# BEYOND THE TWEEDIE RESERVING MODEL: THE COLLECTIVE APPROACH TO LOSS DEVELOPMENT

Michel Denuit
Institute of Statistics, Biostatistics and Actuarial Science
Université Catholique de Louvain (UCL)
Louvain-la-Neuve, Belgium

Julien Trufin
Department of Mathematics
Université Libre de Bruxelles (ULB)
Bruxelles, Belgium

July 6, 2017

**Abstract**

This paper proposes a new loss reserving approach, inspired from the collective model of risk theory. According to the collective paradigm, we do not relate payments to specific claims or policies but we work within a frequency-severity setting, with a number of payments in every cell of the run-off triangle, together with the corresponding paid amounts. Compared to the Tweedie reserving model, that can be seen as a compound sum with Poisson-distributed number of terms and Gamma-distributed summands, we allow here for more general severity distributions, typically mixture models combining a light-tailed component with a heavier-tailed one, including inflation effects. The severity model is fitted to individual observations and not to aggregated data displayed into run-off triangles with a single value in every cell. In that respect, the modeling approach appears to be a powerful alternative to both the crude traditional aggregated approach based on triangles and the extremely detailed individual reserving approach developing each and every claim separately. A case study based on a motor third party liability insurance portfolio observed over 2004-2014 is used to illustrate the relevance of the approach proposed in this paper.

*Key words and phrases:* general insurance, reserving, collective model, mixture models, GAMLSS, solvency evaluation.

# 1 Introduction

In Property and Casualty (P&C) insurance, claims often need several years to be settled. Meanwhile, insurers have to build reserves representing their estimate of outstanding liabilities for claims that occurred on or before the valuation date. Reserving calculation has traditionally been performed on the basis of aggregated data summarized in run-off triangles with rows corresponding to accident years and columns corresponding to development years. Such data exhibit three dimensions: for each accident (or occurrence, or underwriting) year $i$ and development period $j = 1, 2, \ldots$, we read in cell $(i, j)$ inside the triangle the total amount paid by the insurer in calendar year $c = i + j - 1$ for claims originating in year $i$. Techniques dealing with such aggregated triangular arrays of data with a single value in every observed cell go back to the pre-computer era, at a time where the available computing resources, data storage facilities and statistical methodologies were extremely limited. The Chain-Ladder (CL, in short) approach is certainly the most popular technique falling in this category. See for instance Kaas et al. (2008) for an introduction and Wuthrich and Merz (2008) for a detailed account of the topic.

Wuthrich (2003) proposed a reserving model assuming that the amount in cell $(i, j)$ is Tweedie distributed with a mean expressed in function of factors related to accident year $i$ and development $j$. As the Tweedie distribution is a Compound Poisson one, this means that the number of payments in the cell is Poisson distributed, and independent of the amounts of each payment, assumed to be independent and Gamma distributed. This specification is thus intuitively appealing. Moreover, the Tweedie distribution is in the GLM setting, which facilitates the numerical treatment of the data. In particular, the loss prediction based on aggregate data in cell $(i, j)$, i.e. the total payment corresponding to that cell, coincide with the result based on the knowledge of each and every individual payment (recall that in the GLM setting, all responses sharing the same explanatory variables, $i$ and $j$ here, can be summed together without affecting the point estimates, provided the weights are adapted accordingly). See also Alai and Wuthrich (2009), Peters et al. (2009) as well as Rosenlund (2010) for further results.

Boucher and Davidov (2011) proposed a double GLM to let the mean as well as the variance depend on the effects $i$ and $j$. These authors justify this dispersion modelling by the existence of two opposite effects:

- in general, most of the claims are reported early in development so that the frequency component has a decreasing trend throughout developments;
- the average cost per claim often increases through the development years so that the severity component exhibits an increasing trend.

In such a case where frequencies and severities have opposite trends, models with constant dispersion are prone to errors. The approach proposed in this paper is in line with the separate modelling of the frequency and severities, proposed by Boucher and Davidov (2011) as a possible remedy.

Of course, this equivalence between the aggregated approach and the detailed one, based on individual paid amounts, only holds in the Tweedie setting. Even if the Poisson assumption for the number of payments appears to be rather reasonable, the Gamma specification may lead to underestimation of the tail of the loss distribution. This is why we extend here the Tweedie approach to alternative distributions for the amounts paid by the insurer. Typically, we use mixture models combining a light-tailed distribution with a heavier-tailed one. The parameters of this two-component mixture model are explained by means of $i$, $j$ and $c$ effects in a GAMLSS setting. Notice that Gamma distributed amounts are also used in Schiegl (2015) whereas Schiegl (2002) uses Pareto and Negative Exponentially distributed amounts.

As for the three-dimensional (3D) reserving model proposed by Schiegl (2015), the basis of the loss reserving model used in the present paper is the collective model of risk theory (see e.g. Chapters 2-3 in Kaas et al., 2008, for an introduction). Compared to the 3D model, counts are only indexed by two effects in the present paper, as well as amounts. Following the arithmetic and geometric approaches to claim reserving (see e.g. Section 10.3.2 in Kaas et al., 2008), frequencies include accident year $i$ and development $j$ effects, whereas amounts are explained by means of development $j$ and calendar year $c$ effects. The 3D model is more flexible in that it allows to split the total number of claim that are still active during a given calendar year according to their year of occurrence and their year of declaration. Here, we focus on the number of payments in cell $(i, j)$, without reference to characteristics of the claims from which these payments originate. This means that these payments correspond to claims at various stages of development. This heterogeneity is taken into account by means of a mixture model combining light-tailed and heavier-tailed severity distributions. Our approach is closely related to Schiegl (2002), except that we model the number of payments and not the number of open claims (in order to avoid zero payments for some open claims at given developments, which complicates the mixed model for the severities).

As it can be seen from the references listed in the preceding paragraphs, the idea of modeling claim frequency and severity separately from two triangles is not new. Similarly, the use of mixture distributions for claim severity has become rather standard in P&C insurance, in order to capture both the moderate and large values. The contribution of this paper lies in the combination of these methods to propose a new approach to loss reserving calculations based on individual severity data, not aggregated ones. Indeed, mixture distributions cannot be fitted to aggregated severity data displayed into a triangle, with a single value in every observed cell. Individual severities are needed for that purpose. Contrarily to individual loss reserving models which develop each and every claim over time, we work here with detailed severity data to fit a collective model inside each cell $(i, j)$ without tracking individual claim development over time. Under the compound Poisson assumption, these cell-specific distributions can easily be aggregated by convolution to recover the outstanding claim distribution. This collective approach to loss development appears to outperform the Poisson-Gamma, or Tweedie model on the data used in our numerical illustrations (because of the mix between moderate and large severities). It offers enough flexibility to provide accurate answers to most problems encountered in practice as demonstrated in the numerical part of this work.

The remainder of this paper is organized as follows. In Section 2, we present the data set used for the numerical illustrations. Section 3 is devoted to the collective model for insurance losses. In Section 4, we perform reserve calculations and we compare the output with results based on CL and Tweedie loss reserving models. The final Section 5 concludes the paper.

# 2 Notation and data

## 2.1 Accident and development indices

We assume that we have $n$ years of observations. Accident years range from $i = 1$ to $n$ and developments from $j = 1$ to $n$. These data fill a triangle: accident year $i$ is followed from development $j = 1$ (corresponding to the accident year itself) to the last observed development $n - i + 1$ (corresponding to the last calendar year $n$ for which observations are available, located along the last diagonal of the triangle).

Henceforth, we denote as $\omega$ the time needed to settle all the claims occurred during a given accident year $i$, i.e. these claims are closed in calendar year $i + \omega - 1$ at the latest. For business lines with long developments, some claims for accident year 1 may still be open in calendar year $n$ so that $\omega > n$. Precisely, $\omega = n$ if all claims of the first accident year are settled at the end of the observation period. If not, $\omega > n$ and we must introduce a tail factor to account for the last developments before final settlement.

## 2.2 The data

The approach proposed in this paper is applied on a data set extracted from the motor third party liability insurance portfolio of an insurance company operating in the EU. The observation period consists in calendar years 2004 till 2014. The available information concerns accident years 2004 to 2014 so that we have observed developments $j$ up to $n = 11$. Henceforth, we let $i$ belong to $\{1, \ldots, n\}$ or range between $i = 2004$ and 2014, in order to make the numerical results easier to interpret.

There are $52,155$ claims in the data set. Among them, $4,023$ claims are still open at the end of the observation period. Table 1 presents the information available for two claims of the database. Claim #$16,384$ corresponds to an accident occurred in 2009 that has been reported during the same calendar year. Payments have been made in years 2009 to 2013, but no payment has been recorded for 2014. We note that in our approach, all the payments related to the same claim are aggregated over the calendar year. At the end of the observation period, claim #$16,384$ is still open. Claim #$20,784$ corresponds to an accident occurred in 2010 that has been reported during the same calendar year. A payment has been made in 2010, there was no payment in 2011, and the claim has been closed in 2011, one year after its reporting to the insurer. Notice that in our data set, the declaration of a claim corresponds to the first time there is a payment or a positive case estimate for that claim. Hence, late reporting (i.e. at lags 3-4) is due here to the definition adopted for reporting as motor insurance contracts typically impose that policyholders rapidly file the claim against the company.

Table 2 displays descriptive statistics for the payments per accident year $i$ and development lag $j$. We can see there the number of payments, the proportion of claims with no payment, the mean of the payments as well as the standard deviation and skewness per accident year $i$ and lag $j$. Table 2 shows that the standard deviation is often about twice the mean while the large skewness values suggest highly asymmetric distributions. We also see there the typical increase of mean payments with development $j$, together with the corresponding decrease in the number of payments as described in the introduction. These

| Event | No | Year | Amount |
|---|---|---|---|
| Occurrence | 16,384 | 2009 | - |
| | 20,784 | 2010 | - |
| Declaration | 16,384 | 2009 | - |
| | 20,784 | 2010 | - |
| Payments | 16,384 | 2009 | 5,022 |
| | 16,384 | 2010 | 67,363 |
| | 16,384 | 2011 | 903 |
| | 16,384 | 2012 | 6,295 |
| | 16,384 | 2013 | 13,850 |
| | 16,384 | 2014 | 0 |
| | 20,784 | 2010 | 1,605 |
| | 20,784 | 2011 | 0 |
| Closure | 16,384 | Not settled | - |
| | 20,784 | 2011 | - |

Table 1: Information available for claims No $16,384$ and No $20,784$ in the data set. Claim No $16,384$ is still open end of year 2014.

opposite effects are known to invalidate the Tweedie specification with constant dispersion parameters (GLM setting).

# 3 Collective model for losses

## 3.1 Compound sum decomposition

All claims from accident year $i$ are modelled in a collective way. Following Wuthrich (2003), we decompose the total payment $X_{ij}$ at development $j$ (i.e. in calendar year $i+j-1$) for these claims into the compound sum

$$X_{ij} = \sum_{k=1}^{N_{ij}} X_{ijk}, \quad j = 1, 2, \ldots, \omega,$$

where $N_{ij}$ is the number of payments made at lag $j$ for claims originating in accident year $i$, and the $X_{ijk}$s denote the corresponding amounts. All these random variables are assumed to be mutually independent. For given $i$ and $j$, the random variables $X_{ij1}, X_{ij2}, \ldots$ are assumed to be identically distributed. Notice that here, payments related to individual policies are not tracked, only payments for the collective are modelled, and all payments related to the same claim are aggregated in $X_{ijk}$.

*Remark* 3.1. Notice that the proposed approach does not explicitly include the number of reported, or open claims. Denoting as $M_{ij}$ the number of open claims at lag $j$ originating in accident year $i$, and as $Y_{ijl}$ the corresponding yearly payments, $l = 1, 2, \ldots, M_{ij}$, we must account for the cases where no payments have been made for an open claim at that lag,

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2004 | | | | | | | | | | | |
| Num. pay. | 2,848 | 1,459 | 236 | 124 | 68 | 39 | 18 | 18 | 12 | 8 | 7 |
| Mean | 1,133 | 1,877 | 2,713 | 4,349 | 4,446 | 9,894 | 16,765 | 4,422 | 18,072 | 12,314 | 21,263 |
| Std. dev. | 2,378 | 5,317 | 4,861 | 9,405 | 7,918 | 26,576 | 27,037 | 9,768 | 24,203 | 14,436 | 50,490 |
| Skewness | 12.423 | 11.878 | 4.157 | 4.472 | 2.948 | 5.028 | 1.967 | 3.476 | 2.015 | 1.136 | 2.039 |
| 2005 | | | | | | | | | | | |
| Num. pay | 3,001 | 1,492 | 207 | 97 | 53 | 42 | 24 | 21 | 11 | 11 | |
| Mean | 1,112 | 1,659 | 3,168 | 5,455 | 5,132 | 14,882 | 25,781 | 8,997 | 4,230 | 1,347 | |
| Std. dev. | 1,847 | 2,932 | 6,081 | 18,278 | 10,270 | 41,070 | 77,046 | 19,947 | 2,817 | 883 | |
| Skewness | 4.113 | 5.509 | 3.709 | 8.387 | 4.089 | 4.353 | 3.135 | 3.464 | 0.413 | 0.241 | |
| 2006 | | | | | | | | | | | |
| Num. pay | 3,007 | 1,659 | 268 | 117 | 61 | 41 | 21 | 10 | 10 | | |
| Mean | 1,164 | 1,624 | 5,799 | 4,494 | 7,287 | 6,055 | 6,141 | 4,688 | 12,205 | | |
| Std. dev. | 2,972 | 2,932 | 49,737 | 7,632 | 22,190 | 12,682 | 9,173 | 4,594 | 26,907 | | |
| Skewness | 23.651 | 6.020 | 15.840 | 3.104 | 4.485 | 3.248 | 1.624 | 0.506 | 2.440 | | |
| 2007 | | | | | | | | | | | |
| Num. pay | 3,246 | 1,893 | 322 | 170 | 79 | 48 | 24 | 16 | | | |
| Mean | 1,159 | 1,905 | 2,679 | 3,500 | 7,401 | 8,243 | 12,140 | 13,148 | | | |
| Std. dev. | 2,258 | 4,984 | 6,159 | 5,831 | 13,989 | 14,717 | 20,625 | 22,292 | | | |
| Skewness | 9.382 | 13.781 | 6.688 | 3.093 | 3.029 | 2.872 | 2.731 | 1.921 | | | |
| 2008 | | | | | | | | | | | |
| Num. pay | 3,574 | 1,816 | 304 | 125 | 71 | 37 | 22 | | | | |
| Mean | 1,104 | 1,720 | 2,189 | 4,203 | 4,611 | 7,775 | 6,310 | | | | |
| Std. dev. | 1,837 | 3,644 | 3,524 | 8,791 | 9,908 | 12,249 | 7,275 | | | | |
| Skewness | 5.521 | 7.533 | 3.851 | 4.451 | 4.774 | 2.352 | 0.983 | | | | |
| 2009 | | | | | | | | | | | |
| Num. pay | 3,545 | 1,877 | 300 | 131 | 90 | 51 | | | | | |
| Mean | 1,142 | 1,919 | 3,981 | 4,379 | 6,896 | 9,129 | | | | | |
| Std. dev. | 1,926 | 5,710 | 19,797 | 11,584 | 17,446 | 18,474 | | | | | |
| Skewness | 4.610 | 18.270 | 14.896 | 7.229 | 5.379 | 3.265 | | | | | |
| 2010 | | | | | | | | | | | |
| Num. pay | 2,874 | 2,072 | 338 | 161 | 75 | | | | | | |
| Mean | 1,663 | 1,984 | 3,637 | 5,147 | 14,935 | | | | | | |
| Std. dev. | 4,012 | 5,832 | 11,419 | 13,420 | 60,912 | | | | | | |
| Skewness | 18.229 | 16.910 | 11.563 | 5.037 | 5.849 | | | | | | |
| 2011 | | | | | | | | | | | |
| Num. pay | 2,777 | 1,930 | 327 | 119 | | | | | | | |
| Mean | 1,601 | 1,982 | 2,441 | 5,171 | | | | | | | |
| Std. dev. | 2,333 | 4,004 | 4,119 | 14,476 | | | | | | | |
| Skewness | 5.628 | 7.586 | 3.607 | 4.742 | | | | | | | |
| 2012 | | | | | | | | | | | |
| Num. pay | 2,860 | 1,749 | 282 | | | | | | | | |
| Mean | 1,716 | 2,328 | 4,390 | | | | | | | | |
| Std. dev. | 4,587 | 10,085 | 31,803 | | | | | | | | |
| Skewness | 36.917 | 31.363 | 16.083 | | | | | | | | |
| 2013 | | | | | | | | | | | |
| Num. pay | 2,924 | 1,844 | | | | | | | | | |
| Mean | 1,637 | 2,230 | | | | | | | | | |
| Std. dev. | 4,120 | 11,414 | | | | | | | | | |
| Skewness | 31.519 | 35.894 | | | | | | | | | |
| 2014 | | | | | | | | | | | |
| Num. pay | 2,723 | | | | | | | | | | |
| Mean | 1,662 | | | | | | | | | | |
| Std. dev. | 2,360 | | | | | | | | | | |
| Skewness | 7.018 | | | | | | | | | | |

Table 2: Descriptive statistics for payments per accident year $i = 2004, \ldots, 2014$ and lag $j = 1, \ldots, 11$, namely the number of payments (Num. pay.), the mean of the payments as well as the standard deviation and skewness.

i.e. $Y_{ijl}$ must have a probability mass at 0. This is why we work here with the number of payments

$$N_{ij} = \sum_{l=1}^{M_{ij}} \mathrm{I}[Y_{ijl} > 0].$$

Instead of modelling $M_{ij}$ and the probability mass of $Y_{ijl}$ at zero, the approach proposed in this paper directly targets $N_{ij}$. As $\mathrm{E}[N_{ij}] = \mathrm{E}[M_{ij}]\mathrm{P}[Y_{ijl} > 0]$ under suitable independence assumptions, the number of open claims implicitly appears in $\mathrm{E}[N_{ij}]$. The amounts $X_{ijk}$ then correspond to the $k$th aggregate yearly payment, among those open claim with a positive amount paid at lag $j$.

## 3.2 Frequency component

The observed counts $N_{ij}$ are displayed in Table 2. In line with the classical CL model, we use the multiplicative specification

$$\mathrm{E}[N_{ij}] = \alpha_i \beta_j \tag{3.1}$$

subject to the usual identifiability constraint

$$\sum_{j=1}^{\omega} \beta_j = 1.$$

Several distributional assumptions can be used in conjunction with (3.1), such as the Poisson or Negative Binomial, for instance. In this paper, the $N_{ij}$s are assumed to be Poisson distributed with mean (3.1).

The number of reported claims for accident year 2004 is $4,196$. Among them, 9 claims remain open at the end of the observation period, i.e. at the end of 2014, and thus $\omega > n = 11$. Table 3 displays the estimated $\alpha_i$ and $\beta_j$ that have been obtained by Poisson maximum likelihood. Here we assume $\omega = 13$, which is supported by the decreasing trend exhibited by the estimated $\widehat{\beta}_j$.

Intuitively, we would have expected that the $\widehat{\alpha}_i$ would remain roughly stable over time, suggesting that the volume of business stays unchanged, or moderately increases being compensated by the progressive reduction in claim frequencies in motor insurance. In this case, the bell shape can be explained by circumstances specific to this insurance company, including the progressive integration of business previously sold through bank.

| $i$ | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\widehat{\alpha}_i$ | 4,844 | 4,974 | 5,220 | 5,840 | 6,011 | 6,081 | 5,645 | 5,340 | 5,196 | 5,365 | 4,842 |
| $j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| $\widehat{\beta}_j$ | 0.56234 | 0.32635 | 0.05257 | 0.02375 | 0.01287 | 0.00783 | 0.00405 | 0.00311 | 0.00219 | 0.00194 | 0.00144 |
| $j$ | 12 | 13 | - | - | - | - | - | - | - | - | - |
| $\widehat{\beta}_j$ | 0.00109 | 0.00055 | - | - | - | - | - | - | - | - | - |

Table 3: Estimated parameters $\alpha_i$ and $\beta_j$.

## 3.3 Severity component

In order to model the amounts of payments $X_{ijk}$, $k = 1, \ldots, N_{ij}$, in cell $(i, j)$ we need to accommodate for possibly large values. Therefore, we resort to a discrete mixture with two components:

- a lighter-tailed component with probability $1 - \rho_j$ such as Gamma or Inverse Gaussian distributions;
- a heavier-tailed component with probability $\rho_j$ with Pareto type 2 distribution.

The parameters (probabilities assigned to each component as well as distributional parameters) are explained by means of several explanatory variables using appropriate regression models. Specifically, yearly payments are assumed to be mutually independent and explained by means of an inflation effect $\gamma_{i+j-1}$ (in an hedonic approach) and a development effect $\delta_j$. The choice $\gamma_1 = 1$ makes the inflation parameters identifiable and means that the first accident year is taken as the base year for inflation. If needed, the inflation effect may be structured (by specifying a constant inflation rate, for instance). Of course, other effects may also be included. Notice that working with single payments made by the insurer solves the severe identifiability issues faced in the aggregated triangle approach. The parameter $\delta_j$ then represents the average amount paid at lag $j$, corrected for inflation. Notice that allowing for a calendar year effect in $\mathrm{E}[X_{ij1}]$ rules out the classical CL multiplicative model, including only $i$ and $j$ effects. This means that we end up with a more flexible specification, compared to (4.3) in Wuthrich (2003).

In the numerical illustrations, we use a 2-component mixture with a Gamma and a Pareto distributions. The average payment is of the form

$$\delta_{1,j}(1+g_1)^{i+j-2} \tag{3.2}$$

for the Gamma component and of the form

$$\delta_{2,j}(1+g_2)^{i+j-2} \tag{3.3}$$

for the Pareto type 2 component. In these averages, $g_1$ (resp. $g_2$) can be interpreted as a constant inflation rate for the Gamma (resp. Pareto type 2) component and $\delta_{1,j}$ (resp. $\delta_{2,j}$) models the development effect $j$ for the Gamma (resp. Pareto type 2) component.

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\widehat{\rho}_j$ | 0.162 | 0.431 | 0.747 | 0.920 | 0.978 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $1-\widehat{\rho}_j$ | 0.838 | 0.569 | 0.253 | 0.080 | 0.022 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

Table 4: Estimated probabilities $\rho_j$ by development $j$ for a lighter-tailed component Gamma and a heavier-tailed component Pareto type 2.

Such a 2-component mixture of Gamma and Pareto distributions can be fitted with the help of the `GAMLSS` package of the statistical software `R` (http://www.gamlss.org/). GAMLSS (for Generalized Additive Models for Location, Scale and Shape) are regression models where several parameters of the assumed distribution for the response can be modeled as additive functions of the explanatory variables (not only the mean as in GLMs). Estimation is carried out by maximum likelihood. The only restriction is that the individual contribution to the log-likelihood and its first two derivatives with respect to each of the parameters must be computable. We refer the reader to Stasinopoulos et al. (2017) for more details. Thhe `GAMLSS` package supports many continuous, discrete and mixed distributions for modeling the response variable, including the 2-component Gamma-Pareto mixture used here. Development effects can also be smoothed using P-splines, cubic splines or loess smoothing.

Considering the estimations displayed in Table 4, the weights for the Gamma components dominate the weights for the Pareto type 2 components at early developments, here $j \in$

$\{1,2\}$. The weights assigned to the Gamma components rapidly decrease for $j \in \{3,4,5\}$ to become negligible when $j \geq 6$. Since $\omega = 13$ in our case, we still need to determine values for $\rho_{12}$ and $\rho_{13}$. Obviously, we set $\widehat{\rho}_{12} = \widehat{\rho}_{13} = 0$ in line with the weights at previous lags displayed in Table 4.

The estimated inflation rates $g_1$ and $g_2$ are $\widehat{g}_1 = 1.91\%$ and $\widehat{g}_2 = 2.48\%$ and Table 5 shows the estimated $\delta_{1,j}$ and $\delta_{2,j}$ obtained by using cubic spline smoothers. The way these estimations vary with the development $j$ conforms with intuition. The mathematical expectations of the Gamma components decrease with lag $j$ whereas those of the Pareto component increase with $j$, capturing the largest losses with expensive payments and longer development to settlement. The tails factors $\delta_{2,12}$ and $\delta_{2,13}$ are set to $10,150$ because the last estimations $\widehat{\delta}_{2,9}$, $\widehat{\delta}_{2,10}$ and $\widehat{\delta}_{2,11}$ tend to stabilize around that value.

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\widehat{\delta}_{1,j}$ | 678 | 679 | 214 | 63 | 17 | - | - | - | - | - | - |
| $\widehat{\delta}_{2,j}$ | 3,816 | 3,746 | 4,387 | 5,199 | 6,186 | 7,472 | 8,723 | 9,602 | 10,121 | 10,173 | 10,138 |

Table 5: Estimated parameters $\delta_{1,j}$ and $\delta_{2,j}$ by development $j$ for a lighter-tailed component Gamma and a heavier-tailed component Pareto type 2.

The severity distribution mixes a Gamma with a Pareto component, the weight attributed to the Pareto increasing with the development lag $j$. This rules out the Gamma specification underlying the Tweedie reserving model (that can be seen as a compound Poisson sum with Gamma-distributed summands). Notice that the proposed model encompasses the Tweedie model which is thus rejected for the data set under consideration.

# 4 Reserve calculations

## 4.1 Best estimates

For each calendar year $i$, the expected value of the outstanding claims is

$$\mathrm{E}\left[\sum_{j=n-i+2}^{\omega} X_{ij}\right] = \sum_{j=n-i+2}^{\omega} \mathrm{E}\left[N_{ij}\right] \mathrm{E}\left[X_{ij1}\right].$$

Of course, discounting at some appropriate interest rate can be included, if needed. By equations (3.1), (3.2) and (3.3), it comes

$$\mathrm{E}\left[\sum_{j=n-i+2}^{\omega} X_{ij}\right] = \sum_{j=n-i+2}^{\omega} \alpha_i\beta_j\Big((1-\rho_j)\delta_{1,j}(1+g_1)^{i+j-2} + \rho_j\delta_{2,j}(1+g_2)^{i+j-2}\Big).$$

In our case study, we get a reserve estimate of $21,693,829$.

## 4.2 Outstanding loss distribution

In our collective setting, all $X_{ij}$ are independent and compound Poisson distributed. Hence their sum also obeys a compound Poisson distribution and Panjer algorithm can be used to derive the distribution of the outstanding claim amount. However, a computational problem may happen at initialization of the Panjer recursion to compute $\exp(-\lambda)$ for large values of $\lambda$. We refer the reader for instance to Embrechts and Frei (2009) for more details. Here we rather choose to perform Monte-Carlo simulations.

Simulations can be conducted in several ways. The easiest one consists in simulating the number $N_{ij}$ of payments in each future cell $(i, j)$ and then the corresponding amounts. The latter can be generated in two steps, first selecting the Gamma or Pareto component and second inverting the corresponding distribution function. All these cell-specific values are then summed to get the outstanding claim amount. Alternatively, the compound Poisson distributions corresponding to all future cells can first be aggregated by convolution and simulations are then drawn from the resulting distribution. The second approach lowers the number of simulations needed to achieve the same level of accuracy.

Table 6 summarizes the outstanding claim distribution. The Value-at-Risk (VaR) at probability levels 95% and 99.5% are based on $100,000$ simulations of the outstanding claim amount while the reserve estimate is calculated analytically as shown in Section 4.1. We see that about 1,800,000 has to be added to the best estimate of the reserve to reach the 95th percentile. Approximately another 1,500,000 is needed to obtain the 99.5% quantile.

## 4.3 Comparison with CL and Tweedie

To enable benchmarking, we include the estimation results as obtained with standard reserving techniques designed for run-off triangles. We consider the results obtained with the help of an Overdispersed Poisson (ODP) model with CL structure and a Tweedie model (as in the `chainladder` package available in `R`) where in both cases, we apply a tail factor and we perform $10,000$ simulations.

Table 6 also shows the reserve estimates and the VaR at probability levels 95% and 99.5% obtained with these two approaches.

| | Reserve estimate | $\text{VaR}_{0.95}$ | $\text{VaR}_{0.995}$ |
|---|---|---|---|
| Our approach | 21,693,829 | 23,457,061 | 25,017,697 |
| CL | 22,259,690 | 24,142,573 | 25,239,963 |
| Tweedie | 21,992,440 | 23,879,618 | 25,533,633 |

Table 6: Reserve estimates and VaR at probability levels 95% and 99.5%.

As we can see, the three methods give comparable results. The reserve estimate and the VaRs are slightly lower with our approach compared with Tweedie and CL predictions. At probability level 95%, CL gives the highest VaR while at probability level 99.5%, it is the Tweedie model.

Let us explain why the approach proposed in this paper outperforms the two competitors. The Tweedie model appears to be a particular case obtained by setting the probability $\rho_j$ to 0, so that only the Gamma component remains. Hence, the need for a heavier-tailed component invalidates the Gamma assumption inherent to the Tweedie specification. Considering the CL specification fitted to incremental values using the ODP distribution, the model proposed in this paper can recover the multiplicative structure with an accident year and a development year effects as long as no inflation is present in the data (i.e. the parameters $g_1$ and $g_2$ are both equal to 0) so that only a distributional difference remains. As positive inflation parameters have been obtained, this invalidates the CL specification that cannot account for such effects. Notice also that ODP is a limiting case of Tweedie. In the example studied in this paper, we see that the differences remain nevertheless moderate.

# 5 Discussion

A new collective loss reserving approach has been proposed in this paper. Adopting a frequency-severity setting, we consider the number of payments in every cell of the run-off triangle, together with the corresponding claim-specific yearly paid amounts. Compared to the Tweedie reserving model, that can be seen as a compound sum with Poisson-distributed number of terms and Gamma-distributed summands, more general severity distributions are considered, typically mixture models combining a light-tailed component with a heavier-tailed one. A case study based on a motor third party liability insurance portfolio observed over 2004-2014 has been used to illustrate the approach proposed in this paper. A careful comparison with the Tweedie and CL approaches reveals moderate differences even if both models must be rejected in favor of the approach proposed in this paper.

The model that has been proposed in this paper proceeds as follows. The observed numbers of payments are studied in a Poisson regression setting. Moving to more elaborate models, including mixed Poisson specifications, is possible if more appropriate. A discrete mixture regression model is calibrated to paid amounts, with specific inflation effects. Notice that different specifications can be considered for the inflation effects since working with single payments made by the insurer enables to better capture the calendar year effects than in the aggregated triangle approach. Specifically, a two-component mixture model describes the payments at various developments. Such finite mixture models can be fitted to observed loss developments using the `GAMLSS` package of the statistical software `R`.

Clearly, the approach proposed in the present paper can be improved in case additional information is available about the claim, such as the presence of bodily injuries, claimant's or plaintiff's age, whether the case has been submitted to a court, etc. Such characteristics were not available in the data basis used to illustrate the present work. In practice, such information could nevertheless be included in the regression model for the severity component, which enables to work with severity distributions that account for the claim characteristics, making the prediction of amounts more accurate. This is certainly a promising research area.

# Acknowledgements

The authors wish to thank the Addactis team, especially Patrick Hennion, for fruitful discussions and for kind assistance in obtaining the data set used to illustrate the present paper.

The authors would like to express their gratitude to anonymous referees whose comments have been very useful to improve previous versions of the present work.

# References

- Alai, D. H., Wuthrich, M. V. (2009). Taylor approximations for model uncertainty within the Tweedie exponential dispersion family. Astin Bulletin 39, 453-477.

- Boucher, J. P., Davidov, D. (2011). On the importance of dispersion modeling for claims reserving: An application with the Tweedie distribution. Variance 5, 158-172.

- Embrechts, P., Frei, M. (2009). Panjer recursion versus FFT for compound distributions. Mathematical Methods of Operations Research 69, 497-508.

- Kaas, R., Goovaerts, M.J., Dhaene, J., Denuit, M. (2008). Modern Actuarial Risk Theory Using R. Springer, New York.

- Peters, G. W., Shevchenko, P. V., Wuthrich, M. V. (2009). Model uncertainty in claims reserving within Tweedie's compound Poisson models. Astin Bulletin 39, 1-33.

- Rosenlund, S. (2010). Dispersion estimates for Poisson and Tweedie models. Astin Bulletin 40, 271-279.

- Schiegl, M. (2002). On the safety loading for chain ladder estimates: A Monte Carlo simulation study. Astin bulletin 32, 107-128.

- Schiegl, M. (2015). A model study about the applicability of the Chain Ladder method. Scandinavian Actuarial Journal 2015, 482-499.

- Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V., De Bastiani, F. (2017). Flexible Regression and Smoothing: Using GAMLSS in R. Chapman and Hall/CRC.

- Wuthrich, M. V. (2003). Claims reserving using Tweedie's compound Poisson model. Astin Bulletin 33, 331-346.

- Wuthrich, M. V., Merz, M. (2008). Stochastic Claims Reserving Methods in Insurance. Wiley, New York.