Scientific Research Publishing

Open Journal of Statistics, 2023, 13, 425-452
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Modeling Cyber Loss Severity Using a Spliced Regression Distribution with Mixture Components

Meng Sun

Simon Fraser University, Burnaby, Canada
Email: mengsun0205@hotmail.com

How to cite this paper: Sun, M. (2023) Modeling Cyber Loss Severity Using a Spliced Regression Distribution with Mixture Components. *Open Journal of Statistics*, **13**, 425-452.
https://doi.org/10.4236/ojs.2023.134021

Received: June 6, 2023
Accepted: July 8, 2023
Published: July 11, 2023

Abstract

Cyber losses in terms of number of records breached under cyber incidents commonly feature a significant portion of zeros, specific characteristics of mid-range losses and large losses, which make it hard to model the whole range of the losses using a standard loss distribution. We tackle this modeling problem by proposing a three-component spliced regression model that can simultaneously model zeros, moderate and large losses and consider heterogeneous effects in mixture components. To apply our proposed model to Privacy Right Clearinghouse (PRC) data breach chronology, we segment geographical groups using unsupervised cluster analysis, and utilize a covariate-dependent probability to model zero losses, finite mixture distributions for moderate body and an extreme value distribution for large losses capturing the heavy-tailed nature of the loss data. Parameters and coefficients are estimated using the Expectation-Maximization (EM) algorithm. Combining with our frequency model (generalized linear mixed model) for data breaches, aggregate loss distributions are investigated and applications on cyber insurance pricing and risk management are discussed.

Keywords

Cyber Risk, Data Breach, Spliced Regression Model, Finite Mixture Distribution, Cluster Analysis, Expectation-Maximization Algorithm, Extreme Value Theory

1. Introduction

According to Gartner [1], the global market for information security would reach $170.4 billion in 2022. Given the potential economic impact of a successful large-scale cyber attack, cybersecurity risks remain the second-most important

emerging issue risk experts highlight [2]. Beyond traditional technologies, a larger set of risks at the intersection of technology and society is rapidly emerging. AI and big data have led emerging technologies, transforming economic and social structures. The full-scale implications of cyber threats are yet to be experienced, especially since technology is rapidly evolving. COVID-19 has compelled businesses to establish remote workforces and utilize cloud-based platforms. Due to the pandemic, remote work and digital transformation increased the average total cost of a data breach. The FBI reports a 300% increase in reported cybercrimes since the pandemic began. According to IBM [3], data breach costs increased from $3.86 million to $4.24 million in 2021, the highest average total cost in the report's history. Attackers stole $121,000 in bitcoin through nearly 300 transactions due to a Twitter breach that affected 130 accounts [4] resulting in attackers swindling. A security breach disclosed by Marriott compromised the data of more than 5.2 million hotel guests [5]. The Equifax Data Breach, which cost over $4 billion in total, affected 147.9 million consumers [6]. The increasing number of large-scale, widely publicized security breaches suggests that the number of security breaches and their severities is increasing. In 2021, the average cost of a breach will increase by 10%, necessitating quantifying its effects. Regulations and best practices in cyber security hygiene and risk management are changing due to the frequency and severity of cyberattacks. Prominent legislations like the European Union's 2018 General Data Protection Regulation (GDPR)[1], California's 2020 California Consumer Privacy Act (CCPA)[2] and Illinois's 2018 Biometric Information Privacy Act (BIPA)[3] have been passed to enforce severe consequences. To collect, store, process and transfer consumer data, these regulations all have one thing in common: they require organizations to adhere to specific standards.

In addition to reducing vulnerable exposure and increasing technology defence investment, cyber insurance is a fundamental and wildly applicable tool for organizations to maintain their enterprise solvency in light of the rise in cybersecurity threats. Cyber insurance is a type of insurance intended to product against the financial costs associated with the failure or compromise of an organization's information system [7]. Cyber events include a hacking attack by an external party or malware infection, fraud involving debit and credit cards, and the unintentional disclosure of electronic records due to human error. Most cyber insurance providers offer a core set of coverages and various supplement coverages. With the cyber risk insurance market is at an inflection point, it provides an opportunity to embrace a paradigm shift. To safeguard its profitability, the cyber insurance market took four deliberate measures to combat rising loss ratios [8]: cyber premiums increased across the board, regardless of the industry sector or organization size; many carriers imposed sub-limits and coinsurance provisions specific to ransomware claims; carriers wanted to limit their exposure

[1]General Data Protection Regulation https://gdpr.eu/.
[2]California Consumer Privacy Act https://oag.ca.gov/privacy/ccpa.
[3]Illinois Biometric Information Privacy Act https://www.ilga.gov/legislation/.

by limiting capacity; and almost all carries requested more information regarding data security control efforts. An insurance company cannot only rely solely on standard actuarial methods when modeling cyber risk and pricing cyber insurance products. Given how difficult it is to quantify this emerging and evolving risk, new methods and techniques must be developed. In order to identify risk characteristics and assess cyber risk severity, this paper takes an innovated rather than an empirical approach.

A growing number of disciplines are researching Cyber risk. However, acturial risk management is hindered by the need for mature predictive analysis approaches for quantifying and predicting risk severity. We review quantitative research works in actuarial science and describe several research works that focus on loss severity modeling and predictive analysis. [9] combines regression models based on the class of Generalized Additive Models for Location Shape and Scale (GLMLSS), which permits parameters in both the severity and frequency distributions, and a class of ordinal regressions. [10] model hacking data breaches frequency using a hurdle Poisson model and severity using a non-parametric generalized Pareto distribution (GPD). [11] particularly focus mainly on severe claims by combining a generalized Pareto modelling and a regression tree approach for severity analysis. Most of these methods pay special attention to large claims with heavy tail distributions. Traditional actuarial modelling techniques for heavy-tailed insurance loss data concentrate on simple models based on a single parametric distribution that adapts the tail well, such as generalized linear models (GLMs), regression models and quantile regression [12]. Because these techniques are based on a single distribution, they may not be applicable when the behaviour of the tail is inconsistent with the behaviour of the entire loss distribution. It is well known that the actuarial loss distribution is strongly skewed with heavy tails and consists of small, medium and large claims that are difficult to fit with a single parametric distribution. The Extreme Value Theory (EVT) approach, which employs GPD to model excesses over a high threshold ([13] and [14]), gained popularity when dealing with heavy-tailed and high losses data. However, they cannot capture the characteristics across the entire loss distribution range making them unsuitable for use as a global fit distribution [15]. In order to model the complete loss distribution, it is frequently necessary to obtain a global fit for the distribution of losses by splicing [16] several distributions in order to model the complete loss distribution. Several actuarial works proposed splicing models for the application of risk measures. For financial risk analysis, [17] suggest a splicing model with a mixed Erlang (ME) distribution for the body and a Pareto distribution for the tail. [18] suggests a three-component spliced regression model for fitting insurance loss data and demonstrate that spliced results outperform the Tweedie loss model regarding tail fitting and prediction accuracy.

The risk portfolio typically contains unobserved heterogeneity in terms of claim severity, such as workers' compensation and cyber risk data. Given this re-

ality, researchers typically employ a mixture approach to capture the multimodality of the observed loss distribution. [19] designs an optimal Bonus-Malus system in automobile insurance using finite mixture models. [20] models an actual motor insurance claim data set using a mixture Lognormal distribution. [21] proposes a finite mixture of skew-normal distributions that better describes insurance data. [22] suggests a different method for modelling mixture data with heavy tails and skewness in insurance loss distribution that exhibit multimodality. [23] proposes an Erlang loss model using a generalized expectation-maximization (GEM) and clustered method of moments (CMM) algorithm to fit insurance loss data and calculate quantities of interest for insurance risk mixture management. Finally, [24] propose a class of logit-weighted reduced mixtures of experts (LRMoE) models for multivariate claim frequencies or severity distributions and perform the estimation and application to correlated claim frequencies [25].

Upon the above literature review, all the studies on cyber loss severity do not consider excess of zeros loss, spliced composites and mixture models under a global distribution with corresponding sets of covariates. Motivated by cyber risk specific nature, our study aims to fill these gaps using a finite mixture model (FMM) under a non-linear regression framework and a three-component splicing model with a zero-inflated component. We provide four significant contributions overall.

First, we adapt the methodology by providing a flexible mixture distribution model. Instead of modelling the univariate distributions of all mixture components within the same parametric distribution family, we propose a different method that combines different types of distributions such as Gamma, Log-Normal, Weibull, Burr, Inverse Gaussian and Pareto within a single FMM frame. The aim is to determine the degree to which wide cyber loss range of severity distribution and the exhibition of heavy-tailed and skewed. The model's compositional distributions make it a valuable, understandable modelling tool for various risks with heterogeneous performance.

Second, in addition to developing components from parametric non-Gaussian families of distributions, we incorporate FMM into a generalized linear model (GLM) to fully utilize the risk characteristics by treating them as covariates within the regression framework. Traditional actuarial loss distribution analysis examines the loss distribution but rarely considers the impact of individual risk characteristics. Moreover, a generalized linear regression mixture model rarely relates a dependent variable to a set of explanatory variables. Our model divides the unobserved mixture probabilities of observations into subgroups and simultaneously estimates a GLM model for each subgroup.

Third, we built an FMM model together with a zero-inflated regression component for our continuous data type. This adaptable strategy enables using covariates to model both the non-zero mixture distribution and the rate of point mass zero. We demonstrate the adaptability of our method by applying conti-

nuous mixture model components to the zero-inflation component, in contrast to zero-inflated discrete models such as Poisson and Hurdle. Then we create a zero-inflated mixture model with a complete cumulative distribution function by adjusting the mixture proportions following logistic odds.

Last but not least, our work enables cyber risks to be completely quantified under one distribution taking into consider its extreme loss heavy tail. Cyber risk loss exposures permeate every facet of an organization's operations, making the consequences of a data breach potentially catastrophic. Unlike other kinds of property and casualty risk that capping incurred loss at a 95% level could effectively rule out extreme value, cyber risk has the nature that, even upon logarithm, loss distribution is too heavily skewed to be capped at a bell shape distribution. Traditional insurance pricing set up a policy limit and doesn't consider extreme loss when training the model. However, this technique can't be applied to cyber risk as there is no such limit can be set so that cyber loss could be modeled via one single distribution. We brought up a more statistically rigorous attempt to incorporate excess zero, mixture components and heavy tail of cyber risk in a single, statistically consistent step where other estimation process, such as covariates dependence, is also going on.

This paper presents a series of finite mixture regression (FMR) models and discusses their application in cyber risk estimation. In Section 2, we introduce our PRC dataset and conduct cluster analysis on geographical information. Section 3 reviews the definition and composition of FMR models and propose our unique FMR model adjusted by zero-inflated component based on dataset. Next, we introduce the expectation-maximization (EM) algorithm used to estimate coefficients and model parameters in Section 4. Followed by details on how to fit and choose from among these models as well as information about how to assess the goodness of fit of a model in Section 5. Then, we combine the proposed severity model together with our previous frequency model to simulate aggregate cyber loss over a future time interval in Section 6. Finally, we discuss a model application from the insurers' perspective and suggest rate filing and future discussions in Section 7.

2. Chronology of Data Breaches from PRC Dataset

In this section, we perform an empirical data analysis which supports and motivates our data-driven modeling approach and further analysis and application. Several necessary initialization procedures must be investigated. Starting with the explanatory data analysis, we investigate unique features of the dataset through an empirical data analysis in the Section 2.1, followed by a cluster analysis to study a multidimensional location feature of the dataset in Section 2.2.

2.1. Empirical Data Analysis

The dataset serves as a resource for researchers examining the effect of data breaches on the performance of insurance companies. It encourages research on

the loss prediction and premium determination. Our previous work discussed several frequently considered databases from nonprofit corporations. It used a generalized linear mixed model (GLMM) [26] to examine a quarterly frequency modelling approach.

In this paper, we consider a publicly available cyber security dataset, Privacy Rights Clearinghouse (PRC)[4] dataset. Most of the PRC data comes from state attorneys general and the U.S. Department of Health and Human Services. This dataset contains the data breach incidents as well as the number of records breached due to breach incidents. As a sample of the choronology shown in **Figure 1**, the type of cyber event and the victim's information, such as the company's name, type of business, and location, were gathered from breach incidents. Because it contains risk-related characteristics that can be utilized as rating factors, this information is essential for filing insurance rates. In [26], a generalized linear mixed model (GLMM) is proposed to study the quarterly frequency (number of incidents) of the data breaches recorded in this same PRC database and its application to the cyber insurance is discussed. In this study, we are interested in the number of records breached affected by each recorded data breach incident, which is considered as the severity of the breach caused by cyber breach incidents and collected in PRC database. We late convert the breached data record to dollar amount loss in order to get a dollar amount magnitude.

We illustrate in Section 6 an application of our proposed severity methodology in examining aggregate cyber losses by combining the frequency modeling approach proposed in [26] based on the same dataset.

As discussed in the previous section, fitting an adequate loss distribution to the cyber breach data set is difficult due to its nature. Here, we conduct an empirical data analysis of related target and explanatory variables on the PRC data set to demonstrate the necessity of addressing/accounting for several risk features. Our work is based on the most recent download of the PRC data breach chronology, including 9012 data breach incidents observed in the United States. After removing incomplete and inconsistent observations, 8095 incidents from 2001 to 2019 are investigated and modeled. A summary statistics of this data set

Table 1. Sample of PRC Chronology.

Incident Date	Type of Breach	Type of Business	Location	Loss of Records
2018/02/03	CARD	BSF	California	30
2018/05/26	HACK	GOV	Washington	1000
2018/06/30	DISC	MED	Massachusetts	900
2018/09/27	PHYS	EDU	Florida	1500
2018/10/09	INSD	BSR	Texas	700
2018/12/05	PORT	NGO	Ohio	150

[4]Privacy Rights Clearinghouse (PRC) database is available for public download at https://privacyrights.org/data-breaches.

is provided in Table 1.

The first row of Table 2 provides summary statistics for the target variable, the "number of records" breached for a total number of 8095 data breach incidents, rounded to the nearest 100 units, where q_a denotes the empirical α-quantile. We observe from these summary statistics that the number of records has a 32.9% excess of zeros and a very heavy right tail, given that the sample mean is significantly larger than the sample median. This can be revealed by the fact that some types of loss such as competitive advantage and reputation damage can not be measured in digital record units. The breach incident occurred without any reported loss or expenses if there were no records lost in that incident. Loss records range from 0 to 500 billion which is difficult to quantify under one distribution. In this regard, the remainder of the analysis in this paper is based on the logarithm of severity in order to maintain complete low and high loss amounts information.

The PRC data set contains three explanatory variables that can be used as regressors: breach type, organization type, and company location. The first two variables are documented to have 7 subcategories each, while the location is listed in 50 geographical states. We modify on their levels, based on their nature and characteristics to reduce factor dimensions and increase predictive power. Table 3 summarizes the combined model inputs of business and breach types. The level combination of the geographic locations is discussed and introduced in Section 2.2. After obtaining 6 combined levels of information, we investigate their performance on the target variable and find those medical and non-medical organizations behave differently concerning the number of breached records. It can be observed from last two rows of Table 2 the significant differences between the medical data and the non-medical data; the latter refers to the business and non-business types of organizations in Table 3. Although they all process a similar point mass of zero, their zero proportions differ notably to the extent of the heavy tail and maximum amount measured on the non-zero claim amount. In addition, the medical losses are more compact compared to non-medical losses. We hence postulate that the underlying severity distribution features multimodality; in this sense, a multimodal distribution or mixture distribution could be candidates for modeling the overall losses.

The above-mentioned fact can also be observed from Figure 1, where both histograms of logarithmic records breached and incurred by the medical and the non-medical organizations are displayed. The non-zero severity body part of density of medical organization has a peak of around 600 records, and the probability

Table 2. Summary statistics of target variable.

	Number	Zero Prop.	$q_{0.25}$	Mean	Median	$q_{0.75}$	Maximum
Total	8095	32.90%	1000	1,018,488	2800	13,000	5×10^8
Medical	4161	15.66%	1000	69,353	2300	8800	7.88×10^7
Non-medical	3934	51.12%	900	2,750,425	4700	38,900	5×10^8

Table 3. Type of organizations and breaches.

	Original Types	Combined Levels
MED	Healthcare, Medical Providers and Insurance Services	Medical
BSF	Businesses (Financial and Insurance Services)	Business
BSO	Businesses (Other)	
BSR	Businesses (Merchant including Online Retail)	
EDU	Educational Institutions	Non-business
GOV	Government or Military	
NGO	Nonprofits	
CARD	Fraud Involving Debit and Credit Cards	External Malicious
HACK	Hacked by an Outside Party or Infected by Malware	
INSD	Insider	Internal Malicious
PHYS	Physical	
PORT	Portable Device	
STAT	Stationary Computer Loss	Internal Negligent
DISC	Unintended Disclosure	

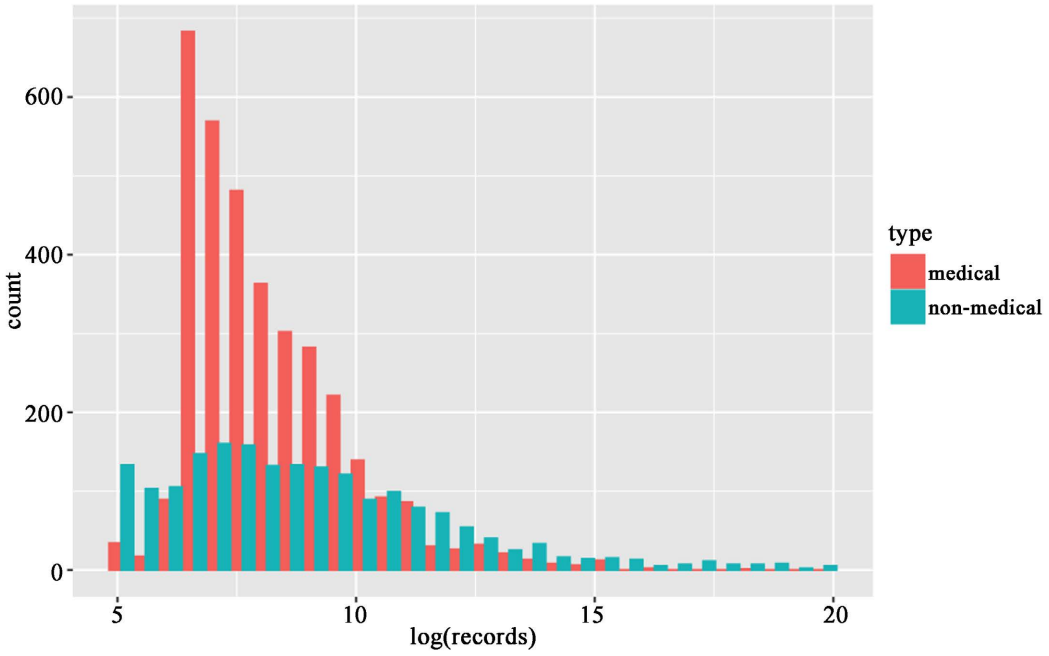

Figure 1. Histograms between medical and non-medical organizations.

for losses being smaller than the mode value is relatively low. Meanwhile, the body part density for non-medical organizations shows a relatively smooth and flat distributional pattern before and after its mode point and relatively a heavier tail. Compared to financial services industry, which has spent the last 20 years focusing on cyber security and protection [27], healthcare organization is not as

frequently attacked by cyber related incidents. Medical organizations form traditionally risk retention group to mitigate huge liability losses caused by cyber breaches, making them reluctant to understand, track, report, and manage threats via open market cyber insurance coverage. Besides, mature incident and vulnerability risk management processes are lacking in most medical organizations [28]. Thus, daily threats are not even reported or managed effectively, which explains the low occurrence of cyber-severity loss of less than 600 incident records. Even though some of the data are already appropriate to model losses with heavy tails, they do not account for this type of multimodality case resulting from data variations observed between medical and non-medical organizations. In this regard, estimating the moderate loss density component with a fixed number of mixed components is advantageous.

2.2. Cluster Analysis

Because the PRC data also contains the geographical location of the victims of cyber attacks, a list of 51 states of U.S. with their latitudes and longitudes serves as the raw data information. It is a common practice that the number of levels in the geographical rating factor are to be reduced in order to provide effective risk measurement for insurance rate-making. For this purpose, we use one of the initialization strategies, cluster analysis [29], to do the analysis. Clustering analysis is a newly developed computer-oriented data analysis which utilize unsupervised machine learning algorithm. Clustering is segmenting a data set based on similarities between the data points. We conduct cluster analysis for three reasons. First, it avoids diluting predictive power caused by the geographical location factor with 51 levels. Second, when states with similar characteristics are grouped, implementing rate-making is simpler. Third, it reduces the likelihood that the rate for one area is drastically differ from that for its neighbouring areas. Cluster analysis divides observations into distinct groups so that the observations within each group are quite similar to one another, as opposed to grouping 51 states into some official government regions, such as those used by the U.S. Census Bureau and the Standard Federal Regions. Cluster analysis divides observations into distinct groups such that the observations within each group are similar to each other. Before clustering, we conduct a cluster analysis using the means of latitude and longitude in each state as representatives. In this regard, we smooth the regression coefficients to make them more reasonable and interpretable, given that clustered groups are based on state average level. Now we have a set of 8095 observations, each with two features, longitude and latitude, that can be used to identify subgroups. We are attempting to discover geographical heterogeneity structures based on the PRC data set, which is an unsupervised problem.

Figure 2 represents the geographical heat map information in a two-dimensional space of longitude and latitude. These are the first two principal components of the data, which summarize the 8095 investigated incidents for location information

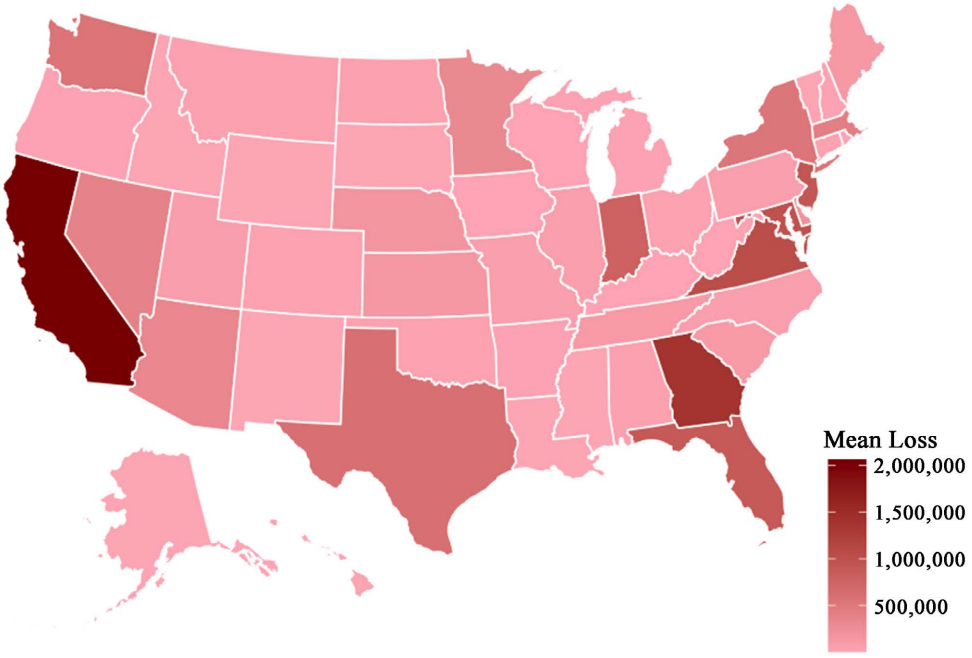

Figure 2. Average severity level among states.

in terms of two geographical dimensions. Each small, closed area corresponds to one of the 51 states, allowing for a visual examination of the average severity level for signs of clustering. There appear to be multiple groups of clusters with similar colour patterns. Two commonly used clustering techniques are *k*-means [30] and hierarchical [31], which have been widely applied in territory studies for finding patterns and investigating the underlying geographical structure of the data. This study uses the *k*-means method with elbow [32] to improve an efficient and effective *k*-means performance. The elbow method is the default method for determining the optimal number of clusters for a characteristic process. The *k*-means clustering algorithm formalizes finding the best similarity grouping where the observations within each cluster are as small as possible, and the variation between clusters is significant. The similarity is measured by the sum of the squared Euclidean distance (SSE) [33], one of the most widely used cluster distance criteria:

$$\mathrm{SSE} = \sum_{k=1}^{K} \left[\sum_{x_i = C_k} \left(x_i - \mu_k \right)^2 \right],$$

where μ_k is the cluster centroid/mean, and C_k represents one of the K clusters. We manually conduct a *k*-means cluster analysis with one to six clusters and calculate the ratio between cluster sum of squares and the total sum of squares for each round. We take this ratio as the *y*-axis and create an elbow plot as inllustrated in **Figure 3(a)**. The plot demonstrates the elbow at $k = 5$, beyond which the gains in the between cluster of sum of squares appear to be minimal as the increase in total sum of squares after $k = 5$ is greatly shrinking down; therefore 5 is the best cluster cut-off point. **Figure 3(b)** depicts the relative

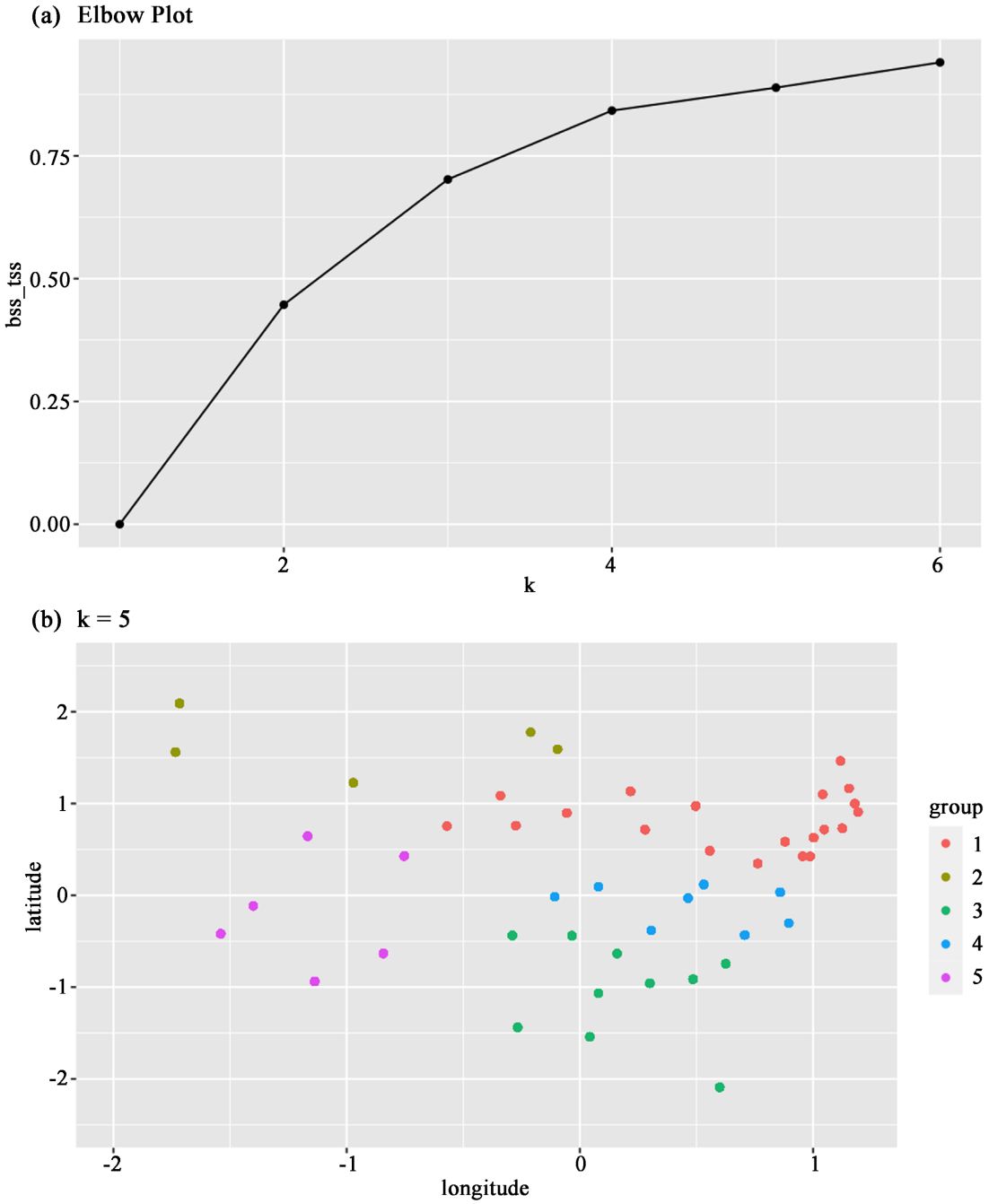

Figure 3. Cluster selection. (a) Elbow plot for clusters; (b) Five geographical clusters.

geographical location of five clusters, while **Appendix A** provides context-specific information about cluster partitioning by state. This way, we can identify the geographical segments of cyber severity and classify them according to similar risk factor factors.

3. Modeling Framework

One of the professional ethics of actuaries is to study loss distributions. Suggested by the empirical data analysis in Section 2.1, the severity distribution of

cyber loss records possesses point masses of zero, over-dispersion and heavy tail on extreme values, and various severity performance types on medical and non-medical organizations, which hardly can be fitted by a single analytic and parametric distribution. Our data set also allows us to examine individual risk characteristics via regression predictors, such as breach type, business type, and location. Based on these characteristics, we propose an FMM with three components integrated into a GLM framework to analyze the severity of a cyber loss.

3.1. Splicing Models

Many risk and loss variables, such as bodily injury costs and cyber losses, have long tails. Therefore, when modelling claim size to set premiums, calculating loss measures, and determining capital requirements for solvency regulations, it is frequently necessary for the actuarial analytic domain to obtain a global fit for risk distributions. In the literature, a splicing model is also called a composite model, in which multiple light-tailed distributions for the body and a heavy-tailed distribution for the tail are combined. The general density form of an m-component spliced distribution is as follows:

$$f(y)=\begin{cases} p_1 f_1(y) & y\in C_1, \\ p_2 f_2(y) & y\in C_2, \\ \vdots & \\ p_m f_m(y) & y\in C_m, \end{cases} \tag{3.1}$$

where f_i is a legitimate density function with all probability on the mutually exclusive and sequentially ordered interval C_i, and positive weights $p_1,\cdots,p_m$ that add up to one, i.e., $\sum_{i=1}^{m} p_i = 1$. In this regard, the density function (3.1) and its corresponding cumulative distribution function can be written as a compact form and as

$$f(y)=\sum_{i=1}^{m} I_{C_i}(y)\, p_i f_i(y), \quad F(y)=\sum_{i=1}^{m} I_{C_i}(y)\left(\sum_{j=1}^{i-1} p_j + p_i F_i(y)\right),$$

where I is an indicator function with $I_{C_i}(y)=1$, if $y\in C_i$, otherwise 0, F_i is the corresponding cumulative distribution function of f_i in the interval C_i.

Based on the empirical analysis results shown in Table 2, we consider a spliced distribution with three components: the first component contains zeros, the second component models the middle segment of the amount of lost data, and the third component models the tail segment. Let Y_j denote the random variable that represents the jth loss amount, c is the non-zero loss threshold, and then the pdf of Y_j can be expressed as

$$f(y_j \mid \boldsymbol{\alpha})=\begin{cases} p_1(y_j;\boldsymbol{\alpha}) & y_j = 0, \\ p_2(y_j;\boldsymbol{\alpha})\dfrac{f_1(y_j;\boldsymbol{\alpha}_1)}{F_1(c;\boldsymbol{\alpha}_1)-F_1(0^+;\boldsymbol{\alpha}_1)} & y_j\in(0,c], \\ \left[1-p_1(y_j;\boldsymbol{\alpha})-p_2(y_j;\boldsymbol{\alpha})\right]\dfrac{f_2(y_j;\boldsymbol{\alpha}_2)}{1-F_2(c;\boldsymbol{\alpha}_2)} & y_j\in(c,\infty), \end{cases} \tag{3.2}$$

where f_1 and f_2 are two density functions with cdf F_1 and F_2, respectively, defined on $(0,\infty)$, $\boldsymbol{\alpha}=\left(\boldsymbol{\alpha}_1^{\mathrm{T}},\boldsymbol{\alpha}_2^{\mathrm{T}}\right)^{\mathrm{T}}$ is a set of parameter vector associated with the distributions for the components. The threshold c is a parameter to be estimated from the data which is investigated in Section 4.3. The remaining unknown parameters p_1, p_2 and $\boldsymbol{\alpha}$ can be estimated using the maximum likelihood estimation (MLE) method by maximizing the log-likelihood function based on observations $y_1, y_2, \cdots, y_n$, which is

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\alpha}) &= \log\left(p_1\left(y_j;\boldsymbol{\alpha}\right)\right)\sum_{j=1}^{n} I_{\{0\}}\left(y_j\right)\\ &+\sum_{j=1}^{n} I_{\left(0^+,c\right]}\left(y_j\right)\left[\log\left(p_2\left(y_j;\boldsymbol{\alpha}\right)\right)+\log f_1\left(y_j;\boldsymbol{\alpha}_1\right)-\log\left(F_1\left(c;\boldsymbol{\alpha}_1\right)-F_1\left(0^+;\boldsymbol{\alpha}_1\right)\right)\right]\\ &+\sum_{j=1}^{n} I_{(c,\infty)(y_j)}\left[\log\left(1-p_1\left(y_j;\boldsymbol{\alpha}\right)-p_2\left(y_j;\boldsymbol{\alpha}\right)\right)+\log f_2\left(y_j;\boldsymbol{\alpha}_2\right)-\log\left(1-F_2\left(c;\boldsymbol{\alpha}_2\right)\right)\right].\end{aligned} \tag{3.3}$$

3.2. Finite Mixture Models

Due to the adaptability in utilizing high-dimensional features, coping with population heterogeneity, and achieving multiple interrelated goals, mixture distributions have gained popularity in recent years. [34] provides a thorough discussion of using the EM algorithm to find maximizers of MLE and the selection of the number of components in finite mixture models. Let $Y_1,\cdots,Y_n$ denote a random sample of size n, and $\boldsymbol{y}=\left(y_1,\cdots,y_n\right)^{\mathrm{T}}$ is the observed value of random vector. Suppose that Y_j follows a finite mixture distribution with density function f on $\mathbb{R}$, which can be written in the form[5]

$$f_M\left(y_j\right)=\sum_{i=1}^{g}\pi_i f_i\left(y_j\right), \tag{3.4}$$

where for $i=1,2,\cdots,g$, f_i is a density function and π_i is a non-negative quantity such that $0\le\pi_i\le 1$ and $\sum_{i=1}^{g}\pi_i=1$. The quantities $\pi_1,\cdots,\pi_g$ are the mixing proportions or weights, and $f_1,\cdots,f_g$ are called the component densities of the mixture. We call density (3.4) as a g-component finite mixture density function and its corresponding distribution F_M as a g-component finite mixture distribution function.

In order to well interpret the mixture models, let $\boldsymbol{Z}_j$ be a g-dimensional component label vector where the ith element $Z_{ij}=\left(\boldsymbol{Z}_j\right)_i$ is defined to be one or zero according to whether the component of Y_j in the mixture is equal to i or not ($i=1,\cdots,g$). Thus this categorical random vector $\boldsymbol{Z}_j$ can be viewed as following a multinomial distribution with probabilities $\pi_1,\cdots,\pi_g$; that is,

$$\mathbb{P}\left\{\boldsymbol{Z}_j=\boldsymbol{z}_j\right\}=\pi_1^{z_{1j}}\pi_2^{z_{2j}}\cdots\pi_g^{z_{gj}} \tag{3.5}$$

according to a multinomial distribution consisting of one draw on g categories with probabilities $\pi_1,\cdots,\pi_g$ We write

[5]In this formulation of the mixture model, the number of components g is considered fixed. In many applications, the value of g is unknown and inferred from the available data, along with the mixing proportions and the parameters in the specified forms of the component densities.

$$\boldsymbol{Z}_j \sim \text{Mult}_g\left(1, \boldsymbol{\pi}\right), \tag{3.6}$$

where $\boldsymbol{\pi} = \left(\pi_1, \cdots, \pi_g\right)^{\mathrm{T}}$. In the interpretation of a mixture model, Y_j is drawn from a population with G with g groups, $G_1, \cdots, G_g$ in proportions $\pi_1, \cdots, \pi_g$, where the density of Y_j in group G_i is $f_i\left(y_j\right)$ The component-indicator variables z_{ij} will be used in finding optimizers under ML estimation via the EM algorithem in Section 4.

Generally, the components can be any exponential family distribution [35]; observations are available from a population known to be a mixture of K sub-populations. In our study, each subpopulation will not necessarily be assumed to have the same type of distribution, which is one of the most significant departures from previous research. For a single observation, the probability density of the exponential family can be expressed as follows:

$$f\left(y_j; \theta_j, \phi\right) = \exp\left\{\frac{y_j\theta_j - b\left(\theta_j\right)}{\phi} + c\left(y_j, \phi\right)\right\} \tag{3.7}$$

where θ_j is a natural parameter or canonical parameter, ϕ is the dispersion parameter or scale parameter[6], $b\left(\theta_j\right)$ is a known function called cumulant function and $c\left(y_j, \phi\right)$ is a normalizing function, thus ensuring that (3.7) is a probability function. The mean and variance of exponential family distributions can be expressed by $b\left(\theta_j\right)$ as follows:

$$\mu = E\left(Y_j\right) = \mu_j = b'\left(\theta_j\right) \text{ and } Var\left(Y_j\right) = \phi b''\left(\theta_j\right). \tag{3.8}$$

Considerations are given to the family of mixtures of generalized linear models (GLMs) because many applications of non-normal mixtures involve components from the exponential family. GLMs are a statistical framework for unifying several significant exponential family models [37]. In this framework, it is permissible for the mixing proportions and the component distributions are allowed to depend on some associated covariates. We have independent sample data $\left\{\boldsymbol{x}_j, y_j\right\}$, $j = 1, \cdots, n$, in which y_j is the response/target variable, n is the sample size, and $\boldsymbol{x}_j = \left(x_{j1}, \cdots, x_{jp}\right)^{\mathrm{T}}$ is a vector of p explanatory variables, GLMs thus fit the following:

$$g\left(\mu_j\right) = \eta_j = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_j = \beta_1 x_{j1} + \cdots + \beta_p x_{jp} \tag{3.9}$$

where η_j is the linear predictor, and μ_j, the mean of an exponential family distribution $f\left(y_j; \theta_j, \phi\right)$, is a known function of the canonical parameter θ_j described in (8), $g\left(\cdot\right)$ is a known link function that connects distribution mean and linear combination together.

From the above methodologies, we propose a finite mixture regression model where mixture components can be from same or different types of parametric family. Our model employs candidate distributions like Gamma, Log Normal,

[6]When ϕ is known, the distribution of Y_j is one-parameter canonical exponential family member. When ϕ is unknown, it is often a nuisance parameter and then it is estimated by the method of moments. In most of GLM theory, the role of ϕ is often treated as an unknown constant but not as a parameter [36].

Inverse Gaussian, and Weibull because loss or severity is typically modelled as continuous random variables.

3.3. Zero Inflated Mixture and Composite Regression Models

In this subsection, we introduce a order to provide a clear understanding of our combined finite mixtures and splicing model, we first perform our model under a general splicing framework with three parts spliced densities jointing with weighting probabilities, followed by a detailed finite mixture section expression of the moderate spliced density part.

Let $Y \in \mathbb{R}^+$ be the claim severity random variable, and let $\boldsymbol{x} \in \mathbb{R}^P$ be the vector of covariate information. The density of the zero-inflated mixture composite regression model written in a spliced form with zero and two densities f_M and f_T and their corresponding cumulative distribution functions (CDFs) F_M and F_T is given by

$$\begin{aligned} & f_Y\left(y_j;\boldsymbol{\alpha},\mathcal{W},\mathcal{B},\boldsymbol{\phi},\boldsymbol{\gamma},\kappa,\boldsymbol{x}_j\right) \\ & = p_1\left(y_j;\boldsymbol{\alpha},\boldsymbol{x}_j\right)\mathbf{1}\left\{y_j=0\right\} \\ & \quad + p_2\left(y_j;\boldsymbol{\alpha},\boldsymbol{x}_j,c\right)\frac{f_M\left(y_j;\mathcal{W},\mathcal{B},\boldsymbol{\phi},\boldsymbol{x}_j\right)}{F_M\left(c;\mathcal{W},\mathcal{B},\boldsymbol{\phi},\boldsymbol{x}_j\right)-F_M\left(0^+;\mathcal{W},\mathcal{B},\boldsymbol{\phi},\boldsymbol{x}_j\right)}\mathbf{1}\left\{0<y_j\le c\right\} \\ & \quad +\left[1-p_1\left(y_j;\boldsymbol{x}_j,\boldsymbol{\alpha}\right)-p_2\left(y_j;\boldsymbol{\alpha},\boldsymbol{x}_j,c\right)\right]\frac{f_T\left(y;\boldsymbol{\gamma},\kappa,\boldsymbol{x}_j\right)}{1-F_T\left(c;\boldsymbol{\gamma},\kappa,\boldsymbol{x}_j\right)}\mathbf{1}\left\{y_j>c\right\}, \end{aligned} \tag{3.10}$$

where $\{p_1, p_2\} \in (0,1)$ are the splicing weights, c is the splicing point which is the threshold separating the moderate and extreme loss values, $\boldsymbol{\alpha}$ is covariate coefficients of zero-inflated weight, $\mathcal{W}$, $\mathcal{B}$ and $\boldsymbol{\phi}$ are parameter vectors of the density of body f_M which is a finite mixture model, and $\boldsymbol{\gamma}$ and κ are coefficients of the density of tail f_T.

In this study, the finite mixture distribution f_M is the density of positively defined continuous distributions with upper truncation at the threshold loss level c.

$$f_M\left(y_j;\mathcal{W},\mathcal{B},\boldsymbol{\phi},\boldsymbol{x}_j\right)=\sum_{i=1}^{g}\pi_{ij}\left(\boldsymbol{x}_j;\mathcal{W}\right)f_i\left(y_j;\exp\left(\boldsymbol{\beta}_i^{\mathrm{T}}\boldsymbol{x}_j\right),\phi_i\right) \tag{3.11}$$

where $\mathcal{B}$ contains the elements of $\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_g$ known a priori to be distinct, $\boldsymbol{\phi}=\left(\phi_1,\cdots,\phi_g\right)^{\mathrm{T}}$ is a vector of fixed dispersion parameters of g distribution components from the exponential family. The parameter π_{ij} is the mixing proportion of the ith function and jth observation which is a function of $\boldsymbol{x}_j$ and commonly modeled by logistic distributions

$$\pi_{ij}=\pi_i\left(\boldsymbol{x}_j;\mathcal{W}\right)=\frac{\exp\left(\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{x}_j\right)}{1+\sum_{h=1}^{g-1}\exp\left(\boldsymbol{\omega}_h^{\mathrm{T}}\boldsymbol{x}_j\right)}, \tag{3.12}$$

where $\mathcal{W}=\left(\boldsymbol{\omega}_1^{\mathrm{T}},\cdots,\boldsymbol{\omega}_{g-1}^{\mathrm{T}},\boldsymbol{\omega}_g^{\mathrm{T}}\right)^{\mathrm{T}}$, with $\boldsymbol{\omega}_g=\mathbf{0}$, contains the logistic regression coefficients. Lastly, f_T is the tail density function from the exponential family with heavy-tailed performance, given by

$$f_T\left(y_j;\boldsymbol{\gamma},\kappa,\boldsymbol{x}_j\right)=f_T\left(y_j;\exp\left(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{x}_j\right),\kappa\right) \tag{3.13}$$

where $\exp\left(\boldsymbol{\gamma}_j^{\mathrm{T}}\boldsymbol{x}_j\right)=\theta_j$, the canonical parameter in (3.7), and κ is the dispersion parameter.

4. Parametric Estimation

Let $\boldsymbol{\Psi}$ represent the set of vectors representing all the unknown parameters in (3.10) that must be estimated, namely,

$$\boldsymbol{\Psi}=\left(\boldsymbol{\alpha}^{\mathrm{T}},\mathcal{W}^{\mathrm{T}},\mathcal{B}^{\mathrm{T}},\boldsymbol{\gamma}^{\mathrm{T}}\right)^{\mathrm{T}}. \tag{4.1}$$

The density of our spliced mixture regression model (3.10) of the jth response variable Y_j can then be written as follows:

$$\begin{aligned} f\left(y_j;\boldsymbol{\Psi}\right)=&\,p_{1_j}\mathbf{1}\left\{y_j=0\right\}+p_{2_j}\frac{\sum_{i=1}^{g}\pi_{ij}f_i\left(y_j;\boldsymbol{\beta}_{ij},\phi_i\right)}{F_M\left(c;\boldsymbol{\Psi}\right)-F_M\left(0^+;\boldsymbol{\Psi}\right)}\mathbf{1}\left\{0<y_j\le c\right\}\\ &+\left(1-p_{1_j}-p_{2_j}\right)\frac{f_T\left(y_j;\boldsymbol{\gamma}_j,k\right)}{1-F_T\left(c;\boldsymbol{\Psi}\right)}\mathbf{1}\left\{y_j>c\right\}. \end{aligned} \tag{4.2}$$

where $p_{1_j}=p_1\left(\boldsymbol{x}_j;\boldsymbol{\alpha}\right)$ and $p_{2_j}=p_2\left(\boldsymbol{x}_j;\boldsymbol{\alpha},c\right)$. In this way, the log likelihood for $\boldsymbol{\Psi}$ can be formed as

$$\begin{aligned} &\log\mathcal{L}\left(\boldsymbol{\Psi}\right)\\ &\propto\sum_{j=1}^{n}\log\left(p_{1_j}\right)\mathbf{1}\left\{y_j=0\right\}+\sum_{j=1}^{n}\left\{\log\left(p_{2_j}\right)+\log\left(\sum_{i=1}^{g}\pi_{ij}f_i\left(y_j;\boldsymbol{\beta}_{ij},\phi_i\right)\right)\right.\\ &\left.-\log\left[F_M\left(c;\boldsymbol{\Psi}\right)-F_M\left(0^+;\boldsymbol{\Psi}\right)\right]\right\}\mathbf{1}\left\{0<y_j\le c\right\}\\ &+\sum_{j=1}^{n}\left\{\log\left(1-p_{1_j}-p_{2_j}\right)+\log f_T\left(y_j;\boldsymbol{\gamma}_j,k\right)-\log\left[1-F_T\left(c;\boldsymbol{\Psi}\right)\right]\right\}\mathbf{1}\left\{y_j>c\right\}. \end{aligned} \tag{4.3}$$

The EM algorithm [38] can be applied to obtain the MLE of $\boldsymbol{\Psi}$ in this spliced mixture regression model. The complete-data log likelihood is given by the following

$$\begin{aligned} &\log\mathcal{L}_c\left(\boldsymbol{\Psi}\right)\\ &\propto\sum_{j=1}^{n}\log\left(p_{1_j}\right)\mathbf{1}\left\{y_j=0\right\}+\sum_{j=1}^{n}\left\{\log\left(p_{2_j}\right)+\sum_{i=1}^{g}z_{ij}\left[\log\left(\pi_{ij}\right)+\log f_i\left(y_j;\boldsymbol{\beta}_{ij},\phi_i\right)\right]\right.\\ &\left.-\log\left[F_M\left(c;\boldsymbol{\Psi}\right)-F_M\left(0^+;\boldsymbol{\Psi}\right)\right]\right\}\mathbf{1}\left\{0<y_j\le c\right\}\\ &+\sum_{j=1}^{n}\left[\log\left(1-p_{1_j}-p_{2_j}\right)+\log f_T\left(y_j;\boldsymbol{\gamma}_j,k\right)-\log\left(1-F_T\left(c;\boldsymbol{\Psi}\right)\right)\right]\mathbf{1}\left\{y_j>c\right\} \end{aligned} \tag{4.4}$$

where z_{ij} denotes the component-indicator variables as defined in (3.5).

4.1. E-Step

The EM algorithm is applied to this problem by treating the z_{ij} as missing data. E (for expectation) and M (for maximization) are the two iterative steps. Given an observed data $\boldsymbol{y}$, we take the conditional expectation of the compe-

lete-data log likelihood (4.4) using the current fit for $\boldsymbol{\Psi}$. We consider $\boldsymbol{\Psi}^{(0)}$ as an initial value of the iterative computation. The E-step computes the conditional expectation of $\log \mathcal{L}_c(\boldsymbol{\Psi})$ given $\boldsymbol{y}$ using $\boldsymbol{\Psi}^{(0)}$ for $\boldsymbol{\Psi}$ on the first EM algorithm iteration, that is

$$Q\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(0)}\right)=\mathbb{E}_{\boldsymbol{\Psi}^{(0)}}\left[\log \mathcal{L}_c(\boldsymbol{\Psi})\,|\,\boldsymbol{y}\right] \tag{4.5}$$

where the subscript $\boldsymbol{\Psi}^{(0)}$ means that using $\boldsymbol{\Psi}^{(0)}$ for $\boldsymbol{\Psi}$ effect the expectation. In this manner, the E-step calculates $Q\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}\right)$ on the $(k+1)$th iteration, where $\boldsymbol{\Psi}^{(k)}$) is the value of $\boldsymbol{\Psi}$) after kth iteration. The E-step requires the calculation of the current conditional expectation of Z_{ij} given the observed data $\boldsymbol{y}$, respectively

$$\mathbb{E}_{\boldsymbol{\Psi}^{(k)}}\left(Z_{ij}\,|\,\boldsymbol{y}\right)=\mathbb{P}_{\boldsymbol{\Psi}^{(k)}}\left\{Z_{ij}=1\,|\,\boldsymbol{y}\right\}=\tau_{ij}\left(y_j;\boldsymbol{Y}^{(k)}\right), \tag{4.6}$$

where

$$\begin{aligned}\tau_{ij}\left(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)}\right)&=\pi_{ij}^{(k)}\frac{f_i\left(y_j;\boldsymbol{\beta}_{ij}^{(k)},\phi_i\right)}{f_M\left(y_j;\boldsymbol{\Psi}^{(k)}\right)}\\&=\pi_{ij}^{(k)}\frac{f_i\left(y_j;\boldsymbol{\beta}_{ij}^{(k)},\phi_i\right)}{\sum_{h=1}^{g}\pi_{hj}^{(k)}f_h\left(y_j;\boldsymbol{\beta}_{hj}^{(k)},\phi_h\right)}\end{aligned} \tag{4.7}$$

for $i=1,\cdots,g$; $j=1,\cdots,n$. The quantity $\tau_{ij}\left(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)}\right)$ is the posterior probability that the jth member of the sample with observed value $\boldsymbol{y}_j$ belongs to the ith component of the mixture. Taking the conditional expectation of (4.4) using (4.6) that

$$\begin{aligned}&Q\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}\right)\\&=\sum_{j=1}^{n}\log\left(p_{1_j}\right)\mathbf{1}\left\{y_j=0\right\}+\sum_{j=1}^{n}\left\{\log\left(p_{2_j}\right)+\sum_{i=1}^{g}\tau_{ij}\left(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)}\right)\left[\log\left(\pi_{ij}\right)\right.\right.\\&\left.\left.+\log f_i\left(y_j;\boldsymbol{\beta}_{ij},\phi_i\right)\right]-\log\left[F_M\left(c;\boldsymbol{\Psi}\right)-F_M\left(0^{+};\boldsymbol{\Psi}\right)\right]\right\}\mathbf{1}\left\{0<y_j\le c\right\}\\&+\sum_{j=1}^{n}\left[\log\left(1-p_{1_j}-p_{2_j}\right)+\log f_T\left(y_j;\boldsymbol{\gamma}_j,k\right)-\log\left(1-F_T\left(c;\boldsymbol{\Psi}\right)\right)\right]\mathbf{1}\left\{y_j>c\right\}\end{aligned} \tag{4.8}$$

We assume $F_M\left(0^{+};\boldsymbol{\Psi}\right)=0$ in the following derivations, which is generally the case.

4.2. M-Step

The M-step on the $(k+1)$th iteration entails solving the following system of four equations:

$$\sum_{j=1}^{n}\frac{\partial}{\partial\boldsymbol{\alpha}}\left[\log\left(p_{1_j}\right)\mathbf{1}\left\{y_j=0\right\}+\log\left(p_{2_j}\right)\mathbf{1}\left\{0<y_j\le c\right\}+\log\left(1-p_{1_j}-p_{2_j}\right)\mathbf{1}\left\{y_j>c\right\}\right] \tag{4.9}$$
$$=\mathbf{0},$$

$$\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_i\left(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)}\right)\frac{\partial}{\partial\mathcal{W}}\log\left(\pi_{ij}\right)\mathbf{1}\left\{0<y_j\le c\right\}=\mathbf{0}, \tag{4.10}$$

$$\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_i\left(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)}\right)\frac{\partial}{\partial\mathcal{B}}\left[\log f_i\left(y_j;\boldsymbol{\beta}_{ij},\phi_i\right)-\log\left[F_M\left(c;\boldsymbol{\Psi}\right)-F_M\left(0^+;\boldsymbol{\Psi}\right)\right]\right]\mathbf{1}\left\{0<y_j\le c\right\} \quad (4.11)$$
$$=\mathbf{0}$$

$$\sum_{j=1}^{n}\frac{\partial}{\partial\boldsymbol{\gamma}}\left[\log f_T\left(\boldsymbol{y}_j;\boldsymbol{\gamma},k\right)-\log\left(1-F_T\left(c;\boldsymbol{\Psi}\right)\right)\right]\mathbf{1}\left\{y_j>c\right\}=\mathbf{0}. \quad (4.12)$$

Equations (4.9) and (4.10) can be solved using a standard algorithm for logistic regression to produce updated estimate $\boldsymbol{\alpha}^{(k+1)}$ and $\mathcal{W}^{(k+1)}$ for the logistic regression coefficients as they both represent the probabilities between 0 and 1. Concerning the computation of $\mathcal{B}$ and $\boldsymbol{\gamma}$ and applying the chain rule of [39], the likelihood equation for $\boldsymbol{\gamma}$ (4.12) can be expressed as

$$\sum_{j=1}^{n}w\left(\mu_j\right)\left(y_j-\mu_j\right)\eta'\left(\mu_j\right)\boldsymbol{x}_j=\mathbf{0}, \quad (4.13)$$

where $\mu_j=\exp\left(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{x}_j\right)$, $\eta'\left(\mu_j\right)=\mathrm{d}\eta_j/\mathrm{d}\mu_j$ and $w\left(\mu_j\right)$ is the weight function defined by

$$w\left(\mu_j\right)=1\Big/\left[\eta_j'\left(\mu_j\right)\right]^2V\left(\mu_j\right). \quad (4.14)$$

The likelihood Equation (4.13) can be solved using iteratively reweighted least squares (IRLS) [37]. The adjusted response variable $\tilde{y}_j$ for the $(k+1)$th iteration is

$$\tilde{y}_j^{(k)}=\eta\left(\mu_j^{(k)}\right)+\left(y_j-\mu_j^{(k)}\right)\eta'\left(\mu_j^{(k)}\right). \quad (4.15)$$

These n adjusted responses are then regressed on the covariates $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ using weights $w\left(\mu_1^{(k)}\right),\cdots,w\left(\mu_n^{(k)}\right)$. This produces an updated estimate $\boldsymbol{\gamma}^{(k+1)}$ for $\boldsymbol{\gamma}$ and, consequently, the updated estimates $\mu_j^{(k+1)}$ for the μ_j for use in the right-hand side of (4.15) to update the adjusted responses, and so on. This procedure is repeated until the variations in the estimates are small enough. Same as (4.13), the likelihood for $\mathcal{B}$ in (4.11) can be written as

$$\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}\left(y_j;\boldsymbol{\Psi}^{(k)}\right)w\left(\mu_{ij}\right)\left(y_j-\mu_{ij}\right)\eta_i'\left(\mu_{ij}\right)\left[\frac{\partial}{\partial\mathcal{B}}\eta_i\left(\mu_{ij}\right)\right]=\mathbf{0} \quad (4.16)$$

where μ_{ij} is the mean of Y_j for the ith component. Given that

$$\frac{\partial}{\partial\boldsymbol{\beta}_h}\eta_i\left(\mu_{ij}\right)=\begin{cases}\boldsymbol{x}_j, & \text{if } h=i\\ 0, & \text{otherwise,}\end{cases} \quad (4.17)$$

Equation (4.16) reduces to solving

$$\sum_{j=1}^{n}\tau_{ij}\left(y_j;\boldsymbol{\Psi}^{(k)}\right)w\left(\mu_{ij}\right)\left(y_j-\mu_{ij}\right)\eta_i'\left(\mu_{ij}\right)\boldsymbol{x}_j=\mathbf{0} \quad (4.18)$$

Separately for each $\boldsymbol{\beta}_i$ to produce $\boldsymbol{\beta}_i^{(k+1)}$, $i=1,\cdots,g$. As (4.15) responses $y_1,\cdots,y_n$ are fitted with weights $\tau_{i1}\left(y_1;\boldsymbol{\Psi}^{(k)}\right),\cdots,\tau_{in}\left(y_n;\boldsymbol{\Psi}^{(k)}\right)$ and fixed dispersion parameter k_i. (4.18) then can be solved using the IRLS approach for a single GLM. The double summation over i and j in (4.16) can be handled by expanding the response vector to have dimension $g\times n$ by replicating each original observation $\left(y_j;\boldsymbol{x}_j^{\mathrm{T}}\right)^{\mathrm{T}}$ g times, with weights $\tau_{i1}\left(y_1;\boldsymbol{\Psi}^{(k)}\right),\cdots,\tau_{in}\left(y_n;\boldsymbol{\Psi}^{(k)}\right)$,

fixed dispersion parameters $k_1,\cdots,k_g$, and linear predictors $\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{\beta}_1,\cdots,\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{\beta}_g$.

4.3. Starting Values

To allow the overall distribution to disjoint at the splicing point, we maintain a density equation at c by setting

$$p_2\frac{f_M\left(c;\boldsymbol{x},\mathcal{B},\boldsymbol{\Phi}\right)}{F_M\left(c;\boldsymbol{x},\mathcal{B},\boldsymbol{\Phi}\right)-F_M\left(0^+;\boldsymbol{x},\mathcal{B},\boldsymbol{\Phi}\right)}=\left[1-p_1\left(y;\boldsymbol{x},\boldsymbol{\alpha}\right)-p_2\right]\frac{f_T\left(c;\boldsymbol{x},\boldsymbol{\gamma},\boldsymbol{\theta}\right)}{1-F_T\left(c;\boldsymbol{x},\boldsymbol{\gamma},\boldsymbol{\theta}\right)}. \quad (4.19)$$

After simplifying 4.19, we derive p_2 as a function of any given p_1

$$p_2=\left[1-p_1\left(y;\boldsymbol{x},\boldsymbol{\alpha}\right)\right]\frac{f_T\left(c;\boldsymbol{x},\boldsymbol{\gamma},\boldsymbol{\theta}\right)}{1-F_T\left(c;\boldsymbol{x},\boldsymbol{\gamma},\boldsymbol{\theta}\right)}\Bigg/\left\{\frac{f_M\left(c;\boldsymbol{x},\mathcal{B},\boldsymbol{\Phi}\right)}{F_M\left(c;\boldsymbol{x},\mathcal{B},\boldsymbol{\Phi}\right)-F_M\left(0^+;\boldsymbol{x},\mathcal{B},\boldsymbol{\Phi}\right)}+\frac{f_T\left(c;\boldsymbol{x},\boldsymbol{\gamma},\boldsymbol{\theta}\right)}{1-F_T\left(c;\boldsymbol{x},\boldsymbol{\gamma},\boldsymbol{\theta}\right)}\right\}$$

Until now, we have considered the fitting of a Zi-MCR for a given value of severity threshold c in the composite model. Typically, where our model is being used to handle overdispersion, c is predetermined from data using extreme value analysis or expert opinion via performing cyber insurance policy limit and similar matters. This is primarily motivated by estimation stability, which is adopted by [17]. Furthermore, conducting formal tests at any stage of this sequential process is chanllenging because regularity conditions for the likelihood ratio test statistic's typical asymptotic null distribution do not hold [34]. Observing the trend in the log-likelihood as c is increased from a sequence of severity levels (1000, 5000, 10,000, 50,000, 100,000) can provide us with a heuristic for determining the optimal value for c. When dealing with a data-driven model, this method for selecting a splicing point makes more sense and is widely used [40].

5. Analysis of Severity of Data Breaches

In this Section, we illustrate the efficiency of the EM algorithm on estimation by fitting a Zi-MCR model, as proposed in Section 3.3, to the PRC dataset. Furthermore, general model form and covariates are discussed, and several distribution combinations are tested to select the best performance.

This study is based upon the PRC cyber breach incident data by stratifying the residuals. The training set fine-tunes all candidate models, and their performance and out-of-sample validation are checked upon the test set. We conduct 5-fold cross-validation and set 80% as the training data to fit the models. Based on a set of breach observations, the problem is to estimate whether the unknown parameters can be contained in the vector $\boldsymbol{\Psi}=\left(\boldsymbol{\alpha}^{\mathrm{T}},\mathcal{W}^{\mathrm{T}},\mathcal{B}^{\mathrm{T}},\boldsymbol{\gamma}^{\mathrm{T}}\right)^{\mathrm{T}}$, as in 4.1. We represent the logarithm rescaled number of loss records of data breach incidents explained in Section 2 as a target or dependent random variable Y_j, and 9 covariates, including intercept, 2 business levels, 2 breach levels, and 4 location area levels, as described in the Sections 2.1 and 2.2 as vector coefficients $\boldsymbol{x}_j$.

Table 4 displays the summaries of three categorical variables; the proportion

of zeros and differences between the mean values of the categories numerically illustrate their distribution patterns. These results demonstrate the importance of letting splicing weights depend on covariates and separately modelling body and tail parts.

In the PRC dataset, a source of the heterogeneity is mainly from businesses that have or do not have high prevention defence systems and active cyber risk managing activities, such as healthcare and financial service organizations. This is explained in Section 2.1 by comparing their kernel plots and enterprise features. The body part component may be viewed as two groups corresponding to whether those incidents happened within medical organizations. The problem is to estimate the medical and non-medical organization mixture rate, that is, the mixing proportion π_1. Given $g=2$, $\boldsymbol{\alpha}^{(k+1)}$ and $\mathcal{W}^{(k+1)}$ can be calculated using binomial error structure with the canonical logit transformation as the link. For illustrative purposes, we fit several popular distribution combinations on a mixture of body and heavy tail parts. To measure the overall goodness of fit of those fitted distributions, we calculate the Akaike information criterion (AIC) statistics. Table 5 reports the global fit distributions overall AIC values on a given $c=5000$ threshold. The fit from Lognormal and Weibull body mixture and Pareto tail outperforms with the lowest AIC. We conduct a simulation study

Table 4. Summary of categorical variables.

Feature	Category	Zeros	Mean	Total Count
Organizations	Medical	652 (15.7%)	58,501	4161
	Businesses	1434 (63.0%)	2,197,387	2275
	Non-businesses	577 (34.8%)	174,932	1659
Breaches	External Malicious	1125 (44.1%)	1,635,354	2549
	Internal Malicious	775 (31.8%)	462,591	2440
	Internal Negligent	763 (24.6%)	75,805	3106
Territories	Area 1	1064 (35.2%)	390,949	3024
	Area 2	143 (26.9%)	301,578	531
	Area 3	449 (27.3%)	666,947	1642
	Area 4	283 (25.9%)	389,733	1093
	Area 5	724 (40.1%)	1,478,791	1805

Table 5. Overall goodness-fit.

Body		Tail	AIC	Body		Tail	AIC
Gamma	Lognormal	Pareto	−46.3333	Gamma	Lognormal	Lognormal	−51.5420
Gamma	Gamma	Pareto	−55.8390	Gamma	Gamma	Lognormal	−49.0340
Lognormal	Weibull	Pareto	−56.6044	Lognormal	Weibull	Lognormal	−45.6592
Gamma	Weibull	Pareto	−55.8390	Gamma	Weibull	Lognormal	−47.3896

based on the entire data set to comprehend the chosen model's adaptability further. The procedure is repeated 200 times to ensure a thorough analysis of the chosen distribution combination. Finally, the estimated parameters are summarized in Table 6 using the above distribution combined with the lowest AIC.

Since the metric is based on data-driven analysis, we can draw a conclusion

Table 6. Parameter estimations.

Vector	Coefficients	Estimation	Vector	Coefficients	Estimation
	α_1	−1.2292		γ_1	5.2892
	α_2	−0.3876		γ_2	0.6577
	α_3	−0.0915		γ_3	−0.1476
	α_4	0.5934		γ_4	−0.6063
$\boldsymbol{\alpha}^{\mathrm{T}}$	α_5	1.6082	$\boldsymbol{\gamma}^{\mathrm{T}}$	γ_5	1.1548
	α_6	−0.1650		γ_6	−0.4896
	α_7	−0.2233		γ_7	0.7618
	α_8	−0.1999		γ_8	1.4402
	α_9	−0.6482		γ_9	1.5340
	β_{11}	1.1170		β_{21}	2.1593
	β_{12}	0.9828		β_{22}	0.1305
	β_{13}	−0.3131		β_{23}	−0.2645
	β_{14}	0.0360		β_{24}	0.0192
$\boldsymbol{\beta}_1^{\mathrm{T}}$	β_{15}	0.0404	$\boldsymbol{\beta}_2^{\mathrm{T}}$	β_{25}	−0.2291
	β_{16}	−0.4804		β_{26}	0.1814
	β_{17}	0.4161		β_{27}	0.2837
	β_{18}	0.2168		β_{28}	0.2892
	β_{19}	0.0734		β_{29}	0.2688
	w_1	0.1704			
	w_2	0.9999			
	w_3	0.5000			
	w_4	0.2856			
$\mathcal{W}^{\mathrm{T}}$	w_5	0.1557			
	w_6	0.6757			
	w_7	1.0000			
	w_8	1.0000			
	w_9	0.0027			

that the selected combination, Lognormal-Weibull and Pareto, has the most ex-

planation power of PRC dataset. If other dataset is given, methodology and algorithm in generating the metric should not change, selected distribution combination would depend on a case by case basis.

6. Cyber Loss Aggregation and Application

The premium calculation algorithm, known as rate order calculation, is applied to categorize segmentation to derive final premium rates. In order to set competitive premiums and develop sustainable underwriting plans, insurers extensively use historical loss data to seek economies of scale and premium balancing. Statistical algorithms and mathematical modelling arguments are used to structure aggregate cyber risk. The purpose of this Section is to describe an aggregate loss model based on the total amount of cyber loss that occurs in a quarter concerning a group of homogeneous risk characteristics and apply this model to determining increased limit factors (ILFs) based on the underlying data in order to balance statistical and economic constraints. A case study is established for two locations and business types (non-business and business, respectively). Their cyber risks are aggregated under all possible policy limits and deductible combinations. One method for addressing this issue is to separately model the individual severity levels and the quarterly incidence rate. These components can then derive a distribution for the total in a period. This approach has several advantages: changes over time can be monitored and attributed to frequency or severity as cyber risk shows a strong seasonal pattern. The other is that methodology can be quickly adapted to find suitable models for the components, as in the previous model selection. According to risk theory [41], a collective risk model with aggregate loss S, which represents the total amount for a quarterly cyber risk, can be defined as follows

$$S = X_1 + X_2 + \cdots + X_N,$$

where loss count N and non-negative severities $X_1, \cdots, X_N$ are random variables with independence assumptions that N does not rely on the severity of loss and $X_i s$ are *i.i.d.* independently with respect to a given count N. Under this assumption, the aggregate loss distribution (ALD) F_S is a compound *cdf* of the following form [16]:

$$F_S(s) = \sum_{n=0}^{\infty} p_N(n) Pr(S \le s \mid N = n) = E_N F_X^{*(n)}(s),$$

where $p_N(n) = Pr(N = n)$ and $F_X^{*(n)}$ is the n-fold convolution of F_X, which is defined by:

$$F_X^{*(n)}(x) = \begin{cases} \int_0^x f_X(y) \mathrm{d}F_X^{*(n-1)}(x-y)\mathrm{d}y & n = 2,3,\cdots \\ F_X(x) & n = 1. \end{cases}$$

Because of the complexity of this claim amount distribution in practical applications, the following approximations for the distribution of the total loss are typically taken into account:

- Pure Risk Premium

$$P = E(N)E(X),$$

- Premium with Safety Loading Factor θ

$$P_{SL}(\theta) = (1+\theta)E(N)E(X) \quad \theta \geq 0,$$

- Premium with Variance Loading Factor a

$$P_V(a) = E(N)E(X) + a\left[E(N)Var(X) + E(X)^2 Var(N)\right] \quad a \geq 0,$$

- Premium with Standard Deviation Loading factor b

$$P_{SD}(b) = E(N)E(X) + b\sqrt{E(N)Var(X) + E(X)^2 Var(N)} \quad b \geq 0,$$

In our case study, where loading factors are not specified, pure premium creates an example, which can be modified once loading information is obtained from businesses. Table 7 displays the total quarterly dollar loss caused by cyber risk. While frequency distribution is determined by our previous investigation of the generalized linear mixed model [26] of the same dataset, severity distribution is the distribution combined with the lowest AIC in Section 5. With dollar amount transferred from units of loss record by applying following rule [42]:

$$\ln(\text{dollar amount loss}) = 7.68 + 0.76 \times \ln(\text{loss records breached})$$

Table 7. Quarterly aggregate loss in dollar amount.

Location	Business Type	Deductible	Max. Coverage	Estimated Loss	$TVaR_{0.95}$
Northeast	Business	-	-	USD 197,891	USD 37,569
		USD 10,000	-	USD 188,469	USD 37,284
		-	USD 1M	USD 197,891	USD 32,182
		USD 10,000	USD 1M	USD 188,469	USD 31,983
	Non-Business	-	-	USD 2,283,023	USD 38,256
		USD 10,000	-	USD 2,273,881	USD 37,965
		-	USD 1M	USD 1,164,335	USD 28,538
		USD 10,000	USD 1M	USD 1,162,902	USD 28,019
West	Business	-	-	USD 1,408,541	USD 24,103
		USD 10,000	-	USD 1,398,568	USD 23,997
		-	USD 1M	USD 1,264,013	USD 19,326
		USD 10,000	USD 1M	USD 1,260,245	USD 19,145
	Non-Business	-	-	USD 14,661,661	USD 104,839
		USD 10,000	-	USD 14,651,699	USD 104,401
		-	USD 1M	USD 1,680,241	USD 43,843
		USD 10,000	USD 1M	USD 1,680,149	USD 43,497

For the collective risk model, the expected value and variance of aggregate

claims S are as follows:

$$E(S) = E(N)E(X),$$

$$Var(S) = E(N)Var(X) + E(X)^2 Var(N).$$

Estimates of quantities such as VaR, TVaR, and ILFs can be analyzed using aggregated results at a given risk tolerance level. Utilizing risk characteristics, our model divides homogeneous risks into segments. Then, product designers can decide whether to implement policy limits or seek reinsurance. It can be broken down into four steps: establishing the base rate, applying risk factors, multiplying discount and surcharge, and factoring in expense retention. The first and second steps lay the groundwork for the entire pricing process. They are frequently carried out using an experienced-based pricing technique with preliminary data for analysis. Our cyber risk loss aggregation results are generated within a Bayesian framework, which proves to be a useful prediction tool for estimating future loss among segmentation with confidence.

7. Conclusions and Discussions

Once an insurance loss model has been constructed, in addition to applying techniques to data sets, we must consider numerous modelling-related factors, such as risk management and pricing decisions (for insurers) or the impact on capital requirements (for enterprises). Our model and findings provide meaningful insights to risk mitigation and risk transfer techniques, which benefit not only the individual organization, but also the overall economy.

Cyber risk exists because computer data is valuable to individuals. Business, and governments; therefore, the data must be protected by organizations that store privileged information. Financial firms receive, maintain, and store large amounts of personally identifiable information. Recent security research [43] indicates that most businesses have unprotected data and inadequate cybersecurity practices, making them susceptible to data loss. As more executives and decision-makers recognize the value and significance of security investments, cybersecurity budgeting has steadily risen to successfully combat potential digital property loss. A systemic cyber event could cost multiple times the current risk retention estimate. As a result of regulatory scrutiny and the need for improved portfolio management, businesses conduct scenario modelling and sensitivity tests regularly based on their changing risk appetite. To reduce cyber risk, organization can adopt threshold limits by monitoring risk with preset limits based on established risk criteria, trigger will be placed in threshold that has been breached. The objective is to achieve and maintain an acceptable level of risk at a reasonable cost. Under the leadership of the Chief Risk Officer, companies must revise their strategies, including changes to their risk appetite and the composition of their hedge products. Due to some businesses' nature or responsibilities, increasing risk appetite or security investments may not be sufficient to achieve the risk management objective. Such limited reserved retention can have dis-

astrous financial consequences if a data breach occurs, forcing the organization to absorb the costs associated with internal remediation and its liability to third parties. In this perspective, cyber insurance has become an effective alternative or backup tool for managing cyber risk.

Our investigation of large claims and an excess of zeros raises the issue of the risk's insurability under various feature characteristics. To eliminate the variance caused by the heavy tail, the non-catastrophe loss will then be used to train a predictive model. Reinsurance will kick in if the loss exceeds the company's tolerance level to ensure that insurers are not severely impacted. Our model offers additional perspective on coping with extreme loss value, such as cyber risk. Risk selection is one of the most crucial processes when designing an insurance product. Since not all customers are equally attracted to an insurance product, segmenting the risks into distinct groups is advantageous to prevent adverse selection. To ensure that cyber insurance products are priced appropriately, we use the results from the Section 6 to divide risks into categorizable segments. In addition, our model can be utilized to perform a preliminary pre-screening of a prospective client to facilitate rate discrimination and the creation of customized contracts. This security audit enables the insurer to capitalize on the profit opportunity presented by the interdependence of cyber risk.

Combining the loss frequency and severity distribution through convolution is the conventional method for estimating the aggregate loss distribution. Given the proposed mixture and composite severity model, aggregate losses can be estimated through simulation since our previous frequency model was semi-parametric with a simulated posterior distribution without a closed form of distribution.

The financial sector faces cybersecurity risks in their daily operations while insuring product providers. Insurers receive personal health and financial information from both policyholders and claimants. The cost of cyber insurance increased by an average of 96% in the third quarter of 2021 as organizations faced a daily onslaught of cyberattacks [44]. To mitigate the premium price increase, policyholders increased their retention. As a result, insurers must improve predictive analysis and cyber risk models to maintain market share and company solvency. Our model provides a method for measuring cyber risk severity, and there are multiple ways to extend this method. As previously stated, all of our results are based on the assumption of equal exposure, whereas exposure is the most crucial factor in determining the pure premium. Cyber risk loss exposures is any condition that presents the possibility of financial loss to an organization from property, net income and liability as a consequence of advanced technology transmissions, operations, maintenance, development and support. Training the predictive model under the assumption of equal exposure in a defined time period would be an important direction for future research once prior experience data with exposure information is obtained.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Gartner Research (2018) Forecast Analysis: Information Security, Worldwide, 2Q18 Update. Gartner Research.

[2] AXA (2019) AXA & Eurasia Group Future Risks Report. AXA.

[3] (2020) IBM: Cost of a Data Breach Report. *Computer Fraud & Security*, **2021**. https://doi.org/10.1016/S1361-3723(21)00082-8

[4] Leswing, K. (2020) Twitter Hackers Who Targeted Elon Musk and Others Received $121,000 in Bitcoin, Analysis Shows. CNBC TECH. https://www.cnbc.com/2020/07/16/twitter-hackers-made-121000-in-bitcoin-analysis-shows.html

[5] Marriott (2020) Marriott International Notifies Guests of Property System Incident. *Marriott International News Center*. https://news.marriott.com/news/2020/03/31/marriott-international-notifies-guests-of-property-system-incident

[6] Equifax (2017) Equifax acquires data-crédito. https://investor.equifax.com/news-events/press-releases/detail/1221/equifax-acquires-data-crdito

[7] Michael, A.B. (2020) Exposure Measures for Pricing and Analyzing the Risks in Cyber Insurance. Casualty Actuarial Society and Society of Actuaries.

[8] Farley, J. (2022) The Cyber Insurance Market Struggles with Continued Hardening Market Conditions. Gallagher.

[9] Malavasi, M., Gareth, P., Shevchenko, P.V., Trück, S., Jang, J. and Sofronov, G. (2021) Cyber Risk Frequency, Severity and Insurance Viability. (Preprint) https://doi.org/10.2139/ssrn.3940329

[10] Sun, H., Xu, M. and Zhao, P. (2021) Modeling Malicious Hacking Data Breach Risks. *North American Actuarial Journal*, **25**, 484-502. https://doi.org/10.1080/10920277.2020.1752255

[11] Farkas, S., Lopez, O. and Thomas, M. (2021) Cyber Claim Analysis Using Generalized Pareto Regression Trees with Applications to Insurance. *Insurance: Mathematics and Economics*, **98**, 92-105. https://doi.org/10.1016/j.insmatheco.2021.02.009

[12] McNeil, A.J. (1997) Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory. *ASTIN Bulletin*: *The Journal of the IAA*, **27**, 117-137. https://doi.org/10.2143/AST.27.1.563210

[13] Allen, D.E., Singh, A.K. and Powell, R.J. (2013) EVT and Tail-Risk Modelling: Evidence from Market Indices and Volatility Series. *The North American Journal of Economics and Finance*, **26**, 355-369. https://doi.org/10.1016/j.najef.2013.02.010

[14] Park, M.H. and Kim, J.H.T. (2016) Estimating Extreme Tail Risk Measures with Generalized Pareto Distribution. *Computational Statistics & Data Analysis*, **98**, 91-104. https://doi.org/10.1016/j.csda.2015.12.008

[15] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J.L. (2004) Statistics of Extremes: Theory and Applications. John Wiley & Sons, Hoboken. https://doi.org/10.1002/0470012382

[16] Klugman, S.A., Panjer, H.H. and Willmot, G.E. (2012) Loss Models: Further Topics. John Wiley & Sons, Hoboken. https://doi.org/10.1002/9781118787106

[17] Reynkens, T., Verbelen, R., Beirlant, J. and Antonio, K. (2017) Modelling Censored Losses Using Splicing: A Global Fit Strategy with Mixed Erlang and Extreme Value Distributions. *Insurance*: *Mathematics and Economics*, **77**, 65-77. https://doi.org/10.1016/j.insmatheco.2017.08.005

[18] Gan, G. and Valdez, E.A. (2018) Fat-Tailed Regression Modeling with Spliced Distributions. *North American Actuarial Journal*, **22**, 554-573. https://doi.org/10.1080/10920277.2018.1462718

[19] Tzougas, G., Vrontos, S. and Frangos, N. (2014) Optimal Bonus-Malus Systems Using Finite Mixture Models. *ASTIN Bulletin*: *The Journal of the IAA*, **44**, 417-444. https://doi.org/10.1017/asb.2013.31

[20] Sattayatham, P. and Talangtam, T. (2012) Fitting of Finite Mixture Distributions to Motor Insurance Claims. *Journal of Mathematics and Statistics*, **8**, 49-56. https://doi.org/10.3844/jmssp.2012.49.56

[21] Bernardi, M., Maruotti, A. and Petrella, L. (2012) Skew Mixture Models for Loss Distributions: A Bayesian Approach. *Insurance*: *Mathematics and Economics*, **51**, 617-623. https://doi.org/10.1016/j.insmatheco.2012.08.002

[22] Miljkovic, T. and Grün, B. (2016) Modeling Loss Data Using Mixtures of Distributions. *Insurance*: *Mathematics and Economics*, **70**, 387-396. https://doi.org/10.1016/j.insmatheco.2016.06.019

[23] Gui, W., Huang, R. and Lin, X.S. (2018) Fitting the Erlang Mixture Model to Data via a GEM-CMM Algorithm. *Journal of Computational and Applied Mathematics*, **343**, 189-205. https://doi.org/10.1016/j.cam.2018.04.032

[24] Fung, T.C., Badescu, A.L. and Lin, X.S. (2019) A Class of Mixture of Experts Models for General Insurance: Theoretical Developments. *Insurance*: *Mathematics and Economics*, **89**, 111-127. https://doi.org/10.1016/j.insmatheco.2019.09.007

[25] Fung, T.C., Badescu, A.L. and Lin, X.S. (2019) A Class of Mixture of Experts Models for General Insurance: Application to Correlated Claim Frequencies. *ASTIN Bulletin*: *The Journal of the IAA*, **49**, 647-688. https://doi.org/10.1017/asb.2019.25

[26] Sun, M. and Lu, Y. (2022) A Generalized Linear Mixed Model for Data Breaches and Its Application in Cyber Insurance. *Risks*, **10**, Article 224. https://doi.org/10.3390/risks10120224

[27] Bell, G. and Ebert, M. (2015) Health Care and Cyber Security: Increasing Threats Require Increased Capabilities. KPMG.

[28] Williams, P.A.H. and Woodward, A.J. (2015) Cybersecurity Vulnerabilities in Medical Devices: A Complex Environment and Multifaceted Problem. *Medical Devices: Evidence and Research*, **8**, 305-316. https://doi.org/10.2147/MDER.S50048

[29] Roberts, S.J. (1997) Parametric and Non-Parametric Unsupervised Cluster Analysis. *Pattern Recognition*, **30**, 261-272. https://doi.org/10.1016/S0031-3203(96)00079-9

[30] Likas, A., Vlassis, N. and Verbeek, J.J. (2003) The Global *k*-Means Clustering Algorithm. *Pattern Recognition*, **36**, 451-461. https://doi.org/10.1016/S0031-3203(02)00060-2

[31] Johnson, S.C. (1967) Hierarchical Clustering Schemes. *Psychometrika*, **32**, 241-254. https://doi.org/10.1007/BF02289588

[32] Bholowalia, P. and Kumar, A. (2014) EBK-Means: A Clustering Technique Based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, **105**, 17-24.

[33] Agrawal, R., Faloutsos, C. and Swami, A. (1993) Efficient Similarity Search in Sequence Databases. In: Lomet, D.B., Eds., *FODO* 1993: *Foundations of Data Organization and Algorithms*, *Lecture Notes in Computer Science*, Vol. 730, Springer, Berlin, 69-84. https://doi.org/10.1007/3-540-57301-1_5

[34] Peel, D. and MacLahlan, G. (2000) Finite Mixture Models. John & Sons, Hoboken. https://doi.org/10.1002/0471721182

[35] Hasselblad, V. (1969) Estimation of Finite Mixtures of Distributions from the Exponential Family. *Journal of the American Statistical Association*, **64**, 1459-1471. https://doi.org/10.1080/01621459.1969.10501071

[36] Yee, T.W. (2015) Vector Generalized Linear and Additive Models: With an Implementation in R. Springer, Berlin.

[37] Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society*: *Series A* (*General*), **135**, 370-384. https://doi.org/10.2307/2344614

[38] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B* (*Methodological*), **39**, 1-22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[39] McCullagh, P. and Nelder, J.A. (2019) Generalized Linear Models. Routledge, New York. https://doi.org/10.1201/9780203753736

[40] Gan, G. and Valdez, E.A. (2018) Regression Modeling for the Valuation of Large Variable Annuity Portfolios. *North American Actuarial Journal*, **22**, 40-54. https://doi.org/10.1080/10920277.2017.1366863

[41] Bühlmann, H. (2007) Mathematical Methods in Risk Theory. Springer Science & Business Media, Berlin.

[42] Jacobs, J. (2014) Analyzing Ponemon Cost of Data Breach. *Data Driven Security*, **11**, 5.

[43] Financial Services Varonis (2021) 2021 Financial Services Data Risk Report.

[44] Marsh McLennan (2021) Cyber Insurance Market Overview: Fourth Quarter 2021. Marsh Cyber Risk Report. https://www.marsh.com/us/services/cyber-risk/insights/cyber-insurance-market-overview-q4-2021.html

Appendix

A. Geographical Location Clusters

Cluster Label	Number of Observations	States
1	3024	CT, DE, DC, IL, IA, ME, MD, MA, MI, NE, NH, NJ, NY, OH, PA, RI, SD, VT, WI, WY
2	531	AK, MN, MT, ND, OR, WA
3	1642	AL, AR, FL, GA, ID, LA, MS, OK, SC, TX
4	1093	IN, KS, KY, MO, NC, TN, VA, WV
5	1805	AZ, CA, CO, HI, NV, NM, UT