FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

RIGORÓZNÍ PRÁCE

Vojtěch Bednárik

Loss reserving for individual claim-by-claim data

Department of Probability and Mathematical Statistics

Supervisor of the master thesis:	RNDr. Michal Pešta, Ph.D.
Study programme:	Mathematics
Study branch:	Financial and Insurance Mathematics

Prague 2018

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In Prague, May 28, 2018 Vojtěch Bednárik

Title: Loss reserving for individual claim-by-claim data

Author: Vojtěch Bednárik

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Michal Pešta, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This thesis covers stochastic claims reserving in non-life insurance based on individual claims developments. Summarized theoretical methods are applied on data from Czech Insurers' Bureau created for educational purposes. The problem of estimation is divided into four parts: occurence process generating claims, delay of notification, times between events and payments. Each part is estimated separately based on maximum likelihood theory and final estimates allow us to obtain an estimate of future liabilities distribution. The results are very promising and we believe this method is worth of a further research. Contribution of this work is more rigorous theoretical part and application on data from the Czech market with some new ideas in practical part and simulation.

Keywords: Claims reserving, non-life insurance, micro-level approach, granular data

Název: Rezervování pro individuální škodní data

Autor: Vojtěch Bednárik

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí práce: RNDr. Michal Pešta, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá stochastickým modelováním škod v neživotním pojištění na základě individuálních škodních průběhů. Shrnuté teoretické metody jsou aplikovány na výuková data od České kanceláře pojistitelů. Problematika odhadování je rozdělena na čtyři části: proces výskytů škod, zpoždění v hlášení, časy mezi událostmi a platby. Každá část je odhadnuta samostatně metodou maximální věrohodnosti a konečné odhady nám umožňují získat odhad rozdělení budoucích závazků. Výsledky jsou velice slibné a věříme, že tato metoda je vhodná pro podrobnější výzkum. Příspěvek této práce spočívá ve formálním odvození teoretické části a aplikaci na datech z českého trhu s několika novými nápady v praktické části a simulaci.

Klíčová slova: Škodní rezervování, neživotní pojištěí, mikro-level přístup, granulární data

I would like to thank my supervisor RNDr. Michal Pešta, Ph.D. for a guidance and help with finding feasible ideas and solutions applicable in the thesis. I thank RNDr. Petr Jedlička, Ph.D. (a team leader of actuarial services at SUPIN s.r.o.) for prepared and provided data used in the thesis. I wish to thank my friend Bc. Petra Kochaniková for all her help with the grammatical part of the thesis. I greatly appreciate possibilities enabled by creators of software `R` and its contributors. Last but not least, my gratitude goes to my parents for their tremendous support during my studies and beyond.

Contents

Introduction

When actuaries in a non-life insurance company estimate insurance liabilities, they usually use chain ladder method, if it is possible. Other widely used methods are probably Bornhuetter-Fergusson method or overdispersed Poisson model. These and many other methods have usually one thing in common: estimation of liabilities is based on aggregate data into development triangles. Generally, all aggregate methods are relatively easy to implement and they are quite understandable. On the other hand, aggregate methods have many disadvantages, namely chain ladder method appears in many articles, where its weaknesses are described.

We briefly mention here few problems with chain ladder method, but some of these problems are related to different methods as well and our list is non-exhaustive. Firstly, one of the assumptions of chain ladder is independence of cumulative claims of different accident years. This assumption can be violated for example by change of laws, here is worth to mention the new civil code effective since January 2014 in the Czech Republic, which affected some claims retrospectively. Secondly, development triangles can be affected by a calendar year effect because of changes in the internal rules and RBNS reserve might be set differently in various years. Thirdly, chain ladder can be easily affected by expert judgment, e.g. by excluding some ratios in calculation of development factors, which are calculated as weighted average of (not excluded) ratios. Finally, two different development triangles can be constructed: the first one is paid triangle which contains paid amounts and the second one is incurred triangle containing paid amounts and RBNS reserve development. In many cases these triangles provide very different results and it must be decided which result should be chosen as final.

The thesis is focused on a different approach. Instead of using aggregate data we try to use individual claims developments to describe a selected part of one line of business by a probability distribution. This is done by splitting the problem into four parts, which are estimated separately. The selected part of the line of business can be then described in terms of occurence process generating claims, delay in notification, times between events and finally, payments. One of the advantages of this approach is a limitation of expert judgment, because we can influence only data used for estimation and such decisions should be appropriately justified. Another influence on results is choice of the most suitable distribution in respective parts, however, all choices can be based on an objective criterion, e.g. comparison of maximized likelihoods.

Chapter 1 deals with the theoretical part, which justifies and explains usage of a claim-by-claim model. We derive formally all needed distributions as probability densities, which is one of contributions of the thesis. Theory described in literature is full of informal notation and derivations and it is not an easy task to understand it, especially for readers who are new in this topic. Notation in the thesis is not completely consistent, it can slightly differ chapter by chapter, but everything should be clear within context. Nevertheless, for a better orientation, we attach a list of notation to the end of the thesis.

Chapter 2 describes few necessary adjustments to the theoretical part in order to overcome insufficiencies of our data. Estimation is based on maximum likelihood theory and final estimates are selected based on the largest maximized likelihood. A simplification concerning amounts of payments is used, specifically they are treated as independent and identically distributed random variables. This simplified part of the model could be improved for example by considering a dependence on previous payments.

Chapter 3 explains our simulation algorithm and describes the practical realization in software `R`. We discuss obtained results and compare them with estimates obtained by chain ladder method. Results based on chain ladder can lead to an estimate of distribution of claims reserve via bootstrap and we can compare volatility of two different methods. It can be easily noticed that our claim-by-claim model implies smaller volatility, which is probably caused by more information used for estimation.

The most important parts of our source code can be found in Appendix. They are attached to ilustrate our practical realization of estimation and simulation in more detail, but note that the overall code with many additional analyses is several times longer. Because all chapters are now briefly described, we can move to the first chapter.

1. Theoretical Part

1.1 Nonhomogeneous Poisson Process

Before we define marked Poisson process, we quickly review a generalization of homogeneous Poisson process, since a great part of marked Poisson process is based on nonhomogeneous Poisson process. In case of further interest see for example Ross (2010, pp. 312–345), where you can find all relevant definitions, derivations, proofs and many examples.

Definition 1. *Counting process $\{N_t, t \geq 0\}$ is said to be nonhomogeneous Poisson process with intensity function $\lambda(t)$, $t \geq 0$, if the following holds true:*

1. $N_0 = 0$;
2. $\{N_t, t \geq 0\}$ *has independent increments;*
3. $\mathsf{P}\left[N_{t+h} - N_t \geq 2\right] = o(h)$;
4. $\mathsf{P}\left[N_{t+h} - N_t = 1\right] = \lambda(t)h + o(h)$;

for all $t \geq 0$ and $h > 0$.

The function $o(h)$ in Definition 1 has its usual meaning, i.e. it satisfies

$$\lim_{h \to 0_+} \frac{o(h)}{h} = 0.$$

For $t \in [0, +\infty]$ we define a cumulative hazard function

$$\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s,$$

which determines expected values of increments, as we can see in Theorem 1, where properties of nonhomogeneous Poisson process are summarized. The first two statements relate only to the process $\{N_t, t \geq 0\}$. From a practical point of view, we assume that $\Lambda(\infty)$ is finite.

Theorem 1. *Let $\{N_t, t \geq 0\}$ and $\{M_t, t \geq 0\}$ be independent nonhomogeneous Poisson processes with respective intensity functions $\lambda(t)$, $\mu(t)$, $t \geq 0$. The following then holds:*

1. *For all $t \geq 0$, $h > 0$ random variable $N_{t+h} - N_t$ follows Poisson distribution with parameter $\Lambda(t+h) - \Lambda(t)$.*
2. *Let T_i be time of the ith event, then for $0 \leq t_1 < \ldots < t_n < \infty$ we have*

$$\lim_{(h_1,\ldots,h_n)\to(0_+,\ldots,0_+)} \frac{\mathsf{P}\left[T_i \in [t_i, t_i + h_i) \text{ for all } i{=}1,\ldots,n,\ T_{n+1} = \infty\right]}{h_1 \cdots h_n} = \\ = \mathrm{e}^{-\Lambda(\infty)} \prod_{i=1}^{n} \lambda(t_i). \quad (1.1)$$

3. *If* $N_t^* = N_t + M_t$, *then* $\{N_t^*, t \geq 0\}$ *is nonhomogeneous Poisson process with intensity function* $\lambda(t) + \mu(t)$. *Given that at time* t *occured an event, it belongs to process* $\{N_t, t \geq 0\}$ *with probability*

$$\frac{\lambda(t)}{\lambda(t) + \mu(t)}.$$

Proof. In the following paragraph we will proof only the second part of Theorem 1 (the other properties are proven in already mentioned Ross (2010)), since it will be used later to derive slightly more complex densities. To determine the probability inside the limit, we need to use the first property of nonhomogeneous Poisson process in Theorem 1, i.e. probability that there is no event during time interval $[a, b)$ is equal to

$$\mathrm{e}^{-[\Lambda(b)-\Lambda(a)]}.$$

Further, Definition 1 states that probability of having one event in time interval $[a, a + h)$ is

$$\lambda(a)h + o(h).$$

We realize that the event, we are interested in, can be equivalently reformulated in the following way: there is no event during $[0, t_1)$, one event during $[t_1, t_1 + h_1)$, no event during $[t_1 + h_1, t_2)$, one event during $[t_2, t_2 + h_2)$ and so on until one event during $[t_n, t_n + h_n)$ and no event during $[t_n + h_n, +\infty)$. Using the independence of increments and already mentioned properties we rewrite the probability inside the limit (for $t_{n+1} = \infty$) as

$$\mathrm{e}^{-\Lambda(t_1)} \prod_{i=1}^{n} [\lambda(t_i)h_i + o(h_i)]\, \mathrm{e}^{-[\Lambda(t_{i+1})-\Lambda(t_i+h_i)]}$$

and after dividing by $h_1 \cdots h_n$ and evaluating the limit we get the declared result. □

1.2 Marked Poisson Process

1.2.1 Notation

In this subsection we introduce several random variables, which together form a marked Poisson process. The notation and theory is mostly based on Norberg (1993), Norberg (1999) and Merz and Wüthrich (2008, pp. 369–377). We consider a homogeneous part of a line of business of an insurance company with a risk exposure per time described by a nonnegative function $w(t)$, $t \geq 0$. The exposure rate $w(t)$ may be thought of as a simple measure of volume of the business, e.g. number of policies in force or earned exposure.

We assume that claims arise at times $T_1, T_2, \ldots$ (accident dates) and no claims occur simultaneously. Claims can be sorted in ascending order with respect to the accident dates, so that the ith claim occurs at time T_i. Number of claims incurred prior to time $t \geq 0$ can be written as

$$N_t = \sum_{i \geq 1} \mathbb{I}(T_i \leq t),$$

where $\mathbb{I}()$ is an indicator function equal to one, when its argument holds true and zero otherwise, i.e.

$$\mathbb{I}(T_i \leq t) = \begin{cases} 1, & \text{if } T_i \leq t, \\ 0, & \text{otherwise} \end{cases}$$

and total number of claims is

$$N = \lim_{t \to \infty} N_t.$$

Note that accident times $\{T_i, i \in \mathbb{N}\}$ determine process $\{N_t, t \geq 0\}$ and vice versa, because time of the ith event is simply

$$T_i = \inf\{t \geq 0 \colon N_t \geq i\}.$$

We assume that $\{N_t, t \geq 0\}$ is a nonhomogeneous Poisson process with a nonnegative intensity function $w(t)\lambda(t)$, $t \geq 0$. To justify the last assumption, we could imagine that occurences of claims on each insurance contract follow a nonhomogeneous Poisson process with a nonnegative intensity function $\tilde{\lambda}(t)$, $t \geq 0$, where $\tilde{\lambda}(t)$ is equal to $\lambda(t)$, when t belongs to a time interval when a selected contract is in force, and zero otherwise. This means that all insurance contracts are risk homogeneous (with respect to occurences of claims) and claims can occur only during the lifetime of the contract. A usual assumption would be independence of the occurence processes, therefore we could use the third part of Theorem 1 and the assumption is justified. Moreover, the first part of Theorem 1 states that the total number of claims follows Poisson distribution with parameter $\Lambda(\infty)$, where

$$\Lambda(t) = \int_0^t w(s)\lambda(s)\,\mathrm{d}s, \quad t \geq 0, \tag{1.2}$$

so the total number of claims is finite with probability one.

Of course, the company receives a notification about a claim (incurred at a nonnegative time T) at a nonnegative time S with the nonnegative delay $U = S - T$. Then the claim is closed at a time $S + V$, where V is a waiting time until a claim settlement after the notification. There are some individual payments within interval $[S, S + V]$ described by a process

$$C = \{C(v), 0 \leq v \leq V\}.$$

Finally, $Z = (U, C)$ is a mark describing the settlement process. A claim can be now described as pair (T, Z). Note that marks can contain even more information about claims, e.g. case estimate. If we denote $\mathcal{Z}$ as a space of possible claims developments of Z, then claim (T, Z) is a random element in set $\mathcal{C} = [0, \infty) \times \mathcal{Z}$.

1.2.2 Payment Process

In this subsection we are interested in a probability distribution of the payment process of a selected claim. To describe the distribution, we will assume that increments of events are independent, hence we will work here with a slightly modified Poisson process. We will also assume that the payment process is independent of time of occurence and delay, or in other words, the claims handling does not change over time. These assumptions are here due to simplifications, so that we are able to derive some properties more formally; nevertheless, they will be applied consistently in the thesis.

The modification lies in considering more types of events, e.g. a payment and a settlement with a payment. We will be referencing to these types of events as events of type 1 and 2. Times of events since their notification are random variables $V_1, V_2, \dots$ and these random variables with types of events determine number of payments $\tilde{N}$ within a selected claim, because an occurence of event 2 means that the claim is closed. It would be much better to work with a third type of event: settlement without a payment. However, this type of event is not taken into account, which will be explained in the next chapter.

We assume that each type of event occurs in accordance with a nonhomogeneous Poisson process with intensity function $\mu_j(t)$ for $t \geq 0$ and $j = 1, 2$. We define cumulative hazard functions

$$M_j(t) = \int_0^t \mu_j(s)\,\mathrm{d}s, \quad t \in [0, +\infty], \quad j = 1, 2$$

and overall cumulative hazard function $M(t) = M_1(t) + M_2(t)$. Moreover, we assume that these processes are independent. Probability that no event occurs during time interval $[a, b)$ is due to the independence equal to product of probabilities

$$\prod_{j=1}^{2} \mathsf{P}\left[\text{No event of type } j \text{ occurs during } [a,b)\right] = \prod_{j=1}^{2} \mathrm{e}^{-[M_j(b) - M_j(a)]}, \tag{1.3}$$

which can be rewritten as

$$\mathrm{e}^{-[M(b)-M(a)]}. \tag{1.4}$$

Now we are able to work with time occurences and types of events. We define number of payments as

$$\tilde{N} = \sum_{k=1}^{\infty} \mathbb{I}(V_k < \infty),$$

i.e. $\tilde{N}$ is number of events with a finite time. Using the same principle as in the proof of Theorem 1, we get (for $n = 1, 2, \dots$ and nonnegative values $v_1, \dots, v_n, h_1, \dots, h_n$)

$$\begin{aligned}
&\mathsf{P}\left[\tilde{N} = n, V_k \in [v_k + h_k), \text{for all } k = 1, \dots, n\right] = \\
&\quad = \mathrm{e}^{-M(v_1)} \left\{ \prod_{k=1}^{n-1} \left[\mu_1(v_k) h_k + o(h_k)\right] \mathrm{e}^{-[M(v_{k+1}) - M(v_k + h_k)]} \right\} \left[\mu_2(v_n) h_n + o(h_n)\right]
\end{aligned}$$

and after dividing by product $h_1 \cdots h_n$ and evaluating a limit of the last expression, where $(h_1, \dots, h_n)$ tends to $(0_+, \dots, 0_+)$, we end up with density

$$f_{\tilde{N}, V_1, \dots, V_n}(n, v_1, \dots, v_n) = \mathrm{e}^{-M(v_n)} \left[\prod_{k=1}^{n-1} \mu_1(v_k)\right] \mu_2(v_n). \tag{1.5}$$

The last expression describes density of some (not specified) n payments at times $v_1, \dots, v_n$ and settlement at time v_n. We assume that the times of events are positive numbers in ascending order, otherwise the density would be (formally) zero.

A small detour from our topic, which might clarify some ambiguities: if we wanted to proove that we really work with a density in equation 1.5 (not taking into account our derivation, which proves it as well), we have to verify that

$$\sum_{n=1}^{\infty}\int_0^{\infty}\int_0^{v_n}\cdots\int_0^{v_2} f_{\tilde{N},V_1,\ldots,V_n}(n,v_1,\ldots,v_n)\mathrm{d}v_1\ldots\mathrm{d}v_{n-1}\mathrm{d}v_n = 1$$

holds, i.e. we have to integrate and sum the density from equation 1.5 over all possible values. The equation above nicely illustrates that $f_{\tilde{N},V_1,\ldots,V_n}$ is a joint density and not a conditional density. We will not discuss the verification in detail, we will just state necessary steps. Firstly, we need to calculate $n-1$ inner integrals (using integration by parts) and obtain as a result

$$\frac{M_1^{n-1}}{(n-1)!}.$$

Secondly, we interchange the sum and the remaining integral and this leads us to

$$\int_0^{\infty}\mathrm{e}^{-M_2(v_n)}\mu_2(v_n)\mathrm{d}v = 1-\mathrm{e}^{-M_2(\infty)} = 1,$$

because

$$M_2(t) = \int_0^t \mu_2(s)\mathrm{d}s = \int_0^t \frac{f_2(s)}{1-F_2(s)}\mathrm{d}s = \log\left(\frac{1}{1-F_2(t)}\right)$$

and this quantity tends to infinity as $t\to\infty$.

Now we return back to the topic at hand, because we are finally getting to the desired density of the payment process. To simplify the situation, we treat payments as independent and identically distributed (iid) random variables $P_1, P_2, \ldots$ with a common density function f_P and independent of the rest of the payment process. To derive the density of the payment process, we will use the well-known formula between joint and conditioned density for two random variables

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y).$$

In our case, X would be paid amounts and Y times and types of events, therefore the density of the payment process can be written as

$$f_{\mathbf{C}}(\mathbf{c}) = \mathrm{e}^{-M(v_n)}\left[\prod_{k=1}^{n-1}\mu_1(v_k)\right]\mu_2(v_n)\left[\prod_{j=1}^{n}f_P(p_j)\right],$$

where $\mathbf{C} = (\tilde{N}, V_1, \ldots, V_{\tilde{N}}, P_1, \ldots, P_{\tilde{N}})$ and $\mathbf{c} = (n, v_1, \ldots, v_n, p_1, \ldots, p_n)$. For completeness, we state possible values of all these variables: $n = 1, 2, \ldots$ and $v_1, \ldots, v_n, p_1, \ldots, p_n$ are nonnegative values. Furthermore, $v_1, \ldots, v_n$ are assumed to be in increasing order, since they correspond to times of events. Note that the density above is again a joint density, in this case for number of events, times of events and payments (it is not a conditional density as it might seem due to the complexity of notation).

When we compare information contained in $\mathbf{C}$ here and in the process C from the previous subsection, we conclude they contain almost equivalent information. There might be a difference in time of settlement, because in the above described approach claims can be closed only with a payment, but in reality it can happen later without any payment. This simplification is a consequence of taking into account only events of type 1 and 2, therefore from now on we will consider the payment process $\mathbf{C}$.

1.2.3 Distribution of Claims Process

For a moment, we return to equation 1.1, where we introduced a density of number of claims and times of occurence. We substitute function $\lambda(t)w(t)$ for function $\lambda(t)$ and therefore we simply write the density in form

$$\frac{[\Lambda(\infty)]^n}{n!} \mathrm{e}^{-\Lambda(\infty)}\, n! \prod_{i=1}^{n} \frac{w(t_i)\lambda(t_i)}{\Lambda(\infty)}, \tag{1.6}$$

where $\Lambda(t)$ is correspondingly defined in 1.2. In this case, $n = 0, 1, \ldots$ and $t_1, \ldots, t_n$ are nonnegative values in increasing order. In a special (and unrealistic) case $n = 0$ we follow a convention that an empty product is equal to one. This form of density provides us a useful interpretation for simulations: number of claims N follows Poisson distribution with finite mean $\Lambda(\infty)$ and when number of claims $N = n$ is given, times of occurence form an ordered sample with a common density for times of occurence

$$f_T(t) = \frac{w(t)\lambda(t)}{\Lambda(\infty)}, \quad t \geq 0. \tag{1.7}$$

For the interpretation we used the well-known formula for density of order statistics and it means that we derived a conditional density of times of occurence when number of claims $N = n$ is given.

Considering only one claim, we can write its density function as

$$f_{T,Z}(t, z) = f_{Z|T}(z|t) f_T(t),$$

where $f_{Z|T}(z|t)$ is a conditional density of delay U and payment process $\mathbf{C}$, which we discussed in the previous subsection, when time of occurence $T = t$ is given. To derive a density of all claims together, we need to append $f_{Z_i|T_i}(z_i|t_i)$ for each claim to the product in 1.6. Before we do it, we summarize necessary assumptions in a definition. Notice that we will use a simpler notation $f_{Z|t}(z)$ instead of the previous notation $f_{Z|T}(z|t)$.

Definition 2. *Marked Poisson process with a nonnegative intensity function $w(t)\lambda(t)$, $t \geq 0$, and position-dependent marks is process denoted by*

$$\{(T_i, Z_i), i = 1, \ldots, N\} \sim \mathrm{Po}\left(w(t)\lambda(t), f_{Z\,|\,t}\right),$$

where accident dates $\{T_i, i \in \mathbb{N}\}$ are determined by nonhomogeneous Poisson process $\{N_t, t \geq 0\}$ with the intensity function $w(t)\lambda(t)$, and marks $Z_i = Z_{T_i} \in \mathcal{Z}$ satisfy the following assumptions:

1. *$\{Z_t, t \geq 0\}$ is a family of random elements in $\mathcal{Z}$ that are mutually independent;*

2. *$\{Z_t, t \geq 0\}$ is independent of process $\{N_t, t \geq 0\}$;*

3. *Marks conditioned by accident date $Z \,|\, T = t$ have density $f_{Z\,|\,t}(z), z \in \mathcal{Z}$.*

Based on Definition 2 we can derive a joint density of all claims together (i.e. number of claims and their developments) from equation 1.6 as

$$f_{N,\mathbf{T},\mathbf{Z}}(n, \mathbf{t}, \mathbf{z}) = \frac{[\Lambda(\infty)]^n}{n!} \mathrm{e}^{-\Lambda(\infty)}\, n! \prod_{i=1}^{n} \frac{w(t_i)\lambda(t_i)}{\Lambda(\infty)} f_{Z|t_i}(z_i),$$

where $\mathbf{T} = (T_1, \dots, T_N)$, $\mathbf{Z} = (Z_1, \dots, Z_N)$ and $\mathbf{t}, \mathbf{z}$ are simply lowercase versions of length n (marks $z_1, \dots, z_n$ are elements of $\mathcal{Z}$ and the rest attains the same values as in equation 1.6). Note that in this notation number of claims N is also indirectly contained in $\mathbf{T}$ as its length, because T_{N+1} is considered equal to infinity. Another possibility of writing the last equation is

$$f_{N,\mathbf{T},\mathbf{Z}}(n, \mathbf{t}, \mathbf{z}) = \mathsf{P}\,[N = n]\, n! \prod_{i=1}^{n} f_T(t_i) f_{Z|t_i}(z_i),$$

which has again a useful interpretation: we can generate number of claims N, then generate times of occurences with marks and sort such claims in ascending order with respect to the generated times of occurence.

1.2.4 Division of Claims

Although we derived the distribution of all claims, we cannot work directly with the density $f_{N,\mathbf{T},\mathbf{Z}}(n, \mathbf{t}, \mathbf{z})$, because we can observe only reported claims. A very natural division of claims brings four groups: settled (S), reported but not settled (RBNS), incurred but not reported (IBNR) and not incurred claims. The last group of claims refers to future periods and is covered by unearned premium reserve (UPR). This reserve is usually calculated on a pro rata temporis basis (per policy).

Our main interest lies in IBNR and reported claims. If τ is the present moment, then we can define these groups more formally. Marks of settled claims incurred at time t is a set

$$\mathcal{Z}_t^{\mathrm{S}} = \{z \in \mathcal{Z} : t + u + v \leq \tau\},$$

marks of RBNS claims incurred at time t is a set

$$\mathcal{Z}_t^{\mathrm{RBNS}} = \{z \in \mathcal{Z} : t + u \leq \tau < t + u + v\},$$

marks of IBNR claims incurred at time t is a set

$$\mathcal{Z}_t^{\mathrm{IBNR}} = \{z \in \mathcal{Z} : t \leq \tau < t + u\},$$

and finally marks of not incurred claims incurred at future time t is a set

$$\mathcal{Z}_t^{\mathrm{UPR}} = \{z \in \mathcal{Z} : t > \tau\}.$$

We introduce marks of reported claims (incurred at time t) as a set

$$\mathcal{Z}_t^{\mathrm{R}} = \mathcal{Z}_t^{\mathrm{S}} \cup \mathcal{Z}_t^{\mathrm{RBNS}} = \{z \in \mathcal{Z} : t + u \leq \tau\}.$$

In the following theorem, we can imagine an arbitrary finite division, but we will consider only one division $G = \{\mathrm{R}, \mathrm{IBNR}, \mathrm{UPR}\}$ referring to the sets defined above. For each $g \in G$ we consider a component process

$$\{(T_i^g, Z_i^g), i = 1, \ldots, N^g\}, \tag{1.8}$$

where the random variables above are constructed in a straightforward way: process counting g-claims is

$$N_t^g = \sum_{i \geq 1} \mathbb{I}(T_i \leq t, Z_i \in \mathcal{Z}^g),$$

times of occurence are

$$T_i^g = \inf(t \geq 0 : N_t^g \geq i)$$

and marks are $Z_i^g = Z_{T_i^g}$.

Theorem 2. *Let G be a finite division of claims, such that for each $t \geq 0$ holds*

$$\sum_{g \in G} \mathsf{P}\left[Z \in \mathcal{Z}_t^g | T = t\right] = 1. \tag{1.9}$$

Then component processes in 1.8 are independent and for each $g \in G$ holds

$$\{(T_i^g, Z_i^g), i = 1, \ldots, N^g\} \sim \mathrm{Po}\left(\lambda^g(t), f_{Z\,|\,t}^g\right),$$

with

$$\lambda^g(t) = w(t)\lambda(t)\,\mathsf{P}\left[Z \in \mathcal{Z}_t^g | T = t\right]$$

and

$$f_{Z\,|\,t}^g(z) = \frac{f_{Z\,|\,t}(z)}{\mathsf{P}\left[Z \in \mathcal{Z}_t^g | T = t\right]} \mathbb{I}(z \in \mathcal{Z}_t^g).$$

Proof. Because of the assumption 1.9, the density $f_{N,\mathbf{T},\mathbf{Z}}(n, \mathbf{t}, \mathbf{z})$ can be rewritten in a form

$$\prod_{g \in G} \left\{ \frac{[\Lambda^g(\infty)]^{n^g}}{n^g!} \mathrm{e}^{-\Lambda^g(\infty)}\, n^g! \prod_{i=1}^{n^g} \frac{\lambda^g(t_i^g)}{\Lambda^g(\infty)} f_{Z\,|\,t_i^g}^g(z_i^g) \right\}, \tag{1.10}$$

where

$$\Lambda^g(t) = \int_0^t \lambda^g(s)\,\mathrm{d}s,$$

it is mostly about reindexing all quantities with respect to their categories. From equation 1.10 follows the independence and the statement that the component processes follow marked Poisson processes. □

Theorem 2 states that g-claims occur with the original intensity multiplied by the probability that a claim incurred at a time t is in a category g. Similarly, the distribution of the marks is determined by the conditional distribution of the marks, given that it is in a category g. This result is a usual property of Poisson process.

To determine distributions of reported and IBNR claims we apply Theorem 2. For reported claims, we need to calculate probability that a claim incurred at a time $t \leq \tau$ has been already reported. Such a probability is simply

$$\mathsf{P}\left[T + U \leq \tau \,|\, T = t\right] = \mathsf{P}\left[U \leq \tau - t \,|\, T = t\right] = F_{U|t}(\tau - t),$$

where $F_{U|t}$ is a conditional cumulative density function of delay U, when time of occurence $T = t$ is given. This means that reported claims occurences have an intensity function

$$\lambda^{\mathrm{R}}(t) = w(t)\lambda(t)F_{U|t}(\tau - t)$$

and the marks have a density

$$f^{\mathrm{R}}_{Z\,|\,t}(z) = \frac{f_{Z|t}(z)}{F_{U|t}(\tau - t)}\mathbb{I}(u \leq \tau - t).$$

Using the complementary probability, IBNR claims occurences have an intensity

$$\lambda^{\mathrm{IBNR}}(t) = w(t)\lambda(t)\left[1 - F_{U|t}(\tau - t)\right]\mathbb{I}(t \leq \tau)$$

and the marks have a density

$$f^{\mathrm{IBNR}}_{Z\,|\,t}(z) = \frac{f_{Z|t}(z)}{1 - F_{U|t}(\tau - t)}\mathbb{I}(t \leq \tau < t + u).$$

For completeness, not incurred claims occurences have an intensity

$$\lambda^{\mathrm{UPR}}(t) = w(t)\lambda(t)\mathbb{I}(t > \tau)$$

and the marks have a density

$$f^{\mathrm{UPR}}_{Z\,|\,t}(z) = f_{Z\,|\,t}(z)\mathbb{I}(t > \tau).$$

Note that sum of the last three mentioned intensities is, indeed, equal to the original intensity $w(t)\lambda(t)$.

To sum up, the density of observed claims is simply the same formula as in 1.10, only without the first product and there is R instead of g. Specifically, the density is

$$f^{R}_{N,\mathbf{T},\mathbf{Z}}(n, \mathbf{t}, \mathbf{z}) = \frac{[\Lambda^{\mathrm{R}}(\infty)]^n}{n!}\,\mathrm{e}^{-\Lambda^{\mathrm{R}}(\infty)}\,n!\prod_{i=1}^{n}\frac{\lambda^{\mathrm{R}}(t_i)}{\Lambda^{\mathrm{R}}(\infty)}f^{\mathrm{R}}_{Z\,|\,t_i}(z_i),$$

where the upper index R is omitted for the number of claims, the times of occurence and their marks, since from the context their category is obvious. The last equation still includes few redundant terms and it can be simplified back to

$$\mathrm{e}^{-\Lambda^{\mathrm{R}}(\infty)}\prod_{i=1}^{n}\lambda^{\mathrm{R}}(t_i)f^{\mathrm{R}}_{Z\,|\,t_i}(z_i), \tag{1.11}$$

where the last density in the product can be further rewritten as

$$f^{\mathrm{R}}_{Z\,|\,t_i}(z_i) = f_{U\,|\,t_i}(u_i)f_{\mathbf{C}\,|\,t_i,u_i}(\mathbf{c_i}) = f_{U\,|\,t_i}(u_i)f_{\mathbf{C}}(\mathbf{c_i}),$$

because we assume that the payment process $\mathbf{C}$ depends neither on time of occurence nor delay.

2. Practical Part

This chapter describes and summarizes all estimated parameters and detailed discussion about appropriateness of accepted decisions is included as well to point out advantages and weaknesses. We start with delay distribution, then we deal with occurence process of claims, times between payments and finally payments distribution. While the previous part was theoretical and quite general, this part deals with specific problems arising from data characteristics and some necessary adjustments are formulated.

First of all we briefly describe data which we use in following sections. The data contain some information about motor third party liability (MTPL) line of business prepared by Czech Insurers' Bureau for educational purposes and are collected from all member insurance companies. Two important parts of MTPL are material damage (MD) and bodily injuries (BI), whose reserves are often calculated separately due to their apparently different nature, as it can be seen in the next subsections. We omit annuities from the analysis, because they have very different nature compared to material damage and bodily injuries and they contain less observations.

Since the data are available only for noncommercial purposes, we do not attach them to the thesis. However, in case of interest, they can be obtained for example via Czech Insurer's Bureau (contact RNDr. Petr Jedlička, Ph.D.) or via Charles university, faculty of mathematics and physics (contact RNDr. Michal Pešta, Ph.D.). For such a case, we attach our source code as an appendix, which can be copied from electronic version of the thesis and everything can be replicated.

The data consist of two files: claims settlements (payments) and RBNS reserve development. Each row of claims settlements contains ID, type of insurance liability, date of insurance accident, date of notification, date of payment and paid amount. RBNS reserve development differs in the last two columns, because there are contained date of change and change of RBNS reserve. All dates are in range from January 1, 2000 to December 31, 2015. The data have relatively good quality, only few rows are additionaly excluded, because of inconsistencies like notification before occurence or date of payment before notification or occurence. In the end we have 47 111 rows for material damage payments, 12 364 rows for bodily injuries payments, 155 500 rows for material damage RBNS reserve changes and 36 978 rows for bodily injuries RBNS reserve changes.

2.1 Delay Distribution

Prior to any analysis, our expectation regarding delay distribution would be a presence of a decreasing trend, i.e. more recent claims are usually reported after a shorter period of time, because of a continuous development of technologies and easier reachability of insurers. After a simple inspection of our data we can immediately conclude that we should not leave out the effect of occurence time on delay. This is indicated by a simple linear regression or a comparison of mean delays in available years. This leads us to a two-stage estimation: we estimate trend in the first stage and then we estimate conditional distribution of delay, which is described below.

We choose the trend for simplicity as a smooth two-parametric function, which would sufficiently capture the decreasing trend of expected value. One possible choice is for example an exponential trend

$$\mathsf{E}\,[U \,|\, T = t] = a\, b^t$$

and another choice might be a logarithmic trend

$$\mathsf{E}\,[U \,|\, T = t] = a + b \log(t),$$

we will be choosing from these two parametrizations. Parameters a, b are unknown and we estimate them in `R` using function `nls()` for estimation by its default nonlinear least squares method. It is very important to note that the estimation is performed only on years prior to 2015, because mainly in the last year are not included all observations yet, especially the larger ones. Time $t = 0$ is set as December 31, 1999 and time is measured in days. Trend for material damage is estimated on 35 856 claims and for bodily injuries on 5 259 claims.

A simple comparison of residual sum of squares indicates that the exponential trend is more suitable in both cases. We can also calculate R squared to back up our choice. We replicate calculation of R squared from linear regression based on total sum of squares and residual sum of squares, specifically

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (u_i - f_i)^2}{\sum_i (u_i - \bar{u})^2},$$

where u_i is ith observation, $\bar{u}$ is arithmetic mean and f_i is smoothed value of u_i (either by the exponential or the logarithmic trend). In case of bodily injuries R squared is indeed larger for the exponential trend (3.26 % against 2.67 % for the logarithmic trend), but for material damage we get lower R squared for the exponential trend (2.94 % against 2.98 % for the logarithmic trend). Nevertheless, the difference of R squared for the exponential trend is negligible and it also lacks the natural interpretation from linear regression and furthermore, it is related to quality of predictions, therefore we choose in this case the exponential trend as well.

Estimated parameters for the chosen exponential trend can be found in Table 2.1. The first parameter a has a meaning of an initial expected value of delay at time $t = 0$ and the second parameter b explains by how much the expected value decreases by one day. For comparison, annual decrease for material damage is approximately 7.5 % and for bodily injuries almost 4 %. A graphical representation of the mentioned trends can be found in Figure 2.1. We can see that the linear and exponential trends are quite similar and significantly better than the logarithmic trend, which is much steeper in the first year and such a decrease is not very reasonable. This is of course another reason for our choice of the exponential trend.

The second stage starts with a transformation of observed delays. We will use a shorter notation U_t instead of $U \,|\, T = t$ to better explain this stage. We assume that random variable U_t is related to the initial delay U_0 via a transformation

$$U_t = \frac{\mathsf{E}\,[U \,|\, T = t]}{\mathsf{E}\,[U \,|\, T = 0]} U_0 = b^t U_0, \tag{2.1}$$

	Material damage	SD	Bodily injuries	SD
a	130.223906	2.065000	291.004888	9.256000
b	0.999784	0.000007	0.999847	0.000012

Table 2.1: Estimated parameters and their standard deviations (SD) of exponential trend of delay

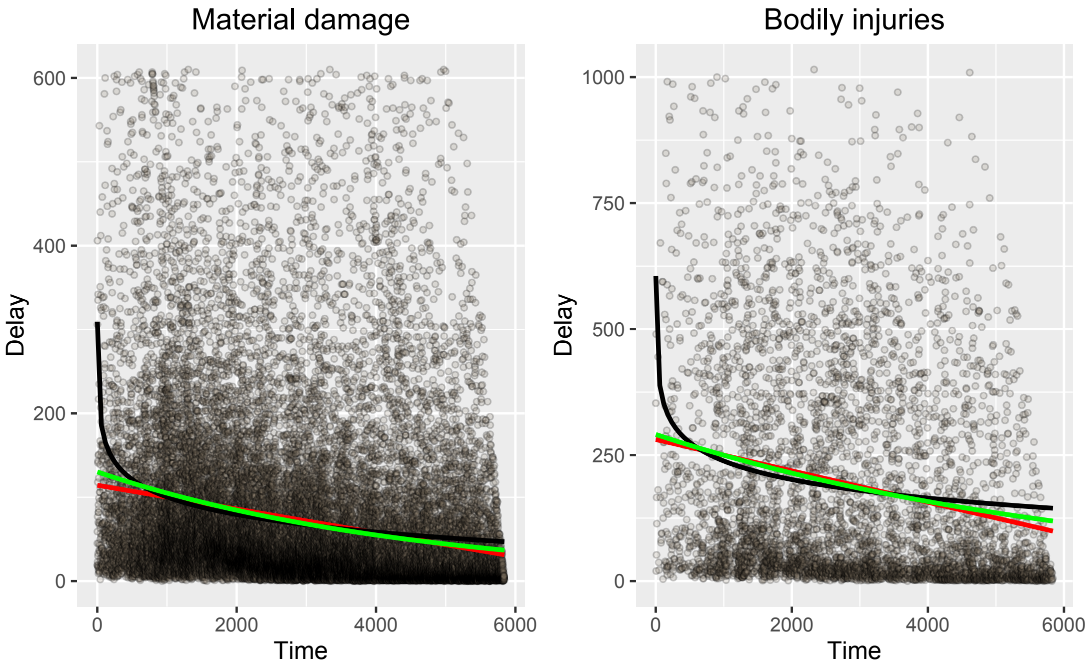

Figure 2.1: Observed delays of notification (dots) with linear trend (red line), exponential trend (green curve) and logarithmic trend (black curve)

which means that U_t has cumulative density function

$$F_{U_t}(u) = \mathsf{P}[U \leq u \,|\, T = t] = \mathsf{P}[b^t U \leq u \,|\, T = 0] = F_{U_0}(b^{-t}u)$$

and density function is obviously

$$f_{U_t}(u) = b^{-t} f_{U_0}(b^{-t}u). \tag{2.2}$$

Because the parameter b is now considered to be known, estimation of delay distribution is reduced to more simple estimation of parameters of f_{U_0}, where observed values u_t are transformed through multiplication by b^{-t}. Such transformation has an advantage that the transformed data form a continuous random sample, opposed to the starting random sample, which has rather a discrete nature. Few

		1st Qu.	Median	Mean	3rd Qu.	SD
Material damage	Observed	13.0	32.0	71.9	77.0	119.3
	Transformed	25.7	56.3	130.0	134.1	223.9
Bodily injuries	Observed	23.0	78.0	187.7	293.0	234.9
	Transformed	36.4	116.5	281.6	438.8	352.5

Table 2.2: Comparison of observed and transformed delay (in days)

	Distribution	Log likelihood	Test statistic	DF	Critical value
MD	Lognormal	-214 295.40	449.58	47	64.00
	Weibull	-217 125.60	4 431.31	47	64.00
	Gamma	-218 310.80	48 502.50	47	64.00
	Burr	-241 525.90	NA	NA	NA
BI	Lognormal	-35 769.19	799.52	47	64.00
	Weibull	-35 750.43	709.47	47	64.00
	Gamma	-35 802.88	812.20	47	64.00
	Burr	-39 626.91	NA	NA	NA

Table 2.3: Logarithmic likelihoods and results of (chi-squared) goodness-of-fit tests for delay distributions

	Parameter	Estimate	Standard deviation
Material damage (lognormal distribution)	μ	4.05578	0.00666
	σ	1.28842	0.00471
Bodily injuries (Weibull distribution)	shape a	0.75504	0.00799
	scale b	210.71940	4.49735

Table 2.4: Estimated parameters of delay distributions

descriptive statistics and a comparison between observed and transformed values of delays are summarized in Table 2.2. Transformed delays are strictly greater than observed values and they have larger standard deviation (SD).

To choose properly f_{U_0} we estimate and compare Burr (this distribution represents a heavy-tailed distribution), gamma, Weibull and lognormal distribution (light-tailed distributions). Estimation is carried out in `R` using package MASS and function `fitdistr()`, which provides maximum likelihood estimates. Note that we excluded zero delays from data, because the chosen distributions can be estimated on positive values only. Number of such observations is relatively small (approximately 0.2 % for material damage and much less for bodily injuries), therefore we consider their influence negligible. Number of observations for estimation is 37 411 for material damage and 5 443 for bodily injuries. Based on the largest likelihood (see Table 2.3) we choose lognormal distribution for material damage and Weibull distribution for bodily injuries as the most suitable distribution for the transformed delays, estimated values of parameters can be found in Table 2.4.

Table 2.3 also shows results of performed goodness-of-fit tests, which we will briefly discuss here. We calculated 50 categories (so that each category contains a reasonable number of observations) for each distribution based on estimated quantiles, which provides us a division for a chi-squared test. Then we calculate the classical chi-squared test statistic

$$\chi^2 = \sum_{i=1}^{50} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is number of observations in ith category and E_i is expected number for ith category, in this case one fiftieth (1/50) of the total number of observations. We observe that both selected distributions also minimize chi-squared test

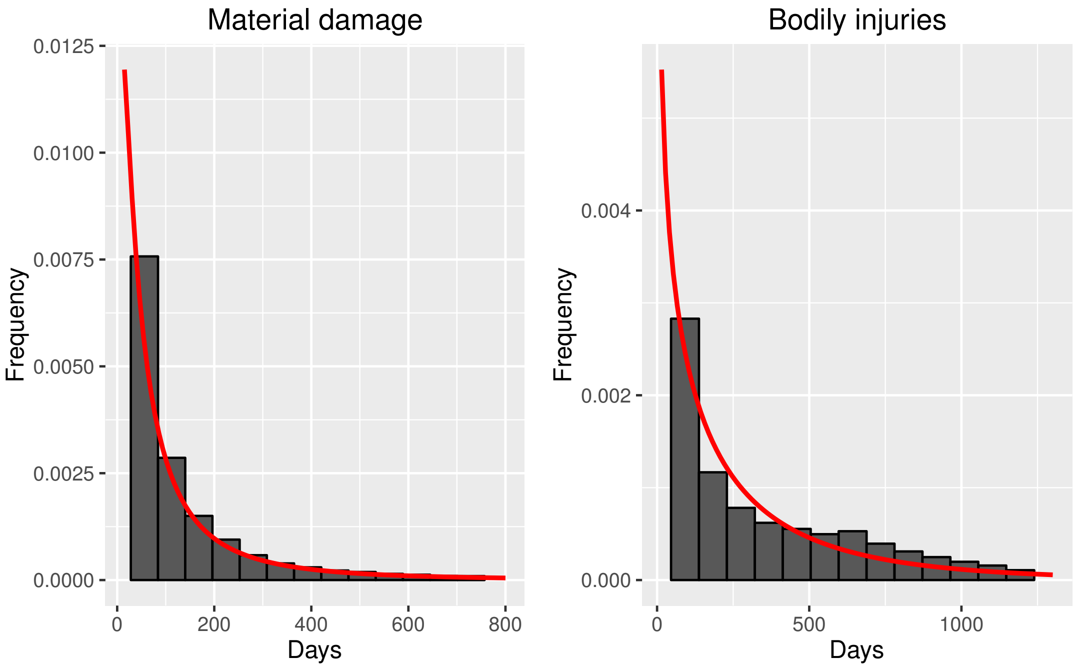

Figure 2.2: Comparison of histograms of transformed delays and estimated density functions

statistic, but all test statistics are too large. However, even a small violation of the null hypothesis is significant with large number of observations, therefore in our case we could have expected prior to any analysis that all p-values will be numerically zeroes. This could be overcome by testing the null hypothesis on a subsample, but this is very dependent on more randomness (the divisions are already based on estimates). From this reason, goodness-of-fit tests do not seem to be appropriate here as a decision rule, because all distributions are formally rejected and that is why we would rather recommend a graphical analysis. Note that goodness-of-fit tests were not calculated for Burr distribution, because this distribution is inappropriate for its small logarithmic likelihood and a graphical comparison of histograms and estimated densities supports this as well.

For completeness, density of lognormal distribution is

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma u} \exp\left\{-\frac{(\log u - \mu)^2}{2\sigma^2}\right\}, \quad u > 0,$$

density of Weibull distribution is

$$f(u) = \frac{\alpha}{\beta}\left(\frac{u}{\beta}\right)^{\alpha-1} \exp\left\{-\left(\frac{u}{\beta}\right)^{\alpha}\right\}, \quad u > 0,$$

density of gamma distribution is

$$f(u) = \frac{a^p}{\Gamma(p)} u^{p-1} \exp\left\{-au\right\}, \quad u > 0$$

and finally density of used Burr distribution is

$$f(u) = ck\frac{u^{c-1}}{(1+u^c)^{k+1}}, \quad u > 0$$

and all parameters are positive values. A graphical comparison of estimated densities with histograms can be found in Figure 2.2. We can see that the log-normal distribution for material damage is very similar to the histogram, while the Weibull distribution for bodily injuries is slightly different. It can be seen that the histogram for bodily injuries is almost flat from 300 to 700 days and a similar flatness can be observed in the original data before transformation as well. This suggests that more complex model might be needed for bodily injuries delays, but for simplicity we will consider Weibull distribution sufficient.

2.2 Occurence Process

In this section we estimate intensity function of the underlying nonhomogeneous Poisson process for occurence of claims. The basic idea is inspired by Antonio and Plat (2014), which is discussed here in more detail. The estimated delay distributions are considered known and fixed and for intensity function $\lambda(t)$ we use a piecewise constant specification. More specifically, we choose division

$$0 = d_0 < d_1 < \ldots < d_m = \tau,$$

where τ is a time difference (in days) between December 31, 2015 and December 31, 1999 and m is a positive integer number. We assume that the exposure function $w(t)$ is also a piecewise constant function and it has the same division as $\lambda(t)$. Note that such assumption is not very restrictive in case of earned exposure. For $t \in (d_{j-1}, d_j]$ the intensity function $\lambda(t)$ is equal to λ_j and the exposure function $w(t)$ is equal to w_j for all $j = 1, \ldots, m$.

To derive an estimate of $\lambda(t)$, we recall equation 1.11 and realize that its likelihood is

$$\mathrm{e}^{-\Lambda^{\mathrm{R}}(\infty)} \prod_{i=1}^{n} \lambda^{\mathrm{R}}(t_i) = \mathrm{e}^{-\Lambda^{\mathrm{R}}(\infty)} \prod_{i=1}^{n} \lambda(t_i) w(t_i) F_{U|t_i}(\tau - t_i),$$

where n is number of observations and since the exposure function and the delay distribution are considered known, they can be excluded from the product above. Observed number of reported claims in the jth interval is

$$N(j) = \sum_{i=1}^{n} \mathbb{I}\big(t_i \in (d_{j-1}, d_j]\big)$$

for $j = 1, \ldots, m$ and with this notation we rewrite the likelihood to

$$\prod_{j=1}^{m} \lambda_j^{N(j)} \exp\left\{ -\lambda_j w_j \int_{d_{j-1}}^{d_j} F_{U|t}(\tau - t)\,\mathrm{d}t \right\}.$$

It is straightforward that the logarithmic likelihood is

$$\sum_{j=1}^{m} N(j) \log(\lambda_j) - \sum_{j=1}^{m} \lambda_j w_j \int_{d_{j-1}}^{d_j} F_{U|t}(\tau - t)\,\mathrm{d}t \tag{2.3}$$

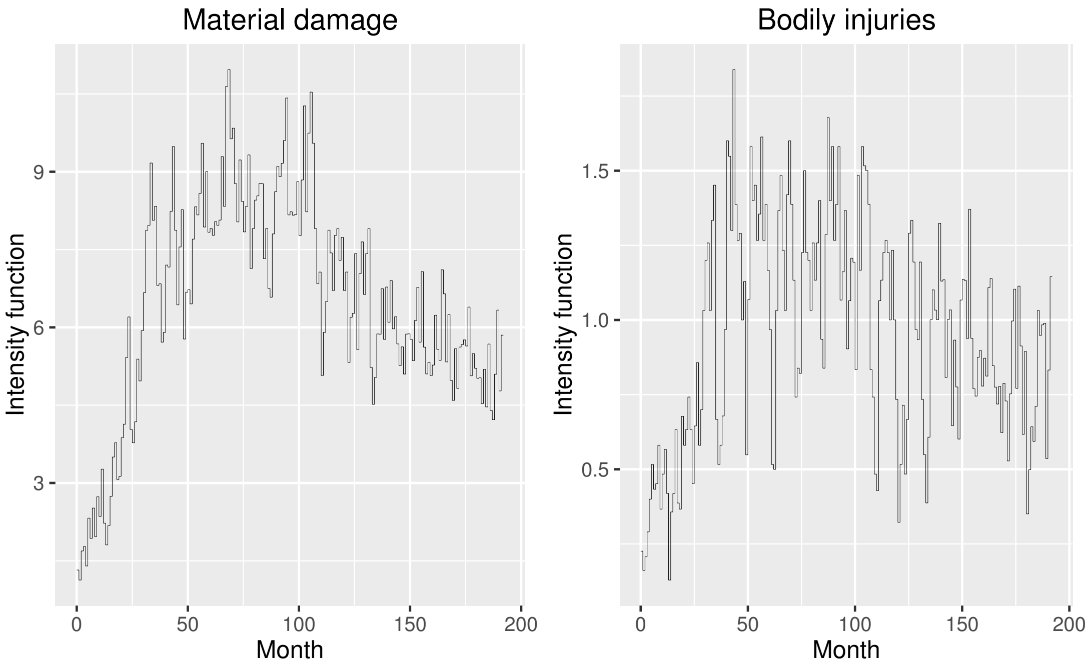

Figure 2.3: Estimated intensity of occurence processes for material damage and bodily injuries

and by setting the first derivatives to zero we get a formula for maximum likelihood estimates

$$\hat{\lambda}_j = \frac{N(j)}{w_j \int_{d_{j-1}}^{d_j} F_{U|t}(\tau - t)\,\mathrm{d}t}$$

for all $j = 1, \ldots, m$.

Note that in this set-up we actually do not need the exposure function. It is because we can rewrite the estimate (using also transformation in equation 2.1) to

$$\hat{\lambda}_j w_j = \frac{N(j)}{\int_{d_{j-1}}^{d_j} F_{U_0}(\frac{\tau - t}{b^t})\,\mathrm{d}t} \tag{2.4}$$

and because only product $\hat{\lambda}_j w_j$ is needed in all calculations, the exposure rate can be omitted and we do not need to evaluate it. We can also consider the overall intensity function as $\lambda(t)$ without $w(t)$ in the first place and with the piecewise constant specification the same estimate, as on the right-hand side of equation 2.4 is derived as maximum likelihood estimator. In any case, the right-hand side of equation 2.4 can be interpreted as intensity function of the whole selected part of line of business. However, with a parametric form of $\lambda(t)$ the exposure function would matter and it would have an influence on estimation of the intensity function.

To obtain the estimate of $\lambda(t)$ we must choose a width of the division and as it could be expected, it will somehow affect the results. It is not quite clear at first whether months, quarters or years should be chosen and how large will be the influence of the choice. However, we realize that with maximum likelihood estimates we can again compare logarithmic likelihoods in equation 2.3 and choose our estimate accordingly. Based on the comparison we choose month intervals in

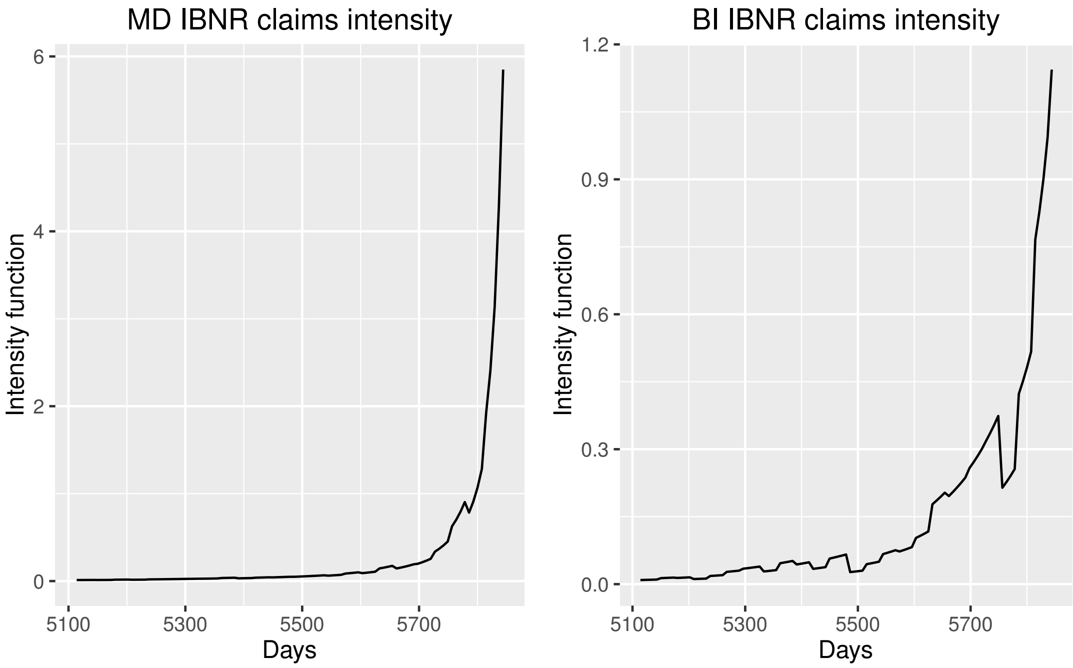

Figure 2.4: Estimated intensity of IBNR claims occurence for material damage and bodily injuries in years 2014 and 2015

both cases for material damage and for bodily injuries. We work here with 37 502 observations for material damage and 5 450 observations for bodily injuries.

Figure 2.3 shows estimated piecewise constant intensities of occurence processes. The intensities show that there might be a seasonal effect and we observe a slight decrease in last few years. Interpretation of the piecewise constant intensity is quite straightforward, for example in December 2015 approximately 5.85 material damage claims are expected to occur per day and similarly in December 2015 approximately 1.15 bodily injuries claims are expected to occur per day. Or more precisely, number of claims occured in one day has Poisson distribution with parameter 5.85 in case of material damage in the last month of 2015 and 1.15 in case of bodily injuries in 2015. Figure 2.4 shows estimated intensity of IBNR claims occurence in the last two years. We remind that equation 1.7 implies that the intensity also determines density, which will be later used for generating times of occurence for IBNR claims, it only needs to be properly scaled.

Integration of the estimated intensities gives us an expected number of IBNR claims, we only need to evaluate

$$\int_0^{\tau} \lambda^{\mathrm{IBNR}}(t)\,\mathrm{d}t,$$

where τ is number of days between December 31, 2015 and December 31, 1999. Using function `integrate()` in `R` we get approximately 219 expected IBNR claims for material damage and 106 for bodily injuries. Influence of different choices of division does not seem to be very large when we compare expected number of IBNR claims for different divisions. We can also compare these numbers with chain ladder method, so that we can somehow assess the estimated quantities. Completed triangles of numbers of material damage and bodily injuries claims are in Table 2.5 and Table 2.6. We used a shorter development history, because with

	0	1	2	3	4	5	6	7
2006	2 477.0	3 033.0	3 056.0	3 064.0	3 065.0	3 065.0	3 065.0	3 065.0
2007	2 512.0	3 018.0	3 046.0	3 049.0	3 055.0	3 055.0	3 055.0	3 055.0
2008	2 723.0	3 195.0	3 232.0	3 238.0	3 239.0	3 239.0	3 239.0	3 239.0
2009	2 135.0	2 525.0	2 554.0	2 561.0	2 562.0	2 562.0	2 562.0	2 562.0
2010	2 034.0	2 404.0	2 460.0	2 474.0	2 474.0	2 474.0	2 474.0	2 474.0
2011	1 767.0	2 110.0	2 148.0	2 159.0	2 160.0	2 160.0	2 160.0	2 160.0
2012	1 791.0	2 110.0	2 136.0	2 139.0	2 140.3	2 140.3	2 140.3	2 140.3
2013	1 803.0	2 076.0	2 093.0	2 098.8	2 100.1	2 100.1	2 100.1	2 100.1
2014	1 716.0	1 962.0	1 986.3	1 991.9	1 993.1	1 993.1	1 993.1	1 993.1
2015	1 646.0	1 947.7	1 971.9	1 977.4	1 978.6	1 978.6	1 978.6	1 978.6

Table 2.5: Completed cumulative development triangle for numbers of reported material damage claims (last two columns omitted)

	0	1	2	3	4	5	6	7	8	9
2006	238.0	373.0	410.0	415.0	415.0	415.0	415.0	415.0	415.0	415.0
2007	248.0	410.0	463.0	470.0	473.0	473.0	473.0	473.0	473.0	473.0
2008	264.0	388.0	438.0	444.0	448.0	448.0	448.0	448.0	448.0	448.0
2009	191.0	311.0	341.0	351.0	352.0	352.0	352.0	352.0	352.0	352.0
2010	183.0	287.0	309.0	312.0	314.0	315.0	315.0	315.0	315.0	315.0
2011	221.0	312.0	334.0	335.0	337.0	337.2	337.2	337.2	337.2	337.2
2012	240.0	321.0	339.0	345.0	346.8	347.0	347.0	347.0	347.0	347.0
2013	230.0	305.0	312.0	316.5	318.1	318.3	318.3	318.3	318.3	318.3
2014	223.0	288.0	313.4	317.9	319.6	319.7	319.7	319.7	319.7	319.7
2015	191.0	280.7	305.5	309.9	311.5	311.6	311.6	311.6	311.6	311.6

Table 2.6: Completed cumulative development triangle for numbers of reported bodily injuries claims

the whole history there were visible trends in residuals. We can easily calculate from these tables that chain ladder results in approximately 372 material damage IBNR claims and approximately 161 bodily injuries IBNR claims. It is possible that expected numbers based on claim-by-claim method underestimate number of IBNR claims, but only overall results of simulations will show whether this detail is important or not.

2.3 Times Between Events

We already briefly discussed this problem in subsection 1.2.2 and in this section we derive likelihoods for continuous distribution of the observed times between events $v_1, v_2, \ldots$ We must realize first that in presence of only one event we observe pairs $(V_1, \delta_1), (V_2, \delta_2), \ldots$, where δ has meaning of a failure indicator, i.e. $\delta_i = \mathbb{I}(V_i \leq W_i)$, where W_i is a time of censoring. Likelihood (or more precisely sublikelihood, because times of censoring are not part of it) can be written (under some usual assumptions) as

$$\prod_{i:\delta_i=0} S(v_i) \prod_{i:\delta_i=1} f(v_i),$$

where f is a continuous density function and S is a survival function, which can be calculated as

$$S(v) = \int_v^\infty f(t)\,\mathrm{d}t, \quad v \geq 0.$$

In presence of two events we assume that the first type has a density f_1 and the second type has a density f_2. Because we assume independent increments, the overall survival function is simply equal to

$$S(v) = S_1(v)\, S_2(v), \tag{2.5}$$

which is implied by equations 1.3 and 1.4. Finally, δ_i is a generalized failure indicator, which can be written as

$$\delta_i = \begin{cases} 0, & \text{if } V_i > W_i, \\ 1, & \text{if } V_i \leq W_i \text{ and the event is of type 1}, \\ 2, & \text{if } V_i \leq W_i \text{ and the event is of type 2}, \end{cases}$$

i.e. zero still means censoring and a positive value refers to the type of event. This leads us to likelihood

$$L = \prod_{i:\delta_i=0} S(v_i) \prod_{j:\delta_j=1} f_1(v_j) \prod_{k:\delta_k=2} f_2(v_k) = L_1\, L_2,$$

where

$$L_m = \prod_{i:\delta_i=0} S_m(v_i) \prod_{j:\delta_j=k} f_m(v_j), \quad m = 1, 2$$

and it means that we can estimate the distributions separately, but censored times are contained in both likelihoods.

Before we progress further, the types of events should be discussed in more detail. We already mentioned that with only two types of events we simplified the situation. Ideally, we would allow settlement without a payment, however, it seems that our data do not contain reliable information about times of settlement. We tried to extract them from the data as the last change of RBNS reserve, but the problem is that many of these changes seem to have an accounting nature. With this suspicion we choose to work only with the reliable part, which are dates of payments. Of course, there are cases where claims are closed some time after the last payment. We can view such a situation in a way that these claims are

	Distribution	Type 1 log likelihood	Type 2 log likelihood
MD	Lognormal	-50 088.35	-100 585.30
	Weibull	-52 511.50	-102 384.80
	Exponential	-52 732.62	-102 536.40
BI	Lognormal	-30 112.09	-11 532.36
	Weibull	-30 963.73	7 969.18
	Exponential	-31 153.56	-11 634.96

Table 2.7: Logarithmic likelihoods for times between payments distributions

	Parameter	Estimate	Standard deviation
Material damage - event 1	μ	3.90580	0.01517
(lognormal distribution)	σ	1.71824	0.01196
Material damage - event 2	μ	4.59656	0.01010
(lognormal distribution)	σ	1.46572	0.00756
Bodily injuries - event 1	μ	4.68635	0.02104
(lognormal distribution)	σ	1.66392	0.01640
Bodily injuries - event 2	μ	5.57852	0.04011
(lognormal distribution)	σ	1.86370	0.03345

Table 2.8: Estimated parameters of times between events

not considered closed at first and another payments are still expected. However, later the claim is closed without any other payment and therefore the last event is updated to type 2 instead of type 1. We note that there might be a problem with more recent claims, where another payment is still expected, but it will be eventually closed later without any other payment.

We should also describe preparation of data used for estimation. Claims with a positive RBNS reserve as at December 31, 2015 are considered as claims where another payment is still expected, therefore their last observed payment is considered as type 1. Other claims with zero RBNS reserve as at December 31, 2015 have some payments of type 1 and the last payment is of type 2. For each claim we calculate time differences in days between payments (or time between the first payment and the notification) while distinquishing type 1 and 2. Finally, observed times of type 0 are calculated as time differences in days between December 31, 2015 and the last observed payment (or the notification, if there is no payment yet). We gather times of type 0 from both files containing payments and RBNS reserve development.

We have an important note regarding data for estimation: before 2013 payments were handled in a different way: each claim settlement was delegated to a member insurance company and date of payment was recorded as date of refundment to the company, while payment from the company had been sent earlier. It implies that development of times between events might not be the same at all times. Because of that, we consider a shorter history for estimation, specifically data concerning claims incurred after December 31, 2007. In case of material damage it would be possible to work with even shorter history, but it might not be a reasonable approach to exclude too many data. All in all, we have 26 609 observations for material damage and 7 464 observations for bodily injuries.

Based on values of logarithmic likelihoods (see Table 2.7), we choose lognor-

	1st Qu.	Median	Mean	3rd Qu.	SD
Material damage	12.4	23.1	39.1	44.9	57.7
Bodily injuries	5.7	17.3	67.5	58.6	169.5

Table 2.9: Descriptive statistics of payments (in thousand CZK)

mal distributions for material damage. For bodily injuries we choose lognormal distributions as well, but we should note that Weibull distributions have significantly larger maximized likelihood when we sum it for Type 1 and Type 2. However, Table 2.7 reveals that logarithmic likelihood for Type 2 is a little bit suspicious. Moreover, a preparation of simulation routine reveals that this choice would be inappropriate, because a comparison of the estimated Weibull hazard functions leads us to conclusion that number of future payments would be very likely overestimated. From these reasons we omit Weibull distribution from our choice for bodily injuries.

Table 2.8 summarizes the estimated distributions and their estimated parameters. We can mention expected values for the estimated distributions: 259 and 233 days for material damage (type 1 and 2 respectively); 323 and 1 633 days for bodily injuries (type 1 and 2 respectively). Later we will need product of cumulative density functions to evaluate probabilities that an event occurs before some selected time, which will be used for sampling times between events.

2.4 Payments

We have few important remarks to payments. Firstly, we do not adjust payments for inflation and take them as they are. Secondly, all payments are positive, i.e. our data do not contain information about salvages and subrogiations. Finally, we noticed the data contain repetitive payments in amounts 500, 1 000, 2 500 and 2 783 CZK, which can be considered as zeroth payments, which are not used anymore. Such payments are actually a remainder of a revoked rule regarding payments, therefore we exclude the mentioned amounts from the data. Table 2.9 contains few descriptive statistics of observed payments and Table 2.11 summarizes estimated distributions for payments.

Because we treat payments as iid random variables, it is relatively easy to estimate this part. We compare exponential, Weibull, Burr and lognormal distribution and based on their maximized likelihoods the most suitable distribution is lognormal in both cases (see Table 2.10). The estimates are based on 41 465 observations for material damage and 11 603 observations for bodily injuries. It is relatively easy to calculate that expected values are 39 021 CZK for material damage and 72 079 CZK for bodily injuries.

We can take a look at Figure 2.5, where is a comparison of theoretical and sample quantiles for logarithms of payments. We observe that the left tail of material damage does not fit our data well, however, this tail is not as important as the right tail. A graphical comparison of estimated densities with histograms can be found in Figure 2.6. We can see that lognormal distribution fits bodily injuries very nicely. Table 2.10 contains also results of performed chi-squared tests in the same way as it was done for delays distributions and the same critique applies here.

	Distribution	Log likelihood	Test statistic	DF	Critical value
MD	Lognormal	-476 464.60	694.70	47	64.00
	Weibull	-479 853.30	5 830.61	47	64.00
	Exponential	-479 858.00	5 872.04	48	65.17
	Burr	-554 627.80	NA	NA	NA
BI	Lognormal	-136 153.70	126.89	47	64.00
	Weibull	-137 003.10	1 362.00	47	64.00
	Exponential	-140 619.40	7 657.25	48	65.17
	Burr	-151 911.40	NA	NA	NA

Table 2.10: Logarithmic likelihoods and results of (chi-squared) goodness-of-fit tests for payments distribution

	Parameter	Estimate	Standard deviation
Material damage	μ	10.06658	0.00494
(lognormal distribution)	σ	1.00526	0.00349
Bodily injuries	μ	9.80910	0.01540
(lognormal distribution)	σ	1.65917	0.01089

Table 2.11: Estimated parameters of payments

This section concludes the practical part, because all we needed has been estimated and we can finally simulate future developments in the next chapter to obtain results.

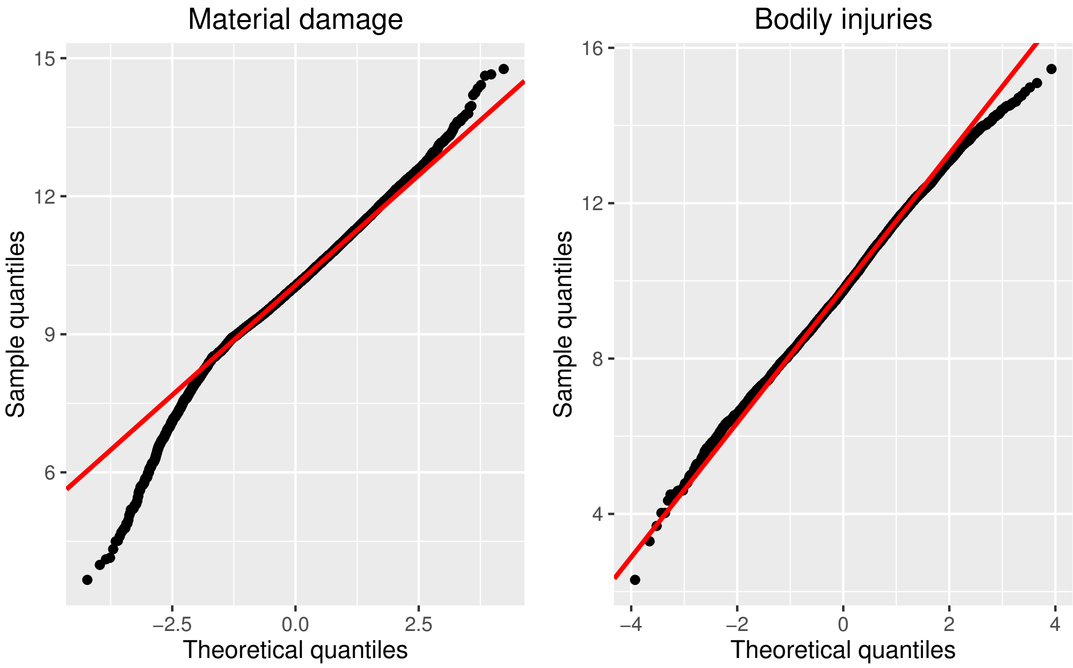

Figure 2.5: Normal Q-Q plots for logarithms of payments

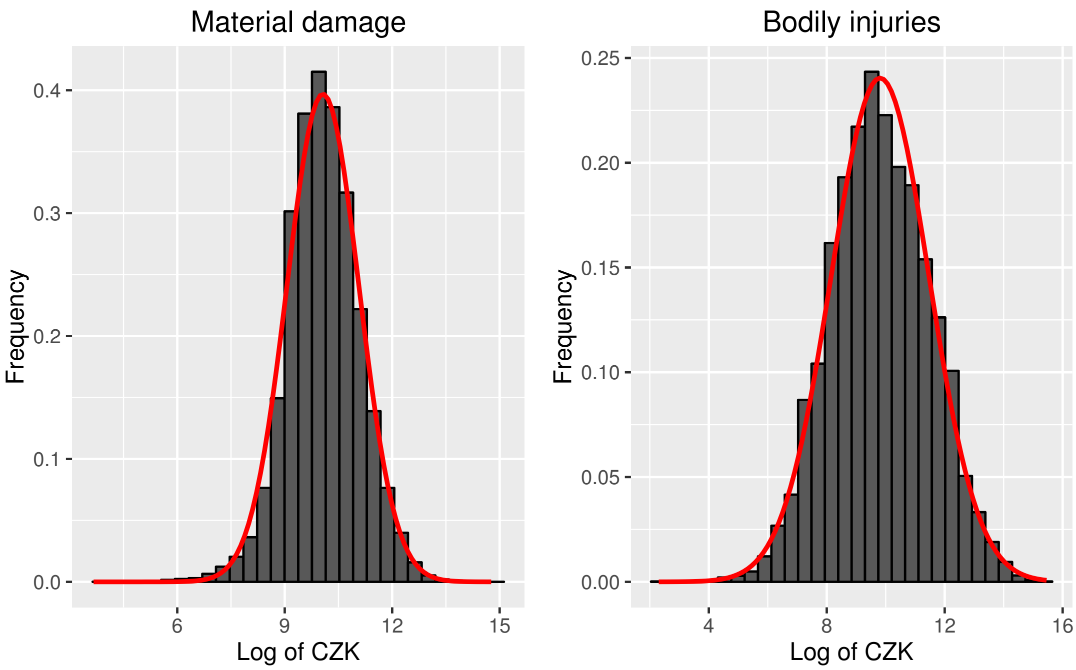

Figure 2.6: Histograms of logarithms of payments and estimated normal density functions

3. Simulation

This chapter utilizes the previous chapters. We describe our simulation algorithm in more detail (especially parts that were not mentioned yet), then we review results and finally compare it with chain ladder method and bootstrap. The simulation algorithm is inspired by Antonio and Plat (2014) with necessary adjustments implemented.

3.1 Simulation Algorithm

At the beginning we prepare a data frame containing RBNS claims, in our set up it is sufficient to keep only information about dates of last payment for each claim. This part can be done only once, other parts are repeated in each simulation and are described in the following subsections. However, the file containing payments contains only part of RBNS claims, because there are still some RBNS claims in the other file, where we need to extract claims with positive RBNS reserve and without any payment yet.

3.1.1 Parameters

In each simulation we sample new parameters for times between events and new parameters for payments from asymptotic distributions implied by maximum likelihood theory. Estimated parameters take role of mean values and variance matrices are obtained as inverse of negative hessian matrices. New parameters are then sampled with function `rmvnorm()` from package mvtnorm. We do not sample new parameters for delay distribution and intensity of occurence process, because we believe that the impact would be relatively small.

3.1.2 Occurence and Delay of IBNR Claims

We recall equations 1.6 and 1.7, their interpretation and add conclusions from the section devoted to division of claims. Number of IBNR claims is a random variable that has Poisson distribution with parameter $\Lambda^{\mathrm{IBNR}}(\infty)$ and this number was already calculated earlier, it is approximately 219 for material damage and approximately 106 for bodily injuries. Random numbers of IBNR claims can be easily generated with function `rpois()` in `R`.

With given number of IBNR claims, we generate each time of occurence from a density function

$$f(t) = \frac{\lambda^{\mathrm{IBNR}}(t)}{\Lambda^{\mathrm{IBNR}}(\infty)},$$

which is simply a rescaled intensity function in Figure 2.4. This is a density, which is not predefined in `R`, therefore we need to define the generating procedure by ourselves. We calculate a set of points $(x, F(x))$, reverse them to $(F(x), x)$ and we finish definition of quantile function by a simple linear interpolation. Since we use the linear interpolation, we need to define the quantile function for too low values. For IBNR claims inccurred three and more years back in the history, we set their time occurrence to three years back (specifically we set it to $\tau - 3 * 365$).

In the current model, this does not affect final estimates of IBNR and RBNS reserve (it can slightly affect only payments during the future one-year window).

For each IBNR claim we also generate a random delay in notification. We generate a random number p from uniform distribution between zero and one and we find u which fulfills

$$\mathsf{P}\left(U_t \leq u \,|\, U_t > \tau - t\right) = p.$$

Firstly, we realize that the left-hand side can be rewritten to

$$\frac{F_{U_t}(u) - F_{U_t}(\tau - t)}{1 - F_{U_t}(\tau - t)},$$

therefore we want to find u which satisfies

$$F_{U_t}(u) = p\left[1 - F_{U_t}(\tau - t)\right] + F_{U_t}(\tau - t). \tag{3.1}$$

Secondly, we recall equation 2.2 and realize that in case of lognormal distribution it means that parameter σ is the same and μ changes to $\mu + t\log(b)$, in case of Weibull distribution it means that shape parameter α is the same and scale parameter β changes to $b^t\beta$. To sum up, we can simply use quantile functions `qlnorm()` and `qweibull()` in `R` with the corresponding parameters to invert equation 3.1 and get the delay u.

3.1.3 Times of Next Event

We put IBNR and RBNS claims together, because the next part of simulation can be easily done for both categories at once. Considering a selected RBNS claim, the last payment was paid t_{last} days from December 31, 1999 and there is a censoring $c = \tau - t_{\text{last}}$. In case of IBNR claims we simply set $c = 0$. In both cases, we want a random number from conditional distribution $V \,|\, V > c$. We recall equation 2.5, which implies that the overall cumulative density function is

$$F(v) = F_1(v)F_2(v).$$

We generate a random number p from uniform distribution between zero and one and the same argument as in equation 3.1 gives us that we want find v such that

$$F(v) = p\left[1 - F(c)\right] + F(c).$$

We use a similar approach for definition of the quantile function as for occurrences of IBNR claims, we only need to define interpolated quantile function for too low and too large values of p. Too low values of p are not a problem, because the quantile function changes from values less than one to approximately one. We believe that the right tail will not cause any problems as well, because large times of a next payment have already similar probabilities of event's type (see the next subsection). This is also supported by a simulation performed with different, more precise, but more time consuming method (in diploma thesis with the same name) - differences are indeed negligible. Furthermore, if a larger precision is desired, it is sufficient to interpolate the quantile function on more dense grid.

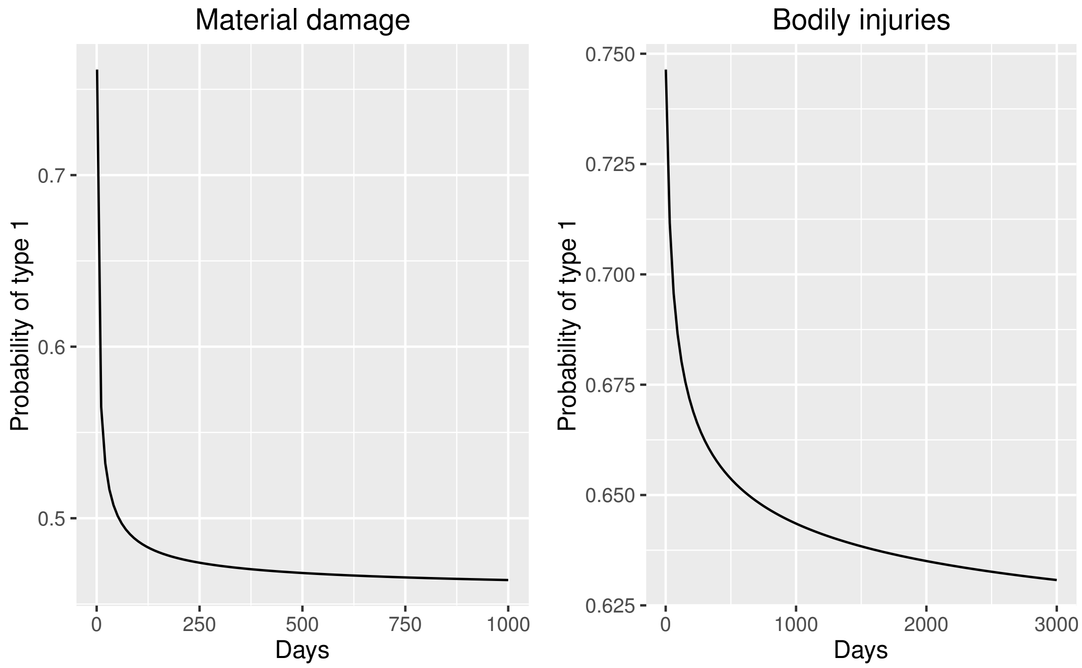

Figure 3.1: Probabilities of event 1 at various times

3.1.4 Types of Next Event

We recall the last statement of Theorem 1: when an event occured at time v, the event is of type 1 with probability equal to

$$p(v) = \frac{\lambda_1(v)}{\lambda_1(v) + \lambda_2(v)},$$

where λ_1, λ_2 are hazard functions, which can be interpreted also as intensity functions. They are calculated (for continuous distributions) as ratio of density and survival function, i.e.

$$\lambda_m(v) = \frac{f_m(v)}{S_m(v)}, \quad m = 1, 2.$$

A possible way to obtain randomly an event with prescribed probability $p(v)$ is the following: we generate a random number p from uniform distribution between zero and one and compare it with $p(v)$. If p is less than or equal to $p(v)$, the next event is of type 1 and we will need to generate another time and type of next event. Otherwise, the next event is of type 2 and it means there will not be another payment in future.

Figure 3.1 shows development of the probability in time. It is possible to calculate that in case of material damage the function is equal to one half at approximately 53 days, i.e. when a next event takes place after a longer time than the mentioned 53 days, it has probability of settlement greater than one half. Probability of settlement for bodily injuries is always smaller than one half, which corresponds to the observed behavior, where usually several payments precede the last payment. We can also see that for large value of days the probabilities of event's type do not change much.

		1st Qu.	Median	Mean	3rd Qu.	SD
MD	No. of payments	3 162	3 206	3 207	3 251	65
	Total liability	122	125	125	128	4
	Liability in 2016	48	50	50	51	2
BI	No. of payments	2 890	2 961	2 963	3 034	106
	Total liability	202	212	214	224	18
	Liability in 2016	24	27	28	31	6

Table 3.1: Summary of results from simulation separately for material damage and bodily injuries (liability values in million CZK)

3.1.5 Payments

Regardless of the type of the last event we generate a payment at time v. This part is very simple, because we use function `rlnorm()` in `R` to generate random numbers from distribution for payments with the corresponding parameters. If the last event is of type 1, we generate another time of event, type and payment, until we get an event of type 2.

3.2 Results

We have few remarks regarding our results. Firstly, we remind that the model is based on few simplifications, mostly implied by data insufficiencies. There is still a space for improvement, for instance in connection with payments, where a more sophisticated model could be used. Times between payments are very important too and other models for them might be worth to examine. Secondly, a comparison of 1 000 and 10 0000 simulations suggests that the results are very similar, but in case of bodily injuries the higher number of simulations is more appropriate. The reason is that in the second case also few scenarios with large severity occured, and therefore it captures the underlying risk and uncertainty better. In case of material damage the smaller number of simulations would suffice, because results are very comparable.

Table 3.1 summarizes results obtained from our simulations performed in `R`. Each simulation starts with random seed set to 22121992, so that both simulations can be repeated with the same result. At the time of finishing the thesis we received new data about payments for material damage and bodily injuries in 2016, hence we are able to compare our results at least with payments in 2016. More specifically, we can calculate paid amounts in 2016 for claims that occured before 2016, which is 29.5 million CZK for material damage and 35.1 million CZK for bodily injuries. The observed number for bodily injuries is still within estimated 95% confidence interval, while the observed number for material damage is not. This particular result of individual claims reserving approach for bodily injuries is much more realistic then estimated payments in 2016 from chain ladder, which are 68.3 million CZK for material damage and 72.4 million CZK for bodily injuries. The last column of Table 3.1, where SD stands for standard deviation, shows us that bodily injuries have more volatile results than material damage, when we compare ratio of standard deviation and mean (coefficient of variation), which could be expected a priori, because bodily injuries claims

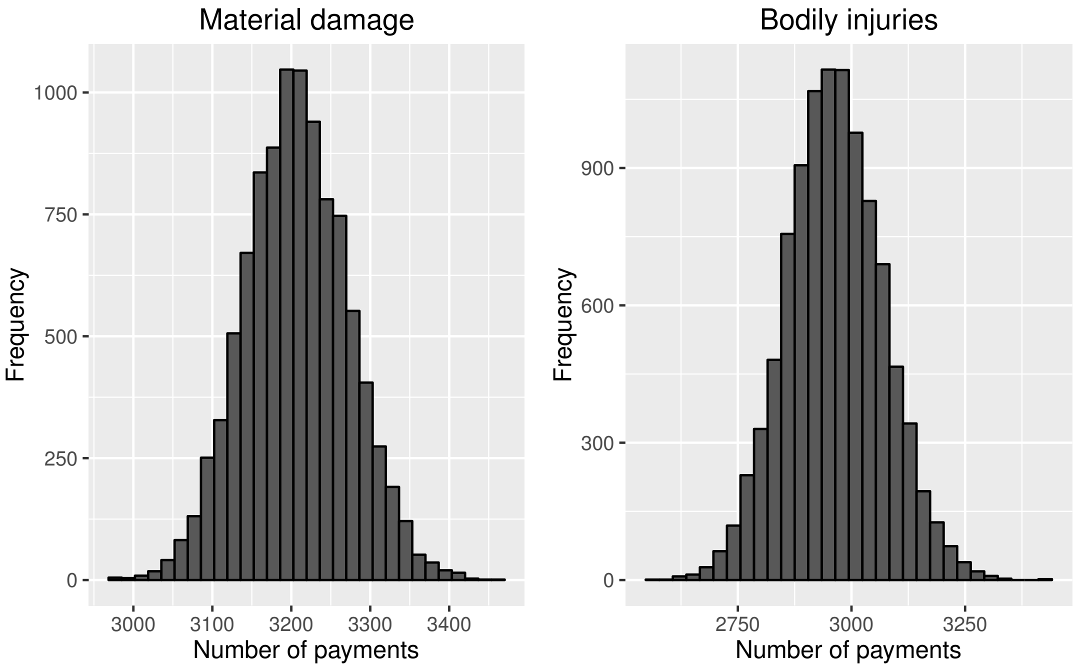

Figure 3.2: Simulated numbers of future payments

have more complicated developments than material damage claims. On the other hand, volatility of material damage seems to be low, considering the magnitude of the average future liability.

Figure 3.2 shows results of simulations with respect to number of future payments. We can easily verify that average number of future payments multiplied by expected value of payments gives us a number very close to average total liability. This should not be surprising, because our simulation almost follows the well-known collective risk model, where the total liability L is considered as

$$L = \sum_{i=1}^{N} P_i,$$

where N is a random number of payments, identically distributed random variables $P_1, P_2, \ldots$ correspond to payments and all random variables are mutually independent. It is easy to derive that in the collective model holds

$$\mathsf{E}\left[L\right] = \mathsf{E}\left[N\right] \mathsf{E}\left[P_1\right].$$

The equality is fulfilled almost precisely in our case, even though we have only estimates of the respective expected values and additionaly some parameters are generated from their asymptotic distribution in order to include parameter uncertainty in the model, which slightly violates the assumptions of the collective model.

Figure 3.3 contains histograms of simulated total liability, the top row corresponds to our claim-by-claim model and the bottom row contains results of bootstrap. We also see a comparison with RBNS reserve, which is particularly interesting in case of bodily injuries (impliying possibly a negative IBNR reserve). Average simulated liabilities in amounts of 125.1 million CZK and 213.8 million

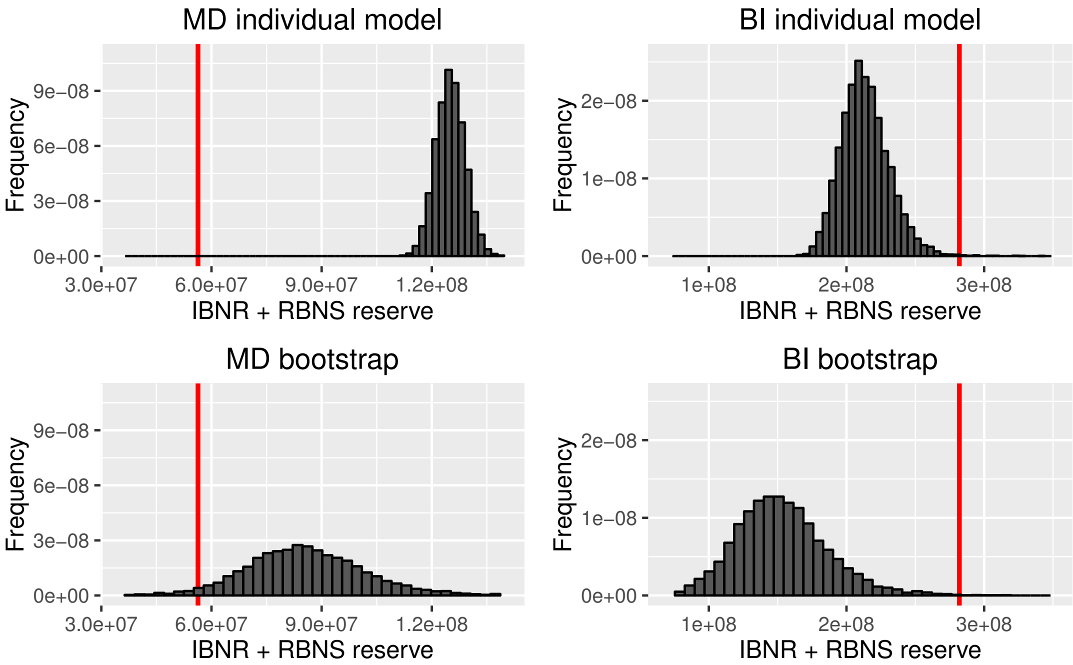

Figure 3.3: Simulated reserve (IBNR + RBNS) based on claim-by-claim data and bootstrap (red line is RBNS reserve)

CZK can be interpreted as best estimates, but they almost coincide with medians, so future liability will be larger with probability roughly one half. We note that 99.5% quantile is 135.4 million CZK for material damage and 269.4 million CZK for bodily injuries.

To sum up, the results for bodily injuries seem to be valid and acceptable, but the results for material damage raise a red flag. One suspicious thing is that payments in 2016 do not fit the reality very well, as we already observed. Nevertheless, this comparison is not available to actuaries, when they calculate technical provisions. However, the low coefficient of variation is also an indicator that something might be wrong.

3.3 Comparison with Chain Ladder

We use chain ladder method to see how it performs on our data and for comparison with our results. We construct development triangles of paid amounts divided by years. We do not use the full available history because of the inconsistency already mentioned in the previous chapter. We omit the first six calendar years for material damage and the first three years for bodily injuries. We should exclude even more periods, but there would arise a problem with too short development history. Paid triangles provide us the following results: 85.7 million CZK for material damage and 152.5 million CZK for bodily injuries, the estimates refer to IBNR and RBNS reserve together. Completed paid triangles can be found in Table 3.2 and Table 3.3. Since incurred triangles produce much lower results, we do not take them into consideration. We note that in case of material damage the corresponding RBNS reserve (see Figure 3.3) is 56.2 million CZK and in case of bodily injuries RBNS reserve is 281.9 million CZK, which is

	0	1	2	3	4	5	6	7	8	9
2006	42.0	125.0	136.3	139.6	140.0	140.0	140.2	140.2	140.4	140.5
2007	56.1	123.2	140.2	143.9	144.5	144.7	144.7	144.7	144.7	144.7
2008	62.3	139.9	149.0	151.7	153.0	153.9	154.1	154.1	154.2	154.3
2009	50.7	104.0	109.8	112.2	113.6	113.7	113.7	113.7	113.8	113.8
2010	55.3	97.9	105.7	108.4	108.9	108.9	109.0	109.0	109.1	109.1
2011	47.5	90.5	94.9	97.6	98.1	98.3	98.4	98.4	98.5	98.5
2012	40.1	77.4	82.1	83.2	83.6	83.8	83.9	83.9	83.9	84.0
2013	44.9	78.3	80.9	82.7	83.2	83.4	83.4	83.4	83.5	83.5
2014	55.4	80.9	87.0	89.0	89.5	89.6	89.7	89.7	89.8	89.8
2015	58.4	117.8	126.6	129.5	130.3	130.5	130.6	130.6	130.7	130.8

Table 3.2: Completed cumulative paid triangle (in million CZK) for material damage

very prudent number when we take into account results of two different methods.

The above mentioned estimates of claims reserve are, however, only point estimates and it is much better to compare two distributions. Table 3.4 contains results of bootstrap applied on paid triangles in `R` calculated with function `BootChainLadder()` from package ChainLadder. We set random seed in both cases to 20121992, number of replicates to ten thousand and we choose gamma process distribution. Figure 3.3 depicts all simulated results by bootstrap in histograms and compares them with our claim-by-claim model. We can easily obtain results with a larger number of replicates, but we choose the same number as in our simulation in order to obtain comparable histograms. On the other hand, mean value and standard deviation do not change drastically with different number of replicates, hence we consider the chosen number of replicates sufficient.

We observe that chain ladder and bootstrap provide lower results, but the approach based on individual claims provides much lower standard deviation. This might be a consequence of using more information in the procedure of estimation, although the standard deviation for material damage is suspicious. We note that in case of chain ladder we were able to discard less older data than in case of individual claims reserving method. The corresponding 99.5% quantiles obtained by bootstrap are 134.3 million CZK for material damage and 260.7 million CZK for bodily injuries. It means that in this case estimates based on claim-by-claim data are placed in the upper tail of chain ladder estimates and are more prudent.

	0	1	2	3	4	5	6	7	8	9	10	11
2003	1.5	16.5	37.5	47.9	52.3	53.1	53.5	53.6	53.7	53.8	53.9	54.0
2004	3.6	28.2	44.3	53.5	57.1	58.0	58.9	59.0	59.1	59.2	59.2	59.2
2005	7.0	37.3	50.7	59.6	62.3	63.8	65.0	65.7	65.8	69.5	69.5	69.6
2006	6.2	33.3	53.3	63.6	65.6	66.6	67.0	67.1	67.5	67.6	67.7	67.7
2007	8.3	41.8	64.2	70.8	73.8	74.3	74.4	74.9	74.9	76.1	76.1	76.2
2008	11.4	47.8	63.8	74.8	84.2	86.0	86.6	87.7	88.0	89.4	89.4	89.5
2009	9.1	34.4	51.6	57.3	61.2	61.8	62.6	63.0	63.2	64.2	64.2	64.3
2010	7.2	28.5	44.0	51.9	53.6	53.8	54.3	54.7	54.8	55.7	55.7	55.8
2011	8.1	29.9	38.0	40.3	40.7	41.3	41.7	41.9	42.0	42.7	42.7	42.8
2012	13.3	36.4	46.3	48.4	51.3	52.0	52.5	52.8	53.0	53.8	53.9	53.9
2013	15.4	32.8	38.4	44.2	46.9	47.6	48.0	48.3	48.4	49.2	49.2	49.3
2014	16.2	46.6	67.6	77.8	82.4	83.6	84.4	84.9	85.2	86.5	86.6	86.6
2015	13.8	53.4	77.4	89.1	94.4	95.8	96.7	97.3	97.5	99.1	99.2	99.2

Table 3.3: Completed cumulative paid triangle (in million CZK) for bodily injuries (with last column omitted)

	1st Qu.	Median	Mean	3rd Qu.	SD
Material damage	75	85	86	96	16
Bodily injuries	131	151	153	173	33

Table 3.4: Summary of results from bootstrap separately for material damage and bodily injuries (values in million CZK)

Conclusion

Objective of the thesis was to summarize theory regarding loss reserving for individual claim-by-claim data, apply it on real data and perform a simulation. Each part is described in its own chapter. Chapter 1 describes the necessary theory; we partially followed theory from the literature, but we derived everything more formally in terms of probability densities. Chapter 2 contains our application on real data from the Czech market. Finally, Chapter 3 presents our simulation algorithm, results and comparison with chain ladder and bootstrap.

Although the final model might appear too complex, it has a nice advantage of flexibility, which can be incorporated in the model. For example, we noticed that delays contain a trend, which we estimated and then we took it into account in later estimations. Another example of flexibility is choice of data for estimation: we realized that times of payments before 2013 were recorded differently, so we restricted the disruptive effect of the older data. We could not restrict it completely, but we were able to discard more older data than in case of chain ladder, where we were limited by not fully developed triangle.

We partially included a parameter uncertainty in the model, where parameters of times between events and parameters of payments are sampled from asymptotic normal distribution implied by maximum likelihood theory. We simplified the situation in a way that delay distribution and intensity of occurence process are treated as given, because we believe that their influence is rather small.

The model is based on few simplifications, so there is still a space for improvement. One simplification concerns intensity function of occurence process, where we used a piecewise constant specification. This approach allowed us to omit exposure, which is not included in our data. However, when exposure is available, a parametric approach might slightly improve the model. Another simplification was incorporated with respect to payments, where we assumed that they are independent and identically distributed. An improvement regarding simulation routine might concern an optimization of the procedure, because 10 000 simulations last more than three hours each (and this increases with number of open claims and expected IBNR claims). Finally, we could sample parameters from asymptotic distribution even for delay and intensity of occurence process, which would lead to full inclusion of parameter uncertainty.

All in all, we believe that individual claims reserving method is very promising and worth of a further research. This conclusion is based on our comparison of results with chain ladder and bootstrap. Another comparison with payments in 2016 suggests that in our case claim-by-claim data provided a more realistic view on payments in the next year than chain ladder. Although individual claims reserving method is more complex than other methods, we believe it is an interesting alternative to methods based on aggregate data.

Bibliography

Antonio, K. and Plat, R. (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, No.7:pp. 649–669.

Merz, M. and Wüthrich, M. V. (2008). *Stochastic Claims Reserving Methods in Insurance*. Wiley Finance, London, UK.

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance. *ASTIN Bulletin*, 23(1):pp. 95–115.

Norberg, R. (1999). Prediction of outstanding liabilities: Model variations and extensions. *ASTIN Bulletin*, 29(1):pp. 5–25.

Ross, S. M. (2010). *Introduction to Probability Models*. 10th edition. Elsevier, Boston, USA.

Appendix

Material damage - highlights of the source code

```
library(MASS)
library(mvtnorm)
library(ggplot2)
library(gridExtra)

# Load data
plneni <- read.csv("plneni.csv", header = TRUE, sep = ";")
plneni$CASTKACZK <- as.numeric(plneni$CASTKACZK)
plneni$DATUMCASSU <- as.Date(plneni$DATUMCASSU)
plneni$DATUMEVIDENCE <- as.Date(plneni$DATUMEVIDENCE)
plneni$den.platby <- as.Date(plneni$den.platby)
sapply(plneni, class)

# create subset of MD
md <- plneni[plneni$DRUH_PLNENI == "vecna skoda", ]

reserve <- read.csv("rezerva.csv", header = TRUE, sep = ";")
reserve$zmena.v.Kc <- as.numeric(reserve$zmena.v.Kc)
reserve$DATUMCASSU <- as.Date(reserve$DATUMCASSU)
reserve$DATUMEVIDENCE <- as.Date(reserve$DATUMEVIDENCE)
reserve$den.zmeny <- as.Date(reserve$den.zmeny)
sapply(reserve, class)

# create subset of MD
mdr <- reserve[reserve$DRUH_PLNENI == "vecna skoda", ]

# Order data
# Aggregate same IDs and sort them with respect to days of payment
md <- md[order(md$IDSKODNIUDALOST,md$den.platby), ]
mdr <- mdr[order(mdr$IDSKODNIUDALOST, mdr$den.zmeny), ]
md$order <- 1
mdr$order <- 1

# Calculate order of payments
for (i in 2:nrow(md)){
  if (md$IDSKODNIUDALOST[i]==md$IDSKODNIUDALOST[i-1]){
    md$order[i] <- md$order[i-1] + 1
  }
}

for (i in 2:nrow(mdr)){
  if (mdr$IDSKODNIUDALOST[i]==mdr$IDSKODNIUDALOST[i-1]){
    mdr$order[i] <- mdr$order[i-1] + 1
  }
}

# Remove 9 claims where notification is before occurence
exclude <- md$IDSKODNIUDALOST[md$DATUMCASSU>md$DATUMEVIDENCE]
md <- md[!md$IDSKODNIUDALOST %in% exclude, ]

# Remove 32 claims where notification is before occurence
exclude <- mdr$IDSKODNIUDALOST[mdr$DATUMCASSU>mdr$DATUMEVIDENCE]
```

```
mdr <- mdr[!mdr$IDSKODNIUDALOST %in% exclude, ]

# Remove 1 claim where payment is before notification
exclude <- md$IDSKODNIUDALOST[md$DATUMEVIDENCE>md$den.platby]
md <- md[!md$IDSKODNIUDALOST %in% exclude, ]

# Remove 2 claims where change of reserve is before notification
exclude <- mdr$IDSKODNIUDALOST[mdr$DATUMEVIDENCE>mdr$den.zmeny]
mdr <- mdr[!mdr$IDSKODNIUDALOST %in% exclude, ]

md[md$CASTKACZK<=0, ] # There are no rows with negative payments

# Intersection of claims
md$include <- 0
for (i in 1:dim(md)[1]){
  if (nrow(mdr[mdr$IDSKODNIUDALOST==md$IDSKODNIUDALOST[i], ])>0){
    md$include[i] <- 1
  }
}
md[md$include==0, ] # One claim not included in mdr
md <- md[md$include==1, ] # We exclude it

mdr$include <- 0
for (i in 1:dim(mdr)[1]){
  if (nrow(md[md$IDSKODNIUDALOST==mdr$IDSKODNIUDALOST[i], ])>0){
    mdr$include[i] <- 1
  }
}

# Add information about current value of RBNS reserve
md$reserve <- 0
for (i in 1:nrow(md)){
  md$reserve[i] <- sum(mdr$zmena.v.Kc[mdr$IDSKODNIUDALOST
                                      ==md$IDSKODNIUDALOST[i]])
}

# few rows have RBNS reserve -2, -1 or 1, we change it to zero
md$reserve[md$reserve<=1] <- 0

mdr$reserve <- 0
for (i in 1:nrow(mdr)){
  mdr$reserve[i] <- sum(mdr$zmena.v.Kc[
    mdr$IDSKODNIUDALOST==mdr$IDSKODNIUDALOST[i]])
}


### Distribution of delay in notification
md$delay <- as.numeric(md$DATUMEVIDENCE - md$DATUMCASSU)
mdr$delay <- as.numeric(mdr$DATUMEVIDENCE - mdr$DATUMCASSU)

## Estimation of acceleration in notification
date0 <- as.Date("1999-12-31")

divisionyearly <- seq.Date(from=date0, by="years", length=16*1+1)
lengths2 <- rep(0, length(divisionyearly))
for (i in 2:length(divisionyearly)) {
  lengths2[i] <- as.numeric(divisionyearly[i]-divisionyearly[i-1])
} # lengths contain lengths of respective years
```

```
divisionyearly2 <- cumsum(lengths2)
divisionyearly2

# present moment in days
tau <- as.numeric(divisionyearly[length(divisionyearly)] - date0)

# Create subset of delays for different claims
delaymd <- mdr[mdr$order==1, ]
summary(delaymd$delay)
sd(delaymd$delay)

delaymd$t <- as.numeric(as.Date(delaymd$DATUMCASSU) - date0)
delaymd <- delaymd[, c("delay", "t")] # Reduce columns

# Quantify effect of time occurence, exclude last year
lmmd <- lm(delay~1+t, data=delaymd[delaymd$t<=divisionyearly2[16],])

# Exponential trend
fexp <- function(a, b, t) {
  a*b^t
}
fitmd1 <- nls(delay ~ fexp(a, b, t),
              data=delaymd[delaymd$t<=divisionyearly2[16], ],
              start=c(a=300, b=1))
cmd1 <- coef(fitmd1)

x <- delaymd[delaymd$t<=divisionyearly2[16], ]
suma1 <- sum((x$delay - fexp(cmd1[1], cmd1[2], x$t))^2)
suma2 <- sum((x$delay - mean(x$delay))^2)
1-suma1/suma2 # R-squared

bmd <- cmd1[2] # Exponential trend is more suitable

# Transform observed values
delaymd$delay2 <- delaymd$delay / bmd^delaymd$t

chisquaretest <- function(observations, quantiles, par) {
  counts <- 1:(length(quantiles)-1)
  for (i in 1:length(counts)) {
    counts[i] <- length(observations[observations>=quantiles[i]
                                     & observations<quantiles[i+1]])
  }
  expected <- length(observations) / length(counts)
  chisqsum <- sum((counts-expected)^2 / expected)
  df <- length(counts) - par - 1
  message("p-value: ", 1 - pchisq(chisqsum, df))
  message("statistic: ", chisqsum)
  message("critical value: ", qchisq(0.95, df))
  print(counts)
  message("expected value in each interval: ", expected)
}

## Exponential distribution
fitexpmd <- fitdistr(delaymd$delay2[delaymd$delay2>0], "exponential")
x <- delaymd$delay2
breaks <- qexp(seq(from=0, to=100, by=2)/100,
               rate=fitexpmd$estimate[1])
chisquaretest(x, breaks, 1)
```

```
## Gamma distribution
fitgammamd <- fitdistr(delaymd$delay2[delaymd$delay2>0],
                       "gamma", start=list(shape=1, rate=1))
x <- delaymd$delay2
breaks <- qgamma(seq(from=0, to=100, by=2)/100,
                 shape=fitgammamd$estimate[1],
                 rate=fitgammamd$estimate[2])
chisquaretest(x, breaks, 2)

## Weibull distribution
fitweibullmd <- fitdistr(x=delaymd$delay2[delaymd$delay2>0],
                         densfun="weibull")
x <- delaymd$delay2
breaks <- qweibull(seq(from=0, to=100, by=2)/100,
                   shape=fitweibullmd$estimate[1],
                   scale=fitweibullmd$estimate[2])
chisquaretest(x, breaks, 2)

## Burr distribution
burrdensity <- function (x, c, k) {
  c * k * x^(c-1) / (1 + x^c)^(k+1)
}
fitburrmd <- fitdistr(x=delaymd$delay2[delaymd$delay2>0],
                      densfun=burrdensity, start=list(c=1, k=1),
                      lower=list(c=0, k=0))

## Lognormal distribution
fitlognormalmd <- fitdistr(x=delaymd$delay2[delaymd$delay2>0],
                           densfun="lognormal")
meanlmd <- fitlognormalmd$estimate[[1]]
sdlmd <- fitlognormalmd$estimate[[2]]
x <- delaymd$delay2
breaks <- qlnorm(seq(from=0, to=100, by=2)/100,
                 meanlog=meanlmd, sdlog=sdlmd)
chisquaretest(x, breaks, 2)


### Piecewise intensity function of the occurence process
division <- seq.Date(from=date0, by="months", length=16*12+1)

# Numbers of reported claims with respect to the given division
countmd <- rep(0, length(division) - 1)
for (i in 1:(length(division) - 1)) {
  countmd[i] <- nrow(mdr[mdr$order==1 &
                         mdr$DATUMCASSU>division[i] &
                         mdr$DATUMCASSU<=division[i+1], ])
}

lengths <- rep(0, length(division))
for (i in 2:length(division)) {
  lengths[i] <- as.numeric(division[i] - division[i-1])
} # lengths of respective intervals and zero in beginning
division2 <- cumsum(lengths)

integrandmd <- function(t) {
  plnorm((tau - t) / bmd^t, meanlog=meanlmd, sdlog=sdlmd)
}
```

```
integratedcdf <- rep(0, length(division2) - 1)
for (i in 1:(length(division2) - 1)) {
  integratedcdf[i] <- integrate(integrandmd, lower=division2[i],
                                upper=division2[i+1])$value
}

# Estimate of intensity of the occurence process
lambdahatmd <- countmd / integratedcdf
intensityRBNSmd <- function(t) {
  lambdahatmd2(t)*integrandmd(t)
}
intensityIBNRmd <- function(t) {
  lambdahatmd2(t)*(1 - integrandmd(t))
}


### Payments
subsetmd <- md$CASTKACZK[md$CASTKACZK!=500 &
                         md$CASTKACZK!=1000 &
                         md$CASTKACZK!=2500 &
                         md$CASTKACZK!=2783]
nrow(md) - length(subsetmd) ### 5 646 excluded observations

fitpmd1 <- fitdistr(subsetmd, densfun="exponential")
breaks <- qexp(seq(from=0, to=100, by=2)/100,
               rate=fitpmd1$estimate[1])
chisquaretest(subsetmd, breaks, 1)

fitpmd2 <- fitdistr(subsetmd, densfun="weibull")
breaks <- qweibull(seq(from=0, to=100, by=2)/100,
                   shape=fitpmd2$estimate[1],
                   scale=fitpmd2$estimate[2])
chisquaretest(subsetmd, breaks, 2)

fitpmd3 <- fitdistr(subsetmd, densfun="lognormal")
breaks <- qlnorm(seq(from=0, to=100, by=2)/100,
                 meanlog=fitpmd3$estimate[1],
                 sdlog=fitpmd3$estimate[2])
chisquaretest(subsetmd, breaks, 2)

fitpmd4 <- fitdistr(subsetmd , densfun=burrdensity,
                    start=list(c=1, k=1), lower=list(c=0, k=0))


### Days between events
md$timeDif <- as.numeric(md$den.platby - md$DATUMEVIDENCE)
for (i in 2:nrow(md)) {
  if (md$order[i] > 1) {
    md$timeDif[i]<-as.numeric(md$den.platby[i]-md$den.platby[i-1])
  }
}

md$event <- 1 # One is for payments
for (i in 2:nrow(md)) {
  if (md$IDSKODNIUDALOST[i-1] != md$IDSKODNIUDALOST[i] &
      md$reserve[i-1] == 0) {
    md$event[i-1] <- 2 # Two means the last payment
```

```
  }
}
if (md$reserve[nrow(md)] == 0) {
  md$event[nrow(md)] <- 2
}

md$survival <- -1
for (i in 2:nrow(md)) {
  if (md$IDSKODNIUDALOST[i-1]!=md$IDSKODNIUDALOST[i] &
      md$reserve[i-1] > 0) {
    md$survival[i-1] <- as.numeric(division[length(division)] -
                                     md$den.platby[i-1])
  }
}
if (md$reserve[nrow(md)]>0) {
  md$survival[nrow(md)] <- as.numeric(division[length(division)] -
                                        md$den.platby[nrow(md)])
}

mdr$survival <- -1
for (i in 1:nrow(mdr)) {
  if (mdr$include[i]==0 & mdr$reserve[i]>0 & mdr$order[i]==1) {
    mdr$survival[i] <- as.numeric(division[length(division)] -
                                    mdr$DATUMEVIDENCE[i])
  }
}

helpmd <- md[md$survival!=-1 & md$DATUMCASSU>divisionyearly[9],
             c("event", "survival")]
helpmd <- rbind(helpmd, mdr[mdr$survival!=-1 &
                              mdr$DATUMCASSU>divisionyearly[9],
                            c("event", "survival")])
helpmd$event <- 0
colnames(helpmd) <- c("event", "timeDif")

helpmd <- rbind(helpmd, md[md$DATUMCASSU>divisionyearly[9],
                           c("event", "timeDif")])

## Lognormal distribution
NLLlognormal <- function(par, x){
  m <- par[[1]]
  s <- par[[2]]
  -sum(log(1-plnorm(x$timeDif[x$event==0], meanlog=m, sdlog=s)))-
    sum(log(dlnorm(x$timeDif[x$event>0], meanlog=m, sdlog=s)))
}

fitlognormaltimesmd1 <- optim(par=c(2, 2), fn=NLLlognormal,
                              x=helpmd[helpmd$event!=2, ],
                              hessian=TRUE)
fitlognormaltimesmd2 <- optim(par=c(2, 2), fn=NLLlognormal,
                              x=helpmd[helpmd$event!=1, ],
                              hessian=TRUE)


### Simulation
## Define Generation of time of the next payment and type of the event
varMatrix1 <- solve(fitlognormaltimesmd1$hessian)
# inverse of negative hessian
```

```
varMatrix2 <- solve(fitlognormaltimesmd2$hessian)
# inverse of negative hessian

parmd1 <- fitlognormaltimesmd1$par
parmd2 <- fitlognormaltimesmd2$par
cdf2 <- function(x, help) {
  plnorm(x, meanlog=parmd1[1], sdlog=parmd1[2]) *
    plnorm(x, meanlog=parmd2[1], sdlog=parmd2[2]) - help
}
x <- seq(from=1, to=5000, length=1000)
qfapprox2 <- approxfun(cdf2(x, 0), x, method = "linear")

nextEventTime2 <- function(censoring) {
  randomNumber <- runif(n=1, min=0.000001, max=0.995)
  pnew <- randomNumber * (1 - cdf2(censoring, 0)) + cdf2(censoring, 0)
  if (pnew < 0.000001) {
    pnew <- 0.000001
  }
  else if (pnew > 0.995) {
    pnew <- 0.995
  }
  qfapprox2(pnew)
}

hazardRatemd1 <- function(t) {
  dlnorm(t, meanlog=parmd1[1], sdlog=parmd1[2]) /
    (1 - plnorm(t, meanlog=parmd1[1], sdlog=parmd1[2]))
}
hazardRatemd2 <- function(t) {
  dlnorm(t, meanlog=parmd2[1], sdlog=parmd2[2]) /
    (1 - plnorm(t, meanlog=parmd2[1], sdlog=parmd2[2]))
}

nextEventType <- function(t) {
  p <- hazardRatemd1(t)/(hazardRatemd1(t)+hazardRatemd2(t))
  randomNumber <- runif(n=1, min=0, max=1)
  if (randomNumber <= p) {
    1
  }
  else {
    2
  }
}

## Define generation of payments
varMatrix3 <- fitpmd3$vcov # variance matrix

par3 <- fitpmd3$estimate
nextPayment <- function() {
  rlnorm(n=1, meanlog=par3[1], sdlog=par3[2])
}

## Prepare RBNS claims
c <- integrate(f=intensityIBNRmd, lower=0, upper=tau,
               subdivisions=2000)$value ### Lambda(infinity)

cdf1 <- function(x, help) {
  integrate(f=intensityIBNRmd, lower=0, upper=x,
```

```
            subdivisions=2000)$value / c - help
}

x <- seq(from = tau - 3*365, to = tau, length = 1000)
cdf1Values <- rep(0, times=length(x))
for (i in 1:length(x)) {
  cdf1Values[i] <- cdf1(x[i], 0)
}

qfapprox1 <- approxfun(cdf1Values, x, method = "linear")

IBNR_occurence <- function() {
  randomNumber <- runif(n=1, min=0, max=1)
  if (randomNumber < cdf1Values[1]) {
    randomNumber <- cdf1Values[1]
  }
  qfapprox1(randomNumber)
}

RBNSclaims <- md[md$reserve>0, c("den.platby", "order")]
RBNSclaims$den.platby <- as.double(RBNSclaims$den.platby - date0)
include <- rep(TRUE, times=nrow(RBNSclaims))
for (i in 1:(nrow(RBNSclaims)-1)) {
  if (RBNSclaims$order[i] < RBNSclaims$order[i+1]) {
    include[i] <- FALSE
  }
}
RBNSclaims <- RBNSclaims[include, ]

# Add open claims with zero payments
helpRBNS<-mdr[mdr$include==0 & mdr$reserve>0 & mdr$order==1,
              c("DATUMEVIDENCE", "order", "reserve")]
helpRBNS <- helpRBNS[helpRBNS$DATUMEVIDENCE>divisionyearly[16],
                     c("DATUMEVIDENCE", "order")]
helpRBNS$DATUMEVIDENCE <- as.double(helpRBNS$DATUMEVIDENCE - date0)
colnames(helpRBNS) <- c("den.platby", "order")
RBNSclaims <- rbind(RBNSclaims, helpRBNS)

## Generate number of IBNR claims, times of occurence and delay
set.seed(22121992)
simulations <- 10000
results <- data.frame(numberOfPayments=numeric(),
                      amountsPaid=numeric(),
                      amountsPaidNextYear=numeric())
tau3 <- as.double(as.Date("2016-12-31")-date0)

for(simulation in 1:simulations) {
  parmd1 <- rmvnorm(n=1, mean=fitlognormaltimesmd1$par, varMatrix1)
  parmd2 <- rmvnorm(n=1, mean=fitlognormaltimesmd2$par, varMatrix2)
  par3 <- rmvnorm(n=1, mean=fitpmd3$estimate, varMatrix3)
  results[simulation, ] <- c(0, 0, 0)
  NSclaims <- RBNSclaims
  rowNumber <- nrow(NSclaims) + 1

  n <- rpois(n=1, lambda=c)
  time <- 0
  u <- 0
  for (i in 1:n) {
```

```
    time <- IBNR_occurence()
    randomNumber <- runif(1, min=0, max=1)
    meanlogcorrected <- meanlmd + time*log(bmd)
    pnew <- randomNumber*(1 - plnorm(tau-time,
                                     meanlog=meanlogcorrected,
                                     sdlog=sdlmd)) +
      plnorm(tau-time, meanlog=meanlogcorrected, sdlog=sdlmd)
    u <- qlnorm(pnew, meanlog=meanlogcorrected, sdlog=sdlmd)
    NSclaims[rowNumber, ] <- c(time+u, 0)
    rowNumber <- rowNumber + 1
  }

  payments <- data.frame(time=numeric(), amount=numeric())
  index <- 1
  for (claim in 1:nrow(NSclaims)) {
    newEvent <- 1
    while (newEvent==1) {
      censoring <- max(0, as.double(tau-NSclaims$den.platby[claim]))
      payments[index, ] <- c(0, 0)
      payments$time[index] <- NSclaims$den.platby[claim] +
        nextEventTime2(censoring)
      newEvent <- nextEventType(payments$time[index])
      payments$amount[index] <- nextPayment()

      NSclaims$den.platby[claim] <- payments$time[index]
      index <- index + 1
    }
  }
  results$numberOfPayments[simulation] <- nrow(payments)
  results$amountsPaid[simulation] <- sum(payments$amount)
  results$amountsPaidNextYear[simulation] <- sum(payments$amount[
    payments$time>tau & payments$time<=tau3])
}
}
```

List of Figures

List of Tables

List of Notation

N_t, M_t ... counting processes
$\lambda(t)$, $\mu(t)$ ... intensity functions
$\Lambda(t)$, $M(t)$ ... cumulative intensity functions
$w(t)$ ... an exposure function
T ... a time of occurence of a claim
S ... a time of reporting of a claim
U_t ... a delay in notification of a claim incurred at time t
V ... a waiting time to a claim settlement (since notification)
V_i ... ith payment for a selected claim
Z ... a mark describing settlement process
$\mathcal{Z}$ ... space of possible claims developments
C ... theoretical payment process
$\mathcal{C}$ ... space of possible claims
$\mathbf{C}$ ... used payment process
N ... number of claims
$\tilde{N}$ ... number of payments
P_i ... ith payment $f(t)$ ... a density
$F(t)$ ... a distribution function
$S(t)$ ... a survival function
DF ... degrees of freedom
SD ... standard deviation
τ ... number of days since date zero (till present)