NON-PARAMETRIC INDIVIDUAL CLAIM RESERVING IN INSURANCE

MAXIMILIEN BAUDRY* & CHRISTIAN ROBERT*

*DAMI Chair & Université Lyon 1 - ISFA

http://chaire-dami.fr/
ACKNOWLEDGEMENTS TO

- The actuarial department of
- And its Data Lab
- And more specifically to Philippe Baudier, Pierre de Sahb and Sebastien Conort
RESERVING 2.0: CHANGING THE WAY OF THINKING ABOUT OUTSTANDING CLAIMS

- Computing reserves by leveraging the use of massive data and artificial intelligence in insurance.
- Taking advantage of Machine Learning algorithms’ predictive power.
- Exploiting the richness of the insurer’s data (any sort of structured and unstructured data).
- Deploying this method on any kind of portfolio (life, non-life).
The current reserving practice consists, in most cases, in using methods based on claim development triangles.

Triangles are organized by origin period (occurrence most of the time or underwriting otherwise) and development period.

Deterministic and stochastic unpaid claim reserving models based on triangles (e.g. Chain Ladder method, Bornhuetter-Ferguson method) have had a great success to manage reserve risk for a variety of lines of business...
... but such models suffer from underlying strong assumptions and give rise to several issues:

- Need for tail factors that may induce over parameterization risk.

- Propagations of errors through the development factors, huge estimation error for the latest development periods.

- Instability in ultimate claims for recent arrival years, uncertainty about the ability to properly capture the pattern of claim development.

- Lack of robustness and need for treatments of outliers.

- Can not separate assessment of IBNR and RBNS claims.
Because triangle-based methods use aggregated data, they don’t use any information on the policy, the claim nor the policy holder.

Natural overcome: Individual claim reserving.

- First approaches: Structural and Parametric models.
- Recent approaches (very few): Non-parametric & Machine Learning models. (Our approach).
WORKS ON MACHINE LEARNING (ML) MODELS IN RESERVING

- Wüthrich (2017)
  - Prediction on the number of RBNS payments with ML (CART algorithm).
  - IBNR prediction with Chain Ladder (CL) method.

- ASTIN ICDML working group (2017)
  - No use of explanatory variables/features (only payments).
  - ‘Cascade’ predictions, i.e. chaining predictions, inducing propagation of errors just like Chain Ladder does.
  - No IBNR prediction at all. The reporting delay is forced to 0.
  - Questionable data simulation: each payment (cumulative) is the product of the ultimate with a noisy coefficient.
WHAT WE PROPOSE

- A new non-parametric and flexible approach to estimate individual claims reserves which handles key effects, such as:

  - Including the key claim characteristics (i.e., explanatory variables) to allow for claims heterogeneity and to take advantage of additional large datasets.

  - Learning the specific development pattern of claims, including their occurrence, reporting and cash-flow features, and detecting potential trend changes.

  - Taking into account possible deviations in the product mix, the legal context or the claims processing over time, to avoid potential biases in estimation and forecasting.

  - Implementing separate and consistent treatments of IBNR and RBNS claims.
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 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 (random forest, gradient boosting,...).
Our prediction flow is able to:

- Compute **RBNS** and **IBNR reserves separately**.

- Run the full reserving process as of **any date** as required (backtesting...).

- Compute reserves along **any time granularity** (monthly, quarterly, semestrial, annual).

- Learn from **any subsample of historical data** specified by user.

- Learn from **any sub-space of features** specified by the user.
THE PROBLEM & OUR APPROACH

We associate with each policy the following quantities:

- \( T_0 \): the underwriting date (\( \Delta \) is the insured period and the contract will expire at \( T_0 + \Delta \)).

- Some features/risk factors are known at \( T_0 \) and may evolve over time: \( (F_t)_{t \geq T_0} \)

Example: For a life insurance policy: 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 possible during the insured period (but it can be easily generalized).
$T_2$: the reporting date.

- We assume that there exists a maximum delay $\Delta_{\text{max},r}$ to report the claims once it has occurred, i.e. $T_2 - T_1 < \Delta_{\text{max},r}$

$T_3$: the settlement date.

- During the settlement period the insurance company receive information on the individual claim like exact cause of accident, type of accident, location of accident, line-of-business and contracts involved, claims assessment and predictions by claims adjusters, payments already done, external expertise, etc.

- We denote this information by $(I_t)_{t \geq T_2}$

- We assume that there exists a maximum delay $\Delta_{\text{max},s}$ to settle the claims once it has been declared, i.e. $T_3 - T_2 < \Delta_{\text{max},s}$
Payment cash flows

- The payments are broken down into $q$ several components: $q - 1$ insurance coverages and the legal and claims expert fees (if any).

- We denote by $(P_t)_{T_2 < t \leq T_3}$ the cumulated payment process. We let $P_t = 0$ for $T_1 < t \leq T_2$

External information may be used to predict reserves. We denote by $E_t$ that information.
THE PROBLEM & OUR APPROACH

Underwriting time

Calendar time

Reserving date

Insured period

Policy $p$:
- $t_{0,p}$: underwriting date
- $t_{1,p}$: occurrence date
- $t_{2,p}$: reporting date
- $t_{3,p}$: settlement date
CATEGORIES OF OUTSTANDING CLAIMS

Note that if $T_1 > T_0 + \Delta$, the insurance company is not liable for this particular claim with the actual insurance policy because the contract is already terminated at claim occurrence.

1. $t < T_1$. There is no outstanding claim.

2. $T_1 < t < T_2$, The insurance claim has occurred but it has not yet been reported to the insurance company.

These claims are called **Incurred But Not Reported (IBNR)** claims. For such claims:

- We do not have individual claim specific information.
- But we can use any external information $E_t$

$$IBNR_t = \mathbb{E} \left[ P_{T_3} 1_{T_1 < t \wedge (T_0 + \Delta)} \bigl| t < T_2, (F_u)_{T_0 \leq u \leq t}, (E_u)_{0 \leq u \leq t} \right]$$
3. $T_2 < t < T_3$. These claims are reported at 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, these claims are not completely settled, yet, and therefore they are called **Reported But Not Settled (RBNS)** claims:

$$RBNS_t = \mathbb{E} [P_{T_3} - P_t | T_2 < t < T_3, T_1 < T_0 + \Delta, (F_u)_{T_0 \leq u \leq t}, (E_u)_{0 \leq u \leq t}, (I_u)_{T_2 \leq u \leq t}]$$

$\Rightarrow$ The individual claims reserve is therefore:

$$ICR_t = IBNR_{t \leq T_2} + RBNS_{t \geq T_2}$$
SCHEME OF OUR FRAMEWORK

CATEGORIES OF OUTSTANDING CLAIMS

- Underwriting time
- Calendar time
- Reserving date
- Insured period
- RBNS losses
- IBNR losses
Let $\delta$ be a fixed timestep and derive a \textbf{grid of times} $t_i = \delta \times i$, $i \geq 0$, for which the insurance company wants to evaluate its liabilities.
We split $RBNS_{t_i}$ in the following way: for $j = 1, 2, 3, ...$ we define the expected increase of the payments between $t_{i+j-1}$ and $t_{i+j}$ given that a claim has been declared such that

$$RBNS_{t_i,j} = \mathbb{E} [P_{t_{i+j}} - P_{t_{i+j-1}} | T_2 < t_i < T_3, T_1 < T_0 + \Delta, (F_{u})_{T_0 \leq u \leq t_i}, (E_{u})_{u \leq t_i}, (I_{u})_{T_2 \leq u \leq t_i}]$$

such that

$$RBNS_{t_i} = \sum_{\text{claim development period } j} RBNS_{t_i,j}.$$
We split $IBNR_{t_i}$ in the following way: for $j = 1, 2, 3, \ldots$ we define the expected increase of the payments between $t_{i+j-1}$ and $t_{i+j}$ given that a claim has been declared such that

$$IBNR_{t_i,j} = \mathbb{E} \left[ (P_{t_{i+j}} - P_{t_{i+j-1}}) 1_{T_1 < t_i \wedge (T_0 + \Delta)} | t_i < T_2, (F_u)_{T_0 \leq u \leq t_i}, (E_u)_{T_0 \leq u \leq t_i} \right]$$

such that

$$IBNR_{t_i} = \sum_{\text{claim development period } j} IBNR_{t_i,j}.$$ 

Moreover, we write $IBNR_{t_i,j}$ in a frequency/severity formula:

$$IBNR_{t_i,j} := IBNR_{freq} t_i,j \times IBNR_{loss} t_i,j$$
DATABASE BUILDING FOR THE MACHINE LEARNING APPROACH AND PREDICTIONS

- Notations:
  - Reserving date: $t_i$
  - Development period: $j$
  - Model: $k$
Case $j = 1$: 1st development period - Test sets

- **Underwriting time**
- **Calendar time**
- **Reserving date**
- All claims are closed

Diagram illustrating the relationship between underwriting time, calendar time, reserving dates, and the closure of claims in a test set.
- **Case** $j = 1, \ k = 1$: 1st development period - Train sets - 1st model
**Case** $j = 1$, $k = 2$: 1\textsuperscript{st} development period - Train sets - 2\textsuperscript{nd} model
Case $j = 2$: 2nd development period - Test sets
**Case** $j = 2, \ k = 1$: 2\textsuperscript{nd} development period - Train sets - 1\textsuperscript{st} model
Final individual claims reserve predictions:

\[
\hat{ICR}_{ti,p} = \hat{IBNR}_{ti,p}1_{t_i < T_2,p} + \hat{RBNS}_{ti,p}1_{t_i \geq T_2,p}
\]

Final claims reserve prediction:

\[
\sum_{p \in RBNS \cup IBNR} \hat{ICR}_{ti,p}
\]
Notations:

- Reserving date: $t_i$
- Development period: $j$
- From underwriting time to occurrence time:
Example: 1st development period
Example: 1\textsuperscript{st} development period prediction
Example: 2\textsuperscript{nd} development period
Example: 2\textsuperscript{nd} development period prediction
The Extra-Trees algorithm builds an ensemble of unpruned regression trees according to the classical top-down procedure.

Its two main differences with other tree-based ensemble methods are that:

- it splits nodes by choosing cut-points fully at random.
- 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.

Remarks:

1. The choice of the prediction algorithm is not the main issue here as long as we use a non-parametric algorithm such as RandomForest, ExtraTree, XGBoost etc...

2. We could use a GLM model, but this needs work on interaction detection, feature engineering and additional assumptions (such as the distribution and link function) which must be tested.
A FEW NUMBERS ON PROGRAMMING

- We used Python to develop our method.
- It runs in 2 steps:
  - Build train and test sets.
  - Predict payment increases for each claim development period.
- ~2k rows of code.
- On the following example, runs in a few minutes for a relatively big portfolio on a huge hardware (for one seed):
  - 80 vCPU.
  - 128 Go RAM.
A CASE STUDY WITH MOBILE PHONE INSURANCE

Note that every characteristic in our portfolio is customizable. This allows us to work with very different kind of portfolio.

We consider a mobile phone insurance which covers the devices for the following damages:

- Theft
- Breakage
- Oxidation

The insurance company provides cover for a range of four phone brands and up to four models by brand with three policy types available for an insured period of one year:

- “Breakage”
- “Breakage and oxidation”
- “Breakage, oxidation and theft”
CASE STUDY - MOBILE PHONE INSURANCE

CENTRAL SCENARIO

For the first generation of policies which will be sold from 2016/01/01 to 2017/12/31, we consider the following central scenario:

- The underwriting Poisson point process has a constant intensity $\lambda_{0,t} = 250,000$ (in yearly unit), i.e. the insurance sells roughly 500,000 policies over the two years.

- Claim severity distributions: Beta distributions

<table>
<thead>
<tr>
<th>Damage type</th>
<th>$\alpha$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breakage</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Oxidation</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Theft</td>
<td>5</td>
<td>.5</td>
</tr>
</tbody>
</table>
CASE STUDY - MOBILE PHONE INSURANCE

• Stationary distribution of the coverage types

<table>
<thead>
<tr>
<th>Coverage type</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breakage</td>
<td>.25</td>
</tr>
<tr>
<td>Breakage + Oxidation</td>
<td>.45</td>
</tr>
<tr>
<td>Breakage + Oxidation + Theft</td>
<td>.30</td>
</tr>
</tbody>
</table>

• Stationary distribution of the brand types

<table>
<thead>
<tr>
<th>Phone brand</th>
<th>Proportion</th>
<th>Base price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brand 1</td>
<td>.45</td>
<td>600</td>
</tr>
<tr>
<td>Brand 2</td>
<td>.30</td>
<td>550</td>
</tr>
<tr>
<td>Brand 3</td>
<td>.15</td>
<td>300</td>
</tr>
<tr>
<td>Brand 4</td>
<td>.10</td>
<td>150</td>
</tr>
</tbody>
</table>
• Multiplicative link between the model and its price

<table>
<thead>
<tr>
<th>Model type</th>
<th>Multiplicative factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1.15</td>
</tr>
<tr>
<td>2</td>
<td>1.15^2</td>
</tr>
<tr>
<td>3</td>
<td>1.15^3</td>
</tr>
</tbody>
</table>

• Claim frequencies assumptions

<table>
<thead>
<tr>
<th>Coverage type</th>
<th>Yearly incidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breakage</td>
<td>.15</td>
</tr>
<tr>
<td>Oxidation</td>
<td>.05</td>
</tr>
<tr>
<td>Theft</td>
<td>.05 x model type</td>
</tr>
</tbody>
</table>
• Reporting delay hazard rate $\alpha = 0.4, \ \beta = 10$:

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

• Payment delay hazard rate $\alpha = 7, \ \beta = 7, \ m = \frac{40}{365}, \ d = \frac{10}{365}$:

$$
\lambda_{2,t+T_1} = \frac{((t - d)/m)^{\alpha-1} (1 - (t - d)/m)^{\beta-1}}{m \int_{(t-d)/m}^1 u^{\alpha-1} (1 - u)^{\beta-1} \, du}, \quad d < t < m + d.
$$

These hazard rates don’t depend neither on the brand, the model, the coverage type nor the occurrence date.
CENTRAL SCENARIO – DESCRIPTIVE STATISTICS
SELECTED FEATURES

- **Features related to the contract:**
  - Brand of the mobile phone.
  - Price of the mobile phone.
  - Type of coverage ("breakage", "breakage and oxidation" and "breakage, oxidation and theft").
  - Underwriting date.

- **Features related to the history of the contract:**
  - Number of days since the underwriting date and exposure.
  - Indicator function whether a claim has been declared or not.
  - Type of damage (breakage, oxydation, theft).
  - Number of days since the claim has been declared.
Means of reserve predictions – central scenario

Uniform scale

- ML
- CL
- GT
Means of relative errors for reserve predictions – central scenario

Uniform scale

- ML
- CL
Feature ranking for IBNR Frequency model

- Time since underwriting
- Exposure
- Coverage type
- Phone brand
- Phone price
- Phone model
OTHER SCENARIOS

- We start from **central scenario**, then **add one of the following:**
  - Monthly scale instead of uniform scale.
  - Time-dependent underwriting rate.
  - Decrease of 10% of the payment delay since January 1\textsuperscript{st} 2017.
  - Increase of 10% of the payment delay since January 1\textsuperscript{st} 2017.
  - Arrivals of new and more expensive mobile phones at the end of the year 2016.
  - Increase of 40% of the claim rate from December 15\textsuperscript{th} 2016 to January 15\textsuperscript{th} 2017.
MONTHLY SCALE SCENARIO
Mean errors of reserve predictions – central scenario

Monthly scale

ML
CL
NON-CONSTANT UNDERWRITING RATE SCENARIO
Means of relative errors for reserve predictions – non-constant underwriting rate

Uniform scale

- ML
- CL
SHOCK ON PAYMENT DELAY: 10% DECREASE
Means of reserve predictions – shock on the payment delay

Uniform scale

- ML
- CL
- GT
Means of relative errors for reserve predictions – shock on the payment delay
SHOCK ON PAYMENT DELAY: 10% INCREASE
Average payment delay by payment months

Average payment delay vs. payment months from 2016-04 to 2019-04.
Means of reserve predictions – positive shock on the payment delay

Uniform scale
Means of relative errors for reserve predictions – positive shock on the payment delay

Uniform scale

- ML
- CL
ARRIVAL OF NEW PHONE MODELS
Means of reserve predictions – arrivals of new mobile phones

Uniform scale

ML
CL
GT
Means of relative errors for reserve predictions – arrivals of new mobile phones

Uniform scale

- ML
- CL
SHOCK ON CLAIM RATE FOR A SHORT PERIOD
Means of reserve predictions – positive shock on the claim rate

Uniform scale

- ML
- CL
- GT
Means of relative errors for reserve predictions – positive shock on the claim rate

Uniform scale

- ML
- CL
CONCLUSION

- We have proposed a **new non-parametric approach for individual claims reserving** using a machine learning algorithm known as Extra-Trees algorithm.

- 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.

- The method provides almost **unbiased estimators of the claims reserves with very small standard deviations** in our simulation study (almost four times smaller than the Mack chain-ladder standard deviation!).

- Our Machine Learning estimators are **more responsive to any changes in the development patterns of claims** including occurrence, reporting, cost modifications,... than the chain-ladder estimator based on aggregated loss data.