risks

MDPI

Article

Classifying Insurance Reserve Period via Claim Frequency Domain Using Hawkes Process

Adhitya Ronnie Effendie [1,*], Kariyam [2], Aisya Nugrafitra Murti [1], Marfelix Fernaldy Angsari [1] and Gunardi [1]

[1] Department of Mathematics, Gadjah Mada University, Sekip Utara BLS 21, Yogyakarta 55281, Indonesia
[2] Department of Statistics, Universitas Islam Indonesia, Jl. Kaliurang Km 14.5, Sleman, Yogyakarta 55584, Indonesia
* Correspondence: adhityaronnie@ugm.ac.id

Abstract: In this paper, the insurance reserve period will be classified according to the claim frequency domain, such as high- or low-frequency periods. We use the clustering method to create and group claims data according to their frequency period. Meanwhile, we use a risk process to mimic and predict the movement of the reserve from time to time in each group of claim period that is formed. The risk process model used here is the Hawkes process, which is a one-dimensional simple point process and a special type of self-exciting process. Based on this process, we will estimate the reserve at a certain date in the future and the average historical reserve for each group period.

Keywords: cluster analysis; risk process; Hawkes process; insurance reserve process

check for updates

Citation: Effendie, Adhitya Ronnie, Kariyam, Aisya Nugrafitra Murti, Marfelix Fernaldy Angsari, and Gunardi. 2022. Classifying Insurance Reserve Period via Claim Frequency Domain Using Hawkes Process. *Risks* 10: 216. https://doi.org/10.3390/risks10110216

Academic Editors: Ingrid-Mihaela Dragotă, Simona Laura Dragos and Cepoi Cosmin Octavian

Received: 6 September 2022
Accepted: 4 November 2022
Published: 14 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In non-life insurance, a huge amount of personal experience exists with some extremely big hazards. A multi-billion-dollar manufacturing company might, for instance, take out workers' compensation, commercial liability, and property damage insurance for its several plants. An insurer may be able to determine the appropriate reserve for these risks using the risk's unique experience.

However, it is not prudent to establish reserves only using the historical behavior of the risk for the entire claim period and ignoring the pattern of claims frequency that may occur within the claim periods. In these situations, it is necessary to analyze risks using classification of risk process according to the claim frequency domain, which involves grouping risks with comparable loss potential and imposing new various manual rates to accommodate for variations in loss potential within the groupings.

Moreover, we can use risk theory to model random processes that will explain the occurrence of claims. The traditional Cramer-Lundberg risk model, developed by Lundberg (1903), has undergone several studies in recent years. Lefevre and Picard (2006) for instance, developed a risk model in which the successive claim amounts are independent and non-identically distributed while Romaniuk (2019) described the application of fuzzy sets to the risk process, especially fuzzy numbers to model the possible penalties which are related to the ruin event. Many studies, including Raducan et al. (2015), describe a risk process with claim costs modeled by Erlang or a mixture of exponential distributions and examine the calculation of ruin probability with independent and non-homogeneously distributed claim costs.

The application of self-exciting Hawkes processes in risk theory, which was first introduced by Hawkes (1971), is also covered in several works. Due of their capacity to reflect endogenously generated clustering, these processes have attracted attention. This method is used by Swishchuk et al. (2021) because it enables the use of practical diffusion approximation results from the risk model with claims arrivals based on a general compound Hawkes process using real empirical data, allowing for the realistic representation of claim costs.

Furthermore, Jeong and Zou (2022) proposed a dynamic credibility model for claim count that extends the benchmark Poisson generalized linear models (GLM) by incorporating self-excitation and exponential decay features from Hawkes processes. Bessy-Roland et al. (2021) proposed a multivariate Hawkes framework for modeling and predicting cyber-attacks frequency. They demonstrated the ability of Hawkes models to capture self-excitation and interactions of data-breaches depending on their type and targets.

In this paper, we provide a novel approach for concurrently categorizing and estimating insurance reserves using the Hawkes process. In this approach, we first categorize the data in accordance with the frequency of claims and then use the Hawkes process risk model for each newly formed group to estimate its reserve. This method has the advantage of allowing us to classify claim data by the risk category involved while also estimating reserves. In addition, there are two phenomena that are common to reserves over time. First, past claims tend to increase reserves, and second, recent claims have a greater impact on reserves than outdated claims. These phenomena are very suitable to be modeled with the Hawkes process which has two main features, namely self-excitation and exponential decay.

This paper is organized as follows: Section 2: Preliminary research and Methodologies, which consists of a discussion of the cluster analysis used in this study to classify the claim period and the Hawkes process that we use in modeling the movement of reserves. Then, in Section 3: Results and Discussions, we describe the data we used in this study. Then we perform a clustering procedure to obtain groups of different claim periods based on their frequency. After that, we continue with the reserve process modeling using the Hawkes process for each claim period formed from the previous steps. At the end of this section, we discuss some of the important findings of our methodology and further study opportunities to develop better models. The last, Section 4, is devoted to the conclusion.

2. Preliminary Research and Methodologies

Two significant methodologies that underpin our study are described in this section: pairing the Hawkes process with cluster analysis.

2.1. Cluster Analysis

Cluster analysis is a part of multivariate analysis that aims to partition objects into several clusters. Each cluster is homogeneous concerning the specific characteristics, i.e., observations in each group are similar to each other. Each group should be different from other groups concerning the same features; that is, the objects of one group should be different from the things of other groups. The principle of clustering objects is to obtain a measure of proximity that explains the similarity between objects. This proximity between objects is calculated based on the type of data.

2.1.1. Hierarchical Method

The hierarchical clustering method moves forward either through a series of subsequent divisions or mergers. The individual items are the first step in the agglomerative hierarchical technique. In the beginning, there are therefore as many clusters as objects. First, groups of the objects that share the most similarities are formed, and then these initial groups are combined based on their commonalities. All the subgroups eventually combine into a single cluster when the similarity lowers. See for example Kaufman and Rousseeuw (1990).

As in Johnson and Wichern (2009), the agglomerative hierarchical clustering for N objects is such as below:

(i) Start with N clusters, each containing a single entity and an $N \times N$ symmetric distances (or similarities) $D = \{d_{ik}\}; i = k = 1, 2, 3, ..., n$. In this paper, we use the

Euclidean distance. Suppose object i and object j on data sets, then the Euclidean distance is formulated as follows:

$$d(i,j) = \sqrt{\sum_{l=1}^{p} (x_l^i - x_l^j)^2}. \tag{1}$$

where p is the number of variables.

(ii) Search the nearest (most similar) pair of objects (e.g., object or cluster U and object or cluster V) from the distance matrix.

(iii) Merge clusters U and V, label them (UV) and then update the distance matrix entries by

(a) deleting the rows and columns corresponding to clusters U and V and

(b) adding a row and column indicating the distance between the merged cluster (UV) and the remaining clusters.

(iv) Repeat steps (ii) and (iii) an additional $N-1$ times. (All objects will be in a single cluster after the algorithm terminates).

This paper uses complete linkage to merge clusters U and V in Step 3. Suppose two cluster U and cluster V have the nearest distance d_{UV}, with d_{UW} is the distance between cluster U and cluster W, while d_{VW} is the distance between cluster V and cluster W. The distance between cluster (UV) and any cluster W is computed by

$$d_{(UV)W} = \max(d_{UW}; d_{VW}). \tag{2}$$

2.1.2. Estimation of the Number of Clusters

One of the problems that often draws attention in cluster analysis is determining the optimal group number. Several methods for estimating the number of clusters were developed using a measure of homogeneity within a group W_k. See for instance Johnson and Wichern (2009), which was formulated as follows:

$$\underset{p\times p}{W_k} = \sum_{g=1}^{k}\sum_{i=1}^{n_g} (\underset{p\times 1}{x_{g_i}} - \underset{p\times 1}{\bar{x}_g})(\underset{p\times 1}{x_{g_i}} - \underset{p\times 1}{\bar{x}_g})^t. \tag{3}$$

where p is the number of variables, n_g is the number of group g, $(g = 1, 2, ..., k)$, x_{g_i} is the ith object $(i = 1, 2, ..., k)$ on the group g and $\bar{x}_g$ is the average of variable in the group g.

Hartigan (1975) proposes the index to determine the number of clusters based on a trace of homogeneity within groups. Hartigan's index is defined as follows:

$$H(k) = \left(\frac{\mathrm{tr}(W_k)}{\mathrm{tr}(W_{k+1})}\right)(n-k-1) \tag{4}$$

where $\mathrm{tr}(W_k)$ is the trace of a square matrix W_k. The number of clusters is the smallest of $k(k \geq 1)$, so Hartigan's index is less than ten $(H(k)) \leq 10$.

2.2. Hawkes Process

2.2.1. Hawkes Process for Claim Arrivals Distribution

Hawkes process is a one-dimensional simple point process and a special type of self-exciting process. It was introduced by Hawkes (1971). Consider $N(t), t \geq 0$, a counting process with history $\mathcal{H}(t), t \geq 0$ that satisfy:

$$\mathbb{P}(N(t+h) - N(t) = m|\mathcal{H}(t)) = \begin{cases} \lambda^*(t)h + o(h), & m = 1 \\ o(h), & m > 1 \\ 1 - \lambda^*(t)h + o(h), & m = 0. \end{cases} \tag{5}$$

A simple and conditional orderly self-exciting point process N on $[0, \infty)$ is called Hawkes process if the conditional intensity function is in the form:

$$\lambda^*(t) = \lambda + \int_0^t \mu(t-s)dN(s). \tag{6}$$

where $\lambda > 0$ is the background rate of the process N and where $\mu(\cdot)$ is a function which governs the clustering density of N. The function $\mu(\cdot)$ is sometimes called the exciting function or the excitation function of N and to avoid treating the homogeneous Poisson process as a trivial case, we assume that $\mu(\cdot) \neq 0$.

In addition, the function (6) also can be written as :

$$\lambda^*(t) = \lambda + \sum_{t_i<t} \mu(t-t_i) \tag{7}$$

with the observed sequence of past arrival $\{t_1, t_2, ..., t_k\}$, basic conditional intensity $\lambda > 0$ and excitation function $\mu(\cdot)$.

One of the common choices for excitation function is the exponential decay function. In this case, $\mu(t) = \alpha e^{-\beta t}$ with $\alpha, \beta > 0$, so that the exponentially decaying intensity function becomes

$$\lambda^*(t) = \lambda + \int_0^t \alpha e^{-\beta(t-s)}dN(s) = \lambda + \sum_{t_i<t} \alpha e^{-\beta(t-t_i)}. \tag{8}$$

In this case, α and β can be interpreted as each arrival in the system instantaneously increases the arrival intensity by α, then over time this arrival's influence decays at rate β.

To model a process from some time before the beginning of the observation period or after it is started, initial condition $\lambda^*(0) = \lambda_0$ can be set. The stochastic differential equation

$$d\lambda^*(t) = \beta(\lambda - \lambda^*(t))d(t) + \alpha dN(t), \quad t \geq 0 \tag{9}$$

must be satisfied for this scenario. If the stochastic equation above applied, the general solution can be written as

$$\lambda^*(t) = e^{-\beta t}(\lambda_0 - \lambda) + \lambda + \int_0^t \alpha e^{-\beta(t-s)}dN(s), \quad t \geq 0, \tag{10}$$

which is an extension of (8).

We restrict ourselves to the Markovian Hawkes process in this study. Generally speaking, the Hawkes process can be non-Markovian. When the excitation function is represented by the sum of exponential functions (8), the Hawkes process is Markovian. See, for instance Gao and Zhu (2017).

2.2.2. Distributional Properties and Auto Correlation Function

Da Fonseca and Zaatour (2013) demonstrated that for a Hawkes process with the conditional intensity function follows (9), the expected value and variance of the number of jumps during a τ-length interval when t goes to infinity are given by

$$\mathbb{E}[N(\tau)] = \lim_{t\to\infty} \mathbb{E}[N(t+\tau) - N(t)] = \frac{\lambda}{1-\frac{\alpha}{\beta}}\tau \tag{11}$$

$$\begin{aligned} \mathrm{Var}[N(\tau)] &= \lim_{t\to\infty} \mathbb{E}[\{N(t+\tau) - N(t)\}^2] - \mathbb{E}[N(\tau)]^2 \\ &= \frac{\lambda\beta}{\beta-\alpha}\left[\tau\left(\frac{\beta}{\beta-\alpha}\right)^2 + \left(1 - \left(\frac{\beta}{\beta-\alpha}\right)^2\right)\frac{1-e^{-\tau(\beta-\alpha)}}{\beta-\alpha}\right]. \end{aligned} \tag{12}$$

Then, as t goes to infinity, the following formula can be used to calculate the covariance of the arrival rate given two τ-length intervals with lag $\delta > 0$ that are not overlapping:

$$\begin{aligned}\mathrm{Cov}(N(\tau),\delta) &= \lim_{t\to\infty}(\mathbb{E}[(N(t+\tau)-N(t))(N(t+2\tau+\delta)-N(t+\tau+\delta))] \\ &\quad -\mathbb{E}[N(\tau)]\mathrm{E}(t+\tau+\delta,\tau)) \\ &= \frac{\lambda\beta\alpha(2\beta-\alpha)(e^{(\alpha-\beta)\tau}-1)^2}{2(\alpha-\beta)^4}e^{(\alpha-\beta)\delta}.\end{aligned} \tag{13}$$

Keep in mind that this necessitates the process's stability, which suggests $\alpha < \beta$. As a consequence of this result, we can obtain the auto correlation function of the process as the following:

$$\mathrm{Acf}(\tau,\delta) = \frac{e^{-2\beta\tau}(e^{\alpha\tau}-e^{\beta\tau})^2\alpha(\alpha-2\beta)}{2(\alpha(\alpha-2\beta)(e^{(\alpha-\beta)\tau}-1)+\beta^2\tau(\alpha-\beta)}e^{(\alpha-\beta)\delta}. \tag{14}$$

2.2.3. Parameter Inference

The maximum likelihood estimation method is one way to fit empirical data to an exponentially decaying Hawkes process. With this approach, as in Laub et al. (2015), we first identify the (log-) likelihood function and estimate the model parameters as the inputs that maximize this function. The log-likelihood function of Hawkes process for the interval $[0, T]$ can be written as

$$\ell = \sum_{i=1}^{N(T)}\log(\lambda+\alpha A(i))-\lambda T+\frac{\alpha}{\beta}\sum_{i=1}^{N(T)}(e^{-\beta(T-t_i)}-1). \tag{15}$$

where

$$A(i) = \begin{cases}0, & i=1\\ \sum_{j=1}^{i-1}e^{-\beta(t_i-t_j)} = e^{-\beta(t_i-t_{i-1})}(1+A(i-1)), & i\in\{2,...,k\}].\end{cases} \tag{16}$$

It is clear that maximum likelihood estimation will usually be very effective for model fitting. However, Filimonov and Sornette (2015) found certain drawbacks to using this method, including problems with bias in small sample sizes but performance issues with large samples. Another numerical disadvantage that might occur with this method is the outcome can be constrained to local optimization. These problems served as the impetus for Da Fonseca and Zaatour (2013) to develop the method of moments for parameter estimation. Zaatour (2014) developed an R package |hawkes|implementing this routine in C++ in an attempt to mitigate the performance issues. The recent trend of performing parameter estimation using the generalized method of moments is primarily due to this "performance bottleneck". Da Fonseca and Zaatour (2013) claim that the process is "immediate" in their test sets. The technique makes use of sample moments and the sample autocorrelation function, which are smoothed using a (ad hoc) user-selected process.

2.2.4. Goodness of Fit by Simulations

With regard to our situation, we will use Hawkes process simulation to assess the model's fitness. Ogata (1981) shows a simulation method of point processes, one of which is the Hawkes' self-exciting process, with the intensity function stated in (6) and the log-likelihood function stated in (15). For the simulations, Ogata (1981) used the likelihood function of the simulated data rewritten as:

$$L_t(\alpha_0,...,\alpha_p,\beta) = \sum_{i-1}^{n}\log\{\mu+\sum_{j=0}^{p}\alpha_jR_j(i)\}-\mu T-\sum_{i=1}^{n}\sum_{j=0}^{p}\alpha_jS_j(T-t_i). \tag{17}$$

The function $R_j(i)$ and $S_j(i)$ are given recursively for $j = 0, 1, 2, ...$ and $i = 2, 3, ...$, as:

$$R_j(i) = A_j(t_i - t_{i-1}) + \sum_{k=0}^{j} {}_jC_k A_{j-k}(t_i - t_{i-1}) R_k(i-1). \tag{18}$$

and

$$S_{j+1}(t) = \frac{\{(j+1)S_j(t) - A_{j+1}(t)\}}{\beta}. \tag{19}$$

where ${}_jC_k$ denotes a binomial coefficient and initial values are set as:

$$\lambda(t_{n+1}|t_1, ..., t_n) = \mu + \sum_{j=0}^{p} \alpha_j R_j(n+1). \tag{20}$$

2.3. Markov Chain for Claim Cost Distribution

In actuarial studies, the claim costs $\{Y_i\}$ are typically assumed to be independent and identically distributed with distribution D. We also assumed $m_1 = \mathbb{E}[Y_1]$ and $m_2 = \mathbb{E}[Y_1^2]$ are the first and the second finite moments of D, respectively. We follow Swishchuk et al. (2021) and apply the theoretical findings for general compound Hawkes processes, which they assumed independence between subsequent claim costs.

Hawkes process could be used to model arrival sequence to a system over time (Ogata (1999)). Hawkes process arrivals and independent and identically distributed claim costs were examined by Stabile and Torrisi (2010) and Cheng and Seol (2020) in relation to risk models. However, for those models, it is difficult to immediately apply the result to empirical data, and the only way to determine ruin probabilities is numerically. Then, based on the empirical distribution of observed claim costs, we approximate an independent and identically distributed sequence using the Markov chain (Z_i) and the function $a(z)$. By expanding the Markov chain's N number of states, the approximation can be made to fit any set of data arbitrarily well.

If G is the largest claim cost that has been observed and $\hat{D}$ is the empirical distribution function of claim costs, then $\hat{D}(G)$ equals 1. We created equally spaced boundaries $(g_1, g_2, ..., g_N = G)$ and specify $p^\star = (p_1^\star, ..., p_N^\star)$ as

$$\begin{aligned} p_1^\star &= \hat{D}(g_1) \\ p_2^\star &= \hat{D}(g_2) \\ &\cdots \\ p_N^\star &= \hat{D}(g_N) - \sum_{i=1}^{N-1} p_i^\star = 1 - \sum_{i=1}^{N-1} p_i^\star \quad . \end{aligned} \tag{21}$$

Hence, by the definition $\sum_{i=1}^{N} p_i^\star = 1$.

Consider that Z_i is a Markov chain on finite state space $\{1, ..., N\}$ with

$$P = \begin{pmatrix} p_1^\star & p_2^\star & \cdots & p_N^\star \\ \cdots & \cdots & \cdots & \cdots \\ p_1^\star & p_2^\star & \cdots & p_N^\star \end{pmatrix} = \begin{bmatrix} p^\star \\ \vdots \\ p^\star \end{bmatrix} \tag{22}$$

as its transition matrix. Please note that because in a finite state space, the elements of P are the same for each row, then (Z_i) is an irreducible Markov chain. This means the process can go from any state to any state, whatever the number of steps it requires.

Furthermore, for each state $k \in Z$ we can see that the columns of P remain constant, thus $\mathbb{P}(Z_{i+1} = k|Z_i = j) = \mathbb{P}(Z_{i+1} = k|Z_i = I) = p^\star_k$ where $\forall j, I \in N$. We can also have the following:

$$\mathbb{P}(Z_{i+1} = k) = \sum_{j \in Z} \mathbb{P}(Z_{i+1} = k|Z_i = j) = p^\star_k \sum_{j \in Z} \mathbb{P}(Z_i = j) = p^\star_k \qquad \forall k \in Z, i \in \mathbb{N} \tag{23}$$

using the total probability rule and Markov property. The probability of realizing one state is also independent of the probability of realizing the preceding state, therefore (Z_i) completely defines an independent and identically distributed sequence. We then let Y be a random variable with distribution function $\hat{D}$, $A_i := \{\omega : Y(\omega) \in (g_i - 1, g_i]\}$, and set

$$\begin{aligned} a(i) &= E[Y|g_{i-1} < Y \le g_i] = E[Y|A_i] \\ &= \frac{\mathbb{E}[Y\mathbf{1}_{A_i}]}{\mathbb{P}(A_i)} = \frac{\mathbb{E}[Y\mathbf{1}_{A_i}]}{p^\star_i}. \end{aligned} \tag{24}$$

Therefore, the expected value of the claim cost can be derived as follows:

$$a^\star = \sum_{i=1}^{N} p^\star_i a(i) = \sum_{i=1}^{N} \mathbb{E}[Y\mathbf{1}_{A_i}] = \mathbb{E}[Y]. \tag{25}$$

In this case, $a(Z_i)$ depicts an independent and identically distributed sequence that close to approximating the distribution $\hat{D}$ in terms of the number of states N goes to infinity.

2.4. *Hawkes Process Implementation in Risk Modeling*

A risk model typically takes the form of the following and aims to describe the capital that an insurance company, or a portion of it, will have available over time:

$$R = u + ct - \sum_{i=1}^{N_t} Y_i. \tag{26}$$

In the above formula, u is the underlying initial capital while c is the premium rate whose value does not change over time. $N(t)$ is a counting process that portraying the quantity of cases happening in the time interval $[0, t]$ while $\{Y_i\}$ is arrangement of non-negative random variables depicting the claim cost. Typically, the $\{Y_i\}$ are independent and identically distributed non-negative random variables with distribution F and $E[Y_1] = m_1$ and $E[Y_1^2] = m_2$. $N(t)$ and $\{Y_i\}$ are also thought to be independent. In the traditional situation, $N(t)$ is a homogeneous Poisson measure; however, going forward, we will consider that $N(t)$ is a fixed Hawkes measure with extremely decaying effect.

Definition 1 (Risk Model with General Compound Hawkes Process Swishchuk (2017))**.** *Let $N(t)$ be any one-dimensional Hawkes process. Let (Z_i) be an ergodic continuous-time finite (or countably infinite) Markov Chain, independent of $N(t)$, with state space Z, and let $a(z)$ be any bounded function*

$$H(t) = H(0) + \sum_{i=1}^{N(t)} a(Z_i). \tag{27}$$

The risk process $R(t)$ based on a general compound Hawkes process can be defined as

$$R(t) = \sum_{k=1}^{n} u_k + ct - \sum_{k=1}^{n} \sum_{i=1}^{N(t)} a_k(Z_i). \tag{28}$$

where n is the number of groups by clustering.

In this definition, (Z_i) is a continuous-time Markov chain on the state space $Z = \{1, ..., n\}$, while $N(t)$ is a Hawkes process. In addition, $a(z)$ is a bounded function on Z while we also assume $N(t)$ and (Z_i) are independent.

Theorem 1 (Law of Large Numbers for General Compound Hawkes Processes Swishchuk (2017))**.** *Let $R(t)$ be the risk model defined in Definition 1, and let (Z_i) be a Markov Chain with state space Z and stationary probabilities $p_i^\star$.*
Suppose that $0 < \hat{\mu} = \int_0^\infty \mu(s)ds < 1$. Then

$$\lim_{t\to\infty} \frac{R(t)}{t} = c - a^\star \frac{\lambda}{1-\hat{\mu}}. \tag{29}$$

where $a^\star = \sum_{i\in Z} a(i)p_i^\star$.

Corollary 1 (Net Profit Condition and Premium Principle Swishchuk (2017))**.** *The net profit condition for general compound Hawkes processes, is given as*

$$c > a^\star \frac{\lambda}{1-\beta}. \tag{30}$$

As a result of the expected value principle, the premium rate for general compound Hawkes processes is provided as follows:

$$c = (1+\theta)a^\star \frac{\lambda}{1-\hat{\mu}}. \tag{31}$$

where θ denotes the safety loading.

3. Result and Discussions

We took claims data from an Indonesian insurance company for the period December 2011 to January 2013, for which data were accessible. We selected a company that is a significant player in Indonesia's insurance market, with total assets exceeding 5 trillion rupiah and membership in companies with a 65% national market share, to ensure we obtain comprehensive data, adequate exposure, and a lengthy data period. Without prejudice to generality, we expected that these data could represent insurance company data in general, both in terms of time and financial context where our approaches might be applied.

Overall, the frequency of claims from December 2011 to January 2013 was 1566 applications approved with a total claim of 11.700 billion Indonesian rupiah. We first clean the data by removing claims with values less than or equal to zero. Then, we group the data by monthly intervals and count the total claims and the total number of claims approved in each interval. This value was approved within a period of 221 days out of 427 days including holidays in the 14 months. This means that an average of 7.09 cases were approved on each claim approval date or an average of 3.67 cases per day.

3.1. Clustering

In this section, we classify the data into groups based on the claim frequency domain. The clustering phases are preprocessed by normalizing the data and using the value range of each variable as the data divisor. Additionally, a hierarchical technique with complete linkage and Euclidean distances will be used to group the 14 monthly data. The Hartigan index is used to calculate the number of groupings. The Hartigan index plot reveals that, as shown in Figure 1, the smallest group with an index below 10 is two groupings. Therefore, there will be two clusters produced in this dataset. We developed our own syntax to calculate and present the Hartigan index with the help of the |`Minitab`20.4| software.

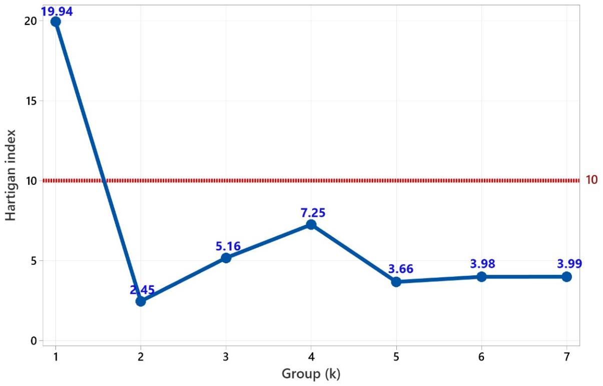

Figure 1. Hartigan Index plot to determine the number of groups from Indonesian Insurance data.

Next, we apply the hierarchical method with complete linkage to produce a dendrogram plot using |`Minitab`20.4| software as follows:

Figure 2. shows the claim data for December 2011, September to December 2012 and January 2013 are in the same group, say Group 1. On the other hand, claims data for the period January to August 2012 are in the second group, say Group 2.

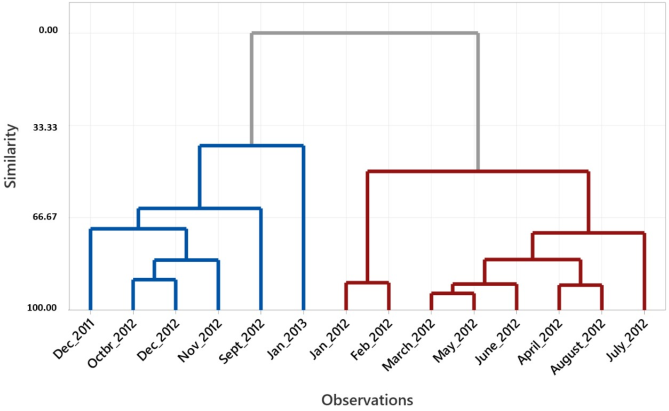

Figure 2. The Dendrogram plot shows that there are two groups of claim data formed.

Group 1 consist of 667 claims worth 5679 billion. Estimated total premiums received in those six months amounted to 602.3 billion. Besides that, in Group 2 there were 899 approved claims with an actual claim of 6021 billion. From January to August 2012, it is estimated that the total gross premiums received were worth 5773 billion. Based on this fact, we can conclude that the risks in Group 2 are greater than Group 1. In the next section, we will model the risk process for each of the groups formed.

3.2. Fitting Hawkes Process to Claim Payments Arrival

We would first draw the data histogram of how many claim payments there were in terms of days for Group 1 and Group 2, using Swishchuk et al. (2021)'s methodology, to determine whether the dataset could be incorporated into the Hawkes process or not with the help of Excel software. The results are shown in Figures 3 and 4.

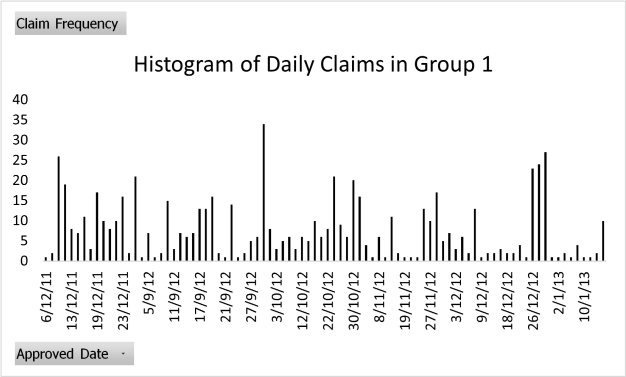

Figure 3. Histogram of Numbers of Claim Payments in Group 1.

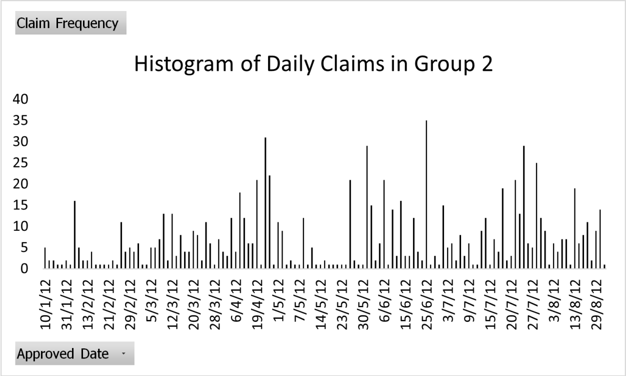

Figure 4. Histogram of Numbers of Claim Payments in Group 2.

Since the Hawkes process is a simple point process, there should only be one event at a particular timestamp. We adjust the arrival time between two occurring events by distributing the occurrences evenly across the same day because our data timestamps are in days. After adjusting the arrival time, we plotted the interarrival time quantile beside the exponential distribution quantile using |R-4.10|software. The results are shown in Figures 5 and 6.

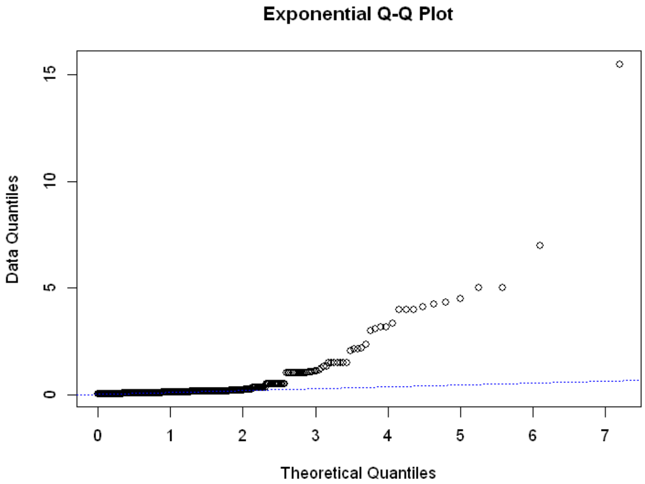

Figure 5. Exponential Q-Q plot of interarrival time between 2 occurring claims in Group 1.

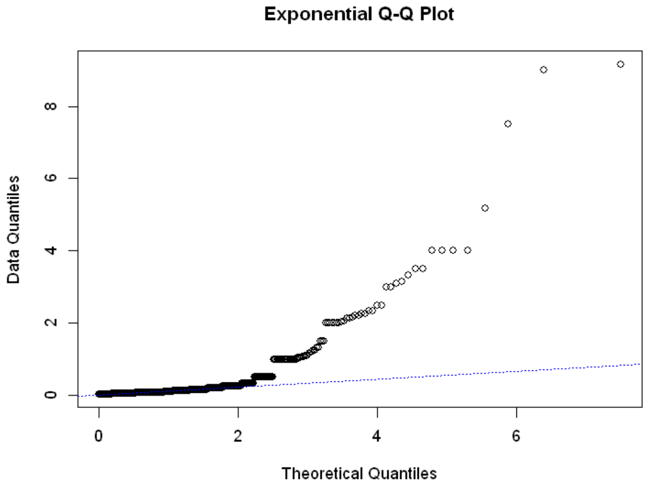

Figure 6. Exponential Q-Q plot of interarrival time between 2 occurring claims in Group 2.

We could observe from Figures 5 and 6 that the interarrival time did not fit the exponential distribution very well. It demonstrates that neither Group 1 nor Group 2's data would match the Poisson distribution well. We then compute the autocorrelation plot of the interval of events with lengths of 7, 14, 21, and 28 using |`R-4.10`|software by repeating the procedures from Swishchuk et al. (2021) and Da Fonseca and Zaatour (2013). The results are shown in Figures 7 and 8.

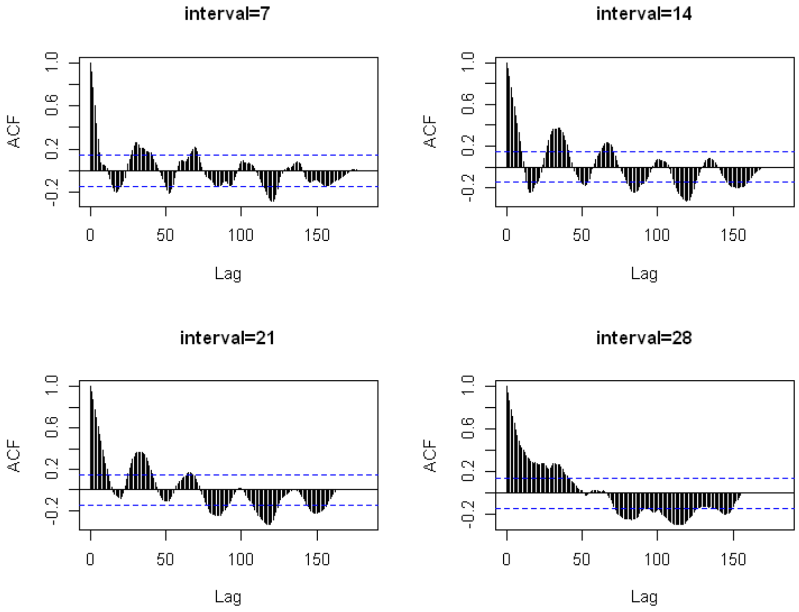

Figure 7. Autocorrelation (ACF) plot of interval length 7, 14, 21 and 28 in Group 1.

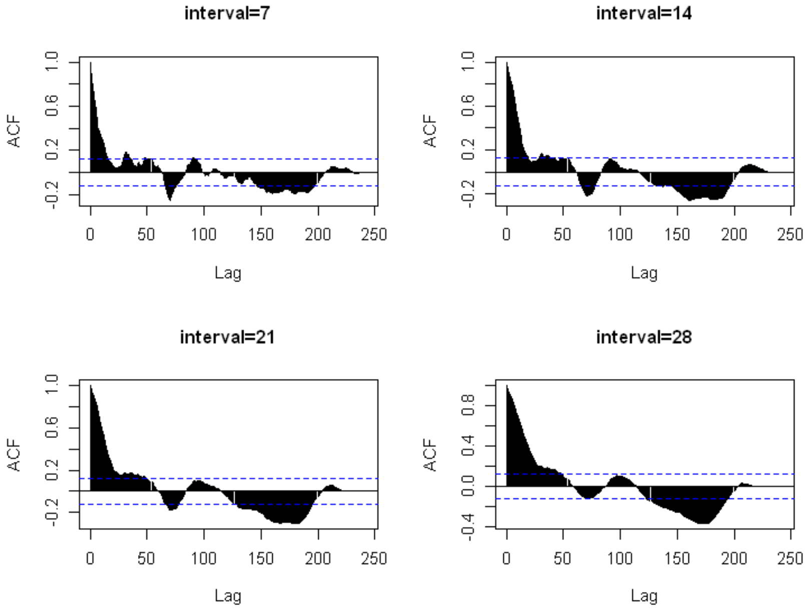

Figure 8. Autocorrelation (ACF) plot of interval length 7, 14, 21 and 28 in Group 2.

Based on the ACF plot above, we could conclude that the claim arrivals decay exponentially over time. Since the claim arrival data have a component of exponential decay, and exponential decay is one of the common choices for the excitation, then it is clear that the claim arrival data have self-excitation and exponential decay component, which is why we can use the Hawkes process fitting for the data.

The next step is to use the procedures from Shi and Odum Institute (2018) and use Peng (2003)'s `PtProc` R package to fit the data to the Hawkes process. The procedures are as follows:

1. Simulate some starting positions by creating uniform random variables numbered from 1 to k for the starting points.
2. Choose the starting point from those simulations that produce the theoretically predicted number of arrivals and variance of length 1 that are most similar to the observed ones. Say this number, c, and then the initial values for the next step are the triplet $(\lambda_0 = c, \alpha_0 = c, \beta_0 = c)$.
3. With the observed time points, the exponential kernel, and the initial values of the parameters established by the previous step, create a Hawkes process using the |`ptproc`| function.
4. Create a penalty for parameter values that are negative so that we can, hopefully, obtain estimates of the parameters that are positive.
5. Call the |`ptproc.fit()`| function on the process we created in Step 3 to perform the estimation.
6. Use (11) to compute the theoretical estimated number of arrivals for various interval duration.
7. Examine the computed parameter's theoretical variance for various interval duration using (12).

In this study, we choose $k = 10$ and the length of arrival duration is 1, 7, 14 and 21 with the help of R software. The outcomes are displayed in Tables 1–4.

Table 1. Hawkes Estimated Parameters and their Empirical and Theoretical Estimated Number of Arrivals of Group 1.

Parameter	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\beta}$	$\mathbb{E}[\widehat{N(1)}]$	$\mathbb{E}[N(1)]$
Estimated Value	0.5180797	2.522569	2.84654	3.625	4.552058

Table 2. Hawkes Estimated Parameters and their Empirical and Theoretical Estimated Number of Arrivals of Group 2.

Parameter	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\beta}$	$\mathbb{E}[\widehat{N(1)}]$	$\mathbb{E}[N(1)]$
Estimated Value	0.7332754	3.024789	3.72142	3.684426	3.917172

Table 3. Comparison of Theoretical Variance of Poisson Process and Hawkes Process vs. Empirical Variance for Different Interval Length Time in Days of Group 1.

Interval Length	Theoretical Poisson Process Variance	Theoretical Hawkes Variance	Empirical Variance
1	4.224193	55.13359	39.58194
7	29.56935	1500.13435	407.49403
14	59.13870	3860.71443	865.84410
21	88.70805	6310.38508	1132.38060

Table 4. Comparison of Theoretical Variance of Poisson Process and Hawkes Process vs Empirical Variance for Different Interval Length Time in Days of Group 2.

Interval Length	Theoretical Poisson Process Variance	Theoretical Hawkes Variance	Empirical Variance
1	3.699588	34.09472	39.12012
7	25.89712	628.83392	300.68076
14	51.79423	1410.15783	845.94836
21	77.69135	2192.64436	1480.71268

As in Tables 3 and 4, it is shown that the result of theoretical variance of Hawkes process is closer to the empirical variance than the result of theoretical variance of Poisson process so that the use of Hawkes process indeed improves the model fit.

We then attempted to simulate a Hawkes process using Zaatour (2014)'s R package `|hawkes|`,to ensure that our data could be matched accurately with the Hawkes process of the determined parameters. The computations in the function are taken from Ogata (1981), which is referenced in Section 2.2.4. In Figures 9 and 10, we then showed the simulation's quantiles and the actual arrival time using R software. With the obtained parameters, we discovered that the Hawkes process could adequately replicate the empirical data for Group 2 data in Figure 10, but for Group 1 data in Figure 9, most of the points in the Q-Q plot are far below the line, it shows that the intensity of the Hawkes process is higher than the intensity of the empirical data.

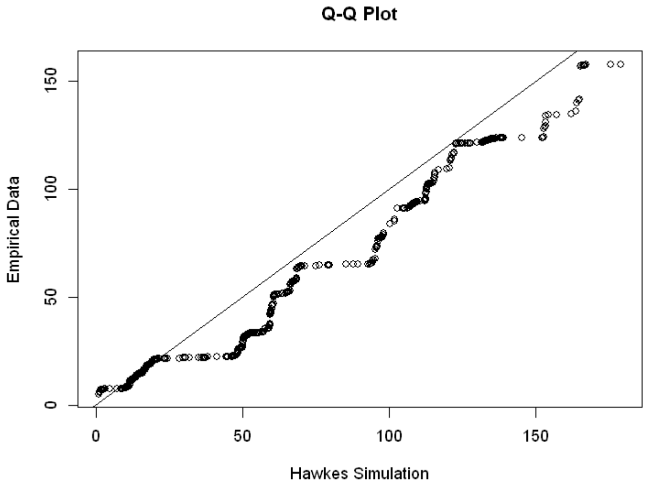

Figure 9. Q-Q plot of arrival time simulated by Hawkes process with parameters and empirical data of Group 1.

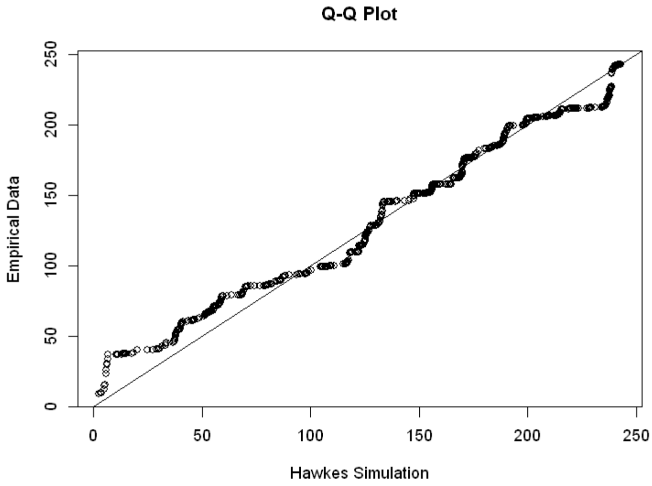

Figure 10. Q-Q plot of arrival time simulated by Hawkes process with parameters and empirical data of Group 2.

3.3. Modeling the Claim Cost

For the claim cost, our empirical data (Y) shows that, for Group 1 we have the minimum claim cost of IDR 120,000, mean of IDR 8,514,808 and maximum (G) of IDR 157,782,000 and for Group 2 we have the minimum claim cost of IDR 50,000, mean of IDR 6,695,487 and maximum (G) of IDR 284,800,000. We use the empirical distribution function $\hat{D}(G)$ as the distribution of the claim cost, and for that we have $\hat{D}(G) = 1$. We then make equally spaced boundaries g by dividing equally G by the number of states we want and count the state values $a(i)$ and the stationary distribution $p^{\star}$ for each state i. Using the formula in (20), (23) and (24), we developed our own syntax to calculate with the help of R software. The expected value was then obtained $a^{\star}$ that is the same as the expected value (mean) of the claim cost. Table 5 shows the result of the 9-States Markov chain we calculated for Group 1 and Table 6 shows the result of the 4-states Markov chain we calculated for Group 2.

Table 5. 9-States Markov chain values for Group 1.

Parameter	Values
$(g_1; g_2; g_3; g_4; g_5; g_6; g_7; g_8; g_9 = G)$	(17,531,333; 35,063,667; 70,125,333; 87,656,667; 105,188,000; 122,719,333; 140,250,667; 157,782,000)
$(p_1^\star; p_2^\star; p_3^\star; p_4^\star; p_5^\star; p_6^\star; p_7^\star; p_8^\star; p_9^\star)$	(0.931034483; 0.023988006; 0.013493253; 0.007496252; 0.008995502; 0.002998501; 0.004497751; 0.004497751; 0.002998501)
$(a(1); a(2); a(3); a(4); a(5); a(6); a(7); a(8); a(9))$	(4,628,737; 22,900,401; 43,237,186; 66,041,056; 80,050,000; 100,400,000; 116,973,515; 124,194,000; 157,291,000)
$a^\star$	8,514,808
E(Y)	8,514,808

Table 6. 4-states Markov chain values for Group 2.

Parameter	Values
$(g_1; g_2; g_3; g_4 = G)$	(71,200,000; 142,400,000; 213,600,000; 284,800,000)
$(p_1^\star; p_2^\star; p_3^\star; p_4^\star)$	(0.979977753; 0.016685206; 0.00224694; 0.001112347)
$(a(1); a(2); a(3); a(4))$	(4,418,278; 102,157,227; 154,791,000; 284,800,000)
$a^\star$	6,695,487
E(Y)	6,695,487

We must create an appropriate state space for the Markov chain to be irreducible in order to have a stationary distribution $p^\star$. For our data, we found that $N = 9$ for Group 1 and $N = 4$ for Group 2. This is due to the fact that, when we attempted to create 10 state spaces, we discovered that only one transitional probability of zero was produced, which rendered the Markov chain irreducible. Figures 12 and 13 show empirical distribution function and Q-Q plot of 9-states Markov chain for Group 1 while Figures 13 and 14 show that the 4-states Markov chain could replicate claim cost data in Group 2 quite well. We use R software to obtain the results in Figures 11–14.

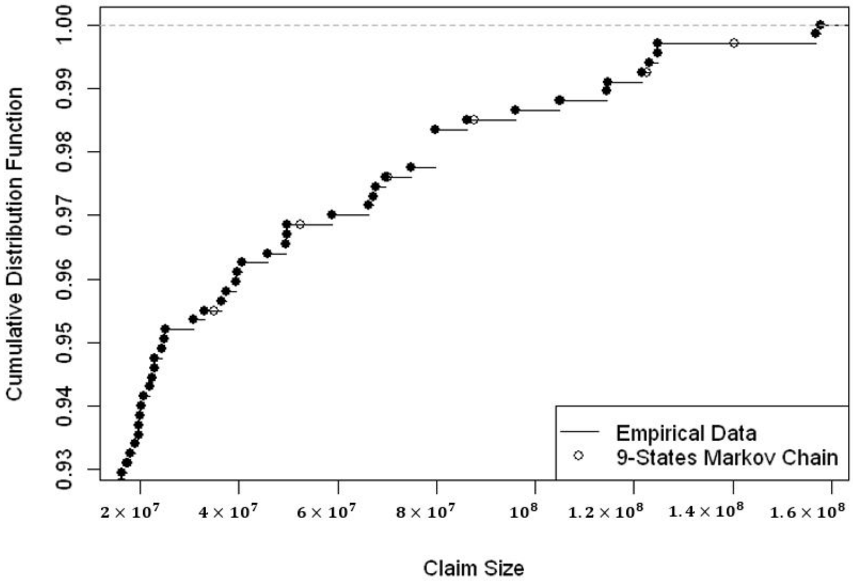

Figure 11. Distribution function of empirical data and 9-states Markov Chain for Group 1 $[\hat{D}_1(g)]$.

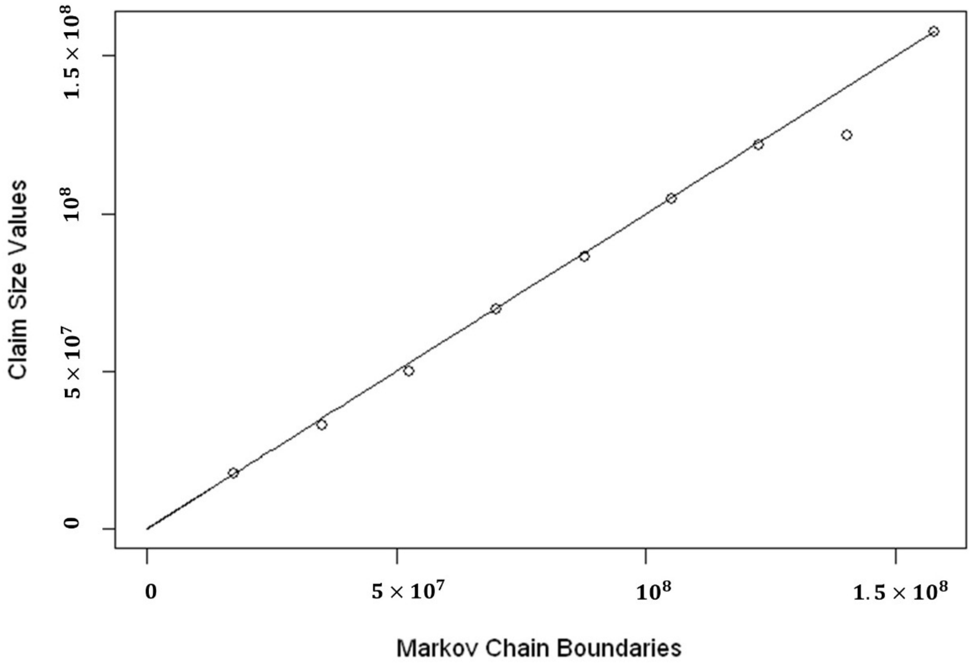

Figure 12. QQ-plot of Markov Chain claim boundaries and empirical claims in the boundaries for Group 1.

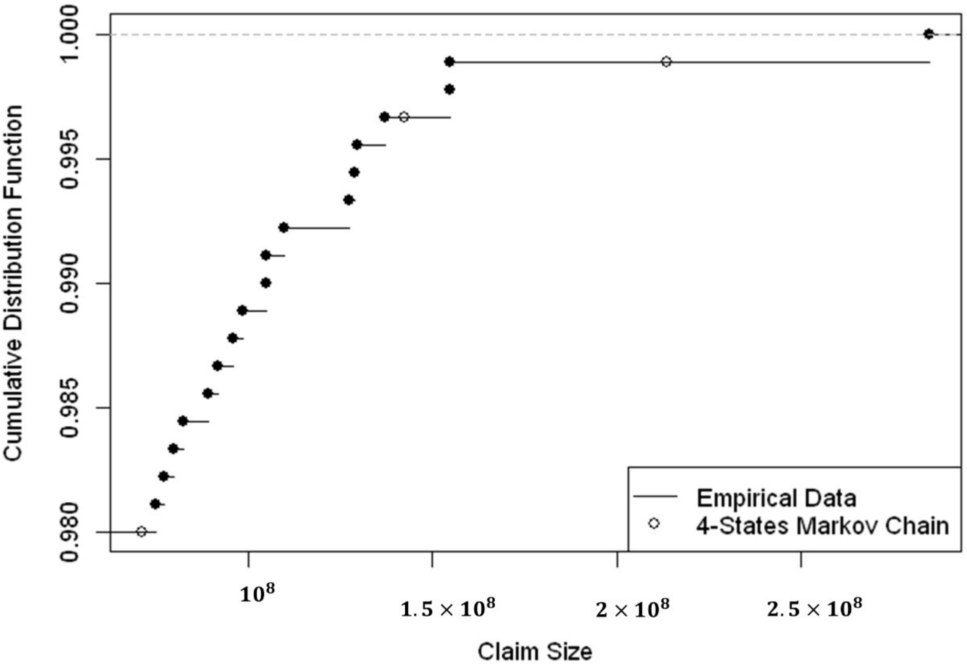

Figure 13. Distribution function of empirical data and 4-states Markov Chain for Group 2 [$\hat{D}_2(g)$].

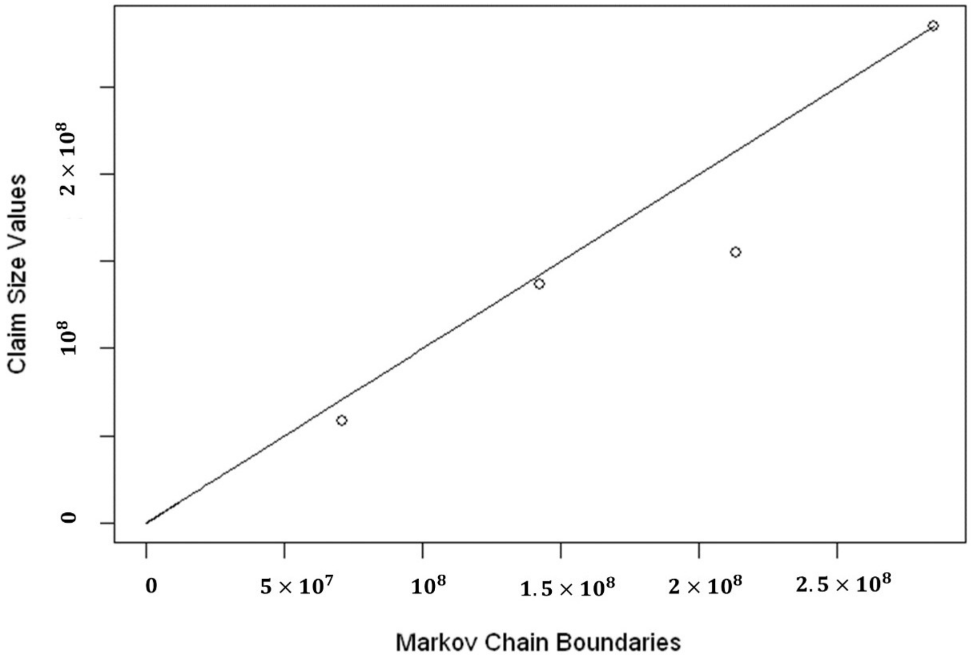

Figure 14. QQ-plot of Markov Chain claim boundaries and empirical claims in the boundaries for Group 2.

3.4. Risk Model Using General Compound Hawkes Processes

After determining that the Hawkes process and the time-continuous Markov chain could be used to represent the counting process and claim severity, we might attempt to describe the risk process using general compound Hawkes process. The initial capital, u, and continuous premium rate must be established before we can model the risk process. Please note that $c = \max(50, 387, 867; 34, 095, 590) = 50, 387, 867$. We use the maximum value of c that is calculated by Equation (31) so that it could cover the higher claims in both two groups.

We use the formula of Net Profit Condition and Premium Principle from (30) to determine the premium rate by taking $\theta = 0.3$ as the safety loading. For the 9-States Markov chain, we obtain the $a^{\star} = 8{,}514{,}808$. With the values we already calculated before, we obtain the premium rate, $c = \max(50, 387, 867; 34, 095, 590) = 50{,}387{,}867$. For the initial capital, we consider the mean and maximum of the claim cost and decided that $u = \text{mean}(\text{claim cost}) \times 70 = 596, 036, 550$ is appropriate. For the 4-states Markov chain, we obtain the $a^{\star}$ =6,695,487. With the values we already calculated before, we obtain the premium rate c =28,850,114. For the initial capital, we consider the mean and maximum of the claim cost and decided that $u = \text{mean}(\text{claim cost}) \times 10 = 66, 954, 873$ is appropriate. In the context of contrasting the fit of Hawkes models with exponential kernels to empirical data, Zhang (2016) proposes a metric, $\hat{S}$ and $\hat{F}$ as the following:

$$\hat{S}(K) = \frac{1}{K} \sum_{i=1}^{L} \frac{\max(\hat{R}_i(t)) - \min(\hat{R}_i(t))}{\max(R(t)) - \min(R(t))}. \tag{32}$$

$$\hat{F}(K) = \frac{1}{K} \sum_{i=1}^{K} \frac{\hat{R}_i(T)}{R(T)}. \tag{33}$$

where $\hat{R}_i(t)$ stands for the simulated risk processes, $R(t)$ stands for the benchmark of empirical process and K stands for the number of simulated paths.

Using the calculated u and c, we then derive the empirical risk process and simulate $L = 1000$ paths of risk process with arrival time following Hawkes process with parameters for group 1 are $\hat{\lambda} = 0.5180797, \hat{\alpha} = 2.522569$ and $\hat{\beta} = 2.84654$ and parameters for group 2 are $\hat{\lambda} = 0.7332754, \hat{\alpha} = 3.024789$ and $\hat{\beta} = 3.72142$. The claim severity in groups 1 and 2, respectively, follows a 9-state Markov chain and a 4-state Markov chain. Then, we simulate with the same value of arrival time generated by Hawkes simulation from `|hawkes|`library in R and randomly generated Markov chain states simulations. Figures 15–17 shows the plot of the empirical risk process and the first 50 simulations. We use R software to obtain the figures. We could see from the plot that the simulations adequately represented the empirical process. Then we use the formulas from Swishchuk et al. (2021) to compute the comparison of fluctuation over time and at time $T = 184$ days for Group 1 and $T = 244$ days for Group 2. Then, we can calculate the average historical insurance reserve with the help of R software. We obtain the results in Tables 7–9.

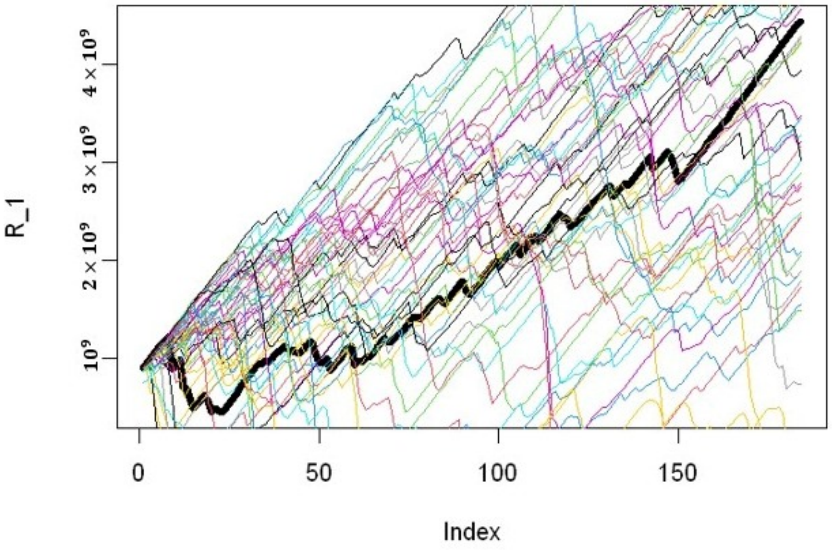

Figure 15. Plot of empirical risk process (bold line) and 50 simulations of the general compound Hawkes process in Group 1.

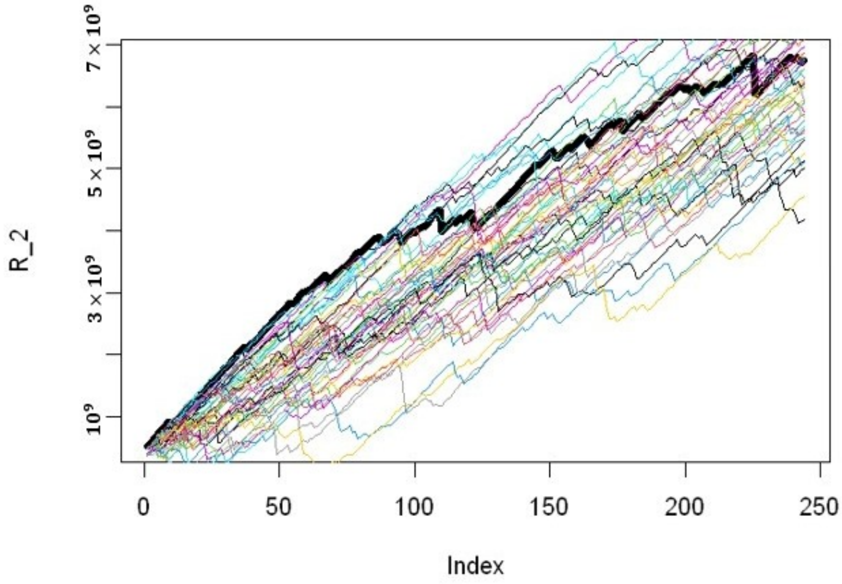

Figure 16. Plot of empirical risk process (bold line) and 50 simulations of the general compound Hawkes process in Group 2.

Table 7. The result of the comparison between an empirical risk process and simulations of the general compound Hawkes process at time $T = 184$ in Group 1.

Parameter	Method	Values
$\hat{S}(K)$	metric	0.9367684
$\hat{F}(K)$	metric	0.6978924
$R(T)$	empirical	4,443,471,513
$E[R(T)]$	simulations	891,786,394
$\sqrt{Var[R(T)]}$	simulations	26,348,096

Table 8. The result of the comparison between an empirical risk process and simulations of the general compound Hawkes process at time $T = 244$ in Group 2.

Parameter	Method	Values
$\hat{S}(K)$	metric	0.9784721
$\hat{F}(K)$	metric	0.9650283
$R(T)$	empirical	6,744,080,615
$E[R(T)]$	simulations	509,061,428
$\sqrt{Var[R(T)]}$	simulations	24,506,352

Table 9. Result of Insurance reserve comparison between period in each Groups.

	Group 1	Group 2	All Data
Average historical insurance reserve	38,161,865	25,635,635	31,020,743

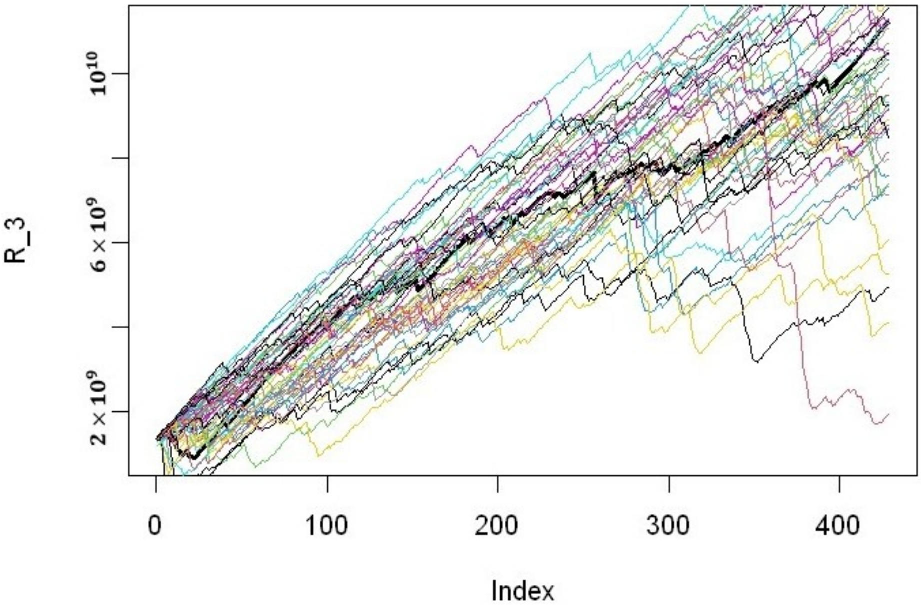

Figure 17. Plot of empirical risk process (bold line) and 50 simulations of the general compound Hawkes process for all data.

3.5. Discussion

This study reveals that a high claim period and low claim period have quite different characteristics in terms of claims volatility. As a result of these differences in characteristics, the required reserves differ significantly. Table 9 shows a considerable difference between the first group, which occurs during the high claim period, and the second group, which is dominated by the low claim period, in terms of the average historical insurance reserve.

In addition, when compared to earlier studies in risk theory that employed the Hawkes process, this study can also help in the development of a novel technique for concurrently classifying and estimating insurance reserves that entails grouping risks with comparable loss potential and imposing new various manual rates to account for variations in loss potential within the groupings.

For further research, for instance, longitudinal data analysis can be used to model the risk with the Hawkes process from claim data that contains observations about different cross sections across time.

4. Conclusions

The Hawkes processes are incredibly intriguing representations of reality. Since many of the common probability models are Markovian, they ignore the process's history. The Hawkes processes are built on the foundation that history matters, which helps to explain why they are used in such a wide variety of applications. The use of the Hawkes process in estimating the calculation of insurance reserves can be used as a good alternative method because of its ability to anticipate behaviors that have spread throughout reserves over time. Claims from the past typically result in higher reserves and new claims have a bigger impact on reserves than outdate claims. The Hawkes process, which primarily exhibits self-excitation and exponential decay, is well suited to describing these occurrences.

The clustering method, which is essentially a type of unsupervised learning method, allows us to categorize risk processes based on claim frequency domain. This involves grouping risks with similar loss potential and imposing a reserve estimation to account for variations in loss potential within the groupings.

This approach may have consequences for changes in the insurer's perspective that have not yet been taken into consideration: the potential for variations between groups in the frequency domain. Finding the risk criteria that divide risks into groups with comparable expected loss experiences is therefore considered to be a step in applying this strategy.

Author Contributions: Conceptualization, A.R.E.; writing—original draft preparation, K., A.N.M. and M.F.A.; writing—review and editing, A.R.E, K., A.N.M. and M.F.A.; project administration, A.R.E.; funding acquisition, G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was fully funded by the project of the Lecturer Publication Acceleration Program of Mathematics Department, Universitas Gadjah Mada 2022 (grant number: 298/J01.1.28/PL.06.02/2022) provided by the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada 2022 towards the World Class University.

Institutional Review Board Statement: This study did not require ethical approval.

Informed Consent Statement: Not applicable. This study did not involve humans.

Data Availability Statement: The corresponding author can be contacted for data availability used in this study.

Acknowledgments: The authors thank reviewers for their careful reading and helpful suggestions.

Conflicts of Interest: The authors declare no conflict of interest in this publication

References

Bessy-Roland, Yannick, Alexandre Boumezoued, and Caroline Hillairet. 2021. Multivariate Hawkes process for cyber insurance. *Annals of Actuarial Science* 15: 14–39. [CrossRef]

Cheng, Zailei, and Youngsoo Seol. 2020. Gaussian approximation of a risk model with stationary Hawkes arrivals of claims. *Methodology and Computing in Applied Probability* 22: 555–71. [CrossRef]

Da Fonseca, José, and Riadh Zaatour. 2013. Hawkes process: Fast calibration, application to trade, clustering and diffusive limit. *Journal of Futures Markets* 34: 548–79. [CrossRef]

Filimonov, Vladimir, and Didier Sornette. 2015. Apparent criticality and calibration issues in the Hawkes self-excited point process model: Application to high-frequency financial data. *Quantitative Finance* 15: 1293–314. [CrossRef]

Gao, Xuefeng, and Lingjiong Zhu. 2017. Limit theorems for Markovian hawkes processes with a large initial intensity. *Stochastic Processes and their Applications* 128: 3807–39. [CrossRef]

Hartigan, John A. 1975. *Clustering Algorithms*. New York: Wiley-Interscience.

Hawkes, Alan G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58: 83–90. [CrossRef]

Jeong, Himchan, and Bin Zou. 2022. A Dynamic Credibility Model with Self-Excitation and Exponential Decay (9 September 2022). Proceedings of the 2022 Winter Simulation Conference. Available online: https://ssrn.com/abstract=4214889 (accessed on 5 May 2022).

Johnson, Richard A., and Dean W. Wichern. 2009. *Applied Multivariate Statistical Analysis*. New York: John Wiley & Sons, pp. 680–95, chp. 12, section 12.3.

Kaufman, Leonard, and Peter J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, Inc. ISBN 9780471878766.

Laub, Patrick J., Thomas Taimre, and Phillip K. Pollett. 2015. Hawkes processes. *arXiv* arXiv:1507.02822.

Lefèvre, Claude, and Phillipe Picard. 2006. A nonhomogeneous risk model for insurance. *Computers and Mathematics with Applications* 51: 325–34. [CrossRef]

Lundberg, Filip. 1903. I. *Approximerad Framstallning af Sannolikhetsfunktionen*: II. *Aterforsakring af kollektivrisker*. Uppsala: Almqvist Wiksell.

Ogata, Yosihiko. 1981. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory* 27: 23–31. [CrossRef]

Ogata, Yosihiko. 1999. Seismicity Analysis through Point-process Modeling: A Review. *Pure and Applied Geophysics* 155: 471–507. [CrossRef]

Peng, Roger D. 2003. Multi-dimensional Point Process Models in R. *Journal of Statistical Software* 8: 1–27. [CrossRef]

Răducan, Anişoara Maria, Raluca Vernic, and Gheorghiţă Zbăganu. 2015. On the ruin probability for nonhomogeneous claims and arbitrary inter-claim revenues. *Journal of Computational and Applied Mathematics* 290: 319–33. [CrossRef]

Romaniuk, Maciej. 2019. Simulation-Based Analysis of Penalty Function for Insurance Portfolio with Embedded Catastrophe Bond in Crisp and Imprecise Setups. In *Information Systems Architecture and Technology: Proceedings of 39th International Conference on Information Systems Architecture and Technology-ISAT 2018*. Advances in Intelligent Systems and Computing. Cham: Springer, vol. 854. [CrossRef]

Shi, Feng, and Odum Institute. 2019. *Learn About the Hawkes Process in R With Data From the DJIA 30 Stock Time Series (2018)*. London: SAGE Publications Ltd. [CrossRef]

Stabile, Gabriele, and Giovanni Luca Torrisi. 2010. Risk processes with non-stationary Hawkes claims arrivals. *Methodology and Computing in Applied Probability* 12: 415–29. [CrossRef]

Swishchuk, Anatoliy. 2017. Risk model based on general compound Hawkes processes. *arXiv* arXiv:1706.09038.

Swishchuk, Anatoliy, Rudi Zagst, and Gabriela Zeller. 2020. Hawkes Processes in Insurance: Risk Model, Application to Empirical Data and Optimal Investment. *Insurance Mathematics and Economics* 101: 107–24. [CrossRef]

Zaatour, Riadh. 2014. |`hawkes`|:Hawkes process simulation and calibration toolkit. R package version 0.0.4. Available online: https://CRAN.R-project.org/package=hawkes (accessed on 5 May 2022).

Zhang, Changyong. 2016. Modeling high frequency data using hawkes processes with power-law kernels. *Procedia Computer Science* 80: 762–71. [CrossRef]