A TREE-BASED ALGORITHM ADAPTED TO MICROLEVEL RESERVING AND LONG DEVELOPMENT CLAIMS

BY

OLIVIER LOPEZ, XAVIER MILHAUD AND PIERRE-E. THÉROND*

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 limited. Through two illustrative examples and introducing an appropriate tree-based algorithm, we show that individual claim reserving can be really promising, especially in the context of long-term risks.

KEYWORDS

Reserving, long tail, censoring, regression tree, disability.

1. INTRODUCTION

Given its greater complexity, is it worth using individual claims reserving in non-life insurance? In this paper, we explain how to adapt the famous CART technique to censored data, and suggest a way to appropriately implement this tree-based algorithm for individual claim reserving purposes. Despite some recent advances, insurance companies still seem to be reluctant to use micro-level reserving as compared to very standard techniques using aggregated data[1], like Chain Ladder and its extensions (Bornhuetter and Ferguson, 1972; Mack, 1993; 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,

*The affiliation for Pierre-E. Thérond has been updated since this article's original publication. An erratum detailing this change has also been published. See https://doi.org/10.1017/asb.2019.21.

Astin Bulletin 49(3), 741-762. doi:10.1017/asb.2019.12

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 can sometimes be of primary importance.

Simultaneously, spectacular improvements to collect historical information and individual features on claims have been made in the insurance industry for more than 15 years, and companies now often have access to comprehensive data sets. 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. Separating RBNS and IBNyR claims allows to perform an advanced risk assessment and monitoring. 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 the characteristics of claims offers many advantages to approximate the reserve. First, it enables to cope with heterogeneity issues that can arise when using aggregated data. Second, 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) is now inputs of the model. Indeed, storing all claims into aggregate run-off triangles makes it impossible to consider neither the crucial risk factors explaining the final claim cost, nor the changes related to claims management, reinsurance programs, legal context, and product mix. Last but not least, micro-level reserving provides individual claims reserves, which is very useful from both a risk management and a claims management perspective (for instance, in order to improve claims management policies).

That being said, a natural question arises: why have such techniques not been widely applied yet? Except that it is harder to implement, the reason seems quite obvious: past contributions on individual claim reserving were mainly focused on parametric models and likelihood maximisation (Haastrup and Arjas, 1993; Larsen, 2007; Zhao et al., 2009; Pigeon et al., 2013; Antonio and Plat, 2014). 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 drivers under study is obviously unknown in reality, and can sometimes be tricky to specify. Lastly, the regulation usually requires the disclosure by companies of gains/losses indicators on a regular basis (e.g., each quarter), the most famous one for reserve estimates being the so-called *Boni–Mali*. For this task, most of actuaries agree to say that parametric models did not lead to any improvement, meaning that their overall quality of prediction is not better than the Chain Ladder's one (at least on the short term). Since one of the main threat for the top management concerns potential urgent need for capital injections, the attractiveness of such techniques thus remains low in practice.

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 important to be aware that many other actuarial applications could use our technique (e.g., the opportunity to decrease costs related to experts involved in claim estimations, the targeting of specific claims causing atypical claim amounts). However, the method presented hereafter strongly depends on the quality and consistency of the available data. This constraint also explains the slow evolution of the insurance market toward micro-level reserving approaches until now. Indeed, developing the methodology that we propose can be limited by the lack of reliable data when their sources are heterogeneous (such as in the case where a company experienced mergers and/or acquisitions). The extension of our approach to such cases is left for future work.

The paper is organized as follows: Section 2 introduces our algorithm to estimate individual reserves, with tools similar to Wüthrich (2018a). However, ultimate individual reserves for RBNS claims are here estimated thanks to an adaptation of the CART algorithm to censored data. Then, two real-life applications are conducted in Section 3. 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 (Baudry and Robert, 2017; Wüthrich, 2018a,b). 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. We summarize below this approach, knowing that further details can be found in 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)). Recall that we wish to estimate the ultimate charge of RBNS claims for individual policies, and then deduce individual predictions of reserves.

2.1. A weighting procedure for duration analysis

As shown in various studies (Lopez, 2018), the time development of a claim seems to be crucial to predict its severity. Not surprisingly, a claim which requires more time to be settled is more likely to be associated with a large amount. Indeed, longest developments often mean that some (costly) underlying procedures are required (this is typical in Third-Party Liabilities insurance,

https://doi.org/10.1017/asb.2019.12 Published online by Cambridge University Press

for instance). Therefore, if M denotes the claim amount, one must provide a model that takes the impact on M 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), some 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 are made up of i.i.d. replications $(N_i, Y_i, \delta_i, \mathbf{X}_i)_{1 \leq i \leq n}$, with n the number of observations. We assume that C is independent of $(M, T, \mathbf{X})$, which implies that 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) weight. For each observation i, this weight is defined by

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

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 following result: 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,

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

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 $\omega_i^* = \delta_i n^{-1}$

$[1 - G(Y_i -)]^{-1}$, because G is usually unknown and has to be estimated. These weights are therefore a convenient way to correct the bias caused by the censoring, since $E[\phi(M, T, \mathbf{X})]$ is consistently estimated by the weighted mean $\sum_{i=1}^{n} \omega_i^* \phi(N_i, Y_i, \mathbf{X}_i)$. Concretely, the KM weight equals 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 $\mu(\mathbf{x}) = E[\phi(M, T) \mid \mathbf{X} = \mathbf{x}]$ has to be estimated. We use the following modified CART algorithm introducing the previous KM weights, computed once for all at the beginning: at each step of the algorithm, one determines "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 wCART algorithm can then be expressed as follows:

Step 1: $R_1(\mathbf{x}) = 1$ for all $\mathbf{x}$, 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 $\delta_i R_j(\mathbf{X}_i) = 1$ 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 $X^{(l)}$ of $\mathbf{X}$ ($l = 1, ..., d$), define the best threshold $x_\star^{(l)}$ to split the data, such that $x_\star^{(l)} = \arg\min_{x^{(l)}} m(R_j, x^{(l)})$, with

$$m(R_j, x^{(l)}) = \sum_{i=1}^{n} \omega_i \big(\phi(N_i, T_i, \mathbf{X}_i) - \bar{n}_{l-}(x^{(l)}, R_j)\big)^2 \mathbf{1}_{X_i^{(l)} \leq x^{(l)}} R_j(\mathbf{x}) + \sum_{i=1}^{n} \omega_i \big(\phi(N_i, T_i, \mathbf{X}_i) - \bar{n}_{l+}(x^{(l)}, R_j)\big)^2 \mathbf{1}_{X_i^{(l)} > x^{(l)}} R_j(\mathbf{x}),$$

  where

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

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

  Then, select the best component index to consider: $\hat{l} = \arg\min_l m(R_j, x_\star^{(l)})$. Define the two new rules $R_{j1}(\mathbf{x}) = R_j(\mathbf{x}) \mathbf{1}_{x^{(\hat{l})} \leq x_\star^{(\hat{l})}}$, and $R_{j2}(\mathbf{x}) = R_j(\mathbf{x}) \mathbf{1}_{x^{(\hat{l})} > x_\star^{(\hat{l})}}$.
- 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 two modalities, they must be transformed into binary variables or the algorithm must be slightly modified, so that the splitting step of each R_j should be done by finding the best partition into 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 (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 trade-off between fit and complexity. Let $K(\mathcal{R})$ denote the number of leaves (or rules) of a subtree. The pruning approach proposed by Breiman et al. (1984), adapted to our framework, consists of minimizing

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

where $\alpha > 0$ is chosen through cross-validation or using a validation sample. Consistency of this approach (i.e., 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 wCART 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, with current observed lifetime equal to k. We wish to estimate the final cost of RBNS claims by $\mathbb{E}[T \mid \delta = 0, Y = k, \mathbf{X}]$ or equivalently $\mathbb{E}[T \mid T \geq k, \mathbf{X}]$. To this aim, we introduce the following algorithm: denote by k_i the i-th censored lifetime,

Step 1: estimate the Kaplan–Meier weights from the whole data.
Step $i + 1$: do successively,

https://doi.org/10.1017/asb.2019.12 Published online by Cambridge University Press

– select claims (potentially censored) with higher lifetime than k_i;

– build the regression tree $(T - k_i) \mid \mathbf{X}, T > k_i$; based on weighted observations;

– prune appropriately the obtained maximal tree;

– estimate the residual lifetime : $E[T - k_i \mid T > k_i, \mathbf{X}]$;

– $i = i + 1$ and go back to step $i + 1$.

Note that the weights are computed from the whole data. For each open claim, a regression tree is then computed following the algorithm of Section 2.2. 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 claims with the largest settlement times. This is essentially due to 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, 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 l-th component of* $\mathbf{X}$ *is time dependent, the function* $t \to X^{(l)}(t)$ *can be discretized by considering some grid of times* $(t_1, ..., t_k)$*. This would not be an obstacle to run our algorithm. However, if we wish to predict the final amount of a censored claim, we would not have knowledge of the evolution of* $X^{(l)}$ *after censorship. One thus needs to develop a prediction model for the evolution of* $X^{(l)}$*, and then plug it into the algorithm of Section 2.2.*

3. Numerical illustration of the CART-based approach

Real-life applications should enable to see whether our non-parametric individual claim reserving algorithm significantly improves the assessment of the overall reserve for RBNS claims. In this view, we make comparisons with Chain Ladder results, based on the *Boni–Mali* indicator (see Section 1). Boni–Mali is useful to backtest the quality of predictions made for the evolution of the expected global reserve.

Claims are usually stored in a database where each record stands for 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 wCART algorithm, and thus updated at given reporting dates. This process enables to compute the *Boni–Mali* between each period. 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. Notice that we also performed the study moving the time step from quarterly to monthly and bi-yearly, but this did not change our conclusions.

3.1. Data description

When looking at aggregate loss triangles, practitioners usually consider that long-term risks are characterized by more than 10 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 to the policyholder for each day in disability state, with a duration limitation of 3 years for one single claim. In local GAAP, claims reserves have to be estimated, on an individual basis, using disability tables. Moreover, IBNyR claims are generally estimated through triangle techniques. Nevertheless, for prudential purposes, best estimate calculations are expected.

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 6 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): 3747 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*," 7471 "*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 claim lifetimes (beyond the deductible), the standard deviation being 166 days. Some descriptive statistics are given in Table 1, as well as boxplots and histograms 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). Indeed, this duration fully explains the claim amount here, like in most of countries for this type of insurance contracts in Europe.

3.2. Data management

As already mentioned, reserves are periodically estimated, say each quarter between 12/31/2009 and 12/31/2010. Concerning either Chain Ladder or our individual claim reserving algorithm, it consists of using the techniques at the following dates: 12/31/2009, 03/31/2010, 06/30/2010, 09/30/2010, and 12/31/2010. Therefore, for every date, we look at the status of the claim (open, closed, and new) since policyholders' health is likely to deteriorate, remain

https://doi.org/10.1017/asb.2019.12 Published online by Cambridge University Press

TABLE 1

STATISTICS ON NUMERICAL VARIABLES AND EVENT DATES, AS OF 12/31/2011.

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

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 individual reserve for coming periods. Table 2 illustrates, for three policyholders, how data are built through the historical pattern of claims. Building the data this way, it is straightforward to get aggregate loss triangles, so as to use Chain Ladder at one given reporting date (see Appendix A to look at the triangles for the five reporting dates considered).

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

When looking at the situation on 12/31/2009, this observation is thus not censored. 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 wCART algorithm is useless (denoted by NA), whatever the reporting date 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 individual predictions. Indeed, the censorship indicator changes as time flies. The global censorship indicator indicates that this observation is fully observed at 12/31/2011 (the claim was settled on 07/29/2010). However, this is not the case when looking to this individual at 12/31/2009, where this employee is now considered as censored. The claim is not closed, and 209$ were already paid. Backtesting shows that there are still 210$ (419–209) to pay for, whereas wCART algorithm predicts that nearly 240$ should be reserved. One quarter later, that is, on 03/30/2010, updates are made: actual payments were increased

TABLE 2

EXAMPLES OF PATTERNS, SHOWING HOW THE DATABASE IS BUILT AT EACH REPORTING DATE.

PH features	Final observed payment and dates		Reporting date and updated information:				
			12/31/09	03/31/10	06/30/10	09/30/10	12/31/10
52 y.o.	Beg: 02/17/2008	Censored claim?	No	No	No	No	No
Employee	End: 04/14/2008	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: 06/05/2009	Censored claim?	Yes	Yes	Yes	No	No
Employee	End: 07/29/2010	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: 04/15/2009	Censored claim?	Yes	Yes	Yes	Yes	Yes
Employee	End: 12/31/2011	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

by 90$ (three months), and wCART reserve prediction equals 226$. 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 estimated. 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 that the mean duration equals 100 days and the maximum equals 1095), notice that the wCART algorithm anticipates that there are still about 200$ to reserve, knowing that 625$ have already been paid.

3.3. Results on Boni–Mali and discussion

How to compute individual reserves has been explained in Sections 2.3 and 3.2. With our method, the aggregation of such reserves leads to approximate the expectation of the global charge over the whole portfolio. This prediction can be compared to the Chain Ladder one, at our five dates of interest (12/31/2009, 03/31/2010, 06/30/2010, 09/30/2010, and 12/31/2010).

Figure 1 shows the evolution of the different estimations (recall that Chain Ladder estimates are provided in Appendix A). Several interesting remarks

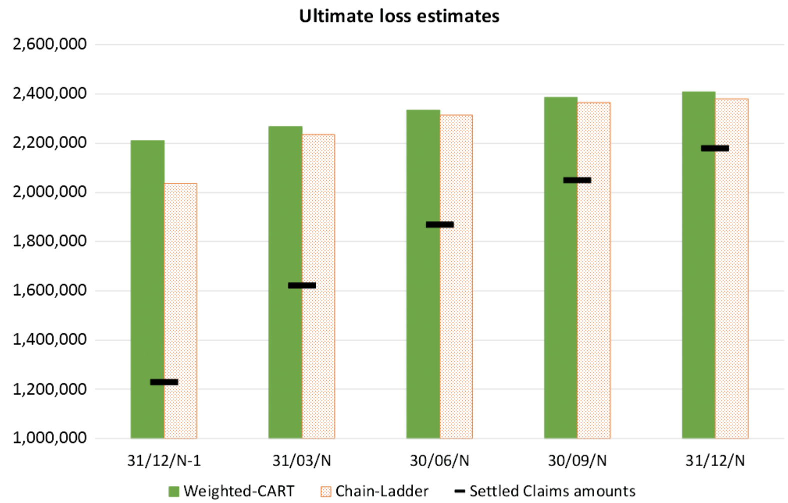


FIGURE 1: Evolution of reserves estimated by both Chain Ladder and the wCART algorithm ("N" refers to the year 2010), depending on the reporting date.

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). This makes sense since the censoring rate is not so high (recall that it roughly equals 7% in 2010), which limits the interest of the comparison between our method and Chain Ladder. Indeed, given that we focus here on RBNS claims to estimate the corresponding reserve, Chain Ladder can access almost the full information here. However, one can expect much more different results in the context of risks with long-tailed developments, simply because the censoring rate would be significantly higher. This should increase the bias of Chain Ladder and leads to poor estimates of the global reserve.

Second, how to reach the ultimate charge 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 wCART algorithm (compare the dotted area to the plain one). 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, wCART 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, as summarized in Table 3. For instance, almost 200,000 US$ more (on top of the reserve) are needed when estimating the global reserve by Chain Ladder at 03/31/2010, as compared to the estimation made at 12/31/2009. This is obviously no good news, as compared to the 60,000 US$ needed when using the wCART algorithm. Clearly, the latter is powerful on such a criterion, and capital injection needs would be impressively decreased (almost 150,000 US$ would be saved

TABLE 3

BONI–MALI INDICATORS CALCULATED BETWEEN THE FIVE REPORTING DATES: *Mali* REQUIRES CAPITAL INJECTIONS. T1 REFERS TO THE EVOLUTION FROM 12/31/2009 TO 03/31/2010, T2 REFERS TO THE PERIOD FROM 03/31/2010 TO 06/30/2010, AND SO ON.

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

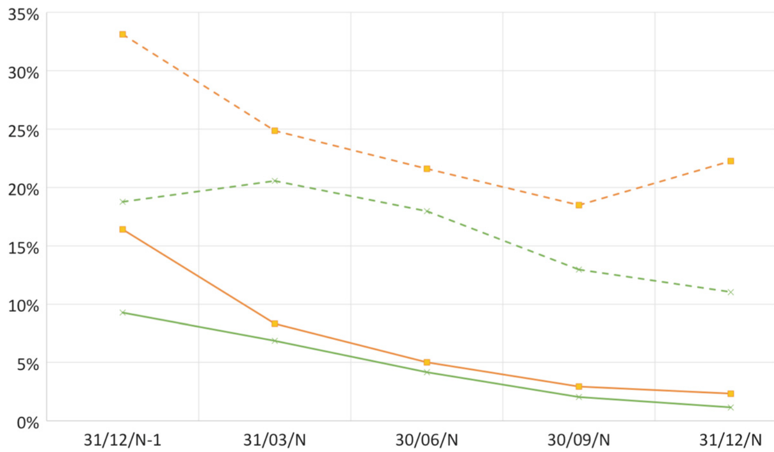


FIGURE 2: Errors (in %) of Chain Ladder method (square marks) and wCART algorithm (cross marks), at different reporting dates. Solid lines correspond to the ultimate claim amounts and dotted lines concern the reserves.

in this case on an annual basis). Notice that this statement is not true for the last quarter under study: this was expected since the percentage of the censored lifetimes decreases as time flies.

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 perform better, especially for the earliest reporting dates. This result is quite interesting since, beyond being able to predict individual reserves, their aggregation leads to better apprehend the overall need of reserves.

4. CONCLUDING REMARKS AND ON-GOING RESEARCH

In this paper, we propose 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. On one hand, Chain Ladder still seem to be a better trade-off to estimate the total reserve when working with very short-term and well-known risks, since it is very simple and exhibits good performance. On the other hand, our estimator is more responsive to any changes in the development patterns of claims, which makes it naturally adapted to unstable business lines or 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 positively correlated to the development time. However, using our algorithm requires a comprehensive database which is not always available in reserving departments, and has to be built by gathering information from different services. 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 to mean absolute error or likelihood maximization, 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." The authors would also like to thank the anonymous referees for their constructive comments that contributed to significantly improve the paper.

Notes

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
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).

References

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

Baudry, M. and Robert, C.Y. (2017) *Non parametric individual claim reserving in insurance*. Working paper.

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

Breiman, L., Friedman, J., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*. Boca Raton, FL: Chapman and Hall.

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

Haastrup, S. and Arjas, E. (1993) Claims reserving in continuous time: A nonparametric Bayesian approach. *ASTIN Bulletin*, **2**, 139–164.

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

Harnau, J. (2017) *Misspecification tests for chain-ladder models*. Technical Report 840, Discussion Paper Series, Department of Economics, University of Oxford.

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

LOPEZ, O. (2018) *A censored copula model for micro-level claim reserving*. Working paper or preprint, February. URL: https://hal.archives-ouvertes.fr/hal-01706935.

LOPEZ, O., MILHAUD, X. and THEROND, P-E. (2016) Tree-based censored regression with applications in insurance. *Electronic Journal of Statistics*, **10**, 2685–2716. dx.doi.org/10.1214/16-EJS1189.

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

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

PIGEON, M., ANTONIO, K. and DENUIT, M. (2013) Individual loss reserving with the multivariate skew normal framework. *ASTIN Bulletin*, **43**(3), 399–428.

QUARG, G. and MACK, T. (2008) 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.

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

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

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

OLIVIER LOPEZ
Sorbonne Université, CNRS
Laboratoire de Probabilités
Statistique et Modélisation
LPSM, 4 place Jussieu
F-75005 Paris, France

XAVIER MILHAUD (Corresponding author)
Université de Lyon
Université Claude Bernard Lyon 1
ISFA, LSAF,
F-69007, Lyon, France
E-Mail: xavier.milhaud@univ-lyon1.fr

PIERRE-E. THÉROND
Université de Lyon
Université Claude Bernard Lyon 1
ISFA, LSAF,
F-69007, Lyon, France
Galea & Associés
25 rue de Choiseul
75002, Paris, France

APPENDIX A. TRIANGLES FOR RBNS CLAIMS

We here give the (non-cumulated) triangles corresponding to RBNS claims at the five reporting dates for the estimations. Recall that there cannot be more than 12 quarters for development since the insured risk is short-term disability (capped at 3 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	173,034	68,439	41,810	31,000	22,771	17,819	14,420	11,649	8458	6355	4221	1730	0
04/01/2006	171,994	66,507	40,492	29,561	21,654	16,433	12,638	9937	8356	6818	4732	2001	0
07/01/2006	154,731	64,126	38,448	28,495	21,676	17,139	12,784	9940	8217	6212	4381	2347	0
10/01/2006	212,830	86,313	51,881	39,540	29,456	22,104	18,492	14,560	10,618	7923	5814	2496	0
01/01/2007	189,416	75,222	45,239	32,859	23,653	19,099	15,653	12,397	9647	7509	5375	1528	
04/01/2007	182,655	72,237	42,465	30,751	22,770	17,245	13,304	10,731	8187	6359	3012		
07/01/2007	176,286	73,766	44,724	34,764	26,256	19,879	15,757	12,931	10,080	3951			
10/01/2007	236,100	96,089	57,422	42,463	31,755	25,174	20,616	16,368	6798				
01/01/2008	204,179	82,000	52,283	37,630	27,601	21,640	17,243	7422					
04/01/2008	207,794	83,240	53,037	39,657	30,581	24,099	9,706						
07/01/2008	185,298	79,989	46,343	35,209	28,109	11,371							
10/01/2008	244,596	101,740	64,641	48,016	19,574								
01/01/2009	207,585	84,078	51,955	19,631									
04/01/2009	217,428	90,243	31,482										
07/01/2009	193,560	45,537											
10/01/2009	155,370												

https://doi.org/10.1017/asb.2019.12 Published online by Cambridge University Press

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	173,034	68,439	41,810	31,000	22,771	17,819	14,420	11,649	8458	6355	4221	1730	0
04/01/2006	171,994	66,507	40,492	29,561	21,654	16,433	12,638	9937	8356	6818	4732	2001	0
07/01/2006	154,731	64,126	38,448	28,495	21,676	17,139	12,784	9940	8217	6212	4381	2347	0
10/01/2006	212,830	86,313	51,881	39,540	29,456	22,104	18,492	14,560	10,618	7923	5814	2496	0
01/01/2007	189,416	75,222	45,239	32,859	23,653	19,099	15,653	12,397	9647	7509	5375	2374	0
04/01/2007	182,655	72,237	42,465	30,751	22,770	17,245	13,304	10,731	8187	6359	5096	1614	
07/01/2007	176,286	73,766	44,724	34,764	26,256	19,879	15,757	12,931	10,080	7186	2530		
10/01/2007	236,100	96,089	57,422	42,463	31,755	25,174	20,616	16,368	12,583	4423			
01/01/2008	204,179	82,000	52,283	37,630	27,601	21,640	17,243	13,723	6115				
04/01/2008	207,794	83,240	53,037	39,657	30,581	24,099	18,545	7680					
07/01/2008	185,298	79,989	46,343	35,209	28,109	21,040	7884						
10/01/2008	244,596	101,740	64,641	48,016	34,920	14,810							
01/01/2009	207,585	84,078	51,955	36,238	13,595								
04/01/2009	217,428	90,243	55,996	21,611									
07/01/2009	193,560	78,175	26,029										
10/01/2009	253,249	58,351											
01/01/2010	130,111												

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	173,034	68,439	41,810	31,000	22,771	17,819	14,420	11,649	8458	6355	4221	1730	0
04/01/2006	171,994	66,507	40,492	29,561	21,654	16,433	12,638	9937	8356	6818	4732	2001	0
07/01/2006	154,731	64,126	38,448	28,495	21,676	17,139	12,784	9940	8217	6212	4381	2347	0
10/01/2006	212,830	86,313	51,881	39,540	29,456	22,104	18,492	14,560	10,618	7923	5814	2496	0
01/01/2007	189,416	75,222	45,239	32,859	23,653	19,099	15,653	12,397	9647	7509	5375	2374	0
04/01/2007	182,655	72,237	42,465	30,751	22,770	17,245	13,304	10,731	8187	6359	5096	2329	0
07/01/2007	176,286	73,766	44,724	34,764	26,256	19,879	15,757	12,931	10,080	7186	4964	1524	
10/01/2007	236,100	96,089	57,422	42,463	31,755	25,174	20,616	16,368	12,583	8743	2650		
01/01/2008	204,179	82,000	52,283	37,630	27,601	21,640	17,243	13,723	11,275	4909			
04/01/2008	207,794	83,240	53,037	39,657	30,581	24,099	18,545	14,289	5765				
07/01/2008	185,298	79,989	46,343	35,209	28,109	21,040	15,540	6261					
10/01/2008	244,596	101,740	64,641	48,016	34,920	27,019	11,479						
01/01/2009	207,585	84,078	51,955	36,238	25,454	9949							
04/01/2009	217,428	90,243	55,996	40,366	16,139								
07/01/2009	193,560	78,175	48,053	18,823									
10/01/2009	253,249	102,356	33,678										
01/01/2010	225,533	49,007											
04/01/2010	141,390												

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	173,034	68,439	41,810	31,000	22,771	17,819	14,420	11,649	8458	6355	4221	1730	0
04/01/2006	171,994	66,507	40,492	29,561	21,654	16,433	12,638	9937	8356	6818	4732	2001	0
07/01/2006	154,731	64,126	38,448	28,495	21,676	17,139	12,784	9940	8217	6212	4381	2347	0
10/01/2006	212,830	86,313	51,881	39,540	29,456	22,104	18,492	14,560	10,618	7923	5814	2496	0
01/01/2007	189,416	75,222	45,239	32,859	23,653	19,099	15,653	12,397	9647	7509	5375	2374	0
04/01/2007	182,655	72,237	42,465	30,751	22,770	17,245	13,304	10,731	8187	6359	5096	2329	0
07/01/2007	176,286	73,766	44,724	34,764	26,256	19,879	15,757	12,931	10,080	7186	4964	2363	0
10/01/2007	236,100	96,089	57,422	42,463	31,755	25,174	20,616	16,368	12,583	8743	5461	1570	
01/01/2008	204,179	82,000	52,283	37,630	27,601	21,640	17,243	13,723	11,275	9056	4033		
04/01/2008	207,794	83,240	53,037	39,657	30,581	24,099	18,545	14,289	10,436	4196			
07/01/2008	185,298	79,989	46,343	35,209	28,109	21,040	15,540	12,430	4923				
10/01/2008	244,596	101,740	64,641	48,016	34,920	27,019	21,061	8936					
01/01/2009	207,585	84,078	51,955	36,238	25,454	19,159	7834						
04/01/2009	217,428	90,243	55,996	40,366	30,478	12,894							
07/01/2009	193,560	78,175	48,053	36,059	14,168								
10/01/2009	253,249	102,356	62,087	24,756									
01/01/2010	225,533	93,119	28,943										
04/01/2010	227,286	53,866											
07/01/2010	122,136												

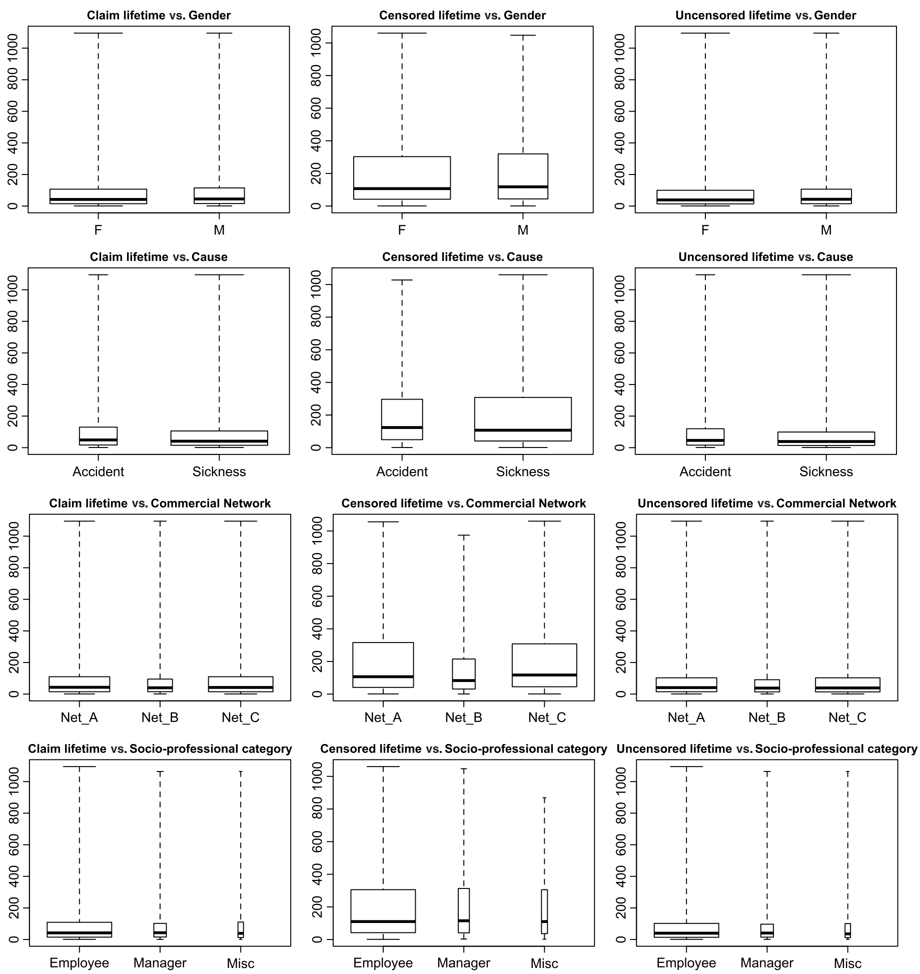
Then at 12/31/2010, leading to Chain Ladder reserve of 862,316$:

	dev1	dev2	dev3	dev4	dev5	dev6	dev7	dev8	dev9	dev10	dev11	dev12	dev13
01/01/2006	173,034	68,439	41,810	31,000	22,771	17,819	14,420	11,649	8458	6355	4221	1730	0
04/01/2006	171,994	66,507	40,492	29,561	21,654	16,433	12,638	9937	8356	6818	4732	2001	0
07/01/2006	154,731	64,126	38,448	28,495	21,676	17,139	12,784	9940	8217	6212	4381	2347	0
10/01/2006	212,830	86,313	51,881	39,540	29,456	22,104	18,492	14,560	10,618	7923	5814	2496	0
01/01/2007	189,416	75,222	45,239	32,859	23,653	19,099	15,653	12,397	9647	7509	5375	2374	0
04/01/2007	182,655	72,237	42,465	30,751	22,770	17,245	13,304	10,731	8187	6359	5096	2329	0
07/01/2007	176,286	73,766	44,724	34,764	26,256	19,879	15,757	12,931	10,080	7186	4964	2363	0
10/01/2007	236,100	96,089	57,422	42,463	31,755	25,174	20,616	16,368	12,583	8743	5461	2572	0
01/01/2008	204,179	82,000	52,283	37,630	27,601	21,640	17,243	13,723	11,275	9056	7152	2680	
04/01/2008	207,794	83,240	53,037	39,657	30,581	24,099	18,545	14,289	10,436	7771	3125		
07/01/2008	185,298	79,989	46,343	35,209	28,109	21,040	15,540	12,430	9482	3887			
10/01/2008	244,596	101,740	64,641	48,016	34,920	27,019	21,061	16,001	6332				
01/01/2009	207,585	84,078	51,955	36,238	25,454	19,159	14,618	5528					
04/01/2009	217,428	90,243	55,996	40366	30,478	23,651	10,269						
07/01/2009	193,560	78,175	48,053	36,059	26,884	10,615							
10/01/2009	253,249	102,356	62,087	44,650	17,932								
01/01/2010	225,533	93,119	55,235	20,466									
04/01/2010	227,286	90,589	32,462										
07/01/2010	200,568	47,162											
10/01/2010	166,502												

APPENDIX B. DESCRIPTIVE STATISTICS: BOXPLOTS AND HISTOGRAMS

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.

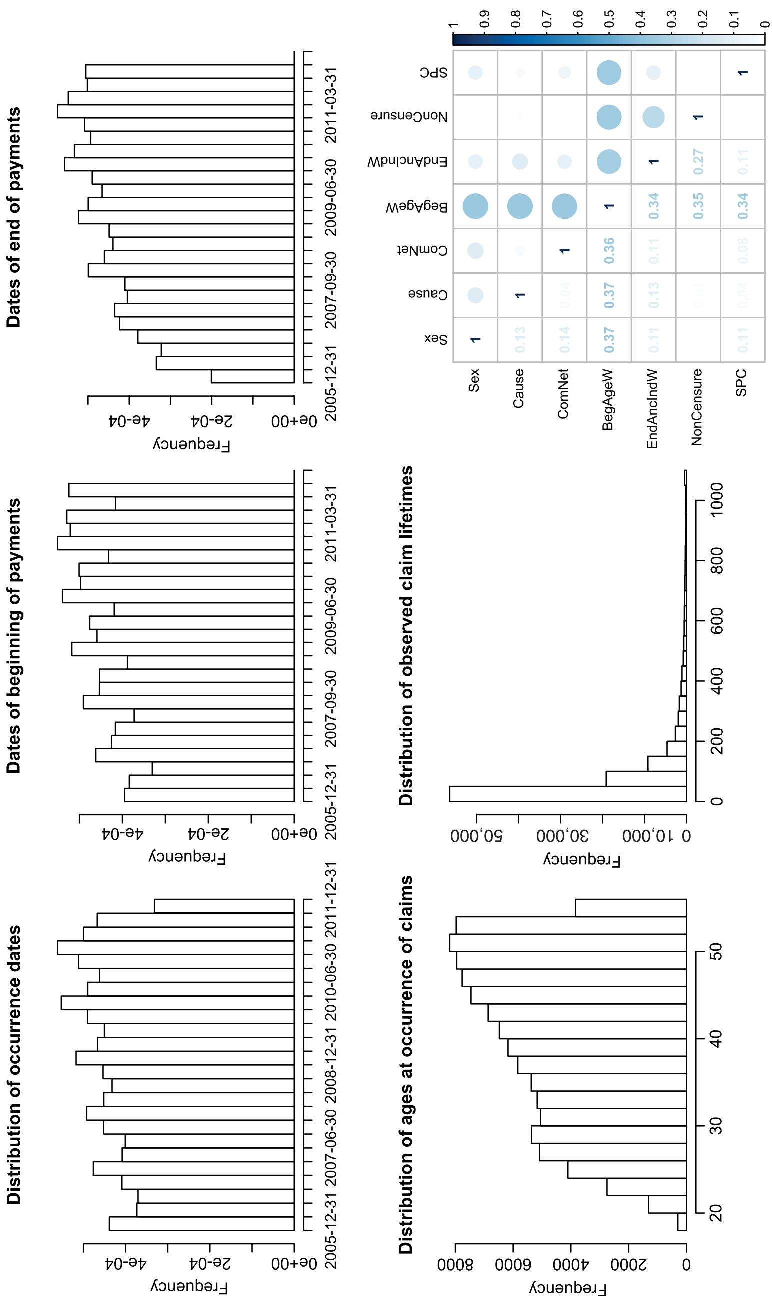


https://doi.org/10.1017/asb.2019.12 Published online by Cambridge University Press

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").

https://doi.org/10.1017/asb.2019.12 Published online by Cambridge University Press


Distribution of occurrence dates
Dates of beginning of payments
Dates of end of payments
Distribution of ages at occurrence of claims
Distribution of observed claim lifetimes
Frequency
Sex
Cause
ComNet
BegAgeW
EndAnclndW
NonCensure
SPC