TWO INCOMPATIBLE OBJECTIVES WITH INDIVIDUAL RESERVE MODELS: AN APPROACH WITH MULTIVARIATE ADAPTIVE REGRESSION SPLINE MODELS

A PREPRINT

Jean-Philippe Boucher
Chaire Co-operators en analyse des risques actuariels
Departement de mathematiques
Universite du Quebec a Montreal
boucher.jean-philippe@uqam.ca

November 3, 2021

ABSTRACT

We show, with simulations, that if the age of the claim is directly used as a covariate in a granular reserving model, the insurer cannot simultaneously expect to have (1) a precise estimate of each individual case reserve and (2) a precise estimate of his actuarial liability. Based on that finding, we develop a new granular reserving model, based on Multivariate Adaptive Regression Spline (MARS) models that are known to have an interesting bias-variance trade-off. By using several hinge functions that can interact with other covariates, this new model proposes a flexible way to model the relationship between the claim's age at closure and the amount of the claim. We show that the MARS granular reserving model generates an analytical form of the computation of each individual case reserve in the portfolio at all times, which is often essential to the regularly use of granular reserve models in practice. Finally, to reduce the effect of the age of the exposure on the granular reserving model, an improved model is proposed, where the residuals of a first regression model is used. All models are applied to real insurance data from a Canadian insurer.

Keywords Micro-reserving, Solvency, Multivariate Adaptive Regression Spline (MARS)

1 Introduction

Historically, P&C insurance companies and actuarial researchers, have used and developed provisioning techniques based on aggregated costs, mainly in the form of runoff triangles (see Wutrich & Merz [2008] or Friedland [2010] for an overview). Those methods usually generate a distribution for the actuarial reserve of an insurance portfolio, from which a final value of the reserve will be computed and then used in financial reports. Those provisioning techniques, however, are not designed to compute a precise estimate of all individual case reserves.

Instead of using collective methods, Arjas (1989) and Norberg (1986) proposed using individual claims data. With the recent availability of granular information about each claim, there are a growing number of publications in actuarial sciences that pursue that idea. These publications generally suppose that the cost of each claim in the portfolio can be individually predicted, see, for example, Pigeon et. al. (2013), Antonio and Plat (2014), or Duval and Pigeon (2019). Techniques that use individual claims data are now seen as good alternatives to collective methods for calculating the reserves that an insurance company must include in its financial statements. For more details, we refer to papers that compare individual and collective approaches, such as Huang et al. (2015), Charpentier and Pigeon (2016), and Wüthrich (2018).

Because they are based on calculations at the individual claim level, granular techniques can also be used to obtain precise individual estimates of the ultimate cost of each open claim. Granular reserving approaches can thus be used in many ways:

1. They can replace claims adjusters in setting individual case reserves for each open claim. Case reserves can then be used as a tool when negotiating with insureds for a claim settlement;
2. Having individual case reserves could help claims managers to choose which action to take for each open claim;
3. Having, at any given time, a precise estimate of the case reserve of each open claim, insurers could even dynamically track the actuarial liabilities of any part of their portfolio by summing each individual case reserve[1];
4. With an adequate estimate of the case reserve of each open claim, actuaries from ratemaking teams could potentially use the whole portfolio, without relying only on closed claims, to perform risk segmentation and compute premiums.

Analyzing all these uses, we see that they do not share the same objectives. Indeed, at time x, by having an instantaneous case reserve of $R_i(x)$, for each claim $i, i =, \ldots, n(x)$, for all $n(x)$ open claims of the portfolio, we can say that a P&C insurance company is willing to have a granular reserving model that satisfies the two following objectives:

O1 : The most accurate estimate of each individual case reserve at any time during the life of the claim.

O2 : An adequate value of the total reserve of the insurance portfolio, at any given time, to determine the company's actuarial liability.

In this paper, we show that if we must use the time since the opening of the claim (which will be called the *age of the claim*), it is possible that the two objectives mentioned above cannot be jointly satisfied by a single granular reserving model.

1.1 Simple Example

To illustrate the conflict between having a correct overall estimation of the actuarial liabilities and an accurate estimate of case reserves for open claims, we use a simple example. Suppose that we have a portfolio with only $n = 3$ claims, all of which occur at the same moment and are all reported at the same time $x = 0$:

[1] If the insurer does not consider incurred but not reported claims, (IBNR claims).

- Claim #1, which will be closed at time $x = \tau_1$ and which will cost \$200;
- Claim #2, which will be closed at time $x = \tau_2$ and which will cost \$500;
- Claim #3, which will be closed at time $x = \tau_3$ and which will cost \$2000;

with τ representing the age of the claim at closure, with $\tau_1 < \tau_2 < \tau_3$. At time $x = 0$, by using standard reserving models, we suppose that the insurer is able to know (approximately) that the total future amount to be paid for those 3 claims will be \$2700. However, obviously, even if the insurer approximately knows the total to be paid, the insurer does not know at $x = 0$ the cost of each individual claim. If the insurer wants to use reserving models to replace claims adjusters, a method for assigning a case reserve for each open claim must be found. Based on this simple example, we can assume that there are two main ways to assign individual case reserves for all open claims in this portfolio: a static and a dynamic estimation.

1.1.1 Static Estimation

The static estimation of the individual reserves means that at time $x = 0$, when all claims occur and are reported, an individual reserve is assigned to each claim and that this reserve does not change over time. We can then separate the analysis into time periods:

1. At $x = 0$, knowing that the total of amount of the claims will be (approximately) \$2700 for 3 claims, each claim has a case reserve of \$900.
2. Between time $[0, \tau_1[$, just before closing claim # 1, the sum of all individual case reserves is \$2700. The sum of all claims to be paid will also be equal to \$2700, meaning that the actuarial liability is thus correctly valued.
3. At $x = \tau_1$, claim # 1, reserved for \$900, is closed and paid at \$200.
4. Between time $[\tau_1, \tau_2[$, the sum of all individual case reserves is now \$1800 (\$900 and \$900). The insurer will have to ultimately pay \$2500 (\$500 for claim #2, \$2000 for claim #3). The sum of all case reserves is then undervalued by \$700.
5. At $x = \tau_2$, claim # 2, reserved for \$900, is paid at \$500.
6. Between time $[\tau_2, \tau_3[$, the sum of all individual reserves is now \$900. The final claim will be paid at \$2000, meaning that the actuarial liability is then undervalued by \$1100.
7. At $x = \tau_3$, claim # 3, reserved for \$900, is paid at \$2000.

In this static estimation of individual claims, just before each settlement, we projected that claim #1 would cost \$900 but was closed at \$200, we projected that claim #2 would cost \$900 but was closed at \$500 and we projected that claim #3 would cost \$900 but was closed at \$2000. Even if the predictions are imprecise, the sum of all individual case reserves at their closure time (\$900 + \$900 + \$900) is equal to the total cost of claims (\$200 + \$500 + \$2000). However, we can see that the objective O2 was not satisfied at all times. More precisely, between τ_1 and τ_3, the sum of the individual case reserves is less than the actuarial liability. If the insurance company wants to estimate how much it will have to pay for open claims in the future, it cannot simply sum all individual case reserves as it would underestimate the overall liability.

1.1.2 Dynamic Estimation

Static estimation cannot satisfy O2. Moreover, this technique is unrealistic in practice because insurers will normally change their individual case reserves over time. For example, after the payment of the first claim at τ_1, knowing that the total amount of future claims to be paid is now \$2500, the insurance company will increase the case reserves for the two remaining claims. This is what we call the dynamic estimation of individual claims. In a scenario like this, we can again separate the analysis into time periods to analyze the impact of this type of reserve approach:

1. At $x = 0$, knowing that the total of amount of claims will be \$2700, each claim is reserved for \$900.
2. Between time $[0, \tau_1[$, the sum of all individual reserves is \$2700, for a total of ultimate cost of claims of \$2700. The actuarial liability is thus correctly valued.

3. At $x = \tau_1$, claim #1, reserved for \$900, is paid at \$200. The remaining claims will cost \$2500, so each open claim will have a case reserve of \$1250.
4. Between time $[\tau_1, \tau_2[$, the sum of all individual case reserves is now \$2500, and the final cost to pay for all open claims is also \$2500. The actuarial liability is thus correctly valued.
5. At $x = \tau_2$, claim #2, reserved for \$1250, is paid at \$500. We know that the remaining claim will cost \$2000, so the last claim is then reserved at \$2000.
6. Between time $[\tau_2, \tau_3[$, the sum of all individual reserves of open claims is now \$2000, and we know that the final claim will be paid at \$2000. The actuarial liabilities are thus correctly valued.
7. At $x = \tau_3$, claim #3, reserved for \$2000, is paid at \$2000.

For objective O2, throughout the period analyzed, the actuarial liability has always been correctly valued. Unlike in static estimation, the insurance company could, at any time, sum the case reserves of all open claims to obtain an accurate estimate of how much it will have to pay for open claims in the future (excluding IBNR claims). However, concerning objective O1, the overall quality of the individual prediction at closure time is incorrect. Each time, the claim was settled for an amount that was less than (or equal to) the case reserve at that time, and the sum of each case reserve just before their time of closure is much higher than the total sum of claims. If the measure of quality of a granular reserving model was based on the difference between the case reserve at time τ and the amount that was finally paid, dynamic estimation models would be rejected.

1.2 Structure of the paper

It is easy to understand the problems raised by this simple example.

1. In the case of a static reserve, when the average cost of claims increases over time, the average individual reserve of the remaining open claims should also increase. Because we are using a static approach, we cannot modify each individual reserves, and this causes a deficit in the actuarial liability.
2. For the dynamic model, knowing that the reserve of a specific claim can be seen as the average of all future claims, and knowing that the claims are in increasing order, the next claim amount will always be lower than the expected average of future claims (until we reach the last claim). That means that the sum of all individual claims will necessarily be less than the sum of the amounts reserved for each claim at the time of closure.

Consequently, the insurer cannot expect to simultaneously have a precise estimate of each individual case reserve, and, by summing the case reserves of all open claims of its portfolio, a precise estimate of actuarial liability. To better understand the problem, and to propose ways to handle it, this paper will be structured as follows. In Section 2, we use simulations to study two more sophisticated examples. In order to obtain an individual claims reserving model that satisfies objectives O1 and O2, the age of the claim cannot be used in the model. If age is used, we see that objectives O1 and O2 cannot be jointly satisfied. We also generalize simulations with covariates, and show that a model that ignores important covariates will be less effective regarding both objectives. We then discuss the implications of the conflicting objectives.

In Section 3, we use a real insurance dataset from a major Canadian insurance company to see how granular reserving models should be handled regarding the conflicting objectives. To illustrate the situation, we develop a new individual reserve model based on Multivariate Adaptive Regression Spline (MARS) models, which are well known to have an interesting bias-variance trade-off. We show that the MARS approach is a way to model the relationship between a random variable and a numeric covariate, and it is used to model the relation between the age of the claim at closure and the value of the claim. We show that the MARS model can be used to obtain an analytical form of the case reserve at all times, a condition that is often essential for the regular use of individual reserve models in practice. In Section 4, we proposed a way to estimate the parameters of the model. The estimated model is then analyzed, explained, and generalized to improve its precision. In conclusion, we suggest some other ideas for research.

2 Simulations

The simple example at the beginning of the paper helps to demonstrate the problem of an expected growth in the cost of a claim as it ages. However, simulations of larger insurance portfolios must also by analyzed in order to grasp how we should model individual reserves in these cases.

2.1 Waiting Times and Expected Value of Truncated Costs

First, to study the impact of cost growth as a function of the age of a claim, we will assume a claim process with the following structure:

- We work with an insurance product where there is only one indemnity payment that occurs on the closing date of the claim, as shown in Figure 1. This type of claims is similar to the claims we will analyze with real insurance data.
- The age of the claim i at closure, τ_i, corresponds to the time between its date of occurrence and its date of closure.
- We assume that the age of a claim at closure, τ_i, follows an exponential distribution of mean 180 days.
- We assume that the cost of a claim i, S_i, follows a gamma distribution of mean $\mu_i = 10,000 + 50\tau_i$, and of variance $\mu_i^2\alpha$, with $\alpha = 10,000$.
- We simulate 10 claims per day, over a calendar time period ranging from 1 to $15,000$ days.

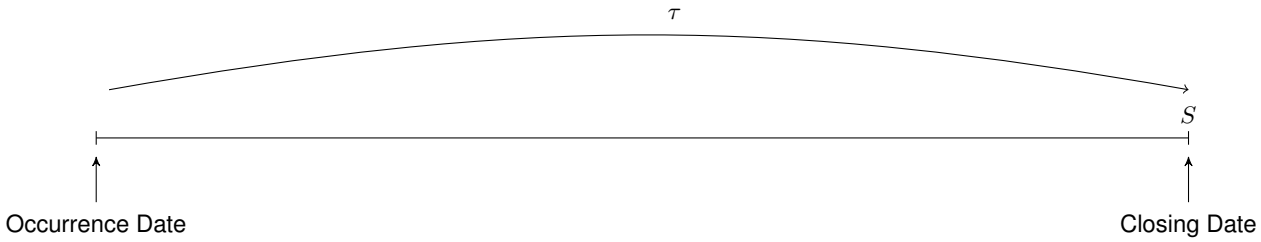

Figure 1: Illustration of a claim's life

We thus create a model where there is a growth of cost as the age of the claim increases. Figure 2 shows the simulated database, showing the link between the age of a claim at closure (τ) and its closing value (S).

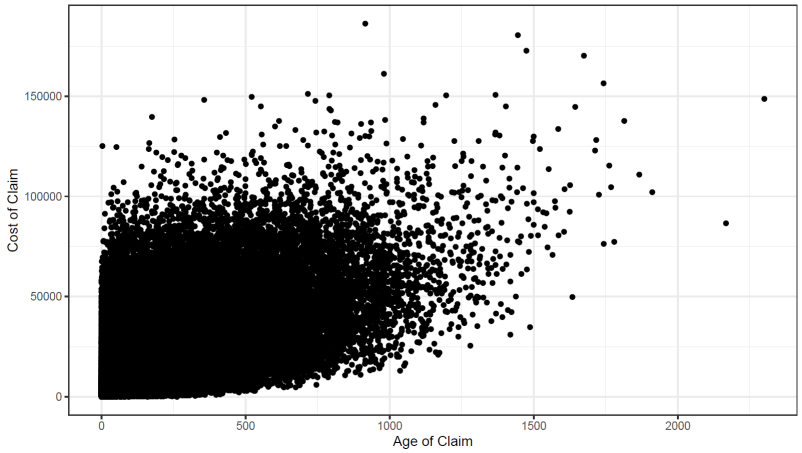

Figure 2: Simulated data

2.1.1 Calculation of Individual Reserves

Based on the static and dynamic approaches in the Introduction, we believe there are at least three possibilities to consider when calculating the reserve for each claim, at any age x:

1. Identical and unique reserves for all claims, at any age x (a.k.a. static reserve). The reserve for each claim i would therefore be equal to:

$$\begin{aligned} R_i^{(1)}(x) = E[S_i|x] = E[S_i] &= E[E[S_i|\tau_i = t]] \\ &= \int_0^\infty (\beta_0 + \beta_1 t)\lambda \exp(-\lambda t)dt = \beta_0 + \frac{\beta_1}{\lambda} \end{aligned} \tag{1}$$

   As we saw in the Introduction, the problem with $R_i^{(1)}$ is that the value of the reserve does not depend on the age of the claim at the evaluation x. Given the value of $\beta_1 > 0$, if we analyze an opened claim at age x, it would be logical to believe that the expected value of the claim is greater than it was when x was equal to 0. This scenario is therefore unrealistic in practice: with the increase of x, the expected value of the claim must also increase.

2. Reserve a claim at any age x, as if x corresponds exactly to the time the claim is closed:

$$R_i^{(2)}(x) = \beta_0 + \beta_1 x \tag{2}$$

   We might be tempted to consider this approach when we fit a basic regression model for claim i, linking the amount paid S_i with the age of payment τ_i. It is, however, a naive approach that has no theoretical foundation. Indeed, the approach only estimates the reserve at age x as if the claim closed exactly at the same age x. Knowing that the value of claim S_i increases over time, we can expect that $E[S_i|x] > R_i^{(2)}$ for all $x > 0$. That's why we did not consider this approach for the example in the Introduction. However, as it may sound logical at first, we wanted to include it in our analysis to show its limitations.

3. Using the conditional expectation of S_i to compute the reserve, knowing that the claim has survived until the time x (a.k.a. dynamic reserve):

$$R_i^{(3)}(x) = E[S_i|\tau_i > x] = E[E[S_i|\tau_i = t]|\tau_i > x] \tag{3}$$

$$= \int_x^\infty (\beta_0 + \beta_1 t)\frac{f(t)}{1 - F(x)}dt = \beta_0 + \beta_1 x + \frac{\beta_1}{\lambda} \tag{4}$$

   This approach seems to be the most logical and coherent form for the individual reserve. However, used in connection with the expression $\beta_0 + \beta_1 x$, which is used to simulate the data, we quickly see that an extra element $\frac{\beta_1}{\lambda}$ is added with $R_i^{(3)}$.

2.1.2 Temporal Analysis of Individual Reserves

Thanks to the simulated data, we have at hand the time of occurrence of each claim, its final cost, and the time that it will take before the claim is closed. We can analyze the evolution of individual reserves $R_i^{(1)}$, $R_i^{(2)}$ and $R_i^{(3)}$ for $i = 1, \ldots, n$, when x corresponds to the time since the opening of the claim (age of claim), but also when x corresponds to calendar time. By discretizing the time per day, Figures 3 and 4 thus correspond to the daily evolution of the reserves when x is the age of the claim, or when x is calendar time.

Figure 3, based on the age of the claim, shows the difference between the various methods of reserving in calculating the actuarial liability, and the estimate of the reserve at the time the claim closes. As expected, method $R_i^{(1)}$ (in purple), which assigns a single and unique reserve to all claims from their opening date, under-evaluates the actuarial liability when the age of the claim increases. When the age of the claim increases, the average amount of claims is increasingly significant, but $R_i^{(1)}$ cannot modify the individual value of the reserves. The figure on the right, showing the total value

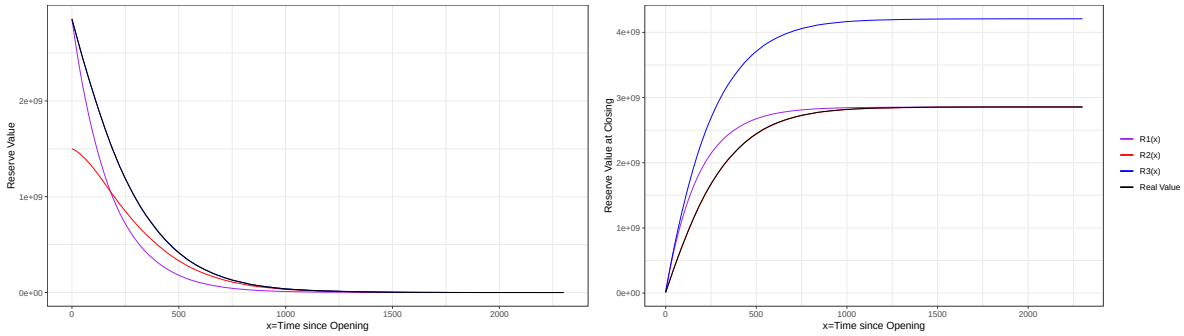

Figure 3: Total value of the individual reserves based on the age of the claim (x). The graph on the left shows the total value of reserves for open claims, while the graph on the right shows the value of the reserve when the claim was closed

of reserves at the time the claim closes, indicates that method $R_i^{(1)}$ is exact: the sum of the individual reserves (at closure) converges towards the total sum of claims. This is not surprising given that the sum $R_i^{(1)}$ is effectively equal to the sum of the claims. We see that $R_i^{(1)}$ satisfies objective O1, but not objective O2.

Analysis of $R_i^{(2)}$, shown in red, reveals that the situation is even worse than $R_i^{(1)}$ for objective O2 because the claims liabilities are always undervalued, even for small values of x. This result is unsurprising because reserve $R_i^{(2)} = \beta_0 + \beta_1 x$ will be constantly undervalued for $x < \tau$, with τ corresponding to the exact closing time of the claim. In the figure on the right, $R_i^{(2)}$ is directly under the black line. This is because $R_i^{(2)}$ is constructed to have an exact correspondence between the reserve and the amount paid when $x = \tau$. We therefore see that $R_i^{(2)}$ satisfies objective O1, but performs even worse for objective O2 than $R_i^{(1)}$.

Finally, $R_i^{(3)}$, in blue, under the black curve in the figure on the left, shows an almost perfect correspondence between the ultimate costs and the sum of the reserves, when x is the age of the claim. Conversely, the figure on the right shows that the sum of the individual reserves at the time of closing is much higher than the total that is really paid. Compared to the method $R_i^{(2)}$ which generated an exact result for objective O1, the difference can be explained by the presence of the element $\frac{\beta_1}{\lambda}$ of the equation of $R_i^{(3)}$. The difference between the total claim amounts and $R_i^{(3)}$ is then simply equal to $n \times \frac{\beta_1}{\lambda}$, with n corresponding to the total number of claims analyzed. In summary, in opposition to $R_i^{(1)}$ and $R_i^{(2)}$, the method $R_i^{(3)}$ satisfies objective O2 but does not satisfy O1.

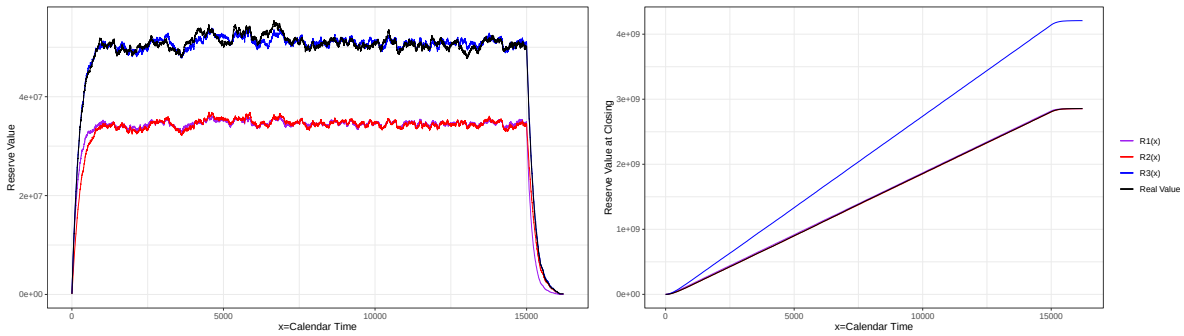

Figure 4: Total value of the individual reserves based on calendar time (x). The graph on the left shows the total value of reserves for open claims, while the graph on the right shows the value of the reserve when the claim was closed

It is also interesting to analyze the evolution of individual reserves as a function of calendar time, as shown in Figure 4. We see a relative similarity between $R_i^{(1)}$ and $R_i^{(2)}$ for objectives O1 and O2: systematic underestimation of actuarial liability at all times x, but an exact estimate of the sum of ultimate claims. Moreover, $R_i^{(3)}$ seems to satisfy the criterion for O2 *on average*, without always being completely sufficient. Indeed, on a few occasions (for example, in the interval

j	Type of Claim	Proportion	λ	β_0	α
1	Minor	40%	$1/30$	$1,000$	$10,000$
2	Moderate	30%	$1/100$	$2,500$	$10,000$
3	Major	25%	$1/600$	$20,000$	$10,000$
4	Catastrophic	5%	$1/2000$	$50,000$	$10,000$

Table 1: Parameters of simulations

$[6250, 6500]$), the sum of the amounts reserved is not sufficient to pay the sum of the expected claims. The random variation nature of claims reserving means that a safety margin greater than the complement $\frac{\beta_1}{\lambda}$ is necessary to satisfy O2, at the cost of reserving more than what the sum of all the claims will ultimately cost.

2.2 Covariates

One of the great advantages of analyzing reserves at the granular level is that precise individual information about a claim, or information from the insured, can be used in the modeling. With this in mind, to complete our analysis of simulations, we will assume the claim process is slightly more complex in order to see the impact of covariates in the pursuit of objectives O1 and O2. We will keep the same working hypotheses as those set out in Section 2.1, with $\tau \sim Exponential(\lambda)$ and $S|\tau \sim Gamma(\mu = \beta_0 + \beta_1\tau, \alpha)$. However, we will now suppose that $\beta_1 = 0$, leading to $\mu = \beta_0$, and we will introduce four types of claims, each having their own distribution parameters as indicated in Table 1.

By design, if the covariate identifying the type of claim is not used in the modelling, the age of the claim at closure becomes significant, and will have to be used to compute each case reserve. Indeed, we can see that the more severe the claim is, the higher its average cost is, and the longer it takes, in average, to close the claim. Figure 5 thus shows the simulated database.

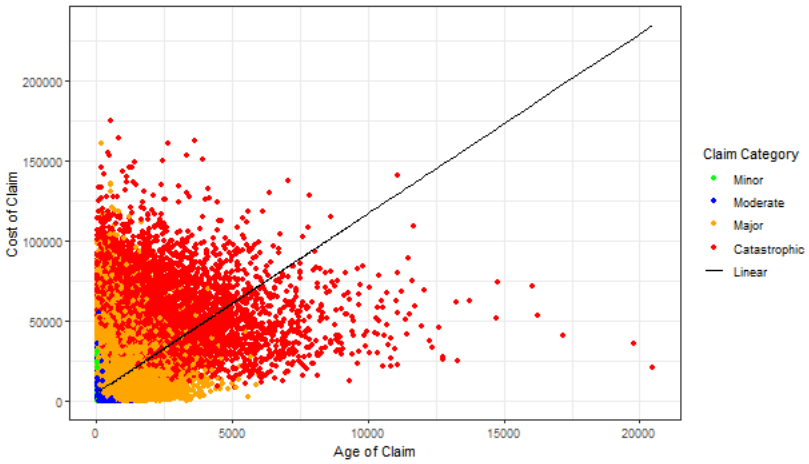

Figure 5: Simulated data with covariates, with relationships to the age of claims

We propose three approaches to better understand the impact of the covariates in our calculation of the reserve:

- Approach A1: We first assume that we do not know the value of each β_0, meaning that the covariates that affect the severity of the claim S are not considered. Instead, the age of the claim will be used to estimate each case reserve. As the link between the age of a claim (τ) and the closing value of the claim (S) is not known, a simple linear relationship is used. This is represented by the solid line of Figure 5. To find a more precise relationship between τ and S, a model with quadratic terms or with splines could be developed. However, for

the sake of illustration, we will use this simple approach. In this case, as shown with equation 4, at time x, the dynamic reserve of claims i, is equal to:

$$R_i^{(A1)}(x) \quad = \quad \widehat{\gamma_0} + \widehat{\gamma_1}\left(x + \frac{1}{\lambda_i}\right)$$

- Approach A2: In this more realistic approach, the covariates that affect the severity of the claim S are still not considered, but the distribution of τ_i is also totally unknown. To account for the mix of the exponential distributions, we will suppose a gamma distribution of mean α/δ and variance α/δ^2, from which both parameters are estimated from the data. In this case, at time x, it can be shown that the case reserve of claims i will be equal to:

$$R_i^{(A2)}(x) \quad = \quad \widehat{\gamma_0} + \widehat{\gamma_1}\frac{\widehat{\alpha}}{\widehat{\delta}}\frac{S_{\tau_i}(x;\widehat{\alpha}+1,\widehat{\delta})}{S_{\tau_i}(x;\widehat{\alpha},\widehat{\delta})}$$

  with $S_\tau(x)$ the cumulative function of τ.

- Approach B: We consider covariates for the modeling. In this case, at time x, the dynamic case reserve of claims i that belong to the type of claims $j \in \{1,2,3,4\}$ is simply equal to:

$$R_i^{(B)}(x) = \beta_0(i),$$

  and does not depend at all on the age of the claim.

As in the analysis of the first simulated data set, it would be interesting to analyze the trend in claim liabilities over calendar time.

2.2.1 Temporal Analysis of Individual Reserves

Figure 6 illustrates the reserve of the portfolio over time and the cumulative reserve value for each claim at the time of closure for each reserve models. The reserve models $R^{(A1)}$ and $R^{(A2)}$ both approximately satisfy objective O2 as the reserve value of both is close to what is left to pay. Reserve model $R^{(B)}$ is also shown in Figure 6 and is represented by the blue line, and it works slightly better than the other models. Overall, we can say that all three models satisfy objective O2, but models $R^{(A1)}$ and $R^{(A2)}$ should be improved, probably by defining a better relation between τ and S. As we saw in Figure 4, the linear approximation is not precise.

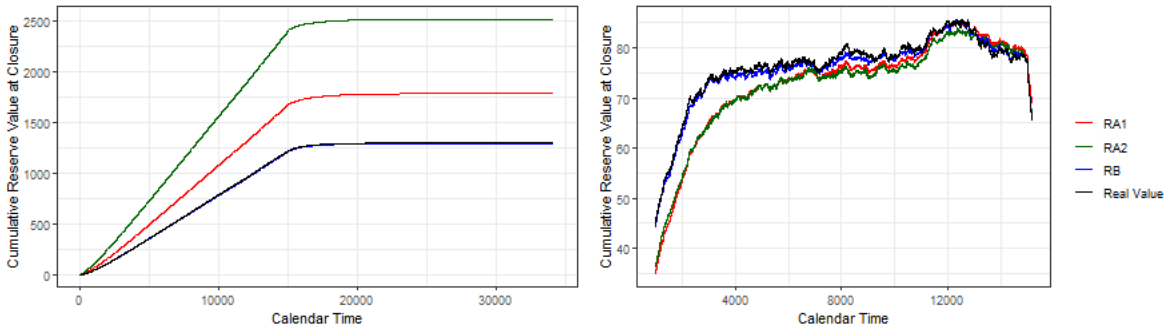

Figure 6: Cumulative reserve value at closure, for each calendar day (left), total reserve for each calendar day (right)

Concerning the precision of the case reserve at closure, i.e. at time τ, the right graph of Figure 6 shows the cumulative difference between models. For reserve models $R^{(A1)}$ and $R^{(A2)}$, there is a significant difference between what was expected to be paid and what was actually paid. This again highlights the fact that objective O1 may be not satisfied when the age of the claim is used directly in the modeling. Even if models $R^{(A1)}$ and $R^{(A2)}$ were almost the same for

objective O2, they are very different for objective O1. Indeed, model $R^{(A2)}$, which is based on an estimated distribution for τ, is far worse than $R^{(A1)}$. Note that there is no easy way to improve the precision of the case reserve at closure: if we reduce the case reserves of models $R^{(A1)}$ and $R^{(A2)}$ to be closer to the total amount paid, objective O2 will no longer be satisfied.

The blue line of reserve model $R^{(B)}$ cannot be easily seen in the figure because it is always hidden under the real value curve, in black. That means that, on average for each days, the total amount paid in closed claims is close to the amounts that were reserved. By ignoring the effect of the age of the claim in the computation of the case reserve, and by relying only on actual covariates (in our example, the type of claims), we see that $R^{(B)}$ can satisfy objectives O1 and O2.

Table 2 summarizes some of those results in relation to objective O1. Indeed, the table shows the sum of the predicted cost of all claims at time τ, and compares it with the true cost. It is interesting that the table also shows the mean square error (MSE) between models. We see that on average, the difference between what was predicted and what was observed at time τ is almost $\$12,000$ for model $R^{(A1)}$ and more than $\$15,000$ for model $R^{(A2)}$. For model $R^{(B)}$, even if Figure 6 does not show any difference by calendar day, an average difference of approximately $9,000 is observed, meaning that even a perfect reserve model shows random variations.

	Objective O1			Objective O2 (by day)			
	Diff (%)	MSE (Close)	MSE (Open)	Diff (%)	+/- 10%	+/- 5%	+/- 1%
$R^{(A1)}$	37.64%	11,886	11,360	-4.54%	84.24%	68.25%	11.80%
$R^{(A2)}$	94.15%	15,283	14,576	-5.14%	83.92%	62.47%	5.25%
$R^{(B)}$	-0.50%	9,303	9,303	-0.73%	100.00%	100.00%	55.59%

Table 2: Statistics for each claims i at time τ_i for different reserve models

For objective O2, we would expect that a good reserving model that uses the age of the claim as a predictor would precisely estimate the total reserve of the portfolio for each calendar day. Models $R^{(A1)}$, $R^{(A2)}$ and $R^{(B)}$ all approximately satisfied that, as we saw earlier. The last three columns of Table 2 shows the proportion of days the total reserve is within 10%, 5% or 1% of the true value.

If we come back to objective O1, we saw that the prediction of any models that use the age of the claim as a predictor is incorrect for a specific date: the closing date with $x = \tau$. If such a difference exists at time $x = \tau$, what does that tell us about the precision of case reserve estimation for the other times x? The fourth column of Table 2 (MSEP - Open) tries to answer that question by showing the average mean square error (MSE) per day for different models. That means that if a claim is open for T days, we computed the average difference between the case reserve and the final cost for each day. The value shown is the average error for all claims in the portfolio. As we may expect, at the individual level, the fitted values of case reserves from models $R^{(A1)}$ and $R^{(A2)}$ are much less precise than the reserve from model $R^{(B)}$. At any time during the life of a claim, when using $R^{(A)}$ models to estimate a case reserve, we should expect an error of $\$11,000 - \$15,000$, while the error of reserve model $R^{(B)}$ is less than $\$10,000$.

2.3 Impact of Incompatible Objectives

Based on the simulations, we see that we should use a reserving model that includes as many covariates as possible to minimize the use of the age of the claim in the modeling. However, if we do not use the age of the claim as a covariate when it is needed, we return to the static model $R^{(2)}$ (see equation [2]), which did not satisfy objective O2. If the age of the claim is shown to be significant in the modeling, we should use it. However, using the age of the claim like in models $R^{(3)}$, $R^{(A1)}$ and $R^{(A2)}$ means that objective O1 cannot be satisfied. By comparing the results of models $R^{(A1)}$ and $R^{(A2)}$, we also saw that a better modeling of τ and of the relationship between the cost of the claim and the age at closure could improve the accuracy of the estimate of each individual case reserve. Consequently, if the age has to used,

it is important to find the best relationship with the cost of the claim.

To conclude, if the age must still be used in the modeling, specific uses of the granular reserving models can be difficult. For example:

1. If actuaries want to use the whole portfolio for ratemaking, using each individual case reserve to to find the final cost of each open claim, we should expect bias in some rating parameters because each individual case reserve will not be accurately estimated.
2. If the insurer uses $R_i^{(3)}(x)$, the sum of case reserves at time $x = \tau$ will necessarily always be higher than what has actually been paid, as opposed to what we wanted to achieve with the objective O1. It is important for the insurer to be aware of this difference and to not judge the quality of their granular reserve models by this criterion. This means that use of granular reserving model will also contradicts what is currently asked of claims adjusters, who subjectively estimate all open claims.
3. Because it is impossible to satisfy objective O1 with a granular reserving model using $R^{(3)}(x)$, it is important for the insurer not to rely solely on the value of the calculated case reserve if the insurer had to negotiate the settlement of a claim with one of its insureds. Indeed, at time x, the objective of the insurer cannot be to close a claim at the case reserve value $R^{(3)}(x)$, but instead at a lower amount. If the granular reserving model is to be used as a negotiation aid when the insurer wants to negotiate a payment at time x, another granular reserving model has to be used: $R_i^{(2)}(x)$, the expected amount of the claim if the claim closes at time x. In this case however, as we saw, objective O2 not longer holds and the insurer cannot use the sum of $R_i^{(2)}(x)$ for each claim $i = 1, \ldots, n(x)$ to estimate its actuarial liability at time x.

3 Application with Real Insurance Data

With real insurance data, we cannot know the impact of covariates on the cost of the claims, or the age of the claim at closure. Some important variables might be missing, the distribution of the age of the claim is not well defined, and the link between the cost of a claim and its age at closure is unknown and can be caused by unobserved variables. In this section, real insurance data will be used to study how to construct an adequate granular reserving model. We now know that if the age of the claim has to be used as a covariate in the reserving model, it means that the best relationship between S and τ_i must be found. Consequently, a new model will be proposed.

3.1 Description of Data

We are using a sample dataset from a specific insurance product of a major Canadian insurance company. We analyze the bodily-injury (BI) coverage. This coverage protects insureds from the financial damages they can cause to a third-party (TP) when they are at least partially (legally) responsible for the loss [2]. BI coverage, compared with other forms of coverage in the Canadian P&C industry, has some particularities:

- The damages are usually paid to a third-party. Each TP involved in an accident has their own damages, and thus can receive separate indemnity. That means that a single accident can have multiple victims. Formally, a TP is called an *exposure*, and each accident is called a *claim*. The micro-level reserving method must estimate the ultimate cost of each exposure.
- The insurance company has a lot of structured and verified information about their own insureds: their date of birth, address, credit score, marital status, address of workplace, etc. However, for a TP, basic information is difficult to obtain and is not automatically verified and validated. Obviously, if large payments are made to a TP, basic information about this TP might be available, but in an unformatted way. In other words, even if the insurer has a lot of information on the TP, and on the injury of the TP, text-mining may be needed to obtain simple information.

[2] Special situations where the insured is not responsible are also possible, but they are not within the score of this paper

- When the damages are paid to a TP, there is often a single and final payment of indemnity. Indeed, throughout negotiations with the TP (or the TP's lawyer) or after a court decision, the insurer will issue a single payment for the damage suffered.

For our modeling study, we will make the following working assumptions:

1. We will only use claims with at least a single payment, and we will suppose that all payments to an exposure are made at the closing date. This is similar to what we illustrated in Figure 1, and also identical to what we supposed in our simulations study. As we will see from the analysis of our dataset, this assumption is realistic.
2. We will not consider the potential dependence between exposures for the same claim. For BI coverage in automobile insurance, many exposures can be injured from the same car accident and we can expect some similarities between all the victims of the same accident. Future analyses where dependence exists between exposures from the same claim, should be considered.
3. Even if we consider all the available covariates of the database, only a few of them are pertinent.
4. We will suppose that the covariates do not change over time, and are fully available at the time the claim is reported to the insurer. For future studies, when this kind of data is available for individual claims reserving methods, dynamic covariates should be included to improve the model, such as the evolution of the injury over time, the hiring of a lawyer by the TP or the amount paid for allocated loss adjustment expenses (ALAE) before the final settlement.

3.1.1 Basic Statistics

We use exposures from claims with a date of loss between 2011 to 2016 (a total of 2000 days). This results in 6,145 exposures. Table 3 shows basic statistics for this database[3]. Indemnity paid corresponds to our interest variable S, which depends on the age of the exposure at closure τ. Only 7.2% of exposures have more than one indemnity payment, and more than 95% of the total amount of the indemnity paid comes from a single payment at the closing date. Figure 7 shows a histogram of S, and the distribution of the age of exposure.

	Average	Std.Err	Minimum	Maximum	25th pct.	50th pct.	75th pct.
Indemnity Paid	69,678	157,755	2	$\approx$ 4,000,000	5,000	25,000	65,000
Age at closure	739.37	599.80	0.00	3,706.00	317.00	557.00	986.00

Table 3: Indemnity paid and age at closure for the BI sample database

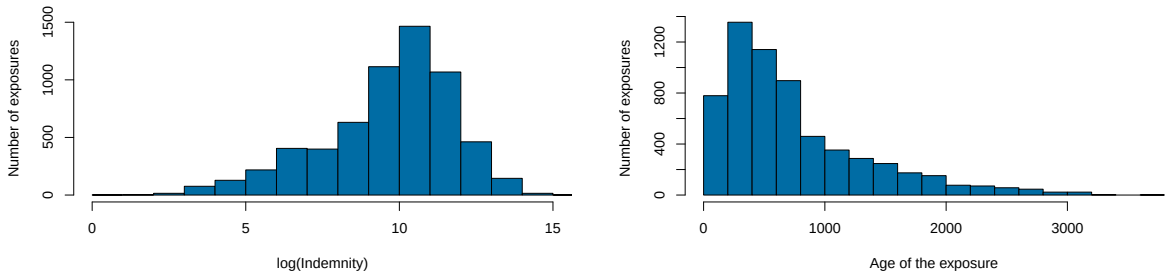

Figure 7: Distribution of the logarithm of the indemnity paid, and the distribution of the age of the exposure for a sample dataset

[3]For confidentially purpose, the maximum value paid for the BI coverage has been rounded.

3.1.2 Covariates

Covariates are available to analyze the cost of the exposure, and the age of the exposure at closure. In Table 4, we refer to reporting delay which represents the time in days between the date of loss and the reporting date, while exposure delay refers to the time in days between the reporting date and the date the claim adjuster opens the exposure for a TP. To detect trends or modification in claim managing, the date of loss is also an important continuous covariate to consider in the modeling.

	Average	Std.Err	Minimum	Maximum	25th pct.	50th pct.	75th pct.
Reporting Delay (RD)	65.87	169.87	0.00	2,343.00	2.00	6.00	28.00
Exposure Delay (ED)	7.94	40.75	0.00	1,432.00	0.00	1.00	4.00

Table 4: Basic statistics for the BI sample database

Figure 8 shows all the other available covariates for the modeling of τ and S, as well as the proportion of each modality. At first sight, some variables might not seem to be predictive, but because we do not have a lot of covariates to include in the models, we kept them to see if they impact.

- Kind of loss (KL);
- How the claim was reported (How);
- Severity of the injury (Sev):
- Fault of the insured (Fault);
- Is the vehicle (of the insured) driveable after the accident (Drv);
- Are the airbags (of the insured's vehicle) deployed during the accident (Airb);
- Total number of exposures on the same claim (NbExp);
- Type of injury (Inj);
- Presence of an ambulance (Amb).

Finally, for the modeling, we separated the database into a training set (70%) and a test set (30%).

3.2 Modeling Approach

For each exposure i, $i = 1, \ldots, n$, to highlight the relation between the time of closure τ_i with the cost to be paid S_i, like we did with the simulations, we are looking to use a model similar to what was defined by equation (3):

$$R_i(x) = E[E[S_i|\tau_i]|\tau_i > x] \quad = \quad \frac{1}{1 - F_{\tau_i}(x)} \int_x^\infty E[S_i|\tau_i = t] f_{\tau_i}(t) dt \tag{5}$$

We do not know the form of $E[S_i|\tau_i]$, or the distribution of the random variable τ_i which corresponds to the age of the exposure at closure. Our modeling strategy is thus a two-step approach, where the age of the exposure at closure (τ_i) will be modeled first, and then the cost of exposure (S_i) will be modeled.

To easily analyze the case reserves at any point of time, an analytical form of $R_i(x)$ would be better. Indeed, in practice, having an analytical equation to compute the case reserve might be important: at any time, a claim adjuster or a claim manager might be interested in the case reserve of a specific exposure or the total liability value of the claim portfolio. As a result, models based on simulations, or on numerical approximations might not be a good solution as they cannot generate a number quickly enough. To obtain an analytic form of (5), we have to put some constraints on the distribution of τ_i as well as on the form of $E[S_i|\tau_i]$.

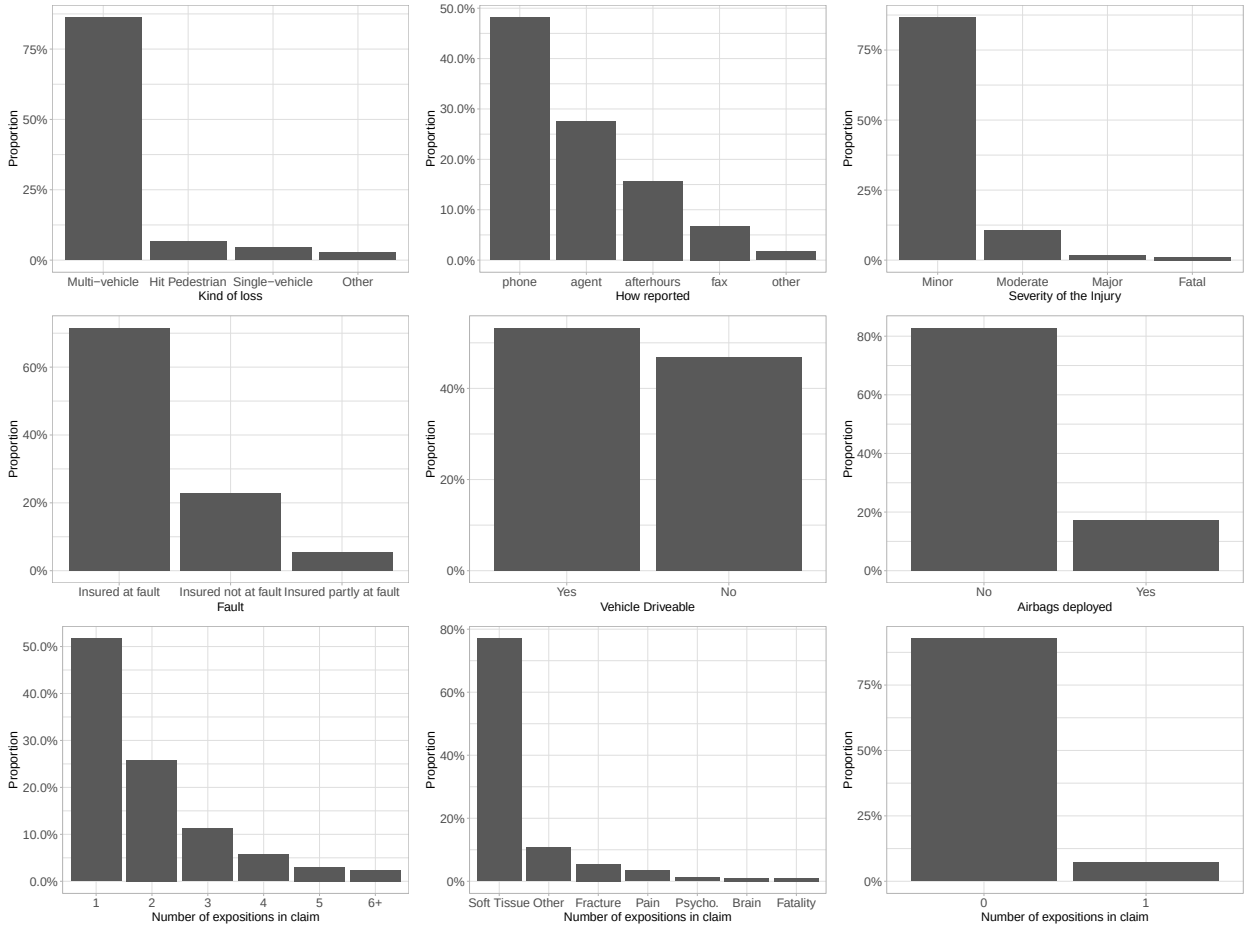

Figure 8: Proportion of different modalities for available categorical variables

3.3 Age of Exposure at Closure (τ)

To model the random variable τ, which represents the age of the exposure at closure, we restricted ourselves to the gamma and the Weibull distribution. Without any covariates, Figure 9 shows the QQ-plots of these three distributions on the training set, with parameters of each distribution estimated by maximum likelihood. A quick look at the QQ-plot tells us that the tail of the distribution does not behave like any of the proposed distributions. To improve the adjustment, covariates are included in the mean parameter of each of the three distributions.

Many approaches can be used to estimate the parameters associated with the covariates. We finally chose to use the random forest approach, with hyper-parameters calibrated by cross-validation. For a default random forest model, the most important variables to explain the age at closure are shown in Figure 10, where we observe that the two most important variables are mainly related to the injury of the victim. This was expected given that a more severe injury will take longer to settle.

3.4 Cost of Exposure (S)

3.4.1 Overview and covariates

To understand the final cost S_i of each exposure i, we need to study the impact of all available covariates on the final indemnity cost of each exposure. To better understand the relationship between these two random variables, a scatter graph is illustrated in Figure 11. Exposures have been grouped together in intervals of 50 days, and each point of the

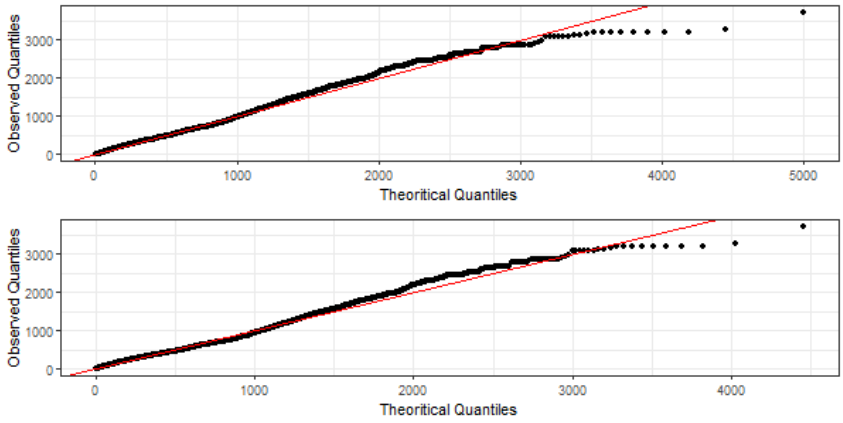

Figure 9: QQ-plots for the modeling of the age of exposure at closure, for the gamma, and Weibull distributions

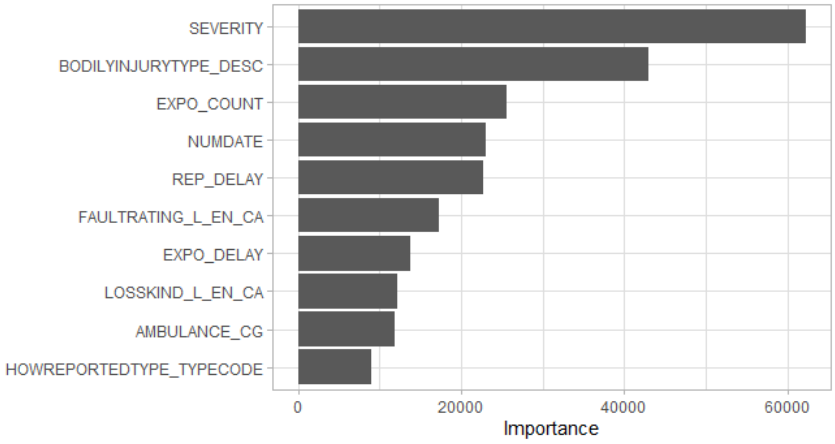

Figure 10: Importance of variables from a random forest model for the age of the exposure at closure (by permutation)

graph corresponds to the average of the indemnity paid for each group (the number of exposures per point is not the same). We see an almost linear relationship between the indemnity paid according to the exposures' age at closure for approximately the first 1250-1500 days. This relationship is not surprising because complex exposures tend take a long time to close. After 1250-1500 days, we see that it is not clear how the age of the claim impacts the final amount.

To understand all available covariates in relation to the cost of each exposure S, another default random forest model was used to show the most important covariates. The result is shown in Figure 12. It shows that the severity and the type of injury are the most important elements to consider to estimate the final cost of each exposure. The age of the exposure at closure τ_i is also an important covariate. Ignoring the age of the exposure when it can provide useful information is similar to the static case reserve approach, and we saw that it cannot give us an adequate reserve total when calculating the actuarial liability. We also saw that using the age of the exposure as a covariate, when cost trends are increasing makes it impossible to accurately estimate each individual case reserve.

Even if we cannot simultaneously satisfy objectives O1 and O2, we saw that a model that uses τ_i as covariate, while adequately capturing the relationship between S_i and τ_i, can generate better results. Consequently, we need to construct a flexible approach to define $E[S_i|\tau_i = t]$.

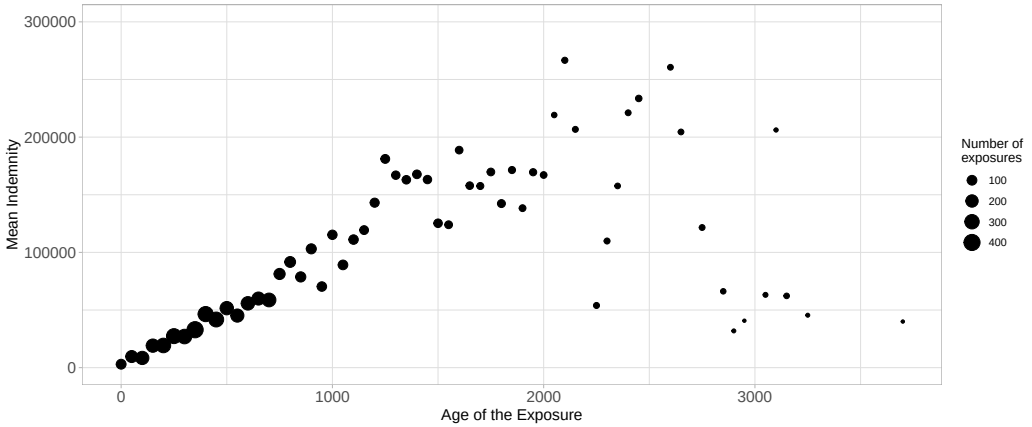

Figure 11: Relation between the indemnity paid at the age of the exposure (at closure)

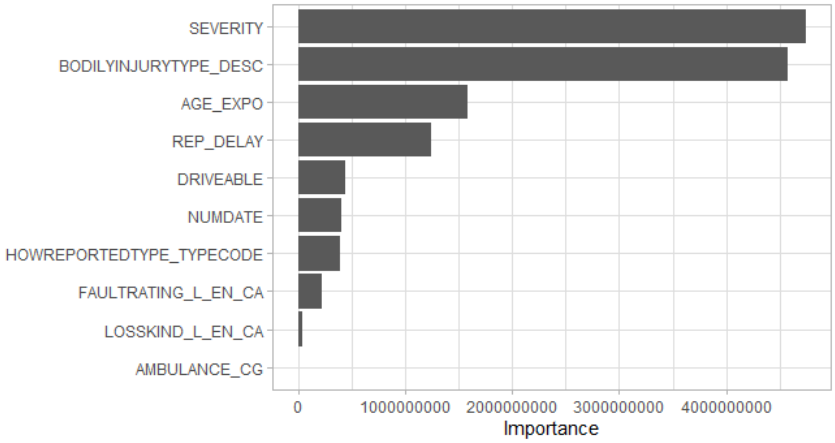

Figure 12: Importance of variables from a random forest model for the cost of the exposure (by permutation)

3.4.2 Parametric Form

According to the form described in equation (5), and to be able to calculate the reserve $R_i(x)$ at any time x, an explicit form of the conditional mean of S_i as a function of x seems desirable. That means that for exposure i, a simple possibility is to use the following model:

$$E[S_i|\tau_i = t] = X_i'\delta + \gamma_1 t$$

where τ_i is the age of the exposure at closure. Other covariates could be included in $X_i'\delta$, while γ_1 would be used to model the linear trend of the age of the exposure at closure like in Section 2. To obtain the reserve at any time x, we simply have to integrate all possible closing times of the exposure. By supposing, for example a specific distribution for the age of the exposure, we can thus obtain:

$$
\begin{aligned}
R_i(x) = E[E[S_i|\tau_i = t]|\tau_i > x] &= \frac{1}{1 - F_{\tau_i}(x;\alpha,\beta)} \int_x^\infty E[S_i|\tau_i = t] f_{\tau_i}(t)dt \\
&= \frac{1}{1 - F_{\tau_i}(x;\alpha,\beta)} (\int_x^\infty (X'\delta + \gamma_1 t) f_{\tau_i}(t)dt) \\
&= \frac{1}{1 - F_{\tau_i}(x;\alpha,\beta)} (X'\delta S_{\tau_i}(x;\alpha,\beta) + \gamma_1 \int_x^\infty t f_{\tau_i}(t)dt) \qquad (6) \\
&= \frac{1}{1 - F_{\tau_i}(x;\alpha,\beta)} (X'\delta S_{\tau_i}(x;\alpha,\beta) + \gamma_1 \frac{\alpha}{\beta} S_{\tau_i}(x;\alpha+1,\beta)) \qquad (7)
\end{aligned}
$$

with $F_\tau(x), S_\tau(x)$, the cumulative and the survival functions respectively. With gamma and Weibull distributions, an analytical form of $\int_x^\infty t f_{\tau_i}(t)dt$ can easily be found. We could add a term τ^2 to the estimation of the model, and we might wonder why we don't also add another term τ^3 or more generally any function $s(\tau)$ to better understand the relationship between the cost of indemnity and the age of the exposure. However, instead of that, we propose using another approach.

3.4.3 Multivariate Adaptive Regression Splines (MARS)

To make our approach more flexible while being able to easily compute $R(x)$, we chose to use multivariate adaptive regression splines (MARS) models (see Duncan et al. [2016] for an application of MARS model in actuarial sciences). The MARS model is a non-parametric technique that extends the linear regression model. The MARS model supposes the following weighted sum for the mean parameter:

$$E[S] = \sum_{i=1}^{p} c_i B_i(y)$$

where each $c_i, i = 1, \ldots, p$ are parameters to be estimated. The term $B_i(y)$ is a function of covariate y. MARS can handle both categorical and numeric covariates. For categorical covariate y, $B_i(y)$ represents a dummy function that equals 1 if y corresponds to a specific value of the covariate, and 0 otherwise. For numeric covariates, such as the age of the exposure at closure τ, $B_i(\tau)$ is a hinge function $h(.)$, with $B_i(\tau) = h(a_i - \tau) = \max(a_i - \tau, 0)$ or $B_i(\tau) = h(\tau - a_i) = \max(\tau - a_i, 0)$ for a specific value of a_i estimated from the data.

Another interesting property of the MARS model is that it can include interaction between covariates. For example, the model can include terms like $B_i(\tau, y) = h(a_i - \tau) \times h(y - b_i)$, where both covariates τ and y are used in a single $B_i(.)$ function. Thus, the relationship between the cost and the age of the claim can be different depending on covariates. This is a useful property of the MARS model because the evolution of cost of the claim over time could vary depending on the characteristics of the insured's injury. As mentioned by Kuhn & Johnson (2013), MARS models have other interesting properties: the model does not need data to be prepared before the modeling, the model automatically calibrates the hinge functions by selecting a_i, which better partitions the data, the MARS model performs automatic variable selection, etc. However, because we are using hinge functions, fitted functions will not be smooth.

More generally, we then use the following structure to model the cost of the exposure:

$$E[S_i|\tau] = \sum_{j=1}^{n_1} c_j^{(1)} B_j(y) + \sum_{j=1}^{n_2} c_j^{(2)} B_j(\tau) + \sum_{j=1}^{n_3} c_j^{(3)} B_j(\tau, y). \qquad (8)$$

Because $B_k(t)$ is a hinge function, it is easy to compute $\int_x^\infty E[S(t)] f_\tau(t)dt$. Indeed, if we exclude $B_j(\tau, y)$, we would have (subscript i is removed for simplicity):

$$
\begin{aligned}
R(x) &= \frac{1}{1-F_\tau(x)} \int_x^\infty E[S|\tau = t] f_\tau(t) dt \\
&= \frac{1}{1-F_\tau(x)} \int_x^\infty \left(\sum_{j=1}^{n_1} c_j^{(1)} B_j(x) + \sum_{j=1}^{n_2} c_j^{(2)} B_j(t) \right) f_\tau(t) dt \\
&= \frac{1}{1-F_\tau(x)} \left(\int_x^\infty \sum_{j=1}^{n_1} c_j^{(1)} B_j(x) f_\tau(t) dt + \int_x^\infty \sum_{j=1}^{n_2} c_j^{(2)} B_j(t) f_\tau(t) dt \right) \\
&= \frac{1}{1-F_\tau(x)} \left(\sum_{j=1}^{n_1} c_j^{(1)} B_j(x) S_\tau(x) + \sum_{j=1}^{n_2} c_j^{(2)} \int_x^\infty B_j(t) f_\tau(t) dt \right). \qquad (9)
\end{aligned}
$$

Each elements of this last equation can be solved easily. For example, we could have:

$$
\begin{aligned}
\int_x^\infty B_k(t) f(t) dt = \int_x^\infty h(a_k - t) f(t) dt &= \int_x^\infty \max(a_k - t, 0) f_\tau(t) dt \\
&= \begin{cases} \int_x^{a_k} (a_k - t) f_\tau(t) dt & \text{if } x < a_k \\ 0 & \text{otherwise} \end{cases} \qquad (10)
\end{aligned}
$$

Depending on the distribution of τ, $R(x)$ can be easily computed for any possible values of x.

3.4.4 MARS model with Bodily Injury Data

If the MARS model does not contain a hinge function, it is reduced to a simple linear model, meaning that parameter estimation is simple. However, when hinge functions are included, all coefficients must still be estimated except values of a, for $h(a - \tau)$ and $h(\tau - a)$, which must be taken from the data. Golub et al. (1979) developed a fitting algorithm for the MARS model that tries all potential values of a. Then, based on the value of each R^2, computed using a Generalized Cross-Validation (GCV) on the training dataset, the best values of a are selected. The *earth* package from R can be used to perform the algorithm and the parameters inference.

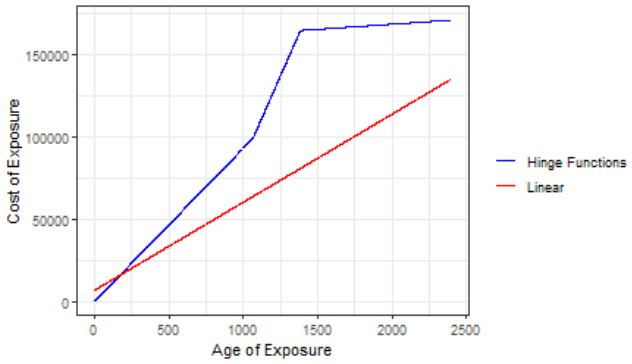

Figure 13: Example of MARS models with hinge functions and linear function on the logarithm of the costs of exposure

To illustrate the way MARS can be used in our case, we directly apply a simple MARS model to the data. Two MARS models with only the age of the exposure at closure have been fitted to the bodily injury data. The first MARS model has a linear trend on the age of the exposure, and the second model uses three hinge functions on the age τ_i. Figure 13 illustrates the link between the age of the exposure and the indemnity cost of both models for specific exposure characteristics. We note the flexibility of the hinge function because it shows different levels of cost increase. This is coherent with what was observed on Figure 11. This result shows the flexibility of the MARS model.

4 Reserving Model

The approaches to modeling the age of the exposure at closure and the cost of the indemnity can both be used to compute individual case reserves. Indeed, by using the case reserve model defined by equation (5), it becomes possible to obtain the case reserve of all exposures i at any time x.

The difficult task in estimating the parameters is to select the hyper-parameters for both the random forest approach for the age of the exposure, and the MARS model for the cost. It is well known that default values for random forest approaches will tend to produce good predictions. However, we still need to calibrate the model if we are looking to find the best model to compute case reserves. Afterwards, for the MARS model, two hyper-parameters are needed: the level of interaction between covariates (*degree*) and the number of terms in the model (*nprune*). For the MARS model, several techniques available in many R packages allow use to easily find those hyper-parameters by cross-validation. However, we did not want to use those techniques because they had to be modified for our specific context. As explained in Section 2.3, our goal is not to compare the final value of the case reserve of the exposure i, $R_i(\tau)$ with the cost of this exposure, S_i, as we would to meet objective O1. We also want to satisfy objective O2, and want the case reserve $R_i(x)$ to be as precise as possible at any time x. Consequently, a more complex calibration procedure was developed.

4.1 Tuning the Model

We found all hyper-parameters of the reserving model in two steps. We first found the hyper-parameters of the random forest model by separating the validation dataset into five binds, where four of the five are used to estimate the parameters and one is used to check for prediction. The hyper-parameters selected from the random forest were those that showed the best predictions.

Our second step was to find the hyper-parameters of the MARS model. Again, we separated the validation dataset into five binds. For each estimation step, the following procedure was done:

1. A random forest approach (with the hyper-parameters we found previously) was fitted. With the fitted value for all exposures, the dispersion parameter of the gamma or the Weibull distribution was estimated by maximum likelihood.
2. For a specific couple (*degree/nprune*), we fitted the MARS model from which estimated values of $c_j^{(1)},\ i = 1, \ldots, n_1$, $c_i^{(2)},\ i = 1, \ldots, n_2$ and $c_i^{(3)},\ i = 1, \ldots, n_3$ from equation (8) were found.
3. Based on equation (9), an automated integration algorithm, flexible enough to be used for all possible values of a for hinge functions $h(a - \tau)$ or $h(\tau - a)$, was developed to compute each case reserve for all of the portfolio's exposures. Obviously, no R packages are available to perform this task and it has to be designed and coded for any values produced by the MARS package.
4. Finally, the reserve value of each exposure i, for any time x, can be computed. Statistics similar to those seen in Table 2 can then be calculated to verify if both objectives are satisfied for each set of hyper-parameters.

Calibration results are shown in Figures 14 and 15 where we suppose a gamma distribution for the age of the exposure at closure. Similar results were obtained for the Weibull distribution.

Figure 14 shows the prediction results related to objective O1. The graph on the left shows the average mean square of prediction (MSEP) of each model calibrated according to the number of degrees of the model (with or without interaction) and the number of terms. We see that models without interaction (degree 1) generally perform better than degree 2 models. We also see that for models without interaction, a number of terms varying between 10 and 20 seems the most appropriate. The right graph, on the other hand, shows the difference between the amounts paid and the individual case reserve at the time of payment. We see that a model with 12 terms, without interaction, seems to be the most accurate.

The criteria related to the O2 objective are shown in Figure 15. The graph on the left shows average difference between the calculated total reserve and the actual payments. As opposed to the results of objective O1, models with interaction seems slightly more precise than degree 1 models. By restricting ourselves to degree 1 models, we see that the models with 12 to 17 terms are the most suitable . Finally, the graph on the right shows the proportion of days for which the reserve was plus or minus 10 % of the ultimate amount. We see that the models with interactions and with many terms (between 20-25) seem the best.

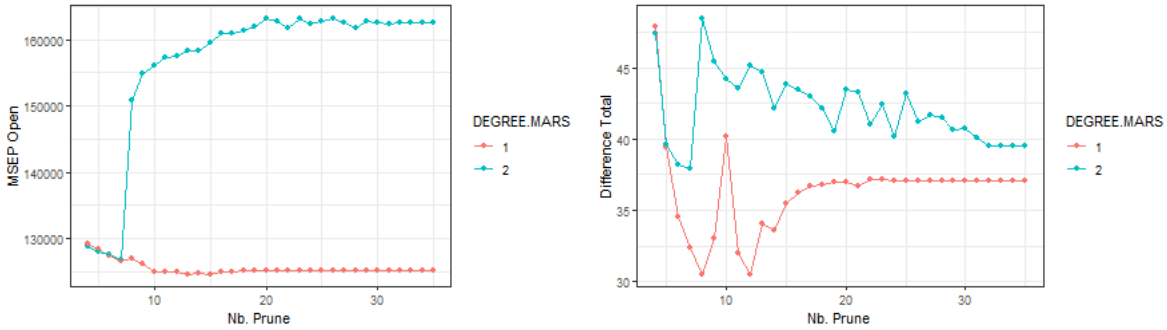

Figure 14: Statistics for Objective O1 of the model based on the number of terms and the number of degrees of the MARS model - with a gamma distribution for τ

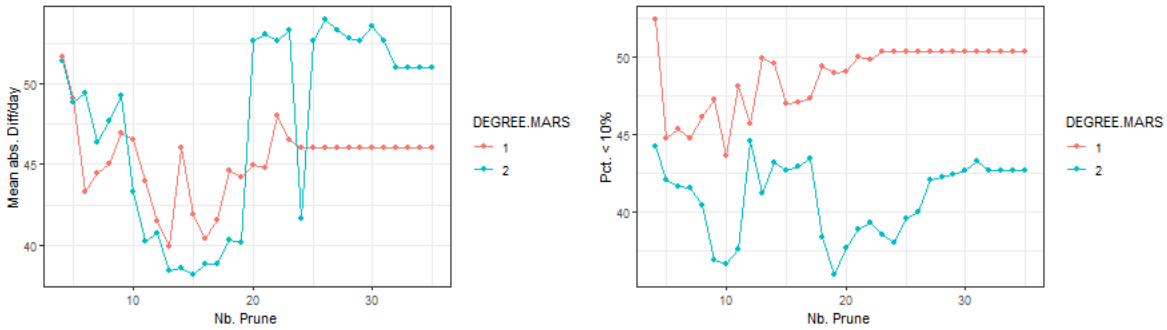

Figure 15: Statistics for Objective O2 of the model based on the number of terms and the number of degrees of the MARS model - with a gamma distribution for τ

We have to select a model that has two conflicting objectives and it is unsurprising to see that some models perform better under a criterion related to the O1 objective while others are better for the O2 objective. We will continue with the model without interaction that has 12 terms: although it is slightly worse than the degree 2 models for the O2 objective, it is much better than these models for the O1 objective. In future research, it might be relevant to define a single selection criterion.

4.1.1 Analyzing the Reserving Model

For the whole training dataset, a shape parameter of 1.843 was estimated with the gamma distribution based on the prediction of the random forest model, while a shape parameter of 1.429 for a Weibull distribution has been estimated. A MARS model with 12 parameters without interaction applied to the cost of exposure generates the estimators shown in Table 5. We can see that four hinge functions are used in the MARS model. Covariates describing the severity of the injury and the type of injury are also included in the model.

Figure 16 shows the sum of all case reserves of open exposures for each calendar day. Case reserves based on a static reserving model where the age of the exposure is not used in the modeling are also shown. On the left graph, we see the results from the training dataset. We see that both the gamma and the Weibull models are similar to the real values. On

Parameter	Coefficient	Parameter	Coefficient
(Intercept)	538,598	-	.
Sev(Moderate)	-259,350	h(τ-2513)	45,643
Sev(Minor)	-293,620	h(τ-2543)	-62,479
Inj(Soft Tissue)	-88,297	h(τ-2631)	16,725
Inj(Pain)	-56,23	h(2543-τ)	-66
Inj(Other)	-57,407	-	.
Inj(Fracture)	-42,259	-	.
Inf(Fatality)	-203,169	-	.

Table 5: Estimated parameters and hinge functions for the MARS model (degree = 1, nb.prune = 12)

the other hand, the static reserve underestimates the reserve for each calendar day. In the graph on the right, the same curves are shown for the test dataset. The static reserve model is again inappropriate. The dynamic reserving models from both the gamma and the Weibull distributions perform better even though the cost is underestimated before day 2000.

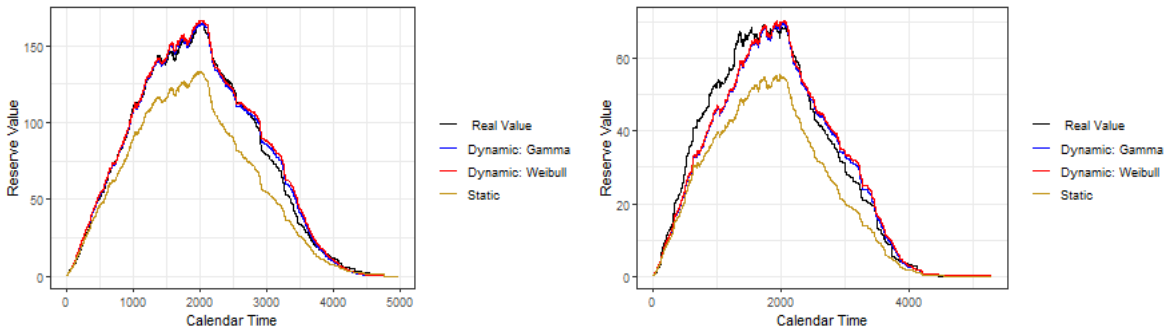

Figure 16: Total reserve for each calendar day (left: training dataset, right: test dataset)

Figure 17 shows the cumulative reserve value at closure for the training dataset (left) and the test dataset (right). As shown in the simulations part of the paper, dynamic models that use the age of the exposure at closure as a covariate cannot expect to satisfy objective O1. On the other hand, the static reserving model is accurate for the training dataset, and almost accurate for the test dataset. Again, this was expected from the simulations study. Finally, the mean individual case reserve for each age is shown in Figure 18. Both dynamic models seem to generally fit the individual reserve over time.

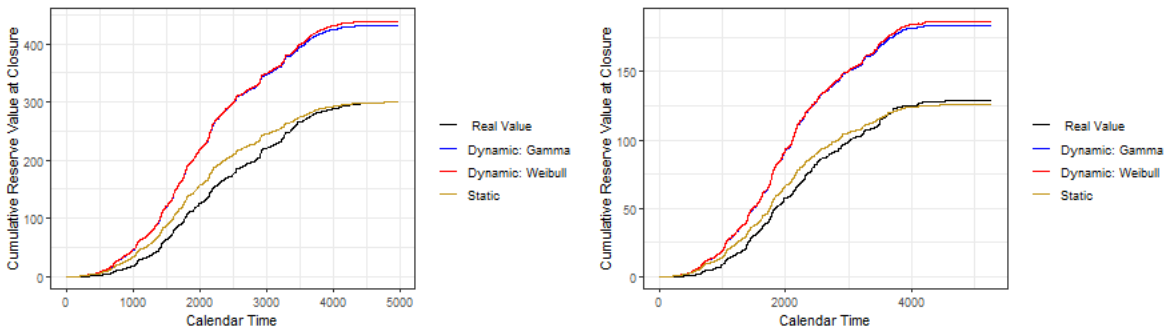

Figure 17: Cumulative reserve value at closure, for each calendar day (left: training dataset, right: test dataset)

Table 6 shows other statistics to compare the models, and verify if both objectives O1 and O2 are satisfied.

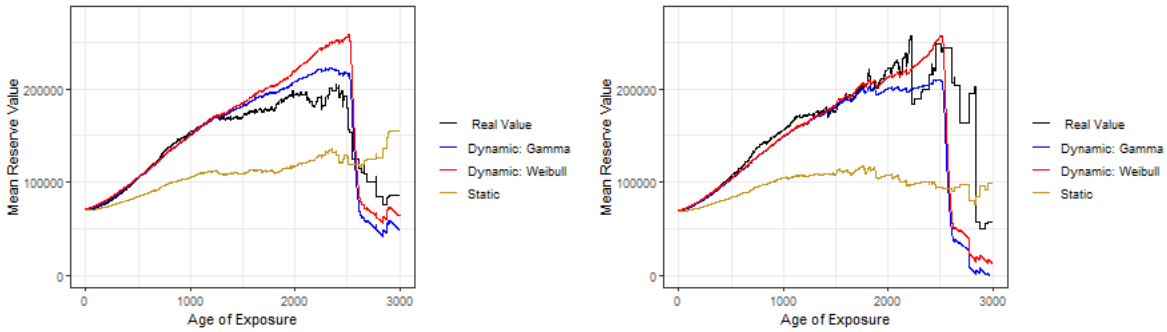

Figure 18: Mean individual case reserve for each age, in days (left: training dataset, right: test dataset)

	Objective O1			Objective O2 (by day)			
	Diff (%)	MSEP (Close)	MSEP (Open)	Diff (%)	+/- 10%	+/- 5%	+/- 1%
Gamma	43.33%	163,077	160,615	11.58%	44.40%	25.51%	7.51%
Weibull	45.61%	162,948	160,624	12.94%	41.82%	24.34%	6.57%
Static	-2.16%	164,432	164,432	-28.80%	0.18%	0.15%	0.00%

Table 6: Statistics for each claim i at time τ_i for different reserve models (test dataset)

4.2 Reducing the Effect of the Age of Exposure

We have seen in the simulations that in order to simultaneously satisfy the objectives O1 and O2 are simultaneously satisfied, it is necessary to minimize the effect of the age of exposure at closure in the model. In the MARS models used in the previous section, covariates were used to describe the exposure and a covariate that uses the age of the exposure was included directly in model $E[S_i|\tau]$. As a result, a joint estimate of all $c_j^{(1)}, j = 1, \ldots, m_1$ (parameters not linked with the age of the exposure), $c_j^{(2)}, j = 1, \ldots, m_2$ (covariates associates with the age) and $c_j^{(3)}, j = 1, \ldots, m_2$ (covariates with interaction with the age) has been performed.

Although the age of exposure at closure is an important variable in the modeling, it seems reasonable to believe that there is no real causal relationship between the time at closure and the final cost of the exposure. Of course, if an exposure takes a long time to close, it will on average cost a lot more than an exposure that was closed quickly. But the increase in the cost of the exposure is not caused by the time itself, as an interest rate would be. Indeed, we think that if we observe an increasing trend, it is simply because it takes more time to close complex exposures that cost, on average, a lot more than other type of exposures. Thus, it is reasonable to believe that the age of the exposure at closure is a proxy for the complexity of the exposure. If the insurer could gather more precise information on each exposure, it will help identify complex exposures, which in turn will diminish the importance of the time to closure in the model. A situation similar to what was observed in Section 2.2 could then emerge.

The insurer's motivation is thus to minimize the impact of the age of the exposure on the estimate of the case reserve, by using all the available information about each exposure. However, apart from using other covariates in the modeling, another approach that seeks to minimize the impact of the age of the exposure seems relevant to us. Indeed, we could first model $E[S_i|\tau]$ with only covariates that do not depend on the age of the exposure:

$$E[S_i] = \sum_{j=1}^{n_1} c_j^{(1)} B_j(y) \tag{11}$$

With the estimated parameters of $c_j^{(1)}, j = 1, \ldots, m_1$, a first estimator of S_i, $\widehat{S}_i$, can be calculated. The idea is to subsequently obtain the residuals of each exposure, $W_i = S_i - \widehat{S}_i$. Secondly, the approach is to fit a model on the

residuals by including only covariates linked with the age of exposure at closure. More formally, we will have the following result:

$$E[W_i|\tau] = \sum_{j=1}^{n_2} c_j^{(2)} B_j(\tau) + \sum_{j=1}^{n_3} c_j^{(3)} B_j(\tau, y). \tag{12}$$

where $B_j(\tau)$ is function of τ, and $B_j(\tau, y)$ is a function of τ and y.

To compare with the results obtained from the first approach, MARS models will be used for these two steps. It is obvious that a two-step MARS approach is less precise in terms of the fit of S_i. However, our objective is not to obtain the best model to estimate $S_i|\tau$, but to have the best reserve model $R(x)$, which is based on $S_i|\tau$. Indeed, we seek to meet objectives O1 and O2. Thus, with the model based on residuals, the individual reserves at time x, $R(x)$, are calculated using the same logic as was expressed in equation (9).

A cross-validation approach similar to that described in Section 4.1 was performed to obtain the hyper-parameters of the model on S_i and of the model on W_i. Estimated parameters from the two-step approach are shown in Table 7.

Parameter	Coefficient	Parameter	Coefficient
(Intercept)	526,117	(Intercept)	100,084
Sev(Moderate)	-243,628	h(τ-2387)	1,368
Sev(Minor)	-294,993	h(τ-2543)	-4,793
Inj(Fatality)	-301,713	h(τ-2631)	3,272
Inj(Psychological)	-178,384	h(2543-τ)	55
Inj(Pain)	-157,926	-	.
Inj(Fracture)	-151,333	-	.
Inf(Other)	156,966	-	.

Table 7: Estimated parameters and hinge functions for two consecutive MARS models

The results of the best model, in relation to objectives O1 and O2, are shown in Table 8. The two-step approach seems to be much superior to the previous MARS approach because the statistics associated with the objectives O1 and O2 are better. Indeed, the sum of each case reserve at the time of closure, compared to what was really paid, is only 25% higher, while the single MARS model showed a 45% difference. We also observe a slightly better MSEP for both closed and open claims. For more than 40% of days, the sum of individual case reserves were within 5% of the true liability value, compared to 25% for the previous model. This shows that although the age of the exposure is necessary in the modeling, at the risk of moving to an inadequate static reserve, it is important to minimize its impact to obtain precise values for the individual reserves.

	Objective O1			Objective O2 (by day)			
	Diff (%)	MSEP (Close)	MSEP (Open)	Diff (%)	+/- 10%	+/- 5%	+/- 1%
Gamma	25.32%	161,390	159,506	10.54%	54.79%	42.95%	11.55%
Weibull	28.66%	160,960	159,489	9.99%	57.65%	44.22%	9.38%

Table 8: Statistics for each claims i at time τ_i for different reserve models (test dataset)

5 Conclusion

The development of micro-reserve methods has increased in recent years and many of these approaches have been developed as alternatives to runoff triangles in the annual calculation of actuarial liabilities. Insurance

companies quickly realized that other uses could be found for these granular reserving models. In this paper, we show that we should be careful before using micro-reserve models for all situations. When available covariates cannot explain the trend in the average cost, it is impossible to have a granular reserving model that provides [O1] an adequate prediction of the amounts paid for each claim, and [O2] an adequate value for the total actuarial liability.

We used simulated data to carefully explain the situation. Using real data, we then proposed a new model based on MARS theory that allows us to generalize the link between the age of the exposures at closure and the cost of the exposure. The hinge functions used in the MARS model allow us to generate an analytical form of the computation of each individual case reserve of the portfolio at all times, which is often essential to the regular use granular reserve models in practice. Hyper-parameters of the model were found by cross-validation with a new automated integration algorithm that was flexible enough to be used for all possible values of a for hinge functions $h(a-\tau)$ or $h(\tau-a)$. However, even if the fit of the MARS model is improved compared to other models, we show that objectives O1 and O2 still cannot be satisfied simultaneously. To reduce the impact of the age of the exposure, a generalization of the MARS reserving model was then proposed, where the age of the exposure is only used on the residuals of a first MARS model based on the characteristics of the exposure. We show that this correction of the MARS model produces better results, and it comes closer to satisfying objectives O1 and O2.

There is much more research on the MARS model for claims reserving, such as generalizing the underlying distribution of the MARS model, or including dependence between exposures from the same accident, that should be studied in the future. For other types of claims than bodily-injury claims, partial information gathered throughout the life of the claim should also be used to minimize the use of the age of the exposure in the model. Other prediction measures to deal with objectives O1 and O2 should also be considered.

Acknowledgements

The author would like to thank Mathieu Pigeon for his comments, as well as Andra Crainic, Alexandre LeBlanc and Veronika Gousseva from Co-operators General Insurance Company for their advice. Finally, the author would like to thank the Co-operators Chair in Actuarial Risk Analysis, and the Natural Sciences and Engineering Research Council (NSERC) of Canada for their financial support.

References

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

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

[3] Charpentier, A. & Pigeon, M. Macro vs. Micro methods in non-life claims reserving (an econometric perspective). *Risks*, 4(2):12, 2016.

[4] Duncan, I., Loginov, M., & Ludkovski, M. Testing alternative regression frameworks for prdictive modelling of health care costs. *North American Actuarial Journal*, 20(1), 65-87, 2016.

[5] Duval, F., & Pigeon, M. Individual loss reserving using a gradient boosting-based approach. *Risks*, 7(3), 79. 2019.

[6] Friedland, J. Estimating unpaid claims using basic techniques. *Casualty Actuarial Society*, vol. 201. 2010.

[7] Golub, G. H., Heath, M., & Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2), 215-223, 1979

[8] Huang, J., Qiu, C. & Wu, X. Stochastic loss reserving in discrete time: individual vs. aggregate data models *Communications in Statistics - Theory and Methods*, 44, 2180-2206. 2015

[9] Kuhn, M. & Johnson, K. Applied Predictive modelling. *New York, NY: Springer New York*, 2013

[10] Norberg, R. A contribution to modelling of IBNR claims. *Scandinavian Actuarial Journal*, 155-203, 1986

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

[12] Wuthrich, M. V. Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, vol. 2018, no 6, p. 465-480. 2018

[13] Wuthrich, M. V., & Merz, M. Stochastic claims reserving methods in insurance *John Wiley & Sons.*, 2008