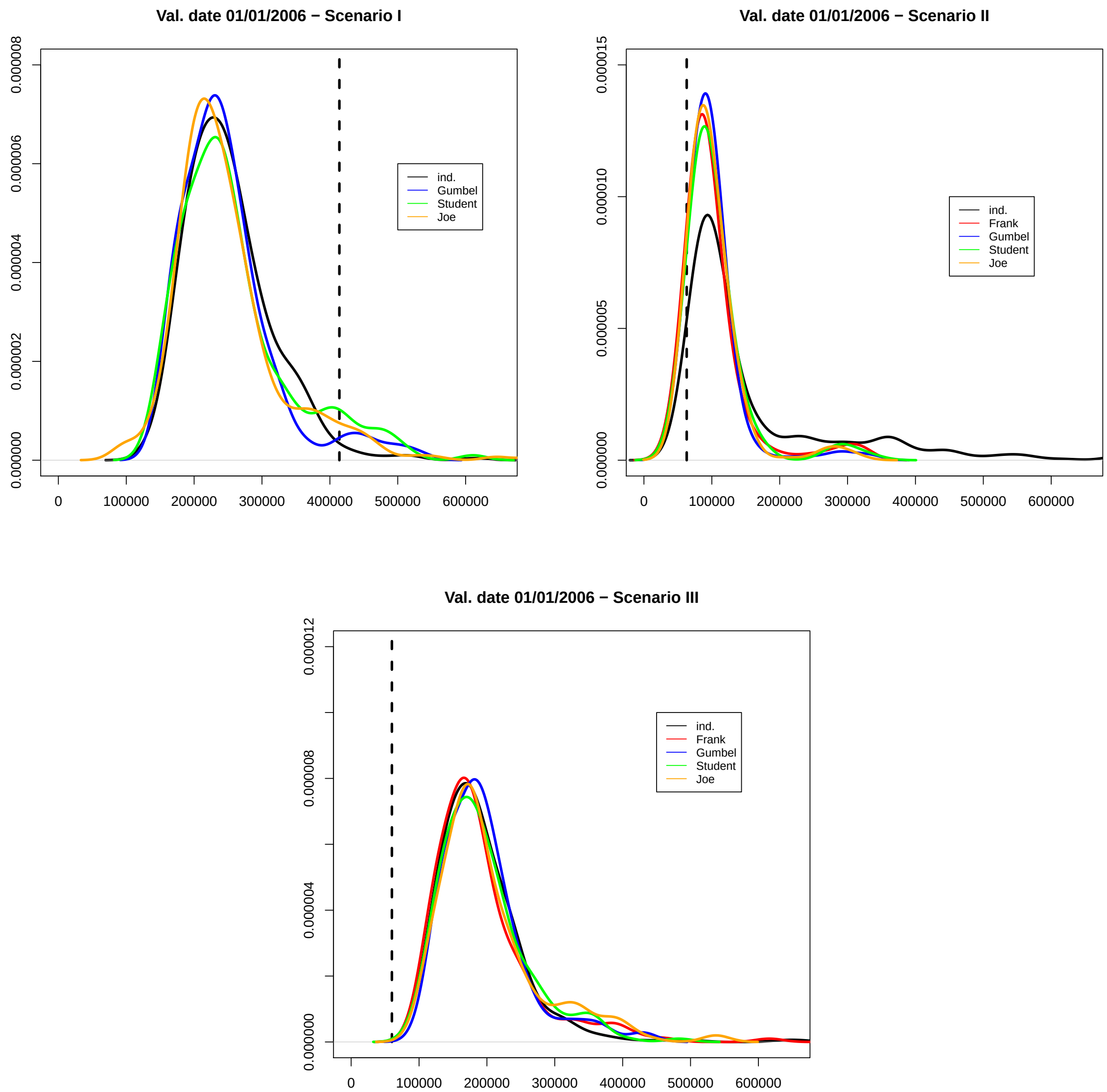

Figure 13: Predictive distribution of the reserve amount. The observed value is $\$414{,}000$ for scenario I (top, left), $\$63{,}000$ for scenario II (top, right) and $\$60{,}000$ for scenario III (bottom)

First observation: it would seem that for the three scenarios, taking dependence into account has a low impact on the predictive distribution except, perhaps, for scenario II where the impact seems a little more visible. This is confirmed with regard to the estimated parameters of the different copulas (see Table 17). Indeed, the values rarely deviate significantly from the values associated with independent random variables. However, in the presence of a dataset where this dependence between severity, reporting lag and settlement lag would be greater, the *Loss Simulator 2.0* tool will allow us to measure this impact on the predictive distribution.

Second observation: we can compare the predictive distribution obtained with the *Loss Simulator 2.0* tool and that obtained with the different models analyzed in the previous subsection. In order to simplify the analysis, we keep only the independence copula model, the Joe copula model (the one selected by the *Loss Simulator 2.0* tool), the $\widehat{M}_i^{(2)}$ model based on survival analysis and the $\widehat{M}_i^{(4)}$ model based on the imputation of missing values. We present results in Figure 14. For scenario I, we note that using a high quantile of the predictive distribution obtained with the $\widehat{M}_i^{(2)}$ and $\widehat{M}_i^{(4)}$ models as a reserve for the portfolio, will make it possible to cover most of the future developments generated by the *Loss Simulator 2.0*. For this scenario, the flattening of the predictive distribution obtained with the $\widehat{M}_i^{(4)}$ model seems unrealistic and suggests favoring the $\widehat{M}_i^{(2)}$ model. For scenarios II and III, the results are similar, except for the $\widehat{M}_i^{(4)}$ model which seems to offer a much more reasonable performance than for scenario I.

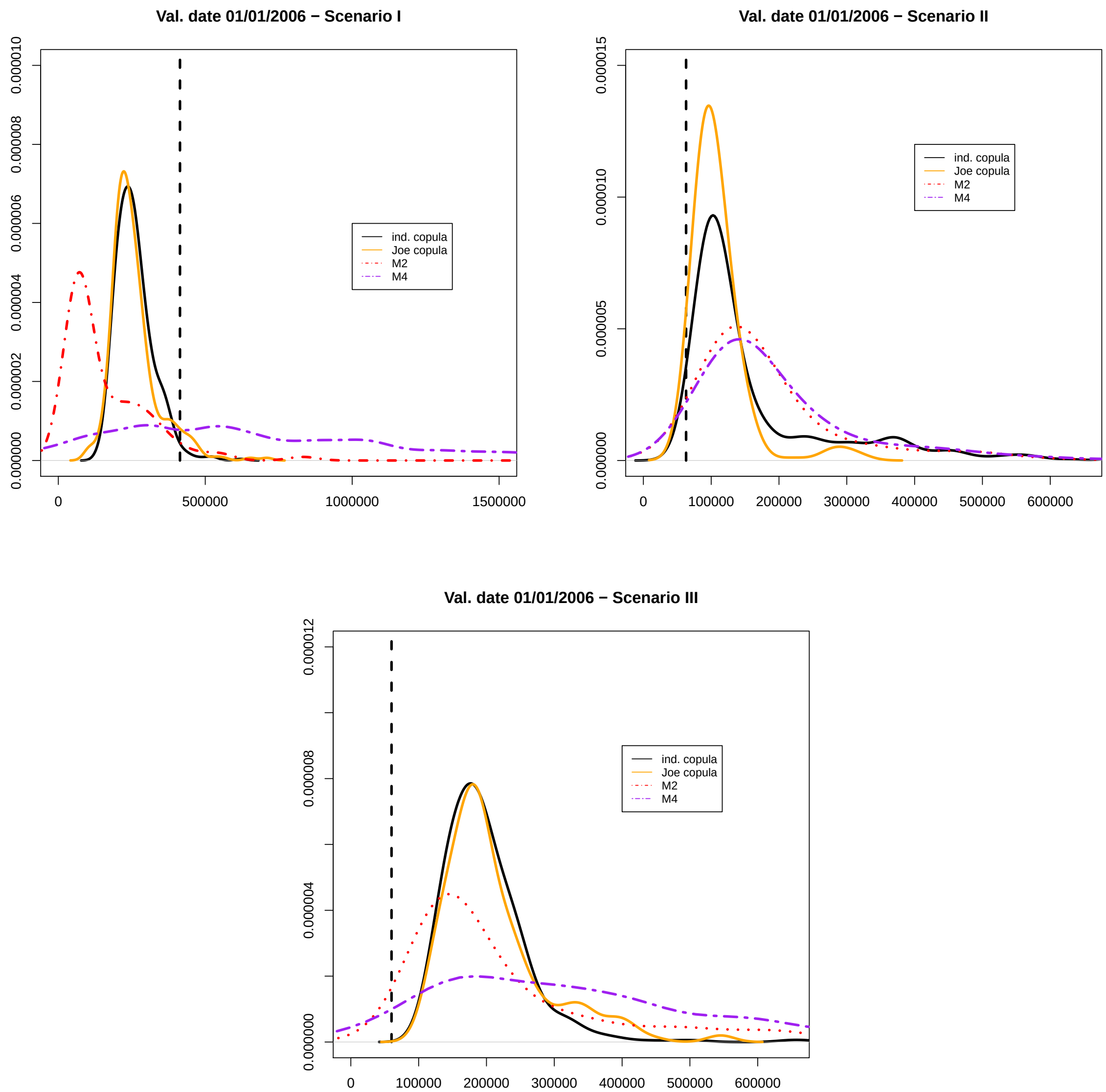

Figure 14: Predictive distribution of the reserve amount for scenario I (top, left), scenario II (top, right) and scenario III (bottom)

Between lines of business dependence. The *Loss Simulator 2.0* tool also makes it possible to include dependence between two, or more, lines of business within a portfolio. However, it allows this dependence to be taken into account only between the monthly claim frequencies. Because we are only interested in open claims and RBNS reserve, this dependence is irrelevant in our analysis. A study of the IBNR reserve could greatly benefit from this tool in order to study the impact of this between lines of business dependence. In addition, the *Individual Claims History Simulation Machine* used to simulate the databases does not allow the generation of several correlated claims in several lines of business. Thus, neither the ICHSM nor the *Loss Simulator 2.0* allows a direct study of the dependence between open claims originating from several lines of business.

To circumvent this limit, we have artificially created a link between some open claims of line of business 1 and line of business 2 (based on scenario II) according to the procedure described below.

- Selection of the proportion p of open claims that will be affected by the dependency.
- Calculation of the number of open files n affected by the dependency

$$n = p\min(n_1, n_2),$$

  where n_1 and n_2 are the number of open claims for lines of business 1 and 2, respectively. If n is not an integer value, it is rounded.
- Random selection of the n open claims in each of the lines of business which will be affected by the dependency.
- For each of the n pairs thus formed, calculation of the total paid amounts using a selected copula with one (or more) parameters, and the marginals selected in Table 17: LN$(5.5955, 1.0612)$ for line of business 1 and LN$(7.2629, 1.132)$ for line of business 2. Note that for each of the open files, the final amount is simulated conditionally on the amount already paid, i.e. it is assumed that the reserve cannot be negative.
- Calculation of the reserve.

We perform this analysis using several proportions p. Because the conclusions were similar, we only present the results for $p = 0.9$ in Table 18. We recall that these results come from simulations obtained using a parametric model (see Table 17) adjusted by the *Loss Simulator 2.0*. We do not consider the probability of a file closing at 0 in this part of our analysis. Combining the dependence between lines of business and the mixed distribution of the total cost per contract (probability mass at 0 and continuous distribution for strictly positive values) would require finer modeling and will be developed in a future project. Thus, the numerical results cannot therefore be directly compared to the different figures presented in the previous subsection. Nevertheless, it is possible to draw general conclusions by considering the independent copula as the base case. We first notice that the expected value is not affected by the dependency structure. Then, a dependency structure increasing the value of Kendall's tau leads to an increase in the variance of the reserve and the Tail Value-at-Risk (TVaR).These results are consistent with the concordance ordering of the chosen copulas, the convex ordering of the sum of risks and its impact on the variance and the TVaR risk measure (see, e.g. Joe (2014) and Denuit *et al.* (2006)).

Table 18: Results for future developments using various copulas between 2 lines of business (based on Scenario II) with 90 % of open claims affected by the dependency

Copula	Par.	Kendall's tau	upper tail dep.	Emp. mean	sdt err.	Emp. TVaR$_{0.95}$	ratio w.r.t. ind.
Ind.	-	0.00	0.00	147,369	31,187	228,626	1.000
Normal	0.2	0.13	0.00	147,318	31,866	230,275	1.007
	0.9	0.71	0.00	147,600	36,160	242,758	1.062
Student	0.4	0.26	0.20	147,480	33,666	235,571	1.030
(df = 4)	0.9	0.71	0.63	147,468	36,089	243,261	1.064
Gumbel	2	0.50	0.59	147,348	35,912	241,605	1.057
	5	0.80	0.85	147,274	36,516	242,869	1.062
	9	0.89	0.92	147,347	36,687	244,296	1.069
Frank	2	0.21	0.00	147,493	32,231	231,197	1.011
	10	0.67	0.00	147,405	34,233	235,214	1.029
Clayton	1	0.34	0.00	147,418	32,156	230,489	1.008
	5	0.71	0.00	147,496	33,827	234,158	1.024
	10	0.83	0.00	147,429	34,532	236,117	1.033

6 Conclusion

The main objective of the project is to analyze how open claims should be integrated into an individual reserve valuation process when statistical learning-based approaches are used. Before tackling this task, we first went through the existing literature on individual loss reserving methodologies using machine learning techniques. We provide a detailed literature review to establish the state-of-the-art regarding machine learning techniques in a loss reserving context. We take from our readings that recent research articles can be divided in two main categories which are granular loss reserve methods relying on some form of decision trees and others based on neural networks. We then pursue with a more detailed analysis of two tree-based methodologies proposed to include open files within the valuation of reserves process. More precisely, we present and discuss the approach of Lopez *et al.* (2016, 2019) using corrective weights based on survival analysis and the one of Duval and Pigeon (2019) using missing data imputation.

After consulting with a few Canadian Property and Casualty insurance companies, we noted a real interest in the development, and use, of granular models for reserves and, more specifically, for methods based on statistical learning techniques. However, these tools are, to our knowledge, mainly at the "research and development" stage. We hope that this report will contribute to the development of these approaches in practice.

With simulated databases obtained using Gabrielli and Wüthrich's simulation machine and for three different scenarios, we compare the performance of these two methodologies, and also with two classical collective loss reserving strategies. From this case-study, we take away the following elements:

- strategy in which open files would be removed from the calibration process is not advisable;
- the two estimators ($\widehat{M}_i^{(1)}$ and $\widehat{M}_i^{(2)}$) proposed in Lopez *et al.* (2016, 2019) behave quite differently in all scenarios. The estimator $\widehat{M}_i^{(2)}$ should be preferred given the stability it has shown compared to $\widehat{M}_i^{(1)}$ which varies greatly;
- the performance of the estimators ($\widehat{M}_i^{(3)}$ and $\widehat{M}_i^{(4)}$) based on Duval and Pigeon (2019) is rather similar in the three scenarios indicating that the individual information embedded in the covariates used in the imputation of missing data does not guide the model to better results;
- the two estimators ($\widehat{M}_i^{(3)}$ and $\widehat{M}_i^{(4)}$) outperform the ones of Lopez *et al.* (2016, 2019) based on Kaplan-Meier weights regarding computation time.

For all three scenarios, we also investigate possible future developments for a claim using the *Loss Simulator 2.0* (Casualty Actuarial Society). Assumptions regarding a possible dependence structure between the claim severity, the reporting delay and the settlement delay using copulas is included. Mainly, we observe the following:

- in all cases, we have not observed a significant dependence between these components. Dependence parameters of all copulas used indicate the independence copula as the appropriate one;
- the comparison of predictive distributions obtained with the *Loss Simulator 2.0* in the previous scenarios does not allow us to favor the estimator $\widehat{M}_i^{(2)}$ or $\widehat{M}_i^{(4)}$ in particular for all of them.

To have an idea of the impact on the reserves of a dependence structure between two lines of business, we used different copulas to include dependency between some losses not related to the same line of business. For every copula, we chose a few values for the dependence parameter to compare results obtained under the assumption of independent lines of business. We find that

- a positive dependence relation between lines of business leads to a small increase in the reserves obtained (here determined through the TVaR risk measure);
- even with strong dependence relations, the impact on the reserves is slight.

In a future work, it would be interesting to reproduce this analysis using a database with more covariates, some of them could be dynamic. In addition, the tools used (*Loss Simulator 2.0* and the Individual Claims History Simulation Machine) clearly limited the study of the impact of a possible dependence within, but mainly between lines of business. We strongly encourage the further development of modules to deal with dependence in these tools.

A Kaplan-Meier Weights

Kaplan-Meier (KM) weights are defined by

$$w_k = \frac{\delta_k}{n(1-\widehat{G}(Z_k^-))}. \tag{9}$$

In order to estimate $G(t) = \Pr(C \leq t)$, Lopez *et al.* (2016) uses the KM estimator given by[4]

$$\widehat{G}(t) = 1 - \prod_{C_i \leq t} \left(1 - \frac{1-\delta_i}{\sum_{j=1}^{n} \mathbb{I}\,(C_j \geq C_i)}\right).$$

[4]Note that because we are estimating the censoring cumulative distribution function, we "observe" $C_i = Z_i$ only for uncensored cases where $\delta_i = 0$.

In order to evaluate KM weights, we can simplify (remember that $Z_1 < Z_2 < \cdots < Z_n$)

$$\begin{aligned}
\widehat{G}(Z_k^-) &= 1 - \prod_{C_i \leq Z_k^-} \left(1 - \frac{1-\delta_i}{\sum_{j=1}^n \mathbb{I}\left(C_j \geq C_i\right)}\right) \\
&= 1 - \prod_{Z_i \leq Z_k^-} \left(1 - \frac{1-\delta_i}{\sum_{j=1}^n \mathbb{I}\left(Z_j \geq Z_i\right)}\right) \\
&= 1 - \prod_{Z_i \leq Z_k^-} \left(\frac{n-i+1-(1-\delta_i)}{n-i+1}\right) \\
&= 1 - \prod_{i=1}^{k-1} \left(\frac{n-i}{n-i+1}\right)^{1-\delta_i}.
\end{aligned}$$

Based on that, the KM weights defined in Lopez *et al.* (2016) are given by

$$\begin{aligned}
w_k = \frac{\delta_k}{n(1-\widehat{G}(Z_k^-))} &= \frac{\delta_k}{n\prod_{i=1}^{k-1}\left(\frac{n-i}{n-i+1}\right)^{1-\delta_i}} \\
&= \left(\frac{\delta_k}{n}\right) \prod_{i=1}^{k-1} \left(\frac{n-i+1}{n-i}\right) \left(\frac{n-i+1}{n-i}\right)^{-\delta_i} \\
&= \left(\frac{\delta_k}{n-k+1}\right) \prod_{i=1}^{k-1} \left(\frac{n-i}{n-i+1}\right)^{\delta_i},
\end{aligned}$$

where $w_1 = \delta_1/n$. Finally, the authors apply the tail correction suggested by Efron (1967):

$$w_n = \prod_{i=1}^{n-1} \left(\frac{n-i}{n-i+1}\right)^{\delta_i}.$$

These weights can be calculated in R using the aft.kmweight function from the imputeYn package.

B Descriptive Statistics

Table 19: Descriptive Statistics

Scenario	LoB	Mean	Median	Sdt. Dev.	Interval	95th quantile	99th quantile
I	–	1,507	292	11,825	(0 − −307,058)	3,188	15,235
II	Overall	1,838	214	26,829	(0 − −863,600)	4,167	10,345
	1(49.2%)	689	276	1,820	(0 − −25,248)	2,285	6,639
	2(50.8%)	2,952	0	37,583	(0 − −863,600)	5,496	13,666
III	Overall	1,748	236	13,662	(0 − −358,335)	4,718	18,229
	1(49.1%)	878	279	2,947	(0 − −38,181)	2,469	11,349
	2(50.9%)	2,587	0	18,902	(0 − −358,335)	5,923	32,446

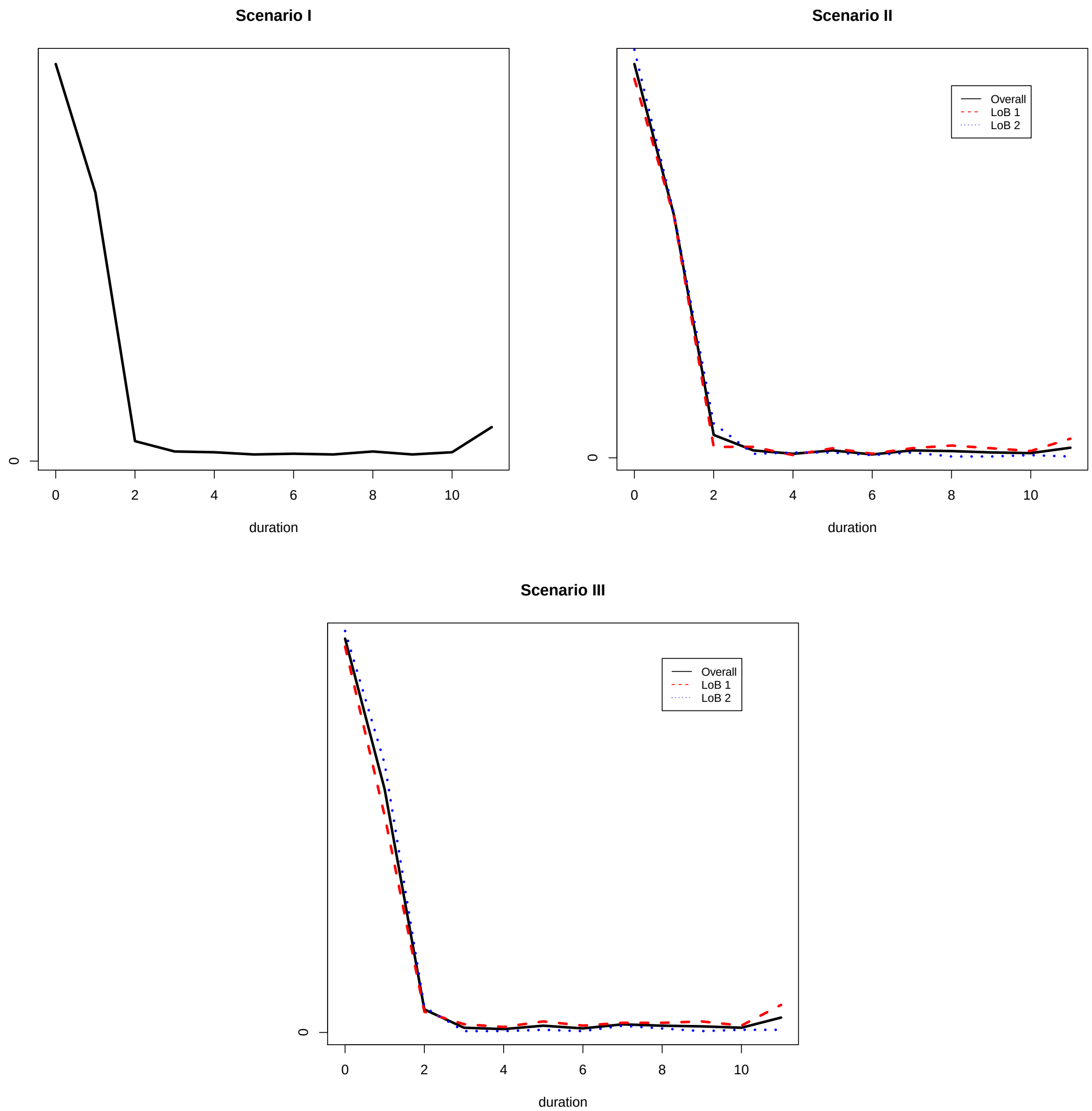

Figure 15: Duration for scenario I (top, left), scenario II (top, right) and scenario III (bottom). The observed mean values are, 1.40, 0.97 and 1.05, respectively.

C R code

```
#########################
### Type.R file        ###
### claimSample method ###
#########################

# Original file
[...]
tmpdata[j,"settlementDate"] <- as.character(as.Date(as.Date(tmpdata[j,"reportDate"])
                               + round(samples[,2])))
tmpdata[j,"ultimateLoss"] <- samples[,1]
[...]

# Modified file
[...]
tmpdata[j,"settlementDate"] <- as.character(as.Date(as.Date(tmpdata[j,"reportDate"])
                               + settlemin + round(samples[,2])))
tmpdata[j,"ultimateLoss"] <- samples[,1]
[...]

##########################
### Distribution.R file ###
### Quantile method     ###
##########################

# Original file (similarly for other distributions)
setMethod("Quantile",signature("Lognormal"), function(object, p) {
if (object@truncated == FALSE) {
qlnorm(p, meanlog =object@p1, sdlog=object@p2)
}else{
qtlnorm(p, meanlog =object@p1, sdlog=object@p2, min=object@min,
        max=max(object@min, object@max))
}
})

# Modified file
setMethod("Quantile",signature("Lognormal"), function(object, p) {
if (object@truncated == FALSE) {
qlnorm(p, meanlog =object@p1, sdlog=object@p2)
}else{
    M <- matrix(c((object@min+1):200000, cumsum(dlnorm((object@min+1):200000,
       meanlog = object@p1, sdlog = object@p2)/plnorm(object@min, meanlog
       = object@p1, sdlog = object@p2, lower.tail = FALSE))), ncol = 2)
    M[which(abs(p - M[,2]) == min(abs(p - M[,2]))), 1]
}
})
```

Figure 16: R code

References

Antonio, K. and Plat, R. (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, 7, 649-669.

Arjas, E. (1989). The claims reserving problem in non-life insurance: some structural ideas. *ASTIN Bulletin*, 19(2), 139-152.

Baudry, M., and Robert, C. Y. (2019). A machine learning approach for individual claims reserving in insurance. *Applied Stochastic Models in Business and Industry*, 35, 1127-1155.

Breiman, L., Friedman J., Olshen R., and Stone C. (1984). *Classification and Regression Trees*. Wadsworth Statistics/Probability Series; New York: Routledge.

Charpentier, A., and Pigeon, M. (2016). Macro vs. micro methods in non-life claims reserving (an econometric perspective). *Risks*, 4, 12.

De Felice, M., and Moriconi, F. (2019). Claim watching and individual claims reserving using classification and regression trees. *Risks*, 7, 102.

Delong, L., and Wüthrich, M.V. (2020). Neural networks for the joint development of individual payments and claim incurred. *Risks*, 8, 33.

Delong, L., Lindholm, M. and Wüthrich, M.V. (2020). Collective reserving using individual claims data, Available at SSRN: `https://ssrn.com/abstract=3582398` (accessed on 15 June 2020).

Denuit, M., Dhaene, J., Goovaerts, M.J., Kaas, R. (2006). *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. Wiley, New York.

Duval, F., and Pigeon, M. (2019). Individual loss reserving using a gradient boosting-based approach. *Risks*, 7, 79.

Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 4, 831-853.

England, P. D., and Verrall, R. J. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal*, 8, 443-544.

Feldblum, S. (1996). NAIC property/casualty insurance company risk-based capital requirements. *Proceedings of the Casualty Actuarial Society*, 83(159), 300-389.

Friedland, J. (2010). Estimating unpaid claims using basic techniques. *Casualty Actuarial Society*, vol. 201.

Gabrielli, A. and Wüthrich, M. V. (2018). An individual claims history simulation machine. *Risks*, 6, 29.

Gabrielli, A., Richman, R. and Wüthrich, M. V. (2020). Neural network embedding of the over-dispersed Poisson reserving model. *Scandinavian Actuarial Journal*, 1, 1-29.

Gabrielli, A. (2019). A neural network boosted double overdispersed Poisson claims reserving model. *ASTIN Bulletin*, 50(1), 25-60.

Harej, B., Gächter, R. and Jamal, S. (2017). Individual claim development with machine learning. *Report of the ASTIN Working Party of the International Actuarial Association.*, Available online: `http://www.actuaries.org/ASTIN/Documents/ASTIN_ICDML_WP_Report_final.pdf` (accessed on 15 June 2020).

Jamal, S. (2018). Machine learning and traditional methods synergy in non-life reserving. *Report of the ASTIN Working Party of the International Actuarial Association.*.

Joe, H. (2014). *Dependence Modeling with Copulas*. Chapman and Hall/CRC.

Kuo, K. (2019). DeepTriangle: A deep learning approach to loss reserving. *Risks*, 7, 97.

Kuo, K. (2020). Individual claims forecasting with Bayesian mixture density networks. *arXiv:2003.02453v1*.

Larsen, C. R. (2007). An individual claims reserving model. *ASTIN Bulletin*, 37, 113-132.

Lopez, O. (2019). A censored copula model for micro-level claim reserving. *Insurance: Mathematics and Economics*, 87(3), 1-14.

Lopez, O., Milhaud, X., and Thérond, P. E. (2016). Tree-based censored regression with applications in insurance. *Electronic Journal of Statistics*, 10(2), 2685-2716.

Lopez, O., Milhaud, X., and Thérond, P. E. (2016). A tree-based algorithm adapted to microlevel reserving and long development claims. *ASTIN Bulletin*, 49(3), 741-762.

Lopez, O., and Milhaud, (2020). Individual reserving and nonparametric estimation of claim amounts subject to large reporting delays *Scandinavian Actuarial Journal*, in press.

Mulquiney, P. (2006). Artificial neural networks in insurance loss reserving. *Paper Presented at the 9th Joint Conference on Information Sciences 2006 - Proceedings, Kaohsiung, Taiwan, 8-11 October; Amsterdam: Atlantis Press.*, Available online: https://www.atlantis-press.com/search?q=mulquiney (accessed on 15 June 2020).

Norberg, R. (1986). A contribution to modeling of IBNR claims. *Scandinavian Actuarial Journal*. 155-203.

Office of the Superintendent of Financial Institutions (2018). Minimum capital test for federally regulated property and casualty insurance companies. url:http://www.osfi-bsif.gc.ca/eng/fi-if/rg-ro/gdn-ort/gl-ld/Pages/mct2018.aspx, (accessed on 15 June 2020).

Parodi, P. (2013). Triangle-free reserving: a non-traditional framework for estimating reserves and reserve uncertainty. url:http://www.actuaries.org.uk/documents/triangle-free-reserving-non-traditional-framework-estimating-reserves-and-reserve, (accessed on 15 June 2020).

Pigeon, M., Antonio, K., and Denuit, M. (2013). Individual loss reserving with the multivariate skew normal framework. *ASTIN Bulletin*, 43, 399-428.

Taylor, G. (2019). Loss reserving models: granular and machine learning forms. *Risks*, 7, 82.

Wüthrich, M. V., and Merz, M. (2008). *Stochastic claims reserving methods in insurance*. Wiley Finance.

Wüthrich, M. V. (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, 6, 465–480.

Wüthrich, M. V. (2018). Neural networks applied to chain-ladder reserving. *European Actuarial Journal*, 8, 407–436.

Zhao, X. B., Zhou, X., and Wang, J. L. (2009). Semiparametric model for prediction of individual claim loss reserving. *Insurance: Mathematics and Economics*, 45, 1-8.