

Mémoire présenté le : 8 mars 2022
pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA
et l'admission à l'Institut des Actuaires

Par : Vincent MAAREK

Titre : Modélisation de la déviation de la prime

Confidentialité : NON (Durée : 1 an 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

*Membres présents du jury de Signature
l'Institut des Actuaires*

Sophie Michon

Ramy Ibrahim

.....

*Membres présents du jury de
l'ISFA*

Stéphane Loisel

.....

.....

Entreprise :

Nom : AXA France

Signature :

*Directeur de mémoire en entre-
prise :*

Nom : Marie Gauvin

Signature :

Invité :


Nom :

Signature :

***Autorisation de publication et
de mise en ligne sur un site de
diffusion de documents actua-
riels (après expiration de l'éventuel
délai de confidentialité)***

Signature du responsable entreprise

Signature du candidat



Modeling the deviation of premiums

Vincent MAAREK

6th February 2022

Abstract

La directive Solvabilité II impose de quantifier les marges de solvabilité par l'intégration de la notion de risque dans les calculs. Notamment, le *Solvency Capital Requirement* (SCR) requiert d'évaluer la distribution des risques auxquels est soumis un assureur, et de mesurer les queues de ces distributions. En particulier, le risque de prime concerne la mesure du risque que le coût des sinistres futurs d'un portefeuille soit supérieur aux primes perçues sur ce portefeuille. Ce mémoire propose une extension du modèle de Jean-Philippe Boucher et Guillaume Couture-Piché [BCP16] afin de mesurer la déviation de la prime encaissée sur un portefeuille donné. L'objectif est centré sur le modèle en lui-même. Des exemples d'applications sont recensés, à la fois dans le contexte du risque de prime, mais également dans un champ actuariel plus large.

The Solvency II directive requires to compute solvency margins by integrating the notion of risk in their computations. More especially, the *Solvency Capital Requirement* (SCR) requires to determine the distribution of all the risks undertaken by an insurer, and to measure the tails of those distributions. In particular, the premium risk deals with the measure of the risk that the cost of future claims for a given portfolio will exceed the premium charged on this portfolio. This thesis extends the model of Jean-Philippe Boucher and Guillaume Couture-Piché [BCP16] in order to measure the deviation of the premium collected on a given portfolio. Focus is made on the model itself. Application examples are provided in a Solvency II context, but also in a wider actuarial context.

Mots clés— Risque de prime, processus de Poisson homogène, processus de Poisson inhomogène, processus de Cox, modèle collectif, files d'attente

Keywords— Premium risk, homogeneous Poisson process, inhomogeneous Poisson process, Cox process, collective model, queuing theory

Acknowledgements

This master thesis is the final point of my studies at the *Institut de Sciences Financières et d'Assurances*, which have been made possible thanks to my former manager at AXA France, Marie Gauvin. I would like to thank her, as well as Pierre-Louis Blanc, who pushed my application at AXA and encouraged me through this journey.

I would also like to thank Antoine Chopineau, my former manager at Mazars, who made my year at Mazars incredible, and who reinforced my taste for the actuarial field. I would also like to thank him for pushing my application at ISFA.

I would also like to thank Frédéric Planchet who took the time to help me when I needed some insights on my model.

I would like to thank my parents, for supporting me through this journey, and especially my mother who, for the third time (after my thesis at Centrale-Supélec and ESCP), reviewed this thesis to avoid language and syntax mistakes.

Finally, I would like to thank my fiancée Julie Blervaque for always be supportive and encouraging.

Introduction

This master thesis originates from my time at AXA France, where I was working as an internal model and reinsurance actuary. I worked specifically on the premium risk, both on the claim side and on the premium side. I wanted to improve the results of the model in place on the premium side, and decided to work on it on my own. I initiated this work way before I started my studies at ISFA. Working on this side project was a way for me to pursue my work on actuarial research and development initiated when I was working at Mazars with Antoine Chopineau.

At that time, I was working on an individual claim reserving model adapted from Alexandre Boumezoued [BD17]. This work showed me the potential of individual data to better capture and measure risk. My idea was then to pursue with this individual approach for developing the model on the premium risk. During the bibliography review, I found the article written by Jean-Philippe Boucher and Guillaume Couture-Piché [BCP16], whose modeling of the number of contracts was conducted in a very elegant and natural way. I decided to work on it to initiate my model.

As this work was initiated before I started my studies at ISFA, the premises of the model were written in English. After starting my studies at ISFA, I thought that this would be an interesting master thesis subject, and I pursued my work more in details on it, and kept English as a working language.

Going back to the thesis itself, its original purpose was to propose an alternative to premium risk modeling, but ended in discovering that it could be applied in many other fields than solvency margins measurements. It also differs from its individual data approach, that enables a much better fit and requires lower historical records than standard aggregate data which are current practice in the market. Thus, this thesis doesn't focus on the application of the proposed model to Solvency II, but rather introduces a technical framework, that can be used either in internal modeling, pricing, customer lifetime value, etc.

It relies strongly on the results introduced in [BCP16], for which I have borrowed the queuing theory and the service time framework. However, this thesis extends the model by working on some more general arrival processes for the

queuing part, and by proposing a semi-parametric approach for the service time, that can also be extrapolated. An innovative algorithm is proposed to simulate the service time.

The idea of the model is to distinguish the number of contracts and the individual premium in the portfolio. The number of contracts modeling is the major part of this thesis, as this is where the main improvements are proposed. The individual premium will be modeled in a straight-forward way. The thesis is constructed in the following way. Homogeneous Poisson, inhomogeneous Poisson and Cox processes are introduced as they are the basis of the number of contracts modeling. Classic results of the queuing framework are also introduced. Then, parameters estimation methods are introduced. Those estimations will be performed, and simulation algorithms will be described and applied. Finally, we will fit the premium straightforwardly, and model the overall portfolio using a collective risk approach. The performance of the different approaches will be compared and discussed.

Contents

1	Modeling the number of contracts	6
1.1	Reminders on Poisson processes	6
1.1.1	Basic results on the homogeneous Poisson process	6
1.1.2	Basic results on the non-homogeneous Poisson process	10
1.1.3	Introduction to Cox Processes	14
1.2	The Queuing Theory	19
1.2.1	Introduction	19
1.2.2	The model	20
1.2.2.1	Preliminaries and notations	21
1.2.2.2	Proof of the output process distribution	23
1.2.2.3	Adaptation to the non-homogeneous case	27
1.2.2.4	An extension to the Cox process	30
2	Estimation	31
2.1	The data	31
2.2	Estimating the intensity function for the arrival process in the homogeneous and inhomogeneous cases	32
2.3	Estimating the subordinator parameters for the arrival process in the Cox case	32
2.4	Estimating the survival function	33
2.4.1	The fully parametric model	34
2.4.2	The semi-parametric model	34
3	Inference and simulation	36
3.1	Presentation of the data	36
3.1.1	General presentation	36
3.1.2	The number of contracts	37
3.1.3	Survival data	42
3.2	Fitting the survival function	44
3.2.1	Likelihood	44
3.2.2	Simulating the survival function	46
3.2.3	Fitting	49
3.3	Fitting the arrival process	51
3.3.1	Fitting the homogeneous Poisson process	52

	Intensity function estimation	52
	Parameter estimation	53
	Simulations	53
3.3.2	Fitting the non-homogeneous Poisson process	54
	Intensity function estimation	55
	Parameters estimation	58
	Simulations	58
3.3.3	Fitting the Poisson-Gamma Cox process	60
	Overdispersion	60
	Shape function	62
	Parameters estimation	63
	Simulation	64
3.3.4	Comparisons with closed formulas from the queuing theory	66
4	Models comparison and performance	71
4.1	Modeling the premium	71
4.1.1	Background on the premium	71
4.1.2	The collective model	72
4.1.3	Model for the individual premium	73
4.2	Performance comparison	75
4.2.1	Simulation and metrics	75
4.2.2	Results	77
5	Discussion and conclusion	83
	Bibliography	86

Chapter 1

Modeling the number of contracts

1.1 Reminders on Poisson processes

As exposed in the introduction, Poisson processes will be at the heart of the number of contracts model. This section depicts basic results on Poisson processes. First, homogeneous Poisson processes will be introduced, as they constitute the most simple way to build a queuing system framework. Then, we will enlarge the framework to non-homogeneous (or inhomogeneous) Poisson processes. Such processes enable to add more flexibility, on the constructed queuing system, which will be discussed later. Finally, doubly-stochastic Poisson processes, often named Cox processes, will be depicted to benefit from a maximum flexibility in the model, but leading to a loss of generality in the queuing system framework, as closed formula won't be available anymore.

Most of this section is comprised of definitions and theoretical results, and the associated proofs. The reader will be guided through these results with illustrating examples in order to see how those results interact with each other.

1.1.1 Basic results on the homogeneous Poisson process

The underlying probability distribution driving Poisson processes is the exponential distribution. Its memoryless property is key in this model. Gamma and Poisson distributions definitions are recalled, as the former will be used as an intermediate result.

Definition 1.1.1 (Exponential distribution). *A random variable S is said to follow an exponential distribution with parameter $\lambda > 0$ if S is positive and its distribution function f_S satisfies*

$$f_S(x) = \lambda \exp(-\lambda x)$$

Proposition 1.1.1 (Memoryless property of exponential distribution). *Let $t > 0, s > 0$. A random variable S exponentially distributed satisfies the memoryless property, that is:*

$$\mathcal{P}(S > x + t \mid S > x) = \mathcal{P}(S > t)$$

Proof. Let S be a random variable exponentially distributed.

$$\begin{aligned} \mathcal{P}(S > x + t \mid S > x) &= \frac{\mathcal{P}(S > x + t \cap S > x)}{\mathcal{P}(S > x)} \\ &= \frac{\mathcal{P}(S > x + t)}{\mathcal{P}(S > x)} \\ &= \frac{\exp(-\lambda(x + t))}{\exp(-\lambda x)} \\ &= \exp(-\lambda t) \\ &= \mathcal{P}(S > t) \end{aligned}$$

□

Let's assume you have been waiting for an event for s units of time. The memoryless property says that the distribution of the remaining waiting time until the event is unchanged by the s units of time you have been already waiting for. This desirable property is shared by a very few classical distributions. The survival curve that will be introduced in section 3.1.3 doesn't share this property, and will make simulations much more difficult.

Definition 1.1.2 (Poisson distribution). *A random variable N is said to be Poisson-distributed with parameter $\lambda > 0$ if N takes its value in \mathbb{N} and its distribution function f_N satisfies*

$$f_N(k) = \frac{\lambda^k \exp(-\lambda)}{k!}$$

Definition 1.1.3 (Gamma distribution). *A random variable G is said to be Gamma-distributed with parameters $n \in \mathbb{N}^*$ and $\lambda > 0$ if G is positive and its distribution function f_G satisfies*

$$f_G(x) = \frac{(\lambda x)^{k-1} \lambda \exp(-\lambda x)}{(k-1)!}$$

Consider a system where individuals enter at time $0 < t_1 < t_2 < \dots < t_n < \dots$. Assume that the increments $s_i = t_i - t_{i-1}$, ($s_1 = t_1$) are independent random variables and exponentially distributed with parameter λ . For any $n \in \mathbb{N}$, denote $S_n = \sum_{i=1}^n s_i$.

This system can be referred as an insurance portfolio, where insured underwrite their contracts at time $(t_i)_{1 \leq i \leq n}$. The increments $(s_i)_{1 \leq i \leq n}$ represent in this case the time difference between two underwritings. Such increments require the need for continuous data, that is, individual data for each insured.

Proposition 1.1.2. S_n is Gamma-distributed with parameters n and λ

Proof. We prove this result by induction. For $n = 1$, $S_1 = s_1$ is exponentially distributed, that is, $f_{S_1}(x) = \lambda \exp(-\lambda x)$, which is a Gamma distribution with parameters $n = 1$ and λ . Hence, the proposition holds for $n = 1$. Assume that the proposition holds for a certain $n \in \mathbb{N}^*$. Hence,

$$\begin{aligned} f_{S_{n+1}}(x) &= \int_0^x f_{S_n, s_{n+1}}(t, x-t) dt \\ &= \int_0^x f_{S_n}(t), f_{s_{n+1}}(x-t) dt \text{ by independence} \\ &= \int_0^x \frac{(\lambda t)^{n-1} \lambda \exp(-\lambda t)}{(n-1)!} \lambda \exp(-\lambda(x-t)) dt \\ &= \frac{\lambda^n \exp(-\lambda x)}{(n-1)!} \int_0^x t^{n-1} dt \\ &= \frac{(\lambda x)^n \lambda \exp(-\lambda x)}{n!} \end{aligned}$$

This concludes the proof. \square

Now define the following counting process:

$$N(t) = \sum_{i=1}^{+\infty} \mathbf{1}_{S_i < t}$$

We have $N(0) = 0$. This process counts the number of arrivals in $[0, t[$. In other words, this process counts the number of contracts that have been underwritten between 0 and t , which is a quantity of interest for our model.

Proposition 1.1.3. For any $t \geq 0$, $N(t)$ is Poisson-distributed with parameter λt .

Proof. We have $\mathcal{P}(N(t) = k) = \mathcal{P}(S_k \leq t \cap S_{k+1} > t)$. Hence:

$$\begin{aligned} \mathcal{P}(S_k \leq t \cap S_{k+1} > t) &= \int_0^t \mathcal{P}(S_{k+1} > t \mid S_k = x) \times f_{S_k}(x) dx \\ &= \int_0^t \mathcal{P}(s_{k+1} > t - x) \times f_{S_k}(x) dx \\ &= \int_0^t \exp(-\lambda(t-x)) \frac{(\lambda x)^{k-1}}{(k-1)!} \lambda \exp(-\lambda x) dx \\ &= \frac{\exp(-\lambda t) \lambda^k}{(k-1)!} \int_0^t x^{k-1} dx \\ &= \frac{(\lambda t)^k}{k!} \exp(-\lambda t) \end{aligned}$$

This concludes the proof. \square

Two key properties of the constructed Poisson process are its independent and stationary increments. Those properties are the direct consequences of the memoryless property of the exponential distribution.

The interest here is that the knowledge of what happened before, namely, the number of underwritten contracts in the past, is not of any use for what is likely to happen in the future.

Both propositions 1.1.4 and 1.1.5 are taken from [Ruw06].

Proposition 1.1.4 (Stationary increments). *Let $t > 0$, $s > 0$, $t \geq s$. Then $N(t) - N(s)$ has the same distribution as $N(t - s)$.*

Proof. Assume that $k \in \mathbb{N}$ arrivals occurred before s . Let's denote by τ_1^s the waiting time between s and S_{k+1} . The inter-arrival times after S_{k+1} remain identical, that is, equal to s_i , $i \geq k + 2$. For simplicity we will write $\tau_i^s = s_{k+i}$ for $i \geq 2$. Due to the memoryless property of the exponential distribution, τ_1^s is exponentially distributed, that is, identically distributed as, and independent from τ_i^s , $i \geq 2$. Define S_n^s , $n \geq 1$ such as $S_n^s = \sum_{i=1}^n \tau_i^s$. Hence, S_n^s is gamma-distributed with parameters n and λ according to proposition 1.1.2. Thus,

$$\mathcal{P}(N(t) - N(s) = i) = \mathcal{P}(S_i^s \leq t - s) - \mathcal{P}(S_{i+1}^s \leq t - s)$$

We know from proposition 1.1.3 and its proof that the above calculation leads to a Poisson distribution of parameters $\lambda(t - s)$, which concludes the proof. \square

Proposition 1.1.5 (Independent increments). *Let $i \geq 2$. For any $0 < t_{i-1} < t_i < t_{i+1}$, $N(t_{i+1}) - N(t_i)$ is independent from $N(t_i) - N(t_{i-1})$.*

Proof. We will use the same notations as proposition 1.1.4. That is, $\tau_1^{t_{i-1}}$ represents the waiting time between t_{i-1} and the first arrival after t_{i-1} . $\tau_1^{t_i}$ represents the waiting time between t_i and the first arrival after t_i . Because of the memoryless property, $\tau_1^{t_{i-1}}$ and $\tau_1^{t_i}$ are exponentially distributed, and independent of the other inter-arrival times. The associated sum of inter-arrival times from t_{i-1} and t_i respectively write $S_n^{t_{i-1}} = \sum_{p=1}^n \tau_p^{t_{i-1}}$ and $S_n^{t_i} = \sum_{p=1}^n \tau_p^{t_i}$. They are Gamma-distributed with parameters n and λ . Denote A_k^i the following event:

$$\{A_k^i\} = \left\{ \left[S_k^{t_{i-1}} \leq t_i - t_{i-1} \right] \setminus \left[S_{k+1}^{t_{i-1}} \leq t_i - t_{i-1} \right] \right\}$$

We get:

$$\mathcal{P}(N(t_i) - N(t_{i-1}) = k \cap N(t_i) - N(t_{i-1}) = p) = \mathcal{P}(A_k^i \cap A_p^{i+1})$$

The events A_k^i and A_p^{i+1} involve disjoint inter-arrival times, that is, independent inter-arrival times. Hence, we can separate those two events, which concludes the proof. \square

This last result will be used in the queuing system framework. Intuitively, it states that conditional on $N(t) = n$ contracts underwritten between 0 and t , the distribution of the underwriting times is uniform on $[0, t[$.

Proposition 1.1.6. *Let $n \in \mathbb{N}^*$ and $0 < t_1 < \dots < t_n < t$ be defined as above. Then, conditional on $N(t) = n$, the joint density of t_1, \dots, t_n is constant over $[0, t]$. That is,*

$$\mathcal{P}(t_1, \dots, t_n \mid N(t) = n) = \frac{n!}{t^n}$$

In particular,

$$\mathcal{P}(t_1 \mid N(t) = 1) = \frac{1}{t}$$

Proof. Using the Bayes rule, we have,

$$\mathcal{P}(t_1, \dots, t_n \mid N(t) = n) = \frac{\mathcal{P}(N(t) = n \mid t_1, \dots, t_n) \mathcal{P}(t_1, \dots, t_n)}{\mathcal{P}(N(t) = n)}$$

To calculate $\mathcal{P}(t_1, \dots, t_n)$, we use,

$$\mathcal{P}(t_1, \dots, t_n) = \prod_{i=1}^n \mathcal{P}(t_i \mid t_{i-1}, \dots, t_1)$$

Recall that the increments are independent and identically distributed random variables, that is, $\mathcal{P}(t_i \mid t_{i-1}, \dots, t_1) = \mathcal{P}(t_i \mid t_{i-1}) = \lambda \exp(-\lambda(t_i - t_{i-1}))$. Hence,

$$\mathcal{P}(t_1, \dots, t_n) = \lambda^n \exp(-\lambda t_n)$$

Since $N(t)$ is Poisson-distributed,

$$\mathcal{P}(N(t) = n) = \frac{(\lambda t)^n \exp(-\lambda t)}{n!}$$

Finally, the probability $\mathcal{P}(N(t) = n \mid t_1, \dots, t_n)$ is the probability that the increment between t_n and t_{n+1} , namely s_{n+1} , is higher than $t - t_n$. Since s_{n+1} is exponentially distributed, we obtain,

$$\mathcal{P}(s_{n+1} > t - t_n) = \int_{t-t_n}^{+\infty} \lambda \exp(-\lambda x) dx = \exp(-\lambda(t - t_n))$$

This concludes the proof. □

1.1.2 Basic results on the non-homogeneous Poisson process

The non-homogeneous Poisson process is a generalization of the homogeneous Poisson process introduced above. The former allows the parameter λ , now called the *intensity function*, to vary with time. The intensity function can be

perceived as the frequency at which contracts are underwritten. In the homogeneous case, this intensity is constant, meaning that contracts are underwritten at the same frequency over time.

This assumption may be quite strong in some cases. One can think of aggressive commercial offers, for which an increase of the underwriting rate is expected. In this case, the intensity rate should be higher during the commercial offer period, than usual. The inhomogeneous Poisson process proposes a solution to take into account such varying rates. Then, in this section, we will consider the intensity as a function of time. Hence $\lambda \rightarrow \lambda(t)$.

Definition 1.1.4 (Non-homogeneous Poisson process). *Consider any collection $\mathcal{I} = (I_1, \dots, I_n)$ of disjoint intervals on \mathbb{R}^+ . That is, consider $0 < t_0 < t_1 < \dots < t_n < +\infty$, and set $I_i = [t_{i-1}, t_i[$. Let $\lambda(t)$ be a positive and integrable function over any finite interval I , that is $\int_I \lambda(t) dt < +\infty$. Let $N(I)$ be a counting process of the number of points on the interval I . $N(I)$ is said to be a non-homogeneous Poisson process if it satisfies the following assertions:*

- $N(I)$ is Poisson-distributed with parameter $\Lambda(I) = \int_I \lambda(t) dt$
- For any set \mathcal{I} defined as above, $N(I_1), \dots, N(I_n)$ are independent. This is the independent increments property.

Note that the stationary increments property is not satisfied anymore. Indeed, let's get back to the aggressive commercial offer introduced above, and say that this offer is active from time t_1 to t_2 , such that $t_2 - t_1 < t_1$. Assume that in normal time, the intensity is given by λ_n , and during the offer, it is given, by λ_o . Then $N(t_2 - t_1)$ and $N(t_2) - N(t_1)$ are not identically distributed. Indeed, the former is Poisson-distributed with parameter $\lambda_n t$, while the latter is Poisson-distributed with parameter $\lambda_o t$.

Fortunately, it is still possible to determine the joint distribution of the occurrence times, that is, a result similar to proposition 1.1.6.

Proposition 1.1.7. *Let $t_1 < \dots, t_n$ be the realizations of n random variables, representing the occurrence times of a point process, such that $0 < t_1 < \dots < t_n < t$. Let $N([0, t])$ be the non-homogeneous Poisson process associated to those occurrence times: $T_n = \min \{t \in [0, +\infty[, N([0, t]) = n\}$. Then the joint distribution of T_1, \dots, T_n writes:*

$$\mathcal{P}(t_1, \dots, t_n) = e^{-\Lambda(t_n)} \prod_{i=1}^n \lambda(t_i)$$

Proof. We will prove this result by induction. We will first work on the cumulative distribution, and derive it to obtain the distribution function.

Let's prove the result for $n = 1$. We need to determine

$$\mathcal{P}(T_1 \leq t_1) = 1 - \mathcal{P}(T_1 > t_1)$$

Saying that the first event occurs after t_1 is equivalent to say that the number of events that occurred before t_1 is zero. Hence:

$$\begin{aligned}\mathcal{P}(T_1 \leq t_1) &= 1 - \mathcal{P}(T_1 > t_1) \\ &= 1 - \mathcal{P}(N([0, t_1]) = 0) \\ &= 1 - e^{-\Lambda(t_1)}\end{aligned}$$

Taking the first derivative of the previous equation, one gets:

$$\mathcal{P}'(t_1) = \lambda(t_1)e^{-\Lambda(t_1)}$$

Let $n \in \mathbb{N}$. Assume that

$$\mathcal{P}(t_1, \dots, t_n) = e^{-\Lambda(t_n)} \prod_{i=1}^n \lambda(t_i)$$

Then one gets:

$$\begin{aligned}\mathcal{P}(t_1, \dots, t_n, T_{n+1} \leq t_{n+1}) &= \mathcal{P}(T_{n+1} \leq t_{n+1} \mid t_1, \dots, t_n) \times \mathcal{P}(t_1, \dots, t_n) \\ &= (1 - \mathcal{P}(T_{n+1} > t_{n+1} \mid t_1, \dots, t_n)) \times \mathcal{P}(t_1, \dots, t_n)\end{aligned}$$

Saying that the $(n+1)^{\text{th}}$ event occurs after t_{n+1} knowing that the n^{th} occurred at t_n is equivalent to say that the number of events that occurred between t_n and t_{n+1} is zero. Hence:

$$\begin{aligned}\mathcal{P}(T_{n+1} > t_{n+1} \mid t_1, \dots, t_n) &= \mathcal{P}(N([t_n, t_{n+1}]) = 0) \\ &= e^{-(\Lambda(t_{n+1}) - \Lambda(t_n))}\end{aligned}$$

Hence,

$$\mathcal{P}(t_1, \dots, t_n, T_{n+1} \leq t_{n+1}) = \left(1 - e^{-(\Lambda(t_{n+1}) - \Lambda(t_n))}\right) \times \mathcal{P}(t_1, \dots, t_n)$$

Taking the first derivative, one gets:

$$\begin{aligned}\mathcal{P}'(t_1, \dots, t_n, t_{n+1}) &= \lambda(t_{n+1})e^{-(\Lambda(t_{n+1}) - \Lambda(t_n))} \times \mathcal{P}(t_1, \dots, t_n) \\ &= e^{-\Lambda(t_{n+1})} \prod_{i=1}^{n+1} \lambda(t_i)\end{aligned}$$

□

Lemma 1.1.1. *Let $t_1 < \dots, t_n$ be the realizations of n random variables, representing the occurrence times of a point process, $t > 0$ and $T > t$ such that $t < t_1 < \dots < t_n < T$. Let $N([t, T])$ be the non-homogeneous Poisson counting process associated to the occurrence times between t and T : $T_n = \min \{\tau \in [t, +\infty[, N([t, \tau]) = n\}$. Then the joint distribution of T_1, \dots, T_n knowing that T_1 occurs after t writes:*

$$\mathcal{P}(t_1, \dots, t_n \mid T_1 > t) = e^{-(\Lambda(t_n) - \Lambda(t))} \prod_{i=1}^n \lambda(t_i)$$

Proof. We will prove the result for $n = 1$, the induction step is the same as the one in proposition 1.1.7.

$$\begin{aligned}\mathcal{P}(T_1 \leq t_1 \mid T_1 > t) &= 1 - \mathcal{P}(T_1 > t_1 \mid T_1 > t) \\ &= 1 - \mathcal{P}(N([t, t_1]) = 0) \\ &= 1 - e^{-(\Lambda(t_1) - \Lambda(t))}\end{aligned}$$

Taking the first derivative of the previous equation, one gets:

$$\mathcal{P}(t_1 \mid T_1 > t) = \lambda(t_1)e^{-(\Lambda(t_1) - \Lambda(t))}$$

□

Proposition 1.1.8. *Using the same notation as proposition 1.1.7, the joint distribution of T_1, \dots, T_n , conditional on $N([0, t]) = n$ writes:*

$$\mathcal{P}(t_1, \dots, t_n \mid N([0, t]) = n) = \frac{n!}{(\Lambda(t))^n} \prod_{i=1}^n \lambda(t_i)$$

Proof. The Bayes rule writes:

$$\mathcal{P}(t_1, \dots, t_n \mid N([0, t]) = n) = \frac{\mathcal{P}(N([0, t]) = n \mid t_1, \dots, t_n) \times \mathcal{P}(t_1, \dots, t_n)}{\mathcal{P}(N([0, t]) = n)}$$

The probability that n events occurred before t knowing the instants of occurrence of n elements before t , namely $\mathcal{P}(N([0, t]) = n \mid t_1, \dots, t_n)$, is the probability that no events have occurred between t_n and t , that is:

$$\begin{aligned}\mathcal{P}(N([0, t]) = n \mid t_1, \dots, t_n) &= \mathcal{P}(N([t_n, t]) = 0) \\ &= e^{-(\Lambda(t) - \Lambda(t_n))}\end{aligned}$$

According to definition 1.1.4,

$$\mathcal{P}(N([0, t]) = n) = \frac{e^{-\Lambda(t)}(\Lambda(t))^n}{n!}$$

Proposition 1.1.7 leads to

$$\mathcal{P}(t_1, \dots, t_n) = e^{-\Lambda(t_n)} \prod_{i=1}^n \lambda(t_i)$$

This concludes the proof. □

Lemma 1.1.2. *Using the same notation as lemma 1.1.1, the joint distribution of T_1, \dots, T_n , conditional on $N([t, T]) = n$ and $T_1 > t$ writes:*

$$\mathcal{P}(t_1, \dots, t_n \mid N([t, T]) = n \cap T_1 > t) = \frac{n!}{(\Lambda(T) - \Lambda(t))^n} \prod_{i=1}^n \lambda(t_i)$$

Proof. The proof is the same as proposition 1.1.8. Using the Bayes rule, we get the following three components:

$$\begin{aligned} \mathcal{P}(N([t, T]) = n \mid t_1, \dots, t_n \cap T_1 > t) &= \mathcal{P}(N([t, T]) = n \mid t_1, \dots, t_n) \\ &= \mathcal{P}(N([t_n, T]) = 0) \\ &= e^{-(\Lambda(T) - \Lambda(t_n))} \end{aligned}$$

We then have,

$$\begin{aligned} \mathcal{P}(N([t, T]) = n \mid T_1 > t) &= \mathcal{P}(N([t, T]) = n) \\ &= \frac{e^{-(\Lambda(T) - \Lambda(t))} (\Lambda(T) - \Lambda(t))^n}{n!} \end{aligned}$$

Lemma 1.1.1 leads to

$$\mathcal{P}(t_1, \dots, t_n \mid T_1 > t) = e^{-(\Lambda(t_n) - \Lambda(t))} \prod_{i=1}^n \lambda(t_i)$$

This concludes the proof. \square

1.1.3 Introduction to Cox Processes

One known result about Poisson processes, is that the mean equals the variance. Often, data exhibits a variance that seems higher than the mean. This phenomenon is called overdispersion. One common workaround to deal with overdispersed data is to work with so-called Cox processes. Often called *Doubly Stochastic Poisson Processes*, Cox processes are a generalization of the Poisson process, letting the intensity or cumulative intensity function to be stochastic.

Definition 1.1.5 (Cox process). *Let $\lambda(t)$ be a non negative stochastic process, referred as the intensity process. Define the cumulative intensity process $\Lambda(t) = \int_0^t \lambda(s) ds$. Let $N(t)$ be a counting process. Let $(\mathcal{F}_t)_{t \geq 0}$ be the natural filtration of $\Lambda(t)$. $N(t)$ is said to be a Cox process if, conditional to $(\mathcal{F}_t)_{t \geq 0}$, $N(t)$ is a non-homogeneous Poisson process satisfying:*

- $N(t)$ has conditional independent increments. That is, for $0 < s < t$, $N(t) - N(s) \mid \mathcal{F}_t$ is independent of $\mathcal{G}_s = \sigma(N_u, u \leq s)$
- Conditional to \mathcal{F}_t , the increment $N(t) - N(s)$ has the following probability distribution:

$$\mathcal{P}(N(t) - N(s) = k \mid \mathcal{F}_t) = \exp(-(\Lambda(t) - \Lambda(s))) \frac{(\Lambda(t) - \Lambda(s))^k}{k!}$$

The previous definition refers to conditional independency of the increments. The following proposition, taken from [Liu12], will derive a condition on the unconditional independency of increments.

Proposition 1.1.9. *Let $N(t)$ be a Cox process with cumulative intensity process $\Lambda(t)$. Then, $N(t)$ has unconditional independent increments if and only if $\Lambda(t)$ has independent increments.*

Proof. The general idea of the proof is presented in [Liu12], but is not detailed. The idea is to work with covariance, and derive an equality that will prove the condition. Let $0 < a < b < c < d$. For simplicity, we will write $N(d) = N_d$

$$\begin{aligned} & \text{Cov}(N_d - N_c, N_b - N_a) \\ &= \mathbb{E}[(N_d - N_c)(N_b - N_a)] - \mathbb{E}[N_d - N_c] \mathbb{E}[N_b - N_a] \\ &= \mathbb{E}[\mathbb{E}[(N_d - N_c)(N_b - N_a) \mid \mathcal{F}_d]] - \mathbb{E}[N_d - N_c] \mathbb{E}[N_b - N_a] \end{aligned}$$

Since $N(t)$ has conditional independent increments, one can write:

$$\begin{aligned} \mathbb{E}[(N_d - N_c)(N_b - N_a) \mid \mathcal{F}_d] &= \mathbb{E}[N_d - N_c \mid \mathcal{F}_d] \mathbb{E}[N_b - N_a \mid \mathcal{F}_d] \\ &= (\Lambda_d - \Lambda_c)(\Lambda_b - \Lambda_a) \end{aligned}$$

Moreover,

$$\mathbb{E}[N_d - N_c] = \mathbb{E}[\mathbb{E}[N_d - N_c \mid \mathcal{F}_d]] = \mathbb{E}[\Lambda_d - \Lambda_c]$$

Hence,

$$\begin{aligned} & \text{Cov}(N_d - N_c, N_b - N_a) \\ &= \mathbb{E}[\mathbb{E}[(N_d - N_c)(N_b - N_a) \mid \mathcal{F}_d]] - \mathbb{E}[N_d - N_c] \mathbb{E}[N_b - N_a] \\ &= \mathbb{E}[(\Lambda_d - \Lambda_c)(\Lambda_b - \Lambda_a)] - \mathbb{E}[\Lambda_d - \Lambda_c] \mathbb{E}[\Lambda_b - \Lambda_a] \\ &= \text{Cov}(\Lambda_d - \Lambda_c, \Lambda_b - \Lambda_a) \end{aligned}$$

Hence, the increments of the Cox process are unconditionally independent if and only if the cumulative intensity process increments are independent. \square

The choice of the cumulative intensity process is crucial. See proposition 1.1.11 for a convenient choice of the cumulative intensity process.

Proposition 1.1.10 (Overdispersion of Cox processes). *Let $N(t)$ be a Cox process. Then $\mathbb{V}(N(t)) \geq \mathbb{E}(N(t))$.*

The proof of the following result is borrowed from [Cas05].

Proof.

$$\begin{aligned} \mathbb{E}(N(t)^2) &= \mathbb{E}(\mathbb{E}(N(t)^2 \mid \Lambda(t))) \\ &= \mathbb{E}(\Lambda(t) + \Lambda(t)^2) \end{aligned}$$

Then:

$$\begin{aligned} \mathbb{E}(N(t))^2 &= \mathbb{E}(\mathbb{E}(N(t) \mid \Lambda(t)))^2 \\ &= \mathbb{E}(\Lambda(t))^2 \end{aligned}$$

Subtracting both equations, one gets:

$$\begin{aligned}\mathbb{V}(N(t)) &= \mathbb{V}(\Lambda(t)) + \mathbb{E}(\Lambda(t)) \\ &= \mathbb{V}(\Lambda(t)) + \mathbb{E}(N(t))\end{aligned}$$

□

Hence, any positive intensity process would enable to solve the overdispersion issue. However, one convenient choice would be to use a Gamma process for the cumulative intensity process. Indeed, a known result about mixing Poisson and Gamma distributions is that it leads to a negative binomial distribution. We will show that the same result applies with Cox processes and Gamma processes.

The following definition is taken from [Mer17b].

Definition 1.1.6 (Gamma process). *Let $\alpha(t)$ be a non-negative function, and $b > 0$. A càdlàg process $G(t)$ is a Gamma process if it satisfies the following conditions:*

- $G(0) = 0$ a.s.
- G has independent increments
- The increment $G(t) - G(s)$, $0 < s < t$ is Gamma-distributed with the following density function:

$$g(x) = \frac{1}{b \int_s^t \alpha(u) du \Gamma\left(\int_s^t \alpha(u) du\right)} x^{\int_s^t \alpha(u) du - 1} \exp\left(-\frac{x}{b}\right) \mathbf{1}_{\mathbb{R}_+}$$

$\alpha(t)$ is called the shape function, while b is called the scale parameter.

In the case $\alpha(t) = a$, the Gamma process is said to be a *homogeneous Gamma process*. Otherwise, it is a non-homogeneous Gamma process. In the homogeneous case, the increments are *stationary*, thus, the homogeneous Gamma process is a Lévy process. However, the non-homogeneous Gamma process is an *additive* process, that is, the stationary increments property is abandoned, see [Mer17b] and [AAAB19].

We now show the above result.

Proposition 1.1.11. *Let $N(t)$ be a Cox process, with a Gamma process as cumulative intensity process. Then, the increment $N(t) - N(s)$ $0 < s < t$ follows a negative binomial distribution:*

$$\mathcal{P}(N(t) - N(s) = k) = \frac{\Gamma(k + \int_s^t \alpha(u) du)}{\Gamma(\int_s^t \alpha(u) du)} \times \frac{1}{k!} \times \left(\frac{b}{b+1}\right)^k \left(\frac{1}{b+1}\right)^{\int_s^t \alpha(u) du}$$

In particular, $N(t)$ is a negative binomial process.

Proof. Let $\Lambda(t)$ be a Gamma process with parameters a and b . Let $N(t)$ be a Cox process with cumulative intensity $\Lambda(t)$. Let \mathcal{F}_t be the natural filtration of $\Lambda(t)$. Let g denote the density function of the increments of $\Lambda(t)$. The law of total probability writes:

$$\mathcal{P}(N(t) - N(s) = k) = \int_{\mathbb{R}} \mathcal{P}(N(t) - N(s) = k \mid \mathcal{F}_t) \times g(x) dx$$

Definition 1.1.5 writes:

$$\mathcal{P}(N(t) - N(s) = k \mid \mathcal{F}_t) = e^{-(\Lambda(t) - \Lambda(s))} \frac{(\Lambda(t) - \Lambda(s))^k}{k!}$$

Definition 1.1.6 writes:

$$g(x) = \frac{1}{b \int_s^t \alpha(u) du \Gamma\left(\int_s^t \alpha(u) du\right)} x^{\int_s^t \alpha(u) du - 1} \exp\left(-\frac{x}{b}\right) \mathbf{1}_{\mathbb{R}_+}$$

Hence, the integral can be written as:

$$\begin{aligned} \mathcal{P}(N(t) - N(s) = k) &= \int_0^{+\infty} e^{-x} \frac{x^k}{k!} \times \frac{x^{\int_s^t \alpha(u) du - 1} e^{-\frac{x}{b}} dx}{b \int_s^t \alpha(u) du \Gamma\left(\int_s^t \alpha(u) du\right)} \\ &= \frac{\int_0^{+\infty} e^{-x(1+\frac{1}{b})} x^{k+\int_s^t \alpha(u) du - 1} dx}{b \int_s^t \alpha(u) du \Gamma\left(\int_s^t \alpha(u) du\right) k!} \end{aligned}$$

Let $v = x(1 + \frac{1}{b})$, then $dv = (1 + \frac{1}{b}) dx$.

$$\begin{aligned} \int_0^{+\infty} e^{-x(1+\frac{1}{b})} x^{k+\int_s^t \alpha(u) du - 1} dx &= \int_0^{+\infty} e^{-v} \left(\frac{v}{1+\frac{1}{b}}\right)^{k+\int_s^t \alpha(u) du - 1} \frac{dv}{(1+\frac{1}{b})} \\ &= \frac{\int_0^{+\infty} e^{-v} v^{k+\int_s^t \alpha(u) du - 1} dv}{(1+\frac{1}{b})^{k+\int_s^t \alpha(u) du}} \\ &= \frac{1}{(1+\frac{1}{b})^{k+\int_s^t \alpha(u) du}} \Gamma\left(k + \int_s^t \alpha(u) du\right) \end{aligned}$$

Hence,

$$\begin{aligned} \mathcal{P}(N(t) - N(s) = k) &= \frac{\Gamma\left(k + \int_s^t \alpha(u) du\right)}{\Gamma\left(\int_s^t \alpha(u) du\right)} \cdot \frac{1}{k!} \cdot \frac{1}{(1+\frac{1}{b})^k} \cdot \left(\frac{1}{b+1}\right)^{\int_s^t \alpha(u) du} \\ &= \frac{\Gamma\left(k + \int_s^t \alpha(u) du\right)}{\Gamma\left(\int_s^t \alpha(u) du\right)} \cdot \frac{1}{k!} \cdot \left(\frac{b}{b+1}\right)^k \cdot \left(\frac{1}{b+1}\right)^{\int_s^t \alpha(u) du} \end{aligned}$$

□

The previous proposition shows that any increments follows a negative binomial distribution. The mean of a negative binomial distribution with shape parameter a and scale parameter b writes:

$$\mathbb{E}(N(t) - N(s)) = \int_s^t \alpha(u) du \cdot \frac{b}{b+1} \cdot \frac{1}{\frac{1}{b+1}} = b \int_s^t \alpha(u) du$$

The variance writes:

$$\mathbb{V}(N(t) - N(s)) = \int_s^t \alpha(u) du \cdot \frac{b}{b+1} \cdot \frac{1}{\left(\frac{1}{b+1}\right)^2} = b(b+1) \int_s^t \alpha(u) du$$

Since the variance of the gamma process writes $\mathbb{V}(\Lambda(t) - \Lambda(s)) = \int_s^t \alpha(u) du \cdot b^2$, proposition 1.1.10 is satisfied.

In this context, the Gamma process can be considered as a time subordinator. That is, the gamma process modifies the time axis. As introduced in [AAAB19], a Lévy process subordinated by a Lévy subordinator is still a Lévy process. Similarly, a Lévy process subordinated by an additive process is still an additive process.

Thus, since a Poisson process is a Lévy process, subordinating it to either a homogeneous or a non-homogeneous Gamma process maintains the Lévy or additive property. Hence, the negative binomial process is a convenient choice to work with our data.

1.2 The Queuing Theory

1.2.1 Introduction

Jean-Philippe Boucher and Guillaume Couture-Piché, in [BCP16], pointed out that the use of queuing theory was well-suited to model the number of policyholders in an insurance portfolio. In this thesis, the framework is borrowed. Improvements are proposed in terms of arrival processes, as Boucher and Couture-Piché restricted to homogeneous Poisson processes for simplicity purposes. An emphasis is made on the variance of the number of policyholders, which is one of the most important metrics in this thesis. We also show that the expectation is consistent between the different models.

In this section, we will introduce the standard results of queuing theory, as applied by Boucher and Couture-Piché, and we will generalize the results in order to take into account non-stationarity in the arrival process, as well as capturing the overdispersion.

The concept of queues is quite intuitive. Assume that you're shopping in a supermarket. If many people proceed to checkout in a very short time interval, the cashier won't be able to handle the flow of customers instantaneously. Thus, a waiting line, or a *queue* will form. A cash desk is called a *server*. The ensemble composed of the queue and servers is called the *system*. Since people are shopping in a different way, the time spent during checkout isn't the same for each customer. Hence the distribution of the time spent at the cash desk will be called the *service time*. The time between each arrival at the cash desk is called the *inter-arrival time*. Finally, the number of people leaving the system at any time is called the *output process*.

The system can be characterized by the Kendall's notations. For the purpose of this thesis, we will restrict to the first three letters $a/s/C$. a represents the probability distribution of the inter-arrival times. s represents the probability distribution of the service-time. C represents the number of servers.

The simplest queue is the $M/M/1$ queue, where M stands for *Markov*. That is, both the inter-arrival times and the service time are characterized by an exponential distribution, while only one server can handle the queue.

This can be generalized by adding c servers, this will form an $M/M/c$ queue. This would be a first model that could fit with our supermarket example, since the supermarket would more likely employ more than one cashier.

The letter c can be infinite. In this case, an individual entering the system is instantaneously taken by a server. Hence no queue is formed. This type of system is called an $M/M/\infty$ queue. This is what happens in an insurance context. If you want to underwrite your insurance contract online, you can do it instantaneously. You don't have to wait for another person to finish its underwriting to make yours. It works as if there were an infinite number of virtual

general agents that can process your request instantaneously.

The advantage of the Markov assumption for the service time is its easy manipulation. However, this might not be realistic. Hence, it's sometimes suggested to transform the second M into a G , meaning *General*. That is, the service time distribution is not exponential anymore, but any distribution with positive support. For instance, an $M/G/c$ queue is a queue where the inter-arrival times are exponentially-distributed, the service time is distributed with any distribution G , while c servers can handle the flow of individuals entering the system. Again, the introduction of a general service time is consistent in an insurance context, as an exponential distribution is not likely to capture the termination behavior of policyholders. One simple counter-example is to recall that policyholders are more likely to cancel their policy by not renewing it, instead of cancelling it at some random time. Hence, the hazard function is expected to show some yearly peaks, and thus violates the exponential assumption.

Compiling what has been said, it is natural to retain a queue with infinite servers and with a general service time for an insurance application. The purpose of the next part is to determine the output process of an infinite server queue and to determine the number of busy servers at any time t . Three different types of arrival process are considered, that are homogeneous Poisson, inhomogeneous Poisson and Cox processes. The section 1.1.1 reminded the basics of Poisson processes properties, that will be used in order to detail the short proof given in [Mir63].

1.2.2 The model

The idea of applying an infinite server queue to model an insurance portfolio is quite intuitive. The arrival process will describe the way policyholders arrive in the portfolio. The service time will describe how long policyholders stay in the portfolio. Some interesting features would be to access the number of underwritten contracts in a given period, or to determine the number of insured policyholders at a given time, and finally to determine the number of terminated contracts during a given period.

This section introduces the main technical aspects of the infinite server queues. First, the main results of the infinite server queue with homogeneous Poisson process arrival are depicted. The proofs of the results are completely taken from [Mir63], as they will be adapted to the non-homogeneous case in the following section. Finally arrivals are adapted with Cox process, which allows for more flexibility in the model, but will result in a lack of generality as no closed formulas can be derived.

1.2.2.1 Preliminaries and notations

An $M/G/\infty$ queue is characterized by exponentially-distributed inter-arrival times, a general distribution for the service time, and an infinity of servers. Since there are infinity of servers, the number of elements in the queue is always zero. Hence, we need to study the number of elements in the system at any time, and the number of elements leaving the system during any time interval. All the notations and proofs are borrowed in their entirety from [Mir63]. The proof of [Mir63] will be detailed, since intermediary results and slight modifications will be needed for the adaptation to the inhomogeneous case.

Let t and T denote two real numbers such that $t > 0$ and $T > 0$. T represents the length of some time-interval. Let $\gamma(t+T)$ represent the number of elements in the system at $t+T$, and $\psi(t, T)$ represent the number of departures of the system between time t and $t+T$.

We assume that the system is initially empty at time $t = 0$. It means that at time $t = 0$, no policyholders are in the portfolio. Since $t = 0$ usually corresponds to the starting time of the analysis, this hypothesis is unverified. That's why it is interesting to distinguish between two types of policyholders: those who arrived before t , and those who arrived after t . The following results are valid for all policyholders that arrived after t , while for the ones who arrived before, it is necessary to determine their residual survival time. This will be the objective of section 4.2.1.

The inter-arrival time of elements in the system is characterized by an exponential distribution with mean $\frac{1}{\lambda}$. Denote by $N(t, t')$ the number of arrivals in the system between time t and t' . According to propositions 1.1.4 and 1.1.5, $N(t, t')$ is Poisson-distributed with parameter $\lambda(t' - t)$ and is independent of $N(s)$, $s < t$. Therefore, the number of arrivals between an interval of length T will be denoted $N(T)$.

The service time is characterized by a cumulative distribution function $H(x)$, with $x \geq 0$. The arrival process and the service time are supposed to be independent, and the service time of an individual is independent of the service time of all the other individuals.

Denote by:

- $p(t, T)$ the probability that an element that entered the system between $[0, t]$ leaves between $[t, t+T]$
- $q(t, T)$ the probability that an element that entered the system between $[0, t]$ leaves after $t+T$
- $r(t, T)$ the probability that an element that entered the system between $[t, t+T]$ leaves between $[t, t+T]$

- $s(t, T)$ the probability that an element that entered the system between $[t, t + T]$ leaves after $t + T$

Proposition 1.2.1. $p(t, T)$, $q(t, T)$, $r(t, T)$ and $s(t, T)$ satisfy:

$$p(t, T) = \frac{1}{t} \int_0^t (H(T + x) - H(x)) dx$$

$$q(t, T) = \frac{1}{t} \int_0^t (1 - H(T + x)) dx$$

$$r(t, T) = \frac{1}{T} \int_0^T H(x) dx$$

$$s(t, T) = \frac{1}{T} \int_0^T (1 - H(x)) dx$$

Hence, $r(t, T)$ and $s(t, T)$ satisfy $r(t, T) + s(t, T) = 1$

Proof. We will prove the result for $p(t, T)$ and $r(t, T)$. The proof is similar for $q(t, T)$ and $s(t, T)$. Consider an arrival in the system at time $s \in [0, t]$. According to proposition 1.1.6,

$$\mathcal{P}(s \mid N(t) = 1) = \frac{1}{t}$$

For this arrival at time s to leave between $[t, t + T]$, its service time must belong to $[t - s, t + T - s]$. The probability of such an event is $H(t + T - s) - H(t - s)$. The law of total probability gives:

$$p(t, T) = \int_0^t \frac{1}{t} (H(t + T - s) - H(t - s)) ds$$

Denote $x = t - s$, we get, $dx = -ds$. Substituting s by x in the previous equations gives:

$$p(t, T) = \frac{1}{t} \int_0^t (H(T + x) - H(x)) dx$$

Consider an arrival in the system at time $s \in [t, t + T]$. According to proposition 1.1.6,

$$\mathcal{P}(s \mid N(T) = 1) = \frac{1}{T}$$

For this arrival at time s to leave between $[t, t + T]$, its service time must belong to $[0, t + T - s]$. The probability of such an event is $H(t + T - s) - H(0) = H(t + T - s)$. The law of total probability gives:

$$r(t, T) = \int_t^{t+T} \frac{1}{T} H(t + T - s) ds$$

Denote $x = t + T - s$, we get, $dx = -ds$. Substituting s by x in the previous equations gives:

$$r(t, T) = \frac{1}{T} \int_0^T H(x) dx$$

□

The multinomial distribution will be used in the proof of the output process. It is a generalization of the binomial distribution.

Definition 1.2.1 (Multinomial distribution). *Assume an experiment where k outcomes are possible. The probability of each outcome i , $1 \leq i \leq k$ is $(p_i)_{1 \leq i \leq k}$. Assume that one repeats this experiment n times, each trial being independent of the others. Denote by $(N_i^n)_{1 \leq i \leq k}$ the random variables that represent the number of time the outcome i was drawn. Hence, the family $(N_i^n)_{1 \leq i \leq k}$ writes*

$$\sum_{i=1}^k N_i^n = n$$

. Then, the family of random variables $(N_i^n)_{1 \leq i \leq k}$ is said to follow a multinomial distribution with parameters n and $(p_i)_{1 \leq i \leq k}$.

Its distribution function satisfies:

$$\mathcal{P}(N_1^n = n_1 \cap \dots \cap N_k^n = n_k) = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k} \mathbf{1}_{(\sum_{i=1}^k n_i = n)}$$

1.2.2.2 Proof of the output process distribution

Let m and n be two integers. We will calculate the joint probability that m elements are in the system at time $t + T$ and that n elements leave the system between time t and $t + T$.

Proposition 1.2.2. *Consider an $M/G/\infty$ queue. The joint probability that m elements are in the system at time $t + T$ and that n elements leave the system between time t and $t + T$ writes:*

$$\begin{aligned} \mathcal{P}(\gamma(t+T) = m \cap \psi(t, t+T) = n) \\ = \frac{e^{-\lambda(sT+qt)} (\lambda(sT+qt))^m}{m!} \frac{e^{-\lambda(rT+pt)} (\lambda(rT+pt))^n}{n!} \end{aligned}$$

Hence, $\gamma(t+T)$ and $\psi(t, t+T)$ are independent random variables, Poisson-distributed, with parameters $\lambda(sT+qt)$ and $\lambda(rT+pt)$ respectively.

Proof. Let $k \in \mathbb{N}$ and $j \in \mathbb{N}$. Assume that k arrivals occurred between time 0 and t , that is $N(t) = k$, and that j arrivals occurred between time t and $t + T$, that is $N(T) = j$. Each of these k arrivals leaves:

- between $[0, t]$ with probability $1 - p(t, T) - q(t, T)$
- between $[t, t + T]$ with probability $p(t, T)$

- after $t + T$ with probability $q(t, T)$

Each of these j arrivals leaves:

- between $[t, t + T]$ with probability $r(t, T)$
- after $t + T$ with probability $s(t, T)$

If m elements are in the system at time $t + T$, then:

- $m - i$ of them are elements that arrived between $[0, t]$, that is, belong to the k elements introduced above
- i of them are elements that arrived between $[t, t + T]$, that is, belong to the j elements introduced above

Then i is an integer satisfying $0 \leq i \leq m$. If n elements leave the system between time t and $t + T$, then:

- necessarily, $j - i$ of them are elements that arrived between t and $t + T$ since the others i are still here. Hence $j - i \geq 0$
- the others $n - j + i$ come from arrivals between 0 and t . Hence $n - j + i \geq 0$

Hence, j is an integer satisfying $i \leq j \leq n + i$. Finally, if k elements arrived between 0 and t , that $m - i$ are still in the system at time $t + T$ and that $n - j + i$ of them have left between t and $t + T$, then k must satisfy $(n - j + i) + (m - i) \leq k$. Mechanically, the $k - ((n - j + i) + (m - i))$ elements left from arrivals between 0 and t have left between 0 and t .

- Let $\gamma_1(t + T)$ be the number of elements in the system at $t + T$ from arrivals between 0 and t
- Let $\gamma_2(t + T)$ be the number of elements in the system at $t + T$ from arrivals between t and $t + T$
- Let $\psi_1(t, t + T)$ be the number of departures between t and $t + T$ from arrivals between 0 and t
- Let $\psi_2(t, t + T)$ be the number of departures between t and $t + T$ from arrivals between t and $t + T$

One can write $\gamma_2(t + T) + \psi_2(t, t + T) = N(T)$. That is, conditional on $N(T)$, the knowledge of $\gamma_2(t + T)$ implies the knowledge of $\psi_2(t, t + T)$. Hence, conditional on $N(T)$, it's necessary to determine $\mathcal{P}(\gamma_2(t, t + T) | N(T))$ to completely characterize the arrivals between t and $t + T$. Similarly, conditional on $N(t)$, the knowledge of both $\gamma_1(t + T)$ and $\psi_1(t, t + T)$ completely characterizes the arrivals between 0 and t .

Then, the law of total probability gives:

$$\begin{aligned}
& \mathcal{P}(\gamma(t+T) = m \cap \psi(t, t+T) = n) \\
&= \sum_{i=0}^m \sum_{j=i}^{n+i} \sum_{k=m+n-j}^{+\infty} \left[\mathcal{P}(\gamma_2 = i, \gamma_1 = m - i, \psi_1 = n - j + i \mid N(t) = k, N(T) = j) \right. \\
&\quad \left. \times \mathcal{P}(N(t) = k, N(T) = j) \right]
\end{aligned} \tag{1.2.1}$$

The next steps will help to conclude on equation 1.2.1. The following equations have been highlighted in different colors to appropriately spot the terms that will vanish and that will be involved in the different sums. Namely, the **orange** terms will be involved in the sum indexed on k , the **magenta** terms will be involved in the sum indexed on j , the **cyan** terms will be involved in the sum indexed on i , and the **olive** terms will vanish.

Proposition 1.1.5 leads to:

$$\begin{aligned}
\mathcal{P}(N(t) = k, N(T) = j) &= \mathcal{P}(N(t) = k) \mathcal{P}(N(T) = j) \\
&= \frac{e^{-\lambda t} (\lambda t)^k}{k!} \frac{e^{-\lambda T} (\lambda T)^j}{j!}
\end{aligned} \tag{1.2.2}$$

Two simplifications can be made:

- The number of elements in the system at $t+T$ from arrivals between t and $t+T$ is independent from the arrivals between 0 and t . Namely, $\gamma_2(t+T)$ is independent from $N(t)$
- The number of elements in the system at $t+T$ from arrivals between 0 and t , and the number of departures between t and $t+T$ from arrivals between 0 and t are independent from the arrivals between t and $t+T$. Namely, $\gamma_1(t+T)$ and $\psi_1(t, t+T)$ are independent from $N(T)$

Hence, one can write:

$$\begin{aligned}
& \mathcal{P}(\gamma_2 = i, \gamma_1 = m - i, \psi_1 = n - j + i \mid N(t) = k, N(T) = j) \\
&= \mathcal{P}(\gamma_2 = i \mid N(T) = j) \times \mathcal{P}(\gamma_1 = m - i, \psi_1 = n - j + i \mid N(t) = k)
\end{aligned}$$

The probability that an element is still in the system at $t+T$ conditional on its arrival between 0 and t is characterized by a Bernoulli distribution with parameter $r(t, T)$. Hence, considering j elements that arrived between t and $t+T$, the probability that i of them are still in the system between at $t+T$ is characterized by a binomial distribution with parameters j and $s(t, T)$. Thus,

$$\begin{aligned}
\mathcal{P}(\gamma_2 = i \mid N(T) = j) &= \frac{j!}{i!(j-i)!} s^i (1-s)^{j-i} \\
&= \frac{j!}{i!(j-i)!} s^i r^{j-i} \\
&= \frac{j!}{i!(j-i)!} \frac{(sT)^i (rT)^{j-i}}{T^j}
\end{aligned} \tag{1.2.3}$$

Similarly, considering k elements that arrived between 0 and t , the probability that $m-i$ of them are still in the system at $t+T$ and that $n-j+i$ of them have left between t and $t+T$ is characterized by a trinomial distribution with parameters k , $p(t, T)$ and $q(t, T)$. Thus, according to definition 1.2.1,

$$\begin{aligned}
&\mathcal{P}(\gamma_1 = m-i, \psi_1 = n-j+i \mid N(t) = k) \\
&= \frac{k!}{(m-i)!(n-j+i)!(k-m-n+j)!} q^{m-i} p^{n-j+i} (1-p-q)^{k-m-n+j} \\
&= \frac{k!}{(m-i)!(n-j+i)!(k-m-n+j)!} \frac{(qt)^{m-i} (pt)^{n-j+i}}{t^{m+n-j}} (1-p-q)^{k-m-n+j}
\end{aligned} \tag{1.2.4}$$

First, let's work on the sum indexed on k . Only equations 1.2.2 and 1.2.4 involve k , and the associated terms have been highlighted in **orange**. Isolating the elements involving k in those equations, one can write:

$$\begin{aligned}
&\sum_{k=m+n-j}^{+\infty} \frac{(\lambda t)^k}{k!} \times \frac{k!}{(k-m-n+j)!} (1-p-q)^{k-m-n+j} \\
&= \sum_{k=m+n-j}^{+\infty} (\lambda t)^{m+n-j} \frac{(\lambda t)^{k-m-n+j}}{(k-m-n+j)!} (1-p-q)^{k-m-n+j} \\
&= (\lambda t)^{m+n-j} \sum_{k=0}^{+\infty} \frac{(\lambda t)^k}{k!} (1-p-q)^k \\
&= \lambda^{m+n-j} t^{m+n-j} e^{\lambda t} e^{-\lambda t(p+q)}
\end{aligned} \tag{1.2.5}$$

The terms in t^{m+n-j} from equations 1.2.4 and 1.2.5 vanish. The terms in $e^{\lambda t}$ from equations 1.2.2 and 1.2.5 vanish. The terms in T^j and $j!$ from equations 1.2.2 and 1.2.3 vanish. Finally, the terms in λ^{m+n-j} in equation 1.2.5 and in λ^j from equation 1.2.2 lead to a term in λ^{m+n} .

Let's now work on the sum indexed on j . Equations 1.2.2, 1.2.3 and 1.2.4 involve j , and the associated terms have been highlighted in **magenta**. One gets:

$$\begin{aligned}
\sum_{j=i}^{n+i} \frac{1}{j!} \times \frac{j!}{(j-i)!} (rT)^{j-i} \frac{(pt)^{n-j+i}}{(n-j+i)!} &= \sum_{j=i}^{n+i} \frac{n!}{n!} \times \frac{(rT)^{j-i}}{(j-i)!} \frac{(pt)^{n-j+i}}{(n-j+i)!} \\
&= \sum_{j=0}^n \frac{n!}{n!} \times \frac{(rT)^j}{j!} \frac{(pt)^{n-j}}{(n-j)!} \\
&= \frac{1}{n!} \times (rT + pt)^n
\end{aligned} \tag{1.2.6}$$

To conclude the proof, let's now work on the sum indexed on i . Equations 1.2.3 and 1.2.4 involve i , and the associated terms have been highlighted in cyan. One gets:

$$\begin{aligned}
\sum_{i=0}^m \frac{(sT)^i}{i!} \times \frac{(qt)^{m-i}}{(m-i)!} &= \sum_{i=0}^m \frac{m!}{m!} \times \frac{(sT)^i}{i!} \times \frac{(qt)^{m-i}}{(m-i)!} \\
&= \frac{1}{m!} \times (sT + qt)^m
\end{aligned} \tag{1.2.7}$$

Grouping the results from equations 1.2.5, 1.2.6 and 1.2.7, one gets the following result:

$$\begin{aligned}
&\mathcal{P}(\gamma(t+T) = m \cap \psi(t, t+T) = n) \\
&= e^{-\lambda T} \lambda^{m+n} e^{-\lambda t(p+q)} \frac{(rT + pt)^n}{n!} \frac{(sT + qt)^m}{m!} \\
&= e^{-\lambda T(r+s)} \lambda^{m+n} e^{-\lambda t(p+q)} \frac{(rT + pt)^n}{n!} \frac{(sT + qt)^m}{m!} \\
&= \frac{e^{-\lambda(sT+qt)} (\lambda(sT + qt))^m}{m!} \frac{e^{-\lambda(rT+pt)} (\lambda(rT + pt))^n}{n!}
\end{aligned} \tag{1.2.8}$$

□

1.2.2.3 Adaptation to the non-homogeneous case

The assumption of a constant arrival rate is often false for some applications. Many systems show that the arrival rate varies with time. For instance, the subway is more likely to be full during peak hours, and not so full at any other time. This variation of people taking the subway could be well captured by a varying arrival rate.

$M/G/\infty$ queues with a varying arrival rate are called $M_t/G/\infty$ queues. In those conditions, we will adapt proposition 1.2.2 and its proof, and see that the result is nearly unchanged.

Let's first adapt the proposition 1.2.1.

Proposition 1.2.3. *In the non-homogeneous case, $p(t, T)$, $q(t, T)$, $r(t, T)$ and $s(t, T)$ satisfy:*

$$\begin{aligned} p(t, T) &= \frac{1}{\Lambda(t)} \int_0^t \lambda(t-x) (H(T+x) - H(x)) dx \\ q(t, T) &= \frac{1}{\Lambda(t)} \int_0^t \lambda(t-x) (1 - H(T+x)) dx \\ r(t, T) &= \frac{1}{\Lambda(t+T) - \Lambda(t)} \int_0^T \lambda(t+T-x) H(x) dx \\ s(t, T) &= \frac{1}{\Lambda(t+T) - \Lambda(t)} \int_0^T \lambda(t+T-x) (1 - H(x)) dx \end{aligned}$$

Hence, $r(t, T)$ and $s(t, T)$ satisfy $r(t, T) + s(t, T) = 1$

Proof. We will prove the result for $p(t, T)$ and $r(t, T)$. The proof is similar for $q(t, T)$ and $s(t, T)$. Consider an arrival in the system at time $s \in [0, t]$. According to proposition 1.1.8,

$$\mathcal{P}(s \mid N([0, t]) = 1) = \frac{\lambda(s)}{\Lambda(t)}$$

For this arrival at time s to leave between $[t, t+T]$, its service time must belong to $[t-s, t+T-s]$. The probability of such an event is $H(t+T-s) - H(t-s)$. The law of total probability gives:

$$p(t, T) = \int_0^t \frac{\lambda(s)}{\Lambda(t)} (H(t+T-s) - H(t-s)) ds$$

Denote $x = t - s$, we get, $dx = -ds$. Substituting s by x in the previous equations gives:

$$p(t, T) = \frac{1}{\Lambda(t)} \int_0^t \lambda(t-x) (H(T+x) - H(x)) dx$$

Consider an arrival in the system at time $s \in [t, t+T]$. According to proposition 1.1.8,

$$\mathcal{P}(s \mid N([t, t+T]) = 1) = \frac{\lambda(s)}{\Lambda(t+T) - \Lambda(t)}$$

For this arrival at time s to leave between $[t, t+T]$, its service time must belong to $[0, t+T-s]$. The probability of such an event is $H(t+T-s) - H(0) = H(t+T-s)$. The law of total probability gives:

$$r(t, T) = \int_t^{t+T} \frac{\lambda(s)}{\Lambda(t+T) - \Lambda(t)} H(t+T-s) ds$$

Denote $x = t+T-s$, we get, $dx = -ds$. Substituting s by x in the previous equations gives:

$$r(t, T) = \frac{1}{\Lambda(t+T) - \Lambda(t)} \int_0^T \lambda(t+T-x)H(x)dx$$

□

The objective of the next proposition is to generalize proposition 1.2.2. The desired output is the one depicted in [DQLD19].

Proposition 1.2.4. *Consider an $M_t/G/\infty$ queue. Let*

$$\mu_i(t, T) = s(t, T) (\Lambda(t+T) - \Lambda(t)) + q(t, T)\Lambda(t)$$

$$\mu_o(t, T) = r(t, T) (\Lambda(t+T) - \Lambda(t)) + p(t, T)\Lambda(t)$$

The joint probability that m elements are in the system at time $t+T$ and that n elements leave the system between time t and $t+T$ writes:

$$\mathcal{P}(\gamma(t+T) = m \cap \psi(t, t+T) = n) = \frac{e^{-\mu_i(t, T)} (\mu_i(t, T))^m}{m!} \frac{e^{-\mu_o(t, T)} (\mu_o(t, T))^n}{n!}$$

Hence, $\gamma(t+T)$ and $\psi(t, t+T)$ are independent random variables, Poisson-distributed, with parameters $\mu_i(t, T)$ and $\mu_o(t, T)$ respectively.

Proof. The proof is similar to the one derived for proposition 1.2.2. Hence, we will simply modify the equations 1.2.2, 1.2.3 and 1.2.4.

Equation 1.2.2 can be rewritten as follows:

$$\begin{aligned} & \mathcal{P}(N([0, t]) = k, N([t, t+T]) = j) \\ &= \mathcal{P}(N([0, t]) = k) \mathcal{P}(N([t, t+T]) = j) \\ &= \frac{e^{-\Lambda(t)} (\Lambda(t))^k}{k!} \frac{e^{-(\Lambda(t+T) - \Lambda(t))} (\Lambda(t+T) - \Lambda(t))^j}{j!} \end{aligned} \quad (1.2.9)$$

Equation 1.2.3 can be rewritten as follows:

$$\begin{aligned} & \mathcal{P}(\gamma_2 = i \mid N([t, t+T]) = j) \\ &= \frac{j!}{i! (j-i)!} s^i (1-s)^{j-i} \\ &= \frac{j!}{i! (j-i)!} s^i r^{j-i} \\ &= \frac{j!}{i! (j-i)!} \frac{(s(\Lambda(t+T) - \Lambda(t)))^i (r(\Lambda(t+T) - \Lambda(t)))^{j-i}}{(\Lambda(t+T) - \Lambda(t))^j} \end{aligned} \quad (1.2.10)$$

Equation 1.2.4 can be rewritten as follows:

$$\begin{aligned}
& \mathcal{P}(\gamma_1 = m - i, \psi_1 = n - j + i \mid N(t) = k) \\
&= \frac{k!}{(m - i)!(n - j + i)!(k - m - n + j)!} q^{m-i} p^{n-j+i} (1 - p - q)^{k-m-n+j} \\
&= \frac{k!}{(m - i)!(n - j + i)!(k - m - n + j)!} \\
&\times \frac{(q\Lambda(t))^{m-i} (p\Lambda(t))^{n-j+i}}{\Lambda(t)^{m+n-j}} (1 - p - q)^{k-m-n+j}
\end{aligned} \tag{1.2.11}$$

Performing the derivation of equation 1.2.1 will lead to result in the exact same way. \square

1.2.2.4 An extension to the Cox process

As for now, no closed formula can be obtained if the arrival process is a Cox process. Hence, the generalization of the queuing framework to Cox arrival processes is only possible through numerical approximations.

That is, one first needs to simulate the Cox arrival process. Then for each arrival, a service time is simulated according to the service time distribution. Then, the number of people in the system can be estimated by counting the number of individuals in the system at any time t . The departure process can be estimated by counting the number of departures inside any time interval.

In this study, this numerical approach will be retained. However, the numerical results will be compared to the theoretical ones for the arrival process only, as they are the only ones available.

Chapter 2

Estimation

In this section, the general framework to fit the data is presented. Specific applications and examples will be presented and derived in section 3.

2.1 The data

The purpose of this model is to represent the behavior of a policyholder inside an insurance portfolio. The insurance portfolio will be the *system*. The underwritten time of a policy will be associated to the *arrival time* in the system. Since no particular "waiting time" is required to underwrite a contract, the use of an infinite server queue is justified. The time spent in the portfolio until termination of the contract will be associated to the *service time*. All of our policyholders are assumed to underwrite their policy independently and at disjoint points in time. Moreover, their service time is assumed to be independent, and independent of their arrival time. Under those conditions, the model described above applies.

Such independence assumptions enable to fit the service time distribution and the arrival time process independently.

The theory introduced above would require the knowledge of each arrival time, provided that these arrival times do not match. This hypothesis is violated in our data since the underwritten date is rounded at the day the contract was signed, and not at the exact time. Hence, even though reality could be represented by a point process, our data show an accumulation of points at certain dates since many policyholders have signed their contract at the exact same day. This accumulation issue has been addressed in quite a few research papers. One turnaround is to consider arrivals in bulks (or batches). The reader could refer to the following papers for further readings on this subject [DP19], [Sha66] and [PW12].

However, our data are not real batch data, since the probability that two policyholders underwrote their contracts at the exact same time is negligible. Thus, to overcome this issue, we will fit our model considering block data, for instance using daily data, and we will count the number of arrivals during each time period.

2.2 Estimating the intensity function for the arrival process in the homogeneous and inhomogeneous cases

Let $[0, T]$ be the observed period, and let $0 = t_0 < t_1 < \dots < t_n = T$ be n disjoint points in time. Let k_1, \dots, k_n be the number of arrivals observed during the period $[t_i, t_{i+1}[$.

Let $\mathcal{L}(\mathbf{p})$ denote the likelihood function, and \mathbf{p} be the parameters of the cumulative intensity function $\Lambda(t)$. According to definition 1.1.4, one gets:

$$\begin{aligned} \mathcal{L}(k_1, \dots, k_n, \mathbf{p}) &= \mathcal{P}(\cap_{i=1}^n \{N([t_{i-1}, t_i[= k_i\})\}) \\ &= \prod_{i=1}^n \mathcal{P}(N[t_{i-1}, t_i[= k_i) \\ &= \prod_{i=1}^n e^{-\Lambda(t_{i-1}, t_i)} \frac{(\Lambda(t_{i-1}, t_i))^{k_i}}{k_i!} \end{aligned}$$

The second equality is obtained using the independence of the events, see definition 1.1.4.

The log-likelihood writes:

$$\ln(\mathcal{L}(k_1, \dots, k_n, \mathbf{p})) = \sum_{i=1}^n [-\Lambda(t_{i-1}, t_i) + k_i \ln(\Lambda(t_{i-1}, t_i)) - \ln k_i!]$$

Under certain regularity assumptions on Λ with respect to \mathbf{p} , the optimal parameters will be obtained by minimizing the negative log-likelihood, that is by solving:

$$\nabla_{\mathbf{p}}(-\ln(\mathcal{L}(k_1, \dots, k_n, \mathbf{p}))) = 0$$

2.3 Estimating the subordinator parameters for the arrival process in the Cox case

We use the same notations as the ones introduced in 2.2. We will consider the non-stationary case, with the time-dependent shape function. Let $\alpha_{\mathbf{p}}(t)$ be

the shape function, b the scale parameter, and \mathbf{p} the parameters of the shape function. According to proposition 1.1.11, the likelihood writes:

$$\begin{aligned}\mathcal{L}(k_1, \dots, k_n, \mathbf{p}, b) &= \mathcal{P}(\cap_{i=1}^n \{N([t_{i-1}, t_i[= k_i]\}) \\ &= \prod_{i=1}^n \mathcal{P}(N[t_{i-1}, t_i[= k_i) \\ &= \prod_{i=1}^n \frac{\Gamma(k_i + \int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt)}{k_i! \cdot \Gamma(\int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt)} \left(\frac{b}{b+1}\right)^{k_i} \left(\frac{1}{b+1}\right)^{\int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt}\end{aligned}$$

The log-likelihood writes:

$$\begin{aligned}\ln(\mathcal{L}(k_1, \dots, k_n, \mathbf{p}, b)) &= \sum_{i=1}^n \left(\ln \left(\Gamma \left(k_i + \int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt \right) \right) - \ln \left(\Gamma \left(\int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt \right) \right) - \ln(k_i!) \right) \\ &+ \sum_{i=1}^n \left(k_i \ln(b) - k_i \ln(b+1) - \ln(b+1) \int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt \right)\end{aligned}$$

The optimal parameters will be obtained by minimizing the negative log-likelihood, that is by solving:

$$\nabla_{\mathbf{p}}(-\ln(\mathcal{L}(k_1, \dots, k_n, \mathbf{p}, b))) = 0$$

2.4 Estimating the survival function

The estimation of the survival function lies on the standard techniques used in life insurance for instance. Indeed, our portfolio comprises policyholders who have cancelled their insurance policy during the observation period, while others have not. Hence, we face right-censored data.

The standard non-parametric estimator of Kaplan-Meier can be used to fit the survival function. The inconvenience of the Kaplan-Meier is that it cannot be used for extrapolation purposes. This is a major issue, since it is mandatory for our model to know the exact departure time of each policyholder, in order to accurately estimate the departure process.

Two workarounds can be used to solve this issue:

- Use a fully parametric model
- Use a semi-parametric model

2.4.1 The fully parametric model

Let $H(x)$ denote the survival function, $h(x)$ the associated distribution function, and \mathbf{q} the parameters of this distribution. The event characterized by the observation of the cancellation is modeled by the indicator function δ_i , where $\delta_i = 1$ if the event is observed for individual i , and $\delta_i = 0$ if the data are censored.

Let X_i be the survival time of individual i provided that the event has been observed, otherwise $X_i = T - u_i$, where u_i denotes the underwriting time for individual i .

For an individual i , the probability of the realization (x_i, δ_i) writes:

$$\mathcal{L}(x_i, \delta_i) = h(x_i)^{\delta_i} H(x_i)^{1-\delta_i}$$

Indeed, if the event is observed ($\delta_i = 1$), then the probability to show the event at time x_i is $h(x_i)$. If the event is not observed ($\delta_i = 0$), we only know that the individual has survived at least x_i , the probability being $S(x_i)$ in that case.

In our case, we have assumed that all our individuals are independent, hence the likelihood writes:

$$\begin{aligned} \mathcal{L}((x_1, \delta_1), \dots, (x_n, \delta_n), \mathbf{q}) &= \prod_{i=1}^n h(x_i)^{\delta_i} H(x_i)^{1-\delta_i} \\ &= \prod_{i=1}^n \lambda(x_i)^{\delta_i} H(x_i) \end{aligned}$$

where $\lambda(x) = \frac{h(x)}{H(x)}$ is called the hazard function, and models the probability to show the event at time $x + dx$ provided that the individual survived up to x .

The negative log-likelihood writes:

$$-\ln(\mathcal{L}((x_1, \delta_1), \dots, (x_n, \delta_n), \mathbf{q})) = -\sum_{i=1}^n [\ln(H(x_i)) + \delta_i \ln(\lambda(x_i))]$$

Under certain regularity assumptions on H and λ with respect to \mathbf{q} , the optimal parameters will be obtained by minimizing the negative log-likelihood, that is by solving:

$$\nabla_{\mathbf{q}}(-\ln(\mathcal{L}((x_1, \delta_1), \dots, (x_n, \delta_n), \mathbf{q}))) = 0$$

2.4.2 The semi-parametric model

In some cases, it might be convenient to use a semi-parametric model to fit the survival data. By semi-parametric, it should be understood a non-parametric survival curve for survival between 0 and some time t , and a parametric survival curve for survival longer than t . Constraints are added in order to ensure the

continuity of the survival curve.

Such solutions have been studied in the medical field in [GGC93]. This article depicts the most simple way to extrapolate non-parametric survival curves with parametric survival curves. The solution has also been formalized more in details in this article [GTP13].

The main idea consists in setting a threshold $s > 0$, and fitting a non-parametric survival curve on $[0, s[$, and fitting a parametric survival curve on $[s, +\infty[$. This threshold s can be set using expert judgements, or using statistical tests, as derived in [GTP13]. The reader is invited to refer to this article for further development on this subject, as the theory won't be reported here.

Formally, this model consists of estimating the survival function as follows:

$$H(t) = \begin{cases} H_{\text{KP}}(t) & \text{for } t < s \\ \frac{H_{\text{KP}}(s)}{H_p(s)} H_p(t) & \text{for } t \geq s \end{cases} \quad (2.4.1)$$

where, H_{KP} denotes the Kaplan-Meier estimator, and H_p denotes the parametric estimator.

In practice, the parametric estimator is estimated the same way as in subsection 2.4.1. This semi-parametric approach will be the one retained in our model. Indeed, the reader will see that the shape of the survival function doesn't exhibit the same behavior for short survival and long survival. The transition is smooth enough to properly fit the parametric tail of the survival curve, which justifies its use.

Chapter 3

Inference and simulation

3.1 Presentation of the data

3.1.1 General presentation

To illustrate the theoretical concepts introduced in the first sections, we will use some home insurance data to fit the model. This database has been anonymized: none of the following figures and results can be used for other purposes than illustrating the model.

The data basis comprises 187,992 contracts underwritten between 01/01/2016 and 01/10/2020.

The data basis contains 4 fields:

The contract number which is unique across our data base.

The underwriting date of the contract Note that it corresponds here to the date of effect of the contract, that is, the date at which the contract is at risk.

The cancellation date of the contract This date can either be a real date, if the cancellation event occurs, or **NA** if we face right-censored data. Same as above, this date is the day at which the contract is not at risk anymore.

The contract annual premium The annual premium is assumed to be constant over the insurance period. This is a strong hypothesis that will be discussed in section 5.

As introduced above, working with a finite time-window leads to right censored data. That is, the cancellation date must not be known for a policyholder that might cancel his policy after the last observation date. Hence, we have added two columns to our data:

The contract duration in days It is defined as $\min(C_i, T_i)$, where C_i denotes the observation time of insured i , and T_i the underwritten time of insured i .

A censoring indicator This is a boolean, returning 1 if the cancellation occurs within the time frame, and 0 if the cancellation of the contract occurs after the last observation date

3.1.2 The number of contracts

Figure 3.1 shows the number of underwritten contracts per day in the portfolio. Some patterns appear. First, it can be appointed from the peaks, that most contracts are effective at the beginning of each month (on the first day), which is quite common in an insurance context. Second, the peaks of both months of May and November seem lower than the ones from other months. One interesting point is that the period going from beginning of 2016 to mid-2017 has a lower underwriting activity than the period going from 2018 to 2020. The transition between both periods seems to be linear (from mid-2017 to beginning of 2018). This could translate some efforts made by the insurer to increase the underwriting rate.

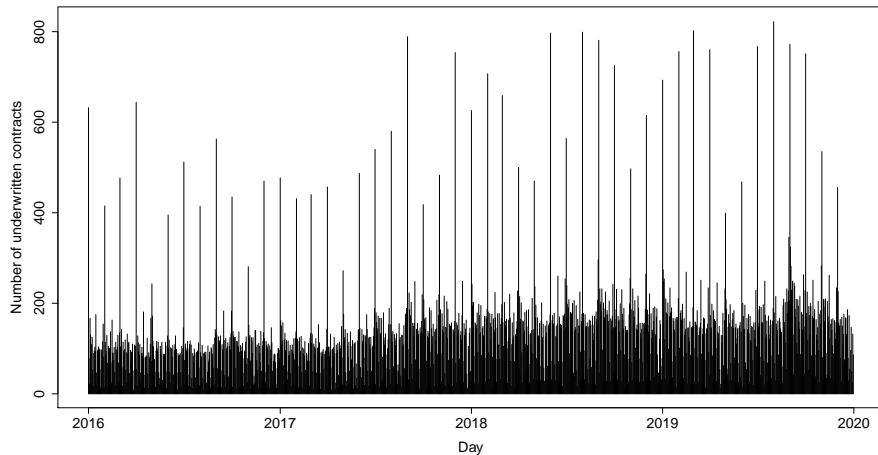


Figure 3.1: Number of underwritten contracts per day

For consistency purposes, it is preferable to work with as-if data. This process requires to transform the data before 2018 to data that look like 2018 and after data. In order to know how the data have evolved over time, it could be interesting to look at the moving average of the number of underwritings per day. Let k , k even, be the width of the rolling window, and $(N_i)_{1 \leq i \leq n}$ be the

number of underwritings per day. Then the rolling average is defined, for $p \geq k$, by:

$$r_p = \frac{1}{k} \sum_{i=p-\frac{k}{2}+1}^{p+\frac{k}{2}} N_i$$

The choice of k is critical. Indeed, some very local variations can be observed in the data set (peaks at the first day of each month, low activity on Sundays, etc.). Similarly, working with a large k would lead to an inaccurate smoothing. Let's choose $k = 100$.

Figure 3.2 shows the rolling window in red. One can see that the linear trend seems to start in May 2017, and that it ends in October 2017. Those two thresholds have been plotted in red on figure 3.2. Recall however that the rolling mean introduces a lag of $\frac{k}{2}$, that is, the red curve is shifted by 50 units to the right. For simplicity purposes, we will then assume that the linear trend starts in March 2017, and ends in August 2017.

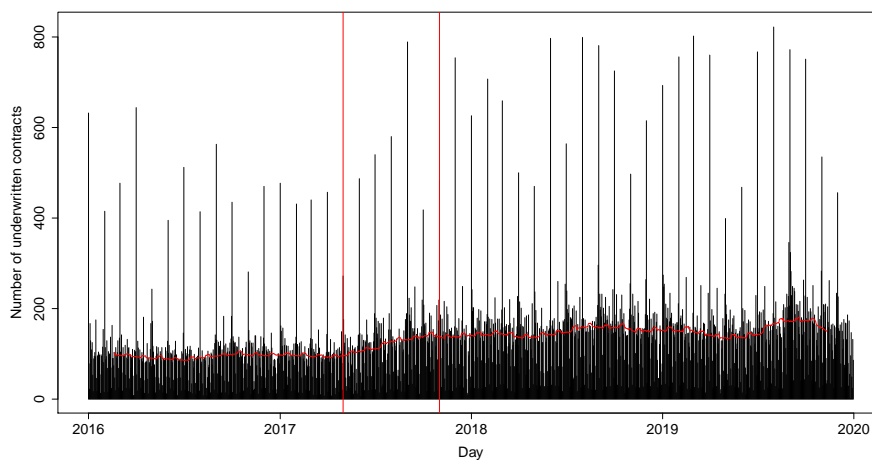


Figure 3.2: Number of underwritten contracts per day and rolling mean with $k = 100$

The idea is then to adjust the number of underwritings to have homogeneous historical data. Let m_1 be the average number of contracts underwritten between 01/01/2016 and 28/02/2017, and m_2 be the number of underwritten contracts between 01/09/2017 and 31/12/2019. Then the adjusted number of contracts between 01/01/2016 and 28/02/2017 will be the original number of contracts multiplied by $\frac{m_2}{m_1}$. Similarly, the adjusted number of contracts between

01/03/2017 and 31/08/2017 will be the original number adjusted by the factor

$$\left(1 - \frac{m_2}{m_1}\right)x + \frac{m_2}{m_1}$$

with $x = 0$ being 28/02/2017 and $x = 1$ being 01/09/2017. The corrected underwriting numbers are depicted in figure 3.3.

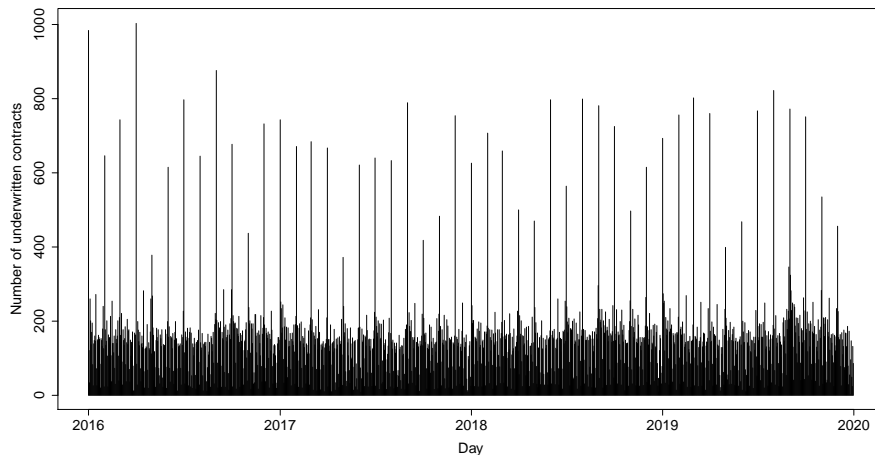


Figure 3.3: Number of adjusted underwritten contracts per day

Finally, looking closer at a specific year through figure 3.4, it seems that the number of underwritten contracts per day of week is quite constant through all business days, with a lower rate during the weekends.

Unsurprisingly, much lower activity is recorded on Saturdays, and even less on Sundays. One could take into account such details for fitting the model, since specifying an intensity function or shape function that could match such patterns would perfectly work. However, one should take care about over-parametrization, which leads to difficult parameter estimations, or over-fitting problems.

Some turnaround would be to work with aggregated, say monthly or yearly data. Indeed, all daily effects depicted in figure 3.4 will completely vanish, as one week approximatively shows the same patterns as another. However, we would only have very few points to fit the model, which would lead to high estimation error. Having this in mind, and for the modeling purpose, it is preferable to work with individual data. It will enable to have much more insight on how customers behave throughout the year. Individual data are getting more and

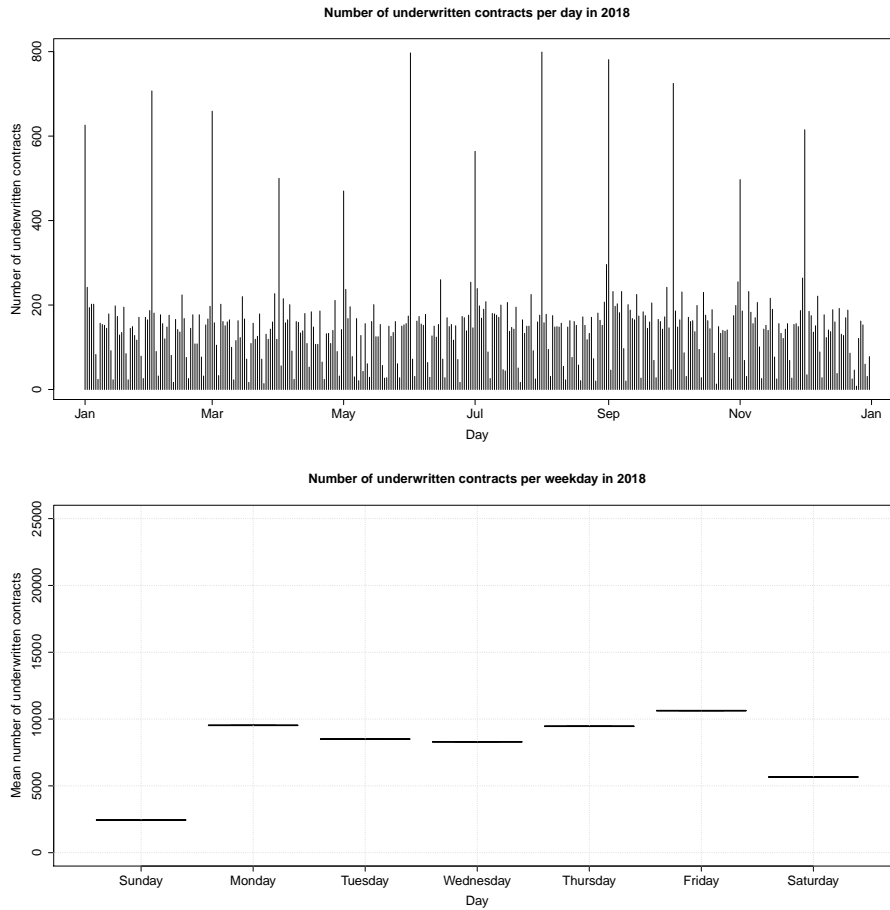


Figure 3.4: Number of underwritten contracts in 2018

more popular, since taking advantage of all available data enables to capture all the information available on the risk. Such a practice is widespread on the claim side, with for instance, individual claim reserving. See for instance [BD17].

Hence, for training the model, we will retain the time frame going from 01/01/2016 to 31/12/2019, that is, 4 years of historical data.

Figures 3.5, and 3.6 respectively show the number of terminated contracts per month, and the number of contracts in the portfolio per day. The number of terminated contracts per day also shows peaks during the first day of each month, which are much higher than the ones on the underwriting side. This can be explained by the following arguments: a policyholder is more likely to cancel his policy on a first day of the month even if he didn't underwrite his contract

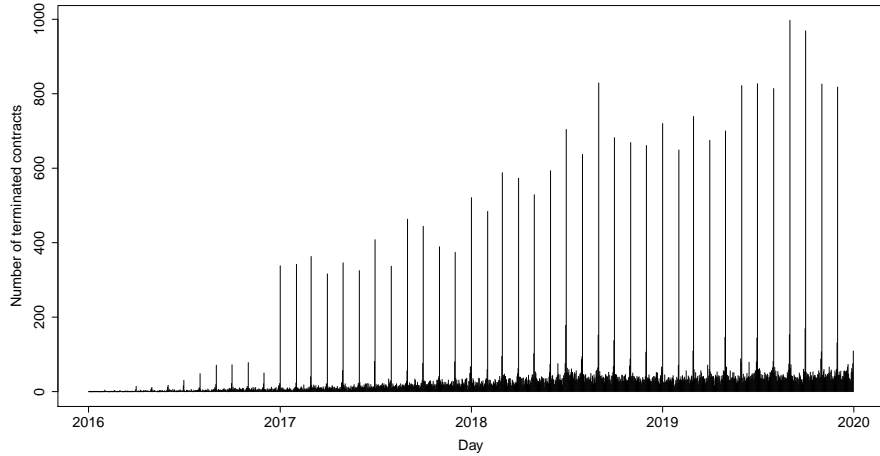


Figure 3.5: Number of terminated contracts per day

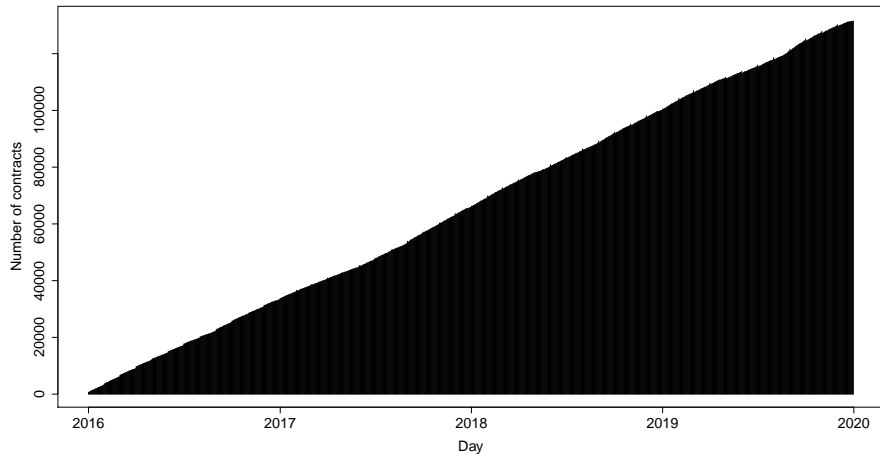


Figure 3.6: Number of contracts per day

on the first day of a month, because premiums are usually paid on a calendar month basis. The idea here is that if a policyholder underwrites a contract, say on 11/09/N, then he will pay $\frac{2}{3}$ of his monthly premium for the month of September. His underlying policy anniversary date will be set on 01/09/N.

Thus, in this thesis, we will call the *anniversary date of a contract* the first day of the month at which the contract was underwritten.

It is possible to create two additional columns in the dataset:

The corrected termination date Let d denote the day at which the contract has been underwritten. The corrected termination date is defined by the actual termination date, to which we add $d-1$. The purpose of this field is to properly identify the policyholders who have cancelled their contracts at their anniversary date, by removing the bias introduced by leap years (except for people who underwrote their contracts on February 29, whose number is negligible). Indeed, to identify those people, it just suffices to check the equality between the underwriting day and the corrected termination day, and the underwriting month and the corrected termination month.

An indicator for anniversary date cancellation It is a direct consequence of the previous point, now that those people can be easily identified.

Going back to figures 3.3 and 3.5, it seems that cancellation peaks are much less volatile than underwriting peaks. That is, the survival curve plays the role of a smoother.

Finally, the number of contracts in the portfolio per day seems to vanish all the variance of the underwritten and termination processes. Indeed, quite a low variability can be observed, and all patterns have completely disappeared from the plot. In that case, it will be interesting to see if the Cox process, used to capture the overdispersion on the underwriting process, has any impact on both the output process and the number of policyholders at risk in the portfolio.

3.1.3 Survival data

Survival data refer, for each policyholder, to the time at risk in the portfolio. It can be known if cancellation is observed during the observation period, or right-censored if such an event has not been observed during the observation period.

Some standard way to get information about survival curves, is to use the Kaplan-Meier estimator, see [KM58]. Often used in life insurance, its use is getting more and more widespread, even in the non-life area. In the context of customer lifetime in an insurance portfolio, survival data have been used by both [BCP16] and [Wan15].

[BCP16] has focused on a fully parametric approach, from which we will borrow the form of the survival function, to fit the tail of our semi-parametric survival curve. On the other side, [Wan15] has focused on using a Cox proportional hazard model, to capture how survival differs on contracts characteristics. Such a study could be very valuable, and will be discussed in section 5. However, for the purpose of this thesis, we will stick to a simple framework, as introduced

in 2.4.2. Complexity could be added in further research on this topic, by taking into account covariates.

Figure 3.7 depicts the overall survival curve, as estimated by the Kaplan-Meier estimator. On this plot, one can see 31 curves, each corresponding to the survival curve for the group of people having underwritten their contracts on day i of the month, $1 \leq i \leq 31$. Those curves reinforce the idea behind which a policyholder who didn't underwrite his contract on the first day of a month, is more likely to cancel his contract the first day of a month rather than an other day.

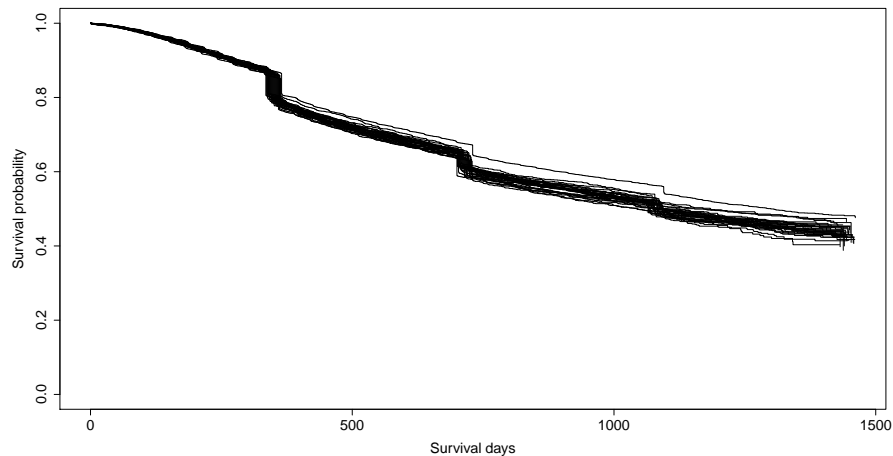


Figure 3.7: Kaplan-Meier estimator by day of underwritings

Two important information can be taken away from this chart:

- One can see vertical steps: usually contracts are annual contracts, hence all people who didn't renew their contract at their anniversary date are gathered in all those vertical steps. This is referred as non-renewal in [Wan15].
- On the opposite side, policyholders are free to cancel their contract at any time they want, which corresponds to the smooth parts of the chart. This is referred as cancellation in [Wan15].

Surprisingly, the smooth part of the first year of survival doesn't show the same pattern as the other smooth parts, as it is concave instead of convex. This particular shape motivates the use of a semi-parametric survival curve, as a fully parametric approach would imply overestimating the cancellation rate.

Both cancellation and non-renewal are well-captured by [BCP16], through their parametric survival function:

$$S(x) = e^{-\gamma x} p^{\lfloor x \rfloor} \tag{3.1.1}$$

In this approach, x represents the duration of the contract in years. The term p represents the shock that is applied to the survival function as soon as the integer part of x changes, that is, at each anniversary date of the contract. This is the non-renewal. Cancellation is modeled through a standard exponential survival curve, parameter γ representing the hazard rate.

This survival function will be retained for modeling the tail of our semi-parametric survival curve.

3.2 Fitting the survival function

3.2.1 Likelihood

Recall that the survival function of equation 3.1.1 is discontinuous. That is, the associated hazard function is not defined when x is an integer. Indeed, for any $x_0 \in \mathbb{N}$,

$$\lim_{\substack{x \rightarrow x_0 \\ x < x_0}} S(x) = e^{-\gamma x_0} p^{x_0-1} \neq \lim_{\substack{x \rightarrow x_0 \\ x > x_0}} S(x) = e^{-\gamma x_0} p^{x_0}$$

Thus, the general setting of section 2.4.1 does not apply here. Fortunately, one of the key assumptions in the model is that the service time (namely the time during which a policyholder is insured) is independent of the arrival process. It means that whatever the arrival process, the insured period is independent of it.

In this case, we can rely on [BCP16]. We will provide a bit more details on how the likelihood is obtained. Let's consider two events:

- A contract is cancelled at the anniversary date of the policy
- A contract is cancelled at a different date than the anniversary date of the policy

Let t_1, t_2, \dots, t_n , $0 < t_1 < \dots < t_n < \dots$ be event times, as introduced above. Borrowing the notations of [BCP16], denote by W_1, W_2, \dots, W_n the number of policyholders in the insurance portfolio right before t_1, t_2, \dots, t_n .

The probability that t_1 is a cancellation at a different date than the anniversary date, is the probability that among the W_1 policyholders, all of them have survived at least t_1 , and a few of them have renewed their contracts if the anniversary date fell in the interval $[0, t_1[$. Let's denote the latter case by $h(t_1)$ using [BCP16] notations. The former case can be rewritten as the probability that the minimum survival time of the W_1 policyholders is t_1 . In this case, what

one needs to determine is the density of the random variable $\min_{1 \leq i \leq W_1} T_i$. Using the cumulative distribution function of this random variable, we get:

$$\begin{aligned} \mathbb{P}\left(\min_{1 \leq i \leq W_1} T_i \leq t_1\right) &= 1 - \mathbb{P}\left(\min_{1 \leq i \leq W_1} T_i > t_1\right) \\ &= 1 - \prod_{i=1}^{W_1} \mathbb{P}(T_i > t_1) \\ &= 1 - \prod_{i=1}^{W_1} e^{-\gamma t_1} \\ &= 1 - e^{-\gamma W_1 t_1} \end{aligned}$$

Taking the first derivative of this cumulative distribution function with respect to t_1 , one gets the distribution function

$$f_{\min_{1 \leq i \leq W_1} T_i}(t_1) = \gamma W_1 e^{-\gamma W_1 t_1}$$

Using the memoryless property of the exponential distribution, one gets in a more general setting, $\forall p \in \mathbb{N}$,

$$f_{\min_{1 \leq i \leq W_p} T_i}(t_1) = \gamma W_p e^{-\gamma W_p (t_p - t_{p-1})}$$

The probability that t_1 is a cancellation at the anniversary date of the contract is the probability that among all the policyholders that have their contract anniversary date that falls between $[0, t_1[$, that is $h(t_1)$, exactly one of them has not renewed it. Since the probability of renewing the contract is p , then the probability of the event of interest is $p^{h(t_1)-1}(1-p)$.

We are now clear on how the likelihood was obtained in [BCP16]. As the end of the calculation is quite clear in the article, we can directly use the final results of the parameters estimators using equations (4.4) and (4.5) of [BCP16], that is:

$$\hat{\gamma} = \frac{A}{\sum_{i=1}^{\xi} T_i} \tag{3.2.1}$$

$$\hat{p} = \frac{\sum_{i=1}^{\xi} \lfloor T_i \rfloor - Q}{\sum_{i=1}^{\xi} \lfloor T_i \rfloor} \tag{3.2.2}$$

where, A denotes the number of contracts that have been cancelled at a different date than the anniversary date, Q denotes the number of contracts that have been cancelled at the anniversary date of the contract, and ξ denotes the total number of contracts in the database.

It is worth commenting on censoring. While taking into account censoring was clear in section 2.4.1, it is not straightforward here. The idea here is that censoring is captured through W_i . Recall that W_i denotes the number of contracts in the portfolio right before an event. It is then a measure of the exposure of the portfolio, and could be understood as the number of individuals at risk in a standard survival analysis. That is, W_i captures the loss of observation of individuals, and thus captures censoring.

3.2.2 Simulating the survival function

Simulating the survival function lies in the standard techniques of random variables simulation. The general idea is to simulate a sample realization of a random variable using a sample realization of a uniform random variable, and inverting the cumulative distribution function.

The method, called the inverse transform method, lies in the following theorem.

Proposition 3.2.1 (Inverse transform method). *Let F be a cumulative distribution function. We define the generalized inverse of the cumulative distribution function by $F^{-1}(u) = \inf\{x, F(x) \geq u\}$. If U is a random variable uniformly distributed on $[0, 1]$, then the cumulative distribution function of X defined by $X = F^{-1}(U)$ is F .*

Proof. See [Sig10]. □

The idea of the general inverse lies in the fact that some cumulative distribution functions might have discontinuities. In that case, the inverse might not be defined, and one needs to rely on the generalized inverse function.

For the exponential distribution, an exact formula exists for F^{-1} . Recall that the cumulative distribution function for an exponential distribution is defined by $F(x) = 1 - e^{-\lambda x}$. That is, F is continuous and strictly increases. According to the intermediate value theorem, F is bijective, and one can find the exact inverse cumulative distribution function for F . Let $u \in [0, 1]$.

$$F(x) = u \iff 1 - e^{-\lambda x} = u \iff \ln(1 - u) = -\lambda x \iff x = -\frac{\ln(1 - u)}{\lambda}$$

Thus

$$F^{-1}(u) = -\frac{\ln(1 - u)}{\lambda} \tag{3.2.3}$$

Simulating an exponential random variable is then quite straightforward and time-efficient, as one simply needs to simulate samples from a uniform random variable, and then apply the function F^{-1} .

Unfortunately, our cumulative distribution function doesn't show such desired properties. It is not continuous, and thus not bijective. Indeed, this survival function shows some steps every year. No closed-form formula can be obtained for the inverse of this function, and the simulation algorithm can be quite complex.

A naïve solution would be to simulate a uniform random variable, and find the associated value of the survival using some dichotomy algorithms. Two major issues may arise from this solution:

- The outcome would not be exact. The dichotomy algorithm would only return an approximate value of the survival time
- The time to run the simulation would be outstanding. The dichotomy algorithm would need to be repeated as many times as the number of simulations we want to perform. Moreover, its time-performance depends on the precision desired by the user. Finally, it also depends on the initial guess.

Of course, one could think of alternative root-finding algorithms, that are much more time-performant than the dichotomy, but the general idea here is to state that repeating such algorithms as many times as the number of samples required would not be suitable for a large portfolio of policyholders. One needs to find an alternative.

Fortunately, our survival function is desirable enough to still be able to perform simulations with:

- An exact result
- Good time performance without relying on root-finding algorithms

The general idea of the algorithm relies on the fact that one doesn't need to simulate up to an indefinite time horizon. For instance, for the purpose of capital requirements, one needs to simulate up to a 1-year time horizon. For others reasons, one could need to project further the simulations, but it would still be a finite horizon. If one is interested in projecting the portfolio until run-off, that would theoretically require to project up to an infinite time horizon. This is actually not the case, as one could actually project up to the maximum number of years a man has ever lived, which still makes the projection time horizon finite. Thus this hypothesis would be valid.

Thus, let's say that we are interested in projecting survival times from 0 years to some arbitrary fixed time horizon h years. That is, F is restricted to $[0, h]$. The main difference between our survival function and a standard exponential cumulative distribution function is the steps. However, between two steps, our survival function is still exponential. Thus, by calculating the upper and lower bound of the survival function at each year $y_i \in [0, h]$, then one

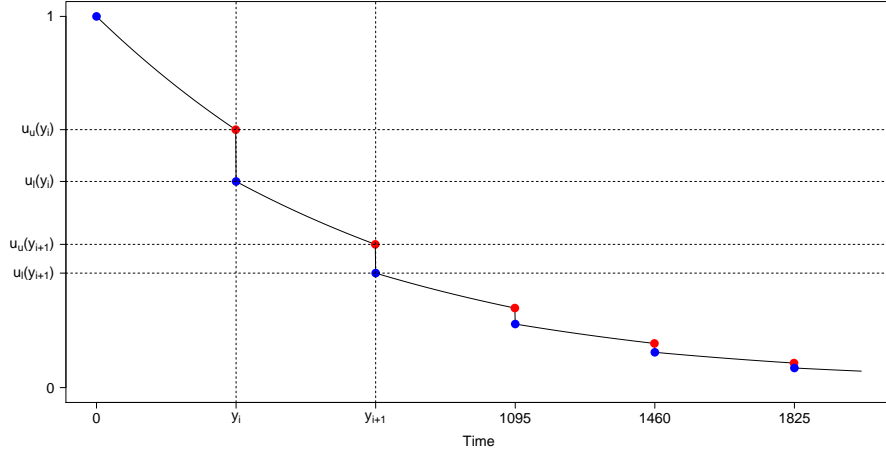


Figure 3.8: The survival function

knows the intervals $[u_u(y_{i+1}), u_l(y_i)]$, $u_l(y_i) = F_{\leftarrow}(y_i)$, $u_u(y_{i+1}) = F_{\rightarrow}(y_{i+1})$ under which the inverse cumulative distribution function is exponential. See figure 3.8. The red points denote the points $F_{\rightarrow}(y_{i+1})$, while the blue points denote the points $F_{\leftarrow}(y_i)$.

We are now able to cut the survival function in $2 \times h$ intervals:

- $[1, u_u(y_1)]$ and $[u_u(y_1), u_l(y_1)]$
- $[u_l(y_1), u_u(y_2)]$ and $[u_u(y_2), u_l(y_2)]$
- etc.

Consider a sample realization $u \in [0, 1]$, from a uniform distribution. Then, if u falls in an interval of the type $[u_l(y_i), u_u(y_{i+1})]$, then one knows that the survival lies in the exponential part of the survival function, and one can determine the associated survival value using the inverse transform method and the exact inverse cumulative distribution function derived in equation 3.2.3. If u falls in an interval of the type $[u_u(y_i), u_l(y_i)]$, then one knows that the survival lies in the step part of the survival function, and one knows exactly the number of years the contract survived in the portfolio.

Such an algorithm is much more powerful, time-efficient and straightforward than the naive one exposed earlier. However, to vectorize such operations, they require to store large matrix, and then a lot of memory.

3.2.3 Fitting

An interesting feature of figure 3.7 is that, depending on the day of underwriting, the survival curves do not match in the exponential part. Intuitively, it means that the γ parameters may differ with respect to the day of underwriting. The fitted curve would look like figure 3.9.

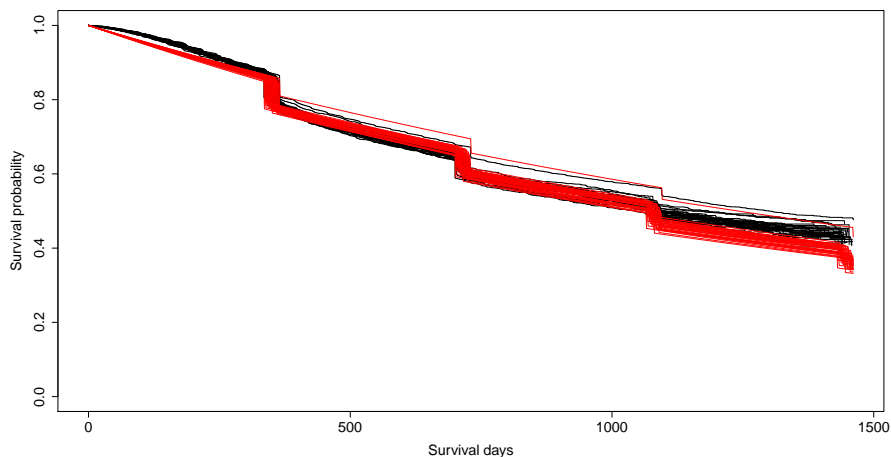


Figure 3.9: Survival function with constant γ

Taking such a specificity into account would lead to the introduction of 31 γ parameters. To avoid an over-parametrization and overfitting issue, it is preferable to use the same γ parameter for all survival curves. The same reasoning could apply to the parameter p . Similarly, only one p will be used.

Recall that the anniversary date of a contract has been defined as the first day of the month at which the contract has been underwritten. That is, for a contract underwritten on 03/02/N, then the anniversary date is set at 01/02/N. The first step to estimate γ and p is to determine A and Q (see equations 3.2.1 and 3.2.2). The determination of A and Q requires the identification of the policyholders in our database that have cancelled their contracts at the anniversary date or not.

To do so, let m be the underwriting month of the contract, and d the underwriting day of the contract. Let t be the termination date of the contract (if the data are not censored), and define the shifted termination date by $t_s = t + d - 1$. Then, if the day of the shifted termination date equals d and if the month of the shifted termination date equals m , then the contract is cancelled at the anniversary date, and this contract contributes to Q . If not, but if the con-

tract is still cancelled, then the contract is terminated at a different date than the anniversary date, and this contract contributes to A . If the contract has not been cancelled yet (censored data), then it doesn't contribute to A nor to Q .

The estimated parameters are given in table 3.1.

Table 3.1: Fitting results of the survival function

$\hat{\gamma}$	\hat{p}	$\sqrt{\mathbb{V}(\hat{\gamma})}$	$\sqrt{\mathbb{V}(\hat{p})}$	$\ln \hat{l}$
0.1620	0.9157	0.000781	0.000669	-170913

The results are also plotted in figure 3.10.

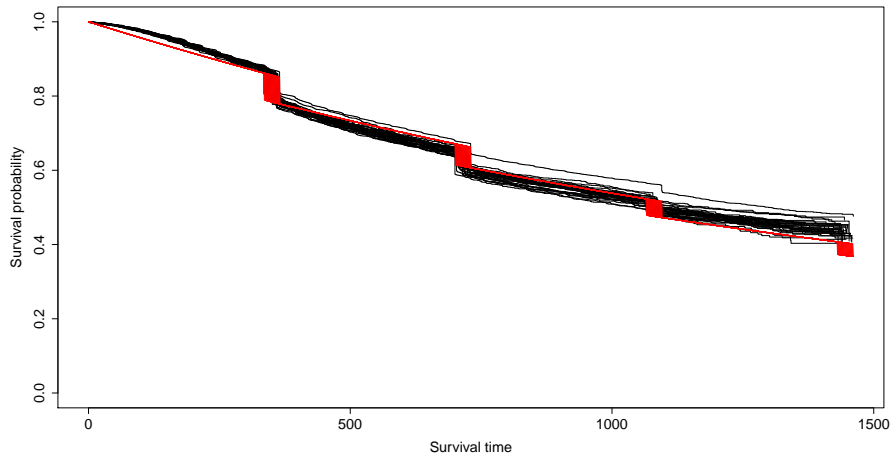


Figure 3.10: Fitted survival function

As predicted, the first exponential part of the survival curve is underestimated. This would lead to a simulation of too many terminations during the first year in the portfolio, which might not be accurate. More precisely, in the context of projecting a portfolio up to a 1-year horizon, figure 3.10 suggests that more than 20% of policyholders who underwrite a contract terminate it within 1-year. The newer the portfolio, the higher the impact of such an error on the number of policyholders still in the portfolio one year later. For much more mature portfolios, the share of new policyholders can be negligible compared to old policyholders, and this impact can be neglected. We might face the following two cases:

- If one wants to study the run-off of the portfolio, then the error on the first exponential part can be neglected, and some simple parametric form of the survival curve can be used.

- If one wants to study the movements of the portfolio in a very short period, the idea of using a semi parametric survival curve, as depicted in section 2.4.2, makes much more sense in this case.

In the semi parametric case, the choice of the threshold can be very tricky, and can lead to severe over or underestimations. Recall that the threshold can be determined statistically using goodness-of-fit criteria. We will however keep the model simple by setting the threshold manually. To do so, finding a range in which we can choose the threshold is key.

A good starting point is to choose it after the first exponential part because, as discussed, we want to use the Kaplan-Meier estimator in this area to avoid any bias.

A second criterion would be to choose a point at which the survival data is not erratic. On the right side of our survival curves, one can see that data are spreading, and that some inconsistencies and irregularities show up.

The last criterion, which is the most important one at some point, is to recall how the parametric curve is "glued" to the Kaplan-Meier estimator. This is simply done by multiplying the parametric curve by a constant factor. That is, the impact of the multiplier will be very important for $S(x)$ close to 1, but much less important for $S(x)$ close to 0. For instance, choosing a threshold $s = 1$ year will lead to an overestimation of the survival curve in the interval $[1, 3]$ years.

An appropriate choice seems then to be $s = 2$ years, because the red curves on figure 3.10 seem to cross the Kaplan-Meier curves approximatively "in the middle". No over nor underestimation is expected at this point, and it enables to use the Kaplan-Meier estimator on $[0, 2]$, which is an interval for which the data are quite stable and not erratic. The results of such a choice are depicted in figure 3.11, for $s = 680$ days. The red curve shows the extension of the Kaplan-Meier estimator, which is in black.

If more precision is required, it is of course possible to determine the threshold statistically, using the method proposed in [GTP13].

3.3 Fitting the arrival process

Sections 1.1.1 enabled to introduce three general ways of modeling the arrival process:

- one based on homogeneous Poisson processes,
- one based on inhomogeneous Poisson processes,
- one based on Poisson-Gamma Cox process.

While section 1.2 showed that the output process and the number of individuals in service were defined in a closed-form for the homogeneous and the

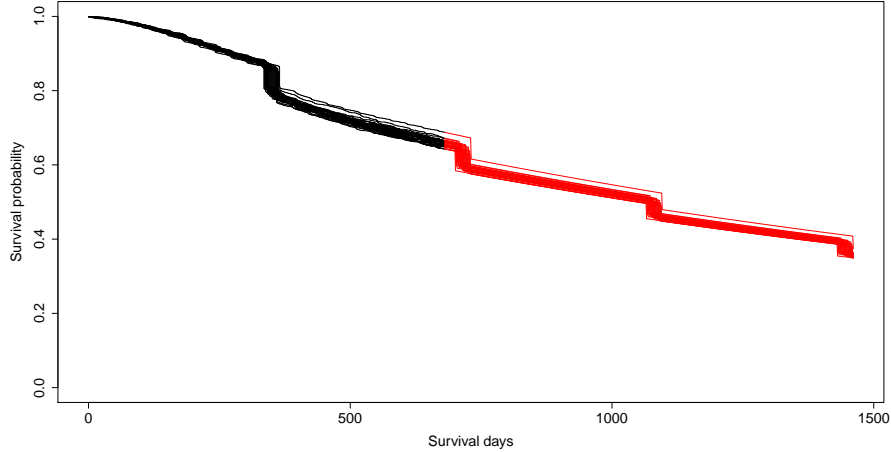


Figure 3.11: Semiparametric estimator of the survival function

non-homogeneous case, it has not been possible to generalize the result for the Poisson-Gamma Cox process. Hence, no closed-form distribution can be applied to backtest the model in the case of overdispersed data. However, the mean and standard deviation can be easily computed for the arrival process.

This section will present the methodology retained for fitting all different kinds of arrival processes, and discuss on the results.

3.3.1 Fitting the homogeneous Poisson process

Intensity function estimation A good starting point of the study would be to keep it simple, by fitting the simple homogeneous Poisson process, and to apply the simple $M/G/\infty$ queue described in section 1.2.2.2. This setting shows the advantage of having a simple distribution for the output process and the number of individuals in service, which are Poisson.

Such assumptions can be simply verified. Recall that the log-likelihood in the Poisson case for the arrival process writes:

$$\ln(\mathcal{L}(k_1, \dots, k_n, \mathbf{p})) = \sum_{i=1}^n [-\Lambda(t_{i-1}, t_i) + k_i \ln(\Lambda(t_{i-1}, t_i)) - \ln k_i!]$$

Let $\lambda > 0$ be the intensity parameter for the arrival process. Then one can write:

$$\Lambda(t_{i-1}, t_i) = \lambda(t_i - t_{i-1})$$

Hence, the log-likelihood writes:

$$\ln(\mathcal{L}(k_1, \dots, k_n, \lambda)) = -\lambda t + \sum_{i=1}^n [k_i \ln(\lambda(t_i - t_{i-1})) - \ln k_i!]$$

Writing the derivative with respect to λ , one gets:

$$\frac{d(-\ln(\mathcal{L}(k_1, \dots, k_n, \lambda)))}{d\lambda} = t - \frac{1}{\lambda} \sum_{i=1}^n k_i$$

Thus,

$$\lambda = \frac{1}{t} \sum_{i=1}^n k_i \tag{3.3.1}$$

This result can be simply interpreted as the mean number of arrivals per day. Hence, no seasonality nor trend will arise from such a model.

Parameter estimation The results are depicted in table 3.2.

Table 3.2: Fitting results of the homogeneous Poisson Process

$\hat{\lambda}$	$\sqrt{\mathbb{V}(\hat{\lambda})}$	AIC	BIC	$\ln \hat{l}$
147.7823	0.3180432	108,792.5	108,797.8	-54,395.24

Simulations A simple algorithm can be found in [Mcq10] to simulate a homogeneous Poisson process. The real number of underwritten contracts can be found on figure 3.12, while the simulated number of contracts can be found in 3.13. The red line on each barplot represents the expected number of underwritten contracts per day, which is constant in the homogeneous case. One clearly sees that the homogeneous Poisson process doesn't fit the real number of contracts per day. This was expected given the seasonality observed in the underwriting scheme. Moreover, one can see that the overdispersion is not captured in the simulation, which was predictable since the variance equals the mean in the homogeneous Poisson case. Even though homogenous Poisson process represents a good first approach for modeling queuing system, it fails when data show overdispersion or non-stationarity.

The expected number of contracts in a homogeneous Poisson process between 0 and t satisfies $\mathbb{E}[N] = \lambda t$, that is, a linear function of time. Figure 3.14 plots the cumulated number of contracts with respect to time.

One sees that the black line doesn't fit with the theoretical red line, which confirms the intuition of non-homogeneity.

The following subsection will deal with the non-homogeneous case. Considering the above discussion, we expect the result to properly capture the expectation, that is non-stationarity, while we expect the overdispersion not to be captured by the model.

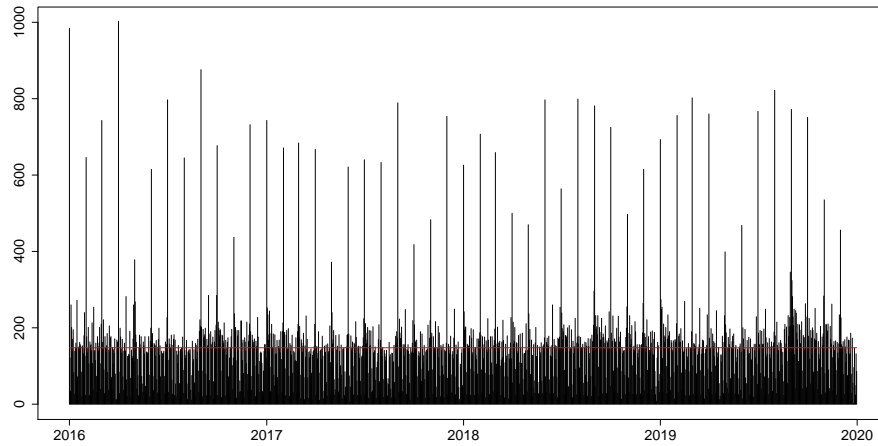


Figure 3.12: Observed number of underwritten contracts per day

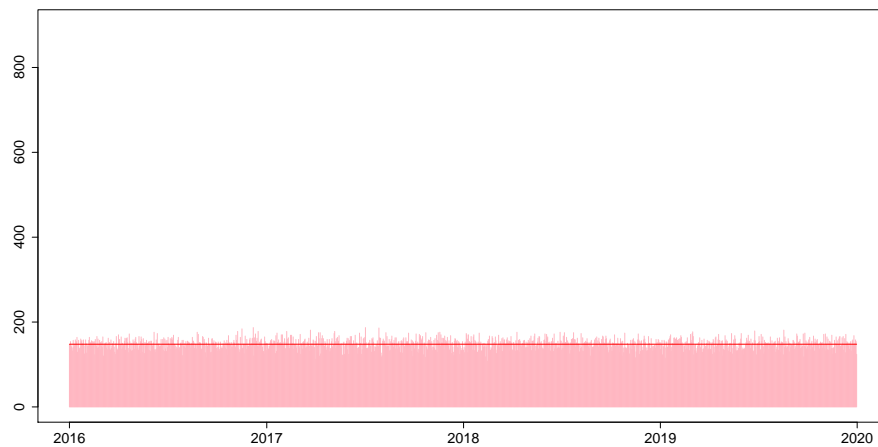


Figure 3.13: Sample realization of the number of underwritten contracts per day, with a homogeneous Poisson process

3.3.2 Fitting the non-homogeneous Poisson process

The previous subsection has shown that the most simple approach for modeling the number of contracts fails in expected value. The studied data show non-stationarity that violates the main underlying results of a homogeneous Poisson process.

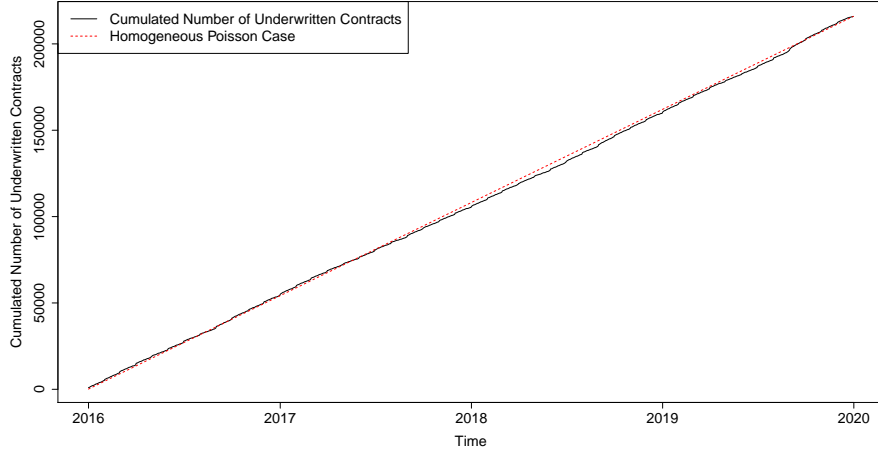


Figure 3.14: Cumulated number of contracts with respect to time

A good alternative to overcome the stationarity problem is to work with non-homogeneous Poisson process, whose intensity function enables to capture the variations in the expected path.

The key point of a non-homogeneous Poisson process is to specify the intensity function. A convenient choice will prevent over-fitting. Going back to figures 3.1 and 3.4, one can see the following patterns:

- Peak activity is recorded during the first day of each month. Intuitively, this is due to people wanting to start their contracts on the first day of the month.
- This peak activity is lower in May and November.
- Flat activity is recorded each day of the week, except on Saturdays and Sundays.

Intensity function estimation Taking into account those considerations, a natural intensity function can be cut in two distinct intensity functions, that are defined depending on the day of the month.

$$\lambda_1(x) = \begin{cases} a & \text{if weekday of } x \text{ is not Saturday nor Sunday} \\ b & \text{if weekday of } x \text{ is Saturday} \\ c & \text{if weekday of } x \text{ is Sunday} \end{cases} \quad (3.3.2)$$

$$\lambda_2(x) = \begin{cases} A & \text{if month of } x \text{ is not May nor November} \\ B & \text{if month of } x \text{ is May or November} \end{cases} \quad (3.3.3)$$

Then the final intensity function is:

$$\lambda(x) = \begin{cases} \lambda_1(x) & \text{if } x \text{ is not the first of the month} \\ \lambda_2(x) & \text{if } x \text{ is the first of the month} \end{cases} \quad (3.3.4)$$

This intensity function is a piecewise constant intensity function, which is very suitable for non-homogeneous Poisson process modeling. It enables to get closed form formulas for the parameters.

Our data set is observed at the end of each day. Each record corresponds to the number of underwritten contracts at the end of each day. Hence, in the likelihood formula in section 2.2, the term $\int_{t_{i-1}}^{t_i} \lambda(t)dt = \Lambda(t_{i-1}, t_i)$ integrates the intensity function for a specific day. That is, $t_{i-1} \in \mathbb{N}$ and $t_i \in \mathbb{N}$, and $t_i = t_{i-1} + 1$. In a more general setting, let $t_i \in \mathbb{R}_+$ and $t_{i-1} \in \mathbb{R}_+$ so that $\lceil t_{i-1} \rceil + 1 = \lceil t_i \rceil$.

In this case, this integral can be re-written in a closed-form. First, let's compute the cumulated intensity function $\Lambda(t_{i-1}, t_i) = \int_{t_{i-1}}^{t_i} \lambda(t)dt$. It is straightforward to say that, depending on the condition introduced with the definition of the intensity function, one gets:

$$\Lambda(t_{i-1}, t_i) = \begin{cases} a(t_i - t_{i-1}) \\ b(t_i - t_{i-1}) \\ c(t_i - t_{i-1}) \\ A(t_i - t_{i-1}) \\ B(t_i - t_{i-1}) \end{cases} \quad (3.3.5)$$

Let $\mathbf{p} = (a, b, c, A, B)$. The log-likelihood writes:

$$\ln(\mathcal{L}(k_1, \dots, k_n, \mathbf{p})) = \sum_{i=1}^n [-\Lambda(t_{i-1}, t_i) + k_i \ln(\Lambda(t_{i-1}, t_i)) - \ln k_i!]$$

Let \mathcal{C}_a be the set of index $i \in \{1, \dots, n\}$ so that t_{i-1} is not the first day of the month and is not during the weekend. Similarly, define \mathcal{C}_b , \mathcal{C}_c , \mathcal{C}_A and \mathcal{C}_B according to the definition of $\lambda(x)$ in equation 3.3.4. Then the likelihood rewrites, for a particular set:

$$\ln(\mathcal{L}(k_1, \dots, k_n, \mathbf{p})) = \sum_{\substack{i=1 \\ i \in \mathcal{C}_a}}^n [-a(t_i - t_{i-1}) + k_i \ln(a(t_i - t_{i-1})) - \ln k_i!]$$

For the other sets, one simply needs to replace the set \mathcal{C}_a in the summation symbol, and the variable a in the term of the sum.

For optimization purposes, one needs to compute the derivative of the log-likelihood with respect of each parameter. It is equivalent to compute the derivative of $\Lambda(t_{i-1}, t_i)$ and $\ln \Lambda(t_{i-1}, t_i)$

Let $p_i \in \mathbf{p}$. One gets:

$$\frac{\partial (\ln \Lambda(t_{i-1}, t_i))}{\partial p_i} = \frac{\frac{\partial \Lambda(t_{i-1}, t_i)}{\partial p_i}}{\Lambda(t_{i-1}, t_i)}$$

For a particular set, say \mathcal{C}_a , then one gets:

$$\frac{\partial}{\partial a} (a(t_i - t_{i-1})) = t_i - t_{i-1}$$

and:

$$\frac{\partial}{\partial a} (\ln a(t_i - t_{i-1})) = \frac{1}{a}$$

In the general form, we get

$$\begin{aligned} & \nabla_{\mathbf{p}}(-\ln(\mathcal{L}(k_1, \dots, k_n, \mathbf{p}))) \\ &= \sum_{i=1}^n \left[\nabla_{\mathbf{p}} \left(\int_{t_{i-1}}^{t_i} \lambda(t) dt \right) - k_i \frac{\nabla_{\mathbf{p}} \left(\int_{t_{i-1}}^{t_i} \lambda(t) dt \right)}{\int_{t_{i-1}}^{t_i} \lambda(t) dt} \right] \end{aligned}$$

For a particular set, say \mathcal{C}_a , we obtain:

$$\nabla_a(-\ln(\mathcal{L}(k_1, \dots, k_n, \mathbf{p}))) = \sum_{\substack{i=1 \\ i \in \mathcal{C}_a}}^n \left[(t_i - t_{i-1}) - \frac{k_i}{a} \right]$$

As stated initially, the advantage of working with a piecewise constant intensity function is to have closed form formula for the parameters. By equating the above equation to 0, we can obtain the formulas for all the parameters:

$$\begin{aligned} \hat{a} &= \sum_{\substack{i=1 \\ i \in \mathcal{C}_a}}^n \frac{k_i}{t_i - t_{i-1}} \\ \hat{b} &= \sum_{\substack{i=1 \\ i \in \mathcal{C}_b}}^n \frac{k_i}{t_i - t_{i-1}} \\ \hat{c} &= \sum_{\substack{i=1 \\ i \in \mathcal{C}_c}}^n \frac{k_i}{t_i - t_{i-1}} \\ \hat{A} &= \sum_{\substack{i=1 \\ i \in \mathcal{C}_A}}^n \frac{k_i}{t_i - t_{i-1}} \end{aligned}$$

$$\hat{B} = \sum_{\substack{i=1 \\ i \in \mathcal{C}_B}}^n \frac{k_i}{t_i - t_{i-1}}$$

The standard error of those estimators writes (example given for a):

$$s_a = \sqrt{\mathbb{V}(\hat{a})} = \sqrt{\frac{\hat{a}}{\#\mathcal{C}_a}}$$

$\#\mathcal{C}_a$ denotes the cardinal of \mathcal{C}_a , that is the number of elements in this set.

Parameters estimation The results of the parameters estimations are depicted in table 3.3. One can see that both AIC and BIC are much lower than in the homogeneous case (see table 3.2).

Table 3.3: Fitting results of the non-homogeneous Poisson process

\hat{a}	\hat{b}	\hat{c}	\hat{A}	\hat{B}
160.60714	83.14778	26.78218	703.95	446.375
$\sqrt{\mathbb{V}(\hat{a})}$	$\sqrt{\mathbb{V}(\hat{b})}$	$\sqrt{\mathbb{V}(\hat{c})}$	$\sqrt{\mathbb{V}(\hat{A})}$	$\sqrt{\mathbb{V}(\hat{B})}$
0.3991647	0.8894764	0.8916753	2.00379	4.4806
AIC	BIC	$\ln \hat{l}$		
23,948.16	23,974.59	-11,969.08		

Simulations The real number of underwritten contracts is depicted in figure 3.15, the simulated number of underwritten contracts is depicted in figure 3.16, and the expected number of underwritten contracts is depicted in figure 3.17. One can see the significative improvement in forecasting the expected number of underwritten contracts per month. The introduction of a varying intensity function for the non-homogeneous Poisson process enables to capture the non-stationarity of our data.

Boucher and Couture-Piché introduced the possibility to take into account non-stationarity in his model, see [BCP16]. They however restrained to considering homogeneous Poisson processes for simplicity purposes. Hence, the first main result of this thesis is the generalization of the work proposed by Boucher and Couture-Piché. That is, it is possible to better estimate the expected number of contracts in the portfolio using a non-homogeneous Poisson arrival process. Moreover, section 1.2.2.3 showed that a generalization of the result from Mirasol [Mir63] was possible with a non-homogeneous Poisson arrival process. Thus, one who is interested in forecasting the mean number of contracts in a portfolio and other expected quantities can perfectly use this model to make

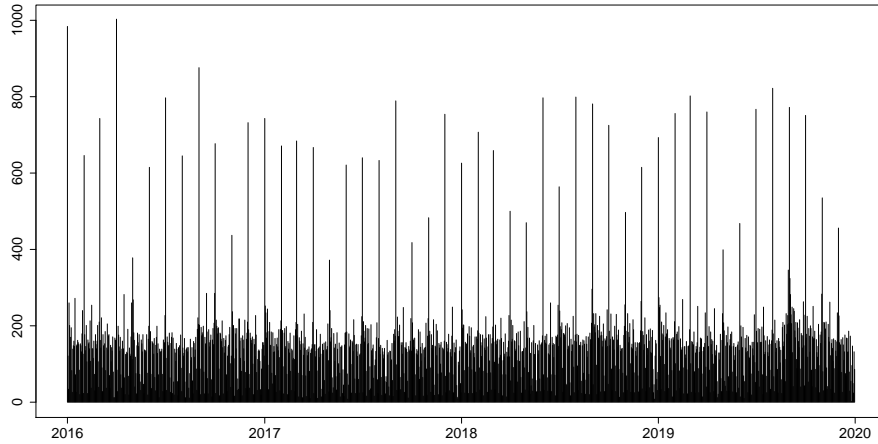


Figure 3.15: Observed number of underwritten contracts per day

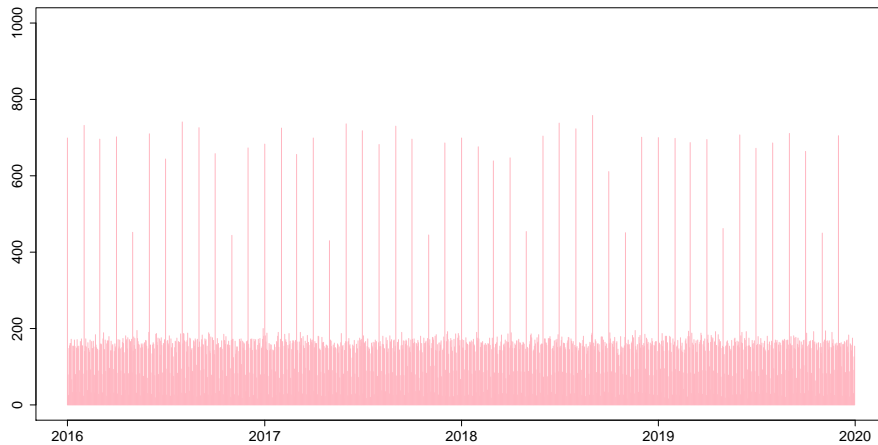


Figure 3.16: Sample realization of the number of underwritten contracts per day, with a non-homogeneous Poisson process

forecasts.

However, overdispersion is still not captured. This thesis focuses on estimating the variability of the premiums. Hence, each component of the model must perfectly fit the variance. In this case, the non-homogeneous Poisson process appears not to be well-fitted for this purpose. The following section will depict

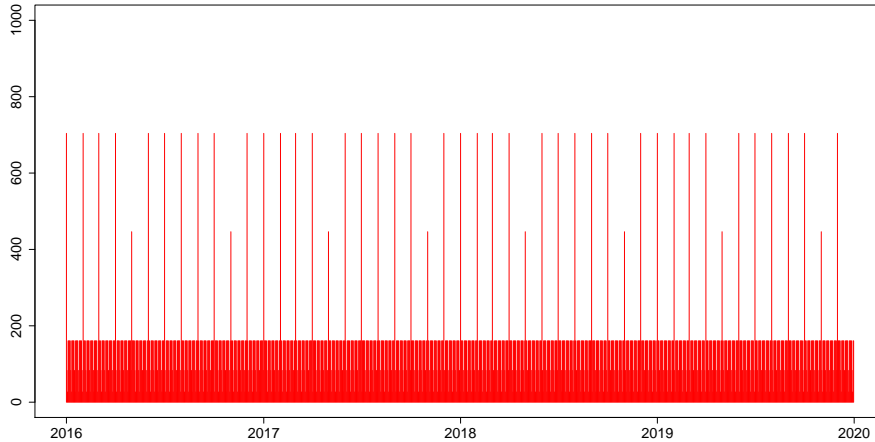


Figure 3.17: Expected number of underwritten contracts per day, with a non-homogeneous Poisson process

the result when using a Cox process for the arrival process. We will lose the generality of the main result proposed in section 1.2.2.2 and 1.2.2.3.

3.3.3 Fitting the Poisson-Gamma Cox process

Overdispersion As discussed earlier, count data usually shows overdispersion. That is, the variance of the process exceeds the mean. A simple calculation of overdispersion can be applied to evidence this behavior in our dataset. Indeed, a simple estimator of the mean and of the standard deviation can be applied to a rolling window, and to calculate both the mean and standard deviation. Then, one can plot the output of the former and the output of the latter. In case no overdispersion arises, the rolling window of the mean should melt with the rolling window of the variance. Indeed, no overdispersion is translated by $\mathbb{E}(X) = \mathbb{V}(X)$. This methodology has been applied with a rolling window of 100 days, that is, a 3-month rolling window. Results are depicted in figure 3.18.

The green line is the rolling standard deviation, that is, the square root of the variance. Since the green line is just below the red line, with values around 100, the variance should equal approximately 10,000, which is much higher than the mean. Overdispersion is present in our dataset.

This kind of behavior has been well pointed out by actuaries, especially on the claim side. For instance, [Liu12] applied a Cox process with shot noise intensity for modeling claim count. Indeed, the shot noise process enables to increase the intensity of claim occurrence when a specific event occurs, and to

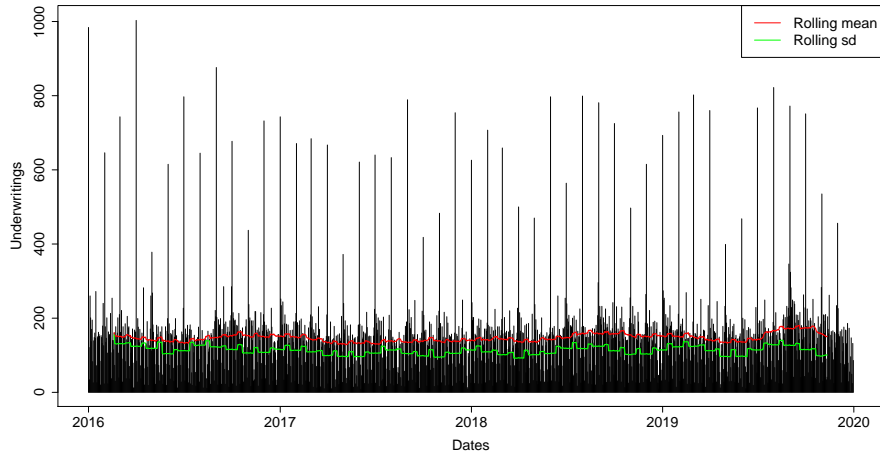


Figure 3.18: Evidence of overdispersion

decrease it exponentially with time. This enables first to better represent reality, since a claim is more likely to be reported to the insurer right after the event, rather than later with time, and secondly to take into account overdispersion, which is a known observation for claim data.

In our case, the shot noise process could be an appropriate way for modeling the arrival process. One could think about special sales or discounts proposed by the insurer which would lead to an increase of the underwritten policies, and which would decrease with time since the discount may have a fixed deadline or simply because advertising is not done anymore for the offer.

However, a strategy would be to make offers regularly in such a way that policyholders would underwrite their contract as regularly as possible, in order to overcome the decrease caused by the shot noise process. That is, an insurer would try to have an underwriting rate as stable as possible. Thus, the shot noise process wouldn't be the best solution to fit the arrival process in the portfolio.

One solution is to adapt the methodology proposed on the claim side by [AAAB19] to our portfolio modeling. Instead of using a shot noise process, the use of a Gamma process seems appropriate. Indeed, it enables first to write the probability of increments with a closed-form formula, that is, a negative binomial distribution (see proposition 1.1.11), and then to model the shape of the arrival process with a specified shape function for the Gamma process. This kind of flexibility is quite suitable for our modeling purposes, since specific patterns can be observed from our data (see section 3.1).

Shape function Hence, let's use for the shape function, the intensity function introduced in section 3.3.2, that is:

$$\alpha(t) = \begin{cases} \lambda_1(t) & \text{if } t \text{ is not the first of the month} \\ \lambda_2(t) & \text{if } t \text{ is the first of the month} \end{cases} \quad (3.3.6)$$

Recall that the log-likelihood in the Poisson-Gamma Cox case writes:

$$\begin{aligned} & \ln(\mathcal{L}(k_1, \dots, k_n, \mathbf{p}, s)) \\ &= \sum_{i=1}^n \left(\ln \left(\Gamma \left(k_i + \int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt \right) \right) - \ln \left(\Gamma \left(\int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt \right) \right) - \ln(k_i!) \right) \\ &+ \sum_{i=1}^n \left(k_i \ln(s) - k_i \ln(s+1) - \ln(s+1) \int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt \right) \end{aligned}$$

Also recall that the integral of $\alpha_{\mathbf{p}}(t)$ has been calculated in equations 3.3.5. For optimization purposes, one could be interested in the gradient of the log-likelihood function. Let $\mathbf{p} = (a, b, c, A, B)$. First, one gets:

$$\nabla_s (-\ln(\mathcal{L}(k_1, \dots, k_n, \mathbf{p}, s))) = \left(\frac{1}{s+1} - \frac{1}{s} \right) \sum_{i=1}^n k_i + \frac{1}{s+1} \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt$$

Since the gradient of the integral of $\alpha_{\mathbf{p}}(t)$ is known thanks to section 3.3.2, one gets:

$$\begin{aligned} & \nabla_{\mathbf{p}} (-\ln(\mathcal{L}(k_1, \dots, k_n, \mathbf{p}, b))) \\ &= \sum_{i=1}^n \nabla_{\mathbf{p}} \left(\int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt \right) \left(\psi \left(\int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt \right) - \psi \left(k_i + \int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt \right) \right) \\ &+ \ln(s+1) \sum_{i=1}^n \nabla_{\mathbf{p}} \left(\int_{t_{i-1}}^{t_i} \alpha_{\mathbf{p}}(t) dt \right) \end{aligned}$$

Where $\psi(x)$ is called the digamma function, and is defined by:

$$\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)} = \frac{\partial \ln \Gamma(x)}{\partial x}$$

In our dataset, it seems that the variance of the number of underwritten contracts on the first day of a month (denoted as N_1) is higher than the number of contracts not underwritten the first day of the month (denoted as N_2). Those two figures are shown in the table 3.4.

That is, the variance is not constant over time, and should then be a function of the time. Such complexification can be taken into account in the model using Extended Gamma Processes (EGP), see [Mer17a] and [AMMV17]. In such

Table 3.4: Evidence of the overdispersion

$\sqrt{\mathbb{V}(N_1)}$	$\sqrt{\mathbb{V}(N_2)}$
150.2979	61.44649

processes, the parameter s is becoming a function of time, that is, $s \mapsto s(t)$.

However, to avoid a too complex model, and to be able to rely on simple results depicted in the previous sections, the model will be constructed as follows. Two arrival processes will be defined: one for the people underwriting their contracts the first day of a month, and one for the others. The shape function of the former will be $\lambda_2(t)$, while the shape function for the latter will be $\lambda_1(t)$. We will then define two scale parameters, s_1 and s_2 . The likelihood remains valid.

Parameters estimation The results of the parameters estimation are depicted in table 3.5 and 3.6.

Table 3.5: Fitting results of the Cox process

\hat{a}	\hat{b}	\hat{c}	\hat{s}_1
16.423002	8.731752	3.062359	9.721608
$\sqrt{\mathbb{V}(\hat{a})}$	$\sqrt{\mathbb{V}(\hat{b})}$	$\sqrt{\mathbb{V}(\hat{c})}$	$\sqrt{\mathbb{V}(\hat{s}_1)}$
0.7086330	0.4171282	0.1647168	0.4222872
AIC	BIC	$\ln \hat{l}_1$	
14,019.75	14,040.89	-7,005.873	

Table 3.6: Fitting results of the Cox process

\hat{A}	\hat{B}	\hat{s}_2
37.45386	24.11130	18.76296
$\sqrt{\mathbb{V}(\hat{B})}$	$\sqrt{\mathbb{V}(\hat{C})}$	$\sqrt{\mathbb{V}(\hat{s}_2)}$
8.051101	5.399656	4.053392
AIC	BIC	$\ln \hat{l}_2$
595.6961	611.5568	-294.8481

Simulation The results of the above derivation are depicted in figures 3.19, 3.20 and 3.21. In this case, both mean and variance seems to be well-fitted. Overdispersion is well taken into account. The other interesting fact is that the expected value process is exactly the same as the one for the non-homogeneous Poisson process.

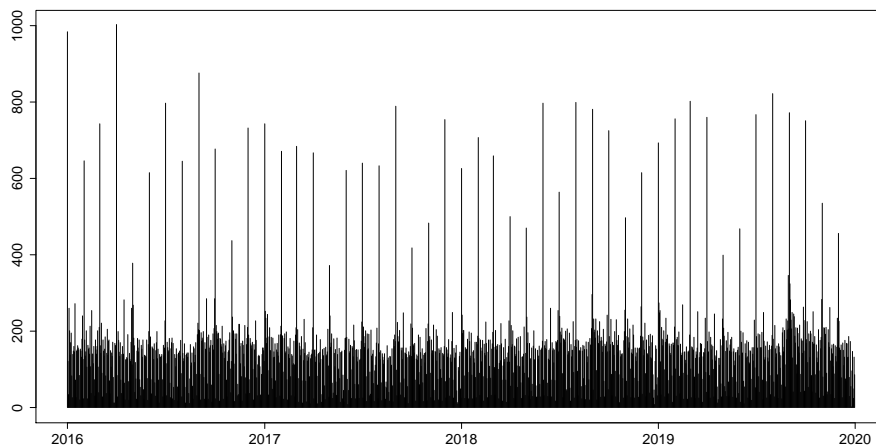


Figure 3.19: Observed number of underwritten contracts per day

Again, AIC and BIC are much lower than the non-homogeneous Poisson case. This improvement shows that the Cox process is suitable to model the overdispersion of underwriting data. The simplicity of the homogeneous Poisson process enables to get a first approximation of the number of contracts in the portfolio, with some bias as peak activity is not captured. The expected value is, in itself, biased. If some raw indications on the number of contracts are required, then the homogeneous Poisson process could be suitable, to avoid too complex computations.

If one wants to overcome the bias issue, a simple turnaround is then to work with non-homogeneous Poisson processes. Indeed, AIC and BIC are approximately five times lower than the homogenous Poisson, meaning a significant improvement in the model. The framework is still simple if one works with piecewise constant intensity functions, as one benefits from the same closed-form formulas for parameter estimations. If one is interested in a better estimation of the expected number of contracts, then this is an appropriate choice. However, no second order measure needs to be derived from this approach.

In this case, it is much better to use the Cox process, that has shown that both the fit is better, but also it enables to properly capture the overdispersion.

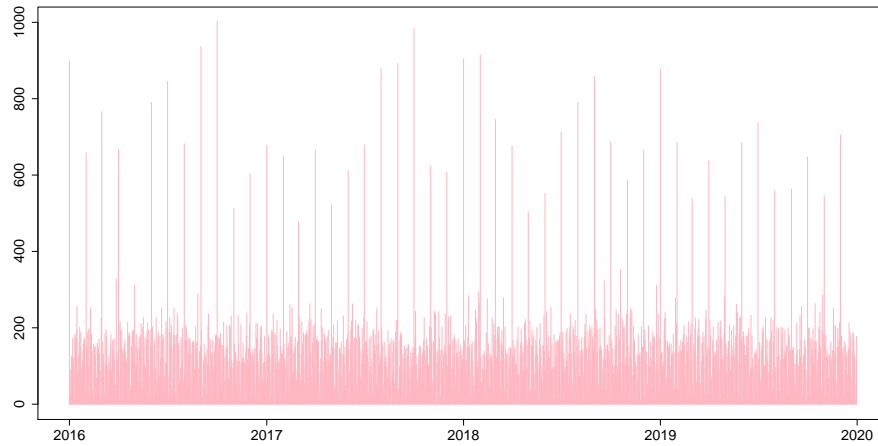


Figure 3.20: Sample realization of the number of underwritten contracts per day, with a Cox process

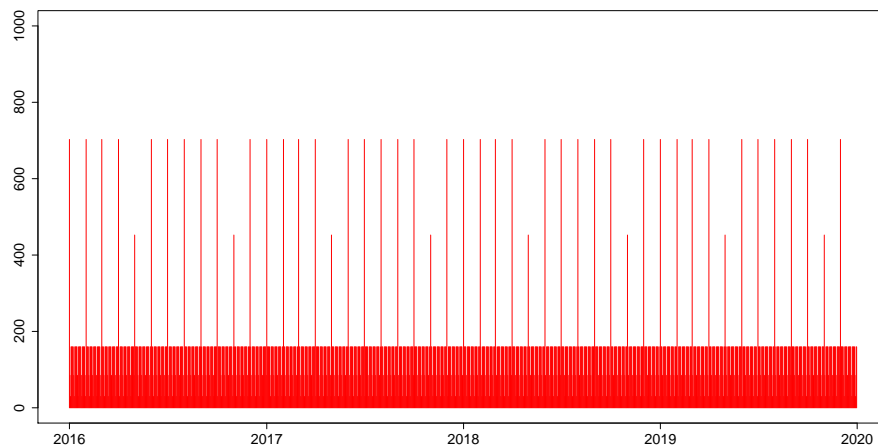


Figure 3.21: Expected number of underwritten contracts per day, with a Cox Process

The expected value is the exact same as the non-homogeneous case, which offers a comprehensive way to model both the expected value and the overdispersion. On the other hand, the simulation algorithm is less efficient, and no closed formula can be derived as stated in section 1.2.2.4.

3.3.4 Comparisons with closed formulas from the queuing theory

Sections 3.3.1, 3.3.2 and 3.3.3 enabled to fit all three types of arrival processes. The expected value of the number of underwritten contracts is fully determined in those three frameworks. However, sections 1.2.2.2 and 1.2.2.3 showed that closed formulas were also available for both the homogeneous and the non-homogeneous Poisson cases, but not for the Cox case.

The purpose of this section is to compute those formulas, and to compare them with our simulated data. The calculations will be made for the non-homogeneous case only, as the homogeneous Poisson process is a particular case ($\lambda(t) = \lambda$).

Recall from proposition 1.2.4 that the number of contracts in the portfolio at time $t + T$ is Poisson-distributed with mean

$$\mu_i(t, T) = s(t, T) (\Lambda(t + T) - \Lambda(t)) + q(t, T)\Lambda(t)$$

and that the number of terminated contracts between time t and $t + T$ is Poisson-distributed with mean

$$\mu_o(t, T) = r(t, T) (\Lambda(t + T) - \Lambda(t)) + p(t, T)\Lambda(t)$$

with $p(t, T)$, $q(t, T)$, $r(t, T)$ and $s(t, T)$ as defined in proposition 1.2.3. Then, one gets:

$$s(t, T) (\Lambda(t + T) - \Lambda(t)) = \int_0^T \lambda(t + T - x)(1 - H(x))dx$$

$$\begin{aligned} q(t, T)\Lambda(t) &= \int_0^t \lambda(t - x)(1 - H(T + x))dx \\ &= \int_T^{T+t} \lambda(t + T - u)(1 - H(u))du \end{aligned}$$

Hence, one gets:

$$\begin{aligned} \mu_i(t, T) &= s(t, T) (\Lambda(t + T) - \Lambda(t)) + q(t, T)\Lambda(t) \\ &= \int_0^T \lambda(t + T - x)(1 - H(x))dx + \int_T^{T+t} \lambda(t + T - x)(1 - H(x))dx \\ &= \int_0^{T+t} \lambda(t + T - x)(1 - H(x))dx \\ &= \int_0^{T+t} \lambda(t + T - x)S(x)dx \end{aligned}$$

The mean number of contracts in the portfolio at time $t + T$ only depends on $t + T$. Moreover, $t + T$ enters in the formula both at the top of the integral, but

also in the intensity function. That is, the knowledge of the number of contracts in the portfolio at any time t_1 doesn't imply the knowledge of the number of contracts in the portfolio at any time $t_2 > t_1$: one needs to compute the integral from $t = 0$. Thus the difference between the number of contracts at time t_2 and at time t_1 must be calculated with $\mu_i(0, t_2) - \mu_i(0, t_1)$, and not with some integral with bounds going from t_1 to t_2 . The only particular case where using an integral from t_1 to t_2 is justified, is for the homogeneous Poisson case, as one gets:

$$\mu_i(0, t_2) - \mu_i(0, t_1) = \lambda \int_{t_1}^{t_2} S(x) dx$$

Some similar results can be obtained for the number of terminated contracts between t and $t + T$.

$$\begin{aligned} r(t, T) (\Lambda(t + T) - \Lambda(t)) &= \int_0^T \lambda(t + T - x) H(x) dx \\ p(t, T) \Lambda(t) &= \int_0^t \lambda(t - x) (H(T + x) - H(x)) dx \\ &= \int_T^{T+t} \lambda(t + T - u) H(u) du - \int_0^t \lambda(t - x) H(x) dx \end{aligned}$$

Hence, one gets:

$$\begin{aligned} \mu_o(t, T) &= r(t, T) (\Lambda(t + T) - \Lambda(t)) + p(t, T) \Lambda(t) \\ &= \int_0^{T+t} \lambda(t + T - x) H(x) dx - \int_0^t \lambda(t - x) H(x) dx \\ &= \int_0^{T+t} \lambda(t + T - x) (1 - S(x)) dx - \int_0^t \lambda(t - x) (1 - S(x)) dx \\ &= \int_0^{T+t} \lambda(t + T - x) dx - \int_0^t \lambda(t - x) dx \\ &\quad - \left(\int_0^{T+t} \lambda(t + T - x) S(x) dx - \int_0^t \lambda(t - x) S(x) dx \right) \end{aligned}$$

This result is very intuitive. The first two integrals (with $\lambda(\cdot)$ only) denote the expected number of underwritten contracts between time t and time $t + T$. The other two sets of integrals can be rewritten as $\mu_i(t, T) - \mu_i(0, t)$, that is, the difference between the number of contracts in the portfolio at time $t + T$ and the number of contracts in the portfolio at time t . Since, intuitively, the expected number of terminations is the difference between the expected number of arrivals minus the expected exposure, the integral makes sense.

Let's now compute the integral $\mu_i(t, T)$. This integral involves the two functions λ and S . Recall from equations 3.3.2, 3.3.3 and 3.3.4 that the intensity

function is defined on a *per-day* basis. Indeed, its value may change depending on the day of the month considered. On the other hand, the survival function is defined on a *per-year* basis: the breaks occur every year (see equation 3.1.1). For the integral to be computed consistently, both functions need to be defined on the same basis. One way is to transform S as:

$$S(x) = e^{-\gamma \frac{x}{365.25}} p^{\lfloor \frac{x+d-1}{365.25} \rfloor}$$

where d denotes the underwriting day to account for the shift discussed in section 3.2.3. The division by 365.25 will introduce a little bias, but it is negligible. Recall also that the survival function has been defined in a semi-parametric way. We will not take this specificity into account to illustrate the closed formulas, as the integral, even if fully computable, would lead to long formulas. The purpose of this section is to illustrate the results of the queuing theory in a simple way. We will write $k = 365.25$.

Recall also that the intensity function being piecewise constant, there exists some intervals $[t_i, t_{i+1}[$ under which $\lambda(t) = \lambda_i$ for $t \in [t_i, t_{i+1}[$. That is, only the integral of the survival function needs to be calculated in $\mu_i(t, T)$, as by cutting the integral using Chasles' segment addition postulate, λ_i will go out of the integral on the appropriate segment.

The integral of the survival function writes:

$$\begin{aligned} \int_a^b S(x) dx &= \int_a^b e^{-\gamma \frac{x}{k}} p^{\lfloor \frac{x+d-1}{k} \rfloor} dx \\ &= \sum_{i=\lfloor \frac{a+d-1}{k} \rfloor}^{\lfloor \frac{b+d-1}{k} \rfloor} p^i \int_{a \vee (i \times k)}^{b \wedge ((i+1) \times k)} e^{-\gamma \frac{x}{k}} dx \\ &= \sum_{i=\lfloor \frac{a+d-1}{k} \rfloor}^{\lfloor \frac{b+d-1}{k} \rfloor} p^i \times \frac{k}{\gamma} \left(e^{-\frac{\gamma}{k}(a \vee (i \times k))} - e^{-\frac{\gamma}{k}(b \vee ((i+1) \times k))} \right) \end{aligned}$$

Then $\mu_i(t, T)$ writes:

$$\begin{aligned} \mu_i(t, T) &= \int_0^{T+t} \lambda(t+T-x) S(x) dx \\ &= \sum_{n=0}^{T+t-1} \lambda(t+T-1-n) \int_n^{n+1} S(x) dx \\ &= \sum_{n=0}^{T+t-1} \lambda(t+T-1-n) \sum_{i=\lfloor \frac{n+d-1}{k} \rfloor}^{\lfloor \frac{n+d}{k} \rfloor} p^i \frac{k}{\gamma} \left(e^{-\frac{\gamma}{k}(n \vee (i \times k))} - e^{-\frac{\gamma}{k}((n+1) \vee ((i+1) \times k))} \right) \end{aligned}$$

For the number of terminations $\mu_o(t, T)$, the calculation is straightforward, as it involves the expected number of underwritten contracts which has been computed in equation 3.3.5, and μ_i , which has just been computed above.

Figure 3.22 shows the simulated and expected number of contracts in the portfolio per day, and figure 3.23 shows the simulated and expected number of terminated contracts per day. One can see that the behavior is perfectly captured *via* the closed formulas. Hence, the queuing framework is very suitable for the homogeneous and inhomogeneous Poisson cases, as all interesting quantities are directly available through closed formulas. If only expected values are desired, but not overdispersion, the non-homogeneous Poisson framework will be of high interest.

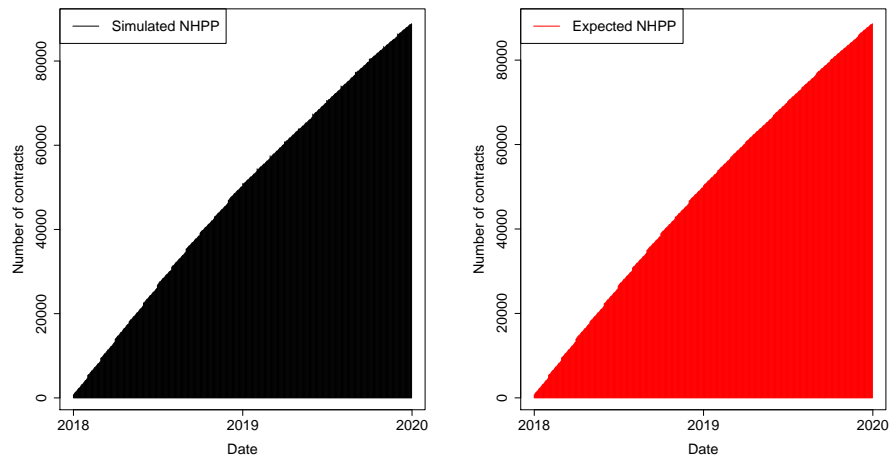


Figure 3.22: Simulated and expected number of contracts in the portfolio per day

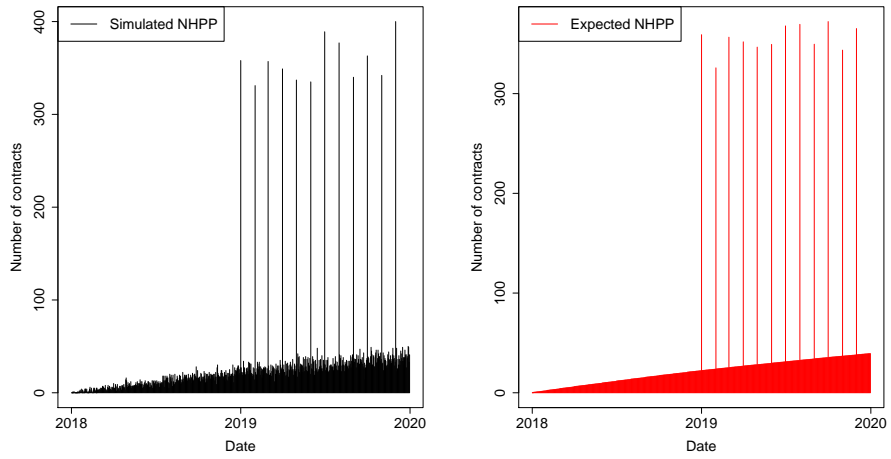


Figure 3.23: Simulated and expected number of terminated contracts per day

Chapter 4

Models comparison and performance

The main development of this thesis was to extend the framework proposed by [BCP16], by proposing an alternative to the homogeneous Poisson process for the arrival process, and a fully parametric survival curve for the service time. All models, going from the simple $M/G/\infty$ queue to the doubly-stochastic Poisson process queue, were discussed in the first section.

The objective of this new part is to compare the performance of the different models, both in terms of expected premium and premium deviation. It has already been shown that the homogeneous Poisson process introduces a bias in the expected number of contracts, but that both non-homogeneous and Cox processes lead to the same expected number of contracts. The next step is to determine how much this impacts the potential bias on the premium.

However, in the context of Solvency II, a measure of the Best Estimate, that is, of an expected value, is not sufficient enough to master the risk of an insurer. The regulator is more interested about the tails of the distributions of the risks that an insurer faces. In this context, the use of a doubly stochastic Poisson process makes much more sense, and we will try to conclude on its performance to properly measure the deviation of the overall insurance premium in the portfolio.

4.1 Modeling the premium

4.1.1 Background on the premium

Premium corresponds to the amount of money that the insurer is ready to accept from the policyholder to undertake its risk. That is, the premium can be understood as a function of insured's risk. Similarly, insured's risk depends

on its characteristics, may they be personal or property, or its past claim history.

Thus, those characteristics are not constant over time, since for instance the age of the policyholder may evolve, or because the policyholder reported some claims. It means that all those characteristics can be modeled as time processes, so is the premium.

Moreover, insurers also capture costs and environmental aspects in their pricing, such as inflation, that is also not constant over time. Finally, low-claim bonuses may also be applied in order to reward good risks in the portfolio, and similarly, bad risks may be penalized. This is for instance the case in the French bonus-malus scale for motor insurance, where good drivers are rewarded with a 5% decrease of their CRM (Coefficient Reduction Majoration) if they didn't face any responsible accident during the year, or be penalized by an increase of 25% of their CRM if they are responsible for a claim during the year. The impact on the premium may be very significant.

Hence, it is unlikely that a policyholder that stayed in the portfolio more than one year sees his premium constant over time. Optimizing the term of contracts is an actuarial field in itself, that requires a proper modeling. Taking such specificities into account in our model would lead to something way too complex, and would not be suitable. Thus, for simplicity purposes, and despite what has just been said, we will assume that the premium is constant over time for one insured. Mathematically speaking, if $(X_t^i)_{t>0}$ denotes the premium process for policyholder i , then $X_t^i \rightarrow X^i$

4.1.2 The collective model

The premium of an insurance contract is usually determined through the use of complex models, and is the subject of many actuarial papers or master thesis. In the P&C industry, Generalized Linear Models are often used, and are a standard in the market. See for instance [GKTG16] and [DJH⁺08].

While insurance pricing's objective is to set the most accurate price for a given risk or individual, this is not our focus here. Indeed, to look at our portfolio in an aggregate way, we don't need to know what is the exact premium of a particular contract, but more how the overall portfolio's premium behaves. Hence, as long as the overall distribution of premium is accurate in our portfolio, we don't give much care to know if the premium we set for an individual was the accurate one.

In this context, the quantity of interest for the premium is not the individual premium $(X_i)_{1 \leq i \leq n}$ for each policyholder in the portfolio, but rather the

aggregate premium of the portfolio defined by:

$$S = \sum_{i=1}^N X_i$$

We retrieve the framework of the classical collective model. The collective model is one of the well-known models for overall claim assessment, or to set a price to an insurance contract. Its idea is quite straightforward:

Claim assessment If S denotes the overall claim amount for a given year, N the overall number of claims during this given year, and $(X_i)_{1 \leq i \leq N}$ the individual amount of each of the N claims, then we get $S = \sum_{i=1}^N X_i$

Pricing If S denotes the overall claim amount for a given policyholder in one year, N the number of claims faced by the policyholder in one year, and $(X_i)_{1 \leq i \leq N}$ the individual amount of each of its N claims, then we get $S = \sum_{i=1}^N X_i$. The pure premium is defined by $\mathbb{E}(S)$.

For an introduction to the collective risk model, see [Dur13].

The purpose of this section is not to apply *stricto sensu* the theoretical results of the collective model, as this would not be suitable, but rather to show the extension possibilities of the overall model developed here, depending on its use. For instance, for customer lifetime value purposes, one could be interested in the behavior of a specific group of individuals. In this case, the premium associated to each individual must be accurate, and could require the use of the collective model to integrate the pricing methodology to X_i .

In our case, the random variable N , which in fact is a process, $N(t)$, has been modeled in the first section. We make the choice here to use a simple distribution for X_i , which will be the same for all policyholders. We keep the standard assumptions of the standard collective model that the X_i are independent and identically distributed, and that they are also independent of $N(t)$. This very last assumption will be discussed in section 5.

4.1.3 Model for the individual premium

The insurance portfolio we work on comprises multiple types of policyholders, that differ according to their characteristics. That is, their premium are not comparable *per se*. One could assume that the data could be modeled using some parametric distributions. Let's first take a look at the distribution of premium on figure 4.1.

The distribution shows one main mode that is slightly skewed to the right. Two classical parametric candidates could fit the data: the gamma distribution and the log-normal distribution. The log-normal is often used to model

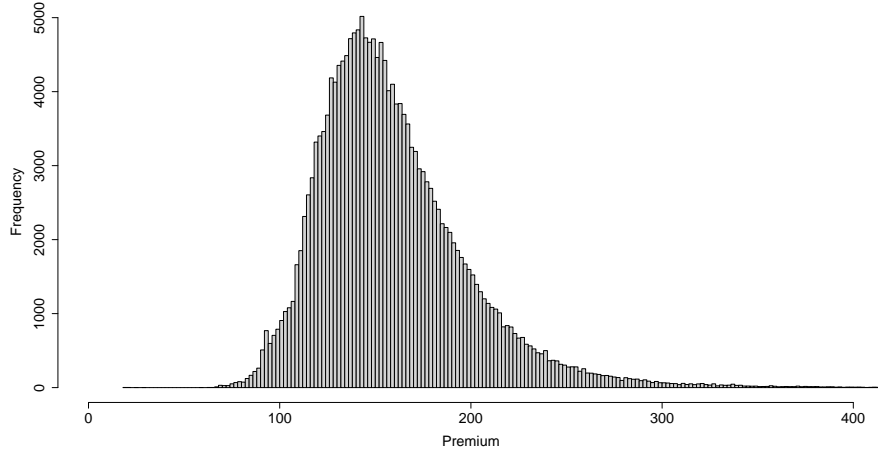


Figure 4.1: Empirical distribution of the premium

intensity for attritional claims. Since premium and claims are usually linked by the expected value, the log-normal may also be a good candidate for individual premium modeling.

Table 4.1 depicts the results of the fit for the log-normal distribution, where μ and σ are the mean-log and sd-log parameters, and table 4.2 depicts the results of the fit for the gamma distribution, where α and β are the shape and rate parameters.

Table 4.1: Fitting results of the log-normal distribution

$\hat{\mu}$	$\hat{\sigma}$	$\sqrt{\mathbb{V}(\hat{\mu})}$	$\sqrt{\mathbb{V}(\hat{\sigma})}$	AIC	BIC	$\ln \mathcal{L}$
5.0411	0.2341	0.00054	0.00038	1883071	1883091	-941533.3

Table 4.2: Fitting results of the gamma distribution

$\hat{\alpha}$	$\hat{\beta}$	$\sqrt{\mathbb{V}(\hat{\mu})}$	$\sqrt{\mathbb{V}(\hat{\sigma})}$	AIC	BIC	$\ln \mathcal{L}$
17.7496	0.1115	0.05725	0.000364	1891701	1891721	-945848.6

When it comes to AIC, the log-normal distribution seems to fit better the data. However, looking at figure 4.2, it seems that the right tail of the distribution is underestimated. The QQ-plot deviates from the straight line around the value 250, which corresponds to the quantile at 97%. The deviation appears to

be quite significant, since some premium might be underestimated by 33%.

One should carefully use this parametric estimate when modeling the premium distribution, as it is known that extremes could lead to significant deviations if they are underestimated. This advantage of working with individual data is that individual premium is available. For such massive portfolios, it can be assumed that most of the distribution is represented in the portfolio, and that a good alternative would be to sample directly from the empirical distribution.

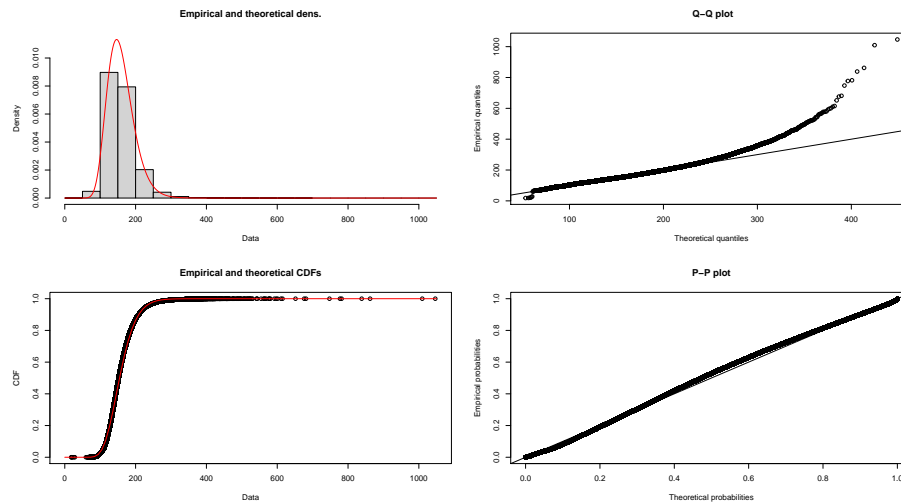


Figure 4.2: Fit of the log-normal distribution

In order to work with a fully parametric model, we will retain the log-normality assumption. However, one needs to carefully verify the hypotheses while using this model.

4.2 Performance comparison

4.2.1 Simulation and metrics

All the elements of the model are now modeled:

- The number of contracts through the process $N(t)$ which will be either homogeneous Poisson, non-homogeneous Poisson, or doubly stochastic
- The individual premium, which will be modeled through a log-normal distribution

The purpose of this section is to examine the distribution of the random variable $S = \sum_{i=1}^N X_i$.

Let $t = t_0$ be the starting point of our study. Let $N_0 = N(t_0)$ be the initial number of policyholders in the portfolio. This number is deterministic. We will also assume that for all N_0 initial policyholders, we know their records in the portfolio, that is, we know for how long they have been in the portfolio, and we know their constant premium $(X_i)_{1 \leq i \leq N_0}$. That is, there is no left-censoring. This hypothesis is realistic, otherwise it would mean that the insurer would have lost the track of its policyholders. Hence, let $(s_i)_{1 \leq i \leq N_0}$ be the current survival time of those N_0 policyholders.

The residual survival time can be defined by $S_{t_0}(t) = \mathbb{P}(T > t \mid T > t_0)$, with $t \geq t_0$. Standard survival analysis computation leads to:

$$\begin{aligned} \mathbb{P}(T > t \mid T > t_0) &= \frac{\mathbb{P}(T > t, T > t_0)}{\mathbb{P}(T > t_0)} \\ &= \frac{\mathbb{P}(T > t)}{\mathbb{P}(T > t_0)} \\ &= \frac{S(t)}{S(t_0)} \end{aligned}$$

Simulating the residual lifetime for a policyholder then requires to compute $S(t_0)$ for each of the N_0 initial policyholders. Then, one needs to restrict $S(t)$ to $[t_0, +\infty[$, and to scale it with $S(t_0)$. Sampling directly from the residual survival distribution using the algorithm depicted in section 3.2.2 would require to compute, for each t_0 , the new breaks of the residual survival function, which will be computationally not efficient. A very simple turnaround is to recall that the inverse transform method consists in finding t so that

$$u = \frac{S(t)}{S(t_0)}$$

for some u uniformly distributed on $[0, 1]$. This is equivalent to find some t so that:

$$u = S(t)$$

for u uniformly distributed on $[0, S(t_0)]$. In this configuration, one simply needs to apply algorithm of section 3.2.2, where u will be drawn from a uniform distribution on $[0, S(t_0)]$.

Then, one needs to simulate the arrival of new policyholders with the choice of the arrival process. All three homogeneous Poisson, non-homogeneous Poisson and Cox processes will be tested and compared. The next step is to simulate the survival for each of those newcomers, using the algorithm depicted in 3.2.2. The last step of the simulation is to simulate their individual premium, by sampling from the fitted log-normal distribution in section 4.1.3.

In order to get some statistics about the desired distribution, this simulation will be performed N times (to be defined), using Monte-Carlo techniques.

Two types of distributions will be considered:

- The number of contracts at the end of a given period
- The total premium collected during a given period

The following metrics will be of particular interest:

Mean : The sample mean should be close to each other since all arrival processes should lead to the same expected number of underwritten contracts.

Standard deviation : The sample standard deviation will be described by a path for the number of contracts simulation, while it will be a number for the total premium collected.

Confidence interval at 5% : This will be characterized by the sample quantiles at 97.5% and 2.5%. Again, it will be a path for the number of contracts simulation, while it will be numbers for the total premium collected.

We set the simulation period between 01/01/2019 and 31/12/2019.

4.2.2 Results

The results of the simulations for the number of contracts are depicted in figures 4.3, 4.4 and 4.5. The results of the simulations for the total premium collected are depicted in figures 4.6, 4.7 and 4.8. On each chart, the red line corresponds to the mean, and the green lines corresponds to the quantiles at 97.5% and 2.5%. 1,000 simulations have been performed.

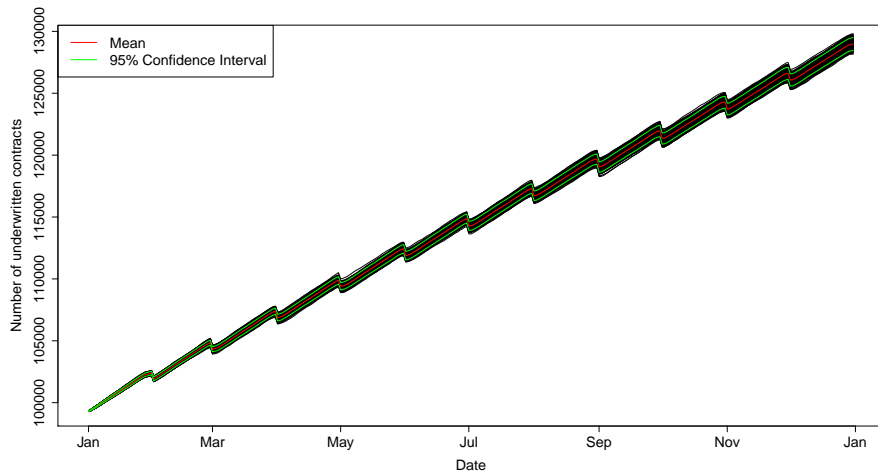


Figure 4.3: 1,000 sample paths of the number of contracts with a HPP

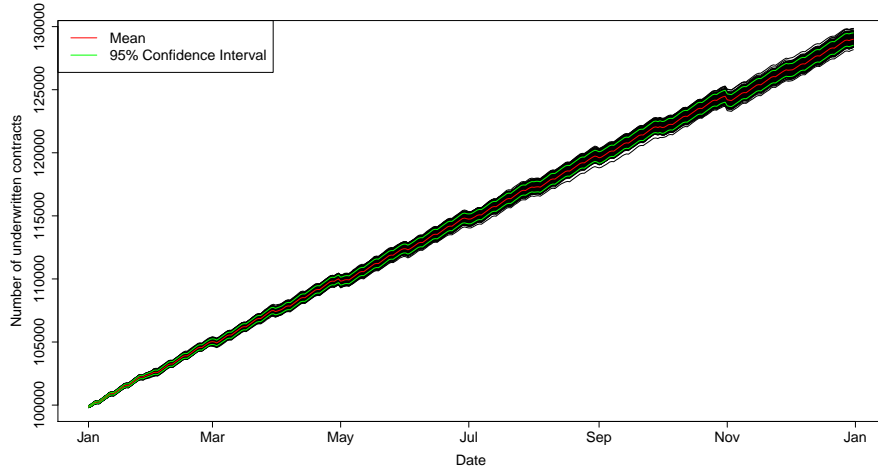


Figure 4.4: 1,000 sample paths of the number of contracts with a NHPP

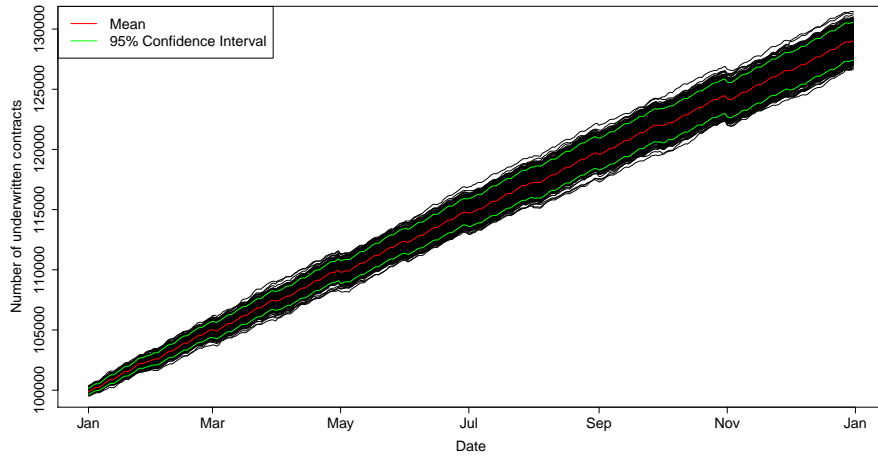


Figure 4.5: 1,000 sample paths of the number of contracts with a Cox Process

It is worth commenting first the aspect of the number of contracts curve. First, one can see steps the first day of each month. This is due to the cancellation of initial policyholders at their anniversary date. Moreover, we can see that there is a general increasing tendency. It means that the number of underwritten contracts exceeds the number of cancelled contracts.

Concerning the homogeneous Poisson process, the number of contracts evolves

nearly linearly between each first day of each month. This linearity is due to the arrival process (recall that the expected number of underwritten contracts is linear with t). The steps are nearly compensated on figure 4.4, which corresponds to the non-homogeneous Poisson process. This is due to the fact that the intensity function captures the higher number of underwritten contracts the first day of each month, which compensates the cancellation from initial policyholders on the first day of each month. The same applies for the Cox process case on figure 4.5. However, one can clearly see that the dispersion of the simulation curves behaves the same in the homogeneous Poisson and inhomogeneous Poisson, but that the spread is much higher in the Cox case. This highlights the fact that the Cox process enables to better capture overdispersion. One important fact is that overdispersion increases with time. That is, uncertainty increases with time. If, for strategic purposes, one needs to increase the projection time frame, then one would need to prefer the Cox process for arrival for risk assessment. One will need to rely on simulation-based estimates as closed-form formulas do not exist for the Cox case. On the other hand, if the quantity of interest is the average expected number, then two cases arise:

- If one projects at the end of one month, then one can simply use the homogeneous Poisson case.
- If one projects at some random time, then one has to use the inhomogeneous Poisson case. Indeed, using the homogeneous Poisson arrival would introduce a bias due to the increased steps at the beginning of each month.

In both cases, the advantage is that one can use closed-form formulas to avoid simulation-based estimates.

All histograms of figures 4.6, 4.7 and 4.8 are constructed with the same x and y axes, and the exact same bins width for matters of comparability.

Those histograms show similarities. First, they all contain one main mode. However, the values of the distributions seem to be concentrated around the main mode for the inhomogeneous Poisson process, while they seem more "normally" spread for the homogeneous Poisson and Cox processes. Indeed, looking at the kurtosis gives:

- $\kappa = -0.07321263$ for the homogeneous Poisson case
- $\kappa = -0.2264731$ for the inhomogeneous Poisson case
- $\kappa = 0.04289296$ for the Cox case

However, the tail is a bit larger in the inhomogeneous Poisson case compared to the homogeneous Poisson case. The significant difference lies in the Cox case, where the main mode is nearly flat, and the tails are thicker. This can be understood by the fact that in the homogeneous and inhomogeneous Poisson cases, the variance of the number of contracts doesn't contribute much

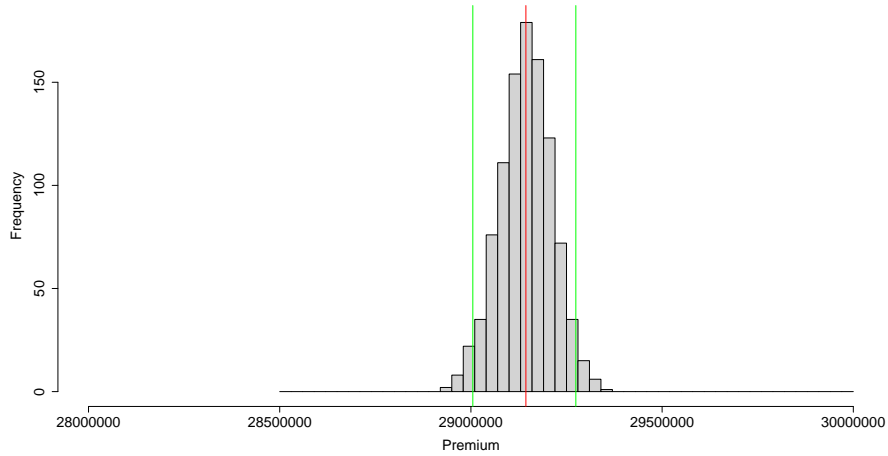


Figure 4.6: Distribution of the total premium collected with a HPP

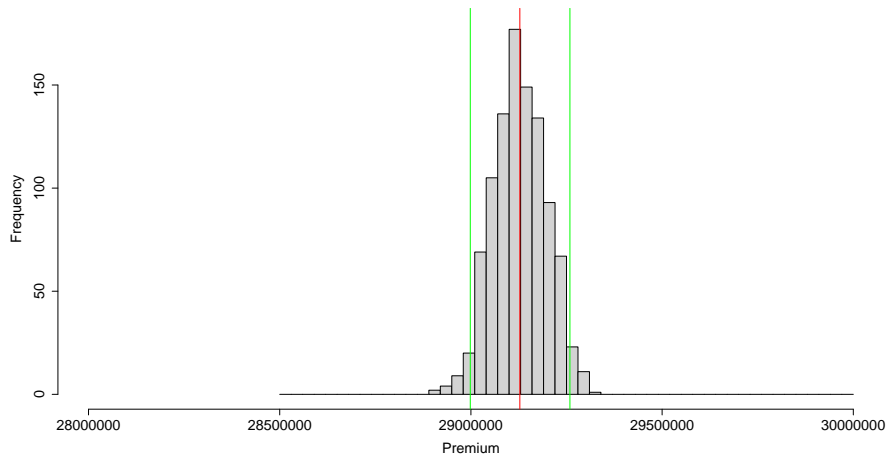


Figure 4.7: Distribution of the total premium collected with a NHPP

to the total premium collected variance. This is the opposite in the Cox case, where the variance of the number of contracts has a real influence on the total premium collected distribution.

Table 4.3 shows the mean, standard deviation, and quantiles at 2.5% and 97.5% for all three different arrival processes.

All means are comparable, which is coherent as the mean of the individual

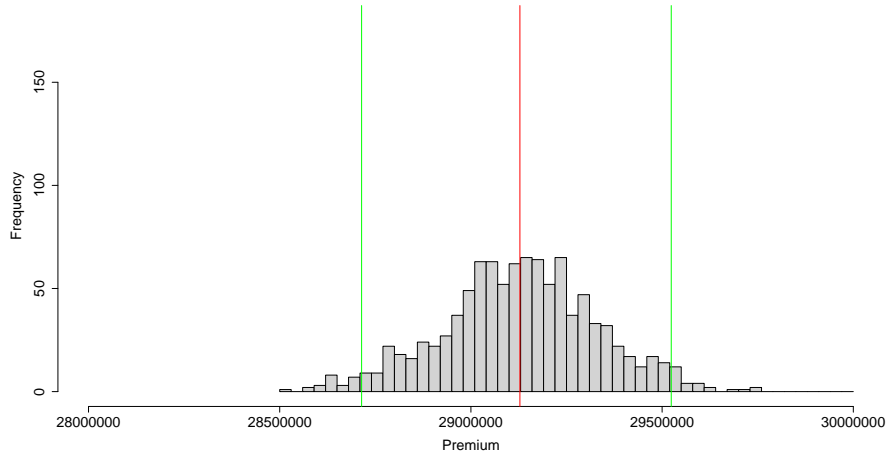


Figure 4.8: Distribution of the total premium collected with a Cox Process

Table 4.3: Total premium collected metrics

Arrival	Mean	Standard deviation	$q_{2.5\%}$	$q_{97.5\%}$
HPP	29,143,773	68,201.03	29,005,007	29,274,393
NHPP	29,127,959	69,153.42	28,998,356	29,258,911
Cox	29,128,120	201,569.8	28,714,173	29,523,873

premium and of the number of contracts are the same in all three models. The standard deviation and quantiles are really comparable between the homogeneous Poisson case and the inhomogeneous Poisson case. It means that making the model more complex using a non-homogeneous Poisson process doesn't improve the results in terms of variance of the total premium collected. However, this is not true for the Cox case, where the standard deviation and quantiles are much higher than the homogeneous and inhomogeneous Poisson cases. The tails are much thicker, and lead to more spread values. In that case, if one is concerned about measuring the risk of deviation of the premium, the Cox arrival process appears to be justified.

However, it is worth noting that the deviation of the premium is limited, in the sense that, even in the Cox case, the difference between the quantiles at 97.5% and 2.5%, and the mean is roughly 400,000, which is $\approx 1.37\%$ of the mean. Thus, the deviation is very limited. Of course, this is approximatively 3 times higher than the deviation measured for the homogeneous and inhomogeneous Poisson case, but still very limited. The use of such a model is thus

questionable, and will be discussed in section 5.

Chapter 5

Discussion and conclusion

The results of the model show significant differences in terms of dispersion. While the homogeneous and inhomogeneous Poisson processes show low dispersion, the Cox process has shown that it better captures the movement of policyholders in the portfolio, and that such movements have significant impacts on both the number of contracts in the portfolio, and the total premium collected. As measuring the premium deviation was the primary objective of the model, the results are satisfying.

On the other hand, the deviation is not significant, whatever the model. This result is questionable, since it is known that the premium of insurance portfolios may vary more than what has been modeled here. One of the main reasons is that the model doesn't take into account external environment behavior, but only the behavior of the portfolio in normal conditions. The external environment plays an important role, as this may be one of the main explanatory factor of underwritings and terminations. External environment may take into account legal and financial aspects. For instance, the Hamon law in France enables policyholders to cancel their insurance contract without penalty as soon as the underwriting period lasted 1 year. This encourages policyholders to always find better contracts, and thus to cancel and underwrite contracts more often. It also forces insurers to adapt their offers, to compensate. This adaptation goes for instance by reducing the premium, which is one of the main, if not the most important criteria for choosing a contract for equivalent guarantees. The reaction of a policyholder to price variation is called price elasticity, and has not been modeled in this thesis.

Indeed, the model simulates first a number of underwritten contracts, and then allocates a premium to each new contract, the arrival process being independent of the individual premium random variable. This hypothesis is extremely strong, and often inaccurate, as the reasoning is usually processed in the other way. It is because the price has varied that the the number of underwritten contracts varies.

Taking into account those two major aspects can be possible, by defining some new processes, say some environment process $(E_t)_{t>0}$ and premium process $(P_t)_{t>0}$. Those time processes would describe the way the environment has evolved over time, and the way the premium has evolved over time. While the environment process would be user-defined because the perception of environmental factors depends on each insurer, the premium process could be directly derived from the insurer records. Indeed, the insurer knows from his portfolio how the premium has evolved over time. Even more interesting, the insurer knows how the premium will evolve in the future, since pricing strongly relies on strategic decisions. That premium process would thus be a predictable process. Having modeled the environment process and premium process, one can then adapt the intensity function of either the non-homogeneous Poisson or the Cox process. Integrating the premium process in the intensity function would then enable to model the elasticity, and better capture the variability.

One of the other strong hypotheses that can be highlighted is the constance of the premium of a policyholder. This hypothesis is very questionable since the premium usually varies yearly, either because of the occurrence or not of a claim, and also because of external factors such as inflation. While inflation can be captured through the environment process (E_t) , it is not the case of revalorization due to the behavior of the policyholder himself. The revalorization mechanism has not been taken into account and plays a major role in the number of contracts behavior. For instance, a policyholder with frequent claims may be forced to cancel his policy, or may quit by himself if his premium increase would be too high. That is, the reason of the cancellation has not been used in the model, and plays a major role. Integrating this aspect in the model would require to model, say, the claims process (both in terms of frequency and intensity), and to take into account the decision process of the insurer to cancel a policy. This would make the model much more complex, and would integrate some "expert judgements", that are usually avoided when possible. Such a hypothesis appears questionable, but seems necessary to avoid an over-parametrization of the model.

In terms of applicability, this model requires the availability of individual data that are, for each policyholder, the underwriting date, the termination date (if applicable) and the individual premium. Normally, those data are easily available for an insurer. That is, the advantage of the model is that it relies on really basic information. A second advantage is that a very few years of data is necessary. While most models that work on an aggregate basis require many historic data, this is not the case here. Indeed, we could benefit from as many observations as the number of policyholders during our 4-year estimating period. Working with yearly or monthly data would have made us work with 4 or 48 points, which would have led to a high estimation bias.

On the other hand, working with individual data requires a lot of comput-

ing capacity. The 1,000 simulations performed on section 4.2.1 required each 20 minutes of running time, for up to 130,000 policyholders simulations. The time required is mostly due to the survival simulation algorithm, that is more complex than Poisson or Cox processes simulation. This is suitable for small portfolios with small projection periods, but it is not for larger portfolios and larger projection periods, which could take hours to run, as well as a lot of memory capacity to store all the simulation results. Choosing simpler survival function may be a good turnaround on the running time, but would lead to a simulation bias. A choice has to be made between the precision of the model and the time to run it.

Another field of application of this model would be to use it for pricing purposes. Indeed, pricing actuaries may be interested in applying the model to predict if the customer would be profitable in the portfolio or not. For instance, the pricing algorithm could propose better prices to potential policyholders for whom it is predicted a long survival in the portfolio, and on the other hand to increase the price for those who are likely to stay a very few time. This can be achieved through the survival function. Indeed, here, a simple parametric function has been used to model the survival behavior. However, if one insurer has access to basic data required to model, it is also likely to have access to a much wider range of variables, that are usually valuable for pricing purposes. Thus, the pricing actuaries may be interested to study the survival depending on the characteristics of each of them. Such a consideration could be possible using a survival model like the proportional hazard Cox model. It integrates covariates that are likely to explain the termination behavior of policyholders, see [Wan15]. Those covariates could also be used to better fit the underwriting behavior, as adverse selection and information asymmetry could also be interpreted from those covariates.

To conclude, this study drafts an extension of the model of Jean-Philippe Boucher [BCP16]. Even if the results are satisfying in the sense that the variability has been better modeled, it is not fully captured as some of the variability may be explained from exogenous factors, or from some of the very strong hypotheses that have been adopted, but that are likely to be unverified empirically. Some of those hypotheses can be relaxed, but require an increased complexity of the overall model. However, some potential for improvement has been identified, and could be the fruit of future work. Finally, this model opens many doors for applying it in many actuarial fields, from internal modeling, to pricing. It also confirms that individual data play a key role in actuarial models, and gives many opportunities.

Bibliography

- [AAAB19] Hansjoerg Albrecher, Jose Carlos Araujo Acuña, and Jan Beirlant. Fitting Nonstationary Cox Processes: An Application to Fire Insurance Data. *North American Actuarial Journal*, 2019.
- [AMMV17] Zeina Al Masry, Sophie Mercier, and Ghislain Verdier. Approximate Simulation Techniques and Distribution of an Extended Gamma Process. *Methodology and Computing in Applied Probability*, 19(1):213–235, 2017.
- [BCP16] Jean-Philippe Boucher and Guillaume Couture-Piché. Modeling the number of insured households in an insurance portfolio using queuing theory. *ASTIN Bulletin*, 46(2):401–430, 2016.
- [BD17] Alexandre Boumezoued and Laurent Devineau. Individual Claims Reserving: a Survey. November 2017. working paper or preprint.
- [Cas05] Roberto Casarin. Stochastic Processes in Credit Risk Modeling. *SSRN Electronic Journal*, 02 2005.
- [DJH⁺08] Piet De Jong, Gillian Z Heller, et al. Generalized Linear Models for Insurance Data. *Cambridge Books*, 2008.
- [DP19] Andrew Daw and Jamol Pender. On the Distributions of Infinite Server Queues with Batch Arrivals. 2019.
- [DQLD19] Li Dongmin, Hu Qingpei, Wang Lujia, and Yu Dan. Statistical inference for $M_t/G/\infty$ queueing systems under incomplete observations. *European Journal of Operational Research*, 279(3):882 – 901, 2019.
- [GGC93] Richard D. Gelber, Aron Goldhirsch, and Bernard F. Cole. Parametric Extrapolation of Survival Estimates with Applications to Quality of Life Evaluation of Treatments. *Controlled Clinical Trials*, 14(6):485–499, 1993.
- [GKTG16] Mark Goldburd, Anand Khare, Dan Tevet, and Dmitriy Guller. Generalized Linear Models for Insurance Rating. *Casualty Actuarial Society, CAS Monographs Series*, 5, 2016.

- [GTP13] Ion Grama, Jean-Marie Tricot, and Jean-François Petiot. Long Term Survival Probabilities and Kaplan-Meier Estimator. September 2013. 14 pages.
- [KM58] E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.
- [Liu12] Chung-Yu Liu. Claim Count Modeling with Shot-noise Cox Processes. *University of South Wales*, 2012.
- [Mcq10] Patrick Mcquighan. Simulating the Poisson process. 2010.
- [Mer17a] Sophie Mercier. Probabilistic Construction And Properties Of Gamma Processes And Extensions. 2017.
- [Mer17b] Sophie Mercier. Probabilistic Construction and Properties of Gamma Processes and Extensions. *MMR 2017 (International Conference on Mathematical Methods on Reliability)*, 2017.
- [Mir63] Noel M. Mirasol. The Output of an $M/G/\infty$ Queuing System is Poisson. *Operations Research*, 11(2):282–284, 1963.
- [PW12] Guodong Pang and Ward Whitt. Infinite Server Queues With Batch Arrivals And Dependent Service Times. *Probability in the Engineering and Informational Sciences*, 26(2):197–220, 2012.
- [Ruw06] Christel Ruwet. Processus de poisson. *Université de Liège*, 2006.
- [Sha66] D. N. Shanbhag. On Infinite Server Queues with Batch Arrivals. *Journal of Applied Probability*, 3(1):274–279, 1966.
- [Sig10] Karl Sigman. Lecture Notes on the Inverse Transform Method. 2010.
- [Đur13] Zlata Đurić. Collective Risk Model in Non-life Insurance. *Ekonomski horizonti*, 15(2):163–172, 2013.
- [Wan15] Hongyuan Wang. Estimating Insurance Attrition Using Survival Analysis. 2015.