# Individual reserving and nonparametric estimation of claim amounts subject to large reporting delays

Olivier Lopez, Xavier Milhaud

# Individual reserving and nonparametric estimation of claim amounts subject to large reporting delays

Olivier Lopez[1], Xavier Milhaud[2]

[1]Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005 Paris, France

[2]Université de Lyon, Université Claude Bernard Lyon1, ISFA, LSAF, F-69007, Lyon, France

April 16, 2020

**Abstract**

Thanks to nonparametric estimators coming from machine learning, microlevel reserving has become more and more popular for actuaries. Recent research focused on how to integrate the whole information one can have on claims to predict individual reserves, with varying success due to incomplete observations. Using the CART algorithm, we develop new results that allow us to deal with large reporting delays and partially observed explanatory variables. Statistically speaking, we extend CART to take into account truncation of the data, and introduce plug-in estimators. Our applications are based on real-life insurance portfolios embedding Income Protection and Third-Party Liability guarantees. The full knowledge of the claim lifetime is shown to be crucial to predict the individual reserves efficiently.

**Keywords** : reserving, reporting delay, truncation, censoring, CART.

## 1 Introduction

In non-life insurance, the final cost is rarely known when the claim occurs. In some cases, many years may pass before the amounts are settled. Standard procedures to predict the reserve, like Chain-Ladder (CL) approaches, rely on aggregated cumulative claim amounts. These techniques have the disadvantage that they do not exploit additional information one may have on the claims, and that may help to improve the prediction. Our paper aims to use this information to better catch the heterogeneity of the data, and

get accurate predictions of the individual reserves. In this view, we propose a CART-based approach (Breiman et al. [1984]) that compensates for the presence of censoring and truncation, naturally present in this kind of data. CART allows us to deal with non-linearities between the response and the explanatory factors while preserving simplicity of interpretation, and has therefore gained much popularity in actuarial sciences in the last decade (see Wüthrich [2018a], Baudry and Robert [2019] for recent papers related to CART and reserving). Note that other machine learning techniques were recently used for reserving applications, see for instance Duval and Pigeon [2019] and Wüthrich [2018b].

This paper is based on two main ideas. First, the time before settlement gives meaningful information that helps to explain the final cost of claims. Second, the reporting delays can strongly affect the prediction of claim amounts. Indeed, right-censoring and left-truncation are known to have a significant impact when there is an underlying time phenomenon (Fleming and Harrington [2011a]). In reserving, most claims handling experts agree to say that the claim lifetime plays a crucial role to explain its final amount. Typically, claims that take a long time to be settled are more likely to be expensive, as pointed by Maegebier [2013], Spierdijk and Koning [2011], Pitt [2007], or Bluhm [1993]. Calibrating a model only based on closed claims would thus lead to underestimations of the average claim amounts, since such a procedure would rely on observations with an overrepresentation of short lifetimes. Trying to better understand the relationship between models based on aggregate data (e.g. Chain Ladder) and those using individual data in the context of survival analysis has been done in numerous recent works; such as Bischofberger et al. [2019], Hiabu [2017], and Miranda et al. [2013]. They discuss the advantages and drawbacks of the most famous reserving techniques, and suggest to link the development of claims to the hazard rate of the cost distribution. This approach bridges the gap between both families of methods, and allow to clarify some assumptions underlying aggregate models. Nevertheless, no explanatory variables were considered in these works, contrary to the paper by Lopez et al. [2019] who propose a tree-based estimator to predict the residual lifetimes of still open claims (Reported But Not Settled, or RBNS claims). We extend the latter work in three ways. Firstly, as we cannot estimate the final claim amount efficiently using partially observed risk factors (such as the claim lifetime), we introduce plug-in strategies to overcome this issue. Secondly, as large reporting delays lead to biased predictions, we must modify the CART estimation procedure. We thus determine consistent weights given to each observation to cancel (asymptotically) the bias caused by censoring and truncation. Finally, we use bootstrap resampling to estimate the

predictive uncertainty of the estimators introduced thereafter.

The remainder of the paper is organized as follows. Section 2 describes the general framework and gives some data processing details. Our main contributions are presented in Section 3, including the theoretical extension that allows us to deal with both partially observed risk factors and large reporting delays. Finally, Section 5 is devoted to real data analyses. We compare our results to two competing approaches used in the insurance industry, namely the Collective Reserving Model (close to Chain Ladder for predicted reserves, see Wahl et al. [2019]), and the semiparametric Cox model (Cox [1972]).

## 2 General framework and data management

Consider $n$ claims, with amounts $(M_i)_{1\leq i\leq n}$. For some of them, the final amount $M_i$ is not observed, since the claim is still open. Indeed, the (random) time before the claim $i$ is fully settled (the so-called claim lifetime), denoted by $T_i$, is censored. Introducing some censoring variables $(C_i)_{1\leq i\leq n}$, we define $(Y_i, \delta_i, N_i, \mathbf{X}_i)$ as i.i.d. replications of

$$\begin{cases} Y &= \min(T, C), \\ \delta &= \mathbf{1}_{T\leq C}, \\ N &= \delta M, \\ \mathbf{X} &= (X^{(1)}, ..., X^{(d)}) \in \mathbb{R}^d, \end{cases}$$

where $\mathbf{X}$ are the claim covariates. Note that defining $M$ and $N$ this way amounts to consider one single payment per claim, which is unrealistic in reality. However, the information on payments already made for unsettled claims is most of time unavailable to actuaries, who are usually not in charge of collecting these information.

**Remark 2.1.** *When one has additional information on $M$ (e.g. partial payments already made for RBNS claims), taking it into account would be straightforward in our setting. There basically exists two possibilities: integrate it as a covariate ($N$ would be the cumulated amount already paid, and we know that $M \geq N$), or consider bivariate censoring.*

These variables are easily built from original data. Let's say that $d_i$ is the date at which the $i$th claim occurs and $s_i$ the date at which it is fully settled. Moreover, introduce $f_i$ the date at which the claim stops being observed. We have $T_i = s_i - d_i$, and $C_i = f_i - d_i$. In practice $f_i$ is often the same date for all the claims, due to data collection. In claim reserving applications, reporting delays must be carefully dealt with. Indeed, the database only reports claims that have been communicated, meaning that "Incurred But Not yet

Reported" (the so-called IBNyR) claims are absent. This phenomenon relates to left-truncation, but is different from the classical left-truncation in the literature. Introducing the reporting delay

$$\tau_i = r_i - d_i$$

where $r_i$ is the reporting date of claim $i$, only claims such that $C_i \geq \tau_i$ are observed. Figure 1 illustrates the case of an unknown claim, and reports the typical information one has on claim history. This truncation phenomenon is not standard, compared to the classical left-truncation model in survival analysis where it is usually assumed that observation only occurs when $T \geq \tau$ (or when $Y \geq \tau$, see respectively Tsai et al. [1987]) and Sellero et al. [2005]). Here, the situation is slightly different: $T < \tau$ means that the claim has been settled before reporting. This can happen for instance if the indemnity has been fixed in advance, and the claim is thus stored in the database. To sum up, the observations are made of $(Y_i, \delta_i, N_i, \tau_i, \mathbf{X}_i)_{1 \leq i \leq n}$ i.i.d. with same distribution as $(Y, \delta, N, \tau, \mathbf{X})$, conditionally to $C \geq \tau$.

As already mentioned, both censoring $C$ and truncation $\tau$ induce bias in the analysis if they are not taken into account. The former leads to under-represent claims with high amounts, due to the positive correlation between the claim lifetime $T$ and the final claim amount $M$. The latter impacts claims that occur just before the extraction of the data. It is thus necessary to introduce a statistical framework in which such phenomenons can be considered, see Section 3. On top of that, our goal is twofold. First, we would like to understand the impact of $T_i$ and $\mathbf{X}_i$ on $M_i$, and typically estimate the final cost of a claim with characteristics $\mathbf{X}_i$ and lifetime $T_i$, that is $E[M_i \,|\, T_i, \mathbf{X}_i]$. Second, we aim to predict the final cost of RBNS claims, i.e. $E[M_i | T_i \geq Y_i, \delta_i = 0, \mathbf{X}_i]$, which would enable us to determine appropriate individual reserves for still open claims.

**Remark 2.2.** *In the same spirit as in other reserving techniques, one has to remove the inflation affecting historical information on $M$ to ensure that the amounts are independent and identically distributed. Various methodologies can be used, among which using external data sources when available. Otherwise, one could follow the procedure by Lopez [2019]. For the sake of conciseness, corresponding technical details are moved to Appendix A.*

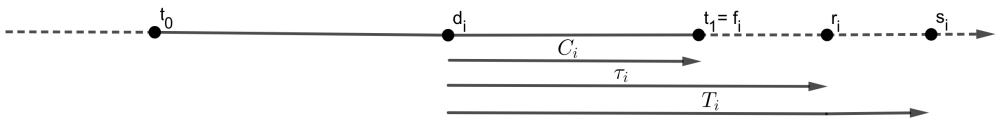


Figure 1: Case of an unknown claim. The observation period begins at $t_0$ and stops at $t_1$.

# 3 Extensions to predict RBNS claims

We introduce in this section the statistical developments for the CART algorithm to be suited to reserving applications.

## 3.1 Managing reporting delays: new "Kaplan-Meier weights"

From a statistical viewpoint, the main difficulty when dealing with censored and truncated data is the fact that the classical empirical means become unadapted. If one wishes to estimate $q_\phi = E[|\phi(T, M, \mathbf{X})|] < \infty$ (for a given measurable function $\phi$), one may write $q_\phi$ as an integral, that is

$$q_\phi = \int \phi(t, m, \mathbf{x}) dF(t, m, \mathbf{x}),$$

where $F(t, m, \mathbf{x}) = \mathbb{P}(T \leq t, M \leq m, \mathbf{X} \leq \mathbf{x})$, and plug a consistent estimator of $F$. In absence of censoring and truncation, a natural estimator of $F$ is the empirical distribution function, leading to an estimation of $q_\phi$ through an empirical mean.

Now, consider that some observations are either partially observed or totally unobserved, due respectively to censoring and truncation. For simplicity matters, consider that we wish to estimate the marginal cumulative distribution function of some lifetime T, i.e. $F_T(t) = \mathbb{P}(T \leq t) = 1 - S(t)$, where $S(t)$ is the survival function. In our case, the fact that truncation occurs when $C \leq \tau$ impacts the way to estimate $F_T$, and makes the estimator differ from the classical one under left-truncation (Tsai et al. [1987]). To introduce this new estimator, one needs to state the following assumption.

**Assumption 1.** *$\tau$ is independent from $(C, M, T, \mathbf{X})$ and $C$ is independent of $(M, T, \mathbf{X})$.*

By analogy to standard techniques like Chain Ladder, this assumption means that we assume homogeneity in terms of time in all dimensions (occurence, development, calendar).

**Proposition 1.** *Provided that Assumption 1 is satisfied, the cumulative distribution function of the lifetime $T$ can be estimated by*

$$\hat{F}_T(t) = 1 - \prod_{Y_i \leq t} \left(1 - \frac{\delta_i \mathbf{1}_{\tau_i < Y_i}}{\sum_{j=1}^n \mathbf{1}_{\tau_j < Y_i \leq Y_j}}\right),$$

*where we assume that there is no ties among the variables $(Y_i)_{1 \leq i \leq n}$.*

*Proof.* For the sake of conciseness, the proof is moved to Appendix B. An alternative formula is also proposed in case of ties. □

Note that $\hat{F}_T$ is piecewise constant, and can therefore be written through the following additive formula:

$$\hat{F}_T(t) = \sum_{i=1}^{n} w_{i,n} \mathbf{1}_{Y_i \leq t},$$

with

$$w_{i,n} = \frac{\delta_i \mathbf{1}_{\tau_i < Y_i}}{\sum_{j=1}^{n} \mathbf{1}_{\tau_j < Y_i \leq Y_j}} \prod_{Y_k < Y_i} \left( 1 - \frac{\delta_k \mathbf{1}_{\tau_k < Y_k}}{\sum_{j=1}^{n} \mathbf{1}_{\tau_j < Y_k \leq Y_j}} \right). \tag{3.1}$$

Since they take truncation into account, these weights allow to deal with the presence of reporting delays. These new weights slightly differ from the ones of the classical product-limit estimator in presence of right-censoring and left-truncation, see Gross and Lai [1996]. Practically speaking, $w_{i,n}$ equals 0 when the observation is censored or truncated. The quantity $w_{i,n}$ can be thought as the weight to put on the $i$th observation to correct the bias caused by censoring and truncation. More weight is given to lowest and largest observations of $T$. Based on the idea of Sellero et al. [2005], one could use the same weight when it comes to estimate the joint cumulative distribution function $F$. Hence, we define

$$\hat{F}(t, m, \mathbf{x}) = \sum_{i=1}^{n} w_{i,n} \, \mathbf{1}_{Y_i \leq t, N_i \leq m, \mathbf{X}_i \leq \mathbf{x}}.$$

A natural estimator of $q_\phi$ is thus given by

$$\hat{q}_\phi = \int \phi(t, m, \mathbf{x}) \, d\hat{F}(t, m, \mathbf{x}) = \sum_{i=1}^{n} w_{i,n} \, \phi(Y_i, N_i, \mathbf{X}_i). \tag{3.2}$$

That being said, we can now introduce the weighting procedure used within the CART algorithm to deal with truncation and censoring.

## 3.2 Weighted regression-tree procedure

To extend the CART algorithm to predict the individual final claim amount $M$, we adopt the same framework as in Lopez et al. [2019]. However, the weighting scheme differs because of the truncation due to reporting delays. Moreover, the use itself of the algorithm (denoted further wCART) for prediction purpose has to be adapted since $T$, considered as an explanatory variable of $M$ thereafter, is partially observed for RBNS claims. For the paper to be self-contained, the complete version of wCART is given in Appendix C. Here, for brevity, we only remind to the reader the main principles.

Suppose that we want to estimate $\pi(\mathbf{z}) = E[\phi(M) \,|\, \mathbf{Z} = \mathbf{z}]$, where $\mathbf{Z} = (T, \mathbf{X})$. At each step of the algorithm, one determines a rule $\mathbf{z} = (t, x^{(1)}, ..., x^{(d)}) \to R_j(\mathbf{z})$ to split

the data and create two partitions of the covariate space (leading to the binary tree structure). That is, for each value of $\mathbf{z}$, $R_j(\mathbf{z})$ equals either 0 or 1 depending on whether some conditions are satisfied by $\mathbf{z}$. Of course $R_j(\mathbf{z})R_{j'}(\mathbf{z})$ equals 0 when $j \neq j'$, and $\sum_j R_j(\mathbf{z})$ equals 1 to ensure that created subsets are exhaustive and disjoint. To select the best rule, we use a weighted quadratic loss so as to minimize within-node variances of the response $\phi(M)$ in created subsets (instead of a quadratic loss in the classical CART algorithm, well-suited to the estimation of an expectation with fully observed variables). Not surprisingly, the weights $w_{i,n}$ given by (3.1) are considered to compensate for censoring and truncation. Finally, each set of rules $\mathcal{R} = (R_1, ..., R_K)$ is associated with a tree-based estimator of the regression function, that is

$$\hat{\pi}^{\mathcal{R}}(\mathbf{Z}) = \sum_{j=1}^{K} \hat{\pi}_j R_j(\mathbf{Z}), \quad \text{where} \quad \hat{\pi}_j = \frac{\sum_{i=1}^{n} w_{i,n}\, \phi(N_i)\, R_j(\mathbf{Z}_i)}{\sum_{i=1}^{n} w_{i,n}\, R_j(\mathbf{Z}_i)}. \tag{3.3}$$

**Remark 3.1.** *Without any stopping rule, this procedure ends up with a complex tree which is very likely to overfit the data. This estimator is thus not satisfactory to estimate $\pi$. To get an estimator with lower dimension, a pruning step is required. Let $K(\mathcal{R})$ denote the number of leaves of a subtree: the pruning approach consists of minimizing the following penalized loss,*

$$\sum_{i=1}^{n} w_{i,n} \left(\phi(N_i) - \hat{\pi}^{\mathcal{R}}(\mathbf{Z}_i)\right)^2 + \alpha\, \frac{K(\mathcal{R})}{n},$$

*where $\alpha > 0$ is a tuning parameter, usually chosen through cross-validation.*

These building and pruning strategies lead to consistent estimators which are asymptotically unbiased, thanks to the properties of Kaplan-Meier estimators (see Lopez et al. [2016] and references therein). Section 4 illustrates such consistency via simulations. Nonetheless, for reserving applications, this estimator cannot be applied directly. Indeed, the explanatory vector $\mathbf{Z}$ is not fully observed, because $T$ can be censored or truncated. We therefore propose some possibilities to overcome this issue.

## 3.3 Strategies to compute the reserves of RBNS claims

Take $\phi(M) = M$. From the selected tree estimator, it is possible to deduce a predictor of $M$ for claims where $\mathbf{Z}$ is fully observed. However, in the context of RBNS claims, $T$ is censored. It means that $\mathbf{Z}$ is partially observed. Denote by $y$ the observed duration. To provide a reasonable estimator of $M$ when $\delta = 0$, one thus needs to take into account that $T \geq y$. Consider the two following approaches.

(A) Bayes - the best prediction given the available data is $E[M \,|\, Y = y, \delta = 0, \mathbf{X} = \mathbf{x}]$, which can also be written as

$$M^\star = E[M \,|\, T \geq y, \mathbf{X} = \mathbf{x}] = \frac{E[M \mathbf{1}_{T \geq y} \,|\, \mathbf{X} = \mathbf{x}]}{E[\mathbf{1}_{T \geq y} \,|\, \mathbf{X} = \mathbf{x}]}. \tag{3.4}$$

Two trees are built using wCART, to estimate the numerator and denominator.

(B) Plug-in - build one tree $\hat{\pi}$ by wCART to estimate $\pi(t, \mathbf{x}) = E[M \,|\, T = t, \mathbf{X} = \mathbf{x}]$. Then fit a model for $T \,|\, T \geq y, \mathbf{X} = \mathbf{x}$, from which a prediction $\hat{T}(y, \mathbf{x})$ can be computed. Finally, predict $M^\star$ using the plug-in principle: $\widehat{M^\star} = \hat{\pi}(\hat{T}(y, \mathbf{x}), \mathbf{x})$.

In the latter approach, several prediction models $\hat{T}$ can be used, among which any machine learning prediction model adapted to censoring and truncation. In the sequel, we consider the three following cases:

(B1) : build two different wCART estimators to estimate $r_{1,y}(\mathbf{x}) = E[T \mathbf{1}_{T \geq y} | \mathbf{X} = \mathbf{x}]$ and $r_{2,y}(\mathbf{x}) = E[\mathbf{1}_{T \geq y} | \mathbf{X} = \mathbf{x}]$, and compute $\hat{T}(y, \mathbf{x}) = \hat{r}_{1,y}(\mathbf{x}) \hat{r}_{2,y}(\mathbf{x})^{-1}$;

(B2) : use a simplified prediction of $T$, assuming that $T$ does not depend on $\mathbf{X}$:

$$\hat{T}(y, \mathbf{x}) = \hat{T}(y) = \frac{\int t \mathbf{1}_{t > y} d\hat{F}_T(t)}{\int \mathbf{1}_{t > y} d\hat{F}_T(t)} = \frac{\sum_{i=1}^n w_{i,n} Y_i \mathbf{1}_{Y_i \geq y}}{\sum_{j=1}^n w_{j,n} \mathbf{1}_{Y_j \geq y}}.$$

Although inconsistent in full generality, the simplicity of this approximation is expected to favor good behavior of this strategy.

(B3) : get $\hat{T}(y, \mathbf{x})$ from the semiparametric Cox model, see Cox [1972].

For comparison purposes, we consider two other naive competitors, where the input information about $T$ is not realistic:

(B4) : predict $M^\star$ by $\widehat{M^\star} = \hat{\pi}(y, \mathbf{x})$, where $y$ is the observed censored duration;

(B5) : predict $M^\star$ by $\widehat{M^\star} = \hat{\pi}(\hat{r}(\mathbf{x}), \mathbf{x})$, where $\hat{r}(\mathbf{x})$ is an estimator of $r(\mathbf{x}) = E[T \,|\, \mathbf{X} = \mathbf{x}]$ obtained by wCART.

Given $\widehat{M^\star}$, the individual reserves as well as the global reserve can easily be obtained. In our setting, this amounts to keep $\widehat{M_i^\star}$ as the estimator of the $i$th individual reserve, since we do not have information on partial payments already made. Strategies (B4) and (B5) are expected to underestimate the individual reserves, since they do not consider the full information about $T$ (i.e. $T > y$, knowing that $E[T \,|\, \mathbf{X}] \leq k + E[T - k \,|\, T \geq k, \mathbf{X}]$ for all $k \geq 0$). On the contrary, strategies (A), (B1), (B2) and (B3) take it into account, and are thus expected to provide well-estimated reserves on average.

# 4 Simulations

First we investigate the practical behaviour of our tree-based estimator with truncated and censored data, where truncation is defined as in Section 2. For the sake of simplicity, we consider the case where we are interested in the distribution of the lifetime $T$, thus focusing on estimating $\pi_0(\mathbf{x}) = E[T \,|\, \mathbf{X} = \mathbf{x}]$. Consider the following simulation scheme:

1. draw $n$ i.i.d. replications $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ of the covariate, with $\mathbf{X}_i \sim \mathcal{U}(0,1)$;
2. draw $n$ i.i.d. lifetimes $(T_1, \ldots, T_n)$ following an exponential distribution such that $T_i \sim \mathcal{E}(\beta = \alpha_1 \mathbb{1}_{\mathbf{X}_i \in [a,b[} + \alpha_2 \mathbb{1}_{\mathbf{X}_i \in [b,c[} + \alpha_3 \mathbb{1}_{\mathbf{X}_i \in [c,d[} + \alpha_4 \mathbb{1}_{\mathbf{X}_i \in [d,e]})$.
3. draw $n$ i.i.d. censoring times (Pareto-distributed: $C_i \sim \mathcal{P}areto(\lambda, \mu)$) and $n$ i.i.d. truncation times, uniformly distributed ($\tau_i \sim \mathcal{U}(\gamma, \delta)$);
4. from the simulated lifetimes, censoring and truncation times, get for all $i$ the actual observed lifetime $Y_i$ and the indicator $\delta_i = \mathbf{1}_{T_i \leq C_i}$;
5. compute the weights $w_{i,n}$ from the entire generated sample $(Y_i, C_i, \tau_i, \delta_i)_{1 \leq i \leq n}$.

Parameter values are stored in Table 1, and descriptive statistics corresponding to the various simulated datasets are available in Table 4 of Appendix D. To each simulated sample, we fit a regression tree with the algorithm of Section 3.2. Then, we compute the weighted squared errors given by $WSE_i = w_{i,n}(\hat{\pi}_{l(i)} - \pi_0(\mathbf{X}_i))^2$, where $\hat{\pi}_{l(i)}$ is deduced from (3.3) for the $i$th observation that belongs to leaf $l(i)$, and where we know that $\pi_0(\mathbf{X}_i) = 1/\beta$. To gain some robustness, we repeated 100 times the simulation scheme above to compute empirical means of $WSE_i$, leading to the $MWSE$. We also considered different values for $(\lambda, \mu)$ and $(\gamma, \delta)$ in the censoring and truncation processes so as to measure the impact of both phenomenons on the procedure's performance (see Table 1). Figure 2 and Table 2 report the results. While the sample size remains low, the censoring and truncation phenomenons seem to significantly impact the performance. This impact seems to be complex when both truncation and censoring are present. However, as soon as there are enough observations, the weighted CART estimator with weights $w_{i,n}$ provides satisfactory results.

| Group-specific means | | | | Component probabilities | | | | Censorship rate | | | Truncation rate | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $[a,b[$ | $[b,c[$ | $[c,d[$ | $[d,e]$ | 10% | 30% | 50% | 5% | 15% |
| 0.08 | 0.05 | 0.16 | 0.5 | $[0,0.3[$ | $[0.3,0.6[$ | $[0.6,0.8[$ | $[0.8,1]$ | $(\lambda,\mu)$ | $(\lambda,\mu)$ | $(\lambda,\mu)$ | $(\gamma,\delta)$ | $(\gamma,\delta)$ |
| 12.5 | 20 | 6.25 | 2 | 30% | 30% | 20% | 20% | (80,1.03) | (20,1.2) | (14,2) | (0,0.3) | (0,1.2) |

Table 1: Parameters involved in the simulation scheme.

| Truncation rate (%) | Censoring rate (%) | Sample size ($n$) | Group-specific MWSE Group 1 | Group 2 | Group 3 | Group 4 | MWSE Global |
|---|---|---|---|---|---|---|---|
| | | 100 | 2.4480 | 20.8320000 | 1.61175 | 2.12600 | 2.23375 |
| | 10% | 1000 | 0.4705 | 11.7646667 | 0.03925 | 0.00975 | 1.10795 |
| | | 5000 | 0.0034 | 0.6306333 | 0.00120 | 0.00020 | 0.05720 |
| | | 100 | 5.9335000 | 53.97467 | 2.1965 | 2.32850 | 5.24670 |
| 15% | 30% | 1000 | 4.4461667 | 15.79833 | 0.0375 | 0.00225 | 1.84215 |
| | | 5000 | 0.1077333 | 1.21450 | 0.0006 | 0.00005 | 0.11877 |
| | | 100 | 5.639833 | 13.9091667 | 2.64475 | 4.27975 | 2.03135 |
| | 50% | 1000 | 2.069667 | 10.3565000 | 0.13850 | 0.15400 | 1.12775 |
| | | 5000 | 0.098900 | 0.9647667 | 0.00100 | 0.00000 | 0.09480 |

Table 2: Mean weighted squared errors w.r.t. the censoring rate and sample size, in the case where the truncation rate equals 15%.

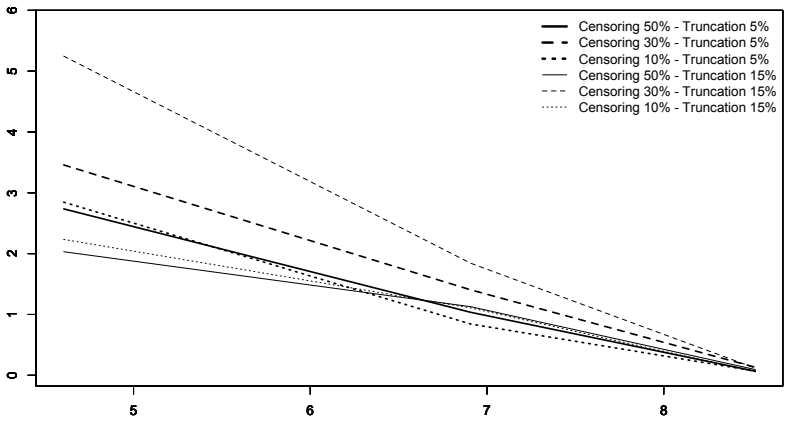


Figure 2: MWSE as a function of the logarithm of the sample size ($n = 100, 1000, 5000$).

# 5 Applications

To be in line with regulatory constraints and current practices in insurance companies, the assessment of reserves is considered on a quarterly basis (note that we changed this time step to check the stability of our conclusions, which was confirmed). Building the database and running the estimations is thus performed every three months, with updated policyholders' features and claim characteristics. Each time, we split our $n$-sized data into two independent subsamples to assess the prediction power of the methods: the learning set with $n^l$ observations (representing two third of the dataset), and the test set containing the remaining $n^t$ observations (from which the predictive power can be assessed). In practice, we backtest our reserve predictions. To do so, we only consider the closed claims (among the $n^t$ observations) at the last observed date $t_1$. This way, the true outcomes $M$ and the true global reserve $P$ are available and can be compared to the predictions. The predictive uncertainty of our estimators $\hat{M}^\star$ and $\hat{P}$, where $\hat{P}$ stands for the global

estimated reserve, is approximated using bootstrap resampling (see Björkwall et al. [2009], Pinheiro et al. [2003], England and Verrall [1999]). Knowing that $\hat{P} = \sum_{i=1}^{n^t}(1-\delta_i)\hat{M}_i^\star$, we will consider:

- the overall standardized error, defined by $\epsilon = (\hat{P} - P)/P$,
- and the root mean squared deviation $RMSD$, defined by

$$RMSD = \sqrt{MSE} = \sqrt{\left(1/\left(\sum_{i=1}^{n^t}(1-\delta_i)\right)\right)\sum_{i=1}^{n^t}(1-\delta_i)(\hat{M}_i^\star - M_i)^2}.$$

The first quantity indicates the overall quality of the reserving method, whereas the second one gives insights about the accuracy of the prediction of individual reserves. We will also report the bootstrap estimate of the relative standard deviation (RSD), defined by

$$RSD = \left(1/\bar{P}\right)\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{P}_b - \bar{P})^2},$$

where $\bar{P} = (1/B)\sum_{b=1}^{B}\hat{P}_b$, with $\hat{P}_b$ the estimated reserve on the $b$th bootstrap sample.

In the sequel, the number of bootstrap samples is set to $B = 1000$. For each of the above indicators, the superscript (.) will be used to precise which strategy was used to make the predictions.

## 5.1 Income Protection

Short-term disability insurance was designed to protect the policyholders against the loss of some revenue. In our context, the coverage can last up to three years, meaning that the duration of payments $T$ is capped. Here, predicting the final claim amount is similar to predicting the residual lifetime in the disability state. Without loss of generality, say that the insurer has to pay 1€ for each insured day (i.e. $M = T$). We wish to predict the global reserve at various settlement dates, by summing predicted remaining claim lifetimes using strategy (A) (with $M$ replaced by $T$, see Section 3.3 and Remark C.1 in Appendix C). Other strategies presented in Section 3.3 are useless, since $T$ is the response in this case. We compare our results to two famous approaches: i) the Cox model (recommended by regulators for such insurance risk), and ii) the Chain Ladder (CL) model applied on RBNS claims only (since there is no reporting delays here, there is no IBNyR claims).

### 5.1.1 About the database

Our database reports 65 670 claims related to income protection guarantees over six years, from 01/01/2006 to 12/31/2011. For each claim, we know the gender of the policyholder (14 455 males, 51 215 females), her socio-professional category (2 406 managers, 62 799 employees and 465 others), her age when the claim occured, the duration in the disability state, the commercial network (three kinds of brokers: 28 662 "*Net-A*", 4 890 "*Net-B*" and 32 118 "*Net-C*"), and the cause (57 131 sicknesses, and 8 539 accidents) that triggered the coverage. The censoring rate equals 7.2% at the end of the observation period. The mean observed duration in the disability state is 100 days (beyond a deductible of 30 days), with a median of 42 days and a standard deviation of 162 days. If necessary, more details about the data can be found in Lopez et al. [2019], Section 3.

### 5.1.2 Evolution of the predicted reserve

We predict the global reserve $P$ every quarter, from 01/01/2008 to 10/01/2009. The overall results are stored in Table 3, with additional details about the datasets given in the top part of Table 8 in Appendix E.4. As an illustration, Figure 3 shows the evolution of the prediction error for each strategy.

First, notice that $P$ is strongly underestimated in 2008, with both (CL) and (A). On 01/01/2008, the error $\epsilon^{(CL)}$ reaches 60%, whereas it roughly equals 44% using (A). Clearly, the (CL) model does not take into account the censoring phenomenon adequately, which causes large underestimations of the final claim amounts. Although (A) is supposed to appropriately deal with censoring, largest observed lifetimes equal two years at the beginning of 2008. Given that asymptotic properties of our tree estimator are guaranteed once the observations almost entirely map the domain of possible values for $T$ (which is not the case here, since some of the claims will last up to three years in practice), it is not surprising to underestimate the individual reserves $M^\star$ (and thus $P$). On the contrary, the Cox model (Cox [1972]) seems to provide satisfactory results for all the settlement dates. Its semiparametric specification allows us to anticipate longer lifetimes since the beginning, thanks to the baseline hazard component. Still, these good results should be moderated since they originate from favourable circumstances: low censoring rate (see Table 8), no indication that the proportional hazards (PH) assumption may not be reasonable (looking at scaled Schoenfeld residuals), and bounded lifetimes that cover a narrow interval ($[0, 12]$ quarters), ensuring that the response $T$ has low variance.

Second, the errors of (CL) and (A) decrease as time passes. This is in line with expec-

| | 01/01/08 | 04/01/08 | 07/01/08 | 10/01/08 | 01/01/09 | 04/01/09 | 07/01/09 | 10/01/09 |
|---|---|---|---|---|---|---|---|---|
| $P$ | 378 817 | 363 899 | 384 703 | 382 289 | 400 365 | 391 806 | 380 500 | 342 298 |
| $\bar{P}^{(CL)}$ | 151 017 | 166 614 | 193 593 | 207 677 | 243 701 | 242 688 | 254 947 | 259 834 |
| $\epsilon^{(CL)}$ | -60.1% | -54.2% | -50% | -45% | -39% | -38% | -33% | -24% |
| $\bar{P}^{(Cox)}$ | 406 559 | 359 710 | 386 701 | 366 381 | 414 068 | 388 272 | 389 268 | 378 820 |
| $\epsilon^{(Cox)}$ | 7.3% | -1.2% | 0.5% | -4.2% | 3.4% | -0.9% | 2.3% | 9.5% |
| $RSD^{(Cox)}$ | 2.1% | 5.6% | 1.9% | 1.8% | 1.6% | 2.4% | 0.9% | 0.8% |
| $RMSD^{(Cox)}$ | 243 | 234 | 241 | 230 | 224 | 210 | 201 | 187 |
| $\bar{P}^{(A)}$ | 211 357 | 227 088 | 263 030 | 312 400 | 402 398 | 384 361 | 387 525 | 374 133 |
| $\epsilon^{(A)}$ | -44.2% | -42% | -31.6% | -18.3% | **0.5%** | -1.9% | 1.8% | 9.3% |
| $RSD^{(A)}$ | 3.2% | 3% | 3.7% | 2.2% | 2.3% | 2.1% | 0.9% | 1% |
| $RMSD^{(A)}$ | 417 | 430 | 444 | 405 | 383 | 376 | 377 | 371 |

Table 3: Actual reserve $P$, prediction $\bar{P}$, and corresponding predictive uncertainties.

tations, as more and more information become available. One year later (on 01/01/2009), $\epsilon^{(A)}$ decreases steeply to reach 0.5%, whereas it still equals 39% with (CL). The prediction of $P$ thus improves much faster with (A) than with (CL). Strategy (A) benefits from the fact that larger observed lifetimes are now included in the learning set, allowing to significantly improve the quality of our estimator. On the contrary, (CL) still suffers from the proportion of censored lifetimes (almost 11.9% on 01/01/2009, see Table 8). The indicators $RSD^{(A)}$ and $RSD^{(Cox)}$ show that the variance of the prediction of $P$ are similar for both strategies, whereas there is a significant difference between them when looking at the $RMSD$ indicator. In particular, the Cox model seems to provide more accurate individual predictions here.

Third, for most recent settlement dates, (A) and Cox tend to overestimate $P$ whereas (CL) still underestimates $P$ ($\epsilon^{(CL)} = -24\%$). These statements make sense, since the portfolio is observed until 12/31/2011. Getting closer to this date, the percentage of

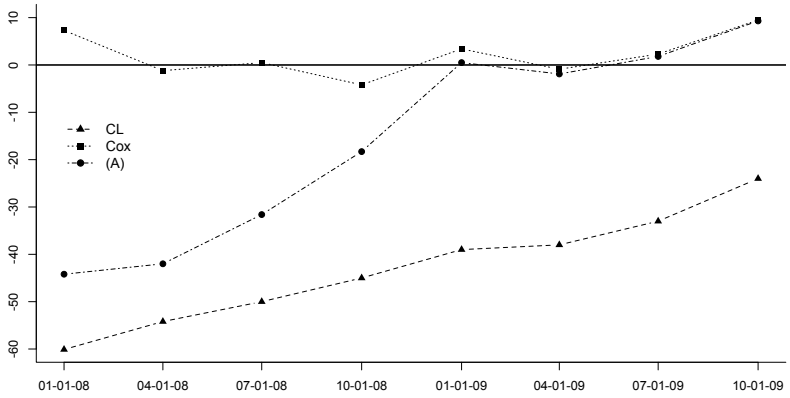


Figure 3: Evolution of the prediction errors (in %) depending on the methodology; for all the settlement dates under consideration.

fully observed claims increases and the actual global reserve decreases. This explains why the (CL) reserve gets closer to the reality. However, such improvements would not be experienced in practice, since $P$ would have no reason to decrease so much (except in case of run-off business). Concerning the results for strategy (A), our backtesting approach implies that considered claims for the prediction of the reserve $P$ are settled on 12/31/2011. The estimations thus face a selection bias, due to an overrepresentation of claims with short developments (whatever $\mathbf{X}$). Since our estimator is based on past information, it anticipates longer developments on average for those claims. Hopefully, this is therefore a non-issue, and the same applies to the Cox model. To make sure about that, we ran the estimations on the whole database (without only selecting closed claims on 12/31/2011), and confirmed that the prediction error remained low and stable.

To sum up, the predictions by strategy (A) and the Cox model give similar results on this example, provided that the main underlying characteristics of the insured risk have been observed (nearly three years of historical information here). In addition, although intermediary results (regression coefficients for Cox, and tree estimators for (A)) were not presented here for conciseness, both models accord with designating the policyholder's age as the most discriminant risk factor to explain $T$. This is good news since this is in line with what risk experts do observe in short-term disability.

## 5.2 Third Party Liability (TPL) insurance

In this application, we aim to estimate individual reserves $M^\star$ based on $T$ and $\mathbf{X}$, where $T$ is censored and where claims are subject to large reporting delays. We compare the methodologies (A) and (B1-5) proposed in Section 3.3 to the Collective Reserving Model (CRM), recently introduced by Wahl et al. [2019]. The CRM enables to split the reserves dedicated to RBNS and IBNyR claims, allowing for a more fair comparison of RBNS reserves than when using Chain Ladder (CL). Moreover, the CRM asymptotically provides similar results than (CL) (excluding tail effects), which is interesting since (CL) remains the benchmark reserving technique in the insurance industry. We consider the `ausautoBI8999` dataset providing claims in motor insurance[1]. A lot of claims have long development times, causing specific claim management processes and atypical loss development triangles (an example of such triangle is provided in Appendix E.2).

[1] Available in the R package `CASdatasets`.

### 5.2.1 Brief description of the database under study

The dataset is made of 22 036 settled automobile bodily injury claims in Australia. These claims arose from accidents occurring from July 1989 to January 1999. The database contains event dates (accident, reporting, closing), operational time (indicator of claim management difficulties), type of injury, number of injured people, potential legal representation of the policyholder, and aggregated settled claim amount. In Appendix E.1, Table 6 summarizes descriptive statistics on these variables, as well as other created by-hand variables useful for our study (e.g. reporting delay, or claim duration). Additionally, Figure 7 shows that reporting started very late as compared to accident dates, which is very likely due to data collection. In particular, there is no settled claims before 01/01/1993. Knowing that the mean claim lifetime equals 558 days and that accident dates start in mid-1989, it looks necessary to omit the data before 01/01/1993 in our study to avoid issues related to data quality. Finally, our database reports 16 822 claims.

### 5.2.2 Prediction of individual reserves for RBNS claims

First, claim amounts have to be inflated to the most recent date (01/01/1999). The annual inflation rate, estimated by the method of Lopez [2019] (Appendix A), equals 0.39%. We wish to give quarterly predictions of the cost of RBNS claims, i.e. $E[M \,|\, T > y, \mathbf{X}]$, between 09/30/1996 and 06/30/1997. We use the strategies (A) and (B1-5) to do so; which obviously lead to different estimators of the individual reserves. More precisely, strategies of type (B) give the same tree estimator $\hat{\pi}(t, \mathbf{x})$, but reserve predictions will differ for one simple reason: rebuilding the information on $T$ is not considered analogously. Anyway, for all the settlement dates studied, looking at the tree $\hat{\pi}$ indicates that the claim lifetime has been detected as the risk factor with the strongest impact on the settled claim amounts. For instance, on 03/31/1997, Figure 4 shows that the most discriminant threshold for claim duration is 838 days, meaning that claims that last more than this threshold before being closed are expected to cause significantly higher final claim amounts (on average 75,000$, as compared to 19,000$ otherwise). Note that this threshold is somewhat stable: it varies by less than 5% depending on the settlement date in practice. Finally, in this example, the population was divided using seven segmentation rules; only based on the claim lifetime, the reporting delay, and the legal representation. Other characteristics such as the number of injured people were not selected, justifying our initial beliefs.

Concerning the reserve predictions, they are in line with our expectations. Figures 5 and 6 give the whole picture of the results. For detailed numerical results and additional

information on the datasets, please refer respectively to Table 7 in Appendix E.3 and bottom part of Table 8 in Appendix E.4. The best prediction method seems to be (B1), since the prediction error remains low and stable as compared to others. The error $\epsilon^{(B1)}$ is about $\pm 7\%$ (except for two settlement dates, but more on this later). Unless this may seem substantial, the explanation is twofold : learning samples are of limited size, and the censoring rate is high (between 30% and 55%, depending on the settlement date). Globally, our reserve estimates are more accurate than insurance practice (CRM model). Indeed, the CRM (closely related to Chain Ladder) systematically underestimates the reserve, which is not surprising in case of claims with long developments. Improvements related to the latest settlement dates are once again fictive, as explained in Section 5.1.2.

The results from other strategies can be further analyzed. Recall that strategy (B4) considers the observed claim lifetime $Y$ as a fully observed input in the modelling, and that the comparison between actual and predicted reserves is made on RBNS claims only (where $T > Y$). Due to the positive correlation between lifetime $T$ and amount $M$ (Kendall's tau equals 0.36), reserves are obviously always underestimated. Due to our backtesting approach, the quality of predictions improves as time passes. This makes sense since the remaining claim durations, not taken into account here, get much lower for most recent settlement dates. In terms of prediction error, strategy (B5) has more or less the same profile. Although the associated estimator uses a prediction of $T$, it does

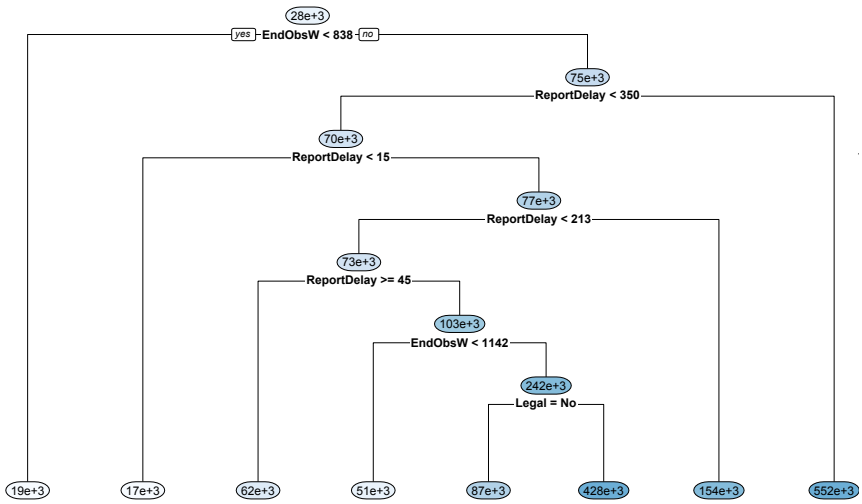


Figure 4: Optimal tree $\hat{\pi}$ following (B)-type strategies, on 03/31/1997. The lifetime ('EndObsW') appears as the most important explanatory variable to predict claim amounts.

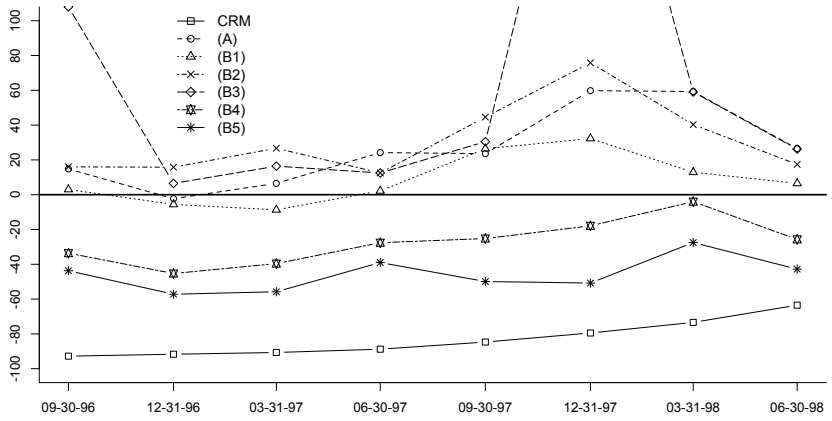


Figure 5: Evolution of the prediction errors for each strategy, for all the settlement dates.

not integrate the information on elapsed time. Given that $\hat{T} = \widehat{E[T \,|\, \mathbf{X}]}$ underestimates $y + E[T - y \,|\, T \geq y, \mathbf{X}]$, these results were clearly expected. This strategy is even worse than (B4) since it systematically makes the same mistake and does not really benefit from newer information coming from latest experience (at most recent settlement dates). Strategy (B3), estimating $T$ (given that $T > Y$) by the semiparametric Cox model before plugging it into the model that predicts $M^\star$, reveals very unstable results. As can be seen in Figure 5, the associated prediction error $\epsilon^{(B3)}$ sometimes explode (see for instance the predicted global reserve on 09/30/96). In fact, a deeper analysis on intermediate results about Cox modelling shows that the crucial underlying PH assumption is strongly violated. In this context, it is very unlikely that predictions on $T$ can be trusted, meaning that our final estimator can not rely on such predictions. Now focusing on strategies (A), (B1) and (B2), the corresponding prediction errors seem to behave the same. In terms of computation power, (B1) is more demanding than (A). However, it is easier to estimate $E[T|T > y, \mathbf{X} = \mathbf{x}]$ than $E[M|T > y, \mathbf{X} = \mathbf{x}]$, since $T$ is a much lower dispersed random variable than $M$. As expected, those methodologies tend to overestimate the global reserve because they suffer from the lack of data related to high values of $y$, which makes the denominator of (3.4) tend to zero. This effect is smoothed when using (B1), thanks to the plug-in step. In addition to being much simpler, strategy (B2) shows good performance, which makes it attractive. However, because it does not integrate the information on $\mathbf{X}$, it should not be recommended if the portfolio composition (in terms of the distribution of $\mathbf{X}$) is subject to significant changes. For latest settlement dates, all strategies taking into account that $T \geq y$ tend to overestimate the reality, because of the selection bias due to backtests (see also the discussion in Section 5.1.2). The bump of the prediction errors for

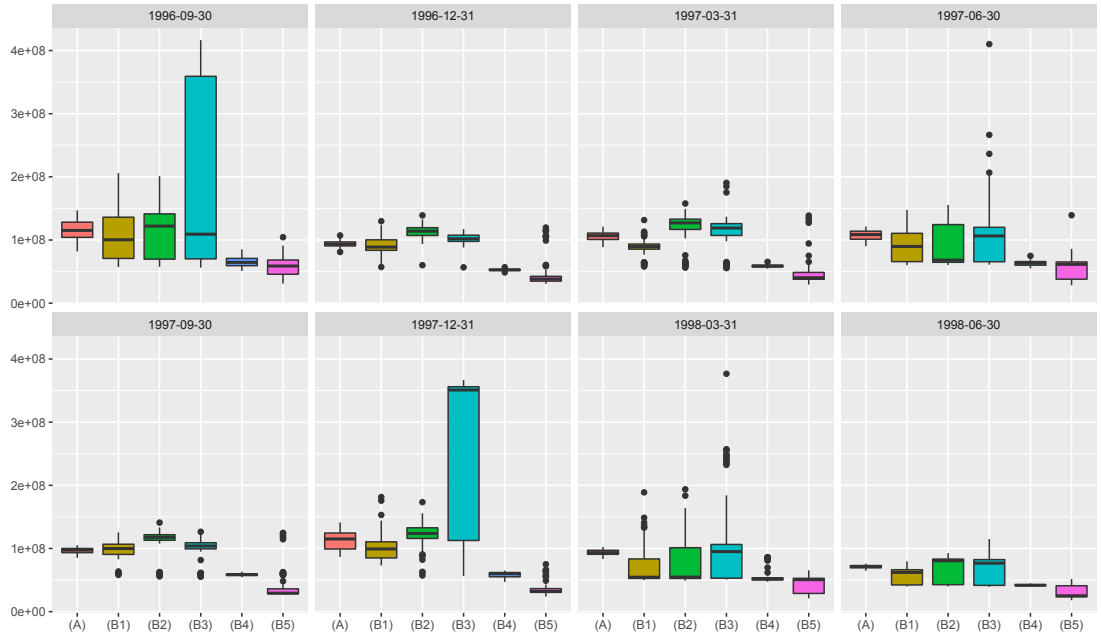


Figure 6: Boxplot of predictions of $P$ at all settlement dates (B=1000 bootstrap samples).

all these strategies on 12/31/97 is caused by significantly different distributions of claim durations in the learning and test sets. To mention an example of such difference, the third quartiles respectively equal 760 and 730 days. This leads to overestimations when predicting $T$ (given that $T \geq y$) on the test set, hence on $\widehat{M}^{\star}$.

Concerning the prediction uncertainty of the strategies considered, Figure 6 highlights that one should have little confidence in predictions coming from (B3). Strategies (A) and (B4) globally seem to provide predicted reserves with little variance, which makes sense since they are one-step prediction methods (without plug-in). Finally, (B1) and (B2) lead to similar prediction uncertainties, higher than when using other strategies (except (B3)).

In a nutshell, the strategy (B1) seems to outperform all other methods in terms of reserve prediction quality, and shows stable results on our data. This estimator is supposed to be asymptotically unbiased, with acceptable variance. We do not pretend it to be the best choice whatever the case, but proved through situations implying different sample sizes and censoring rates that it remains interesting.

# 6 Conclusion

We proposed different methodologies to perform individual claim reserving, based on regression trees. Our contribution is twofold: our modelling enables to take into account the

reporting delays appropriately, and shows the good way to implement plug-in estimators to get reasonable estimations of the reserves. The main features of these approaches are the possibility to use all available information on a claim to predict its final state. In other words, the information on the time since occurrence of the claim is appropriately and fully integrated in the model in our framework. To go further, this work could be improved in several ways. Our applications are mainly a picture of the reserve at some point of time. In particular, no dynamic readjustment of the reserve - due to new information or events that affect the claim - is considered. Nevertheless, our technique may be easily modified to incorporate this, as long as the required information is available. Among other possible improvements, let us mention the possibility to use random forests (i.e. aggregations of regression trees) to stabilize the results, since the CART algorithm is known to be sometimes sensitive to the introduction of new data. The drawback would be a loss of intelligibility of the obtained model. Moreover, some assumptions could be relaxed. In particular, a key assumption in our work is the independence between $\tau$ and the other variables of the model. It is possible to easily relax this assumption by making $\tau$ depend on the covariates $\mathbf{X}$, and then compute a stratified version of the estimator (where the computation of the weights is done separately on groups of observations belonging to the same stratum, as in Galimberti et al. [2002]).

## Acknowledgments

The authors would like to thank the two anonymous referees for their constructive comments that contributed to significantly improve the paper. This work was conducted within the Research Chair DIALog under the aegis of the Risk Foundation, a joint initiative by CNP Assurances and ISFA, Université Claude Bernard Lyon 1 (UCBL).

## References

Maximilien Baudry and Christian Y. Robert. A machine learning approach for individual claims reserving in insurance. *Applied Stochastic Models in Business and Industry*, 35 (5):1127–1155, 2019. doi: 10.1002/asmb.2455.

Stephan M. Bischofberger, Munir Hiabu, and Alex Isakson. Continuous chain-ladder with paid data. *Scandinavian Actuarial Journal*, 0(0):1–26, 2019. doi: 10.1080/03461238.2019.1694973. URL `https://doi.org/10.1080/03461238.2019.1694973`.

Susanna Björkwall, Ola Hössjer, and Esbjörn Ohlsson. Non-parametric and parametric bootstrap techniques for age-to-age development factor methods in stochastic claims reserving. *Scandinavian Actuarial Journal*, 2009(4):306–331, 2009.

W.F Bluhm. Duration-based policy reserves. *Transactions of Society of Actuaries*, 45: 11–53, 1993.

L Breiman, J Friedman, R A Olshen, and C J Stone. *Classification and Regression Trees*. Chapman and Hall, 1984.

D.R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B*, 34(2):187–220, 1972.

F. Duval and M. Pigeon. Individual loss reserving using a gradient boosting-based approach. *Risks*, 7(79):1–19, 2019. doi: 10.3390/risks7030079.

Peter England and Richard Verrall. Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance: Mathematics and Economics*, 25(3):281 – 293, 1999. ISSN 0167-6687. doi: https://doi.org/10.1016/S0167-6687(99)00016-5.

Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011a.

Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011b.

Stefania Galimberti, Peter Sasieni, and Maria Grazia Valsecchi. A weighted kaplan–meier estimator for matched data with application to the comparison of chemotherapy and bone-marrow transplant in leukaemia. *Statistics in Medicine*, 21(24):3847–3864, 2002. doi: 10.1002/sim.1357. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1357`.

Shulamith T Gross and Tze Leung Lai. Nonparametric estimation and regression analysis with left-truncated and right-censored data. *Journal of the American Statistical Association*, 91(435):1166–1180, 1996.

M. Hiabu. On the relationship between classical chain ladder and granular reserving. *Scandinavian Actuarial Journal*, 2017(8):708–729, 2017. doi: 10.1080/03461238.2016.1240709. URL `https://doi.org/10.1080/03461238.2016.1240709`.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

Olivier Lopez. A censored copula model for micro-level claim reserving. *Insurance: Mathematics and Economics*, 87:1 – 14, 2019. ISSN 0167-6687. doi: https://doi.org/10.1016/j.insmatheco.2019.04.001.

Olivier Lopez, Xavier Milhaud, and Pierre-Emmanuel Therond. Tree-based censored regression with applications in insurance. *Electronic Journal of Statistics*, 10:2685–2716, 2016. URL dx.doi.org/10.1214/16-EJS1189.

Olivier Lopez, Xavier Milhaud, and Pierre-E. Thérond. A tree-based algorithm adapted to microlevel reserving and long development claims. *ASTIN Bulletin*, page 1–22, 2019.

A. Maegebier. Valuation and risk assessment of disability insurance using a discrete time trivariate markov renewal reward process. *Insurance: Mathematics and Economics*, 53 (3):802–811, 2013. doi: 10.1016/j.insmatheco.2013.09.013.

María Dolores Martínez Miranda, Jens Perch Nielsen, Stefan Sperlich, and Richard Verrall. Continuous chain ladder: Reformulating and generalizing a classical insurance problem. *Expert Systems with Applications*, 40(14):5588 – 5603, 2013. doi: https://doi.org/10.1016/j.eswa.2013.04.006.

Annette M. Molinaro, Sandrine Dudoit, and Mark J. van der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1):154 – 177, 2004.

Paulo J. R. Pinheiro, João Manuel Andrade e Silva, and Maria de Lourdes Centeno. Bootstrap methodology in claim reserving. *The Journal of Risk and Insurance*, 70(4): 701–714, 2003.

D.G.W. Pitt. Modelling the claim duration of income protection insurance policyholders using parametric mixture models. *Annals of Actuarial Science*, 2(1):1–24, 2007.

C. Sánchez Sellero, W. González Manteiga, and I. Van Keilegom. Uniform representation of product-limit integrals with applications. *Scandinavian Journal of Statistics*, 32(4): 563–581, 2005.

L. Spierdijk and R.H. Koning. Calculating loss reserves in a multistate model for income insurance. Working Paper, 2011.

Wei-Yann Tsai, Nicholas P. Jewell, and Mei-Cheng Wang. A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 74(4):883–886, 1987. ISSN 00063444. URL http://www.jstor.org/stable/2336484.

Felix Wahl, Mathias Lindholm, and Richard Verrall. The collective reserving model. *Insurance: Mathematics and Economics*, 87:34 – 50, 2019.

M.V. Wüthrich. Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, 2018(6):465–480, 2018a. doi: 10.1080/03461238.2018.1428681.

M.V. Wüthrich. Neural networks applied to chain-ladder reserving. *European Actuarial Journal*, 8(2):407–436, 2018b. doi: 10.1007/s13385-018-0184-4.

# A Removing inflation: estimation procedure

Let $M_i^{'}$ denote the observed claim amount (before removing inflation). In general, the data reports $M_i'$ instead of $M_i$. We assume that

$$\log M_i' = \beta d_i + \log M_i, \tag{A.1}$$

where $d_i$ is the fiscal year at which the $i$th claim is observed, and $\beta$ is an inflation factor that is going to be estimated using our data. We assume that $(M_i)_{1\leq i\leq n}$ are i.i.d. and independent of $(d_i)_{1\leq i\leq n}$. The dates $d_i$ take their value in $\{0, ..., D\}$ (where $D+1$ periods, e.g. years, are observed). Then, proceed as follows:

- Compute $m_{i,j}^{'}$, defined as the average of the fully observed claims that occurred on the fiscal year $d_i$, and which are settled after $j$ years (for $(i, j)$ such that $d_i + j \leq t_1$, where $t_1$ is the last observed date). Let $n_{i,j}$ denote the number of such claims.
- Under (A.1), we know that $\log m_{i,j}^{'} \approx \beta_j d_i + \alpha_j$, where $\alpha_j = E[\log M_i | T_i = j]$. For each $j$, we compute $\hat{\beta}_j$ the weighted least-square estimator of the slope $\beta$ based on the points $(m_{i,j}^{'}, d_i)_{i:d_i+j\leq t_1}$. More precisely, one solves

$$(\hat{\alpha}_j, \hat{\beta}_j) = \arg\min_{\alpha_j,\beta_j} \sum_{i:d_i+j\leq t_1} n_{i,j}(\log m_{i,j}^{'} - \alpha_j - \beta_j d_i)^2.$$

- Finally, our estimator of $\beta$ is given by $\hat{\beta} = \frac{\sum_j n_j^{1/2}\hat{\beta}_j}{\sum_j n_j^{1/2}}$, with $n_j = \sum_i n_{i,j}$.

For each claim $i$ such that $\delta_i = 1$ (settled claims), we thus consider $\hat{M}_i = M_i' e^{-\hat{\beta} d_i}$ as an estimator of the final claim amount $M_i$, once removed the inflation effect. In practice, one will use this amount in applications.

# B Estimator of $S(t)$ in our censoring-truncation model

Let us recall that, for a discrete variable $A$ taking value at point $\{a_1, ..., a_k\}$, its survival function $S_A(t) = \mathbb{P}(A \geq a)$ can be written as

$$S_A(t) = \prod_{j=1}^{k} (1 - \lambda_A(a_j)), \tag{B.1}$$

with

$$\lambda_A(t) = -\frac{dS_A(t)}{S_A(t)}.$$

A way to determine an estimator of $S_A$ hence reduces to replace $\lambda_A$ in (B.1) by a consistent estimator obtained from the data. If the variable $A$ is not discrete, it can still be approximated by a discrete distribution where the $(a_j)_{1\leq j\leq k}$ are replaced by the value of the complete observations (in our case, the uncensored observations). This is the basis of the construction of Kaplan-Meier and other product-limit based estimator, see for example Fleming and Harrington [2011b].

Hence, our aim is to determine a consistent estimator of $\lambda_T(t) = -\dfrac{dS_T(t)}{S_T(t)}$.
Let $L(t) = \mathbb{P}(\tau \leq t)$, $S_C(t) = \mathbb{P}(C \geq t)$, $\alpha = P(\tau < C)$, and

$$S_1(t) = E\left[\delta \mathbf{1}_{\tau\leq Y}\mathbf{1}_{Y\leq t}|\tau < C\right].$$

By basic computations, we obtain $dS_1(t) = -\alpha^{-1}S_C(t)L(t)dS_T(t)$.
Indeed,

$$\begin{aligned}
S_1(t) &= \alpha^{-1}E\left[\mathbf{1}_{\tau\leq t}\mathbf{1}_{T\leq T}E[\mathbf{1}_{\max(\tau,T)<C}|\tau,T]\right] \\
&= \alpha^{-1}E\left[\mathbf{1}_{\tau\leq T}\mathbf{1}_{T\leq t}S_C(\max(\tau,T))\right] \\
&= \alpha^{-1}E\left[\mathbf{1}_{\tau\leq T}\mathbf{1}_{T\leq t}S_C(T)\right] \\
&= -\alpha^{-1}\int_0^t S_C(y)L(y)dS_T(y),
\end{aligned}$$

where we used that $(T,\tau)$ is independent of $C$ for the second line, and the independence between $T$ and $C$ for the last line. On the other hand, let

$$S_2(t) = E\left[\mathbf{1}_{\tau<t<Y}|\tau < C\right].$$

We have $S_2(t) = \alpha^{-1}S_T(t)S_C(t)L(t)$. Since

$$\begin{aligned}
S_2(t) &= \alpha^{-1}E\left[\mathbf{1}_{\tau<t}\mathbf{1}_{t<T}\mathbf{1}_{\max(t,\tau)<C}\right] \\
&= \alpha^{-1}E\left[\mathbf{1}_{\tau<t}E\left[\mathbf{1}_{t<T}\mathbf{1}_{\max(t,\tau)<C}|\tau\right]\right] \\
&= \alpha^{-1}E\left[\mathbf{1}_{\tau<t}S_T(t)S_C(\max(t,\tau))\right] \\
&= \alpha^{-1}E\left[\mathbf{1}_{\tau<t}S_T(t)S_C(t)\right],
\end{aligned}$$

hence

$$-\frac{dS_T(t)}{S_T(t)} = \frac{dS_1(t)}{S_2(t)}.$$

The quantities $S_1$ and $S_2$ can be estimated consistently by

$$\hat{S}_1(t) = \frac{1}{n}\sum_{i=1}^n \delta_i \mathbf{1}_{\tau_i<Y_i}\mathbf{1}_{Y_i\leq t} \qquad \text{and} \qquad \hat{S}_2(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\tau_i<t\leq Y_i}.$$

This leads to the following estimator of the survival function,

$$\hat{S}_T(t) = 1 - \hat{F}_T(t) = \prod_{Y_i \leq t} \left(1 - \frac{d\hat{S}_1(Y_i)}{\hat{S}_2(Y_i)}\right),$$

when there is no ties.

In case of ties, let $(t_1, ..., t_k)$ denote the distinct values taken by $(Y_i)_{1 \leq i \leq n}$, leading to

$$\hat{S}_T(t) = \prod_{t_i \leq t} \left(1 - \frac{d\hat{S}_1(t_i)}{\hat{S}_2(t_i)}\right).$$

# C Modification of the CART algorithm

We present here the weighted CART algorithm (wCART), used throughout the paper to take into account censoring and truncation phenomenons.

---

**Step 1:** $R_1(\mathbf{z}) = 1$ for all $\mathbf{z} = (y, \mathbf{x})$, and $n_1 = 1$ (corresponds to the root node).

**Step k+1:** Let $(R_1, ...R_{n_k})$ denote the rules obtained at step $k$. For $j = 1, ..., n_k$,

- if all observations such that $\delta_i R_j(Y_i, \mathbf{X}_i) = 1$ have the same characteristics, then keep rule $j$ as it is no longer possible to segment the population;

- else, rule $R_j$ is replaced by two rules $R_{j1}$ and $R_{j2}$ determined in the following way: for each component $Z^{(l)}$ of $\mathbf{Z} = (Y, \mathbf{X})$ $(l = 1, ..., d+1)$, define the best threshold $z_\star^{(l)}$ to split the data, such that $z_\star^{(l)} = \arg\min_{z^{(l)}} s(R_j, z^{(l)})$, with

$$\begin{aligned} s(R_j, z^{(l)}) &= \sum_{i=1}^{n} w_{i,n}(\phi(N_i) - \bar{n}_{l-}(z^{(l)}, R_j))^2 \mathbf{1}_{Z_i^{(l)} \leq z^{(l)}} R_j(\mathbf{Z}_i) \\ &+ \sum_{i=1}^{n} w_{i,n}(\phi(N_i) - \bar{n}_{l+}(z^{(l)}, R_j))^2 \mathbf{1}_{Z_i^{(l)} > z^{(l)}} R_j(\mathbf{Z}_i), \end{aligned}$$

  where

$$\bar{n}_{l-}(z, R_j) = \frac{\sum_{i=1}^{n} w_{i,n}\phi(N_i)\mathbf{1}_{Z_i^{(l)} \leq z} R_j(\mathbf{Z}_i)}{\sum_{k=1}^{n} w_{k,n}\mathbf{1}_{Z_k^{(l)} \leq z} R_j(\mathbf{Z}_k)}, \quad \bar{n}_{l+}(z, R_j) = \frac{\sum_{i=1}^{n} w_{i,n}\phi(N_i)\mathbf{1}_{Z_i^{(l)} > z} R_j(\mathbf{Z}_i)}{\sum_{k=1}^{n} w_{k,n}\mathbf{1}_{Z_k^{(l)} > z} R_j(\mathbf{Z}_k)}.$$

  Then, select the best component to consider, that is $\hat{l} = \arg\min_l s(R_j, z_\star^{(l)})$. Define the two new rules $R_{j1}(\mathbf{z}) = R_j(\mathbf{z})\mathbf{1}_{z^{(\hat{l})} \leq z_\star^{(\hat{l})}}$, and $R_{j2}(\mathbf{z}) = R_j(\mathbf{z})\mathbf{1}_{z^{(\hat{l})} > z_\star^{(\hat{l})}}$.

- Let $n_{k+1}$ denote the new number of rules.

**Stopping rule:** stop if $n_{k+1} = n_k$.

---

**Remark C.1.** *Sometimes, $M$ may be a deterministic function of $T$. The algorithm is thus easily adapted by replacing $N$ by $Y$, and $\mathbf{Z} = (Y, \mathbf{X})$ by $\mathbf{X}$. In this simpler situation (where only one single censored/truncated variable has to be predicted), competing approaches include survival trees and forests, see Ishwaran et al. [2008] and Molinaro et al. [2004].*

# D Simulations

We give here the characteristics of the simulated samples used in the simulation study.

| Sample | Group-specific exposure | | | | Sample |
|---|---|---|---|---|---|
| size $n$ | Group 1 | Group 2 | Group 3 | Group 4 | mean |
| 100 | 37% | 27% | 16% | 20% | 11.13 |
| 1000 | 26.4% | 31.7% | 20.1% | 21.8% | 11.36 |
| 5 000 | 31.41% | 29.95% | 19.49% | 19.15% | 11.55 |

Table 4: Descriptive statistics of simulated datasets.

# E Applications

## E.1 Descriptive statistics for the motor insurance dataset

Depending on the type of the variable, we give different indicators: for categorical variables, exposure for each category is provided. Concerning numerical variables, the minimum, the maximum, the median, the mean, and the standard deviation are given.

| Variable: | Type | Min. | Median | Mean | Std. | Max. |
|---|---|---|---|---|---|---|
| AccDate | date | 07/01/1989 | 10/01/1994 | 08/02/1994 | | 01/01/1999 |
| ReportDate | date | 09/01/1990 | 03/01/1995 | 05/12/1995 | | 02/01/1999 |
| FinDate | date | 07/01/1993 | 01/01/1997 | 10/11/1996 | | 03/01/1999 |
| Reporting delay | numerical | 0 | 59 | 113 | 173 | 1 430 |
| Claim duration | numerical | 0 | 486 | 558 | 381 | 2 069 |
| Operational time | numerical | 0.1 | 45.9 | 46.33 | 27.1 | 99.1 |
| InjNb | numerical | 1 | 2 | 2.13 | 1.37 | 5 |
| AggClaim | numerical | 10 | 13 854 | 38 367 | 90 981 | 4 485 797 |
| Legal | boolean | No: 8 008 | Yes: 14 028 | | | |
| InjType1 | categorical | | Fatal: | High: | Medium: | Minor: | Severe: | Small: | Not recorded: |
| | frequency | | 256 | 189 | 1 133 | 15 638 | 188 | 3 376 | 1 256 |

Table 5: Descriptive statistics on available information for `ausautoBI8999` dataset.

We also show that reporting started much later than accidents, which led us to remove oldest observations since it is not clear whether these information are reliable.

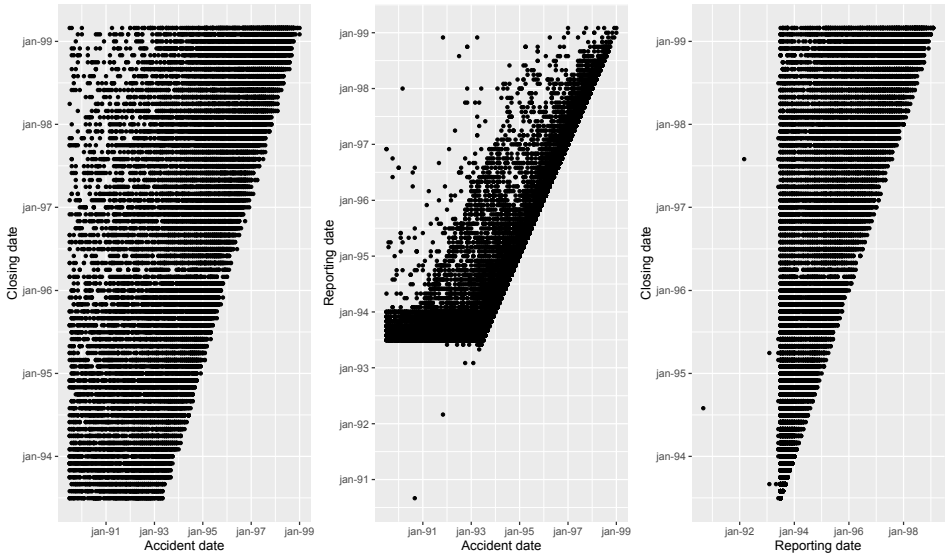


Figure 7: Dates of events in `ausautoBI8999` dataset. Reporting only started mid-1993.

## E.2 Example of loss triangle in the TPL insurance application

We give the (non cumulated) payments for RBNS claims on 12/31/1996 (settlement date), leading to a CL reserve of 26 165 560\$. Numbers are given in thousands and rounded to make the reading easier.

| | d1 | d2 | d3 | d4 | d5 | d6 | d7 | d8 | d9 | d10 | d11 | d12 | d13 | d14 | d15 | d16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 01/01/93 | 0.1 | 27 | 57 | 237 | 343 | 199 | 184 | 524 | 563 | 381 | 428 | 250 | 449 | 651 | 606 | 19 |
| 04/01/93 | 0 | 36 | 171 | 162 | 909 | 332 | 324 | 454 | 208 | 389 | 192 | 862 | 344 | 275 | 119 | |
| 07/01/93 | 5 | 61 | 228 | 395 | 210 | 370 | 353 | 501 | 805 | 188 | 330 | 235 | 849 | 358 | | |
| 10/01/93 | 2 | 49 | 139 | 169 | 98 | 571 | 327 | 647 | 655 | 687 | 829 | 342 | 88 | | | |
| 01/01/94 | 5 | 112 | 101 | 210 | 209 | 299 | 223 | 729 | 264 | 855 | 370 | 120 | | | | |
| 04/01/94 | 2 | 15 | 236 | 372 | 240 | 371 | 792 | 136 | 652 | 604 | 176 | | | | | |
| 07/01/94 | 8 | 152 | 688 | 350 | 437 | 364 | 580 | 327 | 507 | 228 | | | | | | |
| 10/01/94 | 19 | 125 | 304 | 347 | 157 | 439 | 452 | 300 | 143 | | | | | | | |
| 01/01/95 | 61 | 102 | 254 | 166 | 154 | 414 | 492 | 686 | | | | | | | | |
| 04/01/95 | 11 | 82 | 129 | 225 | 310 | 477 | 148 | | | | | | | | | |
| 07/01/95 | 15 | 79 | 67 | 245 | 361 | 719 | | | | | | | | | | |
| 10/01/95 | 0.9 | 43 | 234 | 228 | 139 | | | | | | | | | | | |
| 01/01/96 | 0 | 112 | 263 | 42 | | | | | | | | | | | | |
| 04/01/96 | 5 | 91 | 53 | | | | | | | | | | | | | |
| 07/01/96 | 7 | 16 | | | | | | | | | | | | | | |
| 10/01/96 | 0 | | | | | | | | | | | | | | | |

Table 6: Example of loss triangle.

## E.3 Detailed results on the calibration of reserves

We provide here the detailed results for the application on TPL guarantees. When using the CRM model, we only report the part of the reserve dedicated to RBNS claims.

| | 09/30/96 | 12/31/96 | 03/31/97 | 06/30/97 | 09/30/97 | 12/31/97 | 03/31/98 | 06/30/98 |
|---|---|---|---|---|---|---|---|---|
| $P$ | 100 872 881 | 96 545 407 | 97 351 400 | 87 497 571 | 78 389 452 | 71 489 554 | 58 207 140 | 56 493 734 |
| $P^{CRM}$ | 7 212 119 | 8 014 483 | 9 015 614 | 9 758 563 | 11 935 499 | 14 650 041 | 15 397 966 | 20 617 125 |
| $\epsilon^{(CRM)}$ | -92.8% | -91.7% | -90.7% | -88.8% | -84.7% | -79.5% | -73.4% | -63.5% |
| $P^{(A)}$ | 115 776 538 | 94 284 076 | 103 701 446 | 108 631 498 | 96 881 249 | 114 245 317 | 92 722 102 | 71 456 983 |
| $\epsilon^{(A)}$ | +14.8% | -2.3% | +6.5% | +24.2% | +23.6% | +59.8% | +59.3% | +26.5% |
| $RSD^{(A)}$ | 14.5% | 5.3% | 6.7% | 6.1% | 5.1% | 11.8% | 4.1% | 3.8% |
| $RMSD^{(A)}$ | 131 477 | 118 117 | 103 363 | 74 249 | 61 914 | 82 596 | 67 906 | 149 120 |
| $P^{(B1)}$ | 103 890 209 | 91 167 753 | 88 837 467 | 89 498 892 | 99 104 425 | 94 607 399 | 65 703 004 | 60 192 296 |
| $\epsilon^{(B1)}$ | +3% | -5.6% | -8.7% | +2.2% | +26.4% | +32.3% | +12.9% | +6.5% |
| $RSD^{(B1)}$ | 43.5% | 11.7% | 16.7% | 29% | 16.5% | 19.1% | 26.5% | 18.9% |
| $RMSD^{(B1)}$ | 142 831 | 119 204 | 106 326 | 83 161 | 65 001 | 95 649 | 76 673 | 155 189 |
| $P^{(B2)}$ | 117 001 710 | 111 788 049 | 123 314 660 | 98 413 820 | 113 376 802 | 125 713 715 | 81 641 459 | 66 327 464 |
| $\epsilon^{(B2)}$ | +16% | +15.8% | +26.7% | +12.5% | +44.6% | +75.8% | +40.3% | +17.4% |
| $RSD^{(B2)}$ | 33% | 8.1% | 17.7% | 32.5% | 14.3% | 15.4% | 40.2% | 31.9% |
| $RMSD^{(B2)}$ | 140 472 | 119 905 | 107 727 | 84 696 | 67 241 | 102 941 | 95 601 | 158 678 |
| $P^{(B3)}$ | 210 330 566 | 102 857 540 | 113 318 807 | 98 456 074 | 102 297 911 | 268 349 985 | 92 537 067 | 71 268 961 |
| $\epsilon^{(B3)}$ | +108% | +6.5% | +16.4% | +12.5% | +30.5% | +275% | +59% | +26.2% |
| $RSD^{(B3)}$ | 70% | 6.1% | 20.3% | 39.4% | 12.6% | 42.6% | 58.7% | 28.7% |
| $RMSD^{(B3)}$ | 176 961 | 119 356 | 107 170 | 84 024 | 67 402 | 200 421 | 108 349 | 159 456 |
| $P^{(B4)}$ | 66 954 771 | 52 828 868 | 58 790 353 | 63 305 968 | 58 620 699 | 58 700 374 | 55 884 349 | 41 978 823 |
| $\epsilon^{(B4)}$ | -33.6% | -45.3% | -39.6% | -27.6% | -25.2% | -17.9% | -4% | -25.7% |
| $RSD^{(B4)}$ | 9.5% | 3.2% | 4.6% | 5.1% | 2.8% | 6.3% | 20.5% | 2.5% |
| $RMSD^{(B4)}$ | 132 730 | 120 251 | 106 166 | 78 661 | 62 152 | 81 923 | 93 140 | 154 126 |
| $P^{(B5)}$ | 56 792 306 | 41 315 697 | 43 034 713 | 53 340 431 | 39 257 861 | 35 175 163 | 42 180 221 | 32 297 800 |
| $\epsilon^{(B5)}$ | -43.7% | -57.2% | -55.8% | -39% | -49.9% | -50.8% | -27.5% | -42.8% |
| $RSD^{(B5)}$ | 27% | 42.6% | 21.3% | 39.2% | 62% | 28.5% | 29.5% | 28.8% |
| $RMSD^{(B5)}$ | 132 114 | 122 838 | 110 157 | 81 709 | 68 191 | 81 638 | 69 368 | 157 515 |

Table 7: Quarterly assessment of the global reserve at various settlement dates.

## E.4 Details about subsamples at each settlement date

| **Income protection (Section 5.1)** | 01/01/2008 | 04/01/2008 | 07/01/2008 | 10/01/2008 | 01/01/2009 | 04/01/2009 | 07/01/2009 | 10/01/2009 |
|---|---|---|---|---|---|---|---|---|
| (1) Size of the learning set | 20 542 | 23 370 | 26 214 | 28 740 | 31 962 | 34 796 | 37 700 | 40 344 |
| (2) Size of the validation set | 10 271 | 11 686 | 13 107 | 14 371 | 15 982 | 17 399 | 18 850 | 20 172 |
| (3) Censoring rate in learning set | 16.11% | 13.94% | 12.9% | 11.37% | 11.97% | 10.36% | 9.55% | 8.65% |
| (4) Censoring rate in validation set | 16.24% | 13.66% | 12.8% | 11.4% | 11.89% | 10.32% | 9.27% | 8.25% |
| $\Rightarrow$ Number of backtested claims : (4) $\times$ (2) | 1 688 | 1 596 | 1 677 | 1 638 | 1 900 | 1 795 | 1 747 | 1 664 |
| (5) Total paid amount at settlement date | 818 079 | 955 809 | 1 115 449 | 1 259 591 | 1 448 942 | 1 608 799 | 1 771 356 | 1 955 760 |
| (6) Paid at settlement date (censored claims) | 278 230 | 286 354 | 323 982 | 336 883 | 378 083 | 388 346 | 387 616 | 399 445 |
| (7) Final paid amount (censored claims) | 657 047 | 650 253 | 708 685 | 719 172 | 778 448 | 780 152 | 768 116 | 741 743 |
| (8) Exact (backtested) global reserve $P$ ((7)-(6)) | 378 817 | 363 899 | 384 703 | 382 289 | 400 365 | 391 806 | 380 500 | 342 298 |

| **TPL insurance (Section 5.2)** | 09/30/1996 | 12/31/1996 | 03/31/1997 | 06/30/1997 | 09/30/1997 | 12/31/1997 | 03/31/1998 | 06/30/1998 |
|---|---|---|---|---|---|---|---|---|
| (1) Size of the learning set | 7 960 | 8 513 | 9 022 | 9 515 | 9 995 | 10 436 | 10 782 | 11 033 |
| (2) Size of the validation set | 3 981 | 4 257 | 4 511 | 4 758 | 4 998 | 5 219 | 5 392 | 5 517 |
| (3) Censoring rate in learning set | 53.92% | 49.76% | 46.3% | 43.06% | 38.1% | 33.68% | 28.64% | 22.73% |
| (4) Censoring rate in validation set | 54.43% | 50.69% | 47.21% | 42.45% | 37.99% | 33.01% | 28.52% | 22.77% |
| $\Rightarrow$ Number of backtested claims : (4) $\times$ (2) | 2 167 | 2 116 | 2 099 | 2 046 | 1 995 | 1 785 | 1 595 | 1 230 |
| (5) Total paid amount at settlement date | 32 952 642 | 42 116 625 | 49 585 991 | 57 617 269 | 68 198 761 | 84 406 988 | 91 766 991 | 106 590 794 |
| (6) Paid at settlement date (censored claims) | *unknown* | *unknown* | *unknown* | *unknown* | *unknown* | *unknown* | *unknown* | *unknown* |
| (7) Exact (backtested) global reserve $P$ : | 100 872 881 | 96 545 407 | 97 351 400 | 87 497 571 | 78 389 452 | 71 489 554 | 58 207 140 | 56 493 734 |

Table 8: Actual global reserve at various settlement dates, and other statistics about the data under consideration.