A tree-based algorithm adapted to microlevel reserving and long development claims

Olivier Lopez, Xavier Milhaud, Pierre-Emmanuel Thérond

A tree-based algorithm adapted to microlevel reserving and long development claims

Olivier Lopez[1], Xavier Milhaud[2], Pierre-E. Thérond[2,3]

[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

[3]Galea & Associés, 25 rue de Choiseul, 75002, Paris, France

January 9, 2019

Abstract

In non-life insurance, business sustainability requires accurate and robust predictions of reserves related to unpaid claims. To this aim, two different approaches have historically been developed: aggregated loss triangles and individual claim reserving. The former has reached operational great success in the past decades, whereas the use of the latter still remains very limited. Through two illustrative examples and introducing a new tree-based algorithm, we show, not surprisingly, that individual claim reserving is really promising in the context of long-term risks.

Keywords : reserving, long-tail, censoring, regression tree, disability.

1 Introduction

Given their greater complexity, is it worth using individual claims reserving techniques in non-life insurance? In this paper, we show that adapting the very famous CART algorithm to censored data is not a big deal, and enables to introduce a new tree-based algorithm that shows good performance for individual claim reserving purposes.

Despite some recent advances[1], insurance companies still seem to be reluctant to use micro-level reserving as compared to very standard techniques using aggregated data, like

[1]See the report on non-life reserving practices by ASTIN Working Party (June 2016) at `http://www.actuaries.org/ASTIN/Documents/ASTIN_WP_NL_Reserving_Report1.0_2016-06-15.pdf`

Chain Ladder and its extensions (Mack [1993], Bornhuetter and Ferguson [1972], Quarg and Mack [2008]). In such traditional methods, individual claims are summed and stored into claim development triangles according to a two-dimensional scheme based on origin and development periods. Of course, the success of these models lies in that they are easily understandable, simple to use, and have worked very well in many circumstances in the past. However, practitioners are clearly aware of their limitations[2] and know that they can lead to poor estimates, especially concerning the reserves for the latest development periods. This mainly originates from the fact that these methods *do not capture the pattern of claim development*, which is of primary importance in some cases.

Simultaneously, spectacular improvements to collect historical information and individual characteristics on claims have been made in the insurance industry for more than fifteen years, and companies have now access to very comprehensive datasets. Using these data and regression models, actuaries can use sophisticated statistical procedures to estimate Incurred But Not yet Reported (IBNyR) and Reported But Not Settled (RBNS) claims. RBNS claims correspond to situations where the insurer knows about the existence of the claim, has possibly started to pay for it, but does not know how much the final charge will be. In such a context, taking into account individual features about claims offers many advantages to approximate the reserve. First, it enables to cope with heterogeneity issues that can arise when using aggregated data. Indeed, storing all claims into aggregate run-off triangles makes it impossible to consider changes related to claims management, reinsurance programs, legal context and product mix. It also prevents from integrating key claim characteristics and thus crucial risk factors explaining the final amount to pay. Second, it allows to separate RBNS and IBNyR claims to perform an advanced risk assessment and monitoring. Moreover, the specific development pattern of claims can be considered, which means that the full information about the history of the claim (occurrence, reporting, payments, and closure) are now inputs of the model. And last but not least, these techniques provide individual claims reserves which could be very usefull from both a risk management and a claims management perspective (for instance in order to improve claims management policies).

One could then wonder why such techniques have not been widely applied yet. Except that it is harder to implement, the reason seems quite obvious: past contributions on

[2]Several well-known issues concern propagation of errors through the development factors, instability in ultimate claims for recent arrival periods, necessary previous treatment of outliers, need to integrate tail factors (see for instance Halliwell [2007]). Assumptions underlying such models are also often discussed, as well as corresponding statistical tests (see Harnau [2017]).

individual claim reserving were mainly focused on parametric models and likelihood maximisation (Antonio and Plat [2014], Pigeon et al. [2013], Zhao et al. [2009], Larsen [2007], Haastrup and Arjas [1993]). Due to RBNS claims, deriving the likelihood associated with observed claims is not straightforward, because of truncation and censoring phenomenons. Besides, the parametric relationship existing between claim amounts and risk factors under study can be tricky to specify. As a result, these approaches did not reveal neither convincing nor very effective in practice. Moreover, according to most of actuaries and under regulatory constraints (stating that ultimate reserve estimates should be regularly updated, say each quarter), parametric individual claim reserving models have not really been considered useful so far for one simple reason: quarterly gains/losses indicators (the so-called *boni-mali*) were not improved, which means that the overall quality of prediction of such models was not better than the Chain Ladder's one (at least on the short-term, showing that Chain Ladder remains somewhat effective in most of situations). Since the main threat for the top management concerns potential urgent need for capital injections, this statement diminishes the attractiveness of such techniques. Besides, *mali* can have impacts on the Solvency Capital Requirement, as well as on future premiums. To the best of our knowledge, this paper proposes a new way to anticipate, as soon as possible, the ultimate global reserve by aggregating individual reserve predictions for RBNS claims. We do not claim that our model is better than others, but simply show to which extent individual claim reserving by nonparametric approaches could be beneficial to approximate future payments. Although our application focuses here on claim reserving, it is also important to be aware that many other actuarial applications could use the technique presented in the sequel. Let us mention for instance the opportunity to decrease costs related to experts involved in claim estimations, as well as improving the targeting of specific claims causing atypical claim amounts.

The paper is organized as follows: Section 2 introduces our new method to estimate individual reserves, with tools similar to Wüthrich [2018]. However, ultimate individual reserves for RBNS claims are here estimated thanks to an adaptation of the CART algorithm to censored data. Then, two applications are conducted in Section 3 to answer the initial question. Results are compared to the Chain Ladder method, knowing that its usual stochastic extensions (Mack [1993], England and Verrall [2002]) all provide the same expected ultimate global reserve (the only difference lies in assessing its variance).

2 Proposed individual claim reserving technique

Up to now, very few references exist on individual claims reserving with nonparametric techniques (Wüthrich [2018], Baudry and Robert [2017]). In the case where the insurer can access individual information about the claims, our approach consists in using an extension of the CART algorithm to incomplete observations (Lopez et al. [2016]). This piecewise tree-based estimator allows for nonlinearities in the dependence structure between claim amounts and explanatory risk factors (Olbricht [2012]). We wish to estimate the ultimate amounts of RBNS claims for individual policies, and then deduce individual predictions of reserves.

2.1 A weighting procedure for duration analysis

The time development of a claim is crucial to predict its severity. Roughly speaking, a claim which requires a lot of time to be settled is more likely to be associated with a large amount. Therefore, if M denotes the claim amount, one must provide a model that takes the impact on this variable of the time before settlement.

We are thus interested in a random vector $(M, T, \mathbf{X})$, where $\mathbf{X} = (X^{(1)}, ..., X^{(d)}) \in \mathcal{X} \subset \mathbb{R}^d$ denotes a set of random covariates that may have an impact on T and/or M, and $(M, T) \in \mathbb{R}^{+2}$. In the following, T represents the time before a claim is fully settled, and M the total corresponding amount (only known at the end of the claim settlement process). As we are dealing with a duration T, this variable is subject to censoring, which is a classical issue in survival analysis. This means that, in the database that we use to calibrate the distribution of $(M, T, \mathbf{X})$ (and hence to predict M), all of the claims are not fully settled. To describe this phenomenon, let us introduce a censoring variable $C \in \mathbb{R}^+$, which represents the time between the opening of the claim and the end of observation for any other cause than its settlement. For example, retrocession of a claim leads to a loss of information after some point of time. The observed variables are thus not directly T and M, but $Y = \inf(T, C)$, $\delta = \mathbf{1}_{T \leq C}$, and $N = \delta M$. The covariates $\mathbf{X}$ are considered as always fully observed. The data is made up of i.i.d. replications $(N_i, Y_i, \delta_i, \mathbf{X}_i)_{1 \leq i \leq n}$. We also assume that C is independent of (M, T, X). This assumption implies that the amounts M should be free from inflation, see the discussion in Lopez [2018].

It is important to notice that one should not calibrate a model for M only on the closed claims, that is with $\delta = 1$. Although the closed claims bring a complete information on the variable, this information is biased: indeed, among closed claims, there is an excess

of claims with small time of settlement. Since these claims are more likely to be of small amount, this would lead to an underestimation of the typical values taken by M. The alternative is to correct the bias caused by censoring using an appropriate weighting scheme. For a comprehensive description of the algorithm used hereafter and related properties, the reader is referred to Lopez et al. [2016]. This algorithm is inspired from the well-known CART algorithm, where the problem of incomplete observations forces to introduce the Kaplan-Meier (KM) weights. Those weights are defined by

$$\omega_i = \frac{\delta_i}{n(1-\hat{G}(Y_i-))},$$

with $\hat{G}$ the Kaplan-Meier estimator for the cdf of the censoring variable C, denoted by $G(t) = \mathbb{P}(C \leq t)$. The introduction of such weights is motivated by the fact that, for all function $(m,t,\mathbf{x}) \to \phi(m,t,\mathbf{x})$ (with $\phi(m,t,\mathbf{x}) = 0$ for t s.t. $G(t) = 1$) with finite expectation,

$$E\left[\frac{\delta\phi(N,Y,\mathbf{X})}{(1-G(Y-))} \,|\, \mathbf{X}\right] = E[\phi(M,T,\mathbf{X}) \,|\, \mathbf{X}],$$

under the assumption that $(M,T,\mathbf{X})$ is independent from C. Hence, the weights ω_i can be seen as an approximation of some "ideal" weights $\omega_i^* = \delta_i n^{-1}[1-G(Y_i-)]^{-1}$, since G is usually unknown, and has therefore to be estimated. These weights are hence a convenient way to correct the bias caused by the censoring, since each quantity of the type $E[\phi(M,T,\mathbf{X})]$ will be consistently estimated by the weighted mean $\sum_{i=1}^n \omega_i^* \phi(N_i,Y_i,\mathbf{X}_i)$. Concretely, these KM weights equal 0 when the observation is censored ; otherwise, the greater the fully observed lifetime the higher the weight. This enables to compensate for the fact that very few individuals with high durations are fully observed.

2.2 Weighted regression tree algorithm

Regression trees are a convenient way to estimate a regression function without relying on a linear assumption. Suppose that one wants to estimate a function $\mu(\mathbf{x}) = E[\phi(M,T) \,|\, \mathbf{X} = \mathbf{x}]$. We use the following modified CART algorithm introducing the previous weights that are computed once for all before launching the algorithm. At each step of the algorithm, "rules" $\mathbf{x} = (x^{(1)},...,x^{(d)}) \to R_j(\mathbf{x})$ to split the data, that is, for each possible value of the covariates $\mathbf{x}$, $R_j(\mathbf{x}) = 1$ or 0 depending on whether some conditions are satisfied by $\mathbf{x}$ or not, with $R_j(\mathbf{x})R_{j'}(\mathbf{x}) = 0$ for $j \neq j'$ and $\sum_j R_j(\mathbf{x}) = 1$. The weighted-CART algorithm can be expressed as it follows.

Step 1: $R_1(\mathbf{x}) = 1$ for all $\mathbf{x}$, and $n_1 = 1$.

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 $\delta_j R_j(\mathbf{X}_i) = 1$ have the same characteristics (i.e. if there exists a fixed value of $\mathbf{x}$ such that, for all i such that $\delta_j R_j(\mathbf{X}_i) = 1$, $\mathbf{X}_i = \mathbf{x}$), then keep rule j.

- else, rule j is replaced by two rules R'_{j1} and R'_{j2} determined in the following way: define x_l such that $x_l = \arg\min_x m_l(R_j, x)$, with

$$\begin{aligned} m_l(R_j, x) &= \sum_{i=1}^{n} \omega_i (\phi(N_i, T_i, \mathbf{X}_i) - \bar{n}_{l-}(x, R_j))^2 \mathbf{1}_{X_i^{(l)} \leq x} R_j(\mathbf{x}) \\ &\quad + \sum_{i=1}^{n} \omega_i (\phi(N_i, T_i, \mathbf{X}_i) - \bar{n}_{l+}(x, R_j))^2 \mathbf{1}_{X_i^{(l)} > x} R_j(\mathbf{x}), \end{aligned}$$

  where

$$\bar{n}_{l-}(x, R_j) = \frac{\sum_{i=1}^{n} \omega_i \phi(N_i, T_i, \mathbf{X}_i) \mathbf{1}_{X_i^{(l)} \leq x} R_j(x)}{\sum_{k=1}^{n} \omega_k \mathbf{1}_{X_k^{(l)} \leq x} R_j(x)}, \quad \bar{n}_{l+}(x, R_j) = \frac{\sum_{i=1}^{n} \omega_i \phi(N_i, T_i, \mathbf{X}_i) \mathbf{1}_{X_i^{(l)} > x} R_j(x)}{\sum_{k=1}^{n} \omega_k \mathbf{1}_{X_k^{(l)} > x} R_j(x)}.$$

  Then, select $\hat{l} = \arg\max_l m_l(R_j, x_l)$, and define $R'_{j1}(\mathbf{x}) = R_j(\mathbf{x}) \mathbf{1}_{x^{(l)} \leq x_{\hat{l}}}$, and $R'_{j2}(\mathbf{x}) = R_j(\mathbf{x}) \mathbf{1}_{x^{(l)} > x_{\hat{l}}}$.

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

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

In this version of the CART algorithm, all covariates are continuous or $\{0, 1\}-$valued. For qualitative variables with more than 2 modalities, they must be transformed into binary variables, or the algorithm must be slightly modified so that the splitting step of each $\mathcal{R}_i$ should be done by finding the best partition in two groups on the values of the modalities that minimizes the loss function.

Compared to the classical CART algorithm of Breiman et al. [1984], the splitting criterion (that is the quantity that is minimized at each step to decompose the population into two classes) is a weighted quadratic loss (instead of a quadratic loss) in order to compensate censoring, as explained in section 2.1. The path of the algorithm is a binary tree, whose leaves represent the different rules. Each set of rules $\mathcal{R} = (R_1, ..., R_K)$ is associated with an estimator of the regression function, that is $\hat{\mu}^{\mathcal{R}}(\mathbf{x}) = \sum_{j=1}^{K} \hat{\mu}_j R_j(\mathbf{x})$, where

$$\hat{\mu_j} = \frac{\sum_{i=1}^{n} \omega_i \phi(N_i, Y_i, \mathbf{X}_i) R_j(\mathbf{X}_i)}{\sum_{i=1}^{n} \omega_i R_j(\mathbf{X}_i)}.$$

Of course, this algorithm (called the *growth step*) does not provide a convenient estimate of the regression function $\mu(\mathbf{x})$ (it simply interpolates the data). The final set of rule of the growth step is called the "maximal tree". A *pruning step* must then be performed to extract a subtree from this maximal tree, in order to achieve some compromise between fit and complexity.

Let $K(\mathcal{R})$ denote the number of leaves (i.e. of rules) of a subtree. The pruning approach proposed by Breiman et al. [1984], adapted to our framework, consists of minimizing

$$\sum_{i=1}^{n} \omega_i (\phi(N_i, T_i, \mathbf{X}_i) - \hat{\mu}^{\mathcal{R}}(\mathbf{X}_i))^2 + \frac{\alpha K(\mathcal{R})}{n},$$

where $\alpha > 0$ is chosen through cross-validation or using a validation sample. Consistency of this approach (that is the capacity of this penalization strategy to select the proper subtree) has been shown by Lopez et al. [2016].

2.3 Our algorithm to estimate reserves in practice

We detail here the steps to implement our weighted CART algorithm in the context of individual claim reserving. For the sake of simplicity but without loss of generality, consider that the insurer has to pay 1 US$ each day the claim remains open, which corresponds to the case $M = T$. Consider an open claim, that is $\delta = 0$, and the claim is opened since $Y = k$. We thus aim to estimate RBNS claims by the quantity $\mathbb{E}[T \,|\, \delta = 0, Y = k, \mathbf{X}] = \mathbb{E}[T \,|\, T \geq k, \mathbf{X}]$. In this context, there is a direct link between duration of the claim and final claim amount. In this illustration, our weighted-CART algorithme can be expressed as it follows.

Step 0: Let's denote k_i the i-th right-censored observation.

Step 1: Estimate the Kaplan-Meier weights from the whole data.

Step $i+1$:

- Select claims (potentially censored) with higher lifetime than k_i;
- Build the regression tree $(T - k_i) \,|\, \mathbf{X}, T > k_i$; based on weighted observations ;
- Prune appropriately the obtained tree (see Lopez et al. [2016]);

- Estimate the residual lifetime : $E[T - k_i \,|\, T > k_i, \mathbf{X}]$;
- $i = i + 1$ and go back to step $i + 1$.

Let us note that the weight are computed from the whole data. Once the regression tree is built, the final claim amount can be estimated for each open claim. The behavior of the method is expected to be poorer for the claims with the largest settlement times, which is essentially due to two facts: the lack of claims such that $T > k$; and the erratic behavior of the weights when T becomes too large, which is a classical issue when dealing with the Kaplan-Meier estimator. Nevertheless, the question of extreme claims would require a particular attention, which is not covered by regression trees.

Remark 2.1. *In some situations, time-dependent covariates may be present. If the $j-th$ component of $\mathbf{X}$ is time dependent, the function $t \to X^{(j)}(t)$ can be discretized by considering some grid of times $(t_1, ..., t_k)$. This would not be an obstacle to compute CART algorithms such as the one of section 2.2. However, if we want to predict the final amount of a censored claim, the problem is that we do not have knowledge of the evolution of $X^{(j)}$ after the censoring time. A possibility would be to develop a prediction model for the evolution of $X^{(j)}$, and then plug the predicted evolution in the algorithm of section 2.3.*

3 Application

The goal of this section is to see whether individual claim reserving using our nonparametric approach leads to improve significantly the initial assessment of the global reserve corresponding to RBNS claims, as compared to Chain Ladder. In this view, we make comparisons based on a very simple indicator: the *boni-mali* (see Section 1). *Boni-mali* is useful to backtest the quality of predictions made for the expected global reserve.

Real-life claims are usually stored in a database where each record corresponds to one unique claim, with all corresponding characteristics (in particular the dates of claim occurrence and closure, if available). Then, reserves are regularly estimated using Chain Ladder or the weighted CART algorithm, and thus updated at given dates. Hereafter, reserves are estimated every quarter to remain as close as possible from practice. Indeed, the french regulation states that quarterly reports on reserves must be provided by insurers. This process enables to compute the *boni-mali* between each period. Implementing the Chain Ladder method requires to appropriately aggregate the data (see Section 3.2), whereas we need to define a grid of durations to be studied in the second case (see the

parameter k in the algorithm of Section 2.3). This grid obviously depends on the data, and further details are provided in Section 3.3.

3.1 Data description

When looking at aggregated loss triangles, practitioners usually consider that long-term risks are characterized by more than ten developments periods. Here, liabilities (or guarantees) can last much longer. Indeed, short-term and long-term disability insurance exist to protect the policyholders against the loss of some revenue, due to some accident or disease that prevent them from working. Those type of contracts, mostly sold in collective insurance, can sometimes be assimilated into life annuities.

We focus here on short-term disability insurance. This kind of guarantee is based upon French Social Security guarantees. It provides payments ro the policyholder for each day in disability state, with a duration's limitation of 3 years for a single claims. In local GAAP, claims reserves have to be estimated, on an individual basis, using disability tables. Moreover claims not yet reported are generally estimated through triangle techniques. Nevertheless, for prudential purposes, best estimate calculations are expected. In the following we only focus on IBNER.

To simplify, say that each day corresponds to a payment of 1 US$. The real-life database we consider reports the claims of income protection guarantees over six years, from 12/31/2005 to 12/31/2011. It consists of 103 048 claims, with the following information for each claim: a policyholder ID, cause (89 461 sicknesses, 13 587 accidents), gender (21 912 males, 81 136 females), socio-professional category (SPC): 3 747 managers, 98 577 employees and 724 miscellaneous), age at the claim date, duration in the disability state (perhaps right-censored), commercial network (three kinds of brokers: 44 797 "*Net-A*", 7 471 "*Net-B*" and 50 780 "*Net-C*"). All insurance contracts considered have a common deductible of 30 days, and the overall censoring rate equals 5.5% at 12/31/2011 (of course this rate increases when considering the database at earlier observation dates, see Section 3.2). There is strong dispersion among the observed durations (or claim lifetimes, beyond the deductible), the standard deviation being 166 days. Some descriptive statistics are given in Table 1, and additional information are provided in Appendix B. Our goal is to predict the global capital to reserve, either by Chain Ladder or by our algorithm. In the latter case, it consists in predicting the residual lifetime in the disability state for each policyholder (given the individual features), knowing that this duration fully explains the claim amount here, like in most of countries for this type of insurance contracts in Europe.

Variable:	Type	Min.	Median	Mean	Std.	Max.
Occurrence	date	01/01/2006	02/16/2009	01/21/2009		11/30/2011
Beginning of payments	date	01/01/2006	03/18/2009	02/20/2009		12/30/2011
End of payments	date	01/01/2006	07/08/2009	06/03/2009		12/31/2011
Age at claim	continuous	18.05	41.55	40.43	9.4	55
Censored claim lifetime	continuous	1	110	206.6	223.7	1 060
Uncensored claim lifetime	continuous	1	40	96.5	160.2	1 095
Claim lifetime	continuous	1	42	102.6	166.3	1 095

Table 1: Statistics on numerical variables and event dates, as of 12/31/2011.

3.2 Building the database, and implementing

Reserves are periodically estimated, say each quarter between 12/31/2009 and 12/31/2010. Therefore, for every date, we look at the status of the claim (open, closed, new) since policyholders' health is likely to deteriorate, remain stable, or improve between two consecutive quarters. This process allows to regularly update the characteristics of claims, in particular report the newly declared claims, those that become settled, and the remaining ones (RBNS) requiring an updated computation of the corresponding reserve for coming periods. Table 2 illustrates, for three policyholders, how data are built through the historical pattern of claims. Estimation of the global reserve is made within our two frameworks, namely the Chain Ladder model and our weighted algorithm. Building the data this way, it is straightforward to get classical loss triangles so as to implement Chain Ladder technique given origin periods (quarters). Concerning individual claim reserving, it consists of using the algorithm described in Section 2.3 at the following dates: 12/31/2009, 03/31/2010, 06/30/2010, 09/30/2010, and 12/31/2010.

Let us now comment the different examples given in Table 2. All the three employees are women who suffered from sickness, other policyholders' s characteristics are reported. The first employee declared the sickness on 2008/18/01, and payments started on 2008/17/02. The insured's absence lasted 57 days, terminating on 2008/14/04.

When looking at the situation on 2009/31/12, this observation is thus not censored (this can be seen from the boolean *2009-12-31.NonCensure*, set to 'true'). In this case, there is nothing to reserve since the claim was settled and all payments were made (57$). That is why this observation is never censored and prediction from the weighting CART algorithm is useless, whatever the quarter under consideration.

The second policyholder, with a total sickness lifetime of 419 days, is an interesting example since it will typically enable us to backtest our future predictions. Indeed, the censorship indicator changes as time flies. The global censorship indicator indicates that this observation is fully observed in 12/31/2011 (the claim was settled on 07/29/2010). However, this is not the case when looking for instance on 12/31/2009. At that time, this employee is considered a censored observation: 209 days were already paid, but the claim is not closed. Backtesting shows that there are still 210$ to pay for, whereas weighted CART algorithm predicts that nearly 240$ should be reserved. One quarter later, i.e. on 03/30/2010, updates are made: actual payments were increased by 90$ (three months), and CART prediction equals 226$ for this individual reserve. Six months later (09/30/2010), the observation gets uncensored for the first time. There is thus no further prediction to provide, but this information is used by our algorithm (updating the KM weights given to other uncensored observations to perform the estimation).

Finally, the third example remains censored from the beginning to the end of the period where reserves are calculated (quarters from 12/31/2009 to 12/31/2010). Moreover, the claim is still open on 12/31/2011, and total payments exceed 950$ (990$ exactly). In this case, which seems to correspond to an extreme observation (recall the mean duration equals 100 days, and that the maximum equals 1095), notice that the weighted CART algorithm anticipates that there are still about 200$ to reserve, knowing that 625$ have already been paid. This statement reveals that our algorithm somewhat captured this extreme situation, which is all the more interesting that most expensive claims are often the longest ones in practice.

PH features	Payments and dates		Reporting date and updated information: 09/31/12	10/31/03	10/30/06	10/09/30	10/31/12
52 y.o.	beg: 2008/17/02	Censored claim?	No	No	No	No	No
Employee	end: 2008/14/04	Currently paid (in $):	57	57	57	57	57
Network A	finally paid: 57$	Still to pay (wCART)	NA	NA	NA	NA	NA
43 y.o.	beg: 2009/05/06	Censored claim?	Yes	Yes	Yes	No	No
Employee	end: 2010/29/07	Currently paid (in $):	209	299	390	419	419
Network C	finally paid: 419$	Still to pay (wCART)	239.7	226.4	234.7	NA	NA
50 y.o.	beg: 2009/15/04	Censored claim?	Yes	Yes	Yes	Yes	Yes
Employee	end: 2011/31/12	Currently paid (in $):	260	350	441	533	625
Network C	finally paid: 990$	Still to pay (wCART)	234.5	232.2	225.1	215.9	200.5

Table 2: Three claims with their pattern of payments, showing how we build the database.

3.3 Results on Boni-Mali and discussion

How to compute individual reserves has been explained in Sections 2.3 and 3.2. The aggregation of such reserves leads to approximate the expectation of the total reserve over the whole portfolio, at some given dates. This prediction can be compared to the Chain Ladder one for RBNS claims, for our five dates of interest (12/31/2009, 03/31/2010, 06/30/2010, 09/30/2010, and 12/31/2010). Figures **??** and 1 show the evolution of the different estimations (see also the aggregated data stored in triangle for Chain Ladder estimates in Appendix A), and several interesting remarks can be formulated.

First, the estimation of the ultimate cost of claims over the entire portfolio looks consistent, whatever the technique used (compare each bar to the last one, named "Ultime"). This is not surprising since the censoring rate is not so high (recall that it roughly equals 7%), which limits the interest of our method. Indeed, given that we focus here on RBNS claims to estimate the corresponding reserve, Chain Ladder can access almost the full information. However, it seems that long-tailed risks associated to a higher censoring rate would significantly increase its bias, leading to poor estimates of the ultimate costs and thus the global reserve.

Second, how to reach the ultimate cost is very different, depending on the technique under consideration. When using Chain Ladder, the global reserve provided at 12/31/2009 is clearly underestimated as compared to the one given by the weighted CART algorithm (compare the orange hatched area for Chain Ladder versus the green dotted one for weighted CART). People with high lifetimes were not overweighted with the standard Chain Ladder approach, since the pattern of individual claims is not taken into account. The global reserve is, by consequence, largely underestimated. On the contrary, weighted CART predictions lead to anticipate higher reserves from the beginning, which is interesting since potential future liquidity needs are then decreased. Looking at *boni-mali* indicators between each period confirms this, are summarized in Table 3. Clearly, the weighted CART algorithm is powerful on such a criterion, and capital injection needs would be impressively decreased (almost 150 000 US$ would be saved in this case on an annual basis). Notice that this statement is not true for the last quarter under study, which was expected since the censored part of the lifetimes decreases little by little.

Another way to illustrate errors from both models is given in Figure 2. We compare prediction errors on both ultimate claims amount (solid line) and reserves (dotted line). Our method appears to be dominent, especially for the first estimations. This result is quite interesting since, beyond being able to predict individual reserves, their aggregation

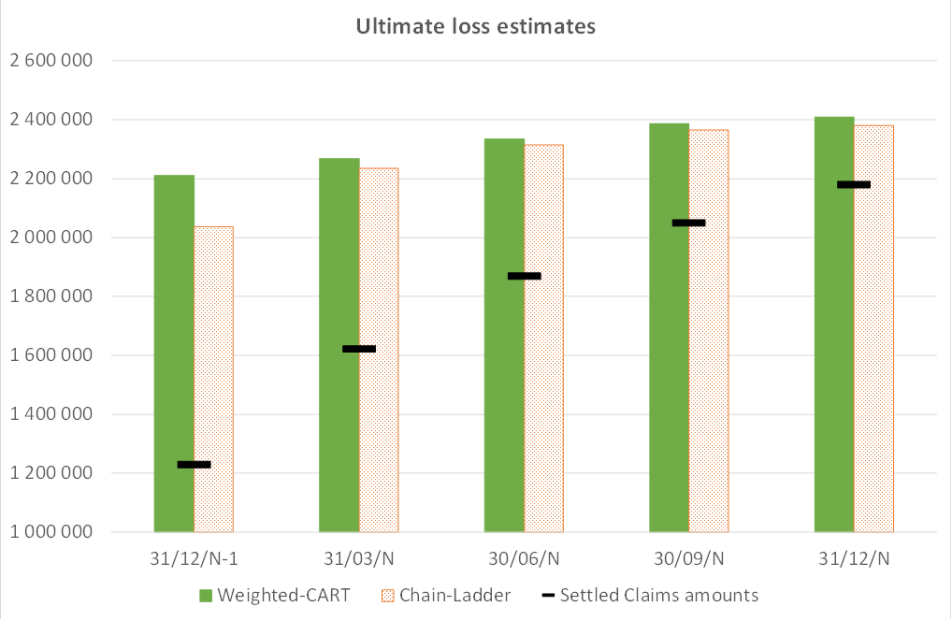


Figure 1: Evolution of reserves estimated by both Chain Ladder and the weighted CART algorithm ('N' refers to the year 2010). For each bar, the filled area is the amount already paid, whereas the hatched one is the estimated reserve.

leads to better apprehend the overall need of reserves.

4 Concluding remarks and on-going research

In this paper, we proposed a simple algorithm based on nonparametric techniques to estimate RBNS claims in non-life insurance. Such techniques have a lot of advantages, the greatest one being that they allow to integrate the history of claims in the final estimation without specifying a parametric relationship. Our estimator is more responsive to any changes in the development patterns of claims, which makes it naturally adapted to long-tailed claim developments (e.g. in Third Party Liability insurance). Practically speaking, this is extremely important since experts know that the final claim amount is highly dependent on the final development time. However, using our algorithm requires

Boni (+) / Mali (-)	T1	T2	T3	T4	Annual
Chain-Ladder	-196 814	-80 173	-51 209	-14 394	-342 591
Weighted CART	-58 515	-65 743	-51 989	-21 801	-198 047

Table 3: Boni-mali indicators: *mali* require capital injections.

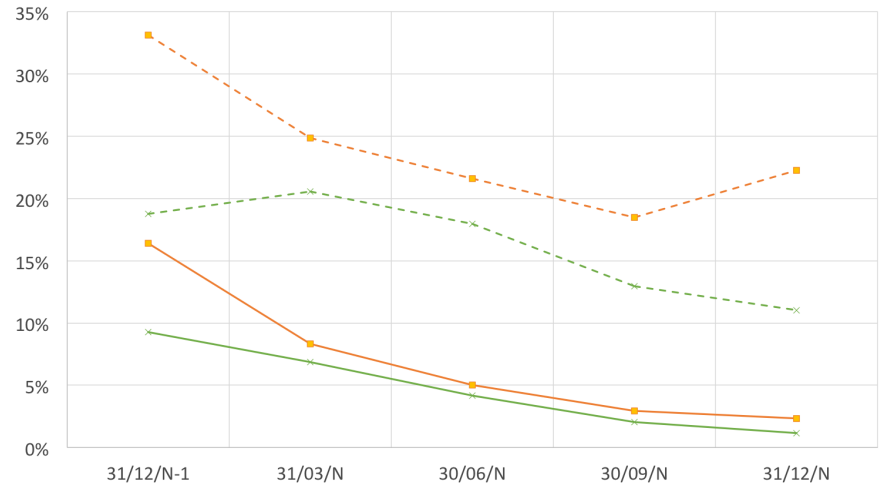


Figure 2: Errors (in %) of Chain Ladder method (lines with square marks) and weighted CART algorithm (lines with cross marks), at different dates of estimation. Solid lines correspond to the ultimate claim amounts, and dotted lines concerns the reserves.

a comprehensive database which is not always available in reserving departments, and has to be built by gathering information from different services. When working with very short-term and well-known risks, Chain Ladder and its extensions still seem to be a better trade-off to estimate the total reserve. Of course, our technique could be improved in several ways. We first think about the extension to the assessment of risk measures, and uncertainty of predictions. Such tasks would require to change the loss function used into the building process of the tree, going from standard Mean Squared Error (MSE) to Mean Absolute Error (MAE) or likelihood maximization (ML) for instance.

Acknowledgments

This work is supported by Institut des Actuaires (French National Actuarial Association), and BNP Paribas through the Research Chair "Data Analytics and Models for Insurance".

A Triangles for RBNS claims

We here give the (non cumulated) triangles corresponding to RBNS claims at the five considered dates for the estimations (some figures have been slightly approximated to print it). Recall that there cannot be more than twelve quarters for development since the insured risk is short-term disability (capped at three years).

We first present the aggregated data at 12/31/2009, leading to a Chain Ladder reserve of 812 862$.

	dev1	dev2	dev3	dev4	dev5	dev6	dev7	dev8	dev9	dev10	dev11	dev12	dev13
01/01/2006	173034	68439	41810	31000	22771	17819	14420	11649	8458	6355	4221	1730	0
04/01/2006	171994	66507	40492	29561	21654	16433	12638	9937	8356	6818	4732	2001	0
07/01/2006	154731	64126	38448	28495	21676	17139	12784	9940	8217	6212	4381	2347	0
10/01/2006	212830	86313	51881	39540	29456	22104	18492	14560	10618	7923	5814	2496	0
01/01/2007	189416	75222	45239	32859	23653	19099	15653	12397	9647	7509	5375	1528	
04/01/2007	182655	72237	42465	30751	22770	17245	13304	10731	8187	6359	3012		
07/01/2007	176286	73766	44724	34764	26256	19879	15757	12931	10080	3951			
10/01/2007	236100	96089	57422	42463	31755	25174	20616	16368	6798				
01/01/2008	204179	82000	52283	37630	27601	21640	17243	7422					
04/01/2008	207794	83240	53037	39657	30581	24099	9706						
07/01/2008	185298	79989	46343	35209	28109	11371							
10/01/2008	244596	101740	64641	48016	19574								
01/01/2009	207585	84078	51955	19631									
04/01/2009	217428	90243	31482										
07/01/2009	193560	45537											
10/01/2009	155370												

Then at 03/31/2010, leading to Chain Ladder reserve of 816 783$:

	dev1	dev2	dev3	dev4	dev5	dev6	dev7	dev8	dev9	dev10	dev11	dev12	dev13
01/01/2006	173034	68439	41810	31000	22771	17819	14420	11649	8458	6355	4221	1730	0
04/01/2006	171994	66507	40492	29561	21654	16433	12638	9937	8356	6818	4732	2001	0
07/01/2006	154731	64126	38448	28495	21676	17139	12784	9940	8217	6212	4381	2347	0
10/01/2006	212830	86313	51881	39540	29456	22104	18492	14560	10618	7923	5814	2496	0
01/01/2007	189416	75222	45239	32859	23653	19099	15653	12397	9647	7509	5375	2374	0
04/01/2007	182655	72237	42465	30751	22770	17245	13304	10731	8187	6359	5096	1614	
07/01/2007	176286	73766	44724	34764	26256	19879	15757	12931	10080	7186	2530		
10/01/2007	236100	96089	57422	42463	31755	25174	20616	16368	12583	4423			
01/01/2008	204179	82000	52283	37630	27601	21640	17243	13723	6115				
04/01/2008	207794	83240	53037	39657	30581	24099	18545	7680					
07/01/2008	185298	79989	46343	35209	28109	21040	7884						
10/01/2008	244596	101740	64641	48016	34920	14810							
01/01/2009	207585	84078	51955	36238	13595								
04/01/2009	217428	90243	55996	21611									
07/01/2009	193560	78175	26029										
10/01/2009	253249	58351											
01/01/2010	130111												

Then at 06/30/2010, leading to Chain Ladder reserve of 835 609$:

	dev1	dev2	dev3	dev4	dev5	dev6	dev7	dev8	dev9	dev10	dev11	dev12	dev13
01/01/2006	173034	68439	41810	31000	22771	17819	14420	11649	8458	6355	4221	1730	0
04/01/2006	171994	66507	40492	29561	21654	16433	12638	9937	8356	6818	4732	2001	0
07/01/2006	154731	64126	38448	28495	21676	17139	12784	9940	8217	6212	4381	2347	0
10/01/2006	212830	86313	51881	39540	29456	22104	18492	14560	10618	7923	5814	2496	0
01/01/2007	189416	75222	45239	32859	23653	19099	15653	12397	9647	7509	5375	2374	0
04/01/2007	182655	72237	42465	30751	22770	17245	13304	10731	8187	6359	5096	2329	0
07/01/2007	176286	73766	44724	34764	26256	19879	15757	12931	10080	7186	4964	1524	
10/01/2007	236100	96089	57422	42463	31755	25174	20616	16368	12583	8743	2650		
01/01/2008	204179	82000	52283	37630	27601	21640	17243	13723	11275	4909			
04/01/2008	207794	83240	53037	39657	30581	24099	18545	14289	5765				
07/01/2008	185298	79989	46343	35209	28109	21040	15540	6261					
10/01/2008	244596	101740	64641	48016	34920	27019	11479						
01/01/2009	207585	84078	51955	36238	25454	9949							
04/01/2009	217428	90243	55996	40366	16139								
07/01/2009	193560	78175	48053	18823									
10/01/2009	253249	102356	33678										
01/01/2010	225533	49007											
04/01/2010	141390												

Then at 09/30/2010, leading to Chain Ladder reserve of 821 319$:

	dev1	dev2	dev3	dev4	dev5	dev6	dev7	dev8	dev9	dev10	dev11	dev12	dev13
01/01/2006	173034	68439	41810	31000	22771	17819	14420	11649	8458	6355	4221	1730	0
04/01/2006	171994	66507	40492	29561	21654	16433	12638	9937	8356	6818	4732	2001	0
07/01/2006	154731	64126	38448	28495	21676	17139	12784	9940	8217	6212	4381	2347	0
10/01/2006	212830	86313	51881	39540	29456	22104	18492	14560	10618	7923	5814	2496	0
01/01/2007	189416	75222	45239	32859	23653	19099	15653	12397	9647	7509	5375	2374	0
04/01/2007	182655	72237	42465	30751	22770	17245	13304	10731	8187	6359	5096	2329	0
07/01/2007	176286	73766	44724	34764	26256	19879	15757	12931	10080	7186	4964	2363	0
10/01/2007	236100	96089	57422	42463	31755	25174	20616	16368	12583	8743	5461	1570	
01/01/2008	204179	82000	52283	37630	27601	21640	17243	13723	11275	9056	4033		
04/01/2008	207794	83240	53037	39657	30581	24099	18545	14289	10436	4196			
07/01/2008	185298	79989	46343	35209	28109	21040	15540	12430	4923				
10/01/2008	244596	101740	64641	48016	34920	27019	21061	8936					
01/01/2009	207585	84078	51955	36238	25454	19159	7834						
04/01/2009	217428	90243	55996	40366	30478	12894							
07/01/2009	193560	78175	48053	36059	14168								
10/01/2009	253249	102356	62087	24756									
01/01/2010	225533	93119	28943										
04/01/2010	227286	53866											
07/01/2010	122136												

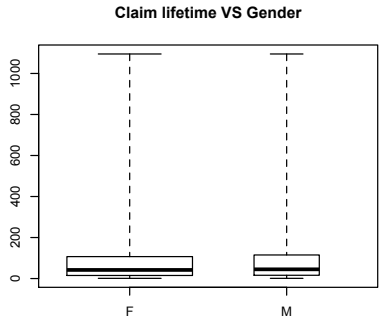
Then at 12/31/2010, leading to Chain Ladder reserve of 862 316$:

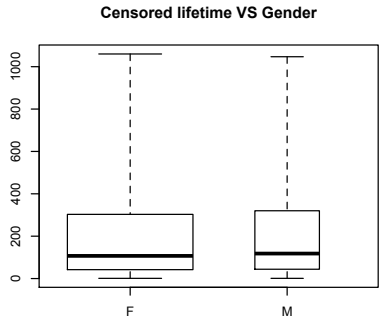
	dev1	dev2	dev3	dev4	dev5	dev6	dev7	dev8	dev9	dev10	dev11	dev12	dev13
01/01/2006	173034	68439	41810	31000	22771	17819	14420	11649	8458	6355	4221	1730	0
04/01/2006	171994	66507	40492	29561	21654	16433	12638	9937	8356	6818	4732	2001	0
07/01/2006	154731	64126	38448	28495	21676	17139	12784	9940	8217	6212	4381	2347	0
10/01/2006	212830	86313	51881	39540	29456	22104	18492	14560	10618	7923	5814	2496	0
01/01/2007	189416	75222	45239	32859	23653	19099	15653	12397	9647	7509	5375	2374	0
04/01/2007	182655	72237	42465	30751	22770	17245	13304	10731	8187	6359	5096	2329	0
07/01/2007	176286	73766	44724	34764	26256	19879	15757	12931	10080	7186	4964	2363	0
10/01/2007	236100	96089	57422	42463	31755	25174	20616	16368	12583	8743	5461	2572	0
01/01/2008	204179	82000	52283	37630	27601	21640	17243	13723	11275	9056	7152	2680	
04/01/2008	207794	83240	53037	39657	30581	24099	18545	14289	10436	7771	3125		
07/01/2008	185298	79989	46343	35209	28109	21040	15540	12430	9482	3887			
10/01/2008	244596	101740	64641	48016	34920	27019	21061	16001	6332				
01/01/2009	207585	84078	51955	36238	25454	19159	14618	5528					
04/01/2009	217428	90243	55996	40366	30478	23651	10269						
07/01/2009	193560	78175	48053	36059	26884	10615							
10/01/2009	253249	102356	62087	44650	17932								
01/01/2010	225533	93119	55235	20466									
04/01/2010	227286	90589	32462										
07/01/2010	200568	47162											
10/01/2010	166502												

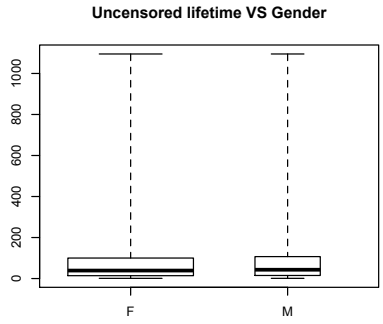
B Boxplots and histograms of our data

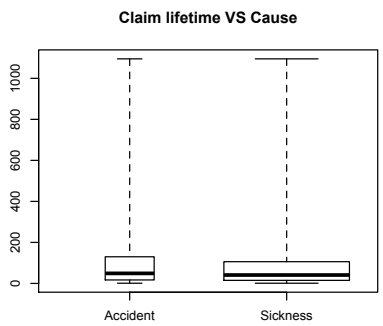
B.1 Boxplots

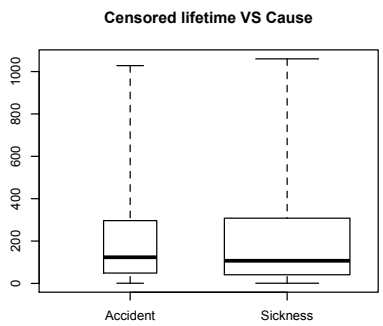
We focus here on the distribution of numerical variables in our database. The boxplots (where the size of each box is proportional to the size of the corresponding population) report the following information: minimum ('whisker' at the bottom), first quartile ('hinge' at the bottom), median, third quartile ('hinge' at the top), and maximum ('whisker' at the top). It enables to easily figure out the dispersion of the variable under study. Here, for each categorical explanatory variables, we study the difference between claim lifetimes depending on the category under study, whatever the status of the claim (still open or closed). Moreover, we also show that these statistics can significantly vary when considering only censored claims, or only uncensored claims.

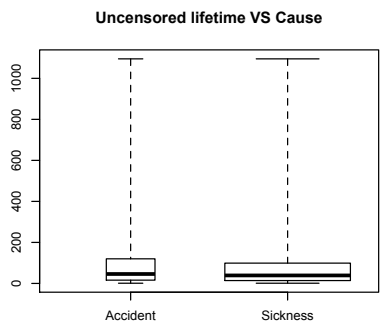

Claim lifetime VS Gender
0
200
400
600
800
1000
F
M

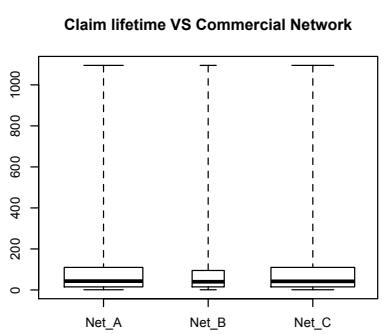

Censored lifetime VS Gender
0
200
400
600
800
1000
F
M

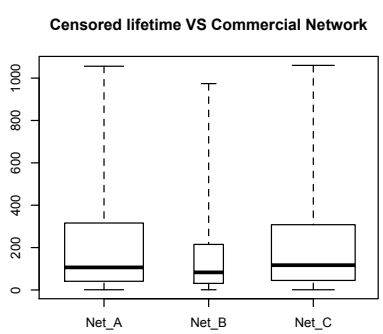

Uncensored lifetime VS Gender
0
200
400
600
800
1000
F
M

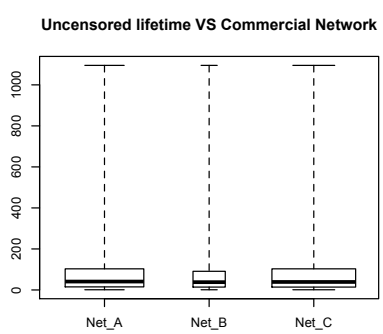

Claim lifetime VS Cause
0
200
400
600
800
1000
Accident
Sickness


Censored lifetime VS Cause
0
200
400
600
800
1000
Accident
Sickness

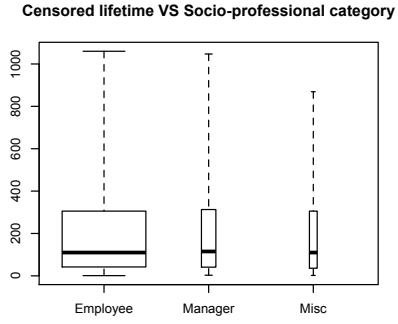

Uncensored lifetime VS Cause
0
200
400
600
800
1000
Accident
Sickness

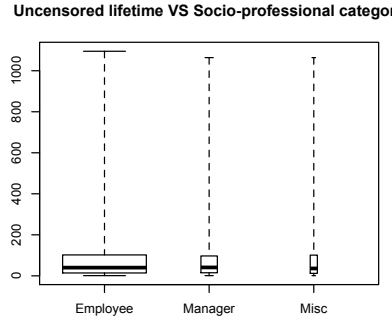

Claim lifetime VS Commercial Network
0
200
400
600
800
1000
Net_A
Net_B
Net_C


Censored lifetime VS Commercial Network
0
200
400
600
800
1000
Net_A
Net_B
Net_C


Uncensored lifetime VS Commercial Network
0
200
400
600
800
1000
Net_A
Net_B
Net_C

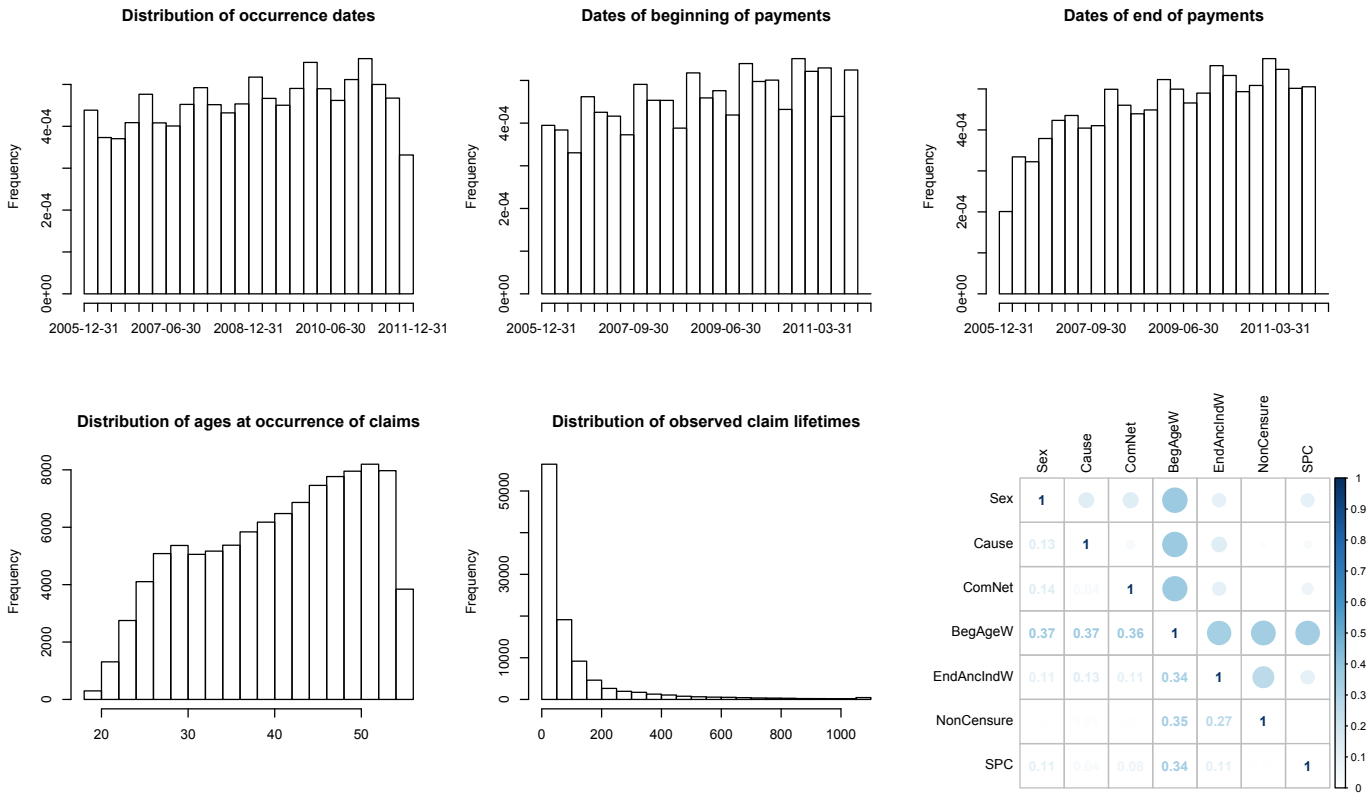

Claim lifetime VS Socio-professional category
0
200
400
600
800
1000
Employee
Manager
Misc


Censored lifetime VS Socio-professional category
0
200
400
600
800
1000
Employee
Manager
Misc


Uncensored lifetime VS Socio-professional category
0
200
400
600
800
1000
Employee
Manager
Misc

B.2 Histograms

We now give some details about the distribution of numerical variables, as well as some information about their association through the V-cramer measure. Notice that the claim lifetime (variable denoted by 'EndAncIndW') is mainly associated with the policyholder's age (variable 'BegAgeW').



References

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

M. Baudry and C.Y. Robert. Non parametric individual claim reserving in insurance. Working Paper, 2017.

R Bornhuetter and R E Ferguson. The actuary and IBNR. *Casualty Actuarial Society*, 59:181–195, 1972.

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

P.D. England and R.J. Verrall. Stochastic claims reserving in general insurance (with discussion). *British Actuarial Journal*, 8(3):443–544, 2002.

S. Haastrup and E. Arjas. Claims reserving in continuous time: a nonparametric bayesian approach. *ASTIN Bulletin*, 2:139–164, 1993.

L.J. Halliwell. Chain-ladder bias: Its reason and meaning. *Variance*, 1(2):214–247, 2007. doi: 10.1080/03461238.2018.1428681.

Jonas Harnau. Misspecification Tests for Chain-Ladder Models. Technical Report 840, Discussion Paper Series, Department of Economics, University of Oxford, 2017.

C. Larsen. An individual claims reserving model. *ASTIN Bulletin*, 37(1):113–132, 2007.

Olivier Lopez. A censored copula model for micro-level claim reserving. working paper or preprint, February 2018. URL `https://hal.archives-ouvertes.fr/hal-01706935`.

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`.

T Mack. Distribution-free calculation of the standard error of chain-ladder reserve estimates. *ASTIN Bulletin*, 23:213–225, 1993.

Walter Olbricht. Tree-based methods: a useful tool for life insurance. *European Actuarial Journal*, 2(1):129–147, 2012. doi: 10.1007/s13385-012-0045-5.

M. Pigeon, K. Antonio, and M. Denuit. Individual loss reserving with the multivariate skew normal framework. *ASTIN Bulletin*, 43(3):399–428, 2013.

G Quarg and T Mack. Munich Chain Ladder: A Reserving Method that Reduces the Gap between IBNR Projections Based on Paid Losses and IBNR Projections Based on Incurred Losses. *Variance*, 2(2):266–299, 2008.

M.V. Wüthrich. Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, 2018, 2018. doi: 10.1080/03461238.2018.1428681.

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