Using Open Files for Individual Loss Reserving in Property and Casualty Insurance

M. Pigeon and H. Cossette

CS-26

Université du Québec à Montréal (UQAM) and Université Laval

Outline

Introduction

Actuarial Loss Reserving

- Determine the outstanding liability for policies issued in the past.
- An estimate of the outstanding liability needs to be recorded in the annual statement.
- It represents the largest liability amount on the balance sheet.
- Two objectives :
  - maximum accuracy ; and
  - better understanding of the underlying components of the risk.

Collective Approaches

Loss reserving is traditionally based on an aggregated dataset (run-off triangle).

Occurrence period	Development period 1	2	3	4	5	6
1	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
2	C_{21}	C_{22}	C_{23}	C_{24}	C_{25}	C_{26}
3	C_{31}	C_{32}	C_{33}	C_{34}		C_{36}
4	C_{41}	C_{42}	C_{43}			C_{46}
5	C_{51}	C_{52}				C_{56}
6	C_{61}					C_{66}

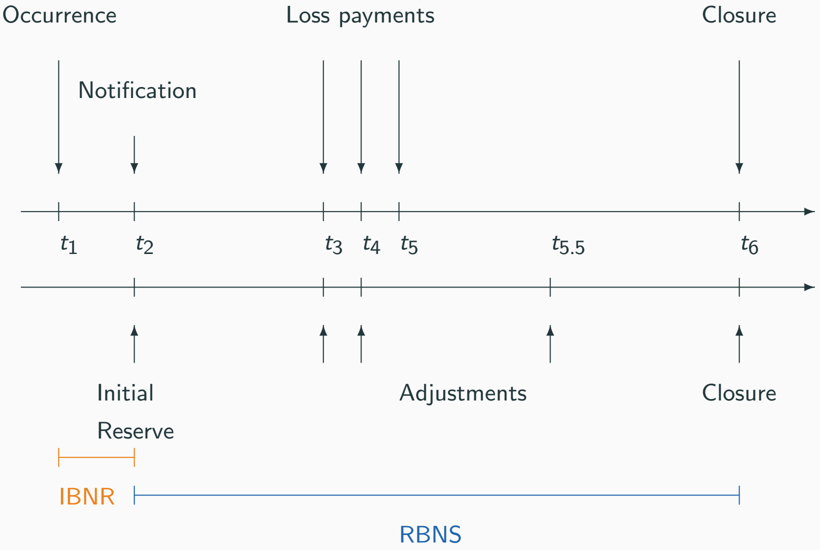
Table 1: Cumulative claims amounts

Collective approaches

- There are many classical (or aggregate, or collective) methods to evaluate reserves.
- Widely discussed in the literature, e.g. *Stochastic claims reserving methods in insurance*[1] by M.V. Wüthrich and M. Merz, or *Estimating unpaid claims using basic techniques*[2] by J. Friedland, J. for an extensive discussion of existing methods.
- While insurance companies always had access to very detailed information, computational and cultural limitations have traditionally prevented their use.
- Nowadays, practitioners have the ability to perform more rigorous reserving models with more detailed information, but traditional collective methods are still dominant in loss reserving practice.

1. Wiley Finance
2. Casualty Actuarial Society, vol. 201

Individual Dynamics



Individual Approaches

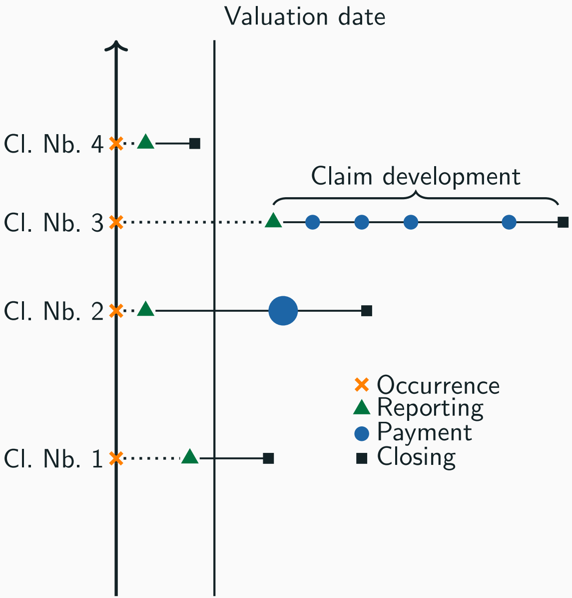
- Individual loss reserving approaches can be traced to the 1980s.
- It is in 2007 that the subject really took off with the availability of detailed data, and the development of computing resources.
- On the one hand, statistical learning techniques are widely used in the field of data analytic.
- On the other hand, only few approaches based on these techniques have been developed in individual loss reserving models.
- In this talk, we focus on tree-based models.

Open Claims

- Almost all individual models assume the availability of many closed files.
- In practice, this assumption is never verified, and the actuary must include open files in the modeling process.
- Two families of approaches : **(A)** strategies based on survival analysis, and **(B)** strategies based on imputation of missing data.
- The main objective of this talk is to investigate both strategies through 2 actuarial models : a tree-based censored regression model from O. Lopez, X. Milhaud and P.E. Thérond (strategy **A**), and an individual loss reserving model using imputation from F. Duval and M. Pigeon (strategy **B**).

2 Strategies

Portfolio


Valuation date
Cl. Nb. 4
Claim development
Cl. Nb. 3
Cl. Nb. 2
Occurrence
Reporting
Payment
Closing
Cl. Nb. 1

No Strategy : Including only Closed Claims

- Actually, there is a third way : including only closed claims in the modeling process.
- Obviously, this is not a good strategy : it leads to building the model using a too high proportion of "simple cases" and underestimating the risk associated with the portfolio.

First Strategy : more Focus on Closed Claims

- Main idea : using a weighted regression (tree) procedure for censored data to correct the selection bias.
- We only keep closed claims in the modeling process, but we associate each claim with a weight according to the duration of the claim.
- Thus, the longest (more complex) claims will have higher weights and vice versa.

Second strategy : Complete all Claims

- Main idea : artificially generating values, or *pseudo-responses*, for all open files in order to "complete" the portfolio.
- We use classical approaches such as Chain-Ladder or (individual) generalized linear models to complete open claims.
- We obtain a predictive distribution for each of the pseudo-responses so we can choose what we will use (mean, quantile, etc.) in the modeling process.

A Toy Example to Illustrate how it Works

Artificial Portfolio

Table 2: Portfolio for the toy example

Claim id	Acc. year	Dev. year 1	Dev. year 2	Dev. year 3	Status (val. date)
1	2000	200	400	100	Closed
2	2000	300	400	150	Closed
3	2001	250	450	–	Open
4	2001	300	500	–	Open
5	2001	350	600	–	Closed
6	2002	400	–	–	Open
7	2002	200	–	–	Open

The valuation date is January 1st, 2003.

Strategy A

- $n = 7$ claims in the portfolio.
- Kaplan-Meier (KM) weights are defined by

$$w_k = \left(\frac{\delta_k}{n-k+1}\right)\prod_{i=1}^{k-1}\left(\frac{n-i}{n-i+1}\right)^{\delta_i}, \qquad k = 2, \ldots, n-1, \tag{1}$$

  with $w_1 = \delta_1/n$, and $w_n = \prod_{i=1}^{n-1}\left(\frac{n-i}{n-i+1}\right)^{\delta_i}$.
- $\delta_k = 1$ for closed claims and 0 otherwise.
- In the CART algorithm, the empirical cdf is replaced by

$$\widehat{F}_Z(x) = \sum_{k=1}^{n} w_k \mathbb{I}(Z_k \leq x).$$

Results with Strategy A

Table 3: Portfolio for a strategy based on survival analysis

Claim id	Paid	Duration (Z)	Status (val. date)	$w^{\text{class.}}$	w^{KM}	Pred. value
1	700	2.9930	Closed	1/7	0.4	–
2	850	3.0040	Closed	1/7	0.4	–
3	700	2.0013	Open	1/7	0	950
4	800	2.0024	Open	1/7	0	950
5	950	1.9911	Closed	1/7	0.2	–
6	400	0.9935	Open	1/7	0	950
7	200	1.0095	Open	1/7	0	950

$$\widehat{R}^{\text{RBNS}} = (950 - 700) + (950 - 800) + (950 - 400) + (950 - 200) = 1{,}700.$$

Strategy B

- We consider a generalized linear model with the over-dispersed Poisson distribution and a logarithmic link function (occurrence and development years as covariates).
- We use a quantile q of the ODP distribution as pseudo-responses. We (should) determine $q = 0.9$ using cross-validation.
- In the CART algorithm, we include all 7 closed, or artificially closed, claims in the portfolio.

Results with Strategy B

Table 4: Portfolio for a strategy based on imputation of missing data

Claim id	Paid	Status (val. date)	Exp. value	Pseudo-resp.	Pred. value
1	700	Closed	–	700	–
2	850	Closed	–	850	–
3	700	Open	857	895	934.5
4	800	Open	957	997	934.5
5	950	Closed	–	950	–
6	400	Open	1,058	1,100	1,100
7	200	Open	858	896	934.5

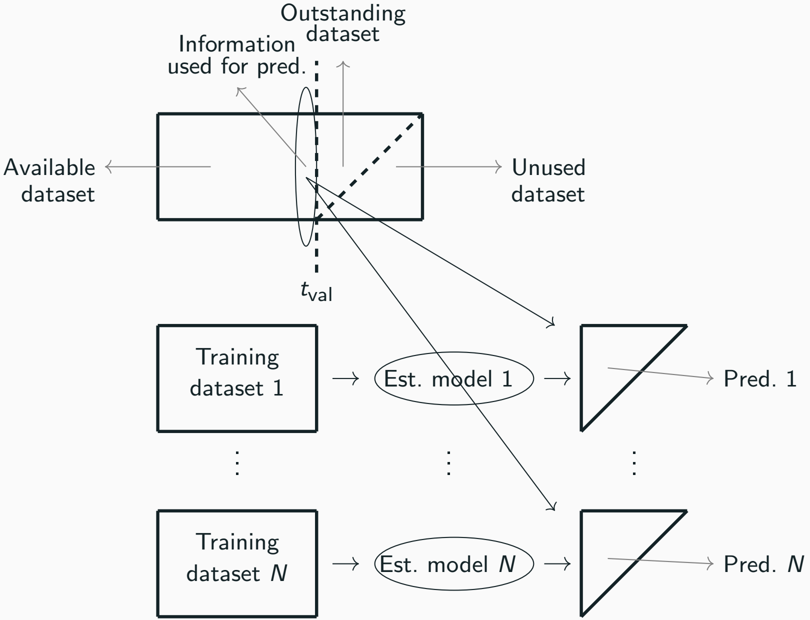
$$\widehat{R}^{\text{RBNS}} = (934.5 - 700) + (934.5 - 800) + (1{,}100 - 400) + (934.5 - 200) = 1{,}788.$$

Numerical Applications

Dataset(s)

- To respect a "replicability" criteria, we use simulated data by the *Individual Claims History Simulation Machine*, or ICHSM, described in
  → A. Gabrielli and M.V. Wüthrich (2018). An individual claims history simulation machine. *Risks*, 6, 29.
- It is a stochastic simulation machine that generates individual claims histories of non-life insurance claims.
- Based on neural networks calibrated on real, but unknown to us and to the public, non-life insurance data.
- Few covariates : lines of business (LoB), labor sector of the injured (cc), age of the injured (age), part of the body injured (inj_part) and reporting delay (RepDel).

General Structure



Analyzes

Using this procedure, we compare the performance of several approaches :

- Mack's model with bootstrap (Gamma distribution) ;
- collective over-dispersed Poisson model for reserves ;
- tree-based model using strategies based on survival analysis (strategy **A**), and
- tree-based model using strategies based on imputation (strategy **B**).

Analyzes

- For strategy **A**, we consider two models :

  $M1$ where the duration and the severity are modeled in a single step, and

  $M2$ where the duration is first modeled, then the severity.

- For strategy **B**, we consider two models :

  $M3$ using only occurrence and developments years as covariates, and

  $M4$ using all covariates.

Analyzes

All approaches are applied to three scenarios

(1) one line of business without inflation (mainly detailed in this talk),

(2) two lines of business without inflation), and

(3) two lines of business with inflation in the frequency.

Scenario I : one Line of Business and no Inflation

We construct a validation dataset containing 1,060 claims, $1{,}060 \times 12 = 12{,}720$ annual photographs and accident years between 1994 and 2005.

Table 5: Validation dataset (in $1,000) for Scenario I

Valuation date	% of censored data	RBNS amount	IBNR amount
01/01/2005	11.9	350	4
01/01/2006	11.7	406	8
01/01/2007	7.7	260	1
01/01/2008	6.6	192	1
01/01/2009	5.4	162	0
01/01/2010	4.2	124	0
01/01/2011	3.7	93	0
01/01/2012	2.6	68	0

Hyperparaters for the Strategy B

- We must first determine the level (quantile) q to be used in the completion of the databases.
- We generate databases of size 2,000 and calculate the mean absolute error of prediction (MAE) for a grid of values of q.
- Selected values are $\widehat{q}^{(2006,3)} = 0.85$, $\widehat{q}^{(2006,4)} = 0.85$, $\widehat{q}^{(2010,3)} = 0.8$, $\widehat{q}^{(2010,4)} = 0.7$ $\widehat{q}^{(2012,3)} = 0.6$ and $\widehat{q}^{(2012,4)} = 0.4$, where $\widehat{q}^{(i,j)}$ is the selected quantile for estimator j ($j = 3$: only occ. and dev. years as covariates and $j = 4$: all covariates) and valuation year i.

Hyperparaters for the Strategy B

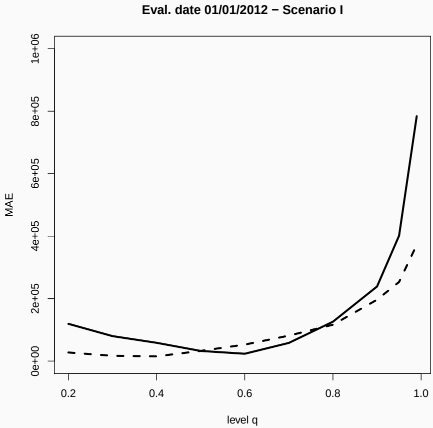


Figure 1: MAE of prediction as a function of the level q for a glm (ODP) using only occurrence and development years as covariates (solid line) and all covariates (broken line).

Scenario I : Results (2006)

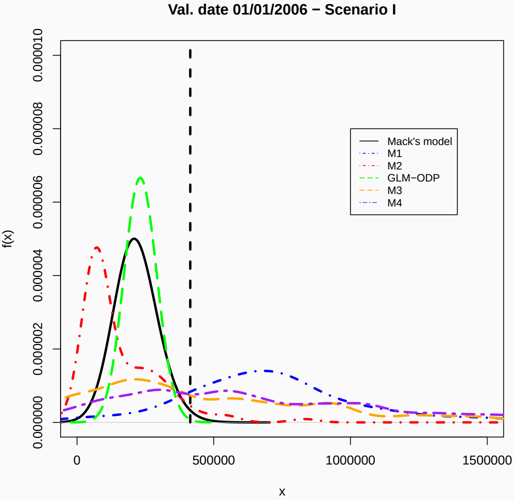


Figure 2: Predictive distribution of the reserve amount. The observed value is $414,000 for 2006.

Scenario I : Results (2010)

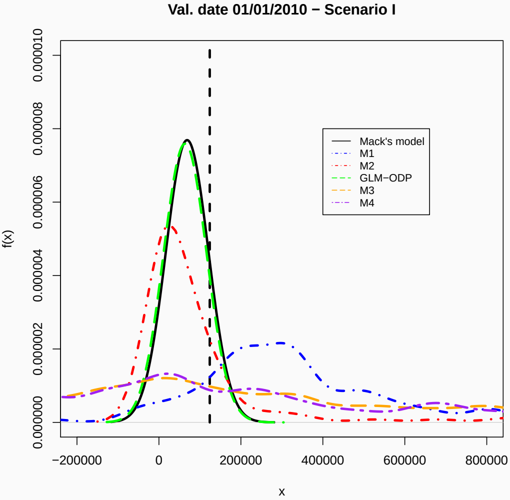


Figure 3: Predictive distribution of the reserve amount. The observed value \$124,000 for 2010.

Scenario I : Results (2012)

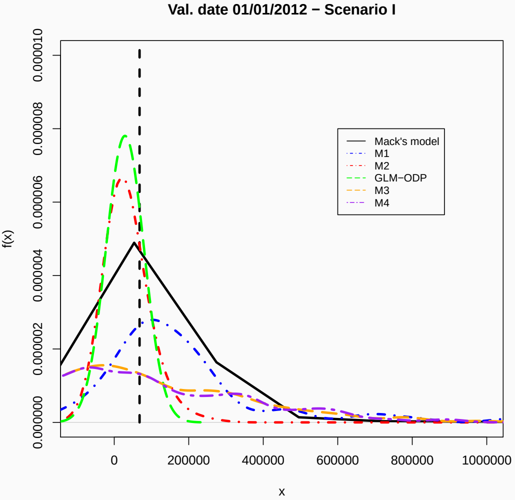


Figure 4: Predictive distribution of the reserve amount. The observed value is $68,000 for 2012.

Remarks

- Tree $M1$ model (blue line) produces very variable reserves resulting in very high expected values and very flattened predictive distributions.
- This effect is less pronounced for a more mature portfolio because there are much fewer open claims.
- Tree $M2$ model (red line) is much more stable, which is mainly due to the fact that there is more data to estimate $\mathbb{P}(Y > z)$ than $\mathbb{P}(M > m, Y > z)$.

Remarks

- Estimators $M3$ and $M4$ offer similar performance, which seems to indicate that the use of individual explanatory variables when imputing missing values does not significantly improve the performance of the model.
- We still add a caveat to this remark due to the small number of micro-level covariates in the database.
- Estimators $M3$ and $M4$ require much shorter computation times than estimators $M1$ and $M2$.

Scenario I : Results (2006 - Strategy A)

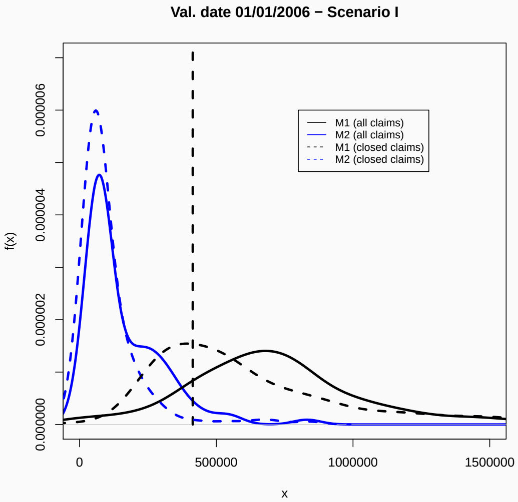


Figure 5: Predictive distribution of the reserve amount using all claims (solid lines) and only closed claims (broken lines) in the calibration process.

Scenario I : Results (2006 - Strategy B)

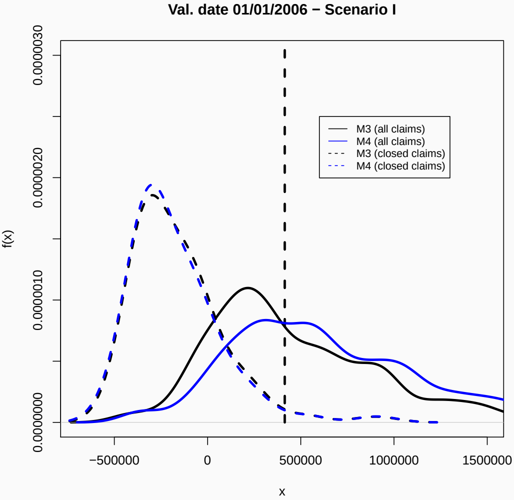


Figure 6: Predictive distribution of the reserve amount using all claims (solid lines) and only closed claims (broken lines) in the calibration process.

Remarks

- We confirm that, in practically all cases, the fact of not considering the open files in the calibration process leads to an underestimation of the risk.
- This underestimation is particularly important for estimators based on strategy **B**.

Conclusion

Conclusion

- Strategy in which open files would be removed from the calibration process is not advisable.
- The two estimators ($M1$ and $M2$) proposed in strategy **A** behave quite differently in all scenarios. The estimator $M2$ should be preferred given the stability it has shown compared to $M1$ which varies greatly.
- The performance of the estimators ($M3$ and $M4$) based on strategy **B** is rather similar in the three scenarios indicating that the individual information embedded in the covariates used in the imputation of missing data does not guide the model to better results.
- The two estimators ($M3$ and $M4$) outperform the ones of strategy **A** based on Kaplan-Meier weights regarding computation time.

References

- Cossette, H. and Pigeon, M. (2021). A Comparison of Two Individual Tree-Based Loss Reserving Methods. *Submitted*.
- Duval, F., and Pigeon, M. (2019). Individual loss reserving using a gradient boosting-based approach. *Risks*, 7, 79.
- Lopez, O., Milhaud, X., and Thérond, P. E. (2016). Tree-based censored regression with applications in insurance. *Electronic Journal of Statistics*, 10(2), 2685-2716.
- Lopez, O., Milhaud, X., and Thérond, P. E. (2019). A tree-based algorithm adapted to microlevel reserving and long development claims. *ASTIN Bulletin*, 49(3), 741-762.