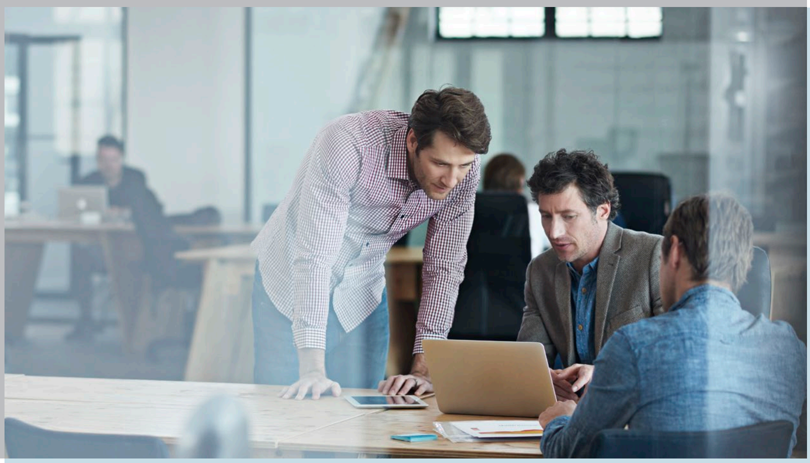
RESEARCH

Swiss Finance Institute Research Paper Series N°16-67

Machine Learning in Individual Claims Reserving



Mario V. WÜTHRICH
ETH Zurich and Swiss Finance Institute

Machine Learning in Individual Claims Reserving

Mario V. Wüthrich[*][†]

November 11, 2016

Abstract

Machine learning techniques make it feasible to calculate claims reserves on individual claims data. This paper illustrates how these techniques can be used by providing an explicit example in individual claims reserving.

Keywords. Individual claims data, individual claims reserving, micro-level stochastic reserving, regression tree, machine learning.

1 Introduction

Traditional calculations of claims reserves are based on *aggregate claims data* structured in claims development triangles. The most commonly used triangular methods on aggregate claims data are the chain-ladder method (see Mack [8]) and the Bornhuetter-Ferguson method [2]. As emphasized in many contributions, these methods on aggregate claims data neglect detailed information about the individual claims behavior. Therefore, recent research strongly promotes claims reserving on *individual claims data*, see, for instance, Antonio-Plat [1], Hiabu et al. [5], Jessen et al. [7], Martínez-Miranda et al. [9], Pigeon et al. [10] or Taylor et al. [11]. Many contributions that are based on individual claims data assume a rather fixed structural form, for example, Pigeon et al. [10] fit a multivariate skew normal distribution to the claims payments. Such fixed structural forms are not very flexible and the consideration of detailed feature information, such as claim code, claim diagnosis, lawyers involved, question of guilt information, etc., is difficult to be implemented in these rigid approaches.

Nowadays, machine learning techniques are very popular in data analytics. These methods are highly flexible and can deal with any sort of structured and unstructured information. The aim of this contribution is to illustrate how machine learning techniques can be used for individual claims reserving. For pedagogical reasons we restrict ourselves to a toy example with the following two main simplifications:

(S1) we only consider the *number of payments* (and not the claim amounts paid);

(S2) we only consider *regression trees* (and no other machine learning techniques).

[*]ETH Zurich, RiskLab, Department of Mathematics, 8092 Zurich, Switzerland

[†]Swiss Finance Institute Professor

Once this pedagogical example is fully understood, we do not see any further difficulty to generalize it to the full real world problem. Concerning item (S1), this requires to generalize the number of payment model to a compound model with an indicator modeling whether there is a payment and with a severity variable modeling the amount in case of a payment. The severity model can either be chosen non-parametric (using the sums of squares as loss function) or with a parametric model (using the corresponding deviance statistics as loss function). In view of item (S2), there are many other powerful machine learning techniques which can easily be adapted to the claims reserving context (bagging, random forest, boosting). For an overview and more insight into the generalization of (S1) and (S2) we refer to Hastie et al. [4] and James et al. [6].

Outline of this paper. In Section 2 we introduce the claims reserving problem and we define the model considered. In Section 3 we specify the feature space used in our example. This feature space describes the available individual claims covariates which are used in the regression tree calibration of the model. In Section 4 we provide an explicit numerical example and in Section 5 we conclude and give an outlook.

2 The problem and the model considered

We give a brief introduction to the claims reserving problem, for an extended description we refer to Chapter 1 of Wüthrich-Merz [15]. A non-life insurance claim occurs at an accident date

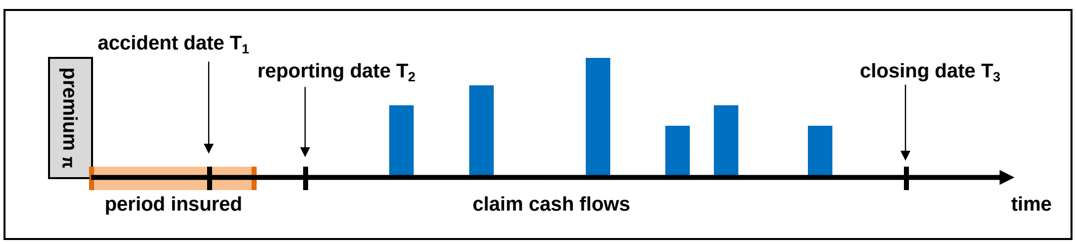


Figure 1: Non-life insurance claim with accident date T_1, reporting date $T_2 > T_1$ and closing date $T_3 > T_2$. There are claim payments (cash flows) during the settlement period $[T_2, T_3]$.

T_1 which lies within the period insured, see Figure 1. This insurance claim is reported to the insurance company at the reporting date $T_2 \geq T_1$. During the time period $[T_1, T_2)$ such a claim is called an *incurred but not yet reported* (IBNYR) claim. After reporting, the claim is settled by paying the compensations for the financial damage occurred. Finally, the claim is closed at the closing date $T_3 \geq T_2$, once all compensations are paid. During the time period $[T_2, T_3]$ such a claim is called a *reported but not settled* (RBNS) claim. This is the time period within which the claim payments are done and during which more and more individual information of that specific claim becomes available. This illustrates that the claim settlement process is a continuous time process. For modeling purposes we restrict ourselves to a discrete time setting. For simplicity we assume an annual discrete time grid.

Assume we have I years of exposures denoted by $1 \leq i \leq I$ and called *accident years*. By $N_{i,j}$ we denote the number of claims occurred in accident year i reported in calendar year $i+j$, thus, $j \in \mathbb{N}_0$ denotes the annualized *reporting delay*. The total number of reported claims of accident year i after reporting delay j is given by

$$M_{i,j} = \sum_{l=0}^{j} N_{i,l}.$$

Typically, one assumes a maximal (possible) reporting delay $J \in \mathbb{N}_0$. Then, $M_{i,J}$ denotes the total number of claims occurred in accident year i. Assume that $\mathcal{F}_t$ denotes the information available at time $t \in \mathbb{N}_0$ and that $(N_{i,j})_{i,j}$ is adapted to that information. Then, we have conditionally expected total number of claims at time t for accident year $i \in \{(t-J)\vee 1, \ldots, t \wedge I\}$

$$\mathbb{E}\left[M_{i,J} \middle| \mathcal{F}_t\right] = M_{i,t-i} + \sum_{j=t-i+1}^{J} \mathbb{E}\left[N_{i,j} \middle| \mathcal{F}_t\right]. \tag{2.1}$$

The first term $M_{i,t-i}$ on the right-hand side of (2.1) gives the reported claims (closed and RBNS claims) and the second term gives the conditionally expected number of IBNYR claims.
We aim at analyzing and predicting the claims payments that are generated by these claims. Fix accident year i and reporting delay j. We have $N_{i,j}$ claims for this tuple (i,j). Every such claim $\nu = 1, \ldots, N_{i,j}$ generates claim cash flows denoted by

$$X_{i,j|0}^{(\nu)},\ X_{i,j|1}^{(\nu)},\ \ldots.$$

For $k \in \mathbb{N}_0$, the variable $X_{i,j|k}^{(\nu)}$ denotes the payments done in calendar year $i+j+k$ for this particular claim ν. If these claims cash flows are also adapted to the information $\mathcal{F}_t$, then we predict the total (nominal) payments for reported claims of accident year i at time t by

$$\sum_{j=0}^{t-i} \sum_{\nu=1}^{N_{i,j}} \left(\sum_{k=0}^{t-(i+j)} X_{i,j|k}^{(\nu)} + \sum_{k>t-(i+j)} \mathbb{E}\left[X_{i,j|k}^{(\nu)} \middle| \mathcal{F}_t \right] \right). \tag{2.2}$$

For closed claims the last summation in (2.2) is equal to 0 (because we do not expect further payments), and for RBNS claims this last term is typically positive. The total (nominal) payments for IBNYR claims of accident year i at time t are predicted by

$$\sum_{j=t-i+1}^{J} \mathbb{E}\left[\sum_{\nu=1}^{N_{i,j}} \sum_{k\geq 0} X_{i,j|k}^{(\nu)} \middle| \mathcal{F}_t \right]. \tag{2.3}$$

In view of simplification (S1) we set for the payments

$$Y_{i,j|k}^{(\nu)} = \mathbb{1}_{\left\{X_{i,j|k}^{(\nu)} \neq 0\right\}}, \tag{2.4}$$

and we replace these variables in (2.2) and (2.3) accordingly. We are now ready to state the model assumptions.

Model Assumptions 2.1 *Assume* $(\Omega, \mathcal{F}, \mathbb{P}, \mathbb{F})$ *is a sufficiently rich filtered probability space with discrete time filtration* $\mathbb{F} = (\mathcal{F}_t)_{t \in \mathbb{N}_0}$. *All processes* $(N_{i,j})_{i,j}$ *and* $(Y^{(\nu)}_{i,j|k})_{i,j,k,\nu}$ *are* $\mathbb{F}$*-adapted for time indexes* $t = i + j$ *and* $t = i + j + k$, *respectively. We make the following assumptions:*

(a1) $(N_{i,j})_{i,j}$ *and* $(Y^{(\nu)}_{i,j|k})_{i,j,k,\nu}$ *are independent;*

(a2) the random variables in $(N_{i,j})_{i,j}$ *and* $(Y^{(\nu)}_{i,j|k})_{i,j,k,\nu}$ *are independent for different accident years* $1 \leq i \leq I$*;*

(a3) the processes $(Y^{(\nu)}_{i,j|k})_{k \geq 0}$ *are independent for different reporting delays* $0 \leq j \leq J$ *and different claims* $\nu \geq 1$*; for the conditional distribution of* $Y^{(\nu)}_{i,j|k+1}$, *given* $\mathcal{F}_{i+j+k}$, *we assume*

$$Y^{(\nu)}_{i,j|k+1} \Big|_{\mathcal{F}_{i+j+k}} \sim \text{Bernoulli}\left(p_{j+k}\left(\mathbf{x}^{(\nu)}_{i,j|k}\right)\right), \tag{2.5}$$

where $\mathbf{x}^{(\nu)}_{i,j|k}$ *is the* $\mathcal{F}_{i+j+k}$*-measurable feature of this claim, contained in the feature space* $\mathcal{X}$, *and* $p_{j+k} : \mathcal{X} \to [0, 1]$ *is a probability function (throughout we assume* $1 \leq i \leq I$, $0 \leq j \leq J$, $1 \leq \nu \leq N_{i,j}$ *and* $k \geq 0$*).*

Remarks. We comment on Model Assumptions 2.1.

- The independence assumption in (a1) is a necessary assumption to receive compound distributions, see Model Assumptions 1.1 in Wüthrich-Buser [14].

- The independence assumption in (a2) implies that claims of different accident years i can be modeled independently. This is a rather common assumption in claims reserving, though a bit weak in view of claims inflation.

- The independence assumption in (a3) implies that we can model single claims cash flows

$$\mathbf{Y}^{(\nu)}_{i,j} = \left(Y^{(\nu)}_{i,j|0},\ Y^{(\nu)}_{i,j|1},\ \ldots\ \right)$$

  independently for all claims ν.

- The crucial assumption for the individual claims modeling is formula (2.5). It says that for every reported claim up to time $t = i + j + k$, there exists an $\mathcal{F}_{i+j+k}$-measurable feature $\mathbf{x}^{(\nu)}_{i,j|k} \in \mathcal{X}$ which completely determines the conditional distribution of $Y^{(\nu)}_{i,j|k+1}$ by

$$\mathbb{P}\left[\left. Y^{(\nu)}_{i,j|k+1} = 1 \right| \mathcal{F}_{i+j+k}\right] = p_{j+k}\left(\mathbf{x}^{(\nu)}_{i,j|k}\right).$$

  At the moment, the (regression) probability function $p_{j+k} : \mathcal{X} \to [0, 1]$ is fully non-parametric and can have any form. We calibrate this regression function p_{j+k} with machine learning techniques. In view of simplification (S2) we use regression trees for this task. Also note that we did not specify an explicit model for the first payments $Y^{(\nu)}_{i,j|0}$, corresponding to index $k = 0$.

- Dropping the independence assumptions in (a1)-(a3) is possible and will lead to more complicated feature spaces $\mathcal{X}$ than the one described in the next section.

3 Feature space and regression tree

3.1 Feature space and dynamic modeling

We first discuss a natural choice of the feature space under Model Assumptions 2.1. The features consist of S-dimensional vectors

$$\mathbf{x}_{i,j|k}^{(\nu)} = \left(x_{i,j|k}^{(\nu)}(s)\right)'_{s=1,\ldots,S} = \left(x_{i,j|k}^{(\nu)}(1), \ldots, x_{i,j|k}^{(\nu)}(S)\right)' \in \mathcal{X},$$

described in the following. Some feature components $x_{i,j|k}^{(\nu)}(s)$ are categorical (for instance, a claim diagnosis), other feature components are either ordered or continuous (for instance, the reporting delay j of a claim ν). As described in Taylor et al. [11], there are static components (that remain constant over the whole life time of the claim) and there are dynamic components (that may (randomly) change over time). Since we collect more and more information over time, typically, the "dimension" of the feature $\mathbf{x}_{i,j|k}^{(\nu)}$ increases over time. This is illustrated by empty components in the following table. In our example below we have the following feature components available, we abbreviate $\mathbf{x}_{i,j|k}^{(\nu)} = \mathbf{x} = (x(1), \ldots, x(S))' \in \mathcal{X}$,

$$\begin{array}{rcl}
x(1) & : & \text{claim code } (\texttt{cc}) \text{ stating type of claim, static categorical feature;} \\
x(2) & : & \text{diagnosis code } (\texttt{diag}) \text{ stating type of injury, static categorical feature;} \\
x(3) & : & \text{lawyer involved } (\texttt{law})\text{, static categorical feature;} \\
x(4) & : & \text{reporting delay } (\texttt{j})\text{, static ordered feature } j \in \mathbb{N}_0; \\
x(5) & : & \text{closed } (\texttt{cl}) \text{ at time } i+j+k\text{, dynamic categorical feature;} \\
x(6) & = & 0 \text{ (only present if } j>0); \\
 & \vdots & \\
x(6+j-1) & = & 0 \text{ (only present if } j>0); \\
x(6+j) & : & \text{payment } (\texttt{Yj+0}) \text{ at time } i+j\text{, dynamic numerical feature;} \\
x(6+j+1) & : & \text{payment } (\texttt{Yj+1}) \text{ at time } i+j+1\text{, dynamic numerical feature;} \\
 & \vdots & \\
x(6+j+k) & : & \text{payment } (\texttt{Yj+k}) \text{ at time } i+j+k\text{, dynamic numerical feature;} \\
x(6+j+k+1) & : & \text{empty;} \\
 & \vdots & \\
x(S) & : & \text{empty.}
\end{array} \tag{3.1}$$

Discussion of (3.1)**.**

- The first three feature components $x(1), x(2), x(3)$ are for simplicity assumed to be static. However, in practice, they may also be dynamic, for instance, lawyers may only be involved in later states of the claim settlement process (which may require a dynamic modeling).

- Feature component $x(5)$ indicates whether the claim is open or closed at time $i+j+k$. This status changes over time and needs stochastic modeling. Here, we assume that only the last state is relevant for prediction (which implicitly assumes a Markov property).

- Feature components $x(6), \ldots, x(6+j+k)$ model the payment history

$$\left(x_{i,j|k}^{(\nu)}(6), \ldots, x_{i,j|k}^{(\nu)}(6+j+k)\right)' = \left(0, \ldots, 0, Y_{i,j|0}^{(\nu)}, \ldots, Y_{i,j|k}^{(\nu)}\right)' \in \{0,1\}^{j+k+1} \tag{3.2}$$

  up to time $i+j+k$ of this claim (where the initial zeros correspond to the reporting delay j and are only present if $j>0$). This is the part of the claim that is $\mathcal{F}_{i+j+k}$-measurable. Feature components $x(6+j+k+1), \ldots, x(S)$ are kept empty, because these are only going to be observed later in the claims development process and are not available for prediction at time $i+j+k$.

- The feature $\mathbf{x}_{i,j|k}^{(\nu)}$ does not include any accident year i specific information. For this reason, the regression function $\mathbf{x}_{i,j|k}^{(\nu)} \mapsto p_{j+k}(\mathbf{x}_{i,j|k}^{(\nu)})$ can be applied to *any* accident year. On the other hand, this regression function already receives some structure, namely, it assumes that the delay parameter $j+k$ is the crucial one (this follows by considering (3.2) as feature components), and reporting delay j is considered as covariate $x(4)$. If this is not an appropriate structure, we need to choose a different feature space $\mathcal{X}$.

From the previous considerations we see that if we would like to make a *multi-period forecast* we need to model stochastically *all* dynamic random feature components. In our example this includes besides the payments $\mathbf{Y}_{i,j}^{(\nu)}$ also an indicator process

$$\mathbf{Z}_{i,j}^{(\nu)} = \left(Z_{i,j|0}^{(\nu)},\ Z_{i,j|1}^{(\nu)},\ \ldots\ \right),$$

which shows whether the claim is closed $Z_{i,j|k}=1$ at time $i+j+k$ or not. For this reason we extend Model Assumptions 2.1 by assuming that in all assumptions we can replace the random variables $Y_{i,j|k}^{(\nu)}$ by the random vectors $(Y_{i,j|k}^{(\nu)}, Z_{i,j|k}^{(\nu)})'$, and (2.5) is replaced by

$$\mathbb{P}\left[Y_{i,j|k+1}^{(\nu)} = y,\ Z_{i,j|k+1}^{(\nu)} = z \middle| \mathcal{F}_{i+j+k}\right] \;=\; p_{j+k}^{(y,z)}\left(\mathbf{x}_{i,j|k}^{(\nu)}\right) \;\geq\; 0, \tag{3.3}$$

for $y, z \in \{0,1\}$ and

$$\sum_{y,z\in\{0,1\}} p_{j+k}^{(y,z)}\left(\mathbf{x}_{i,j|k}^{(\nu)}\right) = 1. \tag{3.4}$$

If we choose feature components (3.1), we implicitly assume that the indicator $\mathbf{Z}_{i,j}^{(\nu)}$ has a Markov property, and we set $x_{i,j|k}^{(\nu)}(5) = Z_{i,j|k}^{(\nu)}$ for the calculation in (3.3).

Our model is now fully specified and we calibrate the regression function (3.3), given by

$$\left(p_{j+k}^{(y,z)}\right)_{y,z\in\{0,1\}} : \mathcal{X} \to [0,1]^4, \tag{3.5}$$

under side constraint (3.4) for time lags $\ell = j+k \geq 0$.

3.2 Regression trees and `rpart` routine

We estimate the regression functions (3.5) in a non-parametric way using classification and regression tree (CART) techniques. These path-breaking techniques in machine learning go back to Breiman et al. [3]. The idea is to successively partition the feature space $\mathcal{X}$ into rectangles by solving *standardized binary split* questions, see Section 2.4.1 and Figure 2.9 in Breiman et al. [3]. The split which reduces a given impurity measure most is the one chosen for the next binary split (in an iterative tree growing algorithm), see Section 2.4 in Breiman et al. [3]. This way, a large binary tree is grown in a first step. In a second step the appropriate tree size is determined by cross-validation and the initially large binary tree is pruned to that size, see Chapter 3 in Breiman et al. [3]. In our example we choose the Gini index as impurity measure and the one-standard-error rule is applied for the choice of the optimal tree size, see Therneau et al. [12] and Section 9.2 in Hastie et al. [4].
The standardized binary split tree growing algorithm is already implemented in R and described in Therneau et al. [12]. It uses the following R command

```
> tree <- rpart(YZ ∼ cc + diag + law + j + cl + Y0 + ... + Yj+k,
                data=dat_t, method=''class'',
                control=rpart.control())
```

This command says that we regress the responses $(Y^{(\nu)}_{i,j|k+1}, Z^{(\nu)}_{i,j|k+1})'$ from the features $\mathbf{x}^{(\nu)}_{i,j|k}$. The data `dat_t` used at time t is given by all observations that are $\mathcal{F}_t$-measurable, i.e. which have indexes $i+j+k+1 \leq t$, see also (4.1) below. The method applied is classification (`''class''`) with the Gini index as impurity measure. The term `control` specifies further parameters, for details we refer to Therneau et al. [12]. In order to apply `rpart` we need to make the responses one-dimensional. This is achieved by setting

$$\texttt{YZ} \;=\; Y^{(\nu)}_{i,j|k+1} + 2Z^{(\nu)}_{i,j|k+1} \;\in\; \{0,1,2,3\}.$$

Note that `YZ` uniquely determines the values of $Y^{(\nu)}_{i,j|k+1}$ and $Z^{(\nu)}_{i,j|k+1}$. The resulting tree can be plotted by `plot.rpart(tree)`, and the optimal tree size can be determined with `plotcp(tree)` and `printcp(tree)`. For more information we refer to the corresponding manual `Package 'rpart'` from the CRAN R server.

4 Individual claims reserving analysis

4.1 Available data

For our data analysis we have 87,944 observed claims in liability insurance over a time period of $I = 7$ accident years and a maximal development delay of 6 years. That is, we have claims information available for indexes $i+j+k \leq I = 7$, which corresponds at time $t = I$ to

$$\mathcal{F}_t \;\stackrel{\text{def.}}{=}\; \sigma\left\{Y^{(\nu)}_{i,j|k}, \mathbf{x}^{(\nu)}_{i,j|k}, Z^{(\nu)}_{i,j|k};\; i+j+k \leq t,\; \nu \leq N_{i,j}\right\}, \tag{4.1}$$

and we aim at predicting

$$\mathcal{F}_t^c = \left\{ Y_{i,j|k}^{(\nu)}, \mathbf{x}_{i,j|k}^{(\nu)}, Z_{i,j|k}^{(\nu)};\ i+j+k > t,\ \nu \le N_{i,j} \right\}. \tag{4.2}$$

Formula (4.1) gives the formal definition of the filtration $\mathbb{F}$ used in our example. The claim code (`cc`) has 11 different values and the diagnosis code (`diag`) has 25 different values; the marginal empirical distributions are given in Figure 4, below. The numbers of reported claims $N_{i,j}$, for $i+j \le I = 7$, are given in Table 7, below. These are exactly the underlying claims that generate the information $\mathcal{F}_I$ in (4.1), and they correspond to the closed and RBNS claims at time $I = 7$.

4.2 Model calibration from closed and RBNS claims

We apply the standardized binary split tree growing algorithm to calibrate the probabilities $p_\ell^{(y,z)}$ of the observed development periods $\ell = 0, \ldots, L-1 = 5$ at time $t = I$, where we set $\ell = j + k$, and $L = 6$ is the maximal observed development delay (time lag). Thus, we consider for development periods $\ell = j + k$

$$\mathbb{P}\left[Y_{i,j|k+1}^{(\nu)} = y,\ Z_{i,j|k+1}^{(\nu)} = z \middle| \mathcal{F}_{i+j+k} \right] \ = \ p_{\ell=j+k}^{(y,z)} \left(\mathbf{x}_{i,j|k}^{(\nu)} \right),$$

with $y, z \in \{0,1\}$ and for reported claims having $i + j + k + 1 \le I = 7$. These claims generate information $\mathcal{F}_I$ in (4.1).

For illustration we fix time lag $\ell = j + k = 1$. This maps the reported claims in column $\ell = 1$ to the next column $\ell + 1 = 2$ of the claims development process. For the calibration of $p_{\ell=1}^{(y,z)}$ we consider the information $(Y_{i,0|2}^{(\nu)}, Z_{i,0|2}^{(\nu)}, \mathbf{x}_{i,0|1}^{(\nu)})$ and $(Y_{i,1|1}^{(\nu)}, Z_{i,1|1}^{(\nu)}, \mathbf{x}_{i,1|0}^{(\nu)})$ provided by the (reported) claims of accident years $1 \le i \le I - 2 = 5$. We grow a very large standardized binary split tree based on this information, see Chapter 5 in Wüthrich-Buser [14]. The results are presented in Figure 2. The left-hand side gives the resulting binary tree with the different colors indicating the likelihoods of the estimated probabilities in the leaves:

$$\max_{y,z \in \{0,1\}} \widehat{p}_{\ell=1}^{(y,z)}(\star),$$

where $\widehat{p}_{\ell=1}^{(y,z)}(\star)$ estimates the probabilities for $y, z \in \{0,1\}$ on the resulting leaves of the tree (abbreviated by $\star$). The right-hand side provides the cost-complexity plot `plotcp(tree)` from which the optimal tree size can be selected, for details we refer to Therneau et al. [12]. If we apply the one-standard-error rule (red line in the plot on the right-hand side of Figure 2) we obtain an optimal tree with 11 leaves. This optimal tree is given in Figure 3. We observe that we ask the following split questions to obtain the 11 leaves:

$$\{\texttt{cl} = \text{yes}\}, \qquad \{\texttt{law} = \text{yes}\}, \qquad \{\texttt{cc} \in \{1,3,4,5,7,8\}\}, \qquad \{\texttt{Y1} = 0\},$$
$$\{\texttt{cc} \in \{1,5,7,8\}\}, \qquad \{\texttt{diag} \in \{3,6,9,10,17\}\}, \qquad \{\texttt{diag} \in \{9,16,21\}\},$$
$$\{\texttt{j} = 1\}, \qquad \{\texttt{cc} \in \{1,3,4,5,6\}\}, \qquad \{\texttt{diag} \in \{7,12,16,23,25\}\}.$$

Thus, except of the feature component `Y0` all information is used in the optimal tree construction for time lag $\ell = 1$, see Figure 3. The estimated probabilities $\widehat{p}_{\ell=1}^{(y,z)}(\star)$ on the 11 leaves $\star$ can be

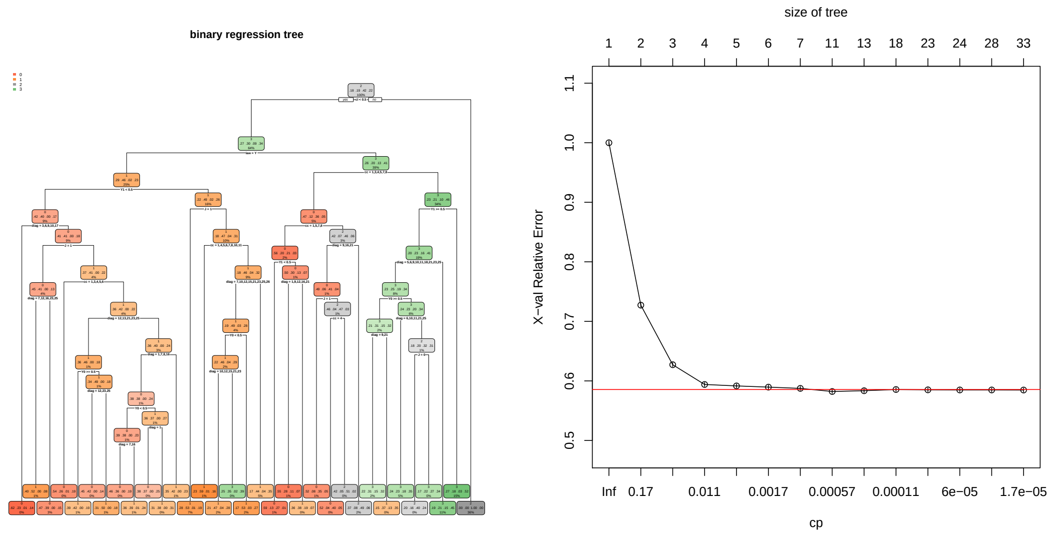


Figure 2: For $\ell = 1$: (lhs) large standardized binary split tree with 33 leaves (the different colors indicate the likelihoods in the leaves: red for $\texttt{YZ} = 0$, orange for $\texttt{YZ} = 1$, gray for $\texttt{YZ} = 2$ and green for $\texttt{YZ} = 3$; (rhs) 10-fold cross-validation results displayed by `plotcp(tree)`, the red line shows the one-standard-error rule.

read off from the lowest level of the resulting tree in Figure 3. Basically, what we can see from this plot is that claims without a lawyer involved are likely closed (green and gray). If lawyers are involved the claims remain open longer. If we would ignore all feature information we would get estimated probabilities (root node of the tree in Figure 3)

$$\left(\bar{p}_{\ell=1}^{(0,0)}, \bar{p}_{\ell=1}^{(1,0)}, \bar{p}_{\ell=1}^{(0,1)}, \bar{p}_{\ell=1}^{(1,1)}\right) = (17.6\%, 19.3\%, 41.5\%, 21.6\%)\,.$$

We call these latter estimates homogeneous estimates (because they do not consider any feature information) and we will denote them by $\bar{p}_\ell^{(y,z)}$.

Having estimates $\widehat{p}_\ell^{(y,z)}(\star)$ for all $y, z \in \{0, 1\}$, $\ell = 0, \ldots, L-1 = 5$ and on all leaves $\star$, the model is calibrated and can be used to predict the future payments of reported claims (RBNS claims and closed claims because the latter may be re-opened under certain circumstances). In Table 1 we present the resulting optimal numbers of leaves for time lags $\ell = 0, \ldots, 5$ which we have obtained in this calibration. We also provide the feature components considered for obtaining these splits. We observe that all feature components are used except the payment indicator `Y0` of the first period $\ell = j + k = 0$. Reporting delay `j` becomes less important for more developed claims. We also observe that we should not assume a Markov property in the payment indicators $(Y_{i,j|k}^{(\nu)})_{k\geq 0}$ because for time lags $\ell = 2, 3$ we use the two last payment observations.

4.3 Individual claims prediction for closed and RBNS claims

The calibration $(\widehat{p}_\ell^{(y,z)})_{y,z,\ell}$ at time $t = I = 7$, obtained in the previous section, can now be used to predict the total number of payments for the reported claims (closed and RBNS claims). This

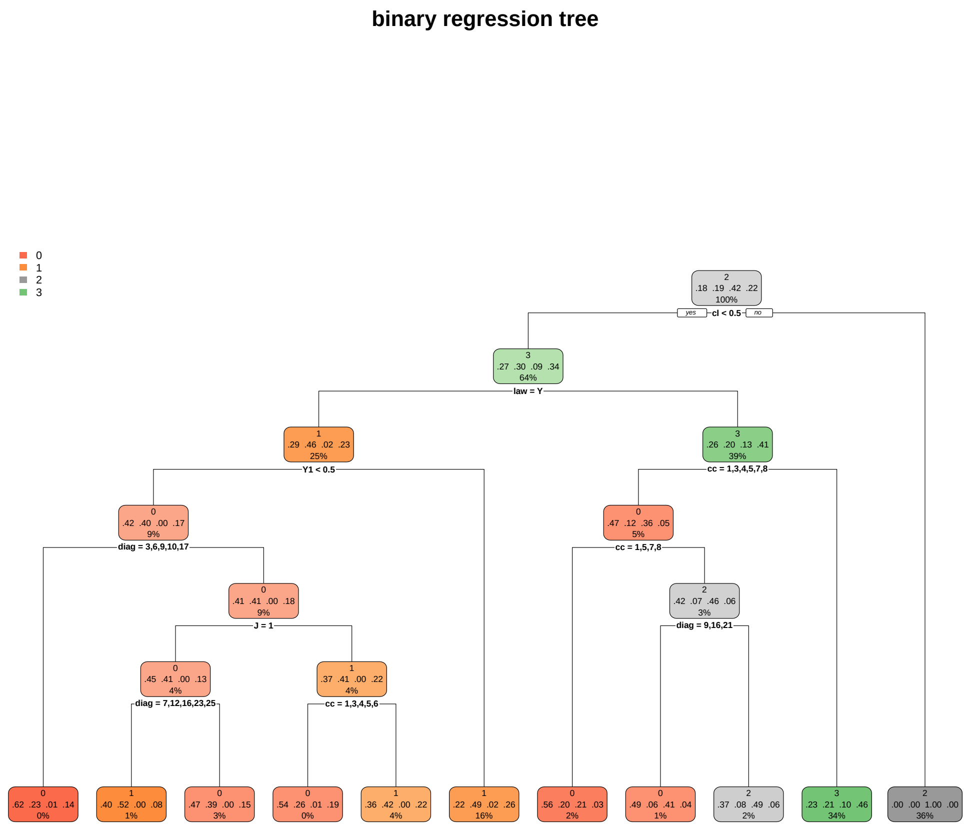


Figure 3: Optimal tree for $\ell = 1$ resulting from the one-standard-error rule.

time lag ℓ	0	1	2	3	4	5
numbers of leaves	8	11	18	12	4	4
	components used for split questions					
	`cl`	`cl`	`cl`	`cl`	`cl`	`cl`
		`law`	`law`	`law`		
	`cc`	`cc`	`cc`	`cc`	`cc`	
	`diag`	`diag`	`diag`	`diag`		`diag`
		`j`	`j`	`j`		
	`Y0`	`Y1`	`Y2`	`Y3`	`Y4`	`Y5`
			`Y1`	`Y2`		

Table 1: Resulting optimal numbers of leaves for time lags $\ell = 0, \ldots, 5$ and feature components considered for these standardized binary splits.

corresponds to the prediction of (we abbreviate closed and RBNS claims by "RBNS")

$$\mathcal{F}_I^{\mathrm{RBNS}} = \left\{ Y_{i,j|k}^{(\nu)}, \mathbf{x}_{i,j|k}^{(\nu)}, Z_{i,j|k}^{(\nu)};\ i+j+k > I,\ \nu \leq N_{i,j},\ i+j \leq I \right\} \ \subset\ \mathcal{F}_I^c.$$

Note that this differs from $\mathcal{F}_I^c$ given in (4.2), because we insert another indicator $i+j \le I$ to ensure that this refers to reported claims. Choose such a reported claim ν with $i+j+k = I = 7$, thus, its feature $\mathbf{x}_{i,j|k}^{(\nu)}$ is observed at time $I = 7$. For this claim we need to predict

$$\left((Y_{i,j|k+1}^{(\nu)}, Z_{i,j|k+1}^{(\nu)})',\ (Y_{i,j|k+2}^{(\nu)}, Z_{i,j|k+2}^{(\nu)})',\ \ldots \right).$$

Choose $k' > k$, then we predict $(Y_{i,j|k'}^{(\nu)}, Z_{i,j|k'}^{(\nu)})'$ at time $t = I$ by

$$\begin{aligned}
\mathbb{E}\left[\left(Y_{i,j|k'}^{(\nu)}, Z_{i,j|k'}^{(\nu)} \right)' \middle| \mathcal{F}_I \right] &= \mathbb{E}\left[\mathbb{E}\left[\left(Y_{i,j|k'}^{(\nu)}, Z_{i,j|k'}^{(\nu)} \right)' \middle| \mathcal{F}_{i+j+k'-1} \right] \middle| \mathcal{F}_I \right] \\
&= \mathbb{E}\left[\sum_{y,z \in \{0,1\}} (y,z)'\ p_{i+j+k'-1}^{(y,z)} \left(\mathbf{x}_{i,j|k'-1}^{(\nu)} \right) \middle| \mathcal{F}_I \right] \\
&= \sum_{y,z \in \{0,1\}} (y,z)'\ \mathbb{E}\left[p_{i+j+k'-1}^{(y,z)} \left(\mathbf{x}_{i,j|k'-1}^{(\nu)} \right) \middle| \mathcal{F}_I \right]. \qquad (4.3)
\end{aligned}$$

This reduces the time index $i+j+k'$ by 1 to $i+j+k'-1$. This is then iterated for the calculation of the last conditionally expectations in (4.3) until we obtain $\mathcal{F}_I$-measurable terms. Replacing the probabilities $p_\ell^{(y,z)}$ by their $\mathcal{F}_I$-measurable calibrations $\widehat{p}_\ell^{(y,z)}$ allows us to estimate (4.3) and predict $(Y_{i,j|k'}^{(\nu)}, Z_{i,j|k'}^{(\nu)})'$, respectively. Since the analytical calculations of (4.3) are a bit cumbersome we evaluate these terms by simulations. This provides for every single reported claim in (2.2) a prediction $\widehat{C}_{i,j}^{(\nu)}$ at time $t = I$ for the total number of payments:

$$\widehat{C}_{i,j}^{(\nu)} = \sum_{k=0}^{I-(i+j)} Y_{i,j|k}^{(\nu)} + \sum_{k=I-(i+j)+1}^{L-j} \widehat{\mathbb{E}}\left[Y_{i,j|k}^{(\nu)} \middle| \mathcal{F}_I \right],$$

where $\widehat{\mathbb{E}}$ denotes the expectation obtained from the calibration $(\widehat{p}_\ell^{(y,z)})_{y,z,\ell}$ at time I. The total number of payments for closed and RBNS claims of accident year $1 \le i \le I$ at time $t = I$ is predicted by

$$\widehat{C}_i^{\mathrm{RBNS}} = \sum_{j=0}^{I-i} \sum_{\nu=1}^{N_{i,j}} \widehat{C}_{i,j}^{(\nu)},$$

and aggregated over all accident years this provides prediction at time I

$$\widehat{C}^{\mathrm{RBNS}} = \sum_{i=1}^{I} \sum_{j=0}^{I-i} \sum_{\nu=1}^{N_{i,j}} \widehat{C}_{i,j}^{(\nu)}.$$

The results are presented in Table 2. The column 'regression tree' considers the full feature space $\mathcal{X}$ with estimated probabilities $(\widehat{p}_\ell^{(y,z)}(\star))_{y,z,\ell}$ on the leaves $\star$, the column 'homogeneous' does not consider any feature information with resulting homogeneous estimates $(\bar{p}_\ell^{(y,z)})_{y,z,\ell}$ (obtained from the root nodes of the corresponding trees). We see clear differences in the results which show that the feature consideration is relevant for individual claims prediction. In fact, in the regression tree model we obtain an *individual prediction* $\widehat{C}_{i,j}^{(\nu)}$ for each claim ν, adapted to its feature. This is important information for insurance pricing if, for instance, a product consists of different hazard covers.

accident year i	regression tree	homogeneous
1	20,311	20,311
2	20,311	20,276
3	21,330	21,336
4	26,500	26,406
5	27,710	28,267
6	26,057	26,986
7	20,347	23,599
total (closed and RBNS)	162,567	167,181

Table 2: Prediction of the total number of payments $\widehat{C}_i^{\mathrm{RBNS}}$ for accident years $1 \le i \le 7$ at time $I = 7$: 'regression tree' considers the full feature space $\mathcal{X}$ and estimates $\widehat{p}_\ell^{(y,z)}(\star)$ on the resulting leaves $\star$; 'homogeneous' does not consider any feature information resulting in estimates $\bar{p}_\ell^{(y,z)}$.

4.4 Prediction for IBNYR claims

In the previous section we have predicted the numbers of payments for closed and RBNS claims. We now turn to the IBNYR claims at time I which are given by

$$\mathcal{F}_I^{\mathrm{IBNYR}} = \left\{ Y_{i,j|k}^{(\nu)}, \mathbf{x}_{i,j|k}^{(\nu)}, Z_{i,j|k}^{(\nu)};\ i+j+k > I,\ \nu \le N_{i,j},\ i+j > I \right\} \ \subset\ \mathcal{F}_I^c.$$

For this set the total numbers of reported claims $N_{i,j}$ are not yet observed. Thus, we also have to predict these random variables, see also (2.3). For this prediction we need an additional model structure. If we assume that the claims occurrence and reporting process can be described by a homogeneous marked Poisson point process, then these numbers of reported claims can be predicted by the chain-ladder method, for details we refer to Section 6.1 in Verrall-Wüthrich [13]. The corresponding results are presented in Table 3 and the predicted values are highlighted

accident year i	reporting delay j: 0	1	2	3	4	5	6
1	5,718	5,222	366	238	31	8	14
2	6,539	4,619	381	275	24	13	*14*
3	6,993	4,322	443	200	19	*11*	*14*
4	8,259	4,904	483	178	*29*	*12*	*17*
5	9,306	4,780	448	*268*	*31*	*13*	*18*
6	9,373	3,957	*466*	*255*	*29*	*13*	*17*
7	10,831	*6,520*	*607*	*332*	*38*	*16*	*22*

Table 3: Observed and predicted numbers of claims $N_{i,j}$, for $1 \le i \le I$ and $0 \le j \le I-1$, chain-ladder predictions are in *italic*.

in *italic*. Each of these IBNYR claims generates more payments. Under Model Assumptions 2.1 and the additional assumptions that the payments $\mathbf{Y}_{i,j}^{(\nu)}$ are independent of $\mathcal{F}_t$ and i.i.d. for

$i + j > t$ we obtain from (2.3)

$$\sum_{j=t-i+1}^{J} \mathbb{E}\left[\sum_{\nu=1}^{N_{i,j}} \sum_{k\geq 0} Y_{i,j|k}^{(\nu)} \middle| \mathcal{F}_t \right] = \sum_{j=t-i+1}^{J} \mathbb{E}\left[N_{i,j} \middle| \mathcal{F}_t \right] \mathbb{E}\left[\sum_{k\geq 0} Y_{i,j|k}^{(1)} \right].$$

In particular, this means that for these claims we cannot provide individual claim specific predictions because we do not have feature information. There remains to provide estimates of the average numbers of payments

$$c_{i,j} = \mathbb{E}\left[\sum_{k\geq 0} Y_{i,j|k}^{(1)} \right],$$

of IBNYR claims at time $t = I$. Assuming homogeneity in accident year $1 \leq i \leq I$, these average numbers of payments can be estimated at time I from the reported claims by

$$\widehat{c}_j \;=\; \frac{1}{I-j} \sum_{i=1}^{I-j} \frac{1}{N_{i,j}} \sum_{\nu=1}^{N_{i,j}} \widehat{C}_{i,j}^{(\nu)}. \tag{4.4}$$

These estimates are given in the last row of Table 4. The total numbers of payments for IBNYR

accident	reporting delay j						
year i	0	1	2	3	4	5	6
1	1.7732	1.7576	1.5902	1.4706	1.4194	1.6250	0.3571
2	1.7284	1.7079	1.7197	1.5278	1.4038	0.8700	
3	1.7995	1.7936	1.5331	1.4837	0.9947		
4	1.9960	1.8215	1.6688	1.5531			
5	1.9661	1.8247	1.5435				
6	2.0592	1.7075					
7	1.8786						
$\widehat{c}_j$	1.8859	1.7688	1.6110	1.5088	1.2726	1.2475	0.3571

Table 4: Estimated average numbers of payments per claim: individual indexes (i, j) correspond to the last terms in (4.4), the last row of the table corresponds to the empirical averages $\widehat{c}_j$.

claims of accident years $1 \leq i \leq I$ at time $t = I$ are predicted by

$$\widehat{C}_i^{\mathrm{IBNYR}} = \sum_{j=I-i+1}^{J} \widehat{N}_{i,j}^{(CL)} \; \widehat{c}_j,$$

if $\widehat{N}_{i,j}^{(CL)}$ denotes the chain-ladder prediction for $N_{i,j}$ at time I; these are the values in *italic* in Table 3. Aggregated over all accident years this provides prediction at time I

$$\widehat{C}^{\mathrm{IBNYR}} = \sum_{i=1}^{I} \widehat{C}_i^{\mathrm{IBNYR}}.$$

In Table 5 we provide these predictions for the IBNYR claims at time I and we compare the tree regression predictors to the homogeneous ones (not considering the feature components).

accident year i	regression tree	homogeneous
1	0	0
2	5	5
3	19	17
4	58	50
5	467	389
6	1,195	1,030
7	13,087	12,616
total (IBNYR)	14,831	14,107

Table 5: Prediction of the total number of payments $\widehat{C}_i^{\mathrm{IBNYR}}$ for the IBNYR claims of accident years $1 \leq i \leq 7$ at time $I = 7$: 'regression tree' considers the full feature model $\widehat{p}_\ell^{(y,z)}(\star)$; 'homogeneous' does not consider any feature information resulting in estimates $\bar{p}_\ell^{(y,z)}$.

Also here we see that the consideration of the features induces substantial differences between the two predictions.

Finally, we aggregate the predictions of reported (closed and RBNS) claims and IBNYR claims to obtain a prediction of all payments of accident years $1 \leq i \leq I$. This is given in Table 6, and

accident year i	regression tree	homogeneous	chain-ladder
1	20,311	20,311	20,311
2	20,316	20,281	20,269
3	21,349	21,353	21,381
4	26,558	26,456	26,829
5	28,177	28,656	29,429
6	27,252	28,015	30,062
7	33,434	36,215	39,667
total	177,398	181,288	187,947

Table 6: Prediction of the total number of payments of all claims of accident years $1 \leq i \leq 7$ at time $I = 7$: 'regression tree' considers the full feature model $\widehat{p}_\ell^{(y,z)}(\star)$; 'homogeneous' does not consider any feature information resulting in estimates $\bar{p}_\ell^{(y,z)}$; 'chain-ladder' is obtained by Mack's chain-ladder model [8] applied directly to the numbers of payments.

it is also compared to the chain-ladder predictions obtained by applying Mack's chain-ladder model directly to the numbers of payments (ignoring the reporting delay feature component $\mathtt{j}$ and also not differentiating between reported and IBNYR claims). The results show substantial differences between the methods. In particular, for the last accident year $I = 7$ the chain-ladder method seems to over-estimate the claims payments: we have a rather high observed payment ratio $\sum_{v=1}^{N_{i,0}} Y_{i,0|0}^{(\nu)}/N_{i,0}$ for accident year $i = 7$ in the first column $j+k = 0$, but the corresponding feature information says that this ratio cannot directly be projected to the future development periods of that accident year (as done by the chain-ladder method) because this ratio is mainly caused by claims that are closed quickly.

This finishes our pedagogical example. There only remains to remark that the considerations in Table 6 do not consider tail development factors for the claims developments beyond the last observed development period $L = 6$.

5 Summary and outlook

We have illustrated an example in individual claims reserving using regression trees. These regression trees are fully flexible and allow us to consider (almost) any kind of feature information. As a result we obtain *claims reserves on individual claims* respecting all available relevant feature information (which also differentiates types of claims). Our considerations were done under the two simplifications (S1) and (S2) mentioned in the introduction.
The following steps are feasible (and necessary) to make our approach useful in practice:

- The restriction to the payment indicators (2.4) was mainly done for illustrative purposes. A next step considers a compound model that also models the claim amounts in case of payments. In the regression tree construction this implies that the Gini index is replaced by other loss functions such as the sums of squares or the deviance statistics.
- A next extension includes the case reserves in the feature information. Since these case reserves are stochastic processes themselves, this will require joint dynamic modeling of the claims payments and the case reserves processes.
- Regression trees are not known to be very robust towards changes in observations. More robust machine learning techniques, like random forests, boosting or neural networks, should be used to overcome this issue.
- Prediction uncertainty is not analyzed, yet. This task can be completed by bootstrap simulation and bagging. The latter is also a useful method for reducing the variance in the resulting parameter estimates.
- Solvency considerations require a more dynamic viewpoint by analyzing the so-called claims development results. This task can be completed by bootstrap simulation as well. However, at this point we may face some issues about computational time.

Acknowledgment. I would like to kindly thank Christoph Buser (AXA-Winterthur), Simone Elmer (Helsana), Gareth Peters (University College London), Peter Reinhard (AXA-Winterthur) and Patrick Zöchbauer (ETH Zurich) for sharing their views on data science and claims reserving. These views provided the essential pieces that are combined in this work.

References

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

[2] Bornhuetter, R.L., Ferguson, R.E. (1972). The actuary and IBNR. *Proceedings of the Casualty Actuarial Society* **59**, 181-195.

[3] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth Statistics/Probability Series.

[4] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* 2nd edition. Springer Series in Statistics.

[5] Hiabu M., Margraff, C., Martínez-Miranda, M.D., Nielsen, J.P. (2016). The link between classical reserving and granular reserving through double chain ladder and its extensions. *British Actuarial Journal* **21/1**, 97-116.

[6] James, G., Witten, D., Hastie, T., Tibshirani, R. (2015). *An Introduction to Statistical Learning. With Applications in R.* Corrected 6th printing. Springer Texts in Statistics.

[7] Jessen, A.H., Mikosch, T., Samorodnitsky, G. (2011). Prediction of outstanding payments in a Poisson cluster model. *Scandinavian Actuarial Journal* **2011/3**, 214-237.

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

[9] Martínez-Miranda, M.D., Nielsen, J.P., Verrall, R.J., Wüthrich, M.V. (2015). The link between classical reserving and granular reserving through double chain ladder and its extensions. *Scandinavian Actuarial Journal* **2015/5**, 383-405.

[10] Pigeon, M., Antonio, K., Denuit, M. (2013). Individual loss reserving with the multivariate skew normal framework. *ASTIN Bulletin* **43/3**, 399-428.

[11] Taylor, G., McGuire, G., Sullivan, J. (2008). Individual claim loss reserving conditioned by case estimates. *Annals of Actuarial Science* **3/1-2**, 215-256.

[12] Therneau, T.M., Atkinson, E.J., Foundation, M. (2015). An introduction to recursive partitioning using the RPART routines. *R Vignettes*, version of June 29, 2015.

[13] Verrall, R.J., Wüthrich, M.V. (2016). Understanding reporting delay in general insurance. *Risks* **4/3**, 25.

[14] Wüthrich, M.V., Buser, C. (2016). Data analytics for non-life insurance pricing. Lecture Notes, ETH Zurich.

[15] Wüthrich, M.V., Merz, M. (2015). Stochastic claims reserving manual: advances in dynamic modeling. *SSRN Manuscript* ID 2649057.

A Individual claims data statistics

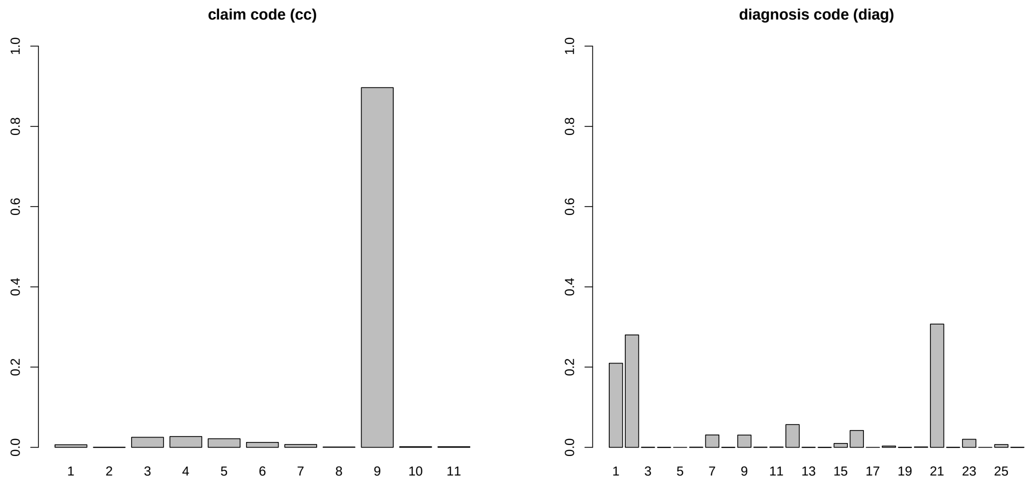


Figure 4: (lhs) marginal distribution of the observed claim codes (`cc`); (rhs) marginal distribution of the observed diagnosis codes (`diag`); the empirical ratio of lawyers involved (`law`) is roughly 25%.

accident	reporting delay j						
year i	0	1	2	3	4	5	6
1	5,718	5,222	366	238	31	8	14
2	6,539	4,619	381	275	24	13	
3	6,993	4,322	443	200	19		
4	8,259	4,904	483	178			
5	9,306	4,780	448				
6	9,373	3,957					
7	10,831						

Table 7: Observed numbers of reported claims $N_{i,j}$, for $i + j \leq I = 7$.

Swiss Finance Institute

Created in 2006 by the Swiss banks, the Swiss Stock Exchange, six leading Swiss Universities and the Swiss Federal government, the Swiss Finance Institute is a unique undertaking merging the experiences of a centuries old financial center with the innovative drive of a frontier research institution. Its goal is to change the research and teaching landscape in areas relevant to banks and financial institutions.

With more nearly 60 full time professors and ca. 80 PhD students, the Swiss Finance Institute represents the premier concentration of expertise in banking and finance across the European continent. The Institute's close affiliation with the Swiss banking industry ensures that its research culture remains in tune with the needs of the financial services sector. Networking events where the participants can meet with local practitioners are therefore also an essential part of the offering.

swiss:finance:institute

c/o University of Geneva, Bd. Du Pont d'Arve 42, CH-1211 Geneva 4
T +41 22 379 84 71, rps@sfi.ch, www.sfi.ch