Non parametric individual claim reserving in insurance

Maximilien BAUDRY* and Christian Y. ROBERT*

November 27, 2017

Abstract

Accurate loss reserves are an important item in the financial statement of an insurance company and are mostly evaluated by macro-level models with aggregate data in run-off triangles. In recent years, a new set of literature has considered individual claims data and proposed parametric reserving models based on claim history profiles. In this paper, we present a non parametric and flexible approach to estimate outstanding liabilities using all the covariables associated to the policy, its policyholder as well as all the informations received by the insurance company on the individual claim since its reporting date. The choice for a non parametric model leads to new issues since the target variables (claim occurrence and claim severity) are right-censored most of the time. The performance of our approach is evaluated by comparing the predictive values of the reserve estimates with their true values on simulated data. We also compare our individual approach with the most used aggregated classical method, Chain Ladder, with respect to the bias and the variance of the estimates.

1 Introduction

With increased requirement on financial reporting and ongoing solvency concerns, actuaries are faced with the need, more than ever, to deliver reliable estimates of claim costs and reserves. A number of deterministic or stochastic methods based on aggregate claims data structured in claims development triangles exist for calculating outstanding claims reserves (e.g. Chain Ladder (CL) method, Bornhuetter-Ferguson method, ...). These methods have had a great success to manage reserve risk for a variety of lines of business but it is known that they can suffer from underlying strong assumptions that give rise to several issues. Among them, one can cite: i) an over parameterization risk induced by a large number of tail factors that have to be estimated as compared to the number of components of the run-off triangles, ii) a risk of error propagation through the development factors associated to a possible huge estimation error for the latest development periods, iii) a potential lack of robustness and the need for appropriate treatments of outliers, iv) the impossibility to separate the assessments of Incurred But Not Reported (IBNR) and Reported But Not Settled (RBNS) claims,...

The existence of these issues and the substantial literature about them indicates that the use of aggregate data is not so fully adequate for capturing the complexities of stochastic reserving for general insurance, mainly because of an important loss of information when aggregating the original individual claim data details (e.g. times of occurrence, reporting delays, times and amounts of payments,...). Further, significant changes in technology (data collection, storage and analysis techniques) may make new approaches like a proper individual claims modeling now accessible.

[*] Université de Lyon, Université Lyon 1, Institut de Science Financière et d'Assurances (50 Avenue Tony Garnier, F-69007 Lyon, France) and Chair of excellence DAMI, Data Analytics and Models for Insurance (chaire-dami.fr/).

Therefore, recent research strongly promotes claims reserving on individual claims data, see, for instance, Antonio and Plat (2014), Badescu et al. (2016), Hiabu et al. (2016), Larsen (2007), Martinez-Miranda et al. (2015), Pigeon et al. (2014), Taylor et al. (2008), Verbelen et al. (2017), Xiaoli (2013), among others... All these contributions assume a fixed and parametric structural form (e.g. Pigeon et al. (2014) assumes a multivariate skew normal distribution to the claims payments). Such fixed structural forms are not very flexible and are sometimes very difficult to estimate due to complex likelihood functions. Moreover the consideration of detailed feature information with a great data diversity is not always compatible with these rigid approaches.

On this basis, it has become crucial to implement more flexible models. Nowadays, machine learning techniques are very popular in data analytics and offer highly configurable and accurate algorithms that can deal with any sort of structured and unstructured information. Wüthrich (2017) has proposed for the first time a contribution to illustrate how the regression tree techniques can be used for individual claims reserving. However, for pedagogical purposes, he only considered the numbers of payments and not the claims amounts paid. Moreover he assumed that the claims occurrences and reporting process can be described by a homogeneous marked Poisson point process, and, as a consequence these numbers of incurred but not reported (IBNR) claims have been predicted by a Chain Ladder method.

On this basis, we have decided to propose a new non-parametric and flexible approach to estimate separately individual IBNR and RBNS claims reserves that can: i) include the key claim characteristics in order to allow for claims heterogeneity and to take advantage of additional large datasets, ii) capture the specific development pattern of claims, including their occurrence, reporting and cash-flow features, and iii) detect potential trend changes, taking into account possible changes in the product mix, the legal context or the claims processing over time, to avoid potential biases in estimation and forecasting.

Our model is estimated on simulated data and the prediction results are compared with those generated by the Chain Ladder model. When evaluating the performance of our approach, we put emphasis on the the impact of using micro-level information on the variances of the prediction errors. We implement our new approach with an ExtraTrees algorithm but many other powerful machine learning algorithms can easily be adapted (RandomForest, XGBoost,...).

The outline of the paper is as follows. In Section 2 we introduce the claims reserving problem and we theoretically define the several types of outstanding claims reserves that we would like to estimate. In Section 3 we explain how we build the train and test sets on which the Machine Learning algorithms will be respectively trained and used to evaluate the individual outstanding claims reserves. Section 4 reminds the Chain Ladder methodology and gives methodological comparisons with our approach. In Section 5, we provide an explicit numerical example based on simulated data. We first describe the simulation scheme, then we list the individual claims covariates which are used by the ExtraTrees algorithm. We finally present and discuss the numerical results, and we compare them with those obtained by the Chain Ladder approach. In Section 6 we conclude and give an outlook.

2 The problem and the categories of outstanding claims

We consider an insurer who has been running business in one line of insurance and who is subjected to claims reserves computation at some time t. Only outstanding claims for which liability has been assumed prior to time t are therefore relevant for this computation.

We follow the fully time-continuous approach developed by Arjas (1989), Norberg (1993, 1999 and Haastrup and Arjas (1996). We assume that the policies' histories from their underwriting (including in particular the development of the claim if one occurs) are generated by a Position Dependent Marked Poisson Process (PDMPP). We equip the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a filtration $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ which satisfies the usual conditions.

We assume that the insurer can use an external information (if it is relevant) to predict the total outstanding payments, like e.g. the economic environment (unemployment rate, inflation rate, financial distress,...), or weather conditions and natural catastrophes (storm, flood, earthquake, etc.), and so on. We denote by $(E_t)_{t\geq 0}$ the (multivariate) continuous-time process that characterizes this information.

We now associate to each policy the following quantities:

- T_0: the underwriting date (Δ will denote the insured period and the contract will expire at $T_0+\Delta$). Some features/risk factors are observed at T_0 by the insurer and can evolve over time: they will be plugged in the (multivariate) continuous-time process $(F_t)_{t\geq T_0}$. For example, for a life insurance policy, such features could be applicant's current age, applicant's gender (if allowed), height and weight of the applicant, health history, applicant's marital status, applicant's children, if any..., applicant's occupation, applicant's income, applicant's smoking habits or tobacco use)...

- T_1: the occurrence date of the claim ($T_1 = \infty$ if there is no claim). Only one claim is considered during the insured period in this paper for expository purpose (but the algorithm implemented in our computer program can deal with any number of claims).

- T_2: the reporting date (T_2 is considered as to be infinite by the insurance company until the claim is reported). We assume that there exists a maximum delay $\Delta_{\max,r}$ to report the claim once it has occurred, i.e. $T_2 - T_1 < \Delta_{\max,r}$.

- T_3: the settlement date. We assume that there exists a maximum delay $\Delta_{\max,s}$ to settle the claim once it has been declared, i.e. $T_3 - T_2 < \Delta_{\max,s}$.

  During the settlement period the insurance company receives information on the individual claim like the exact cause of accident, the type of accident, the claims assessment and the predictions by claims adjusters, the external expertise, etc. We denote by $(I_t)_{t\geq T_2}$ the (multivariate) continuous-time process that characterizes this information.

  The settlement period is the period within transitional claim payments are done. We denote by $(P_t)_{T_2<t\leq T_3}$ the multivariate cumulated payment process. We let $P_t = 0$ for $T_1 < t \leq T_2$. The payments could be broken down into several components in case of several insurance coverages and if some legal and claims expert fees must be paid.

The mark associated to a policy is therefore given by

$$\mathcal{Z} = \left\{(F_t)_{t\geq T_0}, T_1, T_2, T_3, P_{T_1<t\leq T_{3,}}, I_{T_2<t\leq T_3}\right\}.$$

The insurer's portfolio is characterizes by the collection of points $(T_{0,p}, \mathcal{Z}_p)_{p\geq 1}$ where the $\mathcal{Z}_p$ are in the space of policies' marks, as represented in Figure 1.

2.1 Categories of outstanding claims

First it should be underlined that T_1 must be smaller than $T_0 + \Delta$ for the insurance company to be liable for the claim because the contract must not be terminated at the claim occurrence. Let t denote the reserving date. If $t < T_1 < (T_0 + \Delta)$, the (potential) claim has not yet occurred at date t and therefore it is not considered as an outstanding claim.

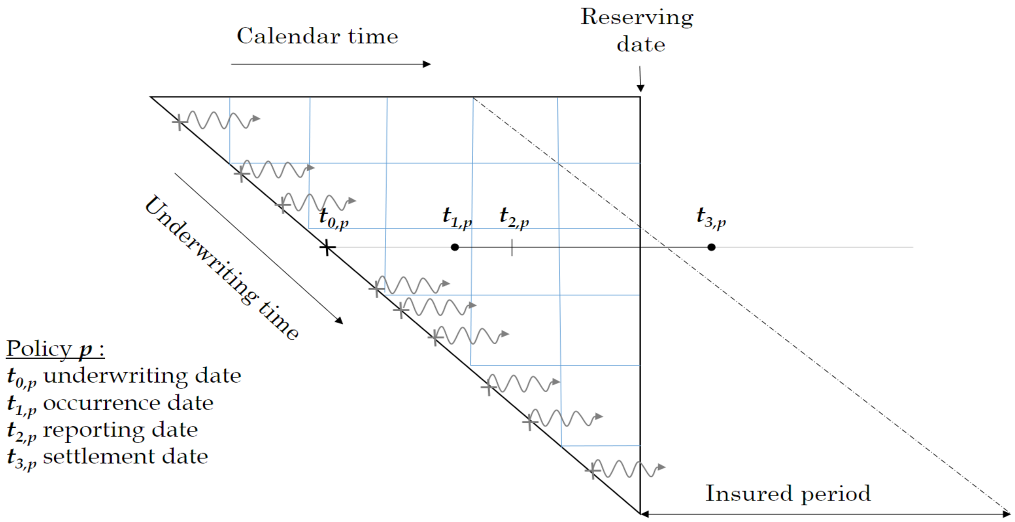

Figure 1: Graphical representation of the PDMPP

Let $a \wedge b$ denote the minimum between the two constants a and b. The claims (that have occurred before t) may be categorized into two categories.

- If $T_1 < t < T_2$, the claim has occurred but it has not yet been reported to the insurance company. This claim is called an Incurred But Not Reported (IBNR) claim. For such a claim we do not have individual claim specific information, but we can use the risk factors as well as external information to predict the likelihood of its occurrence and its cumulated payment. We hence define the $IBNR_t$ for a policy at time t in the following way:

  $$IBNR_t = \mathbb{E}\left[P_{T_3} 1_{T_1 < t \wedge (T_0 + \Delta)} | \mathcal{A}_t\right]$$

  where

  $$\mathcal{A}_t = \{t < T_2, (F_u)_{T_0 \leq u \leq t}, (E_u)_{0 \leq u \leq t}\}.$$

- If $T_2 < t < T_3$, the claim has been reported to the company but the final assessment is still missing. Typically, we are in the situation where more and more information about the individual claim arrives, and the prediction uncertainty in the final assessment decreases. However, this claim is not completely settled, and therefore it is called a Reported But Not Settled (RBNS) claim. We can use the reporting date, the occurrence date, the risk factors, the claim history information as well as external information to predict the cumulated payment from t to the settlement date. We hence define the $RBNS_t$ for a policy at time t for which a claim has been reported in the following way:

  $$RBNS_t = \mathbb{E}\left[P_{T_3} - P_t | \mathcal{B}_t\right]$$

  where

  $$\mathcal{B}_t = \{T_2 < t < T_3, T_1 < T_0 + \Delta, (F_u)_{T_0 \leq u \leq t}, (E_u)_{0 \leq u \leq t}, (I_u)_{T_2 \leq u \leq t}\}.$$

The individual claim reserve is eventually defined as

$$ICR_t = IBNR_t 1_{t < T_2} + RBNS_t 1_{t \geq T_2}.$$

Examples of IBNR and RBNS claims are depicted on Figure 2.

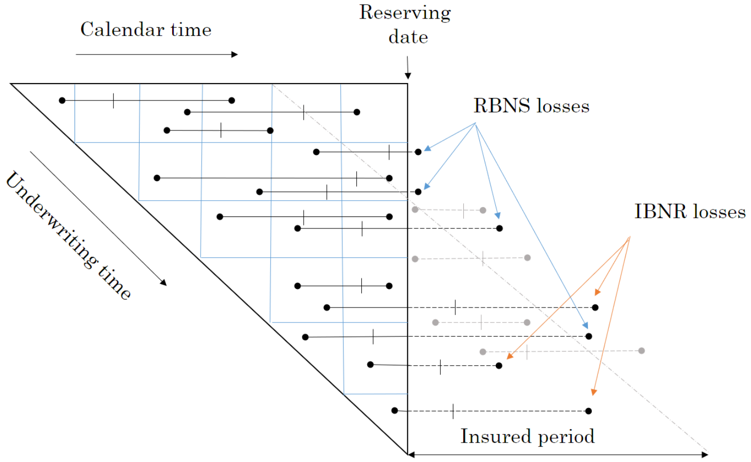

Figure 2: Graphical representation of IBNR and RBNS claims

2.2 Subdivisions of the outstanding claims over development periods

As for the Chain Ladder approach that used several development periods, we will consider a grid of times $t_j = \delta \times j$, where $j \geq 0$ and δ is a fixed timestep. We will assume that the reserving date belongs to this grid and is given by t_i (see Figure 3). We then focus on subdivisions of the outstanding claims over the development periods $(t_{i+j-1}, t_{i+j}]$ for $j \geq 1$.

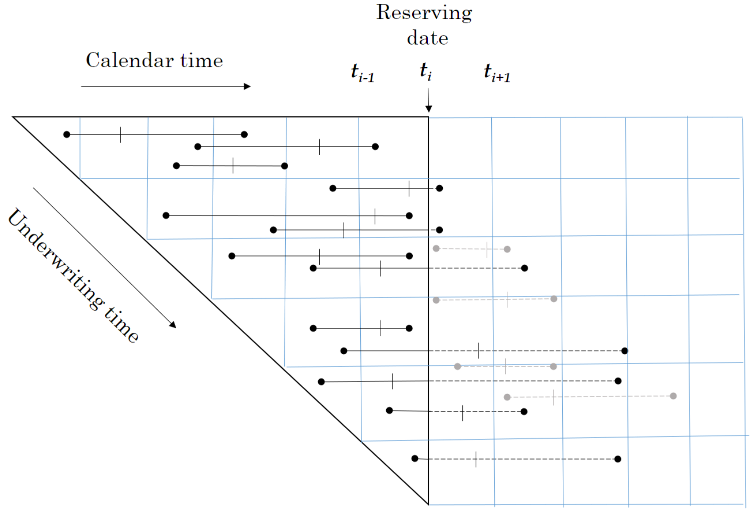

Figure 3: The grid of times $(t_i)_{i\geq 0}$

We define $RBNS_{t_i,j}$, $j = 1, 2, 3, \ldots$, as the expected increase of the payment process between t_{i+j-1} and t_{i+j} given that a claim has been declared and given all the information available on the policyholder and the claim history:

$$RBNS_{t_i,j} = \mathbb{E}\left[P_{t_{i+j}} - P_{t_{i+j-1}} | \mathcal{B}_{t_i}\right]$$

and it follows that

$$RBNS_{t_i} = \sum_{j=1}^{\lceil \Delta_{\max,s}/\delta \rceil} RBNS_{t_i,j}.$$

In the same way, we split $IBNR_{t_i}$ in the following way:

$$IBNR_{t_i,j} = \mathbb{E}\left[(P_{t_{i+j}} - P_{t_{i+j-1}})1_{T_1<t_i\wedge(T_0+\Delta)}|\mathcal{A}_{t_i}\right], \qquad j = 1,2,3,...$$

such that

$$IBNR_{t_i} = \sum_{j=1}^{\lceil(\Delta_{\max,r}+\Delta_{\max,s})/\delta\rceil} IBNR_{t_i,j}.$$

Moreover we break up $IBNR_{t_i,j}$ trough a frequency/severity formula:

$$IBNR_{t_i,j} := IBNR\_freq_{t_i,j} \times IBNR\_loss_{t_i,j}$$

where

$$IBNR\_freq_{t_i,j} = \mathbb{E}\left[1_{(P_{t_{i+j}}-P_{t_{i+j-1}})1_{T_1<t_i\wedge(T_0+\Delta)}>0}|\mathcal{A}_{t_i}\right]$$

and

$$IBNR\_loss_{t_i,j} = \mathbb{E}[(P_{t_{i+j}} - P_{t_{i+j-1}})1_{T_1<t_i\wedge(T_0+\Delta)}|\mathcal{A}_{t_i}, (P_{t_{i+j}} - P_{t_{i+j-1}})1_{T_1<t_i\wedge(T_0+\Delta)} > 0].$$

The quantites $RBNS_{t_i,j}$, $IBNR\_loss_{t_i,j}$, resp. $IBNR\_freq_{t_i,j}$, for $j = 1,2,3,...$, will be estimated using Machine Learning algorithms that will be trained with the quadratic loss function, resp. logistic loss, on specific train tests. We explain the building of these data sets in the next section.

3 Database building for the Machine Learning approach and predictions

3.1 Test sets construction

Let us first recall that the reserving date is denoted by t_i. The sets of policies for which $RBNS_{t_i}$ and $IBNR_{t_i}$ have to be evaluated are then given respectively by

$$\mathcal{P}_{te,RBNS} = \{p : T_{2,p} \leq t_i < T_{3,p}, T_{1,p} < T_{0,p} + \Delta\}$$

and

$$\mathcal{P}_{te,IBNR} = \{p : t_i < T_{2,p} \wedge (T_{0,p} + \Delta + \Delta_{\max,r} + \Delta_{\max,s})\}.$$

To estimate $RBNS_{t_i,j}$, $j = 1,2,3,...$ for a policy p in $\mathcal{P}_{te,RBNS}$, we will use the following vector of variables

$$\left(T_{0,p}, t_i - T_{0,p}, F_{t_i,p}, E_{T_{0,p}}, E_{T_{1,p}}, E_{T_{2,p}}, I_{t_i,p}\right).$$

This vector contains the underwriting date, the exposure of the policy from the underwriting date to the reserving date, the risk factors associated to the policy valued at t_i, the external information observed at the underwriting date, the occurrence date and the reporting date, and the claim history up to t_i.

To estimate $IBNR\_freq_{t_i,j}$ and $IBNR\_loss_{t_i,j}$, $j = 1,2,3,...$ for a policy p in $\mathcal{P}_{te,IBNR}$, we will only use the underwriting date, the exposure of the policy from the underwriting date to the reserving date, the risk factors associated to the policy valued at t_i, the external information observed at the underwriting date,

$$\left(T_{0,p}, t_i - T_{0,p}, F_{t_i,p}, E_{T_{0,p}}\right)$$

since the claim has not yet been reported to the insurance company (if ones occurs).

The matrices

$$\begin{aligned}
\mathrm{X}_{te,RBNS} &= \left(T_{0,p}, t_i - T_{0,p}, F_{t_i,p}, E_{T_{0,p}}, E_{T_{1,p}}, E_{T_{2,p}}, I_{t_i,p}\right)_{p \in \mathcal{P}_{te,RBNS}} \\
\mathrm{X}_{te,IBNR} &= \left(T_{0,p}, t_i - T_{0,p}, F_{t_i,p}, E_{T_{0,p}}\right)_{p \in \mathcal{P}_{te,IBNR}}
\end{aligned}$$

are referred to as the X.TEST sets.

In Figure 4, X.TEST is built by using (co-)variables of policies whose histories begin in the red triangle. Figure 4 illustrates the case of predictions for $j = 1$ and Y.TEST (represented as a green rectangle) then contains the increases of payments (or the indicator function of a positive increase of payments) between t_i and t_{i+1} for all policies in $\mathcal{P}_{te,RBNS}$ or in $\mathcal{P}_{te,IBNR}$.

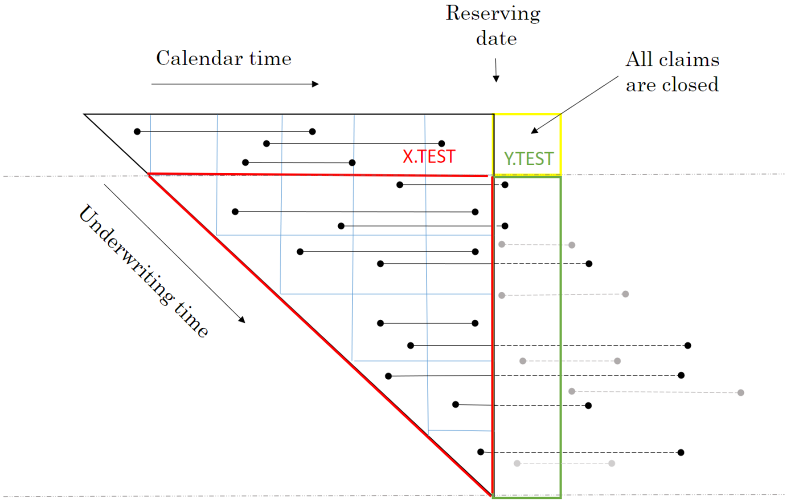

Figure 4: X.TEST and Y.TEST for $j = 1$

3.2 Train sets construction

For each development period j, we consider several prediction models that will be trained on different train sets. We therefore begin by defining three subsets of policies for each development period j and each model k:

- for $RBNS_{t_i,j}$
$$\mathcal{P}^{(j,k)}_{tr,RBNS} = \{p : T_{2,p} \leq t_{i-j-k+1} < T_{3,p}, T_{1,p} < T_{0,p} + \Delta\},$$

- for $IBNR\_freq_{t_i,j}$
$$\mathcal{P}^{(j,k)}_{tr,IBNR\_freq} = \{p : t_{i-j-k+1} < T_{2,p} \wedge (T_{0,p} + \Delta + \Delta_{\max,r} + \Delta_{\max,s})\},$$

- for $IBNR\_loss_{t_i,j}$
$$\mathcal{P}^{(j,k)}_{tr,IBNR\_loss} = \{p : t_{i-j-k+1} < T_{2,p}, (P_{t_{i-k+1}} - P_{t_{i-k}})1_{T_{1,p} < t_{i-j-k+1} \wedge (T_{0,p} + \Delta)} > 0\}.$$

We associate to $\mathcal{P}^{(j,k)}_{tr,RBNS}$, a X.TRAIN set defined by

$$\mathrm{X}^{(j,k)}_{tr,RBNS} = \left(T_{0,p}, t_{i-j-k+1} - T_{0,p}, F_{t_{i-j-k+1},p}, E_{T_{0,p}}, E_{T_{1,p}}, E_{T_{2,p}}, I_{t_{i-j-k+1},p}\right)_{p \in \mathcal{P}^{(j,k)}_{tr,RBNS}}$$

and a Y.TRAIN set defined by

$$\mathrm{Y}^{(j,k)}_{tr,RBNS} = \left(P_{t_{i-k+1},p} - P_{t_{i-k},p}\right)_{p\in\mathcal{P}^{(j,k)}_{tr,RBNS}}.$$

We associate to $\mathcal{P}^{(j,k)}_{tr,IBNR\_freq}$, a X.TRAIN set defined by

$$\mathrm{X}^{(j,k)}_{tr,IBNR\_freq} = \left(T_{0,p}, t_{i-j-k+1} - T_{0,p}, F_{t_{i-j-k+1},p}, E_{T_{0,p}}\right)_{p\in\mathcal{P}^{(j,k)}_{tr,IBNR\_freq}}$$

and a Y.TRAIN set defined by

$$\mathrm{Y}^{(j,k)}_{tr,IBNR\_freq} = \left(1_{(P_{t_{i-k+1,p}} - P_{t_{i-k,p}})1_{T_{1,p}<t_{i-j-k+1}\wedge(T_{0,p}+\Delta)}>0}\right)_{p\in\mathcal{P}^{(j,k)}_{tr,IBNR\_freq}}.$$

We associate to $\mathcal{P}^{(j,k)}_{tr,IBNR\_loss}$, a X.TRAIN set defined by

$$\mathrm{X}^{(j,k)}_{tr,IBNR\_loss} = \left(T_{0,p}, t_{i-j-k+1} - T_{0,p}, F_{t_{i-j-k+1},p}, E_{T_{0,p}}\right)_{p\in\mathcal{P}^{(j,k)}_{tr,IBNR\_loss}}$$

and a Y.TRAIN set defined by

$$\mathrm{Y}^{(j,k)}_{tr,IBNR\_loss} = \left(P_{t_{i-k+1,p}} - P_{t_{i-k,p}}\right)_{p\in\mathcal{P}^{(j,k)}_{tr,IBNR\_loss}}.$$

Figure 5 (resp. 6, 7) illustrates the case $j = 1$, $k = 1$ (resp. $j = 1$, $k = 2$ and $j = 2$, $k = 1$).

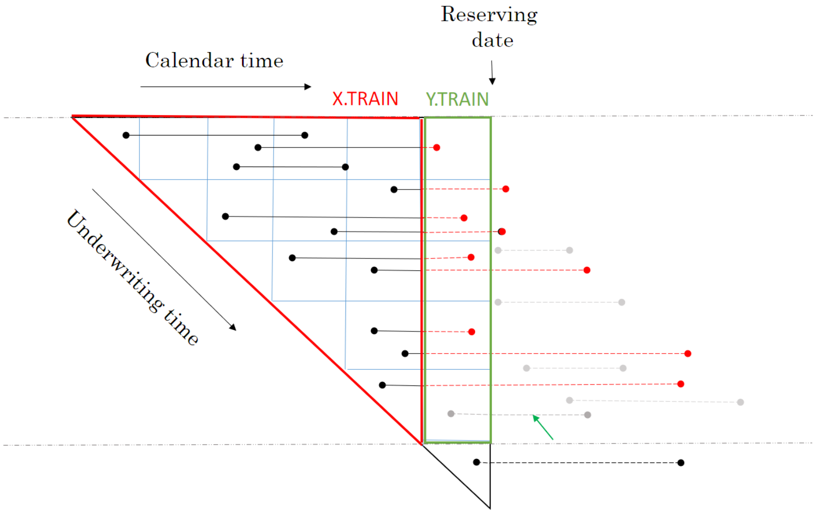

Figure 5: X.TRAIN and Y. TRAIN for $j = 1$, $k = 1$

3.3 Final claim reserves prediction

We will denote the vectors of predictions for development period j and prediction model k as

$$\begin{aligned}
\hat{\mathrm{Y}}^{(j,k)}_{te,RBNS} &= (\widehat{RBNS}^{(k)}_{t_i,j,p})_{p\in\mathcal{P}_{te,RBNS}}\\
\hat{\mathrm{Y}}^{(j,k)}_{te,IBNR\_freq} &= (\widehat{IBNR}\_freq^{(k)}_{t_i,j,p})_{p\in\mathcal{P}_{te,IBNR}}\\
\hat{\mathrm{Y}}^{(j,k)}_{te,IBNR\_loss} &= (\widehat{IBNR}\_loss^{(k)}_{t_i,j,p})_{p\in\mathcal{P}_{te,IBNR}}
\end{aligned}$$

where k denotes the prediction model associated to the k-th train sets.

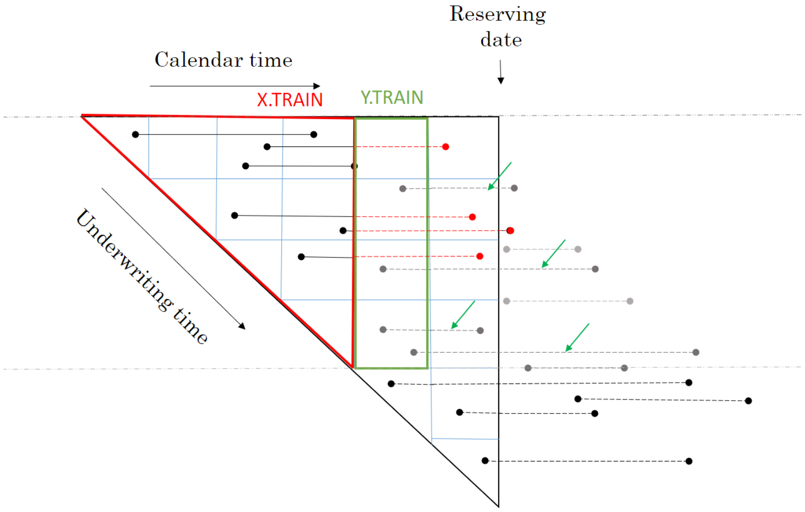

Figure 6: X.TRAIN and Y. TRAIN for $j = 1$, $k = 2$

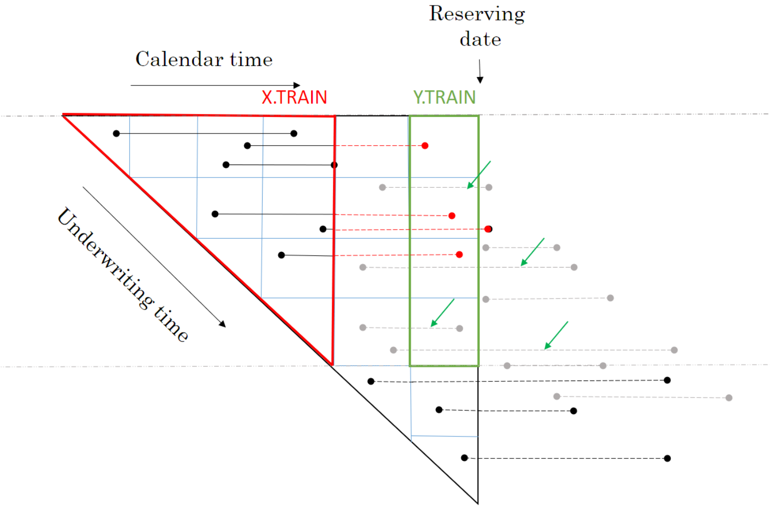

Figure 7: X.TRAIN and Y. TRAIN for $j = 2$, $k = 1$

For a policy p in $\mathcal{P}_{te,RBNS}$, we average predictions over the different models considered:

$$\widehat{RBNS}_{t_i,j,p} = \sum_k p_{j,k}\left(t_{i-k+1}, E_{t_{i-k+1}}\right) \widehat{RBNS}^{(k)}_{t_i,j,p}$$

where $\left(p_{j,k}\left(t_{i-k+1}, E_{t_{i-k+1}}\right)\right)_k$ are sets of positive weights such that $\sum_k p_{j,k}\left(t_{i-k+1}, E_{t_{i-k+1}}\right) = 1$ for each $j = 1, 2, \ldots$.

In the same way, for p in $\mathcal{P}_{te,IBNR}$, we average predictions over the different models considered:

$$\begin{aligned}
\widehat{IBNR}\_freq_{t_i,j,p} &= \sum_k q_{j,k}\left(t_{i-k+1}, E_{t_{i-k+1}}\right) \widehat{IBNR}\_freq^{(k)}_{t_i,j,p} \\
\widehat{IBNR}\_loss_{t_i,j,p} &= \sum_k r_{j,k}\left(t_{i-k+1}, E_{t_{i-k+1}}\right) \widehat{IBNR}\_loss^{(k)}_{t_i,j,p}
\end{aligned}$$

where $\left(q_{j,k}\left(t_{i-k+1}, E_{t_{i-k+1}}\right)\right)_k$ and $\left(r_{j,k}\left(t_{i-k+1}, E_{t_{i-k+1}}\right)\right)_k$ are sets of positive weights such that $\sum_k q_{j,k}\left(t_{i-k+1}, E_{t_{i-k+1}}\right) = 1$ and $\sum_k r_{j,k}\left(t_{i-k+1}, E_{t_{i-k+1}}\right) = 1$ for each $j = 1, 2, \ldots$. We do not discuss the choices of these weights in this paper.

We denote by $ICR_{t_i,p}$ the claim reserve, at date t_i, for a policy p. The estimation of this quantity is computed by summing IBNR and RBNS reserves for p:

$$\widehat{ICR}_{t_i,p} = \widehat{IBNR}_{t_i,p} 1_{t_i < T_{2,p}} + \widehat{RBNS}_{t_i,p} 1_{t_i \geq T_{2,p}}$$

where

$$\widehat{IBNR}_{t_i,p} = \sum_{j=1}^{\lceil (\Delta_{\max,r} + \Delta_{\max,s})/\delta \rceil} \widehat{IBNR\_freq}_{t_i,j,p} \widehat{IBNR\_loss}_{t_i,j,p}$$

and

$$\widehat{RBNS}_{t_i,p} = \sum_{j=1}^{\lceil \Delta_{\max,s}/\delta \rceil} \widehat{RBNS}_{t_i,j,p}.$$

We can finally compute the global claims reserve of the portolio by summing all the reserves predictions:

$$\sum_{p \in \mathcal{P}_{te,RBNS} \cup \mathcal{P}_{te,IBNR}} \widehat{ICR}_{t_i,p}$$

4 Database building for the Chain Ladder approach and predictions

In this section we present the well-known Chain Ladder technique. The first step is to aggregate claims data in a development triangle organized by origin period (occurrence time) and development period. Since we have considered until now the underwriting time as the second axis in our graphical representation (see Figure 1), we must project the claim history lines on this axis such that it now depicts the occurrence time (see Figure 8).

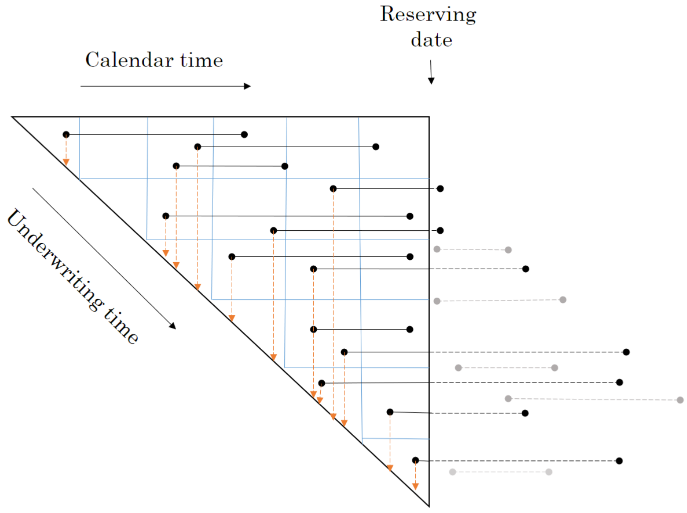

Figure 8: From underwriting time to occurrence time

The second step of the Chain Ladder technique is to compute the development factors for $j = 1, .., J$ where $J = \lceil (\Delta + \Delta_{\max,r} + \Delta_{\max,s})/\delta \rceil$. The set of policies to compute the j-th development factor at time t_i is characterized by:

$$\mathcal{P}_{tr,CL}^{(j|i)} = \{p : T_{2,p} \leq t_{i-j}\}.$$

For a policy p, let $T^{(\delta)}_{1,p} = \inf_{t_j \geq T_{1,p}} t_j$ be the smallest t_j greater than $T_{1,p}$, i.e. the smallest time on the time grid greater than the occurrence date of the claim of the policy. Then the j-th development factor is computed as

$$\hat{F}_{j|i} = \frac{\sum_{p\in\mathcal{P}^{(j|i)}_{tr,CL}} P_{T^{(\delta)}_{1,p}+\delta j,p}}{\sum_{p\in\mathcal{P}^{(j|i)}_{tr,CL}} P_{T^{(\delta)}_{1,p}+\delta(j-1),p}}.$$

Figure 9 (resp. 10) illustrates the computation for the first (resp second) development factor. The numerator of $\hat{F}_{1|i}$ (resp. $\hat{F}_{2|i}$) is the sum of the payments from the occurrence time to $T^{(\delta)}_{1,p} + \delta$ (resp. $T^{(\delta)}_{1,p} + 2\delta$) over $\mathcal{P}^{(1|i)}_{tr,CL}$ (resp. $\mathcal{P}^{(2|i)}_{tr,CL}$) (green squares). The denominator is the sum of the payments from the occurrence time to $T^{(\delta)}_{1,p}$ (resp. $T^{(\delta)}_{1,p} + \delta$) over $\mathcal{P}^{(1|i)}_{tr,CL}$ (resp $\mathcal{P}^{(2|i)}_{tr,CL}$) (red squares).

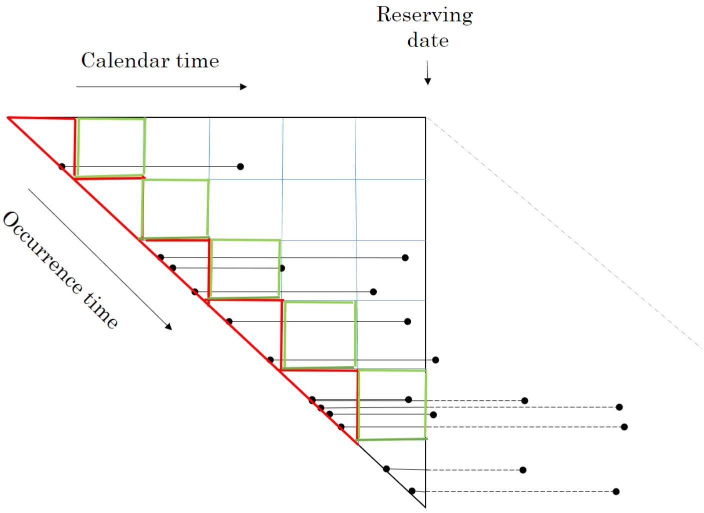

Figure 9: 1-st development period for CL method

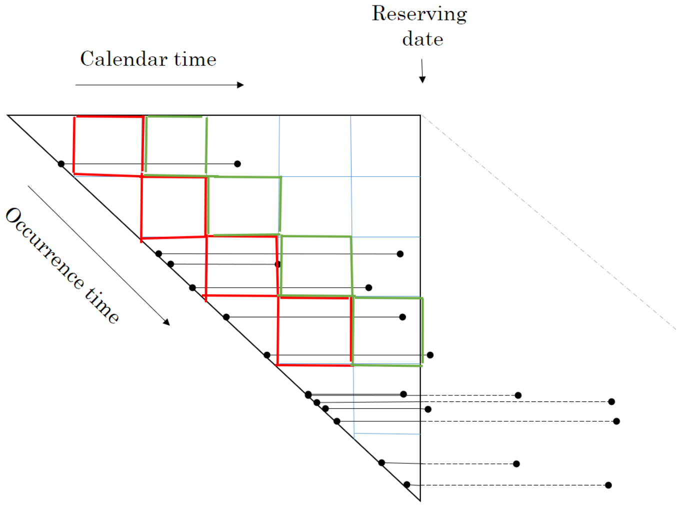

Figure 10: 2-nd development period for CL method

The second step of the Chain Ladder technique is to compute the predictions of the sum the payments step by step by using recursively the previously estimated development factors. The set of policies for which the predictions used the j-th development factor at time t_i is given by:

$$\mathcal{P}^{(i,j)}_{te,CL} = \{p : t_{i-j} \leq T_{2,p} \leq t_i\}\}.$$

Figure 10 (resp. 11) illustrates the computation for the predictions that use the first (resp second) development factor.

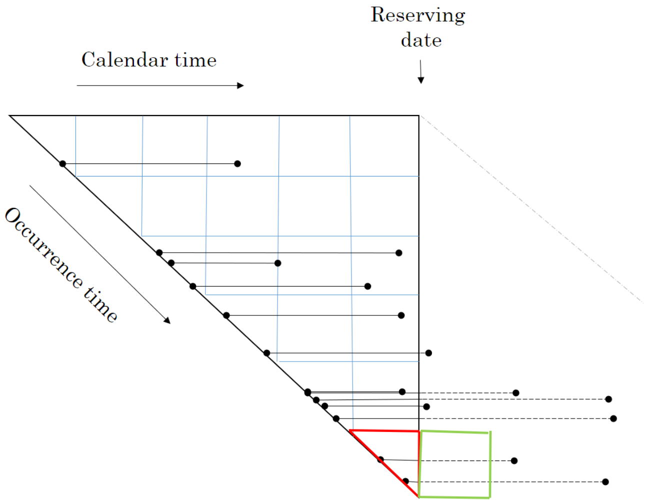

Figure 11: 1-st development period prediction

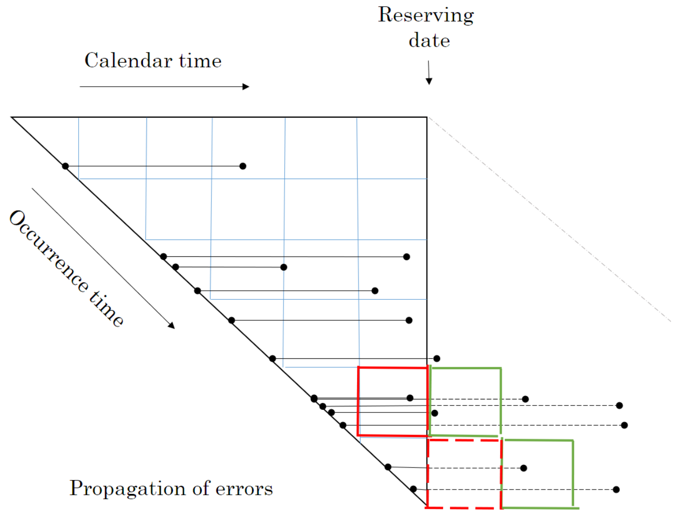

Figure 12: 2-nd development period prediction

The final claims reserve prediction for the whole portfolio with the Chain Ladder method is therefore given by:

$$\sum_{j=1}^{J}\left(\sum_{p\in\mathcal{P}^{(i,j)}_{te,CL}} P_{T^{(\delta)}_{1,p}+\delta(j-1),p}\right)\times\left(\prod_{k=1}^{J-j}\hat{F}_{k+(j-1)|i}-1\right).$$

5 A case study with mobile phone insurance

In this section, we illustrate our method on a synthetic insurance portfolio. Most people now have a mobile phone and they have become an essential part of everyday life. Losing or damaging their mobile phone can be a total nightmare and they may turn down to insurance coverage. For this case study, we consider a mobile phone insurance that covers the devices in the event of theft, breakage or oxidation. The insurance company provides cover for an insured period of one year and for a range of four brands and up to four models by brand with three policy types available: “breakage”, “breakage and oxidation” and “breakage, oxidation and theft”.

To illustrate our method and compare it to the Chain Ladder approach, we will first consider a central scenario with very simplified portfolio assumptions. We will then study several scenarii around this central scenario for which we will modify one of these simplified assumptions.

The underwriting period will be from the first day of 2016 to the last day of 2017 (after this date, the portfolio starts its run-off period). We will perform the reserving exercises for each end of month from September 16 to July 17 and will compute IBNR and RBNS claims reserves with our Machine Learning method, and the global claims reserves with the Chain Ladder method. We will compare these computed reserves with the true reserves and study forecast bias and variance for both methods. Note that the timestep we fixed for this illustration is a uniform monthly timestep (actually we assume that each month has 30 days and that a year has then 360 days). We discuss what happens when a true calendar is used in an additional scenario.

5.1 Technical assumptions of the central scenario

The central scenario aims at being very simple in order to benchmark the Chain Ladder and Machine Learning methods when the portfolio is in an almost stationary environment. We now present its assumptions.

The underwriting Poisson point process has a constant intensity $\lambda_0(t) = 700$ (daily basis), i.e. the insurance company sells roughly 500,000 insurance policies over the two underwriting years. Each policy is generated independently from the other policies.

The three available policy types are assumed to be sold according to the following constant over time distribution

	Probability
Breakage	0.25
Breakage + oxidation	0.45
Breakage + oxidation + theft	0.30

Table 1: Policy types distribution

The policyholders are assumed to insure mobile phone whose brand is distributed according to the following constant over time distribution (note that the basis price of a mobile phone also depends on its brand)

	Probability	Basis price
Brand 1	0.45	600
Brand 2	0.30	550
Brand 3	0.15	300
Brand 4	0.10	150

Table 2: Brand distribution

The price of a mobile phone is equal to a basis price multiplied by a factor that depends on the model type according to the following table (Table 3 also provides the model type distribution)

Model type	Multiplicative factor	Probability
0	1	0.05
1	1.15	0.10
2	1.15^2	0.35
3	1.15^3	0.50

Table 3: Model type distribution and multiplicative factors with respect to the model types

The policy type, the mobile phone brand and the mobile phone model type are independent from each other.

We assume that claim frequencies are generated through a competing model between risks, with constant incidence hazard rates that can depend on the mobile phone model type (but neither on the brand nor the model type), in the following way

	Yearly incidence rate
Breakage	0.15
Oxidation	0.05
Theft	$0.05 \times (1+$ model type)

Table 4: Incidence hazard rates

We assume that the claim amount is a percentage of the buying price of the mobile phone and the distribution of this percentage is a Beta distribution with parameters depending on the claim type

	α	β
Breakage	2	5
Oxidation	5	3
Theft	5	0.5

Table 5: Parameters of the Beta distributions

Once a claim occurs, we assume that the reporting delay hazard rate is given by

$$\lambda_1(t+T_0) = \frac{t^{\alpha-1}(1-t)^{\beta-1}}{\int_t^1 u^{\alpha-1}(1-u)^{\beta-1}\,du}, \qquad 0 < t < 1,$$

with $\alpha = 0.4$, $\beta = 10$ (the mean reporting delay is roughly 14 days). Claims are settled in one payment and the payment delay hazard rate is given by

$$\lambda(t+T_1) = \frac{((t-d)/m)^{\alpha-1}(1-(t-d)/m)^{\beta-1}}{m\int_{(t-d)/m}^1 u^{\alpha-1}(1-u)^{\beta-1}\,du}, \qquad d < t < m+d,$$

with $\alpha = 7$, $\beta = 7$, and with $m = 40/360$, $d = 10/360$ (the mean payment delay is roughly 30 days). These hazard rates do not depend neither on the brand, the model, the coverage type, nor the occurrence date.

5.2 Selected features

The main strength of our Machine Learning approach is to take into account any information about the policy (for IBNR and RBNS claims), and the claim history (for RBNS claims). No external information is considered here. For this case study, we use the following policy-related features (embedded in F_t):

- phone brand,
- phone price,
- phone model type,
- coverage type (“Breakage”, “Breakage and Oxidation”, “Breakage and Oxidation and Theft”).

We moreover use the following claim-related features (embedded in I_t):

- type of damage (“Breakage”, “Oxidation”, “Theft”),
- reporting delay (in days)
- number of days since the claim has been declared.

In case of transitional payments, additional features could be the amount paid and the number of payments made until the reserving date.

5.3 Algorithm used for prediction

Once the train sets have been built (see Section 3), we can train a prediction algorithm. For this case study, we decided to illustrate our methodology by using the ExtraTrees algorithm. This algorithm builds an ensemble of unpruned regression trees according to a classical top-down procedure. Its two main differences with other tree-based ensemble methods are:

1. it splits nodes by choosing cut-points fully at random;
2. it uses the whole learning sample (rather than a bootstrap replica) to grow the trees.

The predictions of the trees are aggregated to yield the final prediction, by majority vote in classification problems and arithmetic average in regression problems. From the bias-variance point of view, the rationale behind the Extra-Trees method is that the explicit randomization of the cutpoint and attribute combined with ensemble averaging should be able to reduce variance more strongly than the weaker randomization schemes used by other methods. The usage of the full original learning sample rather than bootstrap replicas is motivated in order to minimize bias. From the computational point of view, the complexity of the tree growing procedure is like most other tree growing procedures.

Note however that the algorithm used for prediction is not the keystone of our method since we can use other Machine Learning algorithm to make predictions such as RandomForests, XGBoost (which can make better predictions than ExtraTrees). We do not think that Neural Networks (NN) are really adapted here for the following reasons: first they perform better on unstructured data (which is not the case here), second for huge architectures, one may need GPU (Graphical Processing Unit) to fit the NN, which could be a computational bottleneck in some cases.

5.4 Results in the central scenario

We performed 40 simulations of the portfolio over the two years. For each end of month from September 16 to May 18, we computed the outstanding claims reserves with both predictive methods (40 times). Note that for the ML method we only used one model ($k = 1$).

We first averaged the true values of the sum of the cumulated payments called here Ground Truth (GT), the Machine Learning predictions (ML), and the Chain Ladder predictions (CL). These values are represented by solid lines on Figure 13.

We can observe that both methods are almost unbiased except from January 17 to March 17 for the ML method and to July 17 for the CL method. January 17 is actually the first month where the portfolio size and its exposure to risk has begun to be stabilized (i.e. underwritings and terminations are balanced). Both algorithm need time to adjust their predictions, but the ML method does it more quickly.

The vertical dashed segments provide the 95% confidence intervals for each prediction (they are based on the assumption of a Gaussian distribution, i.e. their lengths are approximately equal to four times the standard errors computed over the 40 simulations). We can see that the standard errors of the ML predictions are really lower than the CL predictions (around 4 or 5 times smaller). So one first conclusion here is that the ML method outperforms the CL method in the variance-of-reserves point of view because it takes into account the heterogeneity between claims to provide more accurate predictions.

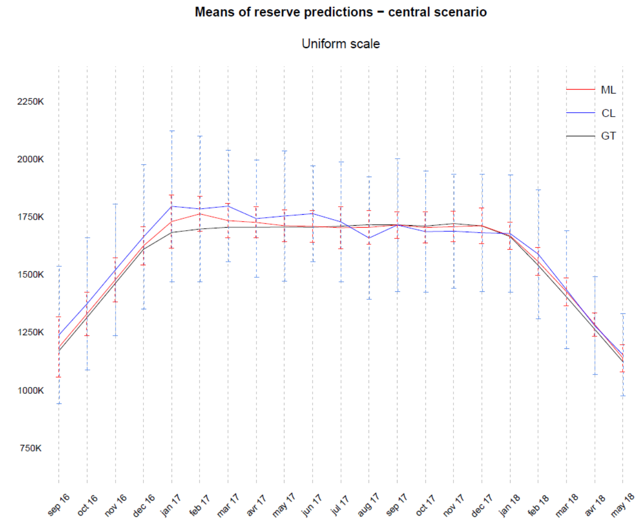

Figure 13: Means of reserve predictions - central scenario

It should be underlined that the variances of the CL predictions are very closed to the variances computed with the Mack's formulas (Mack (1999)) in this case study.

On the other hand, we plot on Figure 14 the averages of the ML RBNS reserves and of the ML IBNR reserves, as well as the averages of their respective true values. One can see that the ML RBNS reserves are really precise since they are unbiased and their standard errors are not very large. The ML IBNR reserves are however a bit biased when the portfolio size starts to be stabilized. This is due to the fact that when the algorithm learns exposures as of the first day of 2017, it has never seen that exposures are actually bounded since the insured period is equal to one year (this means that policies underwrited at the beginning of January 16 are no more exposed to the whole month of January 17, and the algorithm has never learnt such a behavior). This leads to a punctual mis-prediction.

5.5 The other scenarii

We now study several scenarii around this central scenario for which we will modify one (and only one) of the assumptions. The assumptions that we changed are the following ones:

- monthly scale instead of uniform scale,
- time-dependent underwriting rate,
- increase/decrease of 10% of the payment delay,
- arrivals of new phone models (more expensive) at the beginning of the year 2017,
- increase of 40% of the claim rate from mid-December 2016 to mid-January 2017.

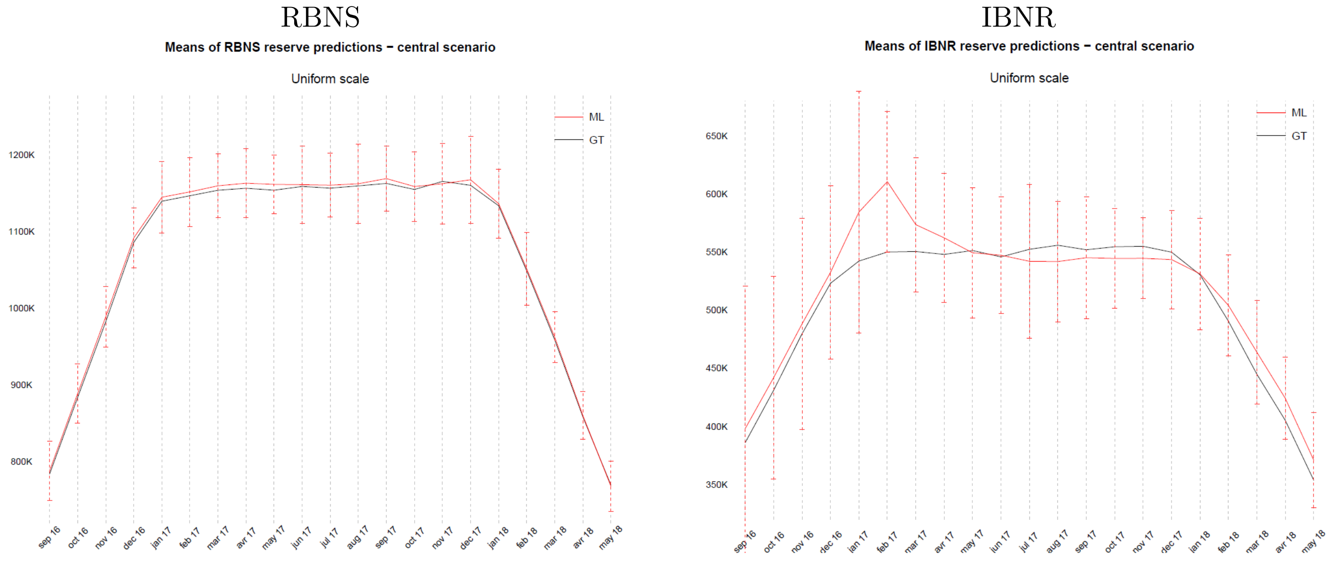

Figure 14: Means of the IBNR and RBNS reserve predictions

5.5.1 Monthly scale

We first wanted to study the consequences of using a true calendar rather than using a uniform timestep and dividing a whole year in twelve equal parts (the timestep size on the central scenario was 30 days). On Figure 15, we plot the average values of the relative errors between the true values of the sum of the increases of payments and respectively the ML predictions and the CL predictions. We notice a strong month-dependent bias for the CL method, but a moderate month-dependent bias for the ML method. These bias are consequences of the fact that months do not have the same length, leading to a mis-learning of the true exposure as well as a mis-prediction for the true future exposure (although this last issue could be fixed by taking into account the true number of days of the month for which the predictions are made). We also see that the ML method still has a lower variance. These are interesting result from the insurer's point of view because the ML method can adapt quite well when the timestep length is not constant over time.

5.5.2 Time-dependent underwriting rate

In this scenario, we let the underwriting rate change over time: we multiplied the constant hazard rate by the density of a Beta distribution with $\alpha = 1.6$ and $\beta = 2.5$, scaled over the two years.

Figure 16 shows the reserves evolution for the Ground Truth (GT), the ML method and the CL method. We can see that both methods perform well, and we still observe a better standard error for the ML method as we observed for the central scenario.

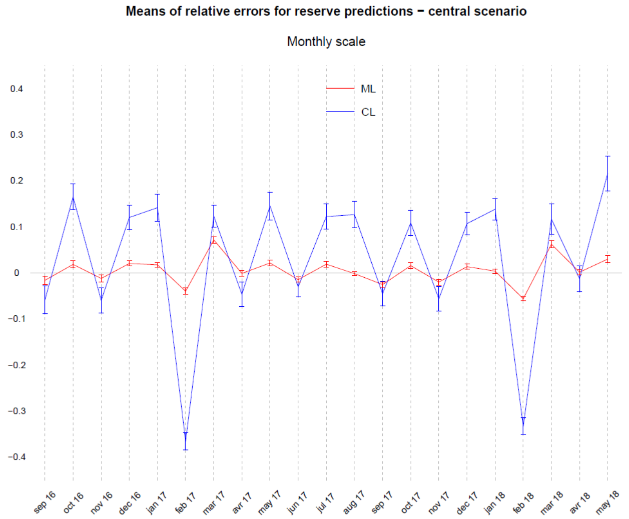

Figure 15: Means errors of reserve predictions

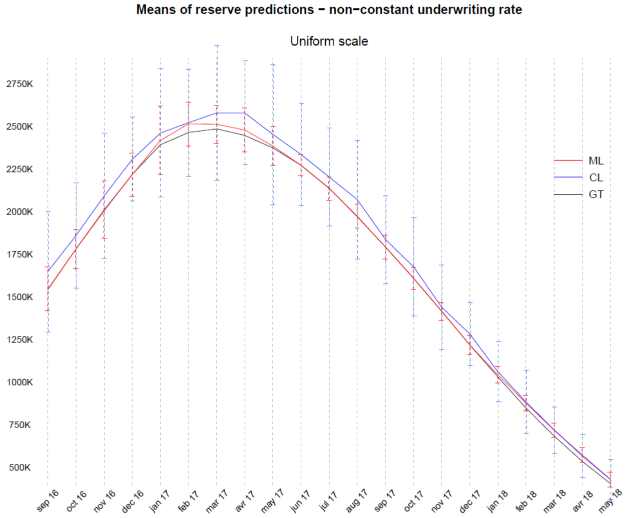

Figure 16: Means of reserve predictions - time-dependent underwriting rate

5.5.3 Payment delay modification

In this scenario, we reduce (resp. increase) all the payments delays by 10% starting from the first day of 2017 until the total run-off of the portfolio, This means that a claim that was settled in 10 days will now be settled in 9 days (resp. 11 days). We can see from Figure 17 and 18 that this change of assumptions leads to a huge effect on the CL reserves, but not on the ML reserves. This is due to a combination of some reasons explained here:

- An increase of the number payments will be observed for the month of January 17. This increase will immediately affects the computation of the first development factor of the CL method. And this factor will have a strong impact on the estimated values of the ultimate cumulated payments for the claims that have occurred and have been reported during the last month. This leads to an important over-estimation of the reserves if the computation of the CL method is done mechanically (which is fortunately not the case in practice, because such a result would alarm the actuary in charge of the reserving exercise). Since the ML prediction is not based on a multiplicative form (that leads to the propagation of the errors), this shock on the payment delay will be softer when using the ML method.

- This phenomenon is amplified by the fact that the average payment delay is around 30 days on the central scenario. This leads to a huge amount of claims which are paid in January compared to December.

In practice, it is the strong difference between the ML and CL computations that should be highlighted. It seems to provide an indicator of important changes whose source could be looked for in the payment delays.

On the other hand, we observe the opposite behavior if the payment delay is lengthened by 10% leading to a huge under-estimation of the reserves.

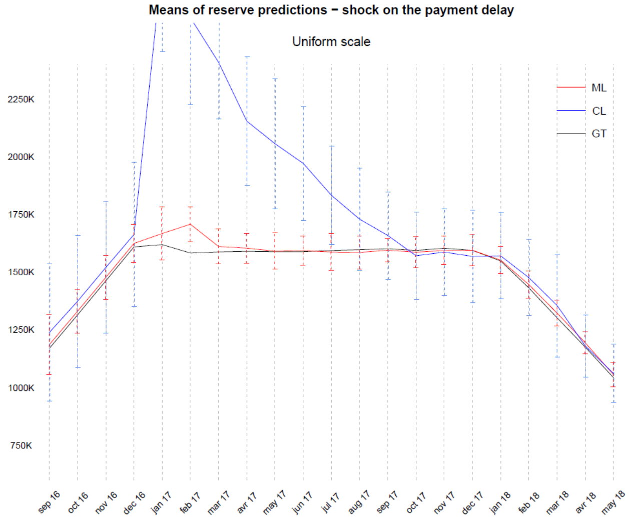

Figure 17: Means of reserve predictions - negative shock on the payment delay

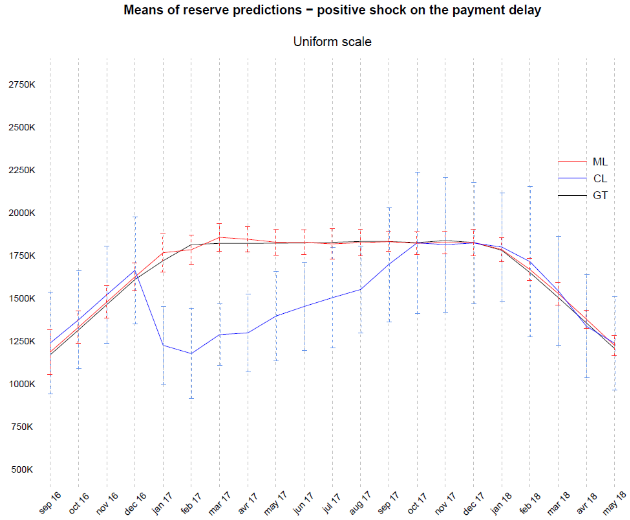

Figure 18: Means of reserve predictions - positive shock on the payment delay

5.5.4 Arrival of new phone models

In this scenario, we decided to change our portfolio structure by letting each brand introduce its Model 3 successively from October 2016. Brand 1 will release its model 3 at the end of October 16, Brand 2 at the end of November 16, etc... The distribution of the model types before (resp. after) the release of Model 3 is given in Table 6.

Model type	Prob dist before	Prob dist after
0	0.15	0.05
1	0.35	0.10
2	0.50	0.35
3	0	0.50

Table 6: Model type distributions

Moreover, it has been assumed that Model 3 of Brand 1 has a claim hazard rate twice as much as the other Model 3 of Brand 2, 3 and 4. Remind moreover that theft damage depends on phone model, so by introducing new phone models over time, we increase the global claim rate of our portfolio. Figure 19 shows that both methods actually handle correctly this change of structure in the portfolio (we also see the same differences between the CL and ML reserves variances):

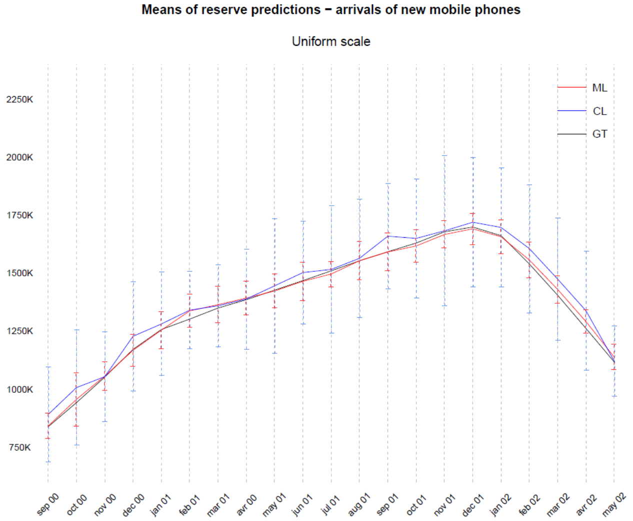

Figure 19: Means of reserve predictions - arrival of new phone models

5.5.5 Temporary claim rate shock

In this scenario, we set a raise of 40% on the claim hazard rate between mid-December 2016 and mid-January 2017. Then the hazard rate just comes back to its normal value. We take from Figure 20 three comments. First, both methods under-estimate reserves for the reserving month of December 2016. Second, for the reserving month of January 2017, the CL method over-estimates the reserves whereas ML method is correct. Third, between February 2017 and June 2017, reserves are over-estimated with the ML method whereas the CL method is back to normal.

We explain these observations in the following way:

1. The CL reserves are computed using the 'stationary' development factor for reserving month December 2016. This factor is indeed under-estimated because we set a sharp raise of hazard rate between mid-December 2016 and the end of December 2016. This leads to an under-estimation of claim reserves in December 16 for January 2017, and this error is then propagated although correct development factors have been used.

   The ML reserves are computed by learning the claim rate observed in December for reserving month December 2016. Hence, the algorithm learns that there is a sharp raise in mid-December. But those additional claims which are happening in December are, for many of them at the end of December, declared in January. This explains the fact that the ML method is able to learn that there is a hazard rate rise, but the fact that those claims are not known yet by the insurer leads to a slight under-estimation of the real reserve (IBNR).

2. The CL reserves are computed using the new development factor, which takes into account the claim rate rise of December. This development factor is over-estimated compared to the GT because we know that the claim rate rise is over by the end of January. Thus, we observe in January the additional claims which occurred in the end of December and in January. Hence, the reserves for the first month are computed using an over-esimated development factor, and on a higher amount of paid claims. This leads to a wide over estimation of reserves.

   The ML reserves are computed by learning the claim rate observed in January, and since it as the same as December, the estimation is just fine.

3. The CL reserves are computed using an over-estimated development factor, but this bias tends to zero over time since the claim rate is back to normal since mid-January 2017.

   The ML reserves are computed using one sub-model (i.e. $k = 1$). Thus, the large claim rate of December and January is learnt several times to compute long-horizon reserves over time. This leads to a bias in reserves calculation, which tends to zero over time, but slower than the CL method. This behavior can be overcome by using more than one sub-model (i.e. $k > 1$), because the bias is time-punctual. This means that we will have a few number of biased sub-models, and averaging the predictions of each sub-model would vanish the bias when k grows.

This scenario illustrates a limit of our method, when we use only one sub-model to compute the ML reserves.

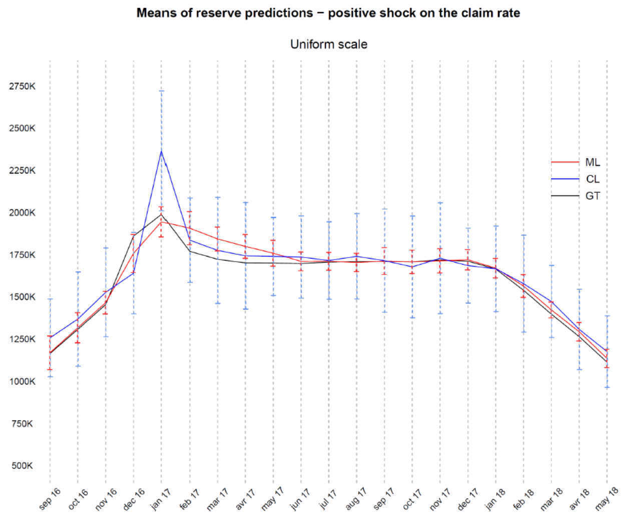

Figure 20: Means of reserve predictions - temporary claim rate shock

6 Summary and outlook

We have proposed a new non-parametric approach for individual claims reserving. Our model is fully flexible and allow to consider (almost) any kind of feature information. As a result we obtain IBNR and RBNS claims reserves for individual policies integrating all available relevant feature information.

In our case study where we used a Machine Learning algorithm known as ExtraTrees algorithm, we observe that the method provides almost unbiased estimators of the claims reserves with very small standard deviations (actually four to five times smaller than the Mack Chain Ladder standard deviation). Moreover the ML estimators are more responsive to any changes in the development patterns of claims including occurrence, reporting, cost modifications,... than the CL estimators based on aggregated loss data.

Acknowledgments

We would like to thank the actuarial department of BNP Paribas Cardif for fruitful discussions about the case study, Philippe Baudier, Pierre de Sahb and Sebastien Conort from the Data Lab of BNP Paribas Cardif for a lot of important discussions about the methodology, the way to build the train and test sets and for checking a large part of our computer code.

References

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

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

[3] Badescu, A.L., Sheldon L.X., Tang, D. (2016) A marked Cox model for IBNR Claims: Model and Theory. Insurance: Mathematics and Economics, 69, 29-37.

[4] Hiabu, M., Margraf, C., Martínez-Miranda, M. D., Nielsen, J.P., (2016). The link between classical reserving and granular reserving through double chain ladder and its extensions. British Actuarial Journal, 21/1, 97-116.

[5] Haastrup, S., Arjas, E. (1996) Claims reserving in continuous time: a nonparametric Bayesian approach. ASTIN Bulletin, 26, 139-164.

[6] Larsen, C. (2007) An individual claims reserving models. ASTIN Bullletin, 37, 113-132.

[7] Martinez-Miranda, M.D., Nielsen, J.P., Verrall, R.J., Wuthrich, M.V. (2015). The link between classical reserving and granular reserving through double chain ladder and its extensions. Scandinavian Actuarial Journal, 5, 383-405.

[8] Mack, T. (1999) The standard error of chain ladder reserve estimates: recursive calculation and the inclusion of a tail factor. ASTIN Bulletin, 29, 361-366.

[9] Norberg, R. (1993) Prediction of outstanding liabilities in non-life insurance. ASTIN Bulletin, 23, 95-115.

[10] Norberg, R. (1999). Prediction of outstanding liabilities ii. Model variations and extensions. ASTIN Bulletin, 29, 5-25.

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

[12] Taylor, G., McGuire, G., Sullivan, J. (2008). Individual claim loss reserving conditioned by case estimates. Annals of Actuarial Science, 3, 215-256.

[13] Verbelen, R., Antonio, K., Claeskens, G. and Crèvecoeur, J. (2017). Predicting daily IBNR claim counts using a regression approach for the occurrence of claims and their reporting delay. Working paper.

[14] Wüthrich, M.V. (2017) Machine Learning in individual claims reserving. Working paper.

[15] Xiaoli, J. (2013) Micro-level loss reserving models with applications in workers compensation insurance. Working paper.