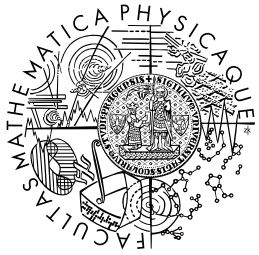

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bc. Daniela Hrbáčová

Truncated marked processes

Department of Probability and Mathematical Statistics

Supervisor of the master thesis: doc. RNDr. Michal Pešta, Ph.D.

Study programme: Mathematics

Study branch: Financial and Actuarial mathematics

Prague 2023

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

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 ..... date ....

Author's signature

I would like to express my sincere thanks to my supervisor doc. RNDr. Michal Pešta, Ph.D., whose guidance, expertise, and motivation were invaluable in completing this thesis. I would also like to extend my gratitude to Czech Insurance Bureau for providing the data and to the creators of the R software and its contributors for enabling me to work with that data. Finally, I am deeply grateful to my boyfriend for his unwavering support and encouragement throughout the writing process.

Title: Truncated marked processes

Author: Bc. Daniela Hrbáčová

Department: Department of Probability and Mathematical Statistics

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

Abstract: This thesis explores the use of marked stochastic processes in the context of delayed reporting of claims in non-life insurance. The focus is on estimating the intensity of the claim occurrence process using the ν-transform of the claim reporting process. The first chapter provides the theoretical background, including the introduction of the Poisson process and the concept of marking. The ν-transform is defined and a special case of the ν-transform is applied in a few examples. In the second chapter the estimation method is described, the goodness-of-fit test for the intensity of the NHPP is proposed and the IBNR claim reserve is discussed. The third chapter applies these theoretical concepts to real-world data from Motor Third Party Liability insurance. The result is a distribution of the IBNR claim reserve prediction for each year. While the approach is computationally intense, it has practical applications in estimating claim reserves for insurance companies. Future work could expand on this approach by considering more complex cases, such as the time-varying conditional distribution of delays or amounts, or including on input a more complex process. Overall, this thesis contributes to the development of claim-by-claim reserving stochastic approaches and provides a well-arranged elaboration of the usage of ν-transform in claims reserving problem with truncated data.

Keywords: truncation, truncated data, stochastic process, marked process, reporting delay, non-life insurance.

Contents

Introduction and Motivation

Non-life insurance is an essential aspect of modern society, providing coverage for individuals and businesses against financial losses resulting from unforeseen events such as accidents, fires, and natural disasters. However, estimating the amount of money that an insurer must reserve to cover future claims is a complex and challenging task. The time between when an event occurs and when a claim is reported and settled can vary widely, making it difficult to estimate the frequency of claims accurately.

One approach to this problem is the claim-by-claim reserving stochastic approach first proposed by Arjas [1989]. This approach involves modeling the occurrence and reporting of claims as stochastic processes and using statistical methods to estimate the intensity of the occurrence process. By estimating the intensity of the claim occurrence process and claim severity, actuaries can better estimate current liabilities of the insurance company to policyholders, allowing them to set aside adequate reserves to cover their obligations.

In recent years, the claim-by-claim reserving stochastic approach has gained popularity among actuaries and insurance companies due to its ability to provide more accurate predictions of future claims. Nevertheless, in practice claim-by-claim approach is used only for modeling extreme claims. Clearly, estimating the intensity of the occurrence process accurately can be challenging, particularly when data is incomplete or truncated.

This thesis focuses on the problem of delayed reporting of claims in non-life insurance and deals with a novel approach published by Maciak et al. [2021], how to estimate the intensity of the occurrence process based on the ν-transform of the claim reporting process. In the context of delayed reporting of claims, marks correspond to the delays between the occurrence of a claim and its reporting. Especially, the ν-transform is a powerful tool for getting the characteristics of the claim occurrence process through the conditional distribution of delays.

The thesis is organized as follows. In the first chapter, we provide the theoretical background on marked stochastic processes, including the Poisson process, marking, and the ν-transform. We also discuss the problem of truncation. In the second chapter we introduce the methods for estimating parameters, formulate the statistical test to verify, whether our model fits the data. In particular, the IBNR claim reserve is expressed mathematically and simulation procedure for predicting the IBNR reserves is mentioned. In the third chapter, we apply the theoretical approach to real-world data from Motor Third Party Liability (MTPL) insurance provided by the Czech Insurers' Bureau (CIB). We demonstrate how the ν-transform can be used to estimate the intensity of the occurrence process in the presence of delayed reportings. Further, we predict the total number of claims based on the estimated intensity and through the distribution of claim amounts we calculate the total claim amount. Finally, the distribution of the IBNR reserve prediction is generated for different years.

Overall, this thesis contributes to the development of the claim-by-claim reserving stochastic approach by providing an unconventional method for estimating the intensity of the claim occurrence process in the presence of truncated data. The proposed approach has practical applications in estimating claim reserves for insurance companies, enabling them to make more reliable estimates of their liabilities and set aside adequate reserves to cover claims that already happened but still are not reported.

1. Marked Processes and Displacement

1.1 Introduction to Claims Processes

There are two parties in insurance policy: the insured and the insurer. The insured pays a deterministic amount of money (called a premium) to the insurer to be protected against the random occurrence of well-specified events. In case such an event happens and the insured reports the claim, the insurer is obliged to pay the amount of money (called claim amount) to cover the damage or injury caused by that event or at least the well-defined amount in case such an event happens.

We distinguish the following moments in time associated with the insured event. Accident time – when the event occurred, it must be during the insured period. Reporting time – when the event was reported to the insurer, the claim is made. Claims closing – when claim payments had been made and no other claim payments are expected associated with this event. However, it can be reopened. Here we think of a simplified situation when the claim amount (sum of all claim payments) is known at the reporting time. Simply, by considering the time of claim closing to be the same as the reporting time.

According to Solvency II, a regulatory framework for insurance and reinsurance companies effective in European Union, insurance companies must estimate the value of their liabilities in order to create technical reserves from the premium paid in the appropriate amount. Here we focus only on incurred but not reported (IBNR) claims reserves. That is the category with the least information we have about it. Neither we know the number of such claims nor the claim amount. Just the risk exposure – the number or volume of policies in force – until this moment could be known.

1.2 Marked Processes and Poisson Process

A Poisson process is the fundamental stochastic process used in various fields of mathematics, statistics, and applied sciences to model and analyze random events occurring over time. It is particularly useful for describing and predicting the behavior of events that occur randomly and independently, such as rare events or events with a low occurrence rate. The Poisson process has wide-ranging applications in areas such as insurance, finance, telecommunications, healthcare, reliability analysis, and queuing theory. In this section, we will explore the key concepts and properties of the Poisson process. We will delve into the mathematical foundations of the Poisson process, including its definition, intensity function, inter-arrival times, and properties of increments. Moreover, we will examine advanced topics such as the nonhomogeneous Poisson processes. However, we are interested in some additional information – not only event times of the process. To each event time is assigned some mark providing valuable context for the data analysis. These marks can represent various characteristics such as

time stamps, categorical labels, numerical measurements, or any other relevant attributes associated with the underlying process. Marked processes find applications in diverse fields, including finance, environmental monitoring, telecommunications, and epidemiology, where the marks enable a deeper exploration of the underlying patterns, dependencies, and relationships. By incorporating the marked information, marked processes offer a powerful framework for capturing and analyzing complex real-world phenomena.

Firstly, let us mention some basic notions, mostly from Tijms [2003], Daley and Vere-Jones [2003], and Jacobsen [2006].

Definition 1. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(S, \mathcal{S})$ a measurable space, and $I \subset \mathbb{R}$. A family of random variables $\{X_t, t \in I\}$ defined on $(\Omega, \mathcal{A}, P)$ with values in S is called a* stochastic process.

I is called the index set of the stochastic process, especially here the $t \in I$ has the meaning of time. S is called the state space of the stochastic process.

Definition 2. *A process $\{X_t, t \in I\}$, where I is an interval, has independent increments, if for every $n \in \{3, 4, \dots\}$ and for any $t_1, t_2, \dots, t_n \in I$ such that $t_1 < t_2 < \cdots < t_n$, the random variables $X_{t_2} - X_{t_1}, \dots, X_{t_n} - X_{t_{n-1}}$ are* independent.

Definition 3 (Simple point process)**.** A simple point process N *is a sequence $\{S_i, i \in \mathbb{N}\}$ of I-valued random variables defined on the probability space $(\Omega, \mathcal{A}, P)$ such that*

- $P[0 < S_1 < S_2 < \cdots] = 1$,
- $P[S_i < S_{i+1}, S_i < \infty] = P[S_i < \infty]$,
- $P[\lim_{i\to\infty} S_i = \infty] = 1$.

S_is are called the points of the simple point process N and take values from the index set of N, for our purpose $I \subset \mathbb{R}$. This thesis deals only with simple point processes, so by a process, we always mean the simple point process.

Definition 4 (Marked point process)**.** *Let us have a point process N with values in S with event times $\{S_i, i \in \mathbb{N}\}$ and a measurable space $(K, \mathcal{K})$, then the point process $\{(S_i, K_i), i \in \mathbb{N}\}$ defined on $S \times K$ is called* marked point process *with state space S and marks in K.*

In our context, the mark could be the claim amount, the delay in reporting the claim, or some information about the insured like the age, the residence, or the car power.

Definition 5 (Independent marks)**.** *The marked point process N has* independent marks *if, given $\{S_i\}$, the $\{K_i\}$ are mutually independent random variables such that the distribution of K_i depends on the corresponding location S_i.*

Definition 6 (Homogeneous Poisson process)**.** *Consider a sequence $T_1, T_2, \dots$ of positive, independent, exponentially distributed random variables with the distribution function $P[T_j \le t] = 1 - e^{-\lambda t}$ for $j = 1, 2, \dots$, positive t and some positive*

constant λ. *Let* $S_0 := 0$ *and* $S_i := \sum_{j=1}^{i} T_j$ *for* $i = 1, 2, \ldots$. Homogeneous Poisson point process with rate λ *is the counting process* $\{N_t, t \geq 0\}$, *where random variable (r. v.)* N_t *is defined as the largest integer* $i \geq 0$ *for which holds that* $S_i \leq t$.

Generally, a counting process is a stochastic process $\{N_t, t \geq 0\}$ with values that are non-negative, integer, and non-decreasing. In Definition 6, think of T_n as a time elapsed between $(n-1)$th and nth occurrence of some specific event, S_n as the time at which the nth event occurs (arrival time of n-th event). Finally, N_t is the number of events up to time t.

Note that there is 1–1 relation between $\{S_i, i \in \mathbb{N}\}$ and $\{N_t, t \geq 0\}$. For homogeneous Poisson process (HPP) is characteristic the memoryless property formulated in Theorem 1.

Theorem 1. *For any* $t \geq 0$, *the r. v.* D_t *representing the waiting time from epoch* t *until the next event has the same exponential distribution as the time elapsed between two consecutive events.*

Proof. Can be found in Tijms [2003] [p. 4].

□

Let us clarify that by the stationarity of increments we mean, that the distribution of the number of events that occur in any time interval depends only on the length of the time interval.

Property 2 (Properties of HPP)**.** *Poisson process* N_t *with constant rate* $\lambda > 0$

- *has* stationary increments,
- *has* independent increments,
- *the number of events in any interval of length* t *is a* Poisson r. v. *with parameter* λt *(interpreted as the mean).*

Obviously, the third property gave rise to the name of the process. Nevertheless, some real-world phenomena do not have the same expected value at every time. There exists a generalization of the Poisson process in this way. We introduce the rate of the Poisson process as a function of time, it must be an integrable function on a bounded interval and the process does not have stationary increments anymore.

Definition 7 (Nonhomogeneous Poisson process)**.** *A counting process* $\{N_t, t \geq 0\}$ *is said to be a nonhomogeneous Poisson process with the intensity function* $\lambda(t)$, *if it satisfies the following properties:*

- $N_0 = 0$,
- *the process* $\{N_t\}$ *has independent increments,*
- $P\left[N_{t+\Delta t} - N_t = m\right] = \begin{cases} 1 - \lambda(t)\,\Delta t + o(\Delta t), & m = 0 \\ \lambda(t)\,\Delta t + o(\Delta t), & m = 1 \\ o(\Delta t), & m \geq 2 \end{cases}$ *as* $\Delta t \to 0$,

where $o(\Delta t)$ *is an arbitrary function such that* $\lim_{\Delta t \to 0_+} \frac{o(\Delta t)}{\Delta t} = 0$ *and* $\lambda(t)$ *is positive locally integrable function.*

For nonhomogeneous Poisson process (NHPP) it holds that the number of events in any time interval is a Poisson random variable; however, its parameter can depend on the location of the interval. More specifically, we can write

$$N_{s+t} - N_s \sim Poisson \left\{ \int_s^{s+t} \lambda(r)\, \mathrm{d}r \right\}.$$

Property 3. *Let* $\{N_t, t \geq 0\}$ *be a nonhomogeneous Poisson process with the specific form of the intensity function* $\lambda(t) = e^{a+bt}$ *for* $a, b \in \mathbb{R}, b \neq 0$*. Given* $N_T = n$*, the times of occurrence of events* $S_1, \ldots, S_n$ *have the same distribution as an ordered sample of* n *observations from the distribution with density function* $\left(e^{bx} \cdot b\right) / \left(e^{bT} - 1\right)$ *on* $[0, T]$*.*

For nonhomogeneous Poisson process observed up to time T with event times $0 < s_1 < \cdots < s_n < T$ we can write the joint density in the form

$$\exp\left\{ - \int_0^T \lambda(t)\, \mathrm{d}t \right\} \lambda(s_1) \times \cdots \times \lambda(s_n),$$

with the above-specified intensity function, it can be rewritten as

$$\exp\left\{ - \frac{e^a \left(e^{bT} - 1\right)}{b} + na + b \sum_{i=1}^{n} s_i \right\}.$$

Just the probability of observing exactly n events of NHPP with above-specified joint density on the time interval $[0, T]$ can be written as

$$\frac{1}{n!} \left(\frac{e^a (e^{bT} - 1)}{b} \right)^n \exp\left\{ - \frac{e^a \left(e^{bT} - 1\right)}{b} \right\}.$$

Finally, the joint conditional density of observed event times given the number of events up to time T can be expressed as

$$n! \prod_{i=1}^{n} \frac{b e^{bs_i}}{e^{bT} - 1},$$

which is the joint density of the order statistic with the marginal density $\left(be^{bx}\right) / \left(e^{bT} - 1\right)$, for $x > 0$.

1.3 Displacement

In what follows, based on Kallenberg [2021], Kingman [1993] and Maciak et al. [2021], we use the ν-transform of the process M to express the intensity of the process N through some assumed conditional distribution of their arrival times.

Let us have a point process M with arrival times $\{Z_i\}$ choosing conditionally independent random elements τ_i with conditional distributions $\nu(\cdot; t)$ given $Z_i = t$, the point process N with arrival times $\{\tau_i\}$ is called *ν-transform* of the process M. The ν-transform randomly displaces the arrival times of a point process and the displaced arrival times form another point process. Especially, in the case of the Poisson process, the Poisson property is preserved as it is stated in the following theorem.

Theorem 4 (Displacement theorem)**.** *Let us have a nonhomogeneous Poisson process with rate function $\lambda_{orig}(t)$. Suppose that the points of the process are randomly displaced, such that the displacements of different points are independent and the distribution of the displaced position given $Z_i = t$ has a density $f(\cdot; t)$. Then the displaced points form again a nonhomogeneous Poisson process with an intensity function*

$$\lambda_{transf}(y) = \int_{\mathbb{R}} \lambda_{orig}(t)\, f(y; t)\, \mathrm{d}t.$$

In particular, if $\lambda_{orig}(t)$ is a constant c, and if $f(y; t)$ is a function of difference $y - t$, then $\lambda_{transf}(y) = c$.

Proof. See Kingman [1993] [p. 61].

□

Having a NHPP M with the parametric intensity $\lambda_{orig}(t; \boldsymbol{\varrho})$ we can express from Theorem 4 the intensity of the claim occurrence process $N_{\cdot}$ as

$$\lambda_{transf}\big(y; \boldsymbol{\varrho}, \boldsymbol{\vartheta}\big) = \int_{\mathbb{R}} \lambda_{orig}(t; \boldsymbol{\varrho})\, f_{S_i|Z_i}(y; t, \boldsymbol{\vartheta})\, \mathrm{d}t, \tag{1.1}$$

where $f_{S_i|Z_i}$ denotes the parametric density of S_i given Z_i. Nevertheless, the integral on the right-hand side of (1.1) can be expressed analytically only for some trivial functions.

Example 1. The easiest case is when we have a homogeneous Poisson process on the real line with the constant intensity $\lambda_{orig}(t) = c$ for $t \in \mathbb{R}$ and the conditional density function is a function of difference $f(y; t) = g(y - t)$ for $y, t \in \mathbb{R}$. Then Theorem 4 says that the transformed process is again a homogeneous Poisson process on the real line with the following intensity

$$\begin{aligned}\lambda_{transf}(y) &= \int_{\mathbb{R}} \lambda_{orig}(t) f(y; t)\, \mathrm{d}t = c \int_{\mathbb{R}} g(y - t)\, \mathrm{d}t \\ &= c \int_{\mathbb{R}} g(v)\, \mathrm{d}v = c \int_{\mathbb{R}} f(v; 0)\, \mathrm{d}v = c.\end{aligned}$$

Example 2. Let us have a homogenenous Poisson process on the interval $[0, T]$ with the intensity $\lambda_{orig}(t) = c$ for $t \in [0, T]$. We are interested in the intensity of a transformed process considering the independent displacement of different points with the conditional density function of the form $f(y; t) = \exp\{-y + 2t\}$ for $t \in [0, T]$ and $y \in (2t, \infty)$. We arrive to

$$\lambda_{transf}(y) = c \int_{y/2}^{T} e^{-y+2t} \, \mathrm{d}t = \frac{c}{2} e^{-y} \left(e^{2T} - e^{y}\right)$$

The transformed process is a nonhomogeneous Poisson on $[0, 2T]$.

Example 3. Having a nonhomogeneous Poisson process with the intensity $\lambda_{orig}(t) = \exp\{-t\}$ defined on $[0, T]$ and the previously mentioned transformation we can calculate the intensity of the transformed process as follows:

$$\lambda_{transf}(y) = \int_{y/2}^{T} e^{-t} \, e^{-y+2t} \, \mathrm{d}t = e^{-y} \left(e^{T} - e^{y/2}\right).$$

The ν-transform can be used also for more complex processes, not necessarily with independent increments, as it is discussed in Maciak et al. [2021].

1.3.1 Car Example

This is a more detailed version of the example from Kingman [1993] [p. 59].

Consider a broad highway with randomly placed cars at time $t = 0$ represented by the real line and points $\{X_i, i \in \mathbb{N}\}$, where each X_i specifies the location of the i-th car. Suppose that cars form a nonhomogeneous Poisson process at time 0 with the intensity $\lambda_{orig}(x)$ and move with constant velocity independently on each other with unrestricted overtaking. Taking snapshots in time we focus on the positions of cars. Denote by $Y^t(X)$ the position at time t of a car that was at the beginning (time 0) at X. It can be rewritten as

$$Y^t(X) = X + Vt.$$

From Theorem 4 follows that cars form a nonhomogeneous Poisson process at every time $t > 0$ with the following intensity

$$\lambda^t_{transf}(y) = \int_{\mathbb{R}} \lambda_{orig}(x) \, f_t(y; x) \, \mathrm{d}x, \tag{1.2}$$

where $f_t(\cdot; X)$ denotes conditional density of the position at time t given the initial position X (at time 0) on the road. The lower index t denotes the time, at which the snapshot is taken.

If the velocities are drawn independently from a probability distribution with density $g(v)$, then it holds that

$$\mathsf{P}\left[Y^t \leq y | X = x\right] = \mathsf{P}\left[x + Vt \leq y | X = x\right] = \mathsf{P}\left[V \leq \frac{y - x}{t}\right]$$

using the independence of V on X in the last step. So the conditional density f_t can be expressed as the scaled density of the velocity distribution

$$f_t(y;x) = \frac{g\left(\frac{y-x}{t}\right)}{t},$$

for $x < y$, otherwise it is zero assuming only positive velocities. Finally, using the substitution rule, equation (1.2) becomes

$$\lambda^t_{transf}(y) = \int_0^{y/t} \lambda_{orig}(y - vt)\, g(v)\, \mathrm{d}v.$$

Note that we can calculate the intensity λ^t_{transf} for as many times t as we want, having just the information that the initial positions form a nonhomogeneous Poisson process with the intensity function λ_{orig} and knowing the density of the velocity distribution.

1.3.2 Truncation

In data analysis, the presence of truncated data is a common challenge that arises when observations are limited to a certain range. Truncated data can occur in various fields, such as finance, economics, or epidemiology, where data collection may be subject to constraints or limitations. Truncation can impact the statistical analysis and requires specialized techniques to handle the inherent biases introduced by the truncation as writes Greene [2002].

Here, the claim occurrence process $\{N_t, t \geq 0\}$ is truncated. Since the insurance company knows about the claims through reporting, where the reporting delay causes that we do not know about some already occurred accidents. Nevertheless, in our case, we handle it by considering a claim reporting process M that is complete – we observe all reporting times $\{Z_i, i = 1, \ldots, M_T\}$ up to time T, and by the displacement, we arrive to the intensity of the claim occurrence process N.

2. Estimation, Goodness-of-fit, and IBNR Prediction

2.1 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is a widely used statistical method for estimating the parameters of a probability distribution based on the observed data. It is a powerful approach that seeks to find parameter values that maximize the joint probability of obtaining the observed data.

Firstly, we define the likelihood function

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}; y_1, \ldots, y_m) = p(y_1, \ldots, y_m; \boldsymbol{\theta}),$$

where p is a joint probability of the observed data viewed as a function of an unknown vector of parameters $\boldsymbol{\theta}$ of a model. The goal of MLE is to find values of model parameters that maximize the likelihood over the parameter space $\Theta \subset \mathbb{R}^d$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}).$$

If we assume independence of observations, the joint probability can be expressed in the form of a product and we can simplify our maximization problem by applying the logarithm on the likelihood function. This is how the log-likelihood function is defined

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}).$$

Since the logarithm is a strictly increasing function, the likelihood and the log-likelihood function have the maximum at the same point.

2.1.1 Joint Likelihood of Marked Process

In the reporting process, we are interested in triplets $\{Z_i, W_i, Y_i, i \in \mathbb{N}\}$, where Z_is are arrival times of the reporting process M and W_i, Y_i are marks representing the $i-$th claim's delay, and amount respectively.

Considering independent marks W_i and Y_i independent of both Z_i and W_i, the joint likelihood of observing triplets $(z_i, w_i, y_i, i = 1, \ldots, m)$ up to time T can be expressed as

$$\begin{aligned} L(\boldsymbol{\varrho}, \boldsymbol{\vartheta}, \boldsymbol{\gamma}) = & L_M\left(\boldsymbol{\varrho}; T, z_i, i = 1, \ldots, m\right) \\ & \times L_{W|Z}\left(\boldsymbol{\vartheta}, z_i; w_i, i = 1, \ldots, m, \right) \\ & \times L_Y\left(\boldsymbol{\gamma}; y_i, i = 1, \ldots, m\right), \end{aligned} \tag{2.1}$$

for $0 < z_1 < \cdots < z_m < T, 0 < w_i, 0 < y_i, i = 1, \ldots, m$. As all parts of the product on the right-hand side of (2.1) depend on different parameters, we can maximize them separately.

2.1.2 Intensity of NHPP

To estimate the intensity $\lambda_{orig}(t; \boldsymbol{\varrho})$ from the data we use the maximum likelihood (ML) estimate of the vector of parameters $\boldsymbol{\varrho}$. The likelihood function of the process M, when the last observable time is T and $M_T = m$, is of the form

$$L_M(\boldsymbol{\varrho}) = \exp\left\{ -\int_0^T \lambda_{orig}(t; \boldsymbol{\varrho})\, \mathrm{d}t \right\} \times \prod_{i=1}^{m} \lambda_{orig}(z_i; \boldsymbol{\varrho}), \tag{2.2}$$

for $0 < z_1 < \cdots < z_m < T$.

Now, consider a special case of a nonhomogeneous Poisson process with the intensity of the form

$$\lambda_{orig}(t; \boldsymbol{\varrho}) = e^{\varrho_1 + \varrho_2 t},$$

where $\boldsymbol{\varrho} = (\varrho_1, \varrho_2)^\top$. The logarithmic likelihood function can be expressed as

$$\ell_M(\varrho_1, \varrho_2; T, z_i, i = 1, \ldots, m) = -\int_0^T e^{\varrho_1 + \varrho_2 t}\, \mathrm{d}t + \sum_{i=1}^{m} (\varrho_1 + \varrho_2 z_i). \tag{2.3}$$

Let us assume that both parameters ϱ_1, ϱ_2 are different from zero, by differentiating of (2.3) with respect to each parameter we obtain

$$\begin{aligned}
\frac{\partial l_M}{\partial \varrho_1} &= \exp\{\varrho_1\} \frac{\exp\{\varrho_2 T\} - 1}{\varrho_2} - m, \\
\frac{\partial l_M}{\partial \varrho_2} &= \exp\{\varrho_1\} \frac{T\varrho_2 \exp\{\varrho_2 T\} - (\exp\{\varrho_2 T\} - 1)}{\varrho_2^2} - \sum_{i=1}^{m} z_i,
\end{aligned}$$

by setting both expressions equal to zero we get the system of two equations with two unknown parameters, which can be solved numerically and the solution we get is the MLE of ϱ_1, ϱ_2 denoted as $\hat{\varrho}_1, \hat{\varrho}_2$, respectively.

2.1.3 Distribution of Delays

Many parametric distributions are suitable for modeling delays $W_i = Z_i - S_i$. They need to have positive support, so modifications of log-normal, Gamma, or Weibull distributions are used in practice. Here we focus on the log-normal distribution with density

$$f(w; \mu, \sigma) = \frac{1}{\sigma w \sqrt{2\pi}} \exp\left\{ -\frac{\big(\log(w) - \mu\big)^2}{2\sigma^2} \right\}, \tag{2.4}$$

for $w > 0$ and zero otherwise. Additionally, time dependence or seasonality can be added by thinking of μ and σ in (2.4) as some parametric time-dependent functions, for example

$$\begin{aligned}
\mu(t, \boldsymbol{\varphi}) &= \varphi_1 + \varphi_2 t + \varphi_3 \cos\left(\frac{\varphi_4 \times 2\pi \times t}{365} \right) + \varphi_5 \sin\left(\frac{\varphi_6 \times 2\pi \times t}{365} \right), \\
\sigma(t, \boldsymbol{\varepsilon}) &= \varepsilon_1 + \varepsilon_2 t,
\end{aligned}$$

where the parameter μ of the previously mentioned distribution depends linearly on the time and takes one seasonality pattern into account and the parameter σ just linearly depends on time.

The likelihood can be expressed as

$$L_{W|Z}\left(\boldsymbol{\vartheta}, z_i; w_i, i = 1, \ldots, m\right) = \prod_{i=1}^{m} f_{W_i|Z_i}\left(w_i; z_i, \boldsymbol{\vartheta}\right).$$

Assuming the easiest case that the delay does not change in time, the conditional density of S_i given $Z_i = t$ has the form

$$f_{S_i|Z_i}(s; t, \vartheta_1, \vartheta_2) = g(t - s; \vartheta_1, \vartheta_2) = \frac{1}{\vartheta_2(t-s)\sqrt{2\pi}} \exp\left\{-\frac{\left(\log\left(t-s\right) - \vartheta_1\right)^2}{2\vartheta_2^2}\right\},$$

for $t \in (0, \infty), s \in (-\infty, t)$ and zero otherwise. Again, the ML estimates we get by solving the problem of maximizing the logarithmic likelihood function.

Simply by plugging in (1.1) we obtain the expression of the intensity of the ν-transformed process

$$\lambda_{transf}\left(s; \boldsymbol{\varrho}, \boldsymbol{\vartheta}\right) = \int_{s}^{\infty} \lambda_{orig}\left(t; \boldsymbol{\varrho}\right) \times g\left(t - s; \boldsymbol{\vartheta}\right) \mathrm{d}t, \tag{2.5}$$

for $t \in (0, \infty), s \in (-\infty, t)$, where $\lambda_{transf}, \lambda_{orig}$ denote the intensity of the claim occurrence process, and reporting process, respectively. Note that the displacement is backward.

Remark. If we assume Y_i to be independent and identically distributed (with distribution depending on the vector of parameters $\boldsymbol{\gamma}$), and independent of reporting times and delays it can be rewritten the likelihood as

$$L_Y\left(\boldsymbol{\gamma}; y_i, i = 1, \ldots, m\right) = \prod_{i=1}^{m} f_Y\left(y_i; \boldsymbol{\gamma}\right).$$

2.2 Goodness-of-fit

The goodness-of-fit test is a fundamental statistical tool used to assess the adequacy of a proposed model in describing observed data. In the context of non-homogeneous Poisson processes, the goodness-of-fit test plays a key role in evaluating the appropriateness of a specified intensity function. The intensity function characterizes the temporal variation in event occurrence rates within the process. By comparing the observed data with the expected events based on the proposed intensity function, the goodness-of-fit test enables us to determine whether the observed data follows the assumed model. This test helps to identify deviations between the observed and expected event patterns. This test is based on Kulich [2017].

Having the observed event data, denoted as $\{z_i, i = 1, \ldots, m\}$, where z_i represents the reporting time of the i-th claim and m denotes the observed number

of events in the time window $[0, T]$. We would like to know, whether indeed the observed data follows NHPP with the intensity function

$$\lambda_0(t) = \exp\left\{\varrho_1^0 + \varrho_2^0 t + \varrho_3^0 t^2 + \varrho_4^0 \cos\left(\frac{2\pi t}{365}\right) + \varrho_5^0 \sin\left(\frac{2\pi t}{365}\right)\right\},$$

for $\varrho_1^0, \varrho_2^0, \varrho_3^0, \varrho_4^0, \varrho_5^0$ real unknown true values of parameters. This is a quite difficult task. Nevertheless, we can test, whether observed counts correspond to expected counts based on the proposed intensity for time intervals of the length l for $l << T$.

Let us have a model $\mathcal{M}_0$: the vector $\boldsymbol{O} = (O_1, \ldots, O_J)^\top$ has the multinomial distribution with parameters $m \in \mathbb{N}, J \in \mathbb{N}$ representing the number of observations, the number of intervals, respectively, and

$$\boldsymbol{p}(\boldsymbol{\varrho}) = \left(\frac{1}{m}\int_{I_1} \lambda(t; \boldsymbol{\varrho})\,\mathrm{d}t, \ldots, \frac{1}{m}\int_{I_J} \lambda(t; \boldsymbol{\varrho})\,\mathrm{d}t\right)^\top$$

the vector of probabilities of falling into the j-th interval for $j \in \{1, \ldots, J\}$ and $\boldsymbol{\varrho} \in \mathbb{R}^d$, where $\int_0^T \lambda(t, \boldsymbol{\varrho})\,\mathrm{d}t = m$ and $d < J$. In our concrete task $d = 5$.

We want to test the null hypothesis (H_0) against the alternative (H_1):

$$H_0 : \exists \boldsymbol{\varrho} \in \mathbb{R}^d : \boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\varrho}),$$
$$H_1 : \forall \boldsymbol{\varrho} \in \mathbb{R}^d : \boldsymbol{p} \neq \boldsymbol{p}(\boldsymbol{\varrho}).$$

Firstly, we must divide the observation period into a set of $J = \lceil T/l \rceil$ non-overlapping time intervals, where $\lceil r \rceil$ denotes the upper whole part of the number r. By the j-th interval we mean

$$I_j = \big[(j-1) \times l, j \times l\big) \text{ for } j \in 1, \ldots, J-1 \text{ and } I_J = \big[J - 1 \times l, T\big).$$

Count the number of observed events that fall within each time interval.

$$O_j = \sum_{i=1}^{m} \mathbb{I}\{z_i \in I_j\} \text{ for } j \in 1, \ldots, J.$$

Then calculate the expected number of events in each time interval based on the proposed intensity function. This can be done by integrating the ML estimate of the intensity function, obtained analogously as in Section 2.1.2, over each interval.

$$E_j = \int_{I_j} \hat{\lambda}(t, \hat{\boldsymbol{\varrho}})\,\mathrm{d}t \text{ for } j \in 1, \ldots, J,$$

where $\hat{\boldsymbol{\varrho}}$ denotes the ML estimate of $\boldsymbol{\varrho}$ and E_j the expected number of claims reported in the j-th interval based on the estimated intensity.

To perform the goodness-of-fit test, we need to formulate the test statistic to assess the discrepancy between the observed and the expected number of events in different time intervals. Calculating the chi-square statistic using the following formula:

$$\chi^2 = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j}.$$

The resulting chi-square test statistic, under the null hypothesis, has an asymptotic chi-square distribution with $J - d - 1$ degrees of freedom. We can use this distribution to calculate the p-value associated with the observed test statistic.

$$p\text{-value} = 1 - F_{\chi^2_{J-d-1}}(\chi^2),$$

where $F_{\chi^2_{J-d-1}}$ denotes the distribution function of chi-square distribution with $J - d - 1$ degrees of freedom and χ^2 denotes the calculated value of the test statistic.

Finally, based on the calculated p-value, we can make a decision regarding the goodness-of-fit. If the p-value is below a predetermined significance level (e.g., 0.05), we reject the null hypothesis of the intensity function being a good fit for our data. Otherwise, we cannot reject the null hypothesis.

2.3 IBNR Reserve Prediction

The IBNR claim reserve for the period $[s, t)$ can be expressed as

$$\text{IBNR}_{[s,t)} = \sum_{i=1}^{N_t} Y_i \mathbb{I}\{s < S_i \leq t\} \mathbb{I}\{Z_i > t\},$$

where

$$\mathbb{I}\{X < t\} = \begin{cases} 1, & X < t, \\ 0, & X \geq t. \end{cases}$$

If the reserves are calculated on a yearly basis, s denotes the beginning of the year for which is the reserve calculated, and t denotes the beginning of the following year. Since we do not take into account the RBNS reserve, the following holds:

$$\text{IBNR}_{[s,t)} = \sum_{i=1}^{N_{[s,t)}} Y_i - \sum_{i=1}^{N'_{[s,t)}} Y_i, \tag{2.6}$$

where $N'_{[s,t)}$ denotes the number of reported claims up to time T that occurred in the time period $[s, t)$ and $N_{[s,t)}$ denotes ultimate number of accidents that occurred in $[s, t)$.

This is one of the possible applications of constructed theoretical approach. Firstly, we estimate parameters $\boldsymbol{\varrho}, \boldsymbol{\vartheta}$, and $\boldsymbol{\gamma}$ for specified parametric distributions through MLE approach. In our specific case

$$\begin{aligned} \boldsymbol{\varrho} &= (\varrho_1, \varrho_2, \varrho_3, \varrho_4, \varrho_5)^\top, \\ \boldsymbol{\vartheta} &= (\vartheta_1, \vartheta_2)^\top, \\ \boldsymbol{\gamma} &= (\gamma_1, \gamma_2)^\top, \end{aligned}$$

assuming the intensity of the reporting process of the exponential form depending on five parameters, log-normally distributed delays, and log-normally distributed claim amounts.

Then we plug in our estimated values of parameters to theoretically expressed results and generate many realizations of the marked occurrence process with

estimated intensity $\hat{\lambda}_{transf}(s; \hat{\boldsymbol{\varrho}}, \hat{\boldsymbol{\vartheta}})$ and claim amounts as marks. "This is called a Monte Carlo approximation, named after a city in Europe known for its plush gambling casinos. Monte Carlo techniques were first developed in the area of statistical physics — in particular, during development of the atomic bomb — but are now widely used in statistics and machine learning as well." (Murphy [2012], p. 52)

Consider K Monte Carlo samples. For every $k = 1, 2, \ldots, K$ we generate a prediction of the number of events

$$N^k_{[s,t)} \sim Po\left(\int_s^t \lambda_{transf}(a)\,\mathrm{d}a\right),$$

and particular claim amounts

$$Y^k_i \sim LN(\gamma_1, \gamma_2) \text{ for } i \in \left\{1, \ldots, N^k_{[s,t)}\right\}.$$

Finally, we calculate $\mathrm{IBNR}^k_{[s,t)}$ as stated in (2.6). The distribution through k provides us the empirical distribution for the prediction of the reserve.

3. Real Data Analysis – IBNR Reserves

Insurance companies face significant challenges in accurately estimating the expected claims that they will receive in the future. These estimates are crucial for calculating appropriate reserves, insurance premiums, managing risk, and ensuring financial stability. Traditional methods for estimating claims amounts and frequencies, such as chain-ladder and Bornhuetter-Ferguson, assume homogeneity in the underlying claims process, which may not always be appropriate in practice. In recent years, there has been growing interest in using stochastic models that allow for more flexible and realistic assumptions about the underlying claims process.

One such approach is stochastic micro claims reserving, which involves dealing with times of claims occurrence as arrival times of some stochastic process. This method takes into account the individual characteristics of each claim and uses statistical techniques to model the underlying process. In particular, the ν-transform of reporting dates can be used to estimate the intensity of the claim occurrence process from the claim reporting process, while considering that we observe the truncated distribution of delays.

The advantage of stochastic micro claims reserving is that it allows for more realistic assumptions about the underlying claims process, which can lead to more accurate estimates of expected claims amounts and frequencies. However, it is important to note that this approach is more computationally intensive and requires more data than traditional methods. As a result, simple methods are still widely used in practice.

In this practical part of my diploma thesis, I will estimate the intensity of the claims reporting process, the distribution of delays and the distribution of claim amounts. Then, by applying the ν-transform I will obtain the estimate of the intensity of the claims occurrence process. I will generate claim amounts for each simulated total number of claims and obtain the IBNR claim reserve prediction by subtracting the observed total claim amount from the corresponding predicted total claim amount. By doing so, I aim to demonstrate the potential benefits of this approach.

3.1 Problem Formulation

Our aim is to estimate the intensity of claims occurrence process N, further denoted as λ_{transf}. We have m triplets of observations

$$(Z_i = z_i, W_i = w_i, Y_i = y_i, i = 1, \ldots, m)\,,$$

where the observed number of events between time 0 and T, M_T, is equal to m, z_i, w_i, y_i denotes the i-th claim's observed reporting time, delay in reporting, and amount, respectively. Reporting times form the claim reporting process M with the intensity λ_{orig}.

3.2 Data Description

Czech Insurers' Bureau collects data about claims from the MTPL insurance that are caused by unidentified or uninsured drivers. We have observations from the beginning of the year 2017 to the end of the year 2019 that contain information about 6 212 material claims characterized by

- claim *ID* (up to 6 digit number),
- claim *amount* paid (in CZK),
- *accident* time (integer part = number of days from January 1, 1900),
- *reporting* time (integer part = number of days from January 1, 1900).

Nevertheless, we are going to use only the data from the first two years for modeling purposes and the information from the last year 2019 serves for comparison to our prediction. From the beginning of the year 2017 till the end of the year 2018, there were 4 237 claims reported. It should be noted that times are provided as decimal numbers, so we can model them as continuous variables.

Amount		Accident		Reporting		Delay	
Min.	278	Min.	-1 217	Min.	0	Min.	0.111
1st Qu.	14 361	1st Qu.	173	1st Qu.	216	1st Qu.	3.160
Median	30 030	Median	350	Median	395	Median	9.876
Mean	50 384	Mean	344	Mean	387	Mean	42.679
3rd Qu.	60 825	3rd Qu.	533	3rd Qu.	571	3rd Qu.	40.001
Max.	2 078 020	Max.	725	Max.	728	Max.	1 404.465

Table 3.1: Summary of the data set (*amount, accident, reporting* rounded to whole numbers and *delay* rounded to three decimal places).

From Table 3.1 we can see that the first claim was reported on January 1, 2017, and the last on December 30, 2018. Looking at accident dates we can see that in the claim occurrence process, the first event occurred on September 2, 2013, and the last was on December 27, 2018. When it comes to the claim amount paid, it is in the range from 278 to more than 2 million CZK. Data contain values of the delay from a few hours up to almost four years.

In Figure 3.1 we can see the trajectory of the claims reporting process.

3.3 Estimating Parameters

In this section we apply the knowledge derived in the previous chapter to get the desired result: an estimate of the intensity of the claim occurrence process. Throughout this section, we assume that

- reporting process forms a nonhomogeneous Poisson process,
- distribution of delays does not change in time.

All data were processed and most calculations were performed in R Core Team [2021] software.

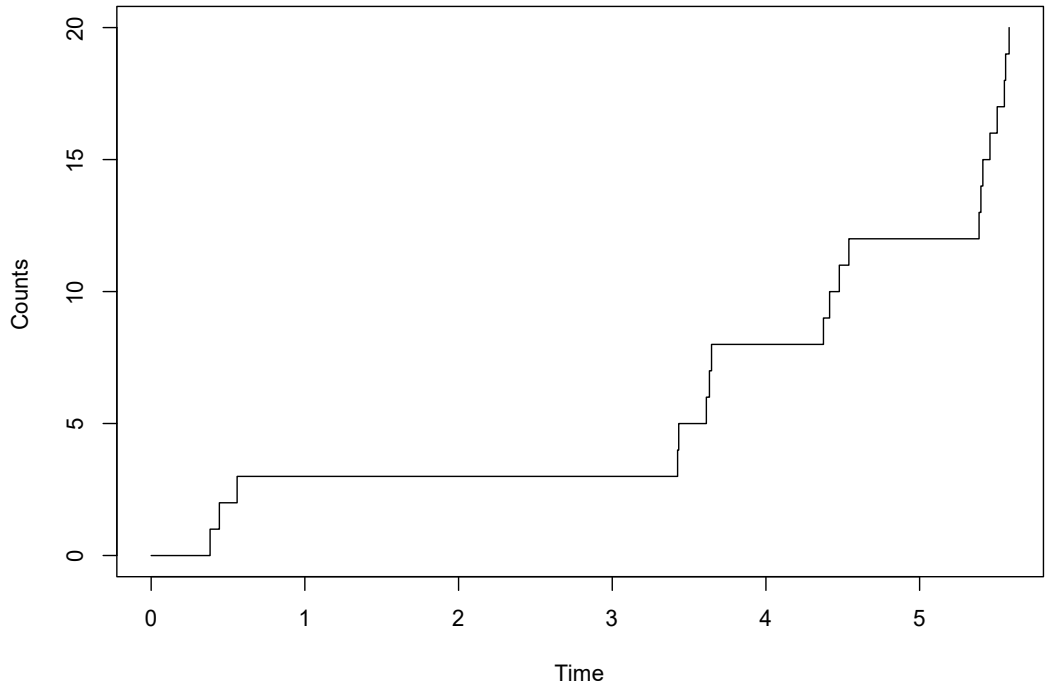

Figure 3.1: Reporting process trajectory, first 20 occurrences (0 denotes January 1, 2017).

3.3.1 Intensity of Reporting Process

Specifically, we think of the intensity λ_{orig} of the form

$$\lambda_{orig}\left(t;\boldsymbol{\varrho}\right)=\exp\left\{\varrho_1+\varrho_2 t+\varrho_3 t^2+\varrho_4\cos\left(\frac{2\pi t}{365}\right)+\varrho_5\sin\left(\frac{2\pi t}{365}\right)\right\},$$

for $t>0$.

We consider having data from the ongoing operations of the insurance company, not from its initial establishment or start of business. Specifically, we take into account all claims that were reported in years 2017-2018, possibly with the accident dates before 2017. The intensity of the NHPP was estimated in R Core Team [2021] package *NHPoisson* by maximization of (2.2). Estimated parameters are stated below.

$$\hat{\boldsymbol{\varrho}}\doteq\begin{pmatrix}1.494\\1.189\times10^{-3}\\-9.978\times10^{-7}\\1.899\times10^{-3}\\-5.630\times10^{-2}\end{pmatrix}. \tag{3.1}$$

Performing the Goodness-of-fit test mentioned in detail in Section 2.2 for 30-day-long intervals gives us the value of the test statistic χ^2 approximately equal to 28.640. The p-value is about 0.072, so on the 5% significance level we cannot reject the null hypothesis that the data are generated by NHPP with the proposed form of the intensity.

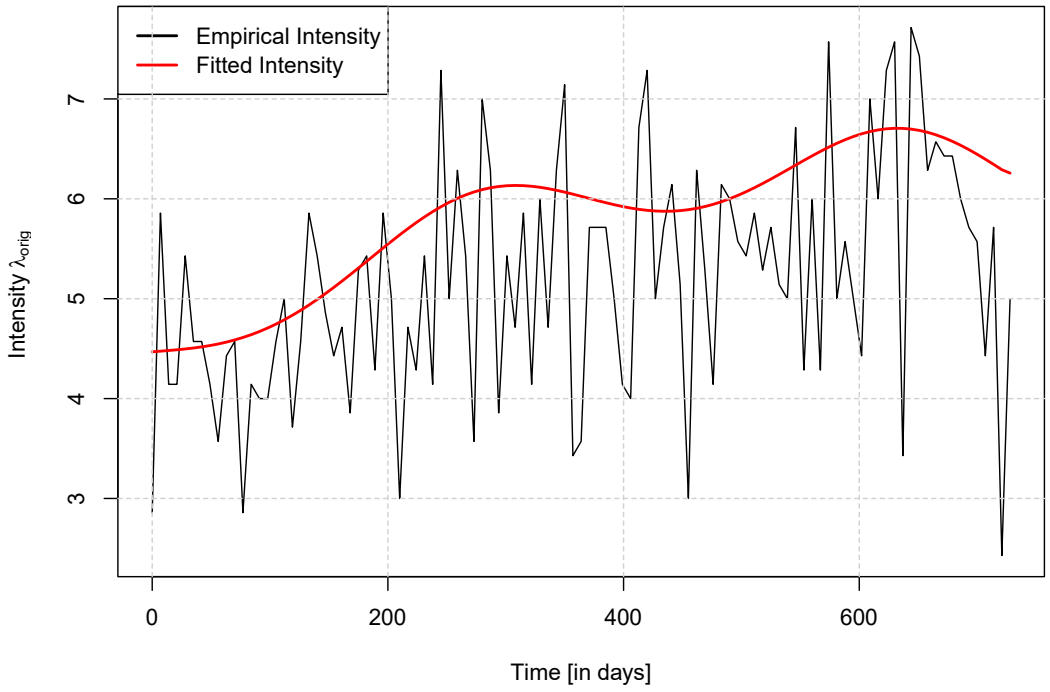

Figure 3.2: Empirical intensity calculated on a weekly basis and fitted parametric intensity of the claims reporting process (0 is 1. 1. 2017).

In Figure **??** we can see the comparison of the empirical and the estimated intensity of the reporting process.

3.3.2 Distribution of Delays

In this section, we focus on the delays W_i – the time between the reporting Z_i and the occurrence S_i of an accident. We assume that delays have a log-normal distribution. Parameters of the distribution can be estimated from the observed data by maximizing the following logarithmic likelihood function

$$\begin{aligned} \log\left(L_{W|Z}(\boldsymbol{\vartheta}, z_i; w_i, i = 1, \ldots, m)\right) &= \ell_W(\boldsymbol{\vartheta}; w_i, i = 1, \ldots, m) \\ &= \sum_{i=1}^{m} \log\left[\frac{1}{\vartheta_2 w_i \sqrt{2\pi}} \exp\left\{-\frac{\left(\log(w_i) - \vartheta_1\right)^2}{2\vartheta_2^2}\right\}\right] \end{aligned} \tag{3.2}$$

where $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)^\top$ in this case. Maximizing expression (3.2) with respect to $\boldsymbol{\vartheta}$ we obtain MLE of $\boldsymbol{\vartheta}$

$$\hat{\boldsymbol{\vartheta}} \doteq (2.427, 1.664)^\top . \tag{3.3}$$

3.3.3 ν-transform

Assuming that reporting times form a nonhomogeneous Poisson process, by Theorem 4 we get that the occurrence process is a nonhomogeneous Poisson process with the intensity of the form (1.1).

Especially, the estimate of the intensity of the occurrence process we get by plug-in approach from formula (2.5), it can be expressed as

$$\hat{\lambda}_{transf}(s) = \int_s^{\infty} \exp\left\{\hat{\varrho}_1 + \hat{\varrho}_2 t + \hat{\varrho}_3 t^2 + \hat{\varrho}_4 \cos\left(\frac{2\pi t}{365}\right) + \hat{\varrho}_5 \sin\left(\frac{2\pi t}{365}\right)\right\} \times \frac{1}{\hat{\vartheta}_2(t-s)\sqrt{2\pi}} \exp\left\{-\frac{\left(\log(t-s) - \hat{\vartheta}_1\right)^2}{2 \times \hat{\vartheta}_2^2}\right\} \mathrm{d}t. \tag{3.4}$$

Figure 3.2 shows the empirical and the ν-transformed intensity of the claims occurrence process. At the end of the observation period, we can see a steep decrease of the empirical intensity of the claim occurrence process. This is partly caused by the truncation. We do not know about many last claims occurred, because of the reporting delay.

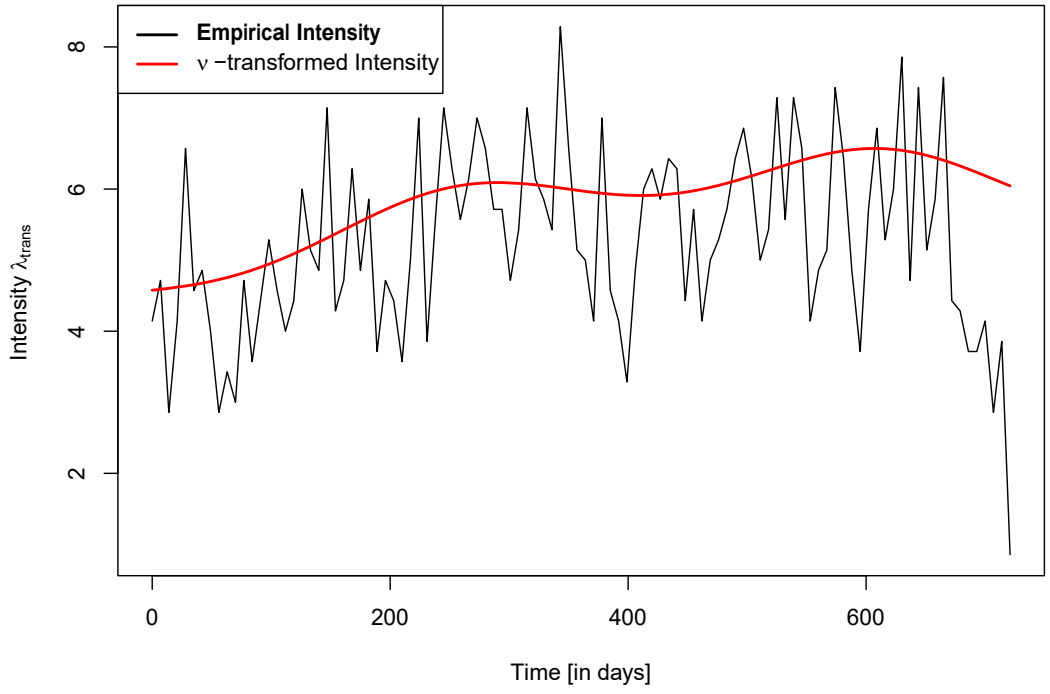

Figure 3.3: Empirical intensity calculated on a weekly basis and the intensity calculated by the ν-transform of the claims occurrence process (0 is 1. 1. 2017).

3.3.4 Distribution of Amounts

Trying to fit the data with log-normal, gamma, and Pareto distribution based on the Akaike Information Criterion (AIC) we opt for the log-normal model as the best. This decision also supports Figure 3.3. Concrete values of the AIC for above-mentioned models can be seen in Table 3.2.

	Log-normal	Gamma	Pareto
AIC	41 349	41 689	41 394

Table 3.2: AIC value for claim amounts data fitted by different distributions.

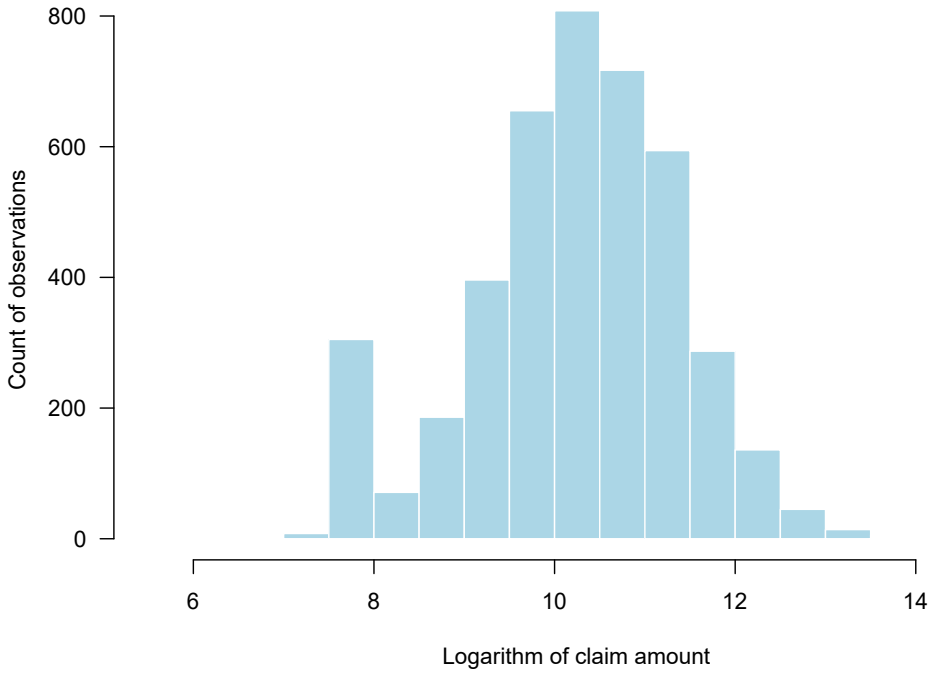

Figure 3.4: Histogram of logarithmic claim amounts.

The ML function is following

$$L_Y(\boldsymbol{\gamma}; y_i, i = 1, \ldots, m) = \prod_{i=1}^{m} \frac{1}{\gamma_2 y_i \sqrt{2\pi}} \exp\left\{ -\frac{\left(\log(y_i) - \gamma_1\right)^2}{2\gamma_2^2} \right\}, \tag{3.5}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^\top$ in this case. It has been maximized expression (3.5) with respect to $\boldsymbol{\gamma}$ and the solution is

$$\hat{\boldsymbol{\gamma}} \doteq (3.316, 1.155)^\top \tag{3.6}$$

In Figure **??** we can see that the fitted density captures the data well.

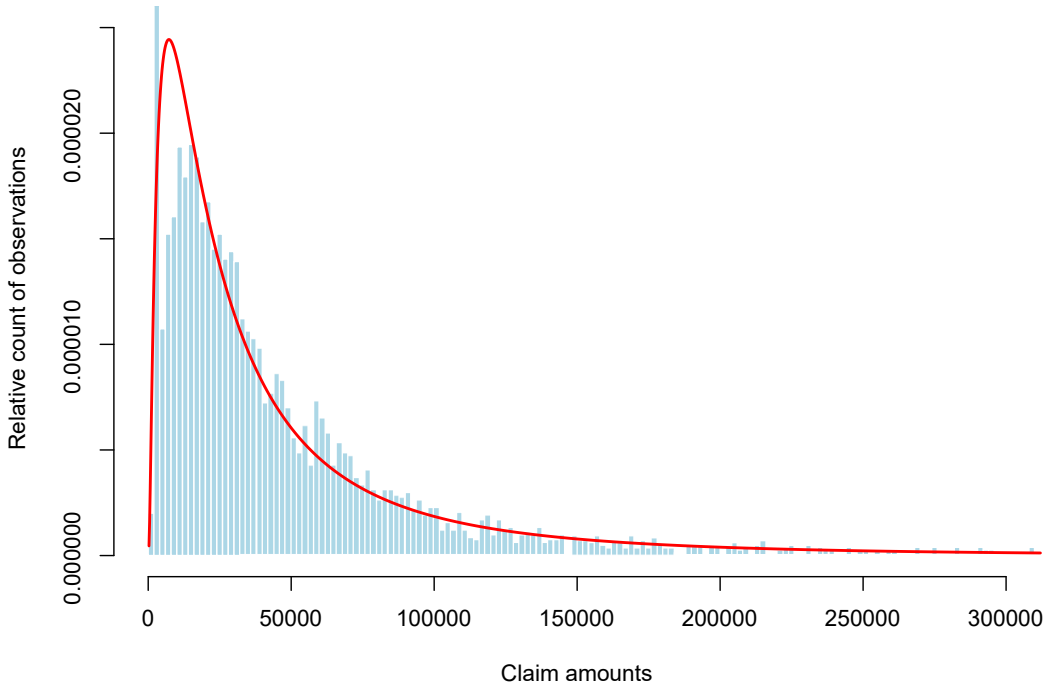

Figure 3.5: Histogram of claim amounts with fitted density (red).

3.4 Generating IBNR

Let us state that by the final model $\mathcal{M}$ we mean the model that assumes:

- $\{Z_i, i = 1, \ldots, m\}$ are arrival times of NHPP with intensity $\lambda(t, \hat{\boldsymbol{\varrho}})$ for $\hat{\boldsymbol{\varrho}}$ stated in (3.1),
- W_i are log-normally distributed with parameter $\hat{\boldsymbol{\vartheta}}$ stated in (3.3),
- the claim occurrence process N is NHPP with intensity $\hat{\lambda}_{transf}(s)$ stated in (3.4),
- Y_i are log-normally distributed with parameter $\hat{\boldsymbol{\gamma}}$ stated in (3.6).

We have generated $K = 1000$ simulations of possible scenarios. Firstly, we look at generated counts of occurred accidents in each year. Figure 3.4 shows the comparison of the mean count of occurred claims generated by the NHPP with intensity (3.4) and observed counts from the data. The mean number of claims per specific year was calculated by numerical integration of

$$\int_s^t \hat{\lambda}_{transf}\,(s)\,\mathrm{d}s,$$

by Wolfram Research, Inc. We can say that the model $\mathcal{M}$ stands up well. Although it does not have the data from the year 2019, it predicts a reasonable

value for that year. As we deal only with material claims, usually the biggest portion of the total claim count is reported in the year of occurrence of the claim. Bearing in mind that the model $\mathcal{M}$ is based on observations from years 2017 and 2018 (the data till 2018).

Let us focus on the yellow and blue bars in Figure 3.4, the model $\mathcal{M}$ predicts almost the same mean total claim count as the claim count observed in the data till 2018 for events that occurred in the year 2017. The total claim count predicted by the model $\mathcal{M}$ for the year 2018 is higher than the observed claim count in the data till 2018, as can be expected. Looking at yellow and green bars in the same figure, we can say that the model $\mathcal{M}$ predicts the total count of claims that occurred in the year 2017, 2018 respectively, only a little bit lower than the number of claims occurred in corresponding years considering the data till the end of the year 2019. Even if the model $\mathcal{M}$ does not know what will happen in the year 2019. Especially, for claims that occurred in the year 2019 the model $\mathcal{M}$ predicts a higher total number of claims than the number of claims observed in 2019.

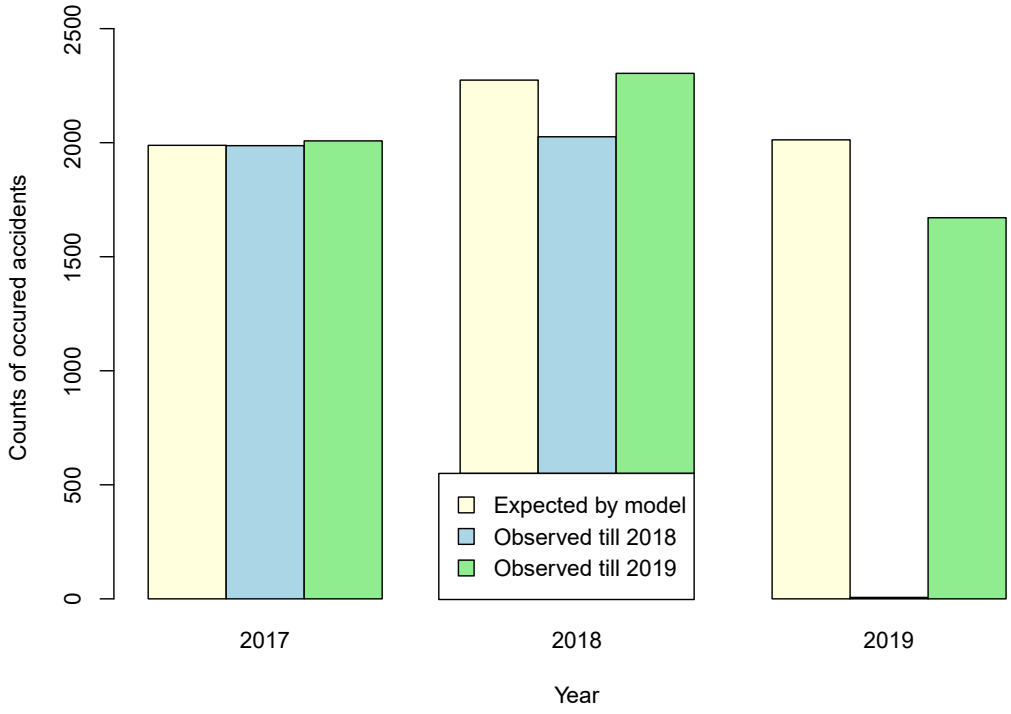

Figure 3.6: Comparison of the mean value of the model $\mathcal{M}$ with reality.

For each claim count N^k generated there were simulated N^k claim amounts from previously estimated distribution in Section 3.3.4. The difference in total claim amount simulated and observed per corresponding year gives us the INBR claim reserve. In Table 3.3 and 3.4 we can see basic characteristics of the IBNR claim reserve distribution for each year. More specifically, in Figure 3.5 and 3.6 are shown histograms of the simulated IBNR claim reserve predicted values. With 99% probability claims reported after the year 2018 that occurred in the year 2017

will not exceed the amount of 12 737 907 CZK. Similarly, for claims reported after the year 2018 that occurred in 2018, we state that their total amount will not be higher than 27 543 360 CZK.

Min.	1st Qu.	Median	Mean	3rd. Qu	Max.
-2 762 734	4 575 612	6 290 720	6 336 770	8 197 856	14 730 793

Table 3.3: Summary of simulated INBR claim reserves per year 2017.

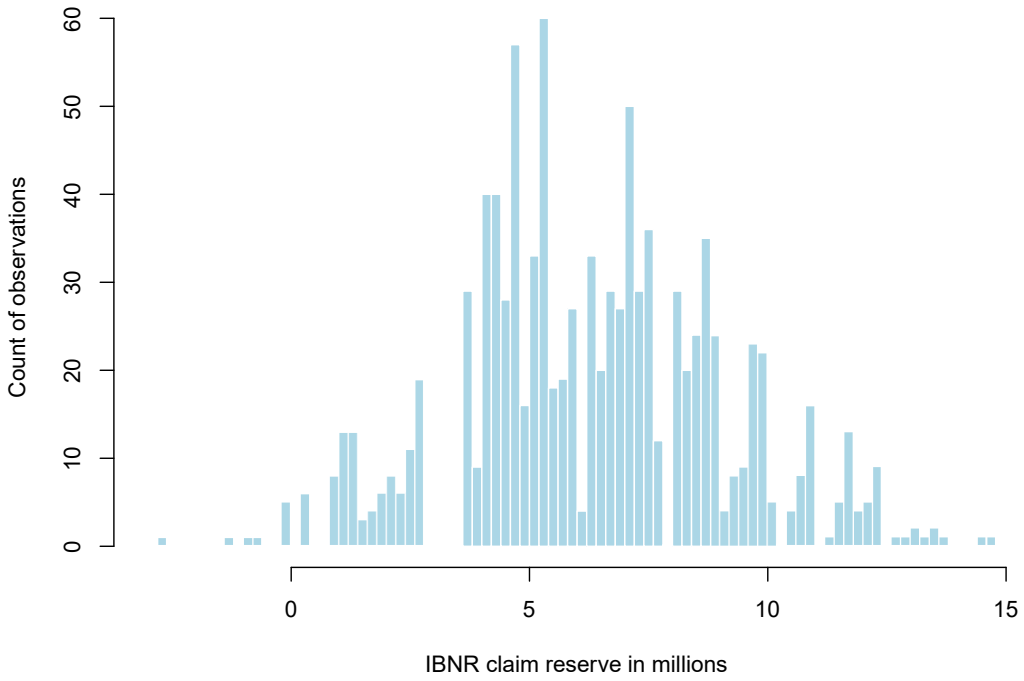

Figure 3.7: Histogram of simulated IBNR claim reserves for 2017.

Min.	1st Qu.	Median	Mean	3rd. Qu	Max.
12 408 250	19 395 995	21 606 516	21 258 683	23 253 522	28 628 612

Table 3.4: Summary of simulated INBR claim reserves per year 2018.

Finally, in Figure 3.7 we can see the mean predicted IBNR claim reserve at time T (end of the year 2018) for each year – the yellow bar. The blue bar represents the total claim amounts observed in the year 2019 for claims that occurred in the previous two years. Note that the total claim amount observed till time T is 104 620 234 CZK for claims that occurred in the year 2017 and 100 706 878 CZK for claims that occurred in the year 2018. Evaluating the model $\mathcal{M}$ we can say that it predicts reasonable values for the ultimate claim amount based on the two-year observation period. We must point out that material claims do not have heavy tails as for example, bodily injury claims do.

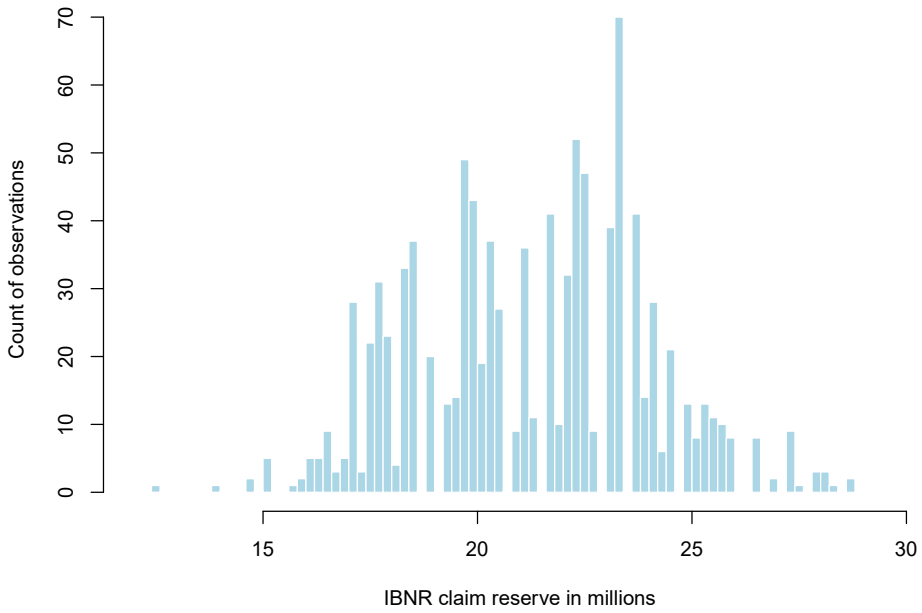

Figure 3.8: Histogram of simulated IBNR claim reserves for 2018.

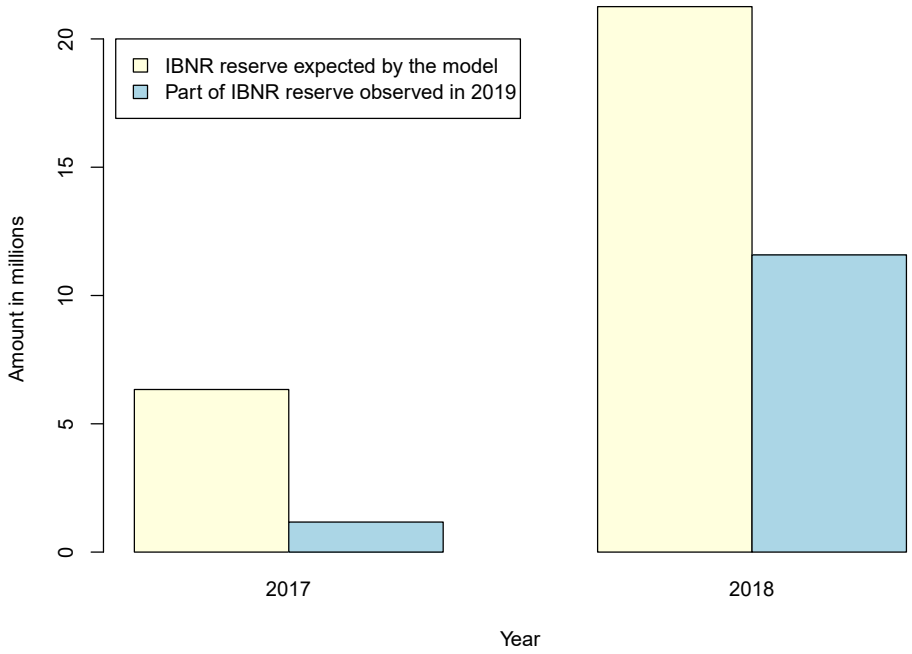

Figure 3.9: Comparison of the mean predicted IBNR claim reserve calculated at the end of the year 2018 (yellow) and the sum of claim amounts for claims reported in 2019 that occurred in the previous years (blue).

Conclusion

In conclusion, the task of dealing with marked stochastic processes, where some realizations of the process are cut off, has been successfully fulfilled in this thesis. Whole problematics has been incorporated in the context of delayed reporting of claims in non-life insurance.

The first chapter provided a thorough theoretical background, introducing the homogeneous Poisson process and its generalization, as well as the concept of marking. The ν-transform was defined for general processes, especially, the intensity of ν-transformed process was expressed in Theorem 4 for the case of the Poisson process, accompanied by an illustrative examples. Additionally, we mentioned the problem of truncation.

In the second chapter, there was expressed the maximum likelihood function for the marked process. The goodness-of-fit test was proposed for nonhomogeneous Poisson process. Moreover, the approach how to generate the IBNR claim reserve distribution was listed.

The third chapter applies the theoretical approach to claim data from the Motor Third Party Liability insurance provided by the Czech Insurers' Bureau. Making specific assumptions a simple model was created, but extensions with time-varying parameters are possible as it was mentioned in the second chapter. The estimated intensity of the occurrence process was obtained assuming that the reporting process forms a nonhomogeneous Poisson process, and that delays are log-normally distributed, and that the distribution of delays does not change in time. Finally, the IBNR claim reserve prediction was made.

The key contribution of this thesis is the well-arranged elaboration of the usage of the ν-transform in the claims reserving problem to estimate the occurrence process intensity while dealing with truncated data. Also, we get the empirical distribution of the IBNR claim reserve prediction. Nevertheless, this method is computationally intense as it uses data that would otherwise be discarded in aggregation.

The methodology presented in this thesis can be extended to more complex cases. Future works can focus on including time-varying conditional distribution of delays or modeling the reporting process by some more complex process. Moreover, there can be incorporated a payment process. The common situation in practice is when there is not just one payment as it was assumed here, but more payments of different amounts made at different times. The practical application to estimate claim reserves for insurance companies can provide significant benefits by making use of the data collected by insurance companies about each claim and not just aggregating them as it is routine nowadays.

Overall, this thesis provides a foundation for future work in this area. This approach has the potential for practical applications in the insurance industry.

Bibliography

E. Arjas. The claims reserving problem in non-life insurance: Some structural ideas. *ASTIN Bulletin: The Journal of the IAA*, 19(2):139–152, 1989.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, New York, New York, 2nd edition, 2003.

W. H. Greene. *Econometric analysis*. Prentice Hall, Upper Saddle River, New Jersey, 5th edition, 2002.

M. Jacobsen. *Point Process Theory and Applications*. Birkhäuser Verlag, Boston, Massachusetts, 1st edition, 2006.

O. Kallenberg. *Probability Theory and Stochastic Modelling: Volume 99: Foundations of Modern Probability*. Springer, Cham, Switzerland, 3rd edition, 2021.

J. F. C. Kingman. *Poisson processes*. Oxford University Press, New York, New York, 1993.

M. Kulich. Poznámky k přednášce. `https://www.karlin.mff.cuni.cz/~kulich/vyuka/ms1/doc/ms1_170112.pdf`, 2017. Accessed: 2023–07-06.

M. Maciak, O. Okhrin, and M. Pešta. Infinitely stochastic micro reserving. *Insurance: Mathematics and Economics*, 100(C):30–58, 2021.

K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, England, 1st edition, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

H. C. Tijms. *A First Course in Stochastic Models*. Wiley, Chichester, England, 2003.

Wolfram Research, Inc. Mathematica, Version 13.2. URL `https://www.wolfram.com/mathematica`. Champaign, Illinois, 2022.

List of Figures

List of Tables

List of Abbreviations and Notation

CIB ... Czech Insurers' Bureau
MTPL ... Motor Third Party Liability
IBNR ... incurred but not reported
HPP ... homogeneous Poisson process
NHPP ... nonhomogeneous Poisson process
MLE ... maximum likelihood estimation
ML ... maximum likelihood
AIC ... Akaike Information Criterion
r. v. ... random variable

$\mathbb{R}$ ... the set of real numbers
$\mathbb{N}$ ... the set of natural numbers
$\mathcal{A}$ ... sigma algebra on a set Ω
$\mathcal{S}$ ... sigma algebra on a set S
$\mathcal{K}$ ... sigma algebra on a set K
P ... probability measure
N_t, M_t ... counting processes
$\lambda(t)$ ... intensity function
$f_{W|Z}$ ... conditional density of r. v. W given the value of r. v. Z