


Article

Dependence Modelling for Heavy-Tailed Multi-Peril Insurance Losses

Tianxing Yan, Yi Lu and Himchan Jeong * 

Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada; tya44@sfu.ca (T.Y.); yi_lu@sfu.ca (Y.L.)

* Correspondence: himchan_jeong@sfu.ca

Abstract: The Danish fire loss dataset records commercial fire losses under three insurance coverages: building, contents, and profits. Existing research has primarily focused on the heavy-tail behaviour of the losses but ignored the relationship among different insurance coverages. In this paper, we aim to model the aggregate loss for all three coverages. To study the pairwise dependence of claims from all types of coverage, an independent model, a hierarchical model, and some copula-based models are proposed for the frequency component. Meanwhile, we applied composite distributions to capture the heavy-tailed severity component. It is shown that consideration of dependence for the multi-peril frequencies (i) significantly enhances model goodness-of-fit and (ii) provides more accurate risk measures of the aggregated losses for all types of coverage in total.

Keywords: dependence modelling; ratemaking; multi-peril insurance; heavy-tail distributions; composite models; copulas; binomial thinning

1. Introduction

Insurance provides financial compensation to individuals or companies after a particular event occurs. Basically, the insurance business relies on the diversification effects after the risks of the policyholders are pooled together and supported by collected premiums from the policyholders, which are primarily determined by the expected amount of claims from each of the policyholders. In this regard, there are two components to be considered for effective risk management purposes: the heavy-tail behaviour of insurance claims and the possible dependence among different insurance coverages. Heavy-tail behaviour can affect the effectiveness of risk mitigation as the risk pooling is inherently based on the total expected claim amounts, whereas an excessively large claim could dampen the solvency of the insurance portfolio due to the heavy-tail behaviour. Such impacts could be more substantial if the insurance portfolio provides multiple types of coverage and the claims from different types of coverage are positively correlated. In this paper, we focus on a reinsurance dataset, Danish multi-peril commercial fire loss, aiming to incorporate the dependency among different insurance coverages and heavy-tailed losses for modelling the monthly aggregate loss for the company.

In this regard, there have been many approaches that could handle heavy-tail behaviours of insurance claims. One example is the peak-over-threshold (POT) approach. However, for a reinsurance company, it is important to consider the financial loss in the entire range instead of focusing on the extreme cases. Additionally, a finite mixture model can be used to handle heavy-tail behaviours, which is a linear combination of multiple distributions such as (Hong and Martin 2018; Miljkovic and Grün 2016). A unique advantage of such models is the ability to construct a multimodal distribution. Different from finite mixture models, a composite model splices and combines random variables (usually continuous random variables) with the consideration of continuity and differentiability at the splicing points. It allows the fitting of different distributions with desirable distributional properties on certain ranges of data, especially to accommodate the heavy-tail



Citation: Yan, Tianxing, Yi Lu, and Himchan Jeong. 2024. Dependence Modelling for Heavy-Tailed Multi-Peril Insurance Losses. *Risks* 12: 97. <https://doi.org/10.3390/risks12060097>

Academic Editor: Olivier Féron

Received: 28 April 2024

Revised: 10 June 2024

Accepted: 13 June 2024

Published: 16 June 2024



Copyright: © 2024 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/>).

nature of the data. For example, (Cooray and Cheng 2015; Pigeon and Denuit 2011) focus on composite lognormal–Pareto models, and (Scollnik and Sun 2012) applied composite Weibull–Pareto models. By considering the model complexity and ability to capture the realistic loss behaviour, we utilize composite models for the loss severities (claim amounts) in this paper.

In addition to the heavy-tail, the dependency among risks is another unignorable behaviour. Two different methods have been applied in this study: the hierarchical and the copula-based modelling frameworks. With the multilevel modelling technique, the hierarchical modelling framework bridges the relationship among different events by sharing the belief that risks from a common environment are not independently distributed. For instance, (Fung et al. 2023) proposed a hierarchical modelling approach that models the number of certain climate events and the associated claim counts subsequently. Recently, (Jeong 2024) considered a multivariate Tweedie distribution where the correlated random effects are modelled only with their moments. An alternative way of studying dependency is the copula method. This methodology allows to flexibly connect random variables using a dependent structure. The authors of (Lee and Shi 2019) suggested a copula-based collective risk model for describing various dependencies in longitudinal insurance claims data. The authors of (Oh et al. 2021) provided a copula-based collective risk model for microlevel multi-year claims data. The authors of (Jeong et al. 2023) considered a factor copula model to capture dependence among claim counts from multiple lines of business.

There are existing studies in the literature that analyze the heavy-tailed behaviour of Danish fire losses. For example, (McNeil 1997) applied the generalized Pareto distribution for the total losses for all coverages and tested the goodness-of-fit. Additionally, (Resnick 1997) suggested some alternative methodologies to study the tail behaviour. However, there is a lack of studies that focus on the dependency among losses under different coverages. They are needed, from the practice perspective, in order to appropriately price and set reserves for multi-peril insurance products. We conduct a comprehensive study to address both issues for effective risk management purposes regarding the aggregate Danish fire loss on a monthly basis. More specifically, under the framework of collective risk models, we model the claim frequency from various types of insurance coverage first and then study the aggregate loss by adding the losses from all insurance coverages and using the composite model for loss amounts.

In this study, we propose three different types of frequency models: a fully independent model as a benchmark model, a hierarchical model, and some copula-based models to model the frequency component. The hierarchical model and copula-based models proposed incorporate the dependency among different coverages. Meanwhile, several two-component composite models are implemented to model the severity component. After conducting statistical and risk analyses, by comparing with the benchmark model, the fully independent model, we conclude that the models with dependency structure significantly improve model goodness-of-fit and provide more accurate risk measures of the aggregate losses for all types of coverages in total. However, there are limitations regarding our proposed models. Our proposed models do not consider the dependency between the frequency and severity. In the literature, for example, (Vernic et al. 2021) proposed a Sarmanov distribution for modelling dependence between the frequency and the average severity of insurance claims. Additionally, there could be models other than copula ones to model the dependence between the claim counts of different types.

The remainder of this article is organized as follows. Section 2 introduces the dataset that motivates our research. Section 3 provides a statistical framework to model the dependent claim frequency from multiple types of insurance coverage. Section 4 provides a framework for the severity component to capture the heavy-tail behaviour of insurance claims. Section 5 provides the estimation results from different models, associated with their implications in risk management. Section 6 concludes this article.

2. Data Exploration

We start with the introduction of a dataset of Danish multi-peril fire losses, which is available in an R library, `CASdataset`. It was recorded by the Denmark's Copenhagen Reinsurance Company and contains 2167 commercial fire loss records from 1980 to 1990. Each recorded claim includes the loss amounts of three sections: building, contents, and profits, which are adjusted by inflation using 1985 as the base year. Below, a few rows of the data are provided. The building, contents, and profits columns show the Danish Krone losses in millions and the total column is the sum of the three. Table 1 shows the first 5 rows of the dataset.

Table 1. Excerpt from the Danish Fire Dataset.

Date	Building	Contents	Profits	Total
3 January 1980	1.09809663	0.58565150	0.00000000	1.683748
4 January 1980	1.75695461	0.33674960	0.00000000	2.093704
5 January 1980	1.73258126	0.00000000	0.00000000	1.732581
7 January 1980	0.00000000	1.30537600	0.47437775	1.779754
7 January 1980	1.24450952	3.36749600	0.00000000	4.612006

We recall that we are interested in analyzing the Danish reinsurance aggregate loss on a monthly basis. The collective risk modelling framework is utilized for each coverage to model the total loss for a single coverage. Then, the adding-up of the losses from three coverages is the aggregate loss we are interested in. In this regard, we aggregate the 132 months claim numbers to obtain the observations using the following notations:

- M_t : Number of reported accidents during month $t = 1, \dots, 132$;
- N_{jt} : Number of claims from the j th lines of insurance during month t , where $j = 1, 2, 3$ represent the building, contents, and profits;
- Y_{jtk} : k th individual loss amounts from j th line of insurance during month t for $k = 1, \dots, N_{jt}$;
- S_{jt} : Aggregate loss amount from the j th line of insurance during month t , which is defined as

$$S_{jt} := \sum_{k=1}^{N_{jt}} Y_{jtk}, \quad N_{jt} > 0 \quad (1)$$

and 0 otherwise. We use a compound risk model (CRM) to describe S_{jt} ;

- S_t : Aggregate loss for all lines of insurance during month t , which is defined as

$$S_{\bullet t} := S_{1t} + S_{2t} + S_{3t}. \quad (2)$$

For the j th lines of insurance, $j = 1$ represents the damage to the building, $j = 2$ is the related contents, and $j = 3$ stands for the profit line. Since we aggregate the data monthly, we consider whether we could assume that the claim numbers S_{1t} , S_{2t} , and S_{3t} are time-independent. Figure 1 shows the boxplots of the claim numbers in three insurance lines for different months.

The plots show no significant seasonal effect. We also checked that claims in the current month do not affect claims in the following month with a separate exploratory analysis not included in this article. Therefore, it could be innocuous to assume that S_1 , S_2 , S_3 , and S_{\bullet} are the aggregate claim random variables that we are interested in, where S_{1t} , S_{2t} , S_{3t} , and $S_{\bullet t}$ are i.i.d. samples of S_1 , S_2 , S_3 , and S_{\bullet} , respectively. However, we can detect some dependence among the claim numbers from three lines of businesses.

We calculate the Pearson correlation coefficient for any two lines of claim numbers. The claim numbers of buildings are highly related to the contents. The Pearson coefficient is 0.876580. For the contents and profits coverage, the coefficient is 0.7454898. The relationship between the buildings and profits is not as strong as the others, and the coefficient is

0.574442. Overall, this exploratory analysis shows the necessity of modelling dependence among the claim counts from multiple coverage.

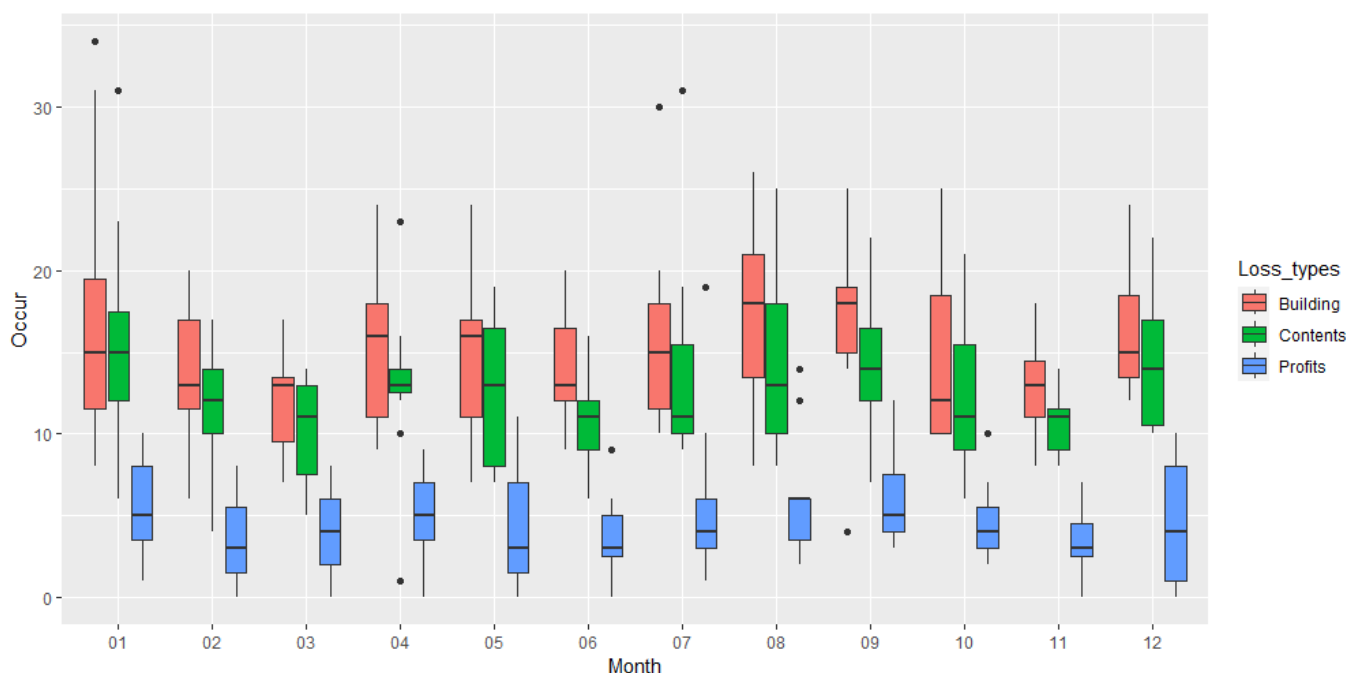


Figure 1. Exploration of seasonal effects with the Danish reinsurance dataset.

It is known that the losses in property insurance are mostly heavy-tailed, whereby the given data is not exceptional. In this regard, (McNeil 1997; Resnick 1997) worked on the extreme value analyses using this dataset, where they applied the peak-over-threshold and estimated the parameters to implement generalized Pareto distributions. In each business line, we observe several data points that have relatively large losses. In Table 2, for all business lines, the averages of the observed losses are higher than the corresponding third quarters.

Table 2. Summary of loss amount for three business lines.

Source	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Building	0.02319	0.96618	1.32013	1.98668	1.97860	152.41321
Contents	0.00083	0.29000	0.57570	1.70178	1.44648	132.01320
Profits	0.00408	0.10011	0.26619	0.85180	0.67929	61.93265

3. Dependence Modelling for Multivariate Claim Frequencies

To model possible dependence among the claim frequencies from the three types of coverage, we consider three types of models: a fully independent model (Section 3.1), a binomial thinning model (Section 3.2), and copula-based models (Section 3.3). The independent model is used as a benchmark model for comparison. Other models consider the dependence among the claim numbers in different lines. The binomial thinning model utilizes a hierarchical framework to model such a dependence. However, it only allows fixed dependency structures between any two margins. Copula-based models are used to construct the joint distribution flexibly.

3.1. Benchmark Model: Independent Frequency Model

The fully independent frequency model assumes independent relationships among the margins of the frequencies (and, subsequently, severities) from multiple types of coverage.

We recall that the claim numbers data are over-dispersed. To capture this behaviour, we model the number of building, content, and profit claims as well as the number of reported accidents using negative binomial random variables, N_1, N_2, N_3 , and M , respectively, instead of using the Poisson distribution that implicitly assumes equi-dispersion. Negative binomial also performs better than Poisson in terms of in-sample goodness-of-fit measures such as AIC, BIC, and the log-likelihood in our dataset. For all $j = 1, 2, 3$, we assume $N_j \sim \mathcal{NB}(\lambda_j, r_j)$ and $M \sim \mathcal{NB}(\lambda, r)$ with the following parameterization:

$$f_{N_j}(N_j = n_j) = \frac{\Gamma(r_j + n_j)}{\Gamma(r_j)\Gamma(n_j + 1)} \left(\frac{\lambda_j}{r_j + \lambda_j}\right)^{n_j} \left(\frac{r_j}{r_j + \lambda_j}\right)^{r_j}, \tag{3}$$

$$f_M(M = m) = \frac{\Gamma(r + m)!}{\Gamma(r)\Gamma(m + 1)} \left(\frac{\lambda}{r + \lambda}\right)^m \left(\frac{r}{r + \lambda}\right)^r, \tag{4}$$

where r and r_j are the size parameters of the negative binomial distributions. Instead of using the probability as the second parameter, we use λ_j and λ , which stand for the means of the random variables N_j and M . The likelihood function of the negative binomial parameters is given by:

$$\begin{aligned} \mathcal{L}(\theta|\mathcal{D}) &= \prod_{t=1}^{132} f_{N_1, N_2, N_3, M}(N_{1t}, N_{2t}, N_{3t}, M_t; \theta) \\ &\stackrel{\text{ind.}}{=} \prod_{t=1}^{132} f_M(M_t; r, \lambda) \cdot f_{N_1}(N_{1t}; r_1, \lambda_1) \cdot f_{N_2}(N_{2t}; r_2, \lambda_2) \cdot f_{N_3}(N_{3t}; r_3, \lambda_3) \\ &= \prod_{t=1}^{132} \left\{ \left[\frac{\Gamma(r + m_t)}{\Gamma(m_t + 1)\Gamma(r)} \frac{\lambda^{m_t}}{(r + \lambda)^{r+m_t}} \right] \cdot \prod_{j=1}^3 \left[\frac{\Gamma(r_j + n_{jt})}{\Gamma(n_{jt} + 1)\Gamma(r_j)} \frac{\lambda_j^{n_{jt}}}{(r_j + \lambda_j)^{r_j+n_{jt}}} \right] \right\}, \end{aligned} \tag{5}$$

where θ is a vector of all the parameters for the negative binomial distributions for the reported accident numbers and the claim numbers from each coverage. We let \mathcal{D} denote the available data.

3.2. Binomial Thinning Model

To investigate the possible dependence of the frequencies of this dataset, we note that the claim numbers ($N_j, j = 1, 2, 3$) cannot be larger than the number of accidents reported (M), by definition. In this case, one can consider the binomial distribution as a natural fit for the N_j given M , while N_j s are conditionally independent given M .

More specifically, we use the negative binomial distribution to model M due to observed over-dispersion, namely $M \sim \mathcal{NB}(\lambda, r)$. We also set $N_j|M = m \sim \mathcal{BN}(m, \lambda_j/\lambda)$ for the three sources of claim numbers so that the joint distribution of (M, N_1, N_2, N_3) can be expressed as:

$$\begin{aligned} &f_{N_1, N_2, N_3, M}(n_1, n_2, n_3, m) \\ &= f_{N_1|M}(n_1|M = m) \cdot f_{N_2|M}(n_2|M = m) \cdot f_{N_3|M}(n_3|M = m) \cdot f_M(m) \\ &= \prod_{j=1}^3 \left[\binom{m}{n_j} \frac{\lambda_j^{n_j} (\lambda - \lambda_j)^{m-n_j}}{\lambda^m} \right] \cdot \left[\frac{\Gamma(r + m)}{\Gamma(m + 1)\Gamma(r)} \left(\frac{\lambda}{r + \lambda}\right)^m \left(\frac{r}{r + \lambda}\right)^r \right]. \end{aligned} \tag{6}$$

We note that the marginal distributions of N_j is a negative binomial random variable with size parameter r and mean λ_j , as shown below:

$$\begin{aligned}
 f_{N_j}(n_j) &= \sum_{m=n_j}^{\infty} [f_{N_j}(n_j|M=m) \cdot f_M(m)] \\
 &= \sum_{m=n_j}^{\infty} \left[\frac{\Gamma(r+m)}{\Gamma(m+1)\Gamma(r)} \left(\frac{\lambda}{r+\lambda}\right)^m \left(\frac{r}{r+\lambda}\right)^r \right] \cdot \left[\binom{m}{n_j} \frac{\lambda_j^{n_j} (m-\lambda_j)^{\lambda-n_j}}{\lambda^m} \right] \\
 &= \frac{\Gamma(r+n_j)}{\Gamma(n_j+1)\Gamma(r)} \sum_{m=n_j}^{\infty} \binom{r+m-1}{m-n_j} \left(\frac{\lambda-\lambda_j}{\lambda}\right)^{m-n_j} \left(\frac{\lambda_j}{\lambda}\right)^{n_j} \left(\frac{\lambda}{\lambda+r}\right)^m \left(\frac{r}{\lambda+r}\right)^r \tag{7} \\
 &= \frac{\Gamma(r+n_j)}{\Gamma(n_j+1)\Gamma(r)} \left(\frac{\lambda_j}{\lambda_j+r}\right)^{n_j} \left(\frac{r}{\lambda_j+r}\right)^r \sum_{m=n_j}^{\infty} \binom{r+m-1}{m-n_j} \frac{(\lambda_j+r)^{n_j+r} (\lambda-\lambda_j)^{m-n_j}}{(\lambda+r)^{m+r}} \\
 &= \frac{\Gamma(r+n_j)}{\Gamma(n_j+1)\Gamma(r)} \left(\frac{\lambda_j}{\lambda_j+r}\right)^{n_j} \left(\frac{r}{\lambda_j+r}\right)^r, \quad n_j = 0, 1, \dots
 \end{aligned}$$

One can write the last step directly from the previous step because we recognize that the part behind the summation is a probability mass function of a negative binomial. Another way to show that the marginal distribution of N_j is negative binomial is to use either probability generating or characteristic functions.

Compared with the independent model, this binomial thinning considers the dependency among three lines of business. However, a very obvious drawback is the unchangeable dependent structure, as the dependent relationship is tied to the marginal distributions.

3.3. Copula-Based Frequency Model

To overcome the drawback of the binomial thinning model, one can use copulas, which were originally defined by (Sklar 1959), where the joint distribution of N_1, \dots, N_k (denoted by H) can be written as a combination of a copula C and the corresponding marginal distributions F_1, \dots, F_k as follows:

$$H(n_1, \dots, n_k) = \mathbb{P}(N_1 \leq n_1, \dots, N_k \leq n_k) = C(F_1(n_1), \dots, F_k(n_k)). \tag{8}$$

As we consider the frequencies from three types of insurance coverage, one can write the joint probability of the claim frequencies via a copula C as follows:

$$\begin{aligned}
 &\mathbb{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3) = \\
 &C(F_1(n_1), F_2(n_2), F_3(n_3)) - C(F_1(n_1 - 1), F_2(n_2), F_3(n_3)) - \\
 &C(F_1(n_1), F_2(n_2 - 1), F_3(n_3)) - C(F_1(n_1), F_2(n_2), F_3(n_3 - 1)) + \\
 &C(F_1(n_1 - 1), F_2(n_2 - 1), F_3(n_3)) + C(F_1(n_1 - 1), F_2(n_2), F_3(n_3 - 1)) + \\
 &C(F_1(n_1), F_2(n_2 - 1), F_3(n_3 - 1)) - C(F_1(n_1 - 1), F_2(n_2 - 1), F_3(n_3 - 1)).
 \end{aligned} \tag{9}$$

To maintain consistency in our analysis, we again assume the same marginal distributions of N_j , which means that $N_j \sim \mathcal{NB}(\lambda_j, r_i)$ for $i = 1, 2, 3$. Regarding the copula families, we use the following three-dimensional copulas:

- Gaussian:

$$C(u_1, u_2, u_3) = \Phi_{3|\Sigma}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3)) \tag{10}$$

where $\Phi_{3|\Sigma}$ is the joint distribution function of the trivariate normal distribution with mean 0 and the (exchangeable) covariance matrix $\Sigma = \sigma J_3 J_3' + (1 - \sigma) I_3$ with $J_3 = (1, 1, 1)'$, I_3 is an identity matrix of size 3 for $\sigma \in (-1, 1)$, and Φ^{-1} is the quantile function of a standard normal random variable. It is implicitly assumed all pairwise correlations in the correlation matrix are the same, which means that the Gaussian copula has an exchangeable structure.

- Gumbel:

$$C(u_1, u_2, u_3) = \exp \left[- \left(-(\log u_1)^{\sigma_G} - (\log u_2)^{\sigma_G} - (\log u_3)^{\sigma_G} \right)^{1/\sigma_G} \right], \quad (11)$$

where $\sigma_G \geq 1$ is the parameter of the Gumbel copula. A larger σ_G value indicates that any pairwise marginals are more positively related.

- Joe:

$$C(u_1, u_2, u_3) = 1 - (1 - [1 - (1 - u_1)^{\sigma_J}][1 - (1 - u_2)^{\sigma_J}][1 - (1 - u_3)^{\sigma_J}])^{1/\sigma_J}, \quad (12)$$

where $\sigma_J \geq 1$ is the parameter the Joe copula. Similar to the Gumbel copula, a larger σ_J constructs stronger positive dependency between any pairwise marginals.

4. Composite Models for Heavy-Tailed Severities

Several positive continuous distributions can be used to study the claim amounts distribution. While some distributions such as gamma and lognormal are good candidates for modelling the low-cost range, they might not be able to capture the heavy-tail behaviour. In this regard, we consider some two-component composite models to model both the body and tail parts in a balanced way.

A two-component composite model combines the body part of a light-tailed distribution with the tail part of a heavy-tailed distribution. Different from mixture distributions, there is no overlap between the supports of these components. By denoting the light-tailed and heavy-tailed densities/distribution functions as $g_1(Y)/G_1(Y)$ and $g_2(Y)/G_2(Y)$, respectively, one can write the density of a composite random variable with two components as follows:

$$g_{comp}(y) = \begin{cases} \frac{1}{1 + \phi} \frac{g_1(y)}{G_1(u)} & y < u; \\ \frac{\phi}{1 + \phi} \frac{g_2(y)}{1 - G_2(u)} & y \geq u, \end{cases} \quad (13)$$

where ϕ is the weight parameter, and u is the threshold to separate the two components. The cumulative distribution function of a composite model can be expressed as follows:

$$G_{comp}(y) = \begin{cases} \frac{1}{1 + \phi} \frac{G_1(y)}{G_1(u)} & y < u; \\ \frac{1}{1 + \phi} + \frac{\phi}{1 + \phi} \frac{G_2(y)}{1 - G_2(u)} & y \geq u. \end{cases} \quad (14)$$

Regarding the estimation scheme, one can use the maximum likelihood estimation to find the optimal parameters for the body and tail distributions. We note that the threshold u and weight parameter ϕ are not estimated but determined to guarantee the continuity and differentiability of the composite distribution at the threshold with the following constraints:

1. Continuity:

$$\lim_{y \rightarrow u^-} g_{comp}(y) = \lim_{y \rightarrow u^+} g_{comp}(y) \implies \phi = \frac{\lim_{y \rightarrow u^-} \frac{g_1(y)}{G_1(u)}}{\lim_{y \rightarrow u^+} \frac{g_2(y)}{1 - G_2(u)}} = \frac{g_1(u)(1 - G_2(u))}{g_2(u)G_1(u)}; \quad (15)$$

2. Differentiability:

$$\frac{1}{1 + \phi} \lim_{y \rightarrow u^-} \frac{d}{dy} \frac{g_1(y)}{G_1(u)} = \frac{\phi}{1 + \phi} \lim_{y \rightarrow u^+} \frac{d}{dy} \frac{g_2(y)}{1 - G_2(u)} \implies \frac{d}{du} \ln \left(\frac{g_1(u)}{g_2(u)} \right) = 0. \quad (16)$$

For example, assume that Y_1 and Y_2 follow gamma and inverse gamma distributions, that is $Y_1 \sim \mathcal{G}(\alpha_1, \theta_1)$ and $Y_2 \sim \mathcal{IG}(\alpha_2, \theta_2)$ with the following density functions:

$$g_1(y_1) = \frac{(y_1/\theta_1)^{\alpha_1} e^{-y_1/\theta_1}}{y_1 \Gamma(\alpha_1)}, \quad y_1 \geq 0, \tag{17}$$

$$g_2(y_2) = \frac{(\theta_2/y_2)^{\alpha_2} e^{-\theta_2/y_2}}{y_2 \Gamma(\alpha_2)}, \quad y_2 \geq 0. \tag{18}$$

With Equations (15) and (16), one can find the threshold and weight parameters as functions of the distribution parameters as follows:

$$\begin{aligned} 0 &= \frac{d}{du} \left[\ln \frac{g_1(u)}{g_2(u)} \right] \\ &= \frac{d}{du} \left[\ln \frac{\frac{(u/\theta_1)^{\alpha_1} e^{-u/\theta_1}}{u \Gamma(\alpha_1)}}{\frac{(\theta_2/u)^{\alpha_2} e^{-\theta_2/u}}{u \Gamma(\alpha_2)}} \right] \\ &= \frac{d}{du} \left[\alpha_1 \ln u - \frac{u}{\theta_1} + \alpha_2 \ln u + \frac{\theta_2}{u} \right] \\ &= (\alpha_1 + \alpha_2)u - \frac{u^2}{\theta_1} - \theta_2 \implies \\ u &= \frac{\alpha_1 + \alpha_2 + \sqrt{(\alpha_1 + \alpha_2)^2 - 4 \frac{\theta_2}{\theta_1}}}{2/\theta_1} \end{aligned} \tag{19}$$

$$\phi = \frac{g_1(u)(1 - G_2(u))}{g_2(u)G_1(u)}. \tag{20}$$

Other composite models considered are listed in Table 3 with corresponding equations to determine the threshold values u . We note that the weight parameter ϕ is given by (15).

Table 3. Threshold values u for various composite models (\mathcal{G} : gamma, \mathcal{E} : exponential, \mathcal{IG} : Inverse-gamma, \mathcal{LN} : lognormal, and \mathcal{Pa} : Pareto).

Name	Head Dist.	Tail Dist.	u
\mathcal{G} and \mathcal{IG}	$\frac{(x/\theta_1)^{\alpha_1} e^{-x/\theta_1}}{x \Gamma(\alpha_1)}$	$\frac{(\theta_2/x)^{\alpha_2} e^{-\theta_2/x}}{x \Gamma(\alpha_2)}$	$u = \frac{\alpha_1 + \alpha_2 + \sqrt{(\alpha_1 + \alpha_2)^2 - 4 \frac{\theta_2}{\theta_1}}}{2/\theta_1}$
\mathcal{G} and \mathcal{LN}	$\frac{(x/\theta_1)^{\alpha_1} e^{-x/\theta_1}}{x \Gamma(\alpha_1)}$	$\exp\left\{-\frac{(\ln x - \mu_2)^2}{2\sigma_2^2}\right\}$	$0 = \alpha_1 - \frac{u}{\theta_1} + \frac{\ln x - u}{\sigma_2^2}$
\mathcal{G} and \mathcal{Pa}	$\frac{(x/\theta_1)^{\alpha_1} e^{-x/\theta_1}}{x \Gamma(\alpha_1)}$	$\frac{\alpha_2 \theta_2^{\alpha_2}}{(x + \theta_2)^{\alpha_2 + 1}}$	$u = \frac{\alpha_1 + \alpha_2 - \frac{\theta_2}{\theta_1} + \sqrt{(\alpha_1 + \alpha_2 \frac{\theta_2}{\theta_1})^2 + 4 \frac{\theta_2}{\theta_1} (\alpha_1 - 1)}}{2/\theta_1}$
\mathcal{E} and \mathcal{IG}	$\frac{e^{-x/\theta_1}}{\theta_1}$	$\frac{(\theta_2/x)^{\alpha_2} e^{-\theta_2/x}}{x \Gamma(\alpha_2)}$	$u = \frac{\alpha_2 + 1 + \sqrt{(\alpha_2 + 1)^2 - 4 \frac{\theta_2}{\theta_1}}}{2/\theta_1}$
\mathcal{E} and \mathcal{LN}	$\frac{e^{-x/\theta_1}}{\theta_1}$	$\exp\left\{-\frac{(\ln x - \mu_2)^2}{2\sigma_2^2}\right\}$	$0 = -\frac{1}{\theta_1} + \frac{1}{u} + \frac{\ln u - \mu_2}{u \sigma_2^2}$
\mathcal{E} and \mathcal{Pa}	$\frac{e^{-x/\theta_1}}{\theta_1}$	$\frac{\alpha_2 \theta_2^{\alpha_2}}{(x + \theta_2)^{\alpha_2 + 1}}$	$u = (\alpha_2 + 1)\theta_1 - \theta_2$

5. Empirical Analysis and Implications for Risk Management

5.1. Estimation Results

The logic of estimating the parameters using the benchmark model is straightforward, so that one can directly maximize the joint log-likelihood with a numerical routine, for example, the `optim` function in R. Table 4 shows the estimated values for the binomial thinning model and the benchmark, independent frequency model. The point estimates of $\lambda, \lambda_1, \lambda_2,$ and λ_3 under the two models are similar. Additionally, the standard errors of the mean parameters are smaller compared with the standard errors of dispersion parameters, $r, r_1, r_2,$ and r_3 . It is observed that there is a significant level of improvement in the log-likelihood by incorporating dependence via the common factor. We note that

the improvements in AIC and BIC are even greater with the dependence modelling as the binomial thinning model is more parsimonious than the independent model.

Table 4. Parameter estimates for the binomial thinning and independent frequency models.

	Binomial Thinning Model				Independent Frequency Model			
	Estimates	CI Lower (95%)	CI Upper (95%)	Std.Err	Estimates	CI Lower (95%)	CI Upper (95%)	Std.Err
λ_1	15.08	14.24	15.91	0.43	15.08	14.21	15.95	0.44
r_1	-	-	-	-	20.74	8.96	32.52	6.01
λ_2	12.72	11.97	13.47	0.2	12.72	11.92	13.52	0.41
r_2	-	-	-	-	17.59	7.55	27.64	5.12
λ_3	4.67	4.27	5.07	0.20	4.67	4.11	5.22	0.28
r_3	-	-	-	-	3.62	1.94	5.30	0.86
λ	16.42	15.53	17.30	0.45	16.42	15.53	17.30	0.45
r	25.32	10.03	40.62	7.80	25.24	10.05	40.43	7.75
$\log \mathcal{L}$		-1183.47				-1516.57		
AIC		2382.94				3043.14		
BIC		2406.00				3057.56		

As mentioned in the previous section, the joint distribution of a copula-based model combines marginal distributions with a copula function. Here, we use the inference by margin (IFM) method, so that the marginal distributions in the independent model are considered as given, while only the copula part is additionally estimated. Table 5 shows the estimated copula parameters and the log-likelihood values of each of the copula models. We note that the parameter estimated with the Gaussian copula model implies positive relationships among the three lines. By comparing the log-likelihood, the Gaussian copula outperforms the others.

Table 5. The estimates and log-likelihood of copula models.

	Gaussian Copula	Gumbel Copula	Joe Copula
Est. parameter	0.70452	1.83147	2.17170
$\log \mathcal{L}$	-1015.953	-1021.079	-1033.461

For the severity components, we use several composite models. For the body part (modelled with a light-tailed distribution), we consider the gamma and exponential distributions. For the tail part (modelled with a heavy-tailed distribution), we use inverse-gamma, Pareto, and lognormal distributions. In the following Tables 6–8, the model selection criteria for various composite models are demonstrated, fitted with the building/content/profit severity data, respectively.

Table 6. Log-likelihoods of composite models for building severity.

	\mathcal{G} and \mathcal{IG}	\mathcal{G} and \mathcal{Pa}	\mathcal{G} and \mathcal{LN}	\mathcal{E} and \mathcal{IG}	\mathcal{E} and \mathcal{Pa}	\mathcal{E} and \mathcal{LN}
$\log \mathcal{L}$	-2800.93	-2771.15	-2771.14	-3181.33	-3220.69	-3220.72
AIC	5609.87	5550.30	5550.29	6368.65	6447.37	6447.43
BIC	5632.25	5572.68	5572.67	6385.44	6464.16	6464.22

Table 7. Log-likelihoods of composite models for content severity.

	\mathcal{G} and \mathcal{IG}	\mathcal{G} and \mathcal{Pa}	\mathcal{G} and \mathcal{LN}	\mathcal{E} and \mathcal{IG}	\mathcal{E} and \mathcal{Pa}	\mathcal{E} and \mathcal{LN}
$\log \mathcal{L}$	-2187.88	-2039.52	-2037.59	-2102.97	-2102.81	-2102.16
AIC	4383.77	4087.04	4083.18	4211.95	4211.61	4210.32
BIC	4405.47	4108.74	4104.88	4228.23	4227.89	4226.60

Table 8. Log-likelihoods of composite models for profit severity.

	\mathcal{G} and \mathcal{IG}	\mathcal{G} and \mathcal{Pa}	\mathcal{G} and \mathcal{LN}	\mathcal{E} and \mathcal{IG}	\mathcal{E} and \mathcal{Pa}	\mathcal{E} and \mathcal{LN}
log \mathcal{L}	−309.19	−297.19	−427.81	−305.93	−304.53	−304.48
AIC	626.38	602.39	863.62	617.86	615.06	614.97
BIC	644.07	620.08	881.31	631.13	628.33	628.24

In the case of building losses, the gamma and lognormal (\mathcal{G} and \mathcal{LN}) and gamma and Pareto (\mathcal{G} and \mathcal{Pa}) distributions are shown to have the best goodness-of-fit. Likewise, we find that gamma and lognormal (\mathcal{G} and \mathcal{LN}) is the best for modelling contents losses, and gamma and Pareto (\mathcal{G} and \mathcal{Pa}) fits the profits losses well. Table 9 shows the point estimates of three composite distributions’ parameters, given the best combinations for each coverage, along with the splicing points distribution and the corresponding weight parameter values based on the parameter estimates. We note that some transformation is required to make the weight parameter meaningful. For example, in the case of building losses, we can interpret $\frac{1}{1+\phi} = 0.2433$ as the proportion of Y that is from the body part, of the gamma distribution. On the other hand, $\frac{\phi}{1+\phi} = 0.7567$ of Y is from the tail part, of the lognormal distribution. Specifically, a larger weight parameter value indicates that the composite model is more heavy-tailed, and vice versa. The splicing point parameter, u , indicates the change of distribution components. For the building coverage, the splicing parameter is 2.08943, which means that the building losses greater than 2.08943 million Danish Krone are modelled by a lognormal distribution. Additionally, we observe that the losses from the profit line are more heavy-tailed compared with the losses from the other two lines.

Table 9. Parameter estimates for the severity components.

	Building: \mathcal{G} and \mathcal{LN}	Contents: \mathcal{G} and \mathcal{LN}	Profits: \mathcal{G} and \mathcal{Pa}
Head Dist.	$\alpha_1 = 3.71085$ $\theta_1 = 0.37198$	$\alpha_1 = 1.98766$ $\theta_1 = 0.21591$	$\alpha_1 = 1.55072$ $\theta_1 = 0.10144$
Tail Dist.	$\mu_2 = -331.88884$ $\sigma_2 = 13.20987$	$\mu_2 = -1.34871$ $\sigma_2 = 1.69228$	$\alpha_2 = 1.41237$ $\theta_2 = 0.37195$
u	2.08943	0.47466	0.11282
ϕ	0.32151	1.34244	2.92302

5.2. Empirical Findings for Risk Management

In the insurance industry, estimating the risk level for a product or portfolio is critical for determining appropriate levels of the premium and reserve. We recall that S_j and S_\bullet mean that the random variable stands for the aggregate loss amount from the j th line and the aggregate loss for all lines of insurance, as defined by (1) and (2). It is also straightforward to see that $\mathbb{E}[S_\bullet] = \sum_{j=1}^3 \mathbb{E}[S_j]$ due to the additivity of expectation. However, such a property generally does not hold for other types of risk measures, so it is important to properly analyze the risk level of total claims S_\bullet , rather than summing up the risk level of S_1, S_2 , and S_3 . For our risk analysis, we use the following well-known risk measures:

1. Value at Risk (VaR) – $\text{VaR}_\alpha(Y) = \min\{y \in \mathbb{R} : F_Y(y) \geq \alpha\}$, $\alpha \in [0, 1]$;
2. Tail Value at Risk (TVaR) – $\text{TVaR}_\alpha(Y) = \mathbb{E}[Y|Y \geq \text{VaR}_\alpha(Y)]$, $\alpha \in [0, 1]$;
3. Proportional Hazard (PH) Risk Measure (Wang 1995)– $\text{PH}_\alpha(X) = \int_0^\infty (1 - F(y))^{1/\alpha} dy$, $\alpha \geq 1$.
4. Dual Power (DP) Risk Measure– $\text{DP}_\beta(X) = \int_0^\infty 1 - (F(y))^\beta dy$, $\beta \geq 1$.

While TVaR is not a coherent risk measure unless the underlying distribution is continuous, it is innocuous to assume that the TVaR of S_1, S_2, S_3 , and S_\bullet are coherent, as we mainly focus on the tail part, where the claim amounts are for sure strictly positive and the underlying distributions are continuous. We also note that (Wang 1994) showed that the integration of the transformed distribution is coherent when the transformation is a concave function, so that PH and DP are both coherent.

For comparison of the calculated risk measures under each of the models, we apply a Monte Carlo simulation to numerically evaluate the values of risk measures. More specifically, we simulated 100,000 data points, number of accidents M , the claim numbers for three business lines N_1, N_2 , and N_3 , and, subsequently, the claim amounts S_1, S_2 , and S_3 under each of the model specifications with the estimated parameters shown in Section 5.1.

The simulation for the independent model is straightforward. Because of the independence, we apply random generations for the negative binomial distributions to simulate claim numbers N_1, N_2 , and N_3 for all business lines. Unlike the independent one, the binomial thinning model requires the simulation of the reported number of accidents M . After that, a binomial random generation with size parameters corresponding to the reported claim numbers is applied to get the claim numbers N_1, N_2 , and N_3 for three lines of business. The logic for the copula models is similar. We first generate trivariate uniform random numbers from the copula functions. With the generated uniform random numbers, we get the claim numbers N_1, N_2 , and N_3 using the inverse of the marginal distributions. Once the claim frequencies N_1, N_2 , and N_3 were generated, the severity components are generated, subsequently. For example, if N_1 is given, then uniform random numbers are generated N_1 times and they are converted to the individual severities via the inverse distribution (or quantile) function of the composite distribution function for building losses. Lastly, these values are summed up as S_1 .

Figures 2–4 show the scatterplots of the combinations of building, content, and profit claims for observed and simulated frequency data. Based on the plots of observed data, there are apparent positive relationships among the marginal frequencies. As we expected, however, the independent model cannot capture such dependent behaviours. In the case of the other models, the binomial thinning model shows a substantial linear relationship between the building and contents claim numbers, which is the most similar to the observed. In the case of Figure 3, however, the Joe copula best captures the relationship between the building and profit frequencies.

Lastly, Table 10 shows the approximated risk measures under different models. The independent model reproduces relatively smaller values of VaR and TVaR for the aggregated claims $S_\bullet = S_1 + S_2 + S_3$, whereas the calculated risk measure values for each coverage, S_1, S_2 , and S_3 are more or less the same, regardless of the chosen model. This is quite natural as, regardless of the (assumed) dependence structure, the marginal distributions for N_1, N_2 , and N_3 (and, subsequently, S_1, S_2 , and S_3) are the same. As a result, the TVaR for S_\bullet under the independent model is severely underestimated compared to the observed (or empirical) TVaR, while the other dependent models are able to reproduce the empirical TVaR for S_\bullet with less deviations. It implies that it is required to consider possible dependence among different types of insurance coverage for an effective enterprise risk management purpose.

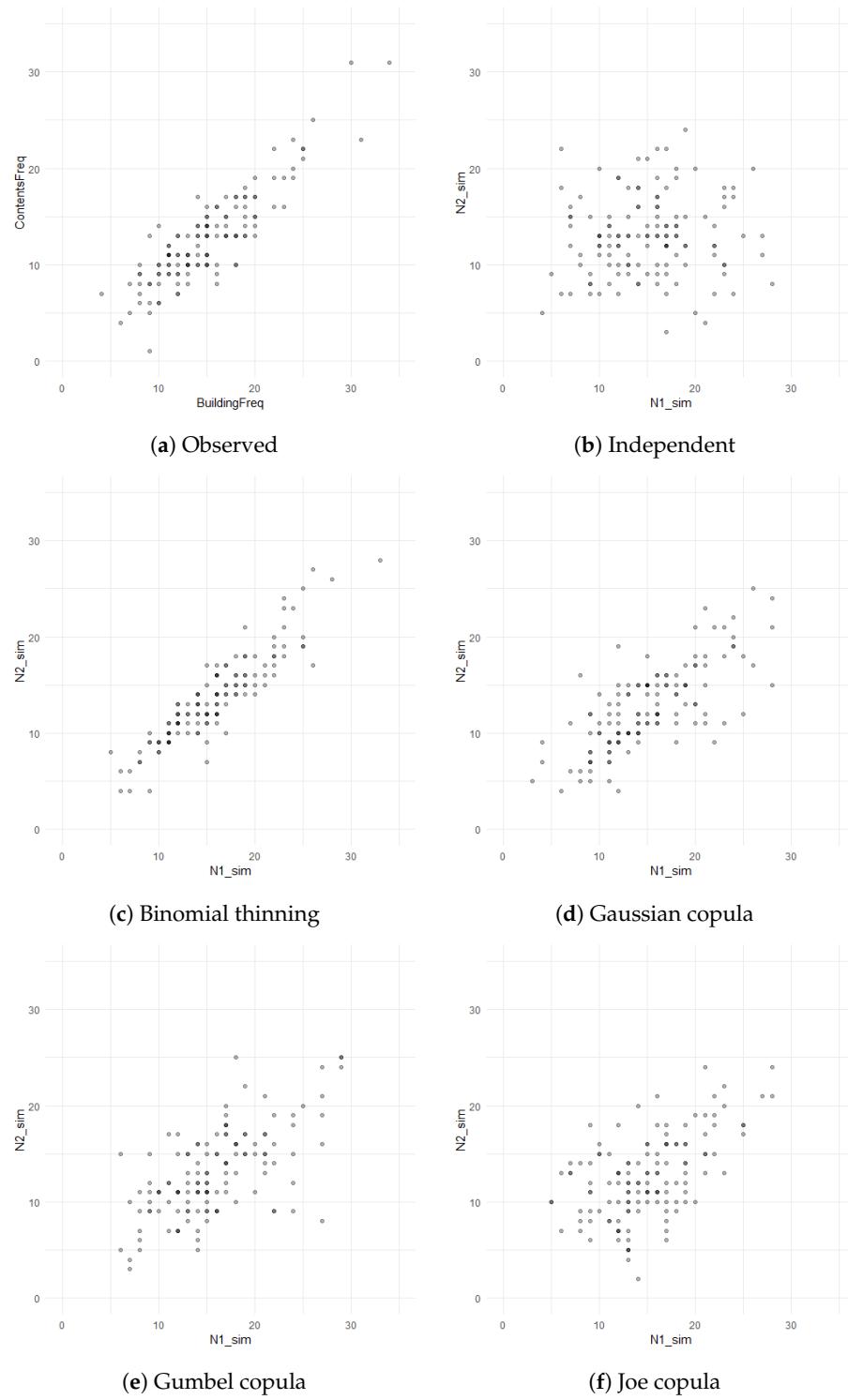


Figure 2. Observed and simulated building vs. contents claims.



Figure 3. Observed and simulated building vs. profits claims.

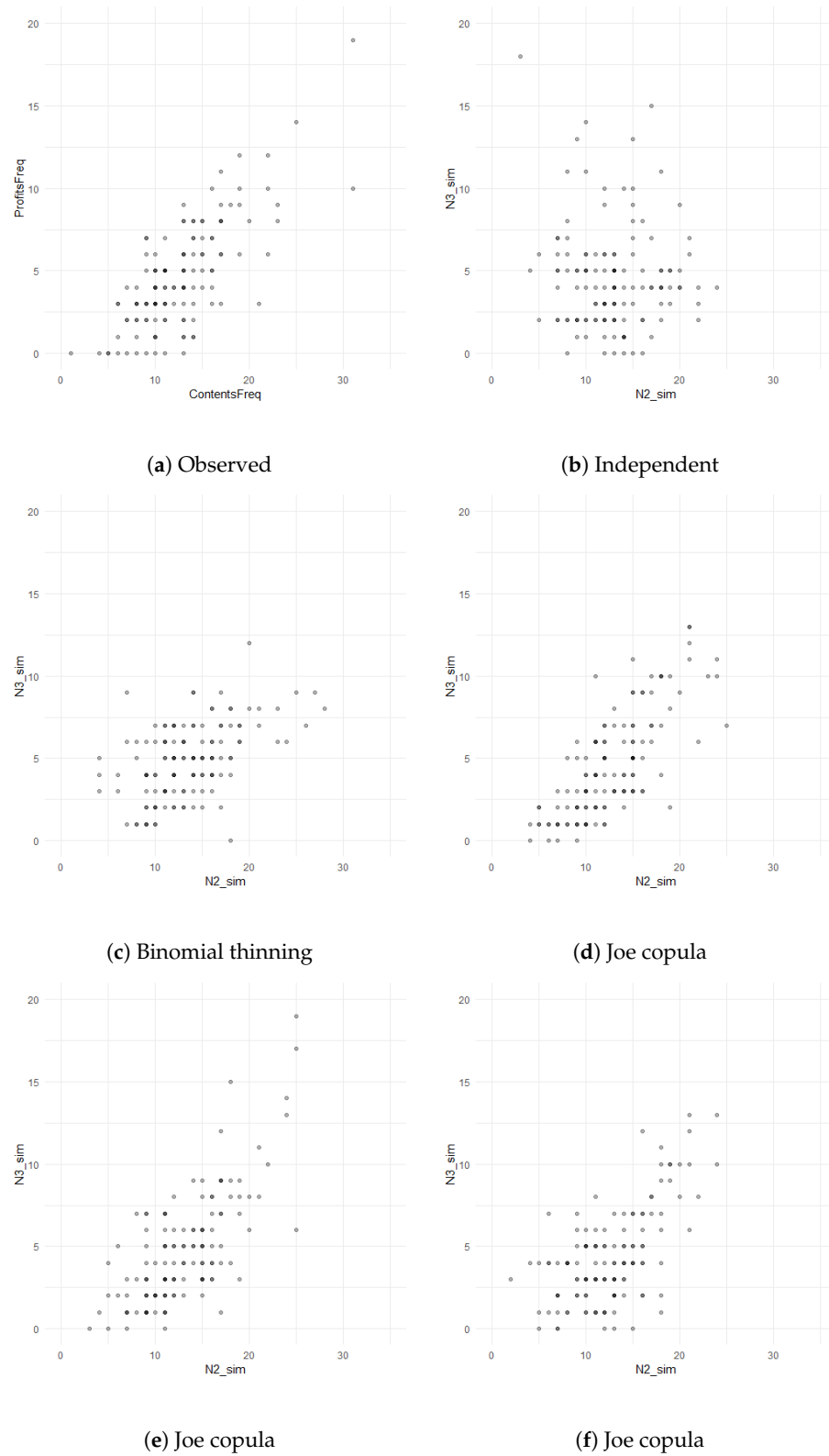


Figure 4. Observed and simulated contents vs. profits claims.

Table 10. Values of risk measures under different models.

Measure	Model	Building (S_1)	Content (S_2)	Profit (S_3)	Aggregate (S_\bullet)
VaR _{0.90}	Observations	43.36753	37.61073	8.927676	84.95829
	Independent	45.83948	40.13501	8.853674	82.54968
	Bin. Thin.	45.55415	39.81174	8.330259	87.44851
	Gaussian	46.00964	39.93148	8.925007	88.03041
	Gumbel	45.94382	40.4155	8.80955	88.64008
	Joe	46.00862	40.25366	8.792391	88.37843
TVaR _{0.90}	Observations	67.63288	61.7309	17.34992	130.9614
	Independent	62.29311	64.19727	21.77994	114.7784
	Bin. Thin.	62.18439	63.46745	22.57028	122.2398
	Gaussian	62.57564	63.71817	24.13904	123.6911
	Gumbel	62.72085	64.09827	24.19948	124.4939
	Joe	63.20349	63.48082	22.36632	122.8164
TVaR _{0.95}	Observations	88.59588	82.65819	24.1481	171.8545
	Independent	75.19079	82.92127	32.76977	140.1614
	Bin. Thin.	75.28895	82.01287	34.9973	149.5173
	Gaussian	75.53303	82.18786	37.40043	151.7549
	Gumbel	75.80739	82.24379	37.66422	152.5675
	Joe	76.64548	81.35734	34.09076	149.5723
PH ₂	Observations	51.93275	42.81760	11.31421	93.62481
	Independent	54.19547	56.37977	28.35901	102.4392
	Bin. Thin.	58.84183	56.89981	40.01965	114.9348
	Gaussian	53.31081	50.2014	63.63473	134.7637
	Gumbel	53.52596	49.71846	55.63991	126.7038
	Joe	60.19573	47.35576	38.58851	111.7715
DP ₃	Observations	43.22733	35.96571	8.24774	82.29961
	Independent	41.57237	35.75419	9.25285	76.3109
	Bin. Thin.	41.50927	35.52621	9.46211	79.59646
	Gaussian	41.69046	35.60265	9.98022	80.09476
	Gumbel	41.69769	35.82457	9.98923	80.20223
	Joe	41.82949	35.57191	9.42534	79.60271

6. Conclusions and Discussions

In conclusion, we bridged the connection among different coverages by considering the loss amounts incurred under different coverages due to a fire accident and taking into account the heavy-tail behaviour. From the insurance aspect, we assessed several risk measures, which can be interpreted differently. For example, the Value at Risk with α can be interpreted as the assets that should be reserved to reduce the bankruptcy possibility to $1 - \alpha$. By comparing with the fully independent model, we found that both dependent modelling frameworks performed better from both statistical and insurance aspects. Specifically, the binomial thinning model captured the behaviour of the observed claim numbers better than the independent model from the calculated model evaluation criterion. All binomial thinning and copula-based models provided more reasonable and consistent risk measures.

Additionally, we presented two modelling frameworks to capture the dependency: binomial thinning and copula-based. Although, from the approximate risk measures results, we could not conclude which is the best, we could still observe the flexibility of copula-based models. The binomial thinning model suggested a certain dependent structure. However, by implying different copula functions, we may capture the dependence based on different joint distributions.

There are some concerns and limitations of the current research. Firstly, while we used copulas with discrete random variables, (Genest and Nešlehová 2007) discussed the limitations of applying copula with discrete random variables. Thus, we shall carefully interpret the results relative to the dependent measures because of the lack of uniqueness

of the copula functions. However, it is still effective when the discrete random variables' probability mass is spread widely enough on their support. Secondly, we also implicitly assumed that the frequency and severity components are independent, whereas some existing literature show the presence of dependence among the frequency and severity components, including, but not limited to, (Jeong and Valdez 2020) and (Vernic et al. 2021).

For future research, (Geenens 2020) proposed a method of incorporating dependency for discrete random variables by using an idea similar to the copula method. Based on and further improving it, we can provide more rigorous analyses for the dependent multivariate discrete data. Additionally, one can also study the dependency between the frequency and severity components on top of the dependence among the claims from multiple types of coverage.

Author Contributions: Conceptualization, H.J. and Y.L.; methodology, H.J. and Y.L.; software, T.Y. and H.J.; validation, T.Y.; formal analysis, T.Y.; investigation, T.Y. and H.J.; data curation, T.Y. and H.J.; writing—original draft preparation, T.Y. and H.J.; writing—review and editing, H.J. and Y.L.; visualization, T.Y.; supervision, H.J. and Y.L.; project administration, H.J. and Y.L.; funding acquisition, H.J. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NSERC Discovery grant number R611467/R611851 and CANSSI GSE, Scholarship number R619645.

Data Availability Statement: The research data used in this article are available with a R package `CASdatasets`. The R codes to reproduce the results in this article are available upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VaR	Value at Risk
TVaR	Tail Value at Risk
PH	Proportional Hazard
DP	Dual Power

References

- Cooray, Kahadawala, and Chin-I Cheng. 2015. Bayesian estimators of the lognormal–Pareto composite distribution. *Scandinavian Actuarial Journal* 2015: 500–15. [\[CrossRef\]](#)
- Fung, Tsz Chai, Himchan Jeong, and George Tzougas. 2023. Investigating the effect of climate-related hazards on claim frequency prediction in motor insurance. *SSRN Electronic Journal* SSRN 4638074. [\[CrossRef\]](#)
- Geenens, Gery. 2020. Copula modeling for discrete random vectors. *Dependence Modeling* 8: 417–40. [\[CrossRef\]](#)
- Genest, Christian, and Johanna Nešlehová. 2007. A Primer on Copulas for Count Data. *Astin Bulletin* 37: 475–515. [\[CrossRef\]](#)
- Hong, Liang, and Ryan Martin. 2018. Dirichlet process mixture models for insurance loss data. *Scandinavian Actuarial Journal* 2018: 545–54. [\[CrossRef\]](#)
- Jeong, Himchan, and Emiliano A. Valdez. 2020. Predictive compound risk models with dependence. *Insurance: Mathematics and Economics* 94: 182–95. [\[CrossRef\]](#)
- Jeong, Himchan, George Tzougas, and Tsz Chai Fung. 2023. Multivariate claim count regression model with varying dispersion and dependence parameters. *Journal of the Royal Statistical Society Series A: Statistics in Society* 186: 61–83. [\[CrossRef\]](#)
- Jeong, Himchan. 2024. Tweedie multivariate semi-parametric credibility with the exchangeable correlation. *Insurance: Mathematics and Economics* 115: 13–21. [\[CrossRef\]](#)
- Lee, Gee Y., and Peng Shi. 2019. A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insurance: Mathematics and Economics* 87: 115–29. [\[CrossRef\]](#)
- McNeil, Alexander J. 1997. Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory. *ASTIN Bulletin* 27: 117–37. [\[CrossRef\]](#)
- Miljkovic, Tatjana, and Bettina Grün. 2016. Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics* 70: 387–96. [\[CrossRef\]](#)
- Oh, Rosy, Himchan Jeong, Jae Youn Ahn, and Emiliano A. Valdez. 2021. A multi-year microlevel collective risk model. *Insurance: Mathematics and Economics* 100: 309–28. [\[CrossRef\]](#)
- Pigeon, Mathieu, and Michel Denuit. 2011. Composite Lognormal–Pareto model with random threshold. *Scandinavian Actuarial Journal* 2011: 177–92. [\[CrossRef\]](#)

- Resnick, Sidney I. 1997. Discussion of the Danish Data on Large Fire Insurance Losses. *ASTIN Bulletin* 27: 139–51. [[CrossRef](#)]
- Scollnik, David P., and Chenchen Sun. 2012. Modeling with Weibull-Pareto models. *North American Actuarial Journal* 16: 260–72. [[CrossRef](#)]
- Sklar, Abe. 1959. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris* VIII: 229–31.
- Vernic, Raluca, Catalina Bolancé, and Ramon Alemany. 2021. Sarmanov distribution for modeling dependence between the frequency and the average severity of insurance claims. *Insurance: Mathematics and Economics* 102: 111–25. [[CrossRef](#)]
- Wang, Shaun. 1994. Premium Calculation by Transforming the Layer Premium Density. *ASTIN Bulletin* 26: 71–92. [[CrossRef](#)]
- Wang, Shaun. 1995. Insurance Pricing and Increased Limits Ratemaking by Proportional Hazards Transforms. *Insurance: Mathematics and Economics* 17: 43–54. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.