Big Data and Information Analytics

http://www.aimspress.com/journal/BDIA

BDIA, 6: 56–73.
DOI:10.3934/bdia.2021005
Received: 14 Janurary 2021
Accepted: 03 June 2021
Published: 16 June 2021

*Research article*

# Aggregate loss model with Poisson-Tweedie frequency

**S. Chen**[1,*], **Z. Wang**[1] **and M. Kelly**[2]

[1] Department of Mathematics, Wilfrid Laurier University, Waterloo ON N2L 3C5, Canada

[2] Lazaridis School of Business & Economics, Wilfrid Laurier University, Waterloo ON N2L 3C5, Canada

* **Correspondence:** Email: chen8470@mylaurier.ca.

**Abstract:** Aggregate loss models are used by insurers to make operational decisions, set insurance premiums, optimize reinsurance and manage risk. The aggregate loss is the summation of all random losses that occurred in a period, and it is a function of both the loss severity and the loss frequency. The need for a flexible model in fitting severity has been well studied in the literature. We extend this work by introducing the Poisson-Tweedie distribution family for the frequency distribution. The Poisson-Tweedie distribution family contains many of the commonly used distributions for modelling loss frequency, thus making loss frequency fitting more flexible and reducing the chance of model misspecification. Using simulation, we show that the sensitivity of percentile based risk measures to different specifications of the frequency distribution. We then apply our proposed model to the Transportation Security Administration (TSA) claims data to demonstrate modelling capacity of the Poisson-Tweedie distribution.

**Keywords:** aggregate loss models; Poisson-Tweedie distribution; distribution simulation; percentile estimation

## 1. Introduction

The aggregate loss model, which describes the distribution of the total loss (typically within a portfolio) within a fixed period of time, is used to make operational decisions, set insurance premiums, optimize reinsurance purchases and manage both solvency and liquidity risk. Regulators, who are charged with ensuring that insurance companies remain solvent, require that insurance companies hold enough capital to provide security against unexpected or extreme losses. Percentile based risk measures, such as value at risk (VaR) or expected shortfall (ES), are derived from aggregate loss models and can be used to calculate "worst case scenarios". For example, the aggregate loss model is adopted by the advanced measurement approaches (AMA) for operational risk, to estimate regulatory capital

which is the $99.9^{th}$ percentile of the aggregate loss distribution: this is stipulated in the BASEL II Accord set by the Basel Committee on Banking Supervision. These measures enable financial institutions to estimate potential losses with credibility. Furthermore, percentiles of the aggregate loss distribution can also be used to calculate the optimal amount of stop-loss or quota-share reinsurance required (see, for example, Lo and Remorov [1] and Tan, Weng and Zhang [2]).

Within a given period, the aggregate loss $L$ can be expressed as a random sum as follows:

$$L = \sum_{j=1}^{N} X_j, \tag{1.1}$$

where $N$ is the total number of claims observed in a certain period, and $X_j$ is the size of loss for the $j^{th}$ claim. The random variable $N$ is the loss frequency and it is modelled using a non-negative discrete distribution such as the Poisson, Negative Binomial, Binomial, Geometric and Panjer class distributions (Griffiths and Mnif [4], Karam and Planchet [4], and Panjer[5]). The random variable $X_j$, for all $j$, is the loss severity. It is typically assumed that the $X_j$ are independent and identically distributed (I. I. D.) random variables and they are often modelled using the exponential density family with positive support (see, for example, Cummins et al. [6], Griffiths and Mnif [3], Jin, Provost and Ren [7], and Shevchenko [8]).

Recent literature has focused on the importance of fitting flexibility and mathematical tractability when choosing a distribution for severity (see, for example, Bae and Ko [9] and Willmot and Lin [10]). We contend that it is also important to consider fitting flexibility and mathematical tractability when choosing a distribution for frequency. A single distribution may not be flexible enough to fit the observed count data well, and such misspecification will lead to poor estimation of the risk measures of the aggregate loss.

In this paper, we use the Poisson-Tweedie distribution, a family with more flexibility on overdispersion and tail behaviour, to model claim frequency. The use of Poisson-Tweedie in analyzing count data has gained increasing support recently with implementation in R package (Esnaola et al. [11]). When employing the three-parameter parameterization introduced in El-Shaarawi, Zhu and Joe [12], this flexible distribution family encompasses several of the commonly used loss frequency distributions including the Poisson, Negative Binomial and Poisson Inverse-Gaussian distributions. One key element of the Poisson-Tweedie family is its property of convolution closedness with regards to its family index parameter. This implies that whether we aggregate data on a daily, weekly, monthly, quarterly or annual basis, the distribution family of loss frequency remains Poisson-Tweedie.

To reduce the impact from frequency model misspecification, we use the Poisson-Tweedie distribution as the frequency distribution in the aggregate loss model. We then derive the estimates of the distribution parameters for both the frequency and the severity distributions. This allows us to derive the moments of the aggregate loss and calculate the quantiles of the aggregate losses. Next, through extensive simulation studies, we examine the feasibility and flexibility of using Poisson-Tweedie distribution family in modelling aggregate losses. We find that, although the estimation of the mean and variance of the aggregate loss does not differ significantly across the different specifications of the frequency distribution, the accuracy of the estimates of tail quantiles of aggregate loss is impacted by the misspecification.

Finally, to examine the practical feasibility of aggregate loss modelling using Poisson-Tweedie loss frequency, we apply our model to the Transportation Security Administration's (TSA's) claims data

(https://www.dhs.gov/tsa-claims-data). The TSA records all liability claims for bodily injury and property damage made against the organization from 2002 to 2017 and is particularly useful to researchers developing statistical models to analyze claim frequency and severity (Kelly and Wang [13]).

The paper is organized as follows. Section 2 describes the Poisson-Tweedie distribution family and discusses some commonly used distributions for severity. Section 3 details the estimation and inference of the parameters for the loss frequency and loss severity distributions and calculates the quantiles of aggregate losses. Section 4 presents the extensive simulation studies. In section 5, the aggregate loss model with Poisson-Tweedie frequency distribution is applied to the TSA's claims data. We complete the paper with conclusion and future works in section 6.

## 2. Modelling

The aggregate loss at time period $i$ for $i = 1, \cdots, T$ is defined as

$$L_i = \sum_{j=1}^{N_i} X_{ij}$$

where $N_i$ is the loss frequency for time period $i$ and $X_{ij}$ is the loss severity of the $j^{th}$ claim in period $i$ . For the loss frequency, we assume $N_1, \cdots, N_T$ are identically and independently distributed with mass function $f_N(n;\boldsymbol{\theta})$, where $n$ is the value of $N_i$ and the support of $N_i$ is the range of non-negative integers. For the loss severity, we assume that, given $N_i = n_i$ for $i = 1, \cdots, T$, $X_{i1}, \cdots, X_{in_i}$ are identically and independently distributed with density function $f_X(x;\boldsymbol{\beta})$, where $f_X(x;\boldsymbol{\beta}) > 0$ for $x > 0$. Note that when $N_i = 0$, no claim is recorded in period $i$ and, as such, there are no observations of loss severity.

It is commonly assumed that the loss frequency and loss severity are independent. This is a reasonable assumption if there is no deductible, but once a deductible is applied to underlying losses, the deductible will impact both claim frequency and claim severity, invalidating the independence assumption.

The loss severity is often modelled using the exponential density family with positive support (see, for example, Cummins et al. [6], Griffiths and Mnif [3], Jin, Provost and Ren [7], and Shevchenko [8]). The density functions of loss severity are typically well defined with closed forms such that the mean and variance are well-defined and are denoted by $\mu_X$ and $\sigma_X^2$, respectively. It is noted that both $\mu_X$ and $\sigma_X^2$ are functions of $\boldsymbol{\beta}$.

In this paper, we use the Poisson-Tweedie distribution family to model the loss frequency. The Poisson-Tweedie family, denoted by $PT(a, b, c)$, is a three-parameter distribution family that envelops several of the commonly used loss frequency distributions including the Poisson, Negative Binomial and Poisson Inverse-Gaussian distributions. Using the parameterization of El-Shaarawi, Zhu and Joe [12], the three-parameter Poisson–Tweedie family $PT(a, b, c)$ has the following probability mass function:

$$\Pr(N = k + 1) = p_{k+1} = \frac{1}{k+1}\Big(bcp_k + \sum_{j=1}^{k} jr_{k+1-j}p_j\Big), \quad k = 1, 2, 3, \ldots,$$

where

$$r_1 = (1 - a)c, \quad r_{j+1} = \Big(\frac{j - 1 + a}{j + 1}\Big)cr_j, \quad j = 1, 2, 3, \ldots, k - 1,$$

for

$$-\infty < a \leq 1, \quad 0 < b < \infty \quad \text{and} \quad 0 < c < 1.$$

That is, the probability mass $p_{k+1}$ is a linear combination of probability mass $p_0, p_1, \cdots, p_k$. Additionally,

$$\Pr(N = 0) = p_0 = \begin{cases} e^{b[(1-c)a-1]/a} & a \neq 0 \\ (1-c)^b & a = 0, \end{cases}$$

and

$$\Pr(N = 1) = p_1 = bcp_0.$$

Note that parameter $a$ in the Poisson-Tweedie distribution, defined in above parameterization, is the family index and it determines the corresponding distribution. For example, the Poisson-Tweedie family includes Poisson ($a = 1$), Poisson Inverse-Gaussian ($a = 0.5$), Negative Binomial ($a = 0$), and Polya-Aeppli ($a = -1$) distributions.

Parameters $b$ and $c$ are associated with the mean and variance of the distribution: the mean is given by

$$\mu_N = \frac{bc}{(1-c)^{1-a}} \tag{2.1}$$

and the variance is

$$\sigma_N^2 = \frac{bc(1-ac)}{(1-c)^{2-a}}. \tag{2.2}$$

The parameterization of El-Shaarawi, Zhu and Joe [12] is a convenient way to study the various distributions covered by the Poisson-Tweedie family, since its recursive representation of probability mass function makes the estimation of parameters possible.

## 3. Estimation and inference

The parameters of the distribution can be estimated using maximum likelihood estimation (MLE). The likelihood of the aggregate loss is given by:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}; n_1, ..., n_T, x_{1,1}, ...x_{T,n_T}) = \left(\prod_{i=1}^{T} f_N(n_i; \boldsymbol{\theta})\right)\left(\prod_{i=1}^{T}\prod_{j=1}^{n_i} f_X(x_{ij}; \boldsymbol{\beta})\right).$$

Therefore, the log-likelihood is

$$\log L(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{n}, \mathbf{x}) = \sum_{i=1}^{T} \log f_N(n_i; \boldsymbol{\theta}) + \sum_{i=1}^{T}\sum_{j=1}^{n_i} \log f_X(x_{ij}; \boldsymbol{\beta}). \tag{3.1}$$

Taking the partial derivatives of the log-likelihood defined in Eq (3.1) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ and setting these equal to zero yields:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{T} \log f_N(n_i; \boldsymbol{\theta}) = 0 \tag{3.2}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{T}\sum_{j=1}^{n_i} \log f_X(x_{ij}; \boldsymbol{\beta}) = 0. \tag{3.3}$$

The solutions to Eqs (3.2) and (3.3) are the maximum likelihood estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$, respectively. Given the independence assumption, the joint estimation of the frequency and severity parameters is equivalent to estimating them separately. Because the means, $\mu_N$ and $\mu_X$, and variances, $\sigma_N^2$ and $\sigma_X^2$, of loss frequency and severity are functions of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the maximum likelihood estimators of $\mu_N$, $\mu_X$, $\sigma_N^2$ and $\sigma_X^2$, denoted by $\hat{\mu}_N$, $\hat{\mu}_X$, $\hat{\sigma}_N^2$ and $\hat{\sigma}_X^2$, can be obtained by substituting $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$, respectively.

### *3.1. Estimation of moments of aggregate loss*

Under the assumption that the loss frequency $N$ and the loss severity $X_j$ are independent in a given period, the mean and variance of the aggregate loss within that period can be calculated as: $\mu_L = \mu_N \mu_X$ and $\sigma_L^2 = \mu_N \sigma_X^2 + \sigma_N^2 \mu_X^2$ respectively. The mean and variance of the aggregate loss depends only on the marginal means and variances of the loss frequency and loss severity. Hence, the maximum likelihood estimators of the expected value and variance of aggregate loss are $\hat{\mu}_L = \hat{\mu}_N \hat{\mu}_X$ and $\hat{\sigma}_L^2 = \hat{\mu}_N \hat{\sigma}_X^2 + \hat{\sigma}_N^2 \hat{\mu}_X^2$.

### *3.2. Estimation of aggregate loss percentiles*

For the management of both solvency and liquidity risk, the right tail percentile of the aggregate loss is used to calculate value at risk (VaR) and expected shortfall (ES). VaR summarizes "the worst loss over a target horizon with a given level of confidence" (Jorion [14], p. 22). It provides a measure of a given risk exposure at a particular point in time with a certain degree of confidence. The relation between VaR and the aggregate loss $L$ is

$$\Pr(L \leq VaR_\alpha) = \alpha$$

where $L$ is the aggregate loss, $VaR_\alpha$ is the loss amount of the VaR statistic and $\alpha \in [0, 1]$ is the confidence level. The time-period length of the VaR statistic is the same as the time length of the aggregate loss. The loss amount $VaR_\alpha$ is then equivalent to the $\alpha \times 100^{th}$ percentile of the aggregate loss.

ES, also known as tail VaR, is the average of losses greater than a given percentile level and it measures the expected loss if losses exceeds the VaR. Similar to VaR, ES is also composed of a time-period, a confidence level $\alpha$ and a loss amount. It is given by

$$E[L|L \geq VaR_\alpha].$$

To estimate VaR and ES, or any other quantile of the distribution, the functional form of the distribution of the aggregate loss is required. Because the aggregate loss distribution is a combination of the frequency and severity distribution, it often has no closed form, or a mathematically intractable form. Commonly used methods for estimating the quantiles of the aggregate loss are simulation, Fast Fourier Transformation (Embrechts and Frei [15]) or Panjer's recursion (Panjer [5]). In this paper, we use simulation to empirically estimate quantiles of the aggregate loss distribution.

## 4. Simulation study

In this section, we investigate the impact of the choice of the loss frequency distribution on the estimation of percentiles, and in particular the tail quantiles, of the aggregate loss distribution. To

examine the estimation of aggregate loss percentiles across different frequency distributions, we model the loss frequency using the Poisson-Tweedie distribution with different values for the family index, $a$. We then choose the parameters $b$ and $c$ such that the frequency distributions have the same means and, where possible, the same variances. The Poisson-Tweedie probability mass function algorithm is programmed according to El-Shaarawi, Zhu and Joe [12]. We use the Log-Normal distribution for our severity distribution in the calculation of aggregate losses. Using the simulation method, we generate the $95^{th}$ percentile and the expected shortfall (above $95^{th}$ percentile) of aggregate loss.

### *4.1. Design of experiment*

We undertake our simulations across 27 different aggregate loss distributions. For claim severity, we use the Log-Normal distribution with three different sets of values for the parameters $\mu$ and $\sigma$: The distributions simulated are the Log-Normal (7, 0.1), Log-Normal (8, 0.2), and Log-Normal (9, 0.3). The means and standard deviations of the severity distributions are given in Table 1.

For frequency, we use the Poisson (Poisson-Tweedie $a = 1$), the Negative Binomial (Poisson-Tweedie $a = 0$) and the Poisson Inverse-Gaussian (Poisson-Tweedie $a = 0.5$) distributions. The remaining parameters, $b$ and $c$, for the Poisson-Tweedie distributions are set so that the loss frequency distributions have means of 2, 10, and 30 claims per period and, for the Negative Binomial (NB) and the Poisson Inverse-Gaussian (PIG) distributions, the variance is 5 times the mean.

In Table 1 we simulate each distribution 10,000 times and repeat the process 1000 times to calculate an empirical variance. We simulate each distribution 100,000 times to obtain the tail quantiles shown.

### *4.2. Results*

Results of the simulations are detailed in Table 1. In particular, we are interested in the last 2 columns of the table, which show the $95^{th}$ percentile and the ES at the $95^{th}$ percentile. Because the restriction on the variance in the Poisson distribution, in all instances, the aggregate loss distributions simulated using a Poisson distribution have lower $95^{th}$ percentiles and associated ES values than the aggregate loss distributions simulated using either the Poisson Inverse-Gaussian (PIG) (Poisson-Tweedie $a = 0.5$) distribution or the Negative Binomial (NB) (Poisson-Tweedie $a = 0$) distribution.

For all severity distributions, when the loss frequency has a mean of 2, the aggregate loss distributions simulated using the Negative Binomial distribution have higher $95^{th}$ percentile values, but lower associated ES values than the aggregate loss distributions simulated using the Poisson Inverse-Gaussian distribution. When the loss frequency has a mean of 10, the $95^{th}$ the percentile values for aggregate loss distributions are very similar when using either the Negative Binomial or the Poisson Inverse-Gaussian as the underlying frequency distribution. However, the ES values for the aggregate loss distribution are greater when the frequency is modelled using the Poisson Inverse-Gaussian distribution. Finally, when the loss frequency has a mean of 30, the aggregate loss distribution simulated using the Poisson Inverse-Gaussian distribution has both greater $95^{th}$ percentiles and associated ES values than the aggregate loss distribution simulated using the Negative Binomial distribution.

**Table 1.** Tail risk statistics of simulated aggregate loss distributions.

| Severity | | | Frequency | | $95^{th}$ Percentile | | $95^{th}$ ES | |
|---|---|---|---|---|---|---|---|---|
| Parameter | Mean | SD | Dist. | Mean | Mean | SD | Mean | SD |
| Log-Norm (7,0.1) | 1102.13 | 110.49 | PT(0,0.5,0.8) (NB) | 2 | 9182.15 | 203.95 | 13497.26 | 275.50 |
| | | | PT(0.5,0.75,0.89) (PIG) | | 8585.72 | 199.28 | 13701.27 | 346.85 |
| | | | PT(1,2,1) (Poisson) | | 5135.20 | 108.36 | 6024.44 | 54.89 |
| | | | PT(0,2.5,0.8) (NB) | 10 | 26002.69 | 281.51 | 32116.74 | 384.01 |
| | | | PT(0.5,3.75,0.89) (PIG) | | 26027.91 | 321.88 | 33443.47 | 476.69 |
| | | | PT(1,10,1) (Poisson) | | 17080.10 | 94.48 | 18841.43 | 106.02 |
| | | | PT(0,7.5,0.8) (NB) | 30 | 57726.73 | 399.58 | 66199.18 | 522.59 |
| | | | PT (0.5,11.25,0.89) (PIG) | | 58185.91 | 443.44 | 67971.08 | 611.78 |
| | | | PT (1,30,1) (Poisson) | | 43350.03 | 141.42 | 46177.74 | 170.28 |
| Log-Norm (8,0.2) | 3041.18 | 614.37 | PT (0,0.5,0.8) (NB) | 2 | 25499.46 | 502.28 | 37349.42 | 758.87 |
| | | | PT (0.5,0.75,0.89) (PIG) | | 23647.37 | 532.65 | 37897.37 | 962.10 |
| | | | PT (1,2,1) (Poisson) | | 14147.21 | 133.71 | 16801.75 | 163.15 |
| | | | PT (0,2.5,0.8) (NB) | 10 | 71887.10 | 767.55 | 88809.65 | 1065.32 |
| | | | PT (0.5,3.75,0.89) (PIG) | | 71949.59 | 885.98 | 92447.93 | 1322.43 |
| | | | PT (1,10,1) (Poisson) | | 47444.42 | 245.73 | 52389.94 | 296.37 |
| | | | PT (0,7.5,0.8) (NB) | 30 | 159505.57 | 1104.46 | 182957.54 | 1450.17 |
| | | | PT (0.5,11.25,0.89) (PIG) | | 160748.00 | 1221.88 | 187822.51 | 1688.13 |
| | | | PT (1,30,1) (Poisson) | | 120084.17 | 402.03 | 128041.67 | 481.29 |
| Log-Norm (9,0.3) | 8476.05 | 2601.12 | PT (0,0.5,0.8) (NB) | 2 | 71365.40 | 1420.97 | 104608.20 | 2125.78 |
| | | | PT (0.5,0.75,0.89) (PIG) | | 66203.55 | 1497.73 | 106066.50 | 2691.74 |
| | | | PT (1,2,1) (Poisson) | | 40168.11 | 380.81 | 47948.84 | 480.64 |
| | | | PT (0,2.5,0.8) (NB) | 10 | 200977.98 | 2158.04 | 248435.37 | 2992.20 |
| | | | PT (0.5,3.75,0.89) (PIG) | | 201114.26 | 2489.59 | 258465.81 | 3712.90 |
| | | | PT (1,10,1) (Poisson) | | 133609.13 | 713.47 | 147957.42 | 851.08 |
| | | | PT (0,7.5,0.8) (NB) | 30 | 445563.92 | 3137.25 | 511326.01 | 4080.62 |
| | | | PT (0.5,11.25,0.89) (PIG) | | 449001.52 | 3403.82 | 524762.05 | 4717.74 |
| | | | PT (1,30,1) (Poisson) | | 336930.07 | 1172.32 | 359855.59 | 1390.08 |

From a practical perspective, using the Poisson distribution to model aggregate claims process would create solvency and liquidity concerns and would result in sub-optimal purchasing of reinsurance. Finally, we note that as the level of frequency average increases, the relative difference between the percentile estimate of aggregate loss with different loss frequency distribution decreases.

Figure 1 plots the $50^{th}$, $80^{th}$ and $95^{th}$ percentiles for each of the 27 distributions defined in section 4.1.. The $y$-axis for each graph is the corresponding percentile of the aggregate loss distribution, and the $x$-axis labels the three different mean loss frequencies.

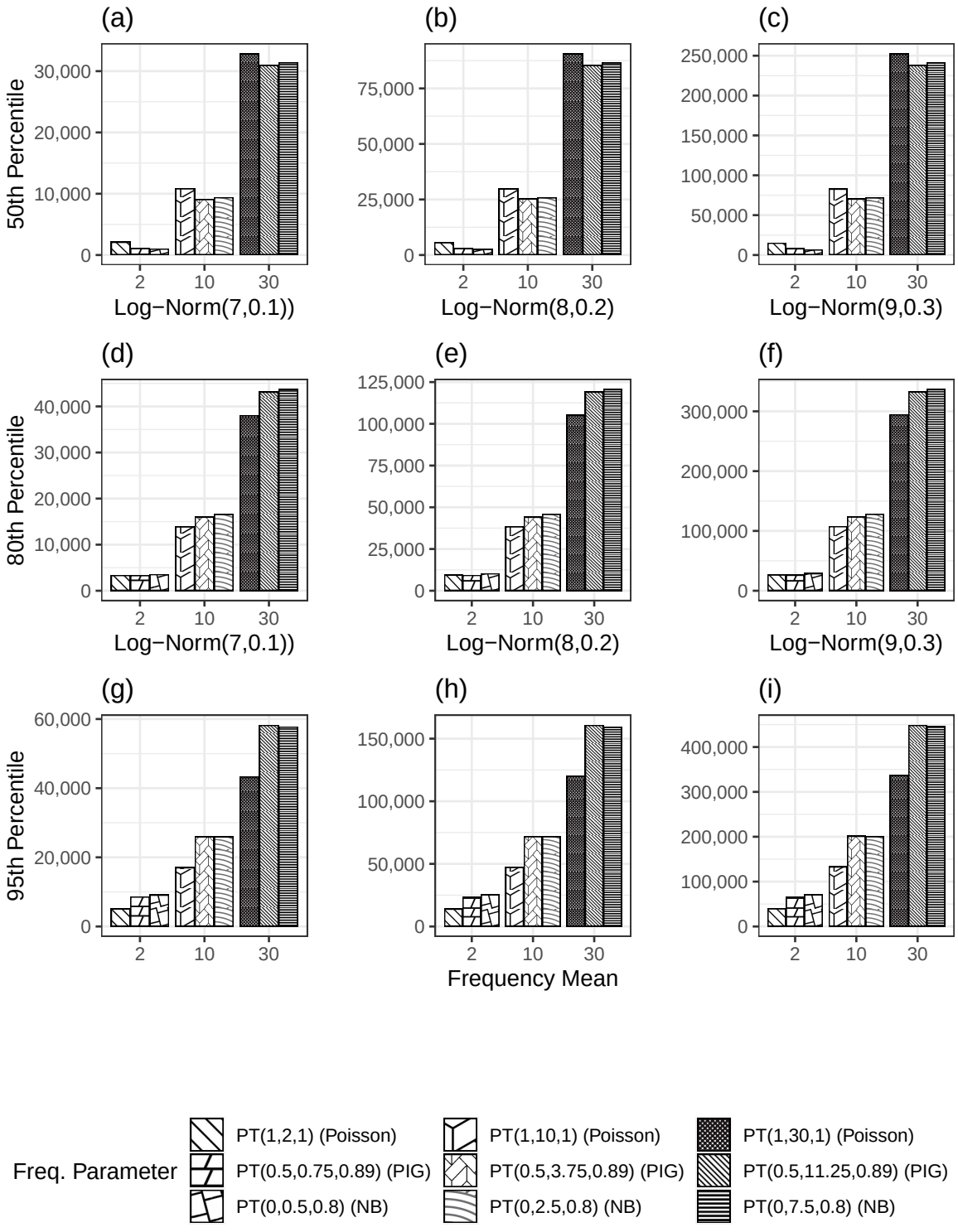


**Figure 1.** Aggregate loss percentiles (50%, 80%, 95%).

The first column for each severity/frequency combination is the aggregate loss percentile using the Poisson distribution, the second column is the Poisson Inverse-Gaussian distribution and the last column is the Negative Binomial distribution.

Of the three frequency distributions, the Poisson Inverse-Gaussian distribution has the fattest right tail, and the Poisson distribution the thinnest. The impact of the right tail is evident in these graphs: The aggregate loss distributions generated using the Poisson distribution have the largest medians, but the smallest $80^{th}$ and $95^{th}$percentiles for all distribution combinations. The aggregate loss distributions generated using the Negative Binomial distribution have the largest $80^{th}$ percentile, whereas the aggregate loss distributions generated using the Poisson Inverse-Gaussian distribution have the largest $95^{th}$ percentile. These graphs show the sensitivity of the aggregate loss distribution to the underlying choice of frequency distribution, and hence the importance of having greater flexibility in fitting real data to a frequency distribution.

## 5. Application

The United States government created the Transportation Security Administration (TSA) in response to the September 11, 2001 terrorists attacks. With the mission of protecting "the nation's transportation system to ensure freedom of movement for people and commerce" (https://www.tsa.gov/about/tsa-mission), TSA agents screen passengers and their luggage at ports of entry in the United States. At airports alone, on a daily basis, TSA agents screen more than 2 million passengers and almost 7 million pieces of luggage. It is inevitable that items could be damaged, lost, or stolen or individuals could be injured. As such, individuals who have suffered a loss may make a claim for losses to the TSA. The federal government reports information on every claim filed at https://www.dhs.gov/tsa-claims-data. We apply our proposed model to this data set to illustrate the use of Poisson-Tweedie as the loss frequency distribution.

TSA data for the years of 2002 to the end of 2015 was obtained from the Department of Homeland Security website (https://www.dhs.gov/tsa-claims-data). A detailed analysis of the data is provided by Kelly and Wang [13]. In total there are 286,952 observations from 2002 to 2015.

A summary of some of the data contained in the database is given in Table 2.

**Table 2.** TSA data variable description.

| Variable Name | Variable Description |
|---|---|
| Claim. Number | Unique claim identification number |
| Date. Received | The date that the claim is received by TSA |
| Claim. Type | The type of damage (e.g., complaints, property damage passenger injury or death) |
| Item. Category | The category of the damaged object (e.g., electronics, clothing, baggage, cameras, jewelry and watches) |
| Claim. Amount | The dollar amount requested for compensation in USD |
| Closed. Amount | The dollar amount paid in compensation in USD |
| Disposition | The status of claim defined as "approve in full", "deny", "settle" or if claim is still open the entry is blank. |

Each claim is uniquely identified by a claim number, and the database provides information on the date the claim is received, the type of loss, the items damaged, the claim amount paid by TSA, and the current disposition of the claim. The variables of interest for this analysis are the date received, the closed amount, and the disposition as defined in Table 2. Claims from 2002 to 2009 also contain information on the claim status (similar to the disposition status, but with information on why a claim was denied and reasons why a claim might still be open) and the claim amount, which is the initial amount requested by the claimant. Our analysis focuses on claims that have been approved in full or settled. We use the closed amount for loss severity. For claims reported between 2002 and 2009, if the closed amount is missing, then the claim amount, if available, is used. Open claims have a blank entry for the disposition of claims and are excluded from our analysis, as are observations with missing numbers. After applying these filters, we have 81,065 observations.

We aggregate claims over different time periods - daily, weekly, monthly, and quarterly. Time plots for the entire time frame display unexplained trends in frequency before 2008 and clear seasonality for the years 2013 to 2015. Because of this, our analysis is focused on the years 2008 to 2012. Our final data has 15,882 observations.

Figure 2 graphs the number of claims on a daily, weekly, monthly, and quarterly basis for the years 2008 to 2012. As can be seen, these 5 years provide a relatively stable period in terms of frequency, however there is a slight dip in frequency between the end of 2009 and early 2010*.

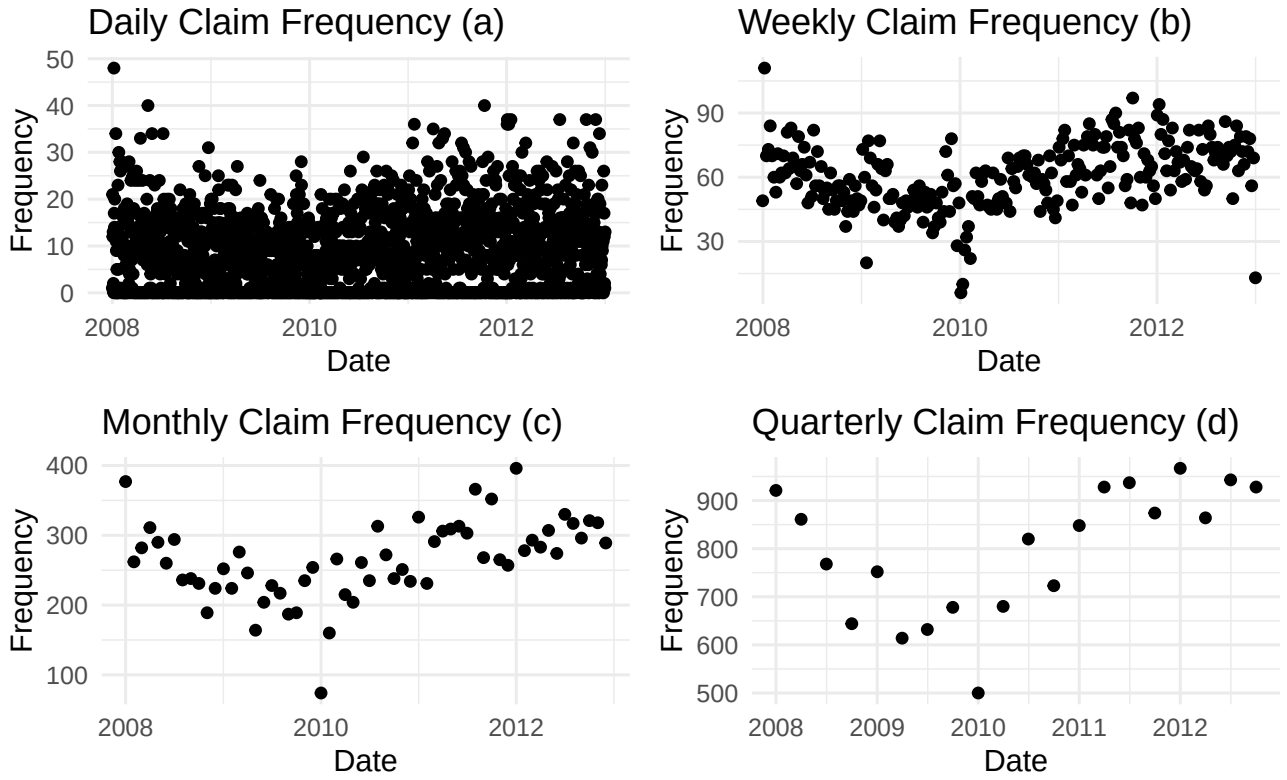


**Figure 2.** Periodic TSA claim frequency scatter plot.

*Although we initially suspected the dip in loss frequency in 2009 was related to the lagged effect of the 2008 financial crisis which negatively affected the US economy, Kelly and Wang [13] show a decline in the number of claims per 100,000 passengers over this time period.

The summary statistics for the number of claims, aggregated on a daily, weekly, monthly, and quarterly basis are given in Table 3.

**Table 3.** Summary Statistics of TSA Claims Frequency (Number of Claims)

| | Min. | 25th Percentile | Median | Mean | 75th Percentile | Max. | SD | Variance |
|---|---|---|---|---|---|---|---|---|
| Daily | 0 | 0.00 | 8.0 | 8.69 | 14.00 | 48 | 8.11 | 65.70 |
| Weekly | 6 | 50.00 | 61.0 | 60.62 | 71.00 | 111 | 15.08 | 227.36 |
| Monthly | 74 | 233.25 | 265.5 | 264.70 | 303.75 | 396 | 55.81 | 3114.79 |
| Quarterly | 500 | 679.50 | 834.0 | 794.10 | 922.75 | 967 | 134.85 | 18,184.94 |

Corresponding histograms for the number of claims over the different aggregation periods are given in Figure 3.

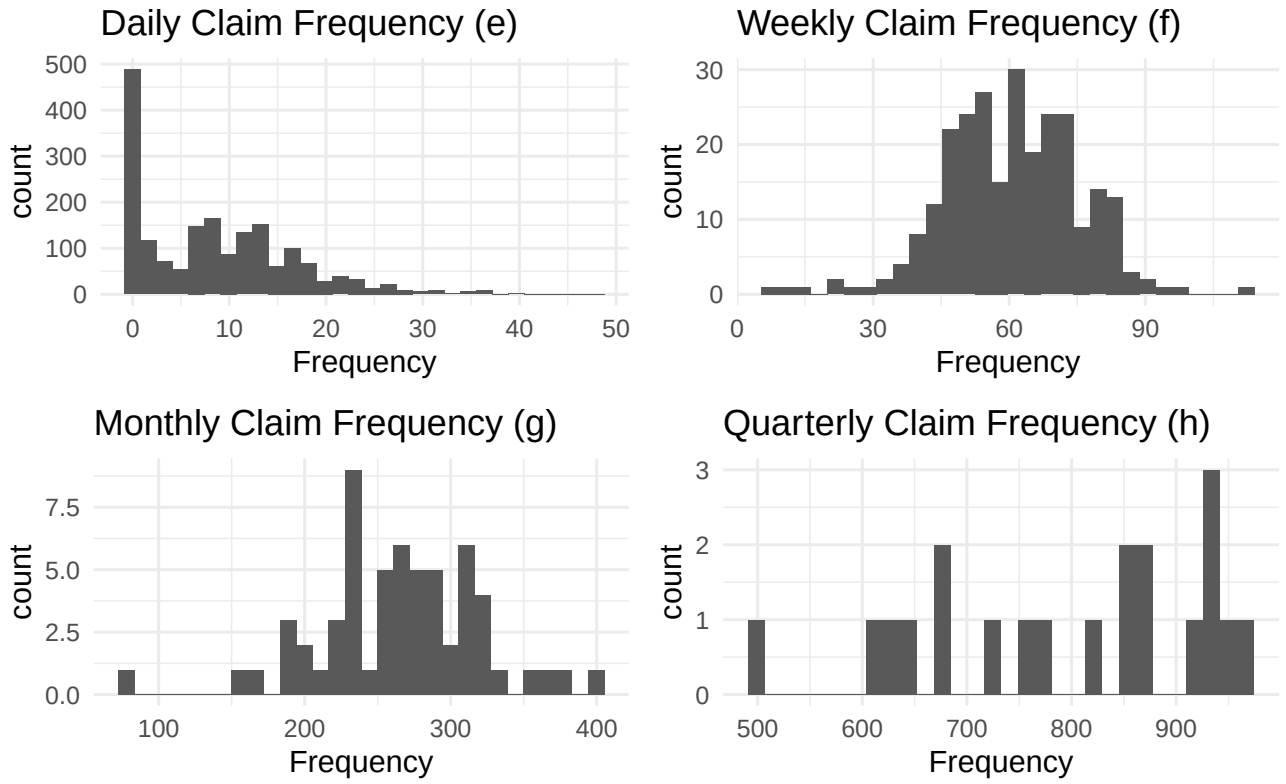


**Figure 3.** Periodic TSA claim frequency histogram.

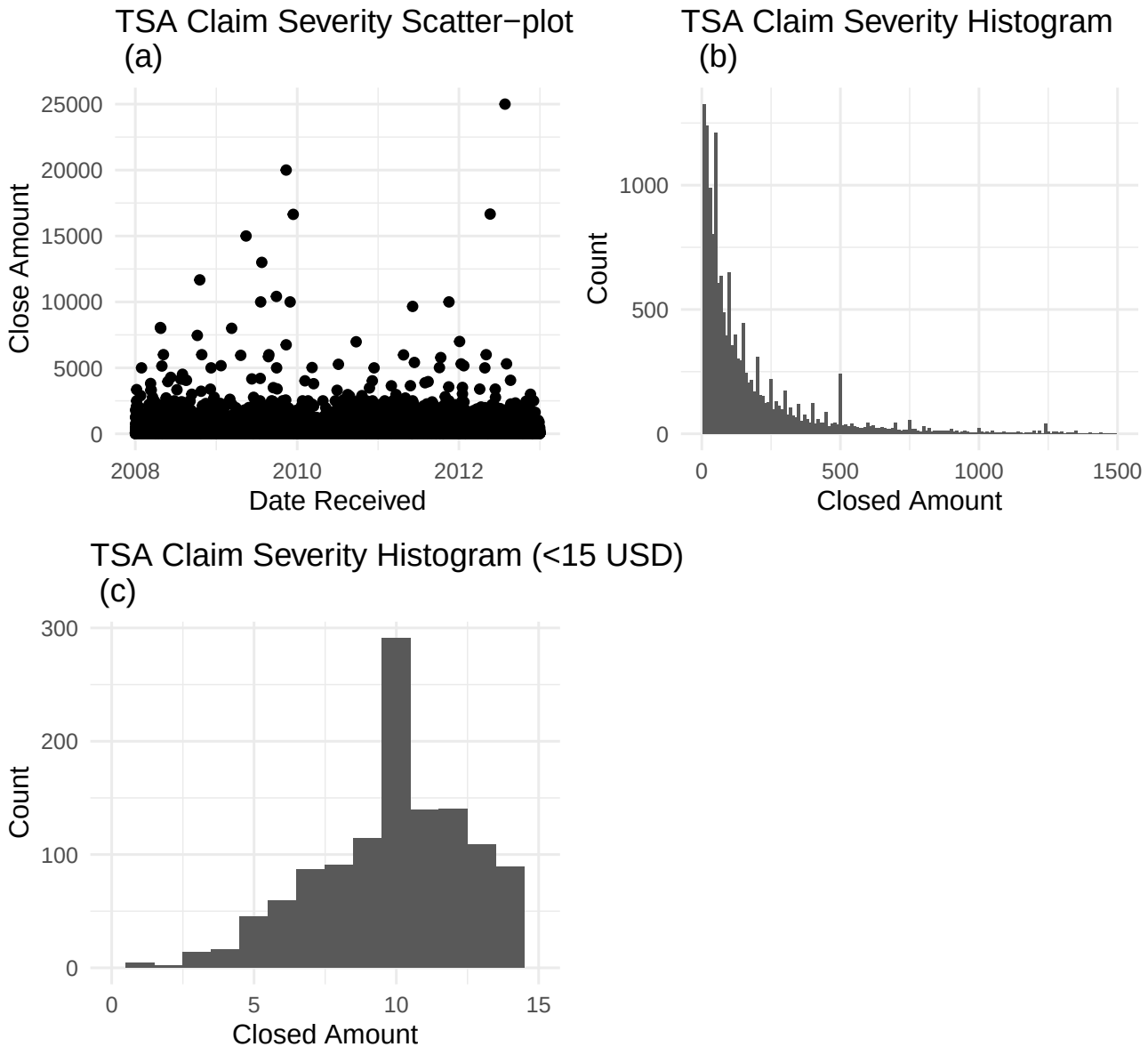


**Figure 4.** TSA claim severity.

The scatter plot for the closed claim amount in Figure 4(a) shows that the claim severity has not changed between 2008 and 2012. Figure 4(b) shows that the severity is positively skewed, and the data are not truly continuous: We see sharp spikes at every 50 USD increment amounts with a very significant spike at the 500 USD amount. The database shows that at the 500 USD amount, the losses are from damaged or lost personal electronic devices and pieces of jewelry. At the 100 USD amount, the common lost or damaged items are luggage, cosmetics, and clothes. We suspect these spikes are related to the claims settlement process. Figure 4(c) shows the overwhelming number of small claims filed against the TSA.

Next, we fit the TSA claim data to the aggregate loss model as defined in (1.1). We first fit the frequency distribution and then severity distribution. We then generate percentiles of the aggregate loss distribution.

*5.1. Analysis of frequency*

We focus on monthly data since daily data are highly zero-inflated (a very large number of days with no claims as shown in Figure 3(e)), weekly data have high auto-correlation at lag 4 (monthly correlation) and quarterly data lack a sufficient number of observations. We use a Poisson-Tweedie distribution $PT(a, b, c)$ to model the number of claims for monthly data.

We estimate parameters using the Maximum Likelihood Estimation method. The fitting is performed using R generic function optim() with the default method "Nelder-Mead" (other optional methods may be feasible). Parameters $b$ and $c$ can be derived from the parameter $a$, and the fitted mean and variance of the loss frequency. Thus, we are most interested in the family index, parameter $a$ in $PT(a, b, c)$ and the estimated mean and variance.

The maximum likelihood estimates for frequency mean, variance and the 95% VaR for the frequency are reported in Table 4. The 95% VaR for the frequency implies that 95% of the time there will be less than 366 claims per month. Knowing the maximum number of claims that could reasonably be expected may help management in workforce planning to process the claims.

**Table 4.** MLE of moments and quantile of TSA claims frequency.

| | MLE Mean | MLE Variance | 95% $VaR(\hat{N})$ |
|---|---|---|---|
| Monthly | 264.21 | 3426.18 | 366 |

The maximum likelihood estimate of parameter $a$ and the 95% confidence interval for parameter $a$, are given in Table 5.

**Table 5.** Estimation results for family index $a$.

| | $\hat{a}$ | 95% C.I. for $a$ |
|---|---|---|
| Estimates | $-1.14$ | $(-2.26, -0.03)$ |

Within the Poisson-Tweedie family, the family parameter "a" for Poisson is 1, 0.5 for PIG, and 0 for Negative Binomial. Since the 95% confidence interval falls into the negative side of real values, Poisson and PIG are clearly rejected. Although, 0 is also excluded from the 95% confidence interval, the upper bound is quite close to 0. We conduct additional likelihood ratio (LR) test for

$$H_0\text{: Negative-Binomial vs. } H_1 : \text{ Poisson-Tweedie}$$

The value of $-2(\text{LogLik}_{NB} - \text{LogLik}_{PT}) = 3.146153$, and the corresponding p-value $= 0.0761 < 10\%$. Thus, the Negative Binomial model is rejected at 10% significance level. This additional test confirms the result of the confidence interval test which excludes the negative binomial distribution. Additionally, the LR tests for PIG and Poisson are also performed. The p-values of all three LR tests are included in Table 6.

Table 6 and Figure 5 provide further evidence of the superiority of the Poisson-Tweedie distribution over the Poisson, Negative Binomial and Poisson Inverse-Gaussian distributions in fitting this data set.

**Table 6.** Goodness-of-fit of monthly distribution fit.

| | Poisson-Tweedie | Negative Binomial | Poisson Inverse-Gaussian | Poisson |
|---|---|---|---|---|
| Negative Log-Likelihood | 328.87 | 330.44 | 334.24 | 595.25 |
| AIC | 663.74 | 664.88 | 672.49 | 1192.50 |
| BIC | 670.02 | 669.07 | 676.68 | 1194.59 |
| LR Test (p-value) | — | 0.0761 | 0.0010 | 0 |

From Table 6, we find that fitting frequency data with Poisson-Tweedie distribution results in the smallest negative log-likelihood for monthly data and the smallest Akaike Information Criterion (AIC). When using the Bayesian Information Criterion (BIC), which has a larger penalty on the number of parameters than the AIC, the BIC for Poisson-Tweedie is very close to the Negative Binomial distribution, which has the lowest BIC for this data. Note that the AIC and BIC are meaningful when sample size goes to infinity. In this data set, the sample size is 60, we regard the AIC and BIC as directional evidence for model selection. Thus, Table 6 shows that Poisson-Tweedie remains competitive against common frequency distributions while having the benefit of model flexibility.

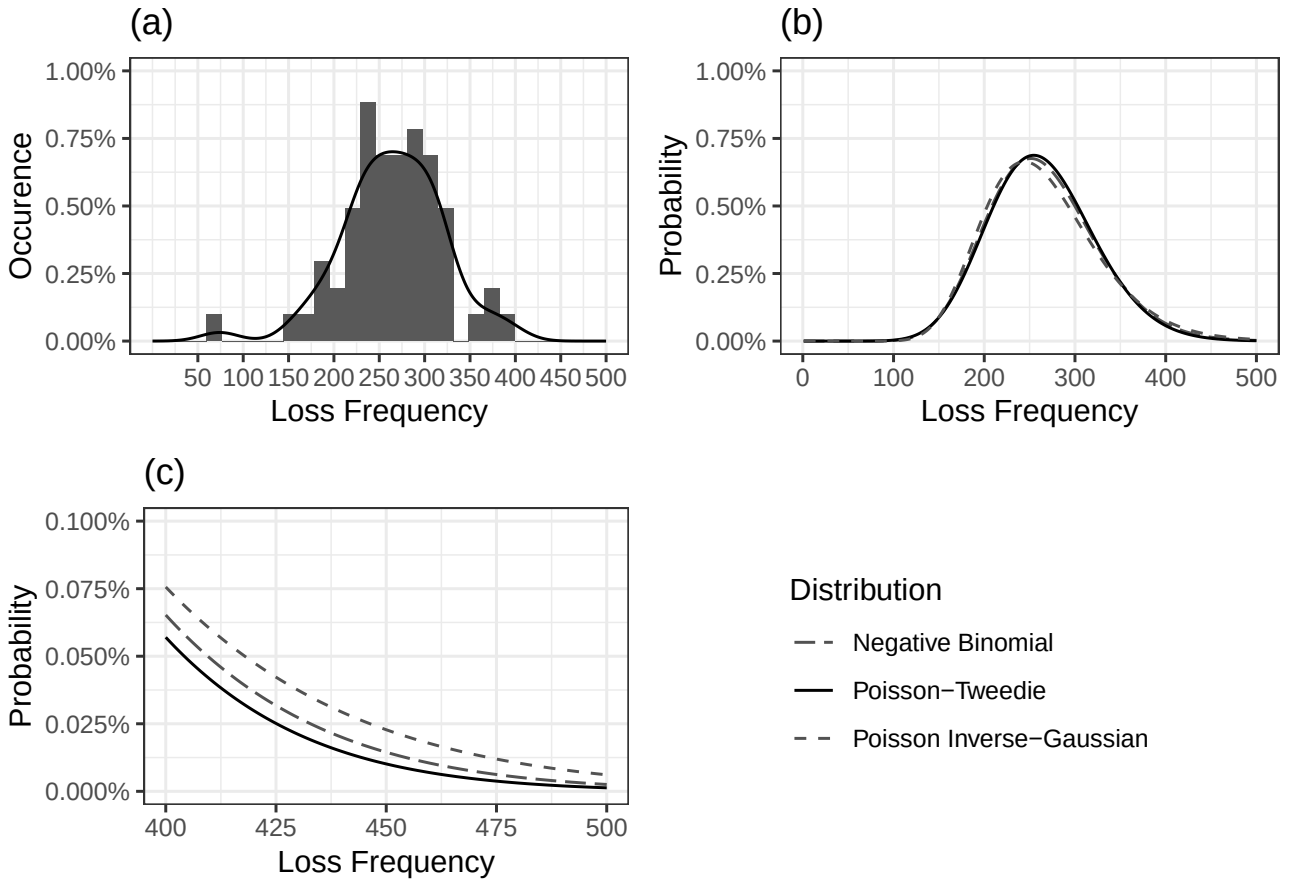


**Figure 5.** Comparison of estimated monthly frequency with different distributions.

Figure 5 gives a graphical We observe that the histogram of the sample monthly loss frequency in Figure 5(a) seems to be symmetric. The sample dispersion (sample variance over the sample mean) is

11.76, indicating that Poisson distribution is not a good fit. From Figure 5(b),(c), we observe that the fitted Negative Binomial and Poisson Inverse-Gaussian are slightly more right-skewed than Poisson-Tweedie.

### *5.2. Analysis of severity*

For severity, Figure 4(b) shows that the severity is positively skewed, and we fit three distributions to the severity data. Our first distribution is the Log-Normal distribution as defined below.

$$f_X(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)$$

for $\mu \in (-\infty, +\infty)$ and $\sigma > 0$.

We also consider the Lomax distribution which is a special case of the Pareto Type II distribution with density

$$f_X(x;\alpha,\lambda) = \frac{\alpha\lambda^\alpha}{(x+\lambda)^{\alpha+1}}$$

for $x \geq 0$, $\alpha > 0$ and $\lambda > 0$.

Additionally, we look at the Gamma distribution defined as

$$f_X(x;k,\theta) = \frac{x^{k-1}e^{-x/\theta}}{\theta^\alpha \Gamma(\alpha)}$$

for $x > 0$, and $a, \theta > 0$.

We fit the data using maximum likelihood estimation and the estimated mean and standard deviation of severity and the goodness-of-fit tests are shown in Table 7. Based on our goodness-of-fit tests, we select the log-normal distribution for use in the calculation of the fitted aggregate loss distribution as it has the lowest AIC, which implies it has the best performance in terms of the TSA data set.

**Table 7.** MLE estimates of severity mean and standard deviation and goodness-of-fit.

| Loss Severity Distribution | MLE Mean | MLE STD | AIC |
|---|---|---|---|
| Log-Normal (4.59,1.31) | 233.83 | 2105.2 | 199579.22 |
| Lomax (2.01,247.35) | 245.64 | 4173.1 | 200517.62 |
| Gamma (0.38,1471.65) | 561.13 | 908.73 | 210687.69 |

### *5.3. Simulation of fitted aggregate loss distribution*

The aggregate loss distribution of monthly data is estimated using the fitted monthly distributions of Poisson-Tweedie and Log-Normal. Using our fitted parameters from sections 5.1 and 5.2, We undertake Monte-Carlo simulation to create 100,000 months. The historical aggregate loss distribution is shown in Figure 6(a), and our simulated distribution is shown in Figure 6(b).

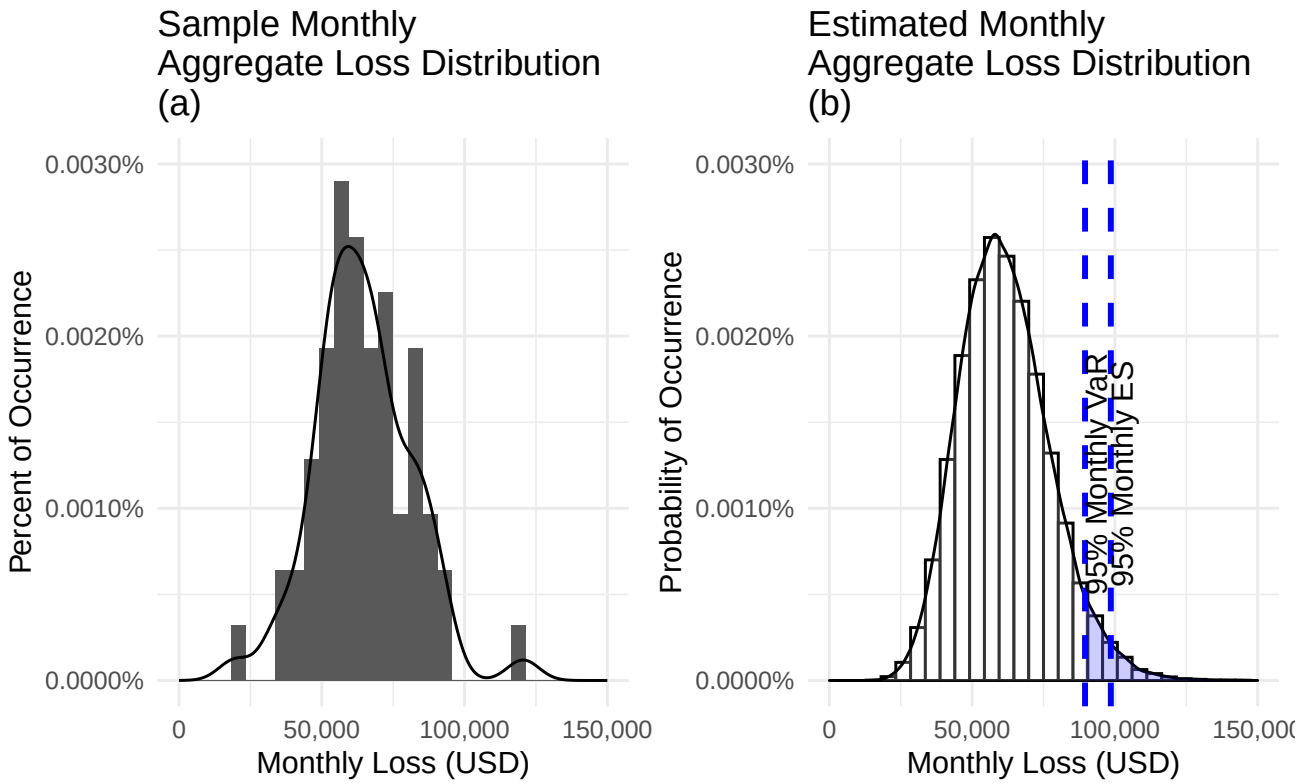


**Figure 6.** Actual vs simulated monthly aggregate loss of TSA claims (aggregated by month).

We find that the estimated monthly aggregate loss distribution has a mean of 61,616.43 and standard error of 15,864.73 using our estimated parameters with the simulation approach. From our model simulation, the estimated $95^{th}$ percentile monthly Value at Risk (VaR) is 89,533.42 USD and the estimated $95^{th}$ percentile monthly Expected Shortfall (ES) is 98,570.69 USD. The fitted Kernel density in Figure 6(a) illustrates the overall shape and skewness of the sample data; the fitted aggregate loss distribution also has similar overall shape and skewness pattern (Figure 6(b)). However, we acknowledge that kernel estimate at tail is not stable, particularly when sample size is small in this case. This is a well-known problem in kernel estimation (Wand and Jones [16]).

## 6. Conclusions

In this paper, we introduce the Poisson-Tweedie distribution as a candidate to model loss frequency. This family of distributions is mathematically tractable and the the ability to fit the family parameter, $a$, increases the fitting flexibility and reduces the possibility of misspecifying the frequency distribution. Assuming the independence of the frequency and severity distributions, the maximum likelihood estimation of the moments of the aggregate loss distribution is straight forward.

We simulate the aggregate distribution of losses where severity is modelled using a Log-Normal distribution and use different distributions within the Poisson-Tweedie family of distributions for the frequency. We find that the choice of the family parameter, $a$, influences the fatness of the right tail of the aggregate loss distribution and therefore affects the Value at Risk and the Expected Shortfall measures. However as the mean frequency increases, the relative difference between these percentile

estimates decreases. Future research should examine the impact of increasing loss frequencies on percentile estimates of the aggregate loss.

We examine the flexibility of the Poisson-Tweedie distribution by fitting the distribution to the number of monthly claims from the Transportation Security Administration's (TSA's) claims database for 2008 to 2012. The fitted family parameter, $\hat{a}$, is significantly different from the family parameters that define the Poisson, Negative Binomial and Poisson Inverse-Gaussian distributions. We show that Poisson-Tweedie distribution performs as well as or better than these three distributions as the goodness-of-fit score using the Poisson-Tweedie distribution is better than or almost the same as the scores from the other three distributions.

To build the aggregate loss model, we fit the the Log-Normal distribution to model TSA claim severity, and then apply simulation to derive the percentiles of the fitted aggregate loss model. Thus, we show that the aggregate loss model with the Poisson-Tweedie frequency family can be applied to real-world data to estimate the aggregate percentile statistics of interest. The resulting aggregate loss percentile estimates based on our proposed model are also similar to the kernel density estimates which supports the validity of our model. In our analysis, we examined only those claims that have closed with a payment. Further research could fit the Poisson-Tweedie distribution to the denied claims in the TSA database since these were excluded from our analysis.

Because the Poisson-Tweedie distribution provides a better fit than commonly used frequency distributions, incorporating the Poisson-Tweedie distribution into aggregate loss modelling could assist insurers in optimizing their reinsurance purchases and improving the metrics needed to effectively manage both solvency and liquidity risk.

## Acknowledgments

Zilin Wang is supported by an NSERC grant. Si Chen is partially supported by CANSSI. The authors would like to thank Dezhao Han and Hongcan Lin for their assistance with debugging and optimizing R and C++ code. We are grateful to the referees for their valuable comments and additional references.

## Conflict of interest

All authors declare no conflicts of interest in this paper.

## References

1. Lo A, Remorov A, (2017) Stop-loss strategies with serial correlation, regime switching, and transaction costs, *J Financ Mark* 34: 1–15.
2. Tan K S, Weng C, Zhang Y, (2009) Var and cte criteria for optimal quota-share and stop-loss reinsurance, *North Am Actuarial J* 13: 459–482.
3. Griffiths R, Mnif W, (2017) Various approximations of the total aggregate loss quantile function with application to operational risk, *J Oper Risk* 12: 23–46.
4. Karam E, Planchet F, (2012) Operational risks in financial sectors, *Adv Decis Sci* 2012: 1–57.

5. Panjer H H, (2006) *Operational Risk: Modeling Analytics*, Wiley Series in Probability and Statistics, Wiley.
6. Cummins J D, Dionne G, Mcdonald J B, et al. (1990) Application of gb2 family of distribution in modeling insurance losses processes, *Insur Math Econ* 9: 257–272.
7. Jin T, Provost S B, J Ren, (2016) Moment-based density approximations for aggregate losses, *Scand Actuarial J* 2016: 216–245.
8. Shevchenko P V, (2011) *Modelling Operational Risk Using Bayesian Inference*, Springer Berlin Heidelberg.
9. Bae T, Ko B, (2020) On the mixtures of length-biased weibull distributions for loss severity modeling, *J Korean Stat Soc* 49: 422–438.
10. Willmot G E, Lin X, (2011) Risk modelling with the mixed erlang distribution, *Appl Stochastic Models Bus Ind* 27: 2–16.
11. Esnaola M, Puig P, Gonzalez D, et al. (2013) A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated rna-seq experiments, *BMC Bioinf* 14: 254.
12. El-Shaarawi A H, Zhu R, Joe H, (2011) Modelling species abundance using the poisson–tweedie family, *Environmetrics* 22: 152–164.
13. Kelly M, Wang Z, (2020) A data set for modeling claims processes—tsa claims data, *Risk Manage Insur Rev* 23: 269–276.
14. Jorion P, (2000) *Value at Risk: The New Benchmark for Managing Financial Risk*, McGraw-Hill.
15. Embrechts P, Frei M, (2009) Panjer recursion versus fft for compound distributions, *Math Methods Oper Res* 69: 497–508.
16. Wand M P, Jones M C, (1995) *Kernel Smoothing*, Monographs on Statistics and Applied Probability, Springer US.


AIMS Press