Non parametric individual claims reserving

Maximilien BAUDRY & Christian Y. ROBERT
DAMI Chair & Université Lyon 1 - ISFA

Workshop on « Data science in Finance and Insurance »
ISBA (UCL), Friday, September 15, 2017
Acknowledgements to:

- the Actuarial Department of

- and its

- and more specifically to Philippe Baudier, Pierre de Sahb, Sebastien Conort!
1 Introduction and motivation

• The current reserving practice consists, in most cases, in using methods based on claim development triangles for projections as well as for capital requirement calculations.

• The triangles are organised 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 form 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,

- existence of a chain-ladder bias,

- lack of robustness and need for treatments of outliers,

- can not include calendar year effects,

- potential different results between projections based on paid losses or incurred losses,

- can not separate assessment of IBNR and RBNS claims,
⇒ These limits are consequences to a loss of information when aggregating the original individual claim data details (time of occurrence, reporting delay, time and amounts of payments,...).

• Recent developments in data collection, storage and analysis techniques implies that a proper individual claims modelling is now accessible.

• Therefore, recent research strongly promotes claims reserving on individual claims data, see, for instance, Antonio and Plat (2014), Arjas (1989), Hiabu et al. (2016), Jessen et al. (2011), Norberg (1993, 1999), Martinez-Miranda et al. (2015), Pigeon et al. (2014), Taylor et al. (2008), Wüthrich (2017), Xiaoli (2013), among others...

... but all contributions that are based on individual claims data, except Wüthrich (2017), 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) proposes for the first a contribution to illustrate how the **regression tree** techniques can be used for individual claims reserving. However, for pedagogical purposes,
  - he only considers the numbers of payments and not the claims amounts paid,
  - he assumes 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 individual IBNR and RBNS claims reserves that can account for 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,

- capturing the specific development pattern of claims, including their occurrence, reporting and cash-flow features, and detecting 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,

- 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,...).
2 The problem and 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 generalised).
- $T_2$: 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 **multivariate cumulated payment process**. We let \( P_t = 0 \) for \( T_1 < t \leq T_2 \).

- The **mark associated to the policy** is

\[
Z = \{(F_t)_{t \geq T_0, T_1, T_2, T_3, P_{T_1 < t \leq T_3}, I_{T_2 < t \leq T_3}}\}
\]

⇒ The **insurer’s portfolio is represented by a collection of points** \((T_{0,p}, Z_p)_{p \geq 1}\) where \( Z_p \) are in the space of policies’ marks.
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 external information (denoted by $E_t$)

$$IBNR_t = \mathbb{E} \left[ P_{T_3} \mathbf{1}_{T_1 < t \wedge (T_0 + \Delta)} | t < T_2, (F_u)_{T_0 \leq u \leq t}, (E_u)_{0 \leq u \leq t} \right]$$
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:

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

⇒ **The individual claims reserve is therefore**

\[
ICR_t = IBNR_t 1_{t<T_2} + RBNS_t 1_{t \geq T_2}
\]
Subdivisions of outstanding claims

Let $\delta$ be a fixed timestep and derive a 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.

\[
RBNS_{t_i,j} = \mathbb{E} \left[ P_{t_{i+j}} - P_{t_{i+j-1}} \mid T_2 < t_i < T_3, T_1 < T_0 + \Delta, (F_u)_{T_0 \leq u \leq t_i}, (E_u)_{u \leq t_i}, (I_u)_{T_2 \leq u \leq t_i} \right].
\]

such that

\[
RBNS_{t_i} = \sum_{j=1}^{\lceil \Delta_{\text{max},s}/\delta \rceil} RBNS_{t_i,j}.
\]
We split $IBNR_{t_i}$ in the following way: for $j = 1, 2, 3, ...$

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

such that

$$IBNR_{t_i} = \left\lceil \frac{(\Delta_{\text{max},r} + \Delta_{\text{max},s})}{\delta} \right\rceil \sum_{j=1} IBNR_{t_i,j}.$$
Moreover we write $IBNR_{t_i,j}$ in a frequency/severity formula:

$$IBNR_{t_i,j} := IBNR_{freq_{t_i,j}} \times IBNR_{loss_{t_i,j}}$$

where

$$IBNR_{freq_{t_i,j}} = \mathbb{E} \left[ 1(P_{t_i+j} - P_{t_i+j-1})1_{T_1 < t_i \wedge (T_0 + \Delta) > 0 | t_i < T_2, (F_u)_{T_0 \leq u \leq t_i}, (E_u)_{T_0 \leq u \leq t_i}} \right]$$

and

$$IBNR_{loss_{t_i,j}} = \mathbb{E} \left[ (P_{t_i+j} - P_{t_i+j-1})1_{T_1 < t_i \wedge (T_0 + \Delta) > 0 | t_i < T_2, (F_u)_{T_0 \leq u \leq t_i}, (E_u)_{T_0 \leq u \leq t_i}, (P_{t_i+j} - P_{t_i+j-1})1_{T_1 < t_i \wedge (T_0 + \Delta) > 0} \right]$$
Database building for the Machine learning approach and predictions

▷ Notations: Reserving date: $t_i$ / development period: $j$ / model: $k$
Case $j = 1$: 1-st development period - Test sets
Case $j = 2$ : 2-nd development period - Test sets
Case $j = 1, k = 1$: 1-st development period - Train sets - 1-st model
Case \( j = 1, k = 2 \) : 1-st development period - Train sets - 2-nd model
Case $j = 2, \ k = 1$ : 2-nd development period - Train sets - 1-st model
Final individual claims reserve predictions:

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

with

\[ \hat{IBNR}_{t_i,p} = \left\lceil \frac{(\Delta_{\text{max},r} + \Delta_{\text{max},s})}{\delta} \right\rceil \sum_{j=1} \hat{IBNR}_{freq_{t_i,j,p}} \hat{IBNR}_{loss_{t_i,j,p}} \]

\[ \hat{RBNS}_{t_i,p} = \left\lceil \frac{\Delta_{\text{max},s}}{\delta} \right\rceil \sum_{j=1} \hat{RBNS}_{t_i,j,p} \]

Final claims reserve prediction:

\[ \sum_{p \in P_{te,RBNS} \cup P_{te,IBNR}} \hat{ICR}_{t_i,p} \]
Extremely randomized trees algorithm

▷ 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.
From the **bias-variance point of view**

- the rationale behind the Extra-Trees method is that the explicit randomization of the cut-point 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.
Table 1  Extra-Trees splitting algorithm (for numerical attributes)

Split_a_node(S)
Input: the local learning subset $S$ corresponding to the node we want to split
Output: a split $[a < a_c]$ or nothing
- If Stop_split($S$) is TRUE then return nothing.
- Otherwise select $K$ attributes $\{a_1, \ldots, a_K\}$ among all non constant (in $S$) candidate attributes;
- Draw $K$ splits $\{s_1, \ldots, s_K\}$, where $s_i = \text{Pick_a_random_split}(S, a_i), \forall i = 1, \ldots, K$;
- Return a split $s_*$ such that $\text{Score}(s_*, S) = \max_{i=1,\ldots,K} \text{Score}(s_i, S)$.

Pick_a_random_split($S, a$)
Inputs: a subset $S$ and an attribute $a$
Output: a split
- Let $a_{\max}^S$ and $a_{\min}^S$ denote the maximal and minimal value of $a$ in $S$;
- Draw a random cut-point $a_c$ uniformly in $[a_{\min}^S, a_{\max}^S]$;
- Return the split $[a < a_c]$.

Stop_split($S$)
Input: a subset $S$
Output: a boolean
- If $|S| < n_{\min}$, then return TRUE;
- If all attributes are constant in $S$, then return TRUE;
- If the output is constant in $S$, then return TRUE;
- Otherwise, return FALSE.
Database building for the chain ladder approach and predictions

▷ Reserving date: $t_i$ / development period: $j$
From underwriting time to occurrence time:
Example: 1-st development period
Example: 1-st development period prediction
Example: 2-nd development period
Example: 2-nd development period prediction
3 A case study with mobile phone insurance

▷ We consider a mobile phone insurance that covers the devices in the event of theft, breakage or oxydation.

▷ The insurance company provides cover for a range of four brands and until four models by brand with three policy types available: “breakage”, “breakage and oxydation” and “breakage, oxydation and theft” and for an insured period of one year.

▷ For the first generation of contracts that 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 contracts over the two years.
- stationary distribution of the coverage types

<table>
<thead>
<tr>
<th>Coverage Type</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breakage</td>
<td>0.25</td>
</tr>
<tr>
<td>Breakage + oxidation</td>
<td>0.45</td>
</tr>
<tr>
<td>Breakage + oxidation + theft</td>
<td>0.30</td>
</tr>
</tbody>
</table>

- stationary distribution of the brand types

<table>
<thead>
<tr>
<th>Brand</th>
<th>Proportion</th>
<th>Basis price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brand 1</td>
<td>0.45</td>
<td>600</td>
</tr>
<tr>
<td>Brand 2</td>
<td>0.30</td>
<td>550</td>
</tr>
<tr>
<td>Brand 3</td>
<td>0.15</td>
<td>300</td>
</tr>
<tr>
<td>Brand 4</td>
<td>0.10</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></th>
<th>Yearly incidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breakage</td>
<td>0.15</td>
</tr>
<tr>
<td>Oxydation</td>
<td>0.05</td>
</tr>
<tr>
<td>Theft</td>
<td>0.05 \times \text{model type}</td>
</tr>
</tbody>
</table>

A competing model between risks is assumed.
- claim amount distributions : beta distribution

<table>
<thead>
<tr>
<th></th>
<th>α</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breakage</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Oxydation</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Theft</td>
<td>5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

- declaration delay intensity : $\alpha = 0.4$, $\beta = 10$,

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

- payment delay intensity : $\alpha = 7$, $\beta = 7$, $m = 40/365$, $d = 10/365$,

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

These intensities don’t depend neither on the brand, the model, the coverage type, nor the occurrence date.
Central scenario - some 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 oxydation" and "breakage, oxydation 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
Means of relative errors for reserve predictions - central scenario

Uniform scale
The other scenarii:

- Monthly scale instead of uniform scale
- Time-dependent underwriting rate
- Decrease of 10% of the payment delay since January 1, 2017
- Increase of 10% of the payment delay since January 1, 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, 2016 to January 15, 2017
Mean errors of reserve predictions – central scenario

Monthly scale

- ML
- CL
Means of relative errors for reserve predictions – non-constant underwriting rate

Uniform scale
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

Uniform scale

Legend:
- ML
- CL
Means of reserve predictions – positive shock on the payment delay
Means of relative errors for reserve predictions – positive shock on the payment delay

Uniform scale

ML
CL
Means of reserve predictions – arrivals of new mobile phones

Uniform scale
Means of relative errors for reserve predictions – arrivals of new mobile phones

Uniform scale

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

Uniform scale

- ML
- CL
4 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 (five times smaller than the Mack chain-ladder standard deviation!).

• 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.
1. **RBNS** - Test step:

- Set of policies:
  \[
  \mathcal{P}_{te, RBNS} = \{ p : T_{2,p} \leq t_i < T_{3,p}, T_{1,p} < T_{0,p} + \Delta \} 
  \]

- \( X_{te, RBNS} \):
  \[
  X_{te, RBNS} = (T_{0,p}, t_i - T_{0,p}, F_{t_i,p}, E_{T_{0,p}}, E_{T_{1,p}}, E_{T_{2,p}}, I_{t_i,p})_{p \in \mathcal{P}_{te, RBNS}}
  \]

- \( Y_{test} \) : predictions
  \[
  \hat{Y}_{te, RBNS}^{(j,k)} = (\overline{RBNS}_{i,j,p}^{(k)})_{p \in \mathcal{P}_{te, RBNS}}
  \]
2. \textit{RBNS} - Train step:

- Set of policies:
  \[ \mathcal{P}_{tr, RBNS}^{(j,k)} = \{ p : T_2,p \leq t_{i-j-k+1} < T_3,p, T_1,p < T_0,p + \Delta \} \]

- \textit{X}.train
  \[ \mathcal{X}_{tr, RBNS}^{(j,k)} = \left( T_0,p, t_{i-j-k+1} - T_0,p, F_{t_{i-j-k+1},p}, E_{T_0,p}, E_{T_1,p}, E_{T_2,p}, I_{t_{i-j-k+1},p} \right)_{p \in \mathcal{P}_{tr, RBNS}^{(j,k)}} \]

- \textit{Y}.train
  \[ \mathcal{Y}_{tr, RBNS}^{(j,k)} = \left( P_{t_{i-k+1},p} - P_{t_{i-k},p} \right)_{p \in \mathcal{P}_{tr, RBNS}^{(j,k)}} \]
3. **IBNR_freq** - Test step:

- Set of policies:

\[ \mathcal{P}_{te,IBNR_freq} = \{ p : t_i < T_{2,p} \land (T_{0,p} + \Delta + \Delta_{max,r} + \Delta_{max,s}) \} \]

- \( X.te,IBNR_freq \)

\[ X_{te,IBNR_freq} = (T_{0,p}, t_i - T_{0,p}, F_{t_i,p}, E_{T_0,p})_{p \in \mathcal{P}_{te,IBNR_freq}} \]

- \( Y.test_freq \) : predictions

\[ \hat{Y}^{(j,k)}_{te,IBNR_freq} = (\overline{IBNR_freq}_{i,j,p})^{(k)}_{p \in \mathcal{P}^{(j)}_{te,IBNR_freq}} \]
4. IBNR_freq - Train step:

- Set of policies:

\[ \mathcal{P}_{tr,IBNR.freq} = \{ p : t_i - j - k + 1 < T_{2,p} \land (T_{0,p} + \Delta + \Delta_{\text{max},r} + \Delta_{\text{max},s}) \} \]

- X.train_freq

\[ X_{tr,IBNR.freq}^{(j,k)} = (T_{0,p}, t_i - j - k + 1 - T_{0,p}, F_{t_i - j - k + 1,p}, E_{T_{0,p}}) \}_{p \in \mathcal{P}_{tr,IBNR.freq}}^{(j,k)} \]

- Y.train_freq

\[ Y_{tr,IBNR.freq}^{(j,k)} = \left( 1(P_{t_i - k + 1,p} - P_{t_i - k,p})1_{T_{1,p} < t_i - j - k + 1, (T_{0,p} + \Delta) > 0} \right) \}_{p \in \mathcal{P}_{tr,IBNR.freq}}^{(j,k)} \]
5. *IBNR_loss* - Test step:

- Set of policies:
  \[ \mathcal{P}_{te,IBNR} = \{ p : t_i < T_{2,p} \land (T_{0,p} + \Delta + \Delta_{\text{max,r}} + \Delta_{\text{max,s}}) \} \]

- *X.test_loss*
  \[ X_{te,IBNR\_loss} = (T_{0,p}, t_i - T_{0,p}, F_{t_i,p}, E_{T_{0,p}})_{p \in \mathcal{P}_{te,IBNR}} \]

- *Y.test_loss*: predictions
  \[ \hat{Y}^{(j,k)}_{te,IBNR\_loss} = (\overline{\text{IBNR\_loss}}_{i,j,p}^{(k)})_{p \in \mathcal{P}_{te,IBNR}} \]
6. \textit{IBNR\_loss} - Train step:

- Set of policies:

\[
\mathcal{P}_{tr,\text{IBNR\_loss}}^{(j,k)} = \{p : t_{i-j-k+1} < T_2.p, (P_{t_{i-k+1}} - P_{t_{i-k}})1_{T_1.p < t_{i-j-k+1} \wedge (T_{0.p} + \Delta) > 0}\}
\]

- \textit{X\_train\_loss}

\[
X_{tr,\text{IBNR\_loss}}^{(j,k)} = (T_{0.p}, t_{i-j-k+1} - T_0.p, F_{t_{i-j-k+1}}, E_{T_0.p})_{p \in \mathcal{P}_{tr,\text{IBNR\_loss}}^{(j,k)}}
\]

- \textit{Y\_train\_loss}

\[
Y_{tr,\text{IBNR\_loss}}^{(j,k)} = (P_{t_{i-k+1}}, P_{t_{i-k}.p})_{p \in \mathcal{P}_{tr,\text{IBNR\_loss}}^{(j,k)}}
\]
Database building for the chain ladder approach and predictions

▷ Reserving date : $t_i$ / development period : $j$

- Set of policies:
  \[ P_{tr,CL}^{(j|i)} = \{ p : T_{1,p} \leq t_{i-j} \}. \]

- $j$-th development factor:
  \[
  \hat{F}_{j|i} = \frac{\sum_{p \in P_{tr,CL}^{(j|i)}} P_{T_{1,p}^{(\delta)} + \delta_j,p}}{\sum_{p \in P_{tr,CL}^{(j|i)}} P_{T_{1,p}^{(\delta)} + \delta(j-1),p}}.
  \]
  with
  \[
  T_{1,p}^{(\delta)} = \inf_{t_j \geq T_{1,p}} t_j.
  \]
Final claims reserve prediction:

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

where

\[
\mathcal{P}_{te,CL}^{(i,j)} = \{ p : t_{i-j} \leq T_{1,p} \leq t_i \}
\]

and

\[
J = \lceil (\Delta + \Delta_{\text{max},r} + \Delta_{\text{max},s})/\delta \rceil.
\]
and

\[
\hat{RBNS}_{t_i,j,p} = \sum_k p_k (t_{i-k+1}, E_{t_{i-k+1}}) \hat{RBNS}_{t_i,j,p}^{(k)}
\]

\[
\hat{IBNR\_freq}_{t_i,j,p} = \sum_k p_k (t_{i-k+1}, E_{t_{i-k+1}}) \hat{IBNR\_freq}_{t_i,j,p}^{(k)}
\]

\[
\hat{IBNR\_loss}_{t_i,j,p} = \sum_k p_k (t_{i-k+1}, E_{t_{i-k+1}}) \hat{IBNR\_loss}_{t_i,j,p}^{(k)}
\]

\((p_k (t_{i-k+1}, E_{t_{i-k+1}}))_k\) is a set of positive weights such that

\[
\sum_k p_k (t_{i-k+1}, E_{t_{i-k+1}}) = 1
\]