Digital insurance and long term risk

Chaire d'Excellence

A tree-based algorithm for individual reserving, with reporting delays and long developments

7th SMIF Conference (03/02/2020)

Work conducted within the Research Chair DIALog under the aegis of the Risk Foundation, a joint initiative by CNP Assurances and ISFA (Universite de Lyon)

Xavier MILHAUD
ISFA, University of Lyon (Lyon)

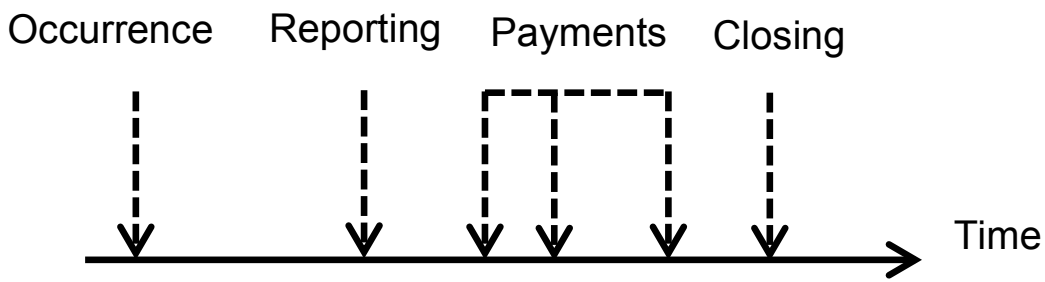
Joint work with O. Lopez (Sorbonne University)

AIM : ESTIMATE NONPARAMETRICALLY THE COST OF RBNS CLAIMS

Estimate some individual claim amount M

given features

- $T \in \mathbb{R}^+$: duration of the claim,
- $\mathbf{X} \in \mathbb{R}^d$: its characteristics.



However, for RBNS claims, we only observe the follow-up time Y (censored) and features $\mathbf{X}$.

Next

Main idea 1 : nonparametric models allows for flexibility.

→ Free relationship b/w response & risk factors.

Main idea 2 : claim lifetime plays a key role to explain claim cost.

→ Well-known from claim handling experts, with positive correlation between duration and amount.

Main idea 3 : significant impact of reporting on final claim amount.

→ And thus on reserves...

⇒ Two last ideas are extensions to [**?**]

▸ Next

DATA AT-HAND

We observe a sample of i.i.d. random variables $(Y_i, N_i, \delta_i, \mathbf{X}_i)_{1 \leq i \leq n}$ with same distribution $(Y, N, \delta, \mathbf{X})$, where

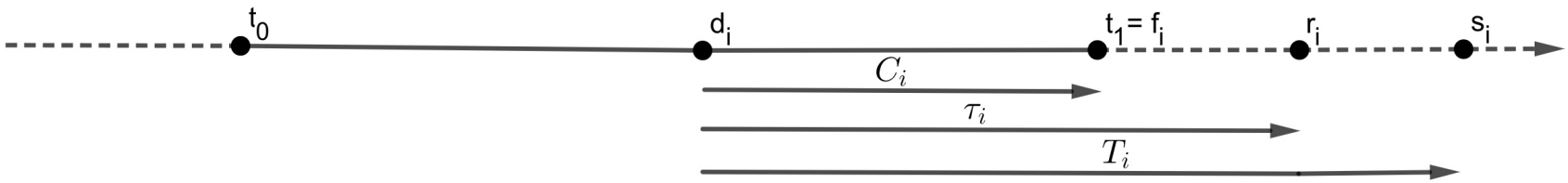
$$\begin{cases} Y & = & \min(T, C), \\ \delta & = & \mathbf{1}_{T \leq C}, \\ N & = & \delta M, \\ \mathbf{X} & \text{is} & \text{the vector of individual characteristics.} \end{cases}$$

C : censoring variable, coming from the censoring mechanism.

$\Rightarrow$ C impacts both T and M at same time.

REPORTING DELAY AND LEFT-TRUNCATION

In the context of reserving, the reporting delay τ can sometimes be large, leading to unknown claims.
$\Rightarrow$ **The claim is observed conditionally to** $C \geq \tau$.



Usually, the phenomenon is truncated when $T \leq \tau$...

Here, $T \leq \tau$ means that claim was closed before being reported, but appears in the database !

REMIND OUR GOAL

We seek the best prediction for M related to still open claims (from available data) :

$$M^* = E\left[M \mid \delta = 0, y, \tau, \mathbf{x}\right],$$

Or equivalently

$$M^* = E\left[M \mid T \geq y, \tau, \mathbf{x}\right].$$

REGRESSION TREES (COMPLETE observations)

$$\pi_0(\mathbf{x}) = E_0[T \,|\, \mathbf{X} = \mathbf{x}] \tag{1}$$

→ Most famous : linear relationship b/w T and $\mathbf{X}$ (restrictive class).
→ General solution : OLS, solve

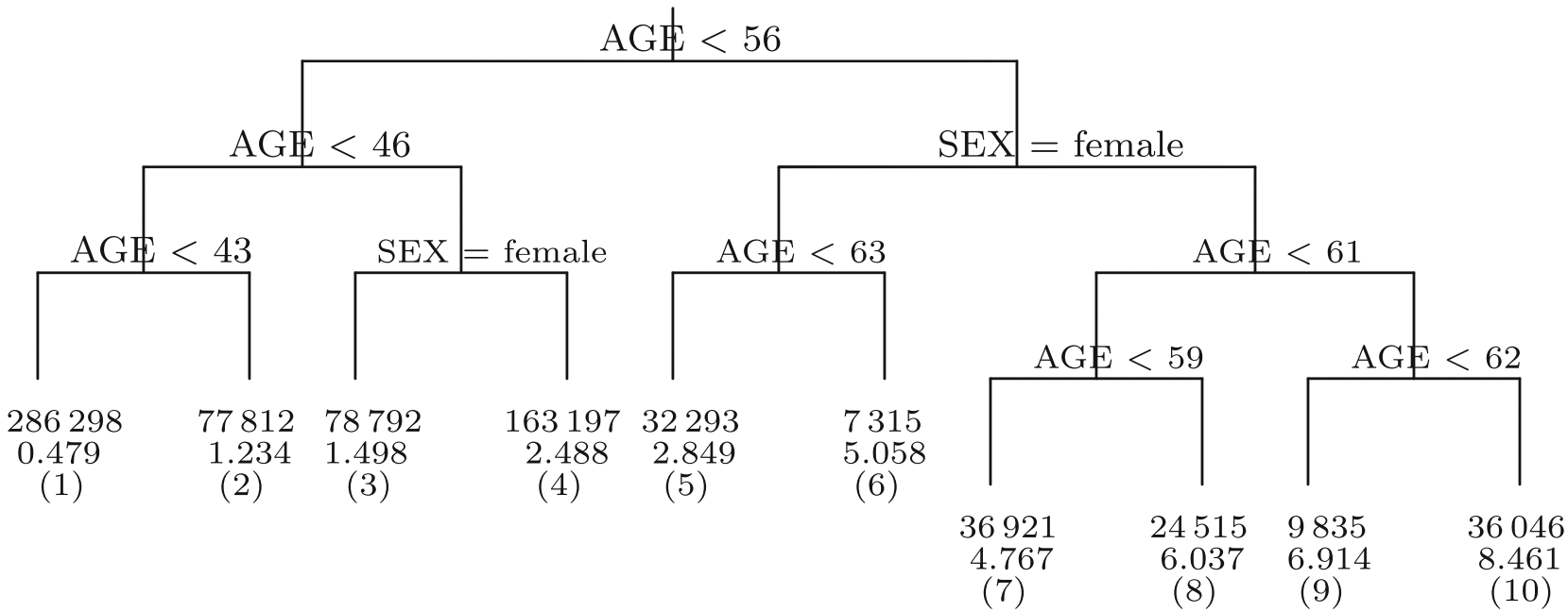
$$\pi_0(\mathbf{x}) = \arg\min_{\pi(\mathbf{x})} E_0\Big[\phi(T, \pi(\mathbf{x})) \,|\, \mathbf{X} = \mathbf{x}\Big] \tag{2}$$

where $\phi(T, \pi(\mathbf{x})) = (T - \pi(\mathbf{x}))^2$.

→ CART : recursive partitioning of covariate space ⇒ minimizes intra-node variances at each step, maximum homogeneity on T following the segmentation rule ⇒ piecewise-constant estimator !

EXAMPLE OF TREE, CLASSIFICATION MORTALITY [Olbricht, 2012])

SwissRe portfolio : $\simeq$ 1.5M indiv. observed over 4y (gender, age).



BACK TO OUR CONTEXT : INCOMPLETE OBSERVATIONS

T is not fully observed... $\Rightarrow$ Adapt CART, first to censoring

1. "Classical KM weights" $\Rightarrow$ additive version of $\hat{F}$: let $\hat{G}$ KM estimator of $G(t) = \mathbb{P}(C \leq t)$, then we have

$$W_{i,n} = \frac{\delta_i}{n[1 - \hat{G}(Y_i-)]} \Rightarrow \hat{F}(t) = \sum_{i=1}^{n} W_{i,n} \mathbf{1}_{Y_i \leq t} \rightarrow F(t)$$

2. IPCW with KM estimator of G :

$$\sum_i W^*_{i,n} \psi(Y_i) = \frac{1}{n} \sum_i \frac{\delta_i \psi(Y_i)}{1 - G(Y_i-)} \xrightarrow[as]{LLN} E\left[\frac{\delta \psi(Y)}{1 - G(Y-)}\right] = E[\psi(T)]$$

Then to reporting delays τ !

$\Rightarrow$ Necessary to modify the classical Kaplan-Meier weights !

Recall that we only get an observation when $C > \tau$...

Assume that $(T, \tau) \perp\!\!\!\perp C$, then one may consider

$$W_{i,n} = \frac{\delta_i \mathbf{1}_{\tau_i < Y_i}}{\sum_{j=1}^n \mathbf{1}_{\tau_j < Y_i \leq Y_j}} \prod_{Y_k < Y_i} \left(1 - \frac{\delta_k \mathbf{1}_{\tau_k < Y_k}}{\sum_{j=1}^n \mathbf{1}_{\tau_j < Y_k \leq Y_j}} \right).$$

Weighted CART, interpretation...

OUR APPLICATION : PREDICT RBNS RESERVE

Recall that we wish to estimate $\mathbb{E}[M \mid T, \mathbf{X}]$.

For RBNS claims, it amounts to estimate

$$\boxed{M^* = \mathbb{E}[M \mid T \geq y, \mathbf{X}]}$$

Problem to use weighted CART (wCART) : T is an incomplete (censored) explanatory variable when considering RBNS claims !

STRATEGIES TO PREDICT RBNS RESERVE

- Use Bayes formula :

$$M^* = \mathbb{E}[M \,|\, T \geq Y, \mathbf{X} = \mathbf{x}] = \frac{\mathbb{E}[M \mathbb{1}_{T \geq Y} \,|\, \mathbf{X} = \mathbf{x}]}{\mathbb{E}[\mathbb{1}_{T \geq Y} \,|\, \mathbf{X} = \mathbf{x}]}$$

  $\Rightarrow$ Yields to build 2 wCART trees (results presented hereafter)

- Use Plug-in principle :
  1. build $\hat{\pi}$ estimator of $\pi(t, \mathbf{x}) = E[M \,|\, T = t, \mathbf{X} = \mathbf{x}]$ from wCART.
  2. then fit a model for $T \,|\, T \geq y, \mathbf{X} = \mathbf{x}$, from which a prediction $\hat{T}(y, \mathbf{x})$ can be computed (e.g. with Algorithm 1).
  3. finally, predict M^* by using plug-in type estimator $\hat{\pi}(\hat{T}(y, \mathbf{x}), \mathbf{x})$.

Once M^* predicted, easy to get the associated individual reserve !

CONTEXT OF APPLICATIONS

Estimate the individual reserves at some given settlement dates.

- Use backtesting :
  1. consider only closed claims (known final claim amount) ;
  2. censoring/truncation variables updated at settlement date ;
  3. define learning / test samples to build / validate estimators :
     → predictions compared to actual data **on test sample only**,
     → get indicators of the predictive power at individual level
  4. sum indiv. reserves to get an indicator of overall performance.

- Compare results to Chain Ladder ([Mack, 1993]) and Cox model ([Cox, 1972]) ;

- Bootstrap resampling to approximate variance of estimators.

TPL INSURANCE

Dataset : `ausautoBI8999` (in `R` package `CASdatasets`) .

Information :

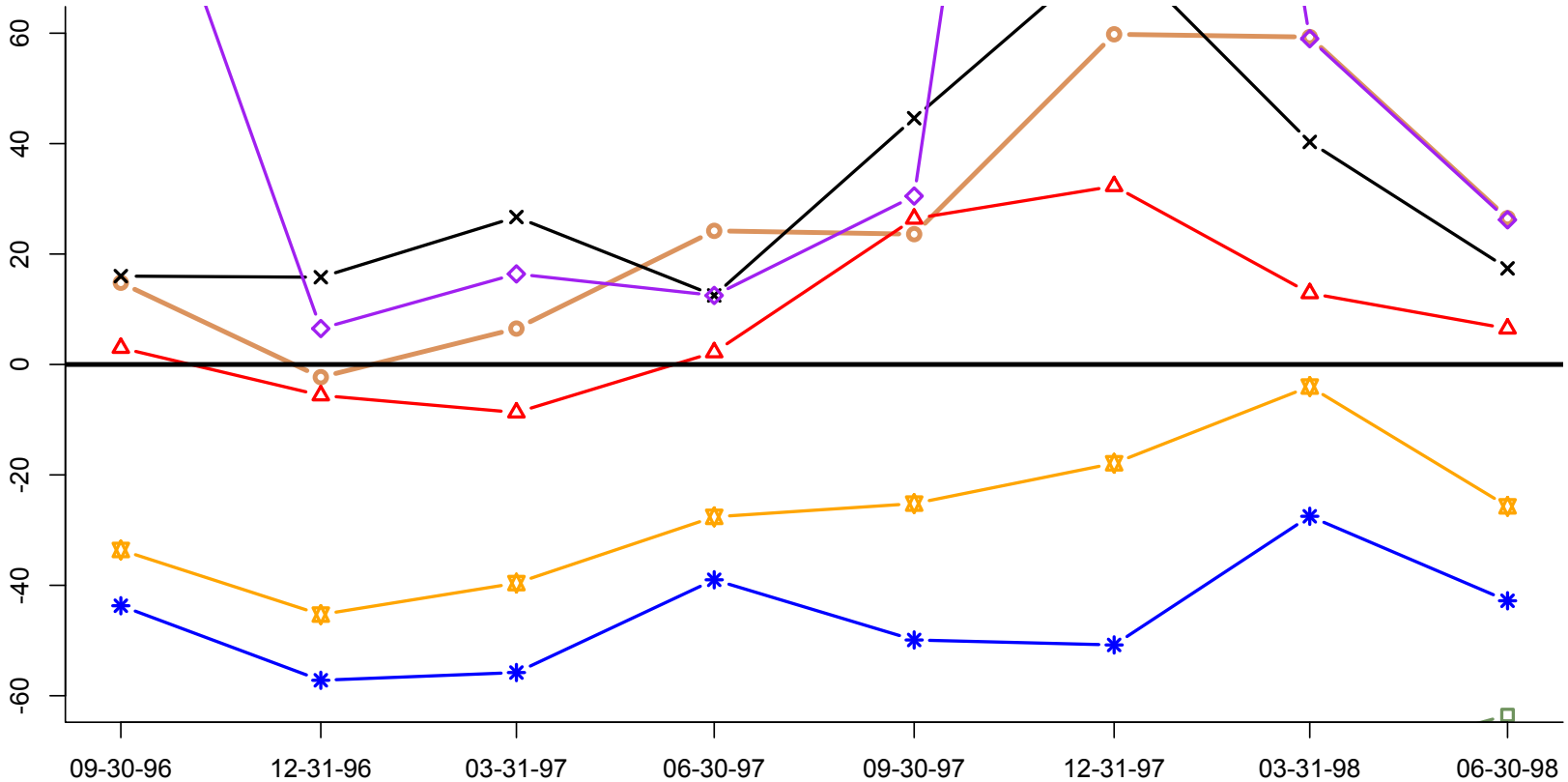
- 22 036 claims in motor insurance over ten years (1989-1999),
- aggregated settled claim amount $\Rightarrow M$
- dates : accident, reporting, closing $\Rightarrow T, \tau$
- Individual claim features $\mathbf{X}$:
  - operational time (indicator for claim management difficulties),
  - type of injury,
  - # of injured people,
  - legal representation of the PH.

$\Rightarrow$ Introduce fictive administrative censoring with settlement dates !

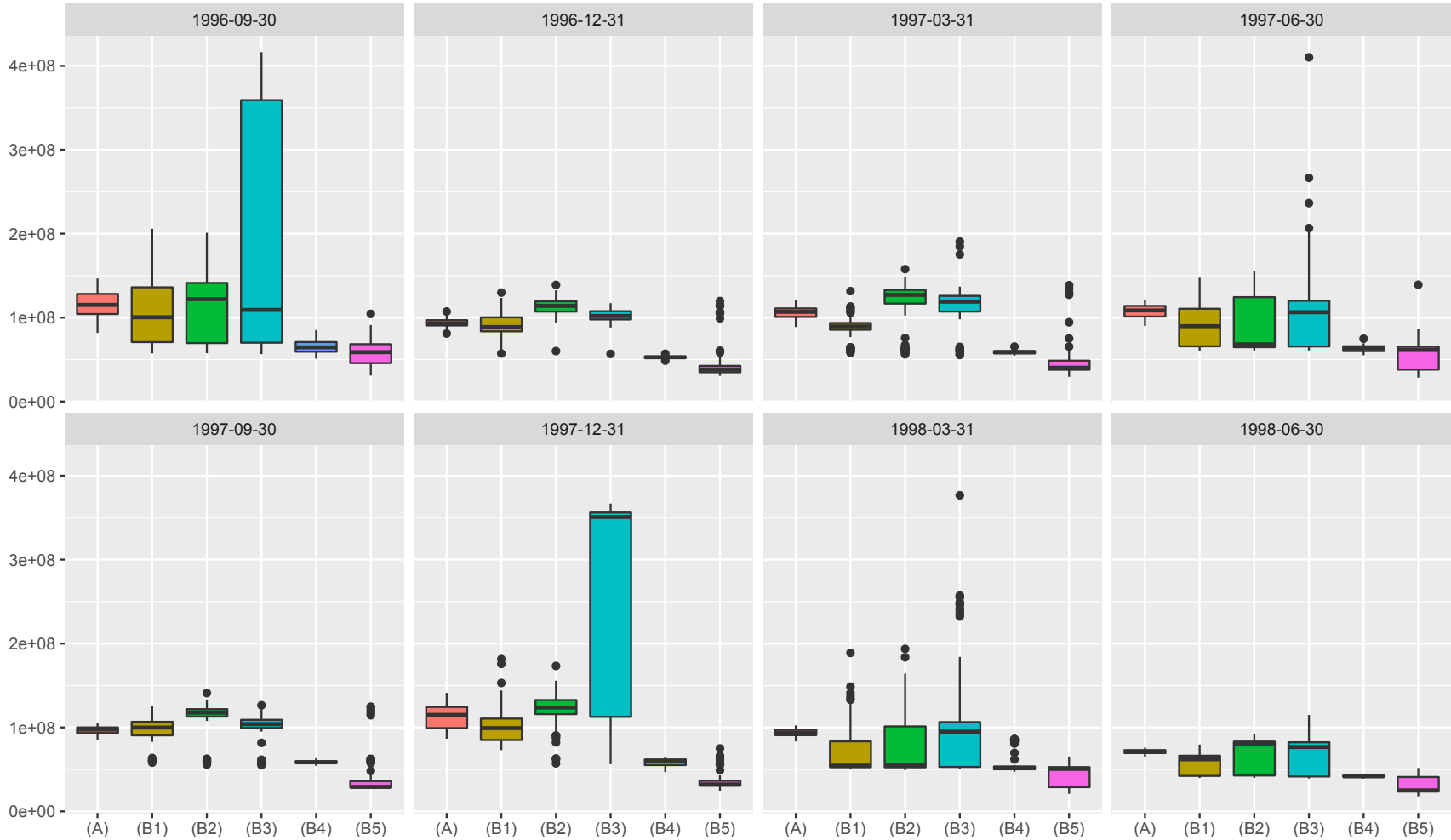
PREDICTION ERRORS ON THE GLOBAL RESERVES

At ≠ settlement dates.

Censoring rate ≃ 50%, 1000 bootstrap samples.

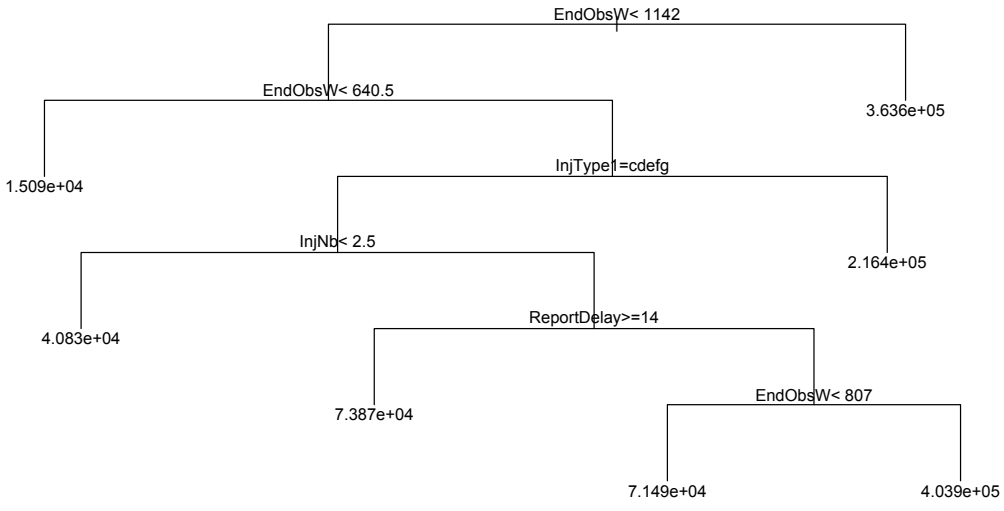


VARIANCE OF THE ESTIMATORS



FINAL REMARKS

- Plug-in strategy confirmed how crucial T is to predict M :



- Our method allows to deal with risk heterogeneity,
- Well-suited to large reporting delays and long developments (without extrapolation).

Cox, D. R. (1972).
Regression models and life tables (with discussion).
J. R. Stat. Soc. Ser. B, (34) :187–220.

Olbricht, W. (2012).
Tree-based methods : a useful tool for life insurance.
Eur. Actuar. J., 2(1) :129–147.