Individual Claims Generator for Claims Reserving Studies: `data simulation.R`

Melantha Wang[*] Mario V. Wüthrich[†]

Prepared for:
Fachgruppe “Data Science”
Swiss Association of Actuaries SAV

Version of June 3, 2022

Abstract

This manuscript explains the individual claims generator “`data simulation.R`”. We use the statistical computing software R [6] to design this individual claims generator. It may be used for developing and back-testing individual claims reserving methods in non-life insurance. We give a descriptive analysis of the generated data, and we provide Mack’s [5] chain-ladder results as a benchmark.

Keywords. Individual claims reserving, granular claims data, individual claims generator, chain-ladder reserving method, Mack’s reserving model, non-life insurance.

0 Introduction and overview

This data analytics manuscript has been written for the working group “Data Science” of the Swiss Association of Actuaries SAV, see

`https://www.actuarialdatascience.org`

The aim of this manuscript is to outline the individual claims generator “`data simulation.R`”[1] that may be used for developing and back-testing individual claims reserving methods in non-life insurance. This individual claims generator is based on the R package `SynthETIC` of Avanzi et al. [1] who provide a simulation environment for individual claims. We have complemented this simulation environment with additional claim features such as the age of the injured person or the type of claim. Such claim information may influence both, the claim frequency and the claim severity, and it is important information for claims prediction in insurance. The purpose of this manuscript is to describe all features of the generated individual claims data, so that users can start to build or test their own individual claims reserving methods based on this data. For our description we mainly use the notation and terminology of Delong et al. [3]; for a general reference on claims reserving in non-life insurance we refer to the monographs Wüthrich–Merz [7, 8].

[*] School of Risk and Actuarial Studies, UNSW Sydney, melantha.wang@unsw.edu.au

[†] RiskLab, Department of Mathematics, ETH Zurich, mario.wuethrich@math.ethz.ch

[1] `https://github.com/JSchelldorfer/IndividualClaimsSimulator`

1 Features of individual claims

A non-life insurance claim is described by an accident date T (claim occurrence), a reporting date U (claim notification), a settlement date V (claim closing), and claim payments $(Y_s)_s$ that are executed between the reporting date U and the settlement date V. Furthermore, the claim may be described by a claim status process $(O_s)_s$, and by further covariates denoted by ϑ; this is going to be described, below. Thus, an insurance claim $\mathcal{C}$ is described by a vector $\mathcal{C} = (T, U, V, (Y_s)_s, (O_s)_s; \vartheta)$. Figure 1 gives an illustration of such a claim $\mathcal{C}$.

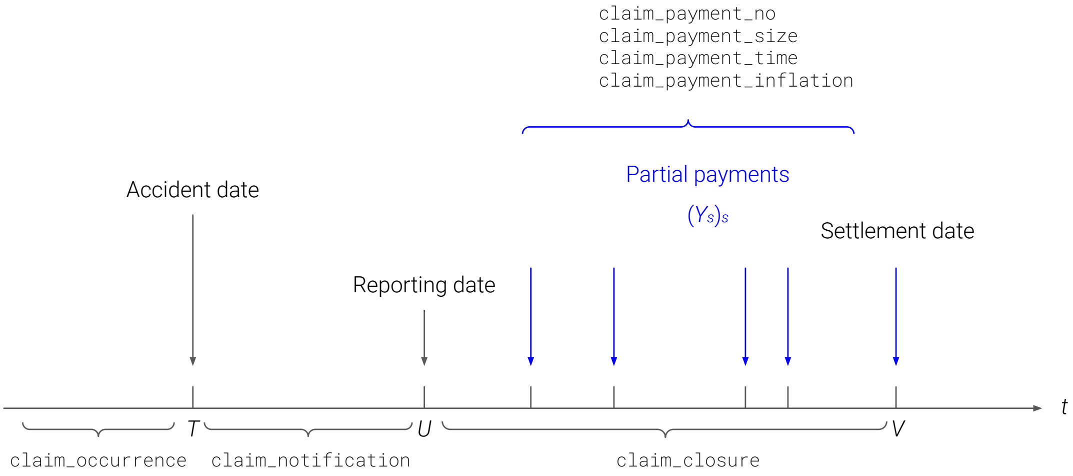

Figure 1: A visual representation of an insurance claim $\mathcal{C}$. The `monospaced` text indicates the associated functions in the `SynthETIC` package [1].

Important in claims reserving, there is a so-called *cutoff date* τ which reflects today's time point, and it determines the available information. Everything that has realized before the time point τ is assumed to be known, and all variables that only occur after time point τ need to be predicted. E.g., a claim may have an accident date and a reporting date $T \le U \le \tau$, which means that this claim has been notified to the insurance company. However, if this claim has not been settled at the time point τ, we need to predict the settlement date $V > \tau$ and the potential claim payments Y_s that only occur after time τ, see Figure 1.

By running the command `data.generation(seed=...)`, one receives such claims $\mathcal{C}_i$, $1 \le i \le n$. We first describe the structure of these claims before going into more detail about the specific nature of the generated claims. A claim has two parts: (1) there is the claim description $(T, U, V; \vartheta)$, and (2) there are the claim payments $(Y_s)_s$ and the claim status process $(O_s)_s$. We describe these two parts. An excerpt of the first part is shown in Listing 1, and it provides the following information:

variable	
Id	claims ID being a unique identifier for each claim $1 \leq i \leq n$
$T =$ AccDate	accident date (in daily units as yyyy/mm/dd) from 2012/01/01 to 2021/12/31
AccMonth	accident date (in monthly units) from 1 to 120
AccWeekday	accident weekday from Mon to Sun
$U =$ RepDate	reporting date (in daily units as yyyy/mm/dd) after 2012/01/01 or NA[2]
RepMonth	reporting date (in monthly units) from 1 to 120 or NA
RepDelDays	reporting delay (in daily units) non-negative or NA
Type	claim type having labels 1 to 6
Age	age of injured person being between 18 and 65
$V =$ SetMonth	settlement date (in monthly units) from 1 to 120 or NA[3]
SetDelMonths	settlement delay (in monthly units) non-negative or NA (w.r.t. RepMonth)
Ultimate*	total claim amount
PayCount*	total number of payments that constitute the total claim amount
Status	claim status at the cutoff date τ being either Closed, RBNS or IBNR
CumPaid	cumulative payments until the cutoff date τ (for Closed and RBNS claims)

This table shows the individual claim descriptions $(T, U, V; \vartheta)$. However, *not all* this information is necessarily available at the cutoff date $\tau = 2021/12/31$ (observe that all claims have an accident date satisfying $2012/01/01 \leq T \leq \tau = 2021/12/31$):

- $\tau < U =$ RepDate: There is no information available about such claims as these are not notified to the insurance company, yet (at the cutoff date τ). I.e., no position of the above table is available for such claims. In particular, the reporting date U is NA. These are so-called incurred but not reported (IBNR) claims, that need to be fully predicted, see Status on the second last line of the above table.

- $U =$ RepDate $\leq \tau < V =$ SetMonth: There is partial information about these claims because they are reported at the cutoff date τ, but they are not (fully) settled, yet. These claims are so-called reported but not settled (RBNS) claims, and the claim development after the cutoff date τ needs to be predicted. In particular, the settlement date V is NA for these RBNS claims, see also Status on the second last line of the above table. For RBNS claims the first 3 blocks of the above table (from Id to Age) and CumPaid are known/observed.

- $V =$ SetMonth $\leq \tau$: For these claims we have full information, they have been Closed as of the cutoff date τ, and no further claim payments are *expected*. Since claims can be re-opened for *unexpected* further claim developments, we still need to keep track of closed claims, or, more mathematically speaking, the closing date V is not a stopping time as it may change at a later time point due to unexpected further developments.

- The items Ultimate* and PayCount* have an asterisk * in the above table. This asterisk indicates that this information is not necessarily available and, in fact, it is the quantity of central interest that should be predicted. Since we have fully simulated claims, this information is available, here, and this helps one to back-test whether a certain claims reserving method works or not on this particular data set.

[2] NA indicates that the claim has a simulated reporting date later than the cutoff date. In other words, we have an incurred but not reported (IBNR) claim.

[3] NA indicates that the claim has a simulated settlement date later than the cutoff date. This can either be an incurred but not reported (IBNR) claim or a reported but not settled (RBNS) claim.

Listing 1: Claim description $(T, U, V; \vartheta)$ of all generated claims $1 \leq i \leq n = 61'288$.

```
1  'data.frame':   61288 obs. of  15 variables:
2  $ Id          : int  1 2 3 4 5 6 7 8 9 10 ...
3  $ Type        : num  5 4 6 1 1 4 4 1 4 5 ...
4  $ Age         : int  23 39 51 40 28 64 49 48 28 36 ...
5  $ AccDate     : Date, format: "2012-01-02" "2012-01-03" ...
6  $ AccMonth    : num  1 1 1 1 1 1 1 1 1 1 ...
7  $ AccWeekday  : Factor w/ 7 levels "Mon","Tue","Wed",..: 1 2 2 5 7 3 5 7 2 1 ...
8  $ RepDate     : Date, format: "2012-01-04" "2012-01-05" ...
9  $ RepMonth    : num  1 1 1 1 1 1 1 1 1 1 ...
10 $ RepDelDays  : int  2 2 2 0 6 3 1 0 5 6 ...
11 $ SetMonth    : num  28 27 1 23 14 4 6 12 8 12 ...
12 $ SetDelMonths: num  27 26 0 22 13 3 5 11 7 11 ...
13 $ Ultimate    : num  38945 3828 724 12658 674 ...
14 $ PayCount    : num  4 1 1 2 1 1 1 1 1 2 ...
15 $ Status      : Factor w/ 3 levels "Closed","RBNS",..: 1 1 1 1 1 1 1 1 1 1 ...
16 $ CumPaid     : num  38945 3828 724 12658 674 ...
```

The information provided in the above table and in Listing 1 describes the features of the individual claims, e.g., we may have information about the age of the injured person (line 4 of Listing 1), the claim type (line 3) and the weekday of the accident (line 7). For claims that are not `Closed` at the cutoff date τ, called *open claims*, we still need to expect payments after the time point τ, and the main goal in claims reserving is to predict these payments (for the `RBNS` and `IBNR` claims at the cutoff date τ). In order to predict these payments we have further information, namely, we know all payments $(Y_s)_s$ that have been done for `Closed` and `RBNS` claims up to the cutoff date τ. Since there is the (rare) case of `Closed` claims being re-opened after the cutoff date τ, with further unexpected claim payments after τ, we also need to keep track of `Closed` claims and their claim status (open/closed), this is encoded in the claim status process $(O_s)_s$.

Listing 2: Observed claim payments $(Y_s)_s$ and claim status $(O_s)_s$ up to the cutoff date τ.

```
'data.frame':   86917 obs. of  6 variables:
 $ Id        : int  1 1 1 1 1 2 3 4 4 5 ...
 $ EventId   : int  1 2 3 4 5 1 1 1 2 1 ...
 $ EventMonth: num  9 15 21 27 28 27 1 10 23 14 ...
 $ Paid      : num  11139 10863 11831 0 5112 ...
 $ PayInd    : num  1 1 1 0 1 1 1 1 1 1 ...
 $ OpenInd   : num  1 1 0 1 0 0 0 1 0 0 ...
```

Listing 2 shows the claim payments $(Y_s)_s$, called `Paid`, and the claim status process $(O_s)_s$, called `OpenInd`, up to the cutoff date τ. Firstly, all payments are recorded on an individual claim basis, and the variable `Id` allows one to identify the claims of Listing 1, their payments Y_s and their claim status $O_s \in \{0, 1\} = \{\text{closed}, \text{open}\}$ at calendar month s given in Listing 2. Secondly, payments are aggregated within calendar months, termed *monthly payments*, and Listing 2 shows all calendar months in which such a monthly payment occurs (up to the cutoff date τ). Moreover, whenever the claim status changes either from open $O_{s-1} = 1$ to closed $O_s = 0$ (claim closing) or from closed $O_{s-1} = 0$ to open $O_s = 1$ (claim re-opening) there is an entry in Listing 2. An "event" can be either a payment or a change in claim status. E.g., the claim with `Id` $= 1$ has four monthly payments (with event identifications `EventId` $= 1, 2, 3, 5$), see Listing 3. These four payments have taken place in calendar months (`EventMonth`) 9, 15, 21 and 28 (1 corresponds to January 2012), and the sizes of these monthly payments are 11'139, 10'863, 11'831 and 5'112. This claim has been `Closed` in calendar month 21, see `OpenInd`, and this claim has been re-opened in calendar month 27, and it has been closed again in calendar month 28 with an additional payment of 5'112. Claim `Id` $= 1$ is a `Closed` claim at the cutoff date τ, see Listing 1, and the total claim amount of this claim is `Ultimate` $= 38'945$.

Listing 3: Observed claim payments $(Y_s)_s$ and claim status $(O_s)_s$ of claim `Id = 1` up to the cutoff date τ; this is a `Closed` claim that went through a re-opening.

```
Id EventId EventMonth  Paid PayInd OpenInd
 1       1          9 11139      1       1
 1       2         15 10863      1       1
 1       3         21 11831      1       0
 1       4         27     0      0       1
 1       5         28  5112      1       0
```

Summary. We conclude that "`data simulation.R`" outputs the data of Listings 1 and 2. The goal is to design claims reserving methods that predict the claim payments after the cutoff date τ, such that these predictions provide total claim estimates that are as close as possible to the true claim, called `Ultimate`, and shown on line 13 of Listing 1.

2 Descriptive analysis

We provide a descriptive analysis of the outputs of Listings 1 and 2 in this section. Note that in a real claims reserving situation, not all information that is displayed in these two listings is available, e.g., the `Ultimate` claim amounts are only available for `Closed` claims (supposed that they are not re-opened). Nevertheless, we use this extended information, here, to illustrate the properties of our individual claims generator, and we indicate in each graph whether this is available or not in a typical claims reserving situation.

2.1 Claim counts and claims reporting

We start by analyzing claim occurrence and claim reporting.

Figure 2 illustrates the accident dates T on the y-axis vs. the reporting delays $W = U - T$ on the x-axis; the units are calendar weeks. Each dot represents a single claim, with the exception that a green dot indicates more than four claims within the same accident and reporting week and a blue dot between two and four claims within the same accident and reporting week. Each graph corresponds to a different claim `Type`. The orange diagonal line shows the cutoff date τ with reported claims above this orange diagonal line, and `IBNR` claims (to be predicted) below this orange diagonal line. This is the typical information available about claims reporting in claims reserving. We have the following observations: (1) The maximally observed reporting delay $W = U - T$ is 156 weeks which corresponds to 3 years. (2) Some claim types suffer more claims and some less claims, this is indicated by the density (and color) of the dots. (3) Claim type 5 has an annual seasonal pattern, the others do not, however, they may still increase or decrease in volume/frequency (i.e., have a trend).

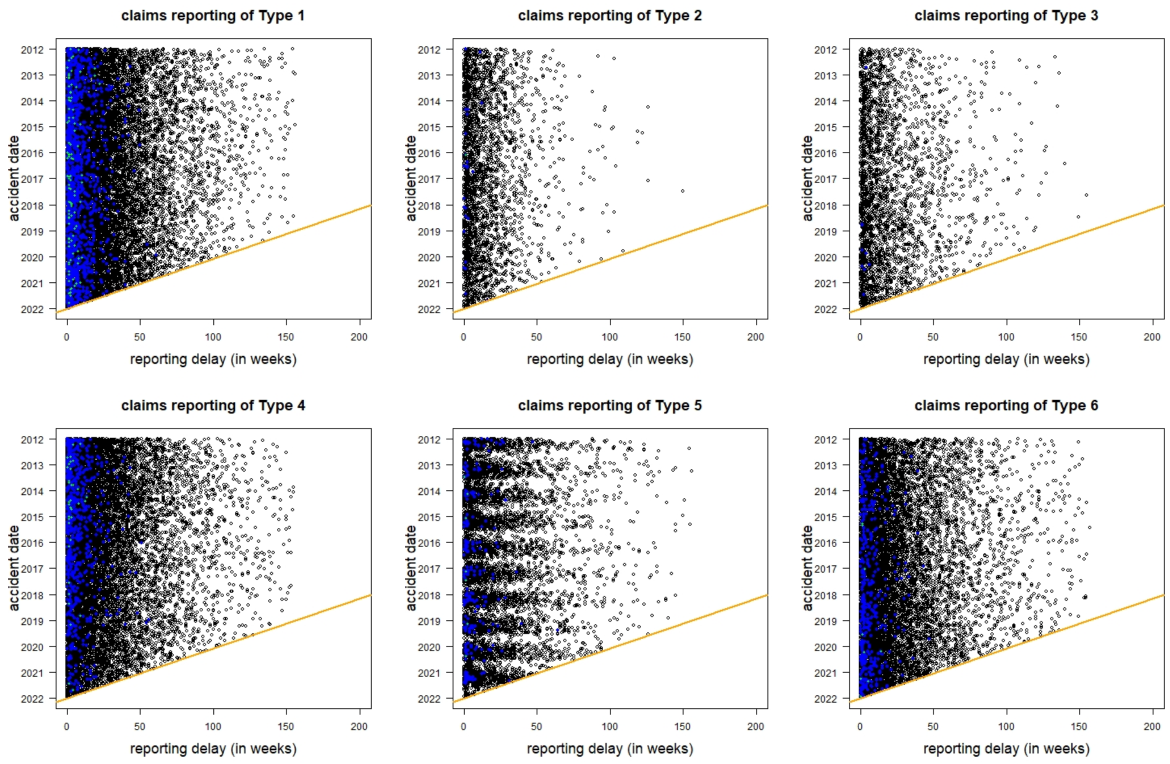

Figure 2: Accident dates T (in weekly units) vs. reporting delays $W = U - T$ (in weekly units) for each of the 6 claim types; this information is available in a typical claims reserving situation.

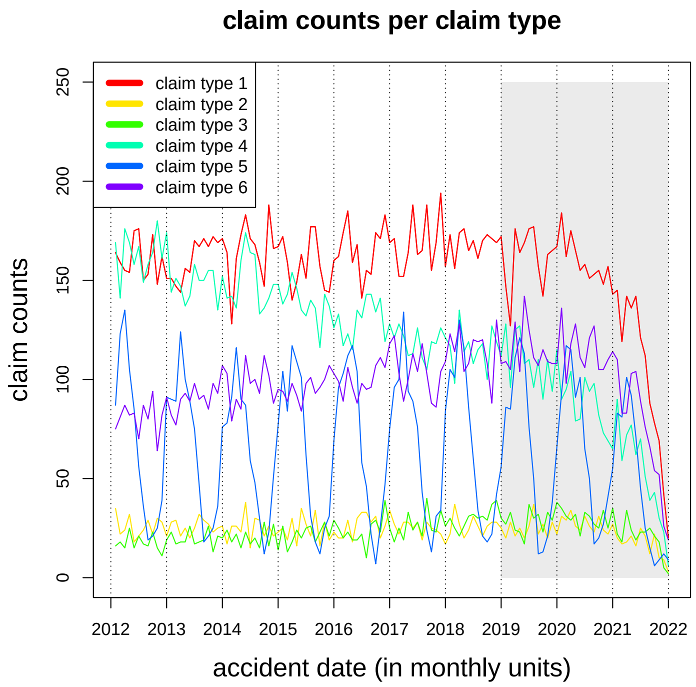

Figure 3: Claim counts w.r.t. accident dates for the 6 different claim types; this information is available in a typical claims reserving situation.

Figure 3 shows the number of reported claims for each claim `Type` and aggregated per accident date (in monthly units). This figure only shows the claims that are reported by the cutoff date τ. `IBNR` claims with reporting dates $U > \tau$ are missing in this plot, which explains the drop in observed claim counts over the most recent months. From Figure 2 we concluded that the maximal reporting delay $W = U - T$ is 3 years. Assuming that this observation generalizes to the future, we can still expect `IBNR` claims with accident dates between 2019/01/01 and 2021/12/31 to come through after τ. This is illustrated by the gray shaded area in Figure 3, where we expect the claim counts still to increase. Figure 3 confirms that claim type 5 has a strong seasonal component. We can also conclude that the number of types 3 and 6 claims is increasing over the accident years, while this trend is reversed for claim type 4 (i.e., decreasing claim counts).

2.2 Ultimate claim sizes

Next we analyze the total ultimate claim sizes. We therefore use the `Ultimates` of Listing 1. Since this information is only available for `Closed` claims, the graphs in this section are typically *not available* in a real claims reserving situation. It is possible to analyze the distribution of the ultimate claim size of `Closed` claims that have occurred and have been settled in the past. However, as pointed out earlier, these claims may be subject to re-openings and further claim developments after the cutoff date, and consequently do not provide a complete picture of the ultimate claims as presented below. Moreover, if we only consider `Closed` claims we typically receive a bias because smaller and simpler claims are settled faster than bigger and more complicated claims.

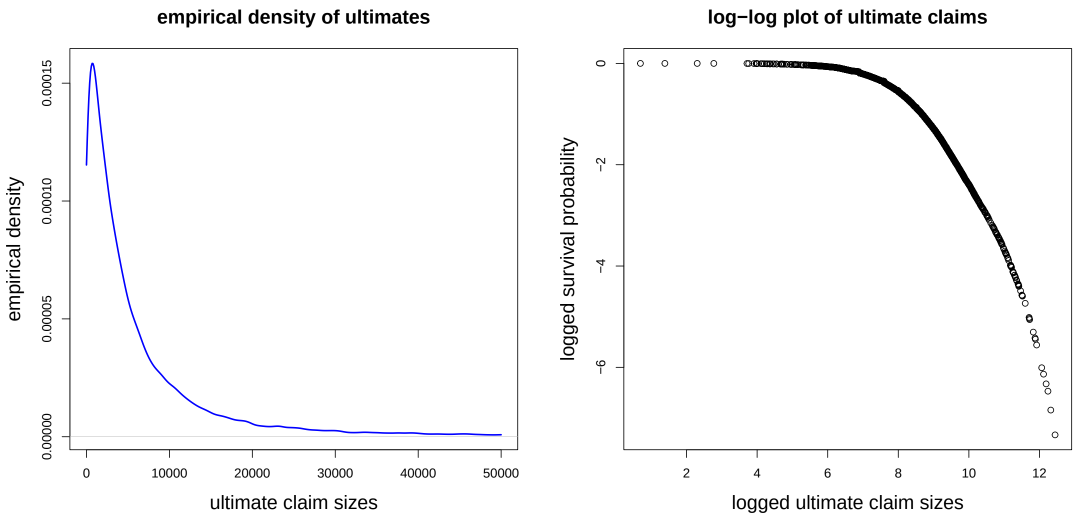

Figure 4: Ultimate claim sizes (lhs) empirical density (truncated at 50'000) and (rhs) log-log plot; this information is usually not available.

Figure 4 shows the empirical density and the log-log plot of the ultimate claim sizes. We have a unimodal density with most claims below 50'000, the biggest claim is 819'279 (not shown on the graph). A certain heavy-tailedness is also indicated by the log-log plot of Figure 4.

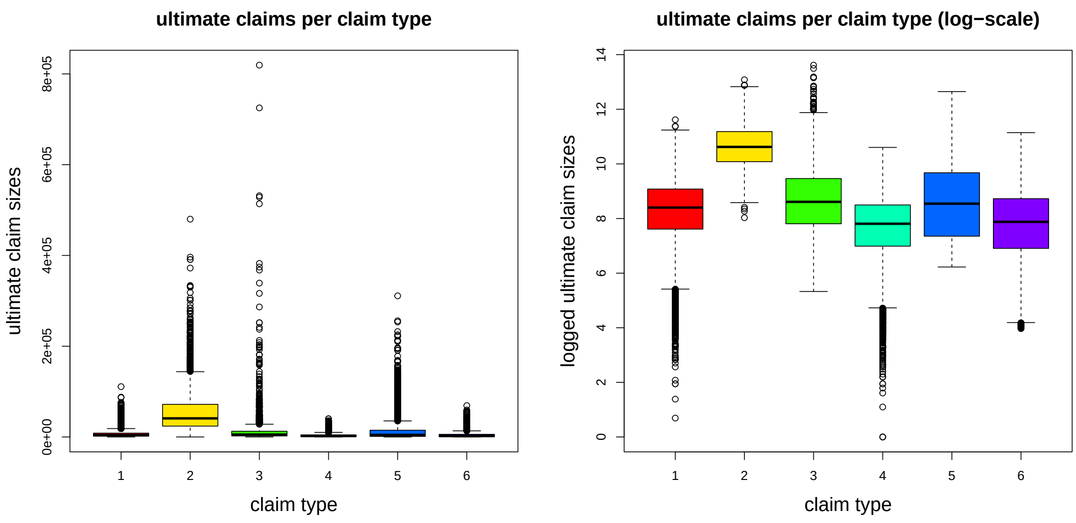

Figure 5: Box plots of ultimate claim sizes per claim type (lhs) original scale and (rhs) log-scale; this information is usually not available.

Figure 5 shows the box plots of the ultimate claim sizes per claim type, both on the original scale and on the log-scale. These plots support heavy-tailedness, and there are substantial differences between the different claim types, e.g., claim type 2 having bigger claims than the other claim types (also recall from Figure 3 that they have lower claim counts, which is quite typical). This is also verified by Table 1 showing empirical statistics over all strictly positive claims.

claim type	mean	std.dev.	minimum	maximum
type 1	6’710	7’228	2	110’921
type 2	56’652	49’412	3’070	479’926
type 3	14’711	37’956	206	819’279
type 4	3’734	4’011	1	40’299
type 5	13’748	22’768	505	310’977
type 6	4’701	5’870	53	68’336

Table 1: Empirical means and standard deviations, observed minimum and maximal ultimate claims per claim type; this information is usually not available.

Figure 6 shows the ultimate claim sizes as a function of the age of the injured person. For claim type 1 we see an increasing trend and for claim type 3 a decreasing trend in the age variable. Interestingly, claim type 4 exhibits a cyclic structure, with ages in the 50s having the highest claim sizes.

Figure 7 shows that there are further dependencies of the ultimate claim sizes on the covariate information, e.g., claim type 5 has an annual seasonal component, and claim type 6 a weekday component, letting accidents on weekends be more severe than accidents on working days. Thus, the covariate information of Listing 1 should generally help to improve the accuracy of predictions because it has some predictive power for the ultimate claim sizes.

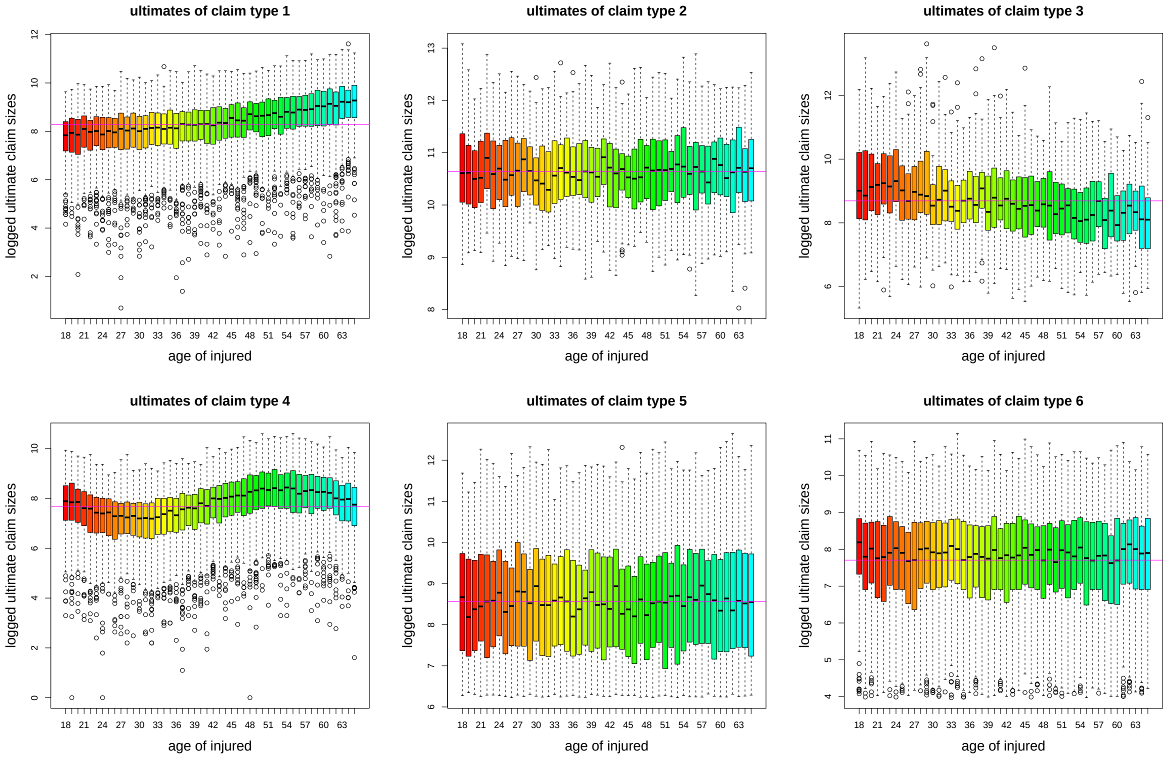

Figure 6: Logged ultimate claim sizes as a function of the age of the injured person for the 6 different claim types; this information is usually not available.

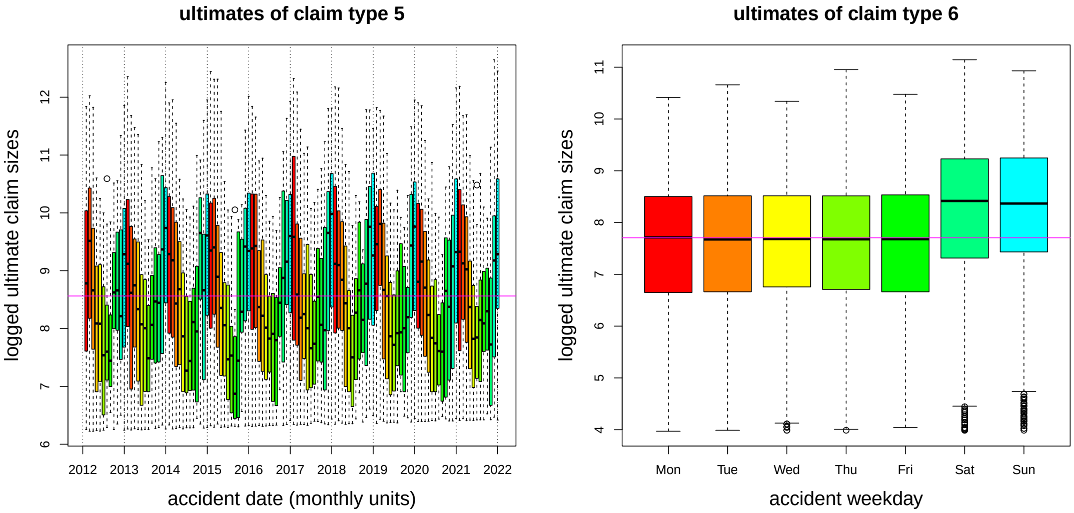

Figure 7: Logged ultimate claim sizes as a function of the accident date (in monthly units; for claim type 5; lhs) and as a function of the weekday of the accident (for claim type 6; rhs); this information is usually not available.

2.3 Analysis of reporting delays

Next we study the reporting delays. These can only be studied for reported claims, as this information is not available for IBNR claims, and, as a consequence, the following plots typically have a (systematic) negative bias, i.e., the missing information from IBNR claims will typically increase the reporting delays as they usually represent late reported claims.

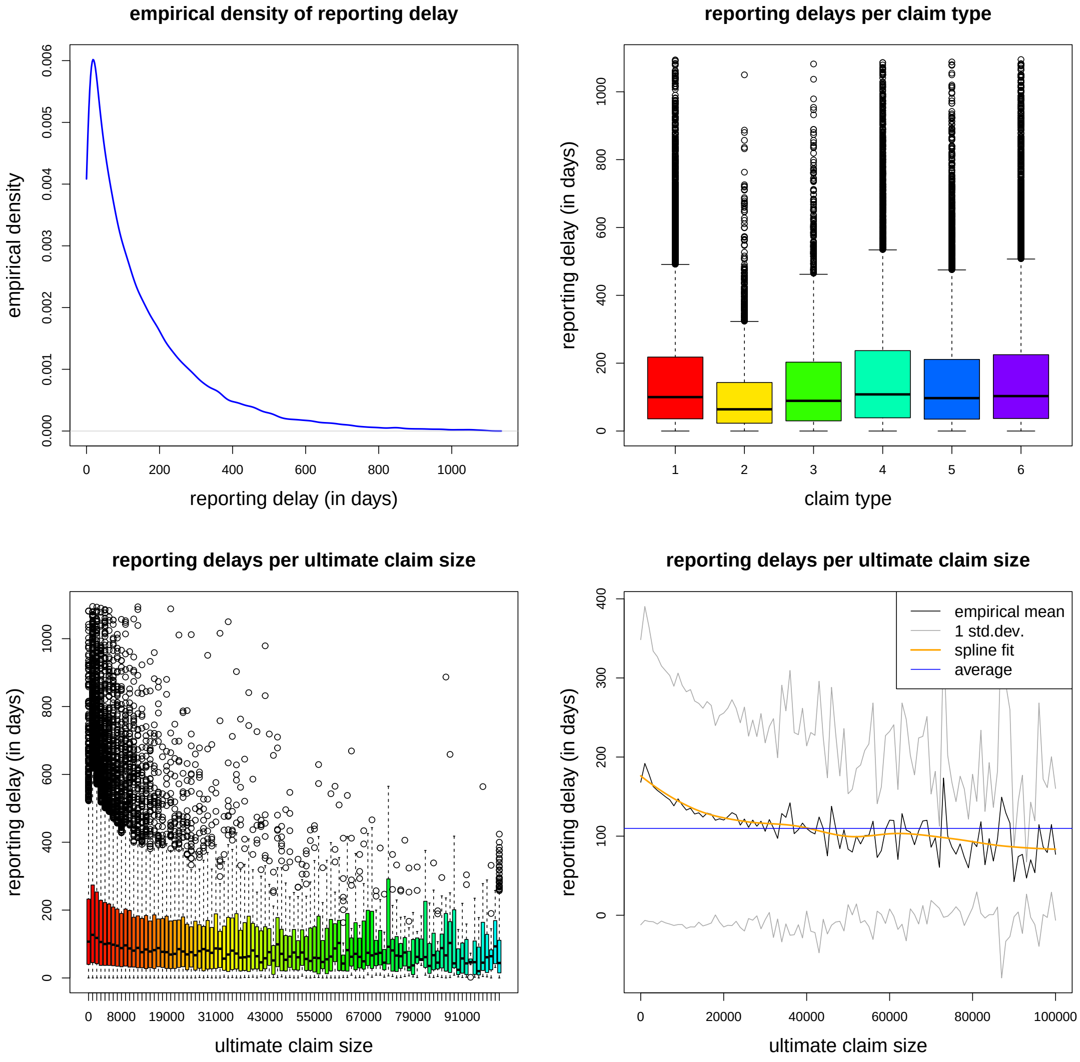

Figure 8: Reporting delays of reported claims (top-lhs) empirical density, (top-rhs) as a function of claim type, (bottom) as function of ultimate claims; the bottom plots are usually not available.

Figure 8 shows the reporting delays of the reported claims (i.e., excluding IBNR claims). We observe that these reporting delays depend on the claim types, and there is a negative correlation with the ultimate claim sizes, with bigger ultimate claims having smaller average reporting delays. The latter statement uses the Ultimate information which is typically only available for Closed claims. The average reporting delay across all reported claims is roughly 100 days (horizontal blue line bottom-right plot). Again, we remark that this is likely an underestimate of the true reporting delay for all claims (i.e., including IBNRs).

2.4 Analysis of settlement delays

In this section we illustrate settlement delays, see SetDelMonths of Listing 1. Settlement delays can only be studied for Closed claims, as this information is not available for RBNS and IBNR claims. Since a Closed claim may still be re-opened after the cutoff date τ, some of these settlement delays may still increase due to a later (final) settlement date. That is, the following settlement delay plots have a negative bias on Closed claims due to the potential re-openings. This negative bias may become even bigger if we additionally account for RBNS and IBNR claims, as this missing information from open claims typically further increases settlement delay figures (unless IBNR claims have very small settlement delays in general).

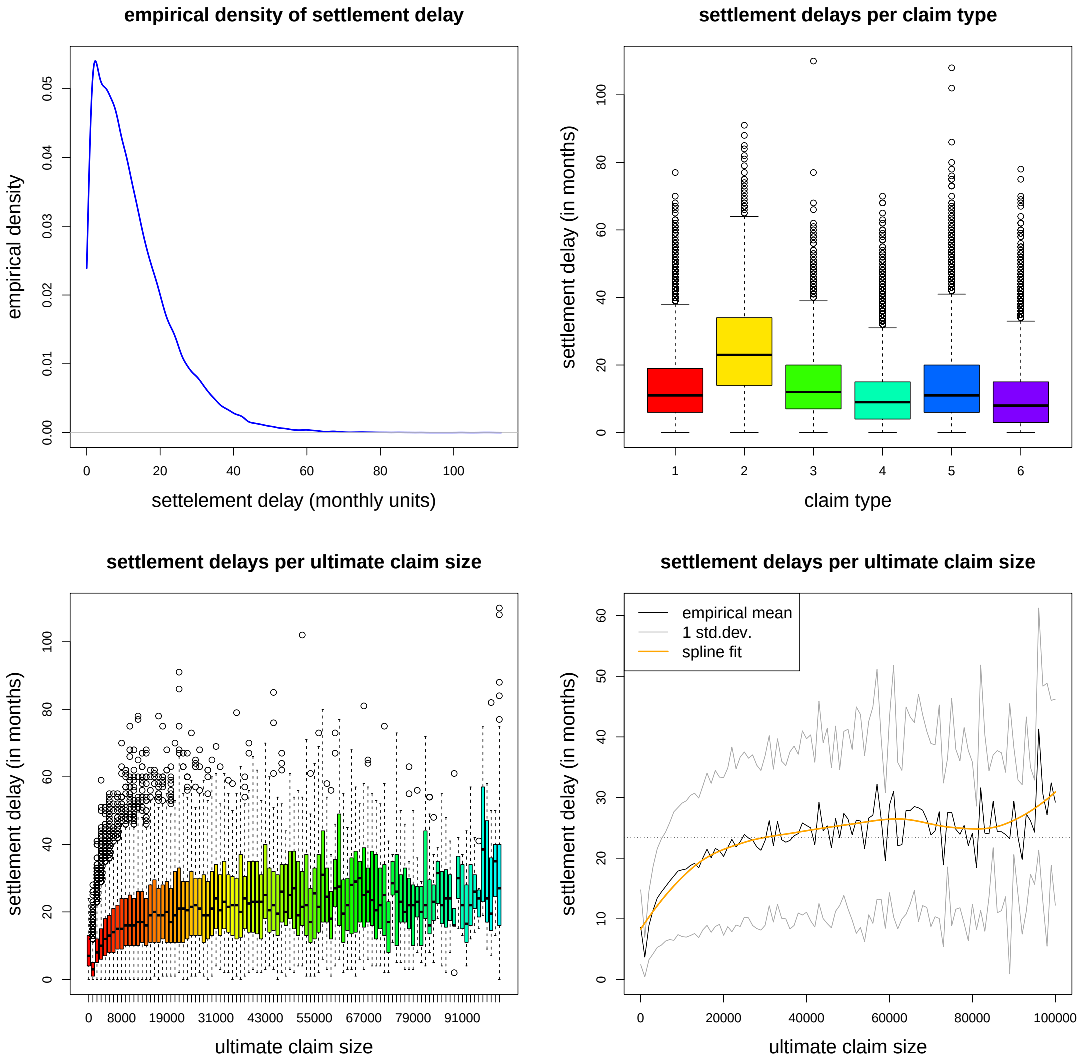

Figure 9: Settlement delays of Closed claims (top-lhs) empirical density, (top-rhs) as a function of claim type, (bottom) as function of ultimate claims; this information is available.

Figure 9 shows the graphs for the settlement delays of Closed claims. The density of the settlement delay of Closed claims is unimodal with most claims being settled within 5 years (60 months) after reporting. Settlement delay seems claim type dependent, see Figure 9 (top-rhs), and, not surprising, smaller claims are settled faster, see bottom of Figure 9.

Thus, we have seen that claim features, such as the age of the injured, claim type and weekday of accident, influence claim counts, reporting delays, claim sizes and settlement delays, and a good claims reserving model should be able to model these interactions.

2.5 Re-openings of closed claims

As discussed in Section 1, claims that were marked Closed may still re-open for unexpected further claim developments. In Figure 10, we plot the proportion of closed claims that are re-opened as a function of time elapsed since the first claim closure (in calendar months), separately for each claim Type. This information will not be available in a typical claim reserving situation as we do not know which of the claims marked Closed as of the cutoff date τ will re-open in the future.

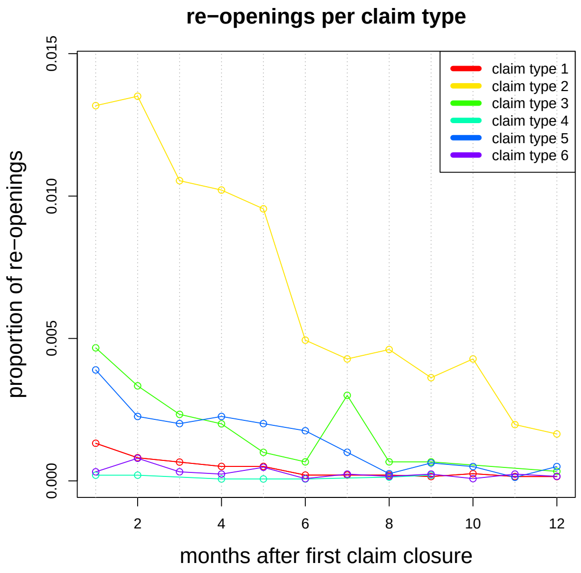

Figure 10: Proportion of re-openings w.r.t. months after the first claim closure for the 6 different claim types; this information is not available in a typical claims reserving situation.

In this simulated portfolio, slightly less than 1% of the Closed claims are re-opened, with some variations between the different claim types. Claim type 2, e.g., experiences more frequent re-openings than any other claim types, while claim type 4 observes almost zero re-openings. All re-openings occur within 12 months of the first claim closure, and the number of re-openings reduce roughly exponentially with the time elapsed after closing.

2.6 Monthly claim payments

The claims of Listing 1 are settled by claim payments $(Y_s)_s$. The payments made up to the cutoff date τ are given as in Listing 2. This includes payments for `Closed` claims and payments for `RBNS` claims. In general, we expect further payments for `RBNS` claims, and there might be some unexpected payments for `Closed` claims that are re-opened after the cutoff date τ. Moreover, we need to predict payments for `IBNR` claims. The general goal of the reserving task is to predict all these payments based on the available information at the cutoff date τ.

Claims can be settled by multiple payments, but there may also be some claims that do not have any payments. In this simulated data, roughly 5.1% of all claims are settled without any payments, and 94.9% of the claims have at least one payment. In fact, 52.4% of all claims are settled with exactly one monthly payment, 27.3% with two monthly payments and 15.2% with more than two monthly payments; this information is typically not available in a real claims reserving situation.

Noteworthy, some claims of our simulated data have recovery payments, i.e., payments with a negative sign. Recovery payments may have different reasons. Often, an insurance company pays the full claim in a first step, and, in a second step, it can recover part of these (full) payments. E.g., the insurance policyholder has a deductible that can be recovered by the insurance company, or there is a subrogation through another insurance contract. Thus, if we model and predict claim payments as in Listing 2, we need to select a stochastic model that can also cope with recovery payments, otherwise, we will receive a biased prediction.

This finishes the small empirical analysis of the generated individual claims data.

3 Mack's chain-ladder model

The most popular claims reserving method is Mack's [5] distribution-free chain-ladder (CL) model which has been introduced in 1993. Mack's CL model (only) considers claim payments in an aggregated version, and it does not consider any additional claim features, nor does it distinguish between `Closed`, `RBNS` and `IBNR` claims.

Denote by $X_{i,j}$ all claim payments that have been done in accident year $2012 \leq i \leq 2021$ with payment delay $0 \leq j \leq 9$, payment delay is measured in yearly units and w.r.t. the accident year. Cumulative payments for a fixed accident year i in development year j are then defined by

$$C_{i,j} = \sum_{l=0}^{j} X_{i,l}.$$

The observed cumulative payments $C_{i,j}$, $i+j \leq 2021$, are given by the upper triangle in Table 2, and claims reserving aims at completing the corresponding lower triangle (after the cutoff date $\tau = 2021/12/31$). Mack's [5] CL method completes this lower triangle and assigns a prediction uncertainty to this prediction. For a mathematical description of the CL method we refer to Mack [5] and our monographs Wüthrich–Merz [7, 8]. For the present outline, we use the `R` package `ChainLadder` of Gesmann et al. [4] to perform the CL analysis and to calculate Mack's CL uncertainty formula.

i / j	0	1	2	3	4	5	6	7	8	9
2012	4’172	20’325	35’126	42’687	45’962	47’818	48’545	48’955	48’991	49’131
2013	4’983	23’047	36’622	44’973	48’927	50’426	50’937	50’946	50’946	
2014	4’698	24’127	38’991	46’334	50’044	51’423	51’808	51’900		
2015	5’467	23’948	38’796	46’885	49’679	51’059	51’608			
2016	5’985	25’933	42’458	50’943	54’559	56’307				
2017	6’729	26’970	43’165	51’640	55’206					
2018	5’561	26’614	42’950	51’918						
2019	6’338	27’358	44’036							
2020	5’563	26’843								
2021	6’267									

Table 2: Observed cumulative payments $C_{i,j}$ in the upper triangle $i + j \leq 2021$ (in 1’000).

i	CL predicted	true `Ultimate`	difference	RMSEP	in %
2012	49’131	49’131	0	–	–
2013	51’092	50’957	134	4	3350.0%
2014	52’067	51’934	133	33	403.0%
2015	51’950	51’767	182	260	70.0%
2016	57’292	57’513	-221	336	65.8%
2017	57’944	57’126	818	453	180.6%
2018	58’514	59’386	-871	713	122.2%
2019	59’852	58’759	1’092	972	112.3%
2020	59’262	59’298	-35	1’781	2.0%
2021	62’937	57’190	5’748	5’227	110.0%
total	560’041	553’061	6’980	5’947	117.4%

Table 3: Mack’s CL prediction (in 1’000).

The results of Mack’s CL method are presented in Table 3. We observe that in this example the CL method over-estimates the true claims in the lower triangle by 6’980K. This corresponds to 117.4% of the square-rooted mean squared error of prediction (RMSEP), i.e., the CL prediction deviates more than one (predictive) standard deviation from the true total claim amount. Table 3 provides a first benchmark that should be beaten by any more sophisticated claims reserving method.

In a final analysis we study the volatility of these CL results under the assumption that the data has been simulated by the same model but with a different seed for the random number generator. For this we just change the `seed` in `data.generation(seed=...)`, and all other model parameters and assumptions are kept fixed.

Figure 11 shows the results over 20 different `seeds`. The blue box plot shows the true outstanding loss liabilities, which correspond to all payments after the cutoff date τ. In average we have 111’583K payments after the cutoff date. The red box plot shows the corresponding CL predictions. In average we have a positive bias of roughly 4’000K (the orange horizontal line shows the average prediction of 115’530K), and the predictions are more volatile (the red box is bigger than the blue box). This is not surprising because the true outstanding loss li-

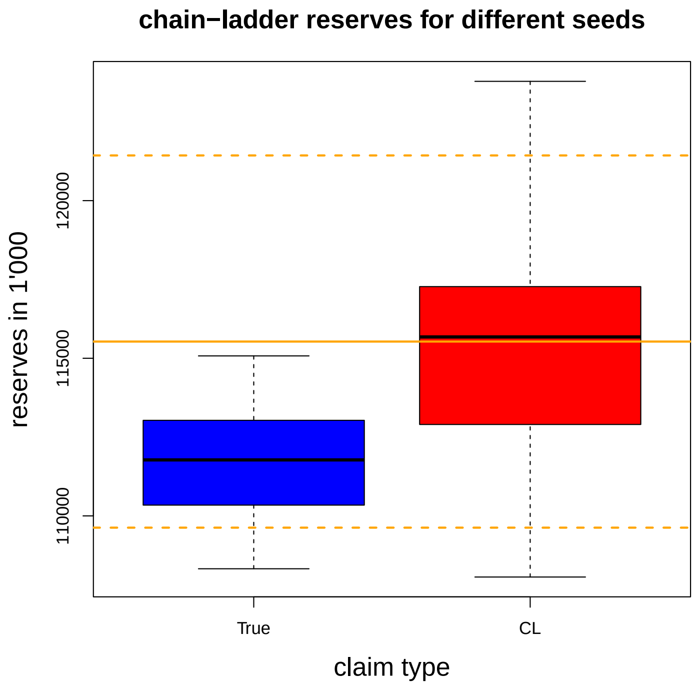

Figure 11: Box plot of 20 simulations with different seeds: blue shows the true outstanding payments, red shows the CL reserves, the orange straight line gives the average CL reserves over the 20 different runs and the dotted lines correspond to the average RMSEP.

abilities in blue only contain pure randomness (irreducible risk), whereas the CL method also contains fluctuations from model uncertainty (from parameter estimation). The orange dotted lines show the average RMSEP of 5'903K (below and above the average CL reserves). Thus, the CL method is biased for this individual claims generator, but in average this bias is smaller than one RMSEP.

4 Summary

The R code "`data simulation.R`" provides an individual claims generator that generates individual claims data based on a selection of claim features. Our generator extends the `SynthETIC` simulator by Avanzi et al. [1] by explicitly allowing for claim covariates and re-openings. We also note that Avanzi et al. [2] have recently provided an extension to the original package that allows for the simulation of incurred losses (in addition to paid losses).

In our generator "`data simulation.R`" we specify the portfolio volumes, the claim frequencies, the claim payments and other claim features. The only parameter that can be changed is the `seed` of the random number generator; alternatively one can dive into the code of the R function `data.generation(seed=...)` to, e.g., modify the claim size distributions or the portfolio volumes and the claim frequencies.

The output of this generator includes: (1) individual claims that are established with features and, (2) claim payments (and including changes in claim status, e.g., re-openings). The output provides all information that is typically available at the cutoff date τ for claims reserving. Beyond that it provides more information like the `Ultimate` size of the claims, which is typically not available in real world situations. This additional information is useful for back-testing claims reserving algorithms. As an example, we provide Mack's [5] distribution-free chain-ladder method, see Table 3. The back-test provides a non-negligible over-estimation of the true claims by the chain-ladder method on our data. This bias seems systematic as simulating claims with the same model but with a different initial seed provides similar results, see Figure 11. The chain-ladder model is the simplest but still rather robust method. It can serve as a benchmark that should be beaten by any more sophisticated claims reserving method.

References

[1] Avanzi, B., Taylor, G., Wang, M., Wong, B. (2021). SynthETIC: an individual insurance claim simulator with feature control. *Insurance: Mathematics and Economics* **100**, 296-308.
`https://doi.org/10.1016/j.insmatheco.2021.06.004`

[2] Avanzi, B., Taylor, G., Wang, M. (2022). SPLICE: a synthetic paid loss and incurred cost experience simulator. *Annals of Actuarial Science* (in press).
`hhttps://doi.org/10.1017/S1748499522000057`

[3] Delong, Ł., Lindholm, M., Wüthrich, M.V. (2022). Collective reserving using individual claims data. *Scandinavian Actuarial Journal* **2022/1**, 1-28.
`https://www.tandfonline.com/doi/full/10.1080/03461238.2021.1921836`

[4] Gesmann, M., Murphy, D., Zhang, Y., Carrato, A., Wüthrich, M.V., Concina, F., Dal Moro, E. (2022). ChainLadder: statistical methods and models for claims reserving in general insurance. R package version 0.2.15.
`https://CRAN.R-project.org/package=ChainLadder`

[5] Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin* **23/2**, 213-225.

[6] R Core Team (2021). R: a language and environment for statistical computing. R *Foundation for Statistical Computing*, Vienna, Austria. `https://www.R-project.org/`

[7] Wüthrich, M.V., Merz, M. (2008). *Stochastic Claims Reserving Methods in Insurance*. Wiley.

[8] Wüthrich, M.V., Merz, M. (2015). *Stochastic Claims Reserving Manual: Advances in Dynamic Modeling*. SSRN Manuscript 2649057.
`https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2649057`

A File “data simulation.R”

The file “data simulation.R” illustrates our data simulator. It initializes the necessary parameters, and it calls the file

```
source("./Tools/functions simulation.R")
```

which contains all important functions and procedures. The function

```
claims_list <- data.generation(seed=1000, future_info=FALSE)
```

then generates the individual claims data. seed sets the seed for the random number generator, and for future_info=FALSE exactly the data in Listings 1 and 2 is produced (as a list with two objects). If one sets future_info=TRUE, then a list of five objects is returned, the latter three objects contain the full information (also after the cutoff date τ) of the claim description $(T, U, V; \theta)$, the claim payments $(Y_s)_s$ and the re-opening claim records $(O_s)_s$.