The role of covariates in cyber risk ratemaking using GAMLSS

Alana K. Azevedo[1], Agnieszka I. Bergel[2], Alfredo D. Egídio dos Reis[2],
[1]*ISEG-School of Economics and Management, Universidade de Lisboa; REM - Research in Economics and Mathematics, CEMAPRE and Universidade Federal do Ceará*
alanakna@phd.iseg.ulisboa.pt
[2]*ISEG-School of Economics and Management, Universidade de Lisboa; REM - Research in Economics and Mathematics, CEMAPRE*
alfredo@iseg.ulisboa.pt, agnieszka@iseg.ulisboa.pt

Abstract

Relying on the SAS OpRisk Global Data which is the world's largest collection of publicly reported operational losses, we performed a detailed analysis of the frequency and severity of cyber risk claims with the intention of generating information that helps in the ratemaking process for insurance of this type of risk. We developed two types of analysis, the first considering the Loss Distribution Approach (LDA) and the second using Generalized Additive Models for Location, Scale and Shape (GAMLSS). For both approaches, we adjusted two models, one for the frequency and another for the severity of claims. Additionally, through GAMLSS, we could estimate the coefficients that compose the calculation model for a risk premium, analysing intensity, the covariates effect, allowing to generate *a priori* estimates for the premium to be calculated for each policy based on the individual risk profile of the insured. Our work shows how much the introduction of covariates can increase the financial need to be charged as well as how much the premium value changes depending on the risk class. Insurance premium estimates may become too high leading to disinterest in both parties, for the insurer in accepting such a risk and on the policyholder due to the high cost.

Keywords: Cyber risk; ratemaking; LDA; GAMLSS.

1 Introduction and aims

Cyber risks, with other sorts of operational risks, have become a recurrent topic when it comes to proper management. As this is a challenge in relation to the risk classification and prediction of loss amounts, there is an interest in studying techniques that can satisfactorily manage this risk category. Cyber attacks have become more and more frequent both at the individual level and in organizations, especially with the increase in digital transactions. Cyber risk has commonly been related to a risk category that comprises operational risks, mainly after Basel II regulatory requirements, see Gleeson (2010).

Operational risks can become challenging. Cyber risks are linked to these as they involve losses resulting from human errors or deficiencies in systems or controls. Cyber risks also resemble operational risks in terms of their attributes. According to Karam (2014), operational risk is driven by high-frequency low impact events, constituting the shape of the distribution and refering to expected losses; and cyber risk is driven by low-frequency high-impact events, constituting the tail of the distribution and referings to unexpected losses.

One way of analyzing losses due to cyber risks is through the Extreme Value Theory, as it has the property of focusing its analysis only on the tail area, dealing with large losses. This fact is confirmed in this work by concluding that the best adjustment for the severity of losses is the Weibull distribution. Another feature that characterizes cyber risks and makes their management even more challenging is the dispersion of data. Measures such as asymmetry and kurtosis of a probability distribution may improve estimate regression models that can help in the ratemaking process.

One way to input dispersion measures in the cyber risk analysis process is through the Generalised Additive Models for Location, Scale and Shape (briefly GAMLSS). It is a general framework for fitting regression type models that include highly skew and kurtotic continuous and discrete distributions. This regression model allows us identify factors and key risk drivers that may influence aggregate loss. For Malavasi *et al.* (2022) this is a powerful framework in better understanding the factors that are influential in the central moments of studied loss processes.

What sets GAMLSS apart from other regression models is the fact that the assumed distribution can be any parametric distribution not only exponential, and that all the parameters of the distribution can be modelled as functions of the explanatory variables. These models make it possible to identify regression factors that could influence the frequency and severity of processes in the mean, variance, skewness and kurtosis.

Recently, GAMLSS has been applied in various fields, including actuarial science in some of its research areas. Bonus-Malus systems (Tzougas *et al.*, 2015), operational risk (Hambuckers *et al.*, 2018), crash data (Zou *et al.*, 2011), etc. Regarding cyber risk, the use of GAMLSS in cyber risk analysis is scarce. We can cite only Malavasi *et al.* (2022) as a starting point in the joint research of cyber risk and the use of GAMLSS.

We propose an analysis of cyber risk losses, introducing a framework of the GAMLSS which can model not only the mean but all other parameters. Our main goal is to seek for an actuarial model for the coverage of cyber risks losses using all available information in the estimation of aggregate loss distribution. We intend to identify significant risk classification variables, determine tariff classes and calculate premiums considering an *a priori* model.

We shed new light on the cyber risk topic, by estimating insurance pure premium tables as a function of past incidents (prior ratemaking) and the significant variables. To do this, we use a dataset of a world collection of publicly reported operational losses to model frequency and severity of claims and performe a fit considering first the Loss Distribution Approach (briefly LDA), a statistical approach for computing aggregate loss distributions, without specifying covariates in order to compare to the results of the GAMLSS.

The manuscript is organized as follows. In the next section we survey related work. Section 3 is devoted to the presentation of the database, including description and descritive statistics. Section 4 brings frequency and severity modelling through the LDA) Section 5 present the application of a GAMLSS in cyber risk ratemaking. It also includes definitions and discussion. Last section ends our study by setting some concluding remarks.

2 Background and motivation

The biggest challenge in terms of cyber risk is the management and an efficient modeling by bringing for organizations and insurers the benefit of making a healthy contractual relationship. For organizations, the objective is the protection against financial losses from cyber attacks, for insurers the guarantee of solvency despite the occurrence of claims.

For this, the correct classification and analysis of the costs of this type of risk becomes essential for the implementation of a pricing methodology. Romanosky (2016) examined the composition and costs of cyber events, and verified existence of incentives for firms to improve their security practices in order to reduce the risk of attack. This results suggested that there is an excessive concern about increasing violations, contrasting with a relatively modest financial impact for companies that suffer from this type of event.

Eling and Wirfs (2016), considering an operational risk database and using the peaks-over-threshold method from extreme value theory, analyzed cyber incidents identifying "cyber risks of daily life" and "extreme cyber risks". These authors showed that human behavior is the main source of cyber risk and that cyber risks are very different compared with other risk categories. Peters *et al.* (2018) explored the different approaches adopted by cyber research and cyber crime agencies to classify cyber crime loss event types. The manuscript presented the emerging market of cyber risk insurance and the challenges faced by this market, discussing regulator and industry responses to cyber risk management, mitigation and insurance.

In case of cyber risk management, Carfora *et al.* (2019) pointed out the peculiarities of cyber insurance contracts with respect to the classical non life insurance. The author analyzed data breaches, examining suitable distributions to represent the frequency and the severity of the reported cyber incidents. Mukhopadhyay *et al.* (2013) proposed, through a Copula-aided Bayesian Belief Network (CBBN) for cybervulnerability assessment (C-VA) and the computation of the expected loss, a utility based preferential pricing (UBPP) model to compute the premium that a cyber risk insurer can charge to indemnify cyber losses.

Peng *et al.* (2018) initiated the study of modeling multivariate cybersecurity risks developing the first statistical approach, which was centered at a Copula-GARCH model that uses vine copulas to model the multivariate dependence exhibited by real-world cyber attack data. The authors found that ignoring the due multivariate dependence can cause a severe underestimation of cybersecurity risks and that the proposed approach leads to accurate predictions of multivariate cybersecurity risks.

Concerning Cyber insurance and its market, Böhme and Schwartz (2010) proposed a comprehensive formal framework to classify all market models of cyber-insurance by taking into account specific properties of cyber risk such as interdependent security, correlated risk, and information asymmetries.

Barracchini *et al.* (2014) suggested two hypothesized typologies of computer policies such as activities of daily cyber and good cyber use. The paper formulated the actuarial premises for coverage in case of cyber damage. Biener *et al.* (2015) investigated the insurability of cyber risk by analysing 994 cases of cyber losses from an operational risk database. The results exposed the distinct characteristics of cyber risks compared to other operational risks.

Romanosky *et al.* (2019) performed a qualitative research for examining cyber insurance policies describing what losses are covered by cyber insurance policies, what questions do insurance companies pose to applicants in order to assess risk and how are cyber insurance premiums determined. Yang and Lui (2014) contributed in providing a fundamental understanding on how "network externality" with "node heterogeneity" may affect security adoption. The authors also studied cyber insurance and how the presence of a competitive insurance market can affect the security adoption.

As this is a very challenging risk for the insurance market, the use of methodologies that can better assess cyber risk becomes essential. GAMLSS methods has been used in several areas of study, including insurance. First, Stasinopoulos and Rigby (2008) introduced a tutorial of the generalized additive models for location, scale and shape for performing regression analysis. Hambuckers *et al.* (2018) presented a GAMLSS to describe the dynamics of operational losses, and to correctly estimate tail-related risk indicators.

Zou *et al.* (2011) developed a Sichel GAMLSS for modeling highly dispersed crash data. The authors used several goodness-of-fit metrics to attest that the Sichel GAMLSS may offer a viable alternative to the traditionally GLMs for analyzing highly dispersed crash datasets. In our study, we considered a Weibull-generalized Poisson GAMLSS to develop a cyber risk analysis. The use of this distribution for cyber risk studies is scarce. One application, however not about cyber risk, of this distribution was brought by Gupta and Huang (2014) that demonstrate the usefulness of the Weibull-generalized Poisson model in the analysis of survival data.

Tzougas *et al.* (2015) presented the design of optimal Bonus-malus Systems using GAMLSS showing that the employment of advanced models can provide a measure of uncertainty regarding the credibility updates of claim frequency/severity of each specific risk class. In another case, Ganegoda and Evans (2013) modeled the severity of operational losses by using a regression analysis based on the GAMLSS framework to model the scaling properties of these losses.

Pitt *et al.* (2020) investigated a GAMLSS framework for estimating the frequency and severity of losses associated with catastrophic risks finding a superior fit to empirical loss data in comparison to generalized linear regression models. Emmanuel *et al.* (2021) deployed a GAMLSS to model a typical automobile insurance portfolio.

Malavasi *et al.* (2022) investigated the viability of insurance for cyber risk using a utility modelling framework with premium calculated by classical certainty equivalence analysis utilising GAMLSS and a class of ordinal regressions.

We introduce a new perspective to the study of cyber risk pricing with GAMLSS. The particular strengths and differential of the current study, are: (i) Within the framework of the GAMLSS, the use of risk classes in order to compare with the LDA ratemaking process to check the differences in tariff values, and (ii) The use of real data from a world collection of publicly reported operational losses.

3 Cyber risk data description

For our empirical analysis of cyber risk we rely on the SAS OpRisk Global Data which is the world's largest collection of publicly reported operational losses. The database gives information about 37,429 incidents of operational loss in the period between March 1971 and January 2021. For each incident, besides the amount of the loss, the database reports the description of the event, the business lines and industry sectors, the risk category, country of incident (that could be all over the world) and other informations about the firms involved.

All losses, given in US$, are presented in current value for proper comparison.

Regulators of insurance and financial markets categorize cyber risk as operational risk, to identify cyber risk in the database, we considered the categorization of CRO (2016) that enumerates the following, we quote:

- Any risks emanating from the use of electronic data and its transmission, including technology tools such as the internet and telecommunications networks;
- Physical damage that can be caused by cyber attacks;
- Fraud committed by misuse of data;
- Any liability arising from data use, storage and transfer;
- The availability, integrity and confidentiality of electronic information (be it related to individuals, companies or governments).

We decided to consider two subcategories for cyber risk: (1) Actions of people, and (2) Systems and technical failure. Considering information of the SAS database with complete records from 2004, a total of 974 cyber risk incidents were identified.

3.1 Descriptive statistics

To first understand our choice of considering data from the year 2004, Figure 1 illustrates that the number of cyber risk incidents was small before that year.

Figure 1: Cyber risk incidents

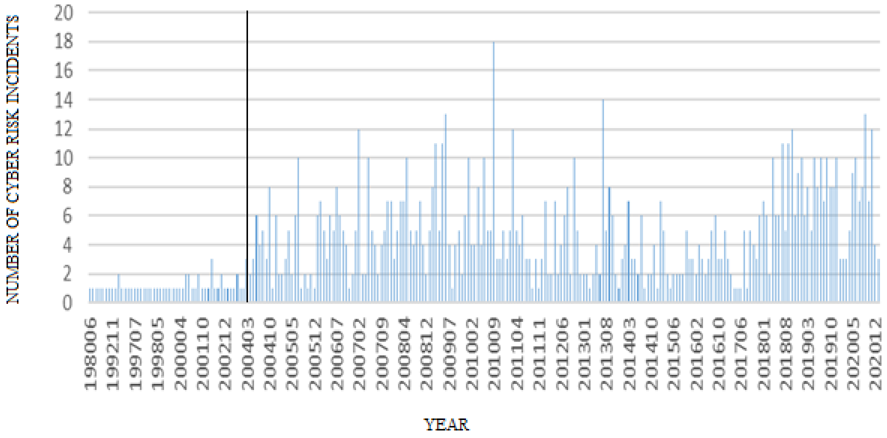


Table 3.1 provides a summary of the cyber risk sample from the year 2004 into 5 panels: panel A-D, with several statistics as number of incidents, total loss, average loss, standard deviation, median, skewness and kurtosis. Panel B of Table 3.1 shows that, in 89.6% of the cases, human behavior is the main source of cyber risk incidents. The average loss amount is quite different too. Losses by systems and technical failure are, in million US$, 21.71 greater than the total average loss amount.

Table 3.1: Cyber risk losses (in million US$) descriptive statistics

	Incidents	%	Total loss	Average loss	Std. dev.	Median	Skewness	Kurtosis
Panel A: Total sample								
Total	974	100%	21312.87	21.88	86.11	1.62	9.05	104.68
Panel B: Subcategories								
Actions of people	873	89.6%	16910.45	19.37	84.96	1.44	10.02	122.21
Systems and technical failure	101	10.4%	4402.43	43.59	93.07	6.09	3.25	11.38
Panel C: Geographic Region								
Asia	135	13.9%	5050.51	37.41	158.97	1.13	6.76	49.10
Europe	241	24.7%	5144.00	21.34	62.56	1.92	4.62	22.97
North America	522	53.6%	10519.17	20.15	73.52	1.71	7.84	73.23
Other	76	7.8%	599.19	7.88	17.63	1.29	3.37	11.38
Panel D: Industry								
Non-financial	291	29.9%	11643.82	40.01	136.91	4.43	6.52	48.66
Financial	683	70.1%	9669.05	14.16	49.06	1.06	6.61	53.47
Panel E: Company size by number of employees								
Small	325	33.4%	7934.74	24.41	117.00	1.69	8.66	82.09
Medium	325	33.4%	6203.58	19.09	60.62	1.50	6.78	61.41
Large	324	33.2%	7174.55	22.14	70.02	1.82	5.75	40.84

Note: Size-classification is based on the lower, middle and upper 33% quantiles of number of employees; Small ($\leq$ 7,000 employees); Medium (between 7,000 and 57,129 employees); Large ($\geq$ 57,129 employees).

Concerning the geographic region, Panel C shows that firms of North America presents more than half of the incidents (53.6%). Europe comes second with 24.7% of cyber risk incidents. Despite a higher incidence, North America has a lower average value of losses than Asia and Europe. In a similar analysis, Biener *et al.* (2015) suggests that North American firms are more capable of and willing to invest in risk mitigating measures for extreme losses.

In Panel D of Table 3.1 we can see that finantial service industries hold 70.1% of the total cyber risk incidents. This can be explained by the type of information that are managed by those firms. The more valuable the information, the greater the risk and the greater the awareness of exposure to certain risks, then the better should be the mitigation of those risks. This could explain why even with more incidents the average of loss amount for financial service firms is much smaller than the average for non-financial firms.

Separating the firms by size based on quantiles, Panel E of Table 3.1, the loss severity is similar for each category. The next step of our work is the analisys of the distributions for the loss frequency and severity.

4 Ratemaking through the Loss Distribution Approach (LDA)

To perform an analisys about loss frequency and severity, we consider the Loss Distribution Approach (LDA), it is a parametric technique primarily based on historic observed loss data. Karam (2014) explains that LDA consists of separately estimating a frequency distribution for the occurrence of losses and a severity distribution for the economic impact of the losses. After established these two distributions we combine both to obtain an aggregate loss distribution. Goodness-of-fit tests were applied in order to identify the best models. Figure 2 and Table 4.1 presents a preliminar analisys showing some descritive statistics of loss frequency and loss severity, measured monthly.

Figure 2: Claim frequency and severity histograms and box-plots

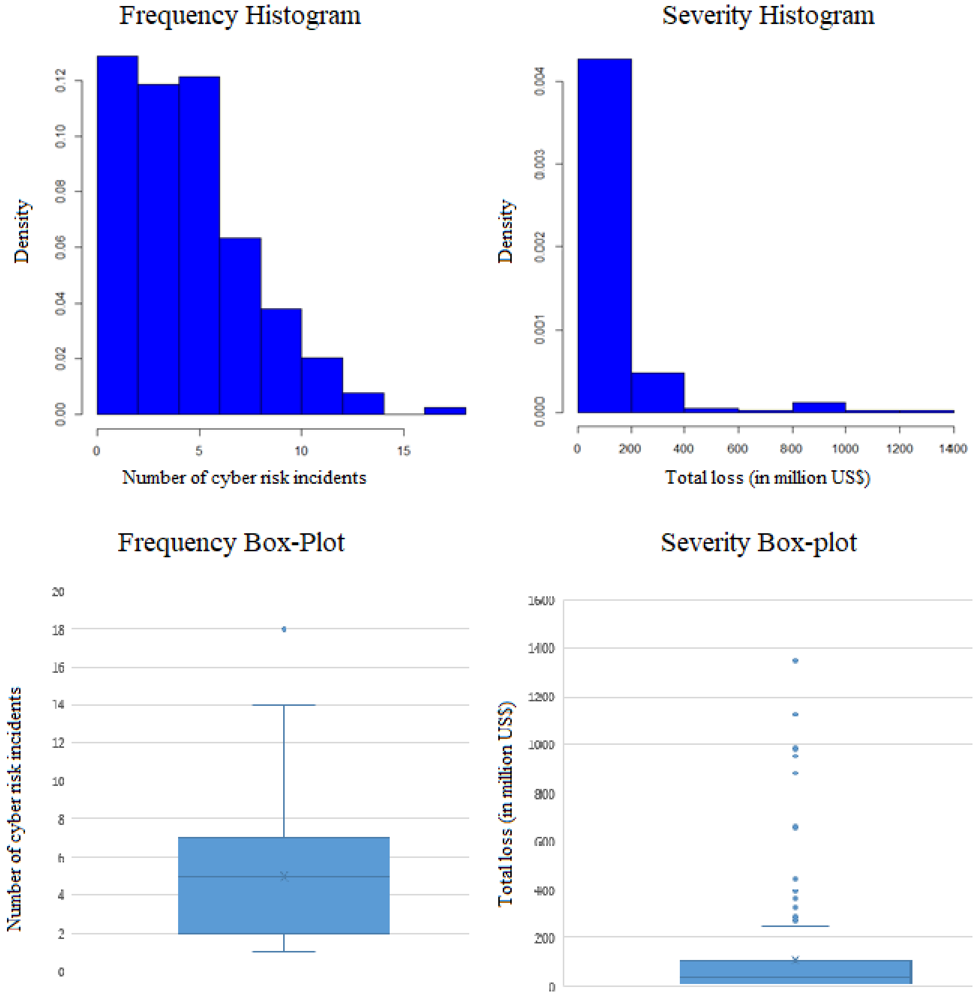


Table 4.1: Claim frequency and severity summary statistics

Statistic	Frequency	Severity
Min.	1	0.137
1st Qu.	2	10.638
Median	5	34.262
Mean	4.985	107.781
3rd Qu.	7	107.337
Max.	18	1346.444
Estimated sd.	3.152	204.438
Estimated skewness	1.016	3.694
Estimated kurtosis	3.975	18.040

To model the claim frequency we denoted N as the number of events over the time period.

Table 4.2: Claim frequency goodness-of-fit

Distribution	df	AIC
Generalised Poisson	2	974.26
Negative Binomial	2	974.90
Double Poisson	2	986.79
Zero Adjusted Poisson	2	1025.76
Zero Inflated Poisson	2	1028.53

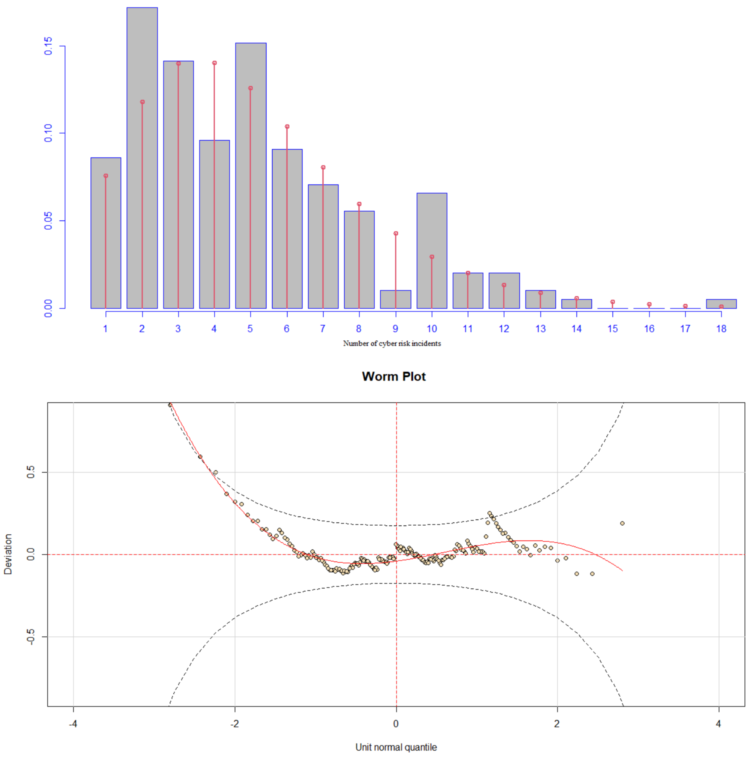
Our intention in a goodness-of-fit test was to find the best distribution to represent the claims frequency with at most two parameters due to the fact that we wanted an aggregate claims distribution with a maximum of four parameters. Several studies that use count data restrict the analysis considering only the Poisson or the Negative Binomial distributions. Our interest was to expand into new possibilities. To do that we considered the Akaike Information Criterion (AIC), a method that allows comparing models with different families of distributions and that does not need further inferences about the model to corroborate its result (Burnham and Anderson, 2004). The best model is the one with the lowest AIC value. The AIC values of Table 4.2 indicate that for the loss frequency, measured monthly, the Generalized Poisson distribution (GP) provides better fit than the Poisson and the Negative Binomial distributions.

The probability function of the GP distribution can be written as:

$$P(n;\mu,\upsilon) = \frac{\mu(\mu+\upsilon n)^{n-1}e^{-\mu-\upsilon n}}{(1-e^{-\mu})n!}, \tag{4.1}$$

where $n \in \mathbb{N}$, $\mu \in \mathbb{R}^+$, $max(-1,-\mu/m) \leq \upsilon \leq 1$ and $m(\geq 4)$ is the largest positive integer for which $\mu + m\upsilon > 0$.

Figure 3: Claim frequency GP fit



The GP distribution when compared with the Poisson distribution has one more parameter that identifies the underdispersion as well as overdispersion of the data, see (Rigby and Stasinopoulos, 2005, 317-320). The worm plot in Figure 3 gives a diagnostic regarding the residuals. The closer the points representing deviations of the residuals to the centerline, the closer the distribution of residuals is to a standard normal distribution. If the model is correct then approximately 95% of the points will be within the two elliptic curves that denote the confidence interval (Buuren and Fredriks, 2001). Points far outside the limits indicate outliers. In terms of claim frequency, the randomised quantile residuals, mean and variance, were 0.0174 and 0.9631, respectively, which shows a good fit for the GP distribution since such values are similar to the values of a standard normal distribution.

The parameters in Table 4.3 were estimated using the maximum likelihood method (briefly ML). This information is essential for calculating the expected number of claims, denoted by $E(N)$, the first moment of the GP distribution.

Table 4.3: Result of the fitted GP Distribution

	Estimate	Std.Error	t value	Pr($>|t|$)
μ	1.6064	0.0444	36.1712	$<$ 2.22e-16 ***
υ	-2.5347	0.1870	-13.5521	$<$ 2.22e-16 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 ¨ 0.1 ˙ 1

$$E(N) = \frac{\mu}{1-\upsilon}. \tag{4.2}$$

To model claim severity, denoted by X, with X_i meaning the i-th severity, it is necessary to assume that all losses are positive and considered independent and identically distributed random variables from a continuous distribution. We also considered a skewness-kurtosis analysis, ordering distributions by assessing the heaviness of their tails. As we considered a distribution with two parameters for the frequency of claims, in order to have a distribution for aggregate claims with a maximum of four parameters, the choice of the best distribution for the severity adjustment must also have two parameters max. The Weibull distribution provides the best results for cyber losses severity. This distribution is part of a set of distributions that are used to model extreme values, which is quite pertinent when dealing with cyber risks.

Table 4.4: Claim severity goodness-of-fit

Distribution	df	AIC
Weibull	2	2146.50
Generalised Pareto	2	2153.60
Log Normal	2	2154.04
Gamma	2	2161.42
Inverse Gaussian	2	2303.41

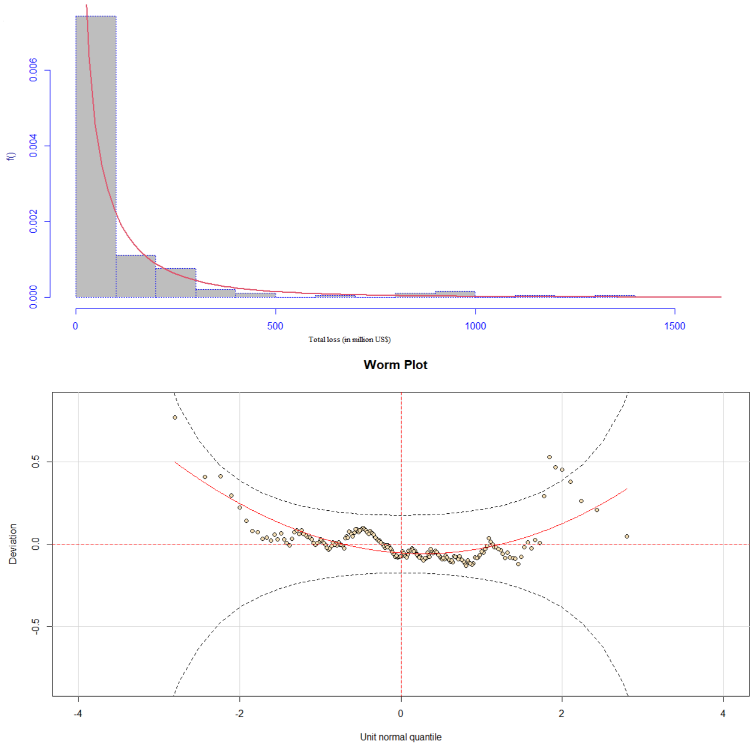
The AIC values of Table 4.4 indicate that for the loss severity, measured monthly, the Weibull distribution provides the better fit. Figure 4, upper graph, shows the plot obtained by the best model for the severity of claims along with the original data.

We may say that $\{X_i\}_{i=1}^{N}$ is Weibull distributed with scale parameter σ, shape parameter τ and density function

$$f(x;\sigma,\tau) = \tau\sigma^{\tau}x^{\tau-1}e^{-(\sigma x)^{\tau}}, x \geq 0, \tag{4.3}$$

where $\sigma \geq 0$ and $\tau \in \mathbb{R}$.

Figure 4: Claim severity fit



The randomised quantile residuals, mean and variance, were 0.0057 and 0.9514, respectively, which shows a good fit for the Weibull distribution. The parameters in Table 4.5 were estimated using the ML. Again, this information is essential for calculating the expected amount of claims, denoted by $E(X)$, the first moment of the Weibull distribution.

Table 4.5: Result of the fitted Weibull Distribution

	Estimate	Std.Error	t value	Pr($>|t|$)
σ	4.2683	0.1217	35.0776	$<$ 2.22e-16 ***
τ	-0.4829	0.0537	-8.9934	$<$ 2.22e-16 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 ˙˙ 0.1 ˙ 1

$$E(X) = \sigma\Gamma\left(1+\frac{1}{\tau}\right) \tag{4.4}$$

By the classic actuarial method of premium calculation, the net premium is given by the expected value of the insurer's payout $E(S)$. The random variable S represents the aggregate amount of claims arising from cyber risk. Let N denote the number of claims produced by a portfolio of policies in a time period, X_i is the amount of the ith claim. Then

$$S = X_1 + X_2 + ... + X_N \tag{4.5}$$

represents the aggregate claims generated by the portfolio for the period under study. Suppose that $\{X_i\}_{i=1}^{\infty}$ is a reference of identically distributed random variables and that N and $\{X_i\}_{i=1}^{\infty}$ are independent.

The distribution of the random variable S is a compound distribution, considering the claim severity X and the claim frequency N. The expected claim amount and variance, $E(S)$ and $Var(S)$ respectively, assuming that X and N are independent, are given by:

$$E(S) = E(N)E(X). \tag{4.6}$$

$$Var(S) = E(N)Var(X) + Var(N)(E(X))^2. \tag{4.7}$$

Considering the first moments of the distributions here presented (GP and Weibull), the calculation below quantifies the estimated net premium as 26.77 (in million US$).

$$E(N) = \frac{\mu}{1-\upsilon} = \frac{1.6064}{1-(-2.5347)} = 0.4545. \tag{4.8}$$

$$E(X) = \sigma\Gamma\left(1+\frac{1}{\tau}\right) = 4.2683\Gamma\left(1+\frac{1}{-0.4829}\right) = 58.9004 \tag{4.9}$$

$$E(S) = (0.4545)(58.9004) = 26.7702 \tag{4.10}$$

In next subsection we formally present a new distribution for analyzing aggregate claims.

4.1 The Weibull-generalized Poisson Distribution

After defining the distributions for frequency and severity, we can model aggregate claims. When a GP distribution is chosen for N, S is compound generalized Poisson distributed.

The Weibull-generalized Poisson distribution (WGP) is a compound generalized Poisson (CGP) where the mixing distribution of the generalized Poisson (GP) rate is a Weibull distribution. The cumulative distribution of S, developed in the work of Gupta and Huang (2014), can be written as:

$$F(s;\theta) = \frac{1-e^{-(\mu/\upsilon)W(g)}e^{-\mu}}{1-e^{-\mu}}, s > 0, \tag{4.11}$$

where $W(g)$ is the Lambert W function defined as $W(g)e^{W(g)} = g$ and $g = -\upsilon e^{-\upsilon-(\sigma s)^{\tau}}$. S is a random variable having the WGP distribution with parameter space $\theta = \theta(\mu,\, \sigma,\, \upsilon,\, \tau)$. The density function is given by

$$f(s;\theta) = \frac{\lambda\tau\sigma^{\tau}s^{\tau-1}e^{-(\mu/\upsilon)W(g)-\mu}(-W(g))}{(1-e^{-\lambda})(1+W(g))\upsilon}, s > 0. \tag{4.12}$$

Both cumulative and density functions are expressed in terms of the Lambert W function. The formal solution of the Lambert W function can be expressed as

$$f(g;\theta) \approx ln\left(\frac{g}{ln\left(\frac{g}{ln\left(\frac{g}{\cdots}\right)}\right)}\right) \tag{4.13}$$

This requires an infinite iterative calculus which is not easy to compute. Vazquez-Leal *et al.* (2019) define the Lambert W function as a multibranched function. $W_0(g)$ and $W_{-1}(g)$ are the branches considering only real numbers of g. $W_0(g)$ is called the upper branch (also named the principal branch) and satisfies the condition $W(g) \geq -1$. $W_{-1}(g)$ is called the lower branch and satisfies the condition $W(g) \leq -1$. Boyd (1998) developed analytical approximations to $W_0(g)$ for all $g \geq -e^{-1}$. In our work, we considered the following approximations, taken from Vazquez-Leal *et al.* (2019)

$$W_0^{Boyd} = W_0^B\left(1 + \frac{[ln(y) - (7/5)]e^{[-(3/40)(ln(y)-(7/5))^2]}}{10}\right), \tag{4.14}$$

where

$$y = 1 + e^1 g \tag{4.15}$$

and

$$W_0^B = \tanh\left(\frac{2y^{1/2}}{ln(10) - ln[ln(10)]}\right)[ln(y+10) - ln(ln(y+10))]. \tag{4.16}$$

The k-th moment of the WGP distribution is given by

$$E(S^k) = \frac{\Gamma(1+k/\tau)}{\sigma^k}\sum_{n=1}^{\infty} n^{-k/\tau} P_N(N=n), \tag{4.17}$$

where $P_N(.)$ is the probability function of zero-truncated GP distribution, see Gupta and Huang (2014).

5 Ratemaking through Generalized Additive Models for Location, Scale and Shape (GAMLSS)

First introduced by Rigby and Stasinopoulos (2005), GAMLSS have been applied in several fields, including actuarial science. It is a framework that consider a single response variable, allowing many explanatory valuables. The dependence of the response variable in relation to the explanatory variables could be linear, non-linear parametric function or non-parametric smoothing functions.

Stasinopoulos and Rigby (2008) explain that GAMLSS allows modeling not only the mean but all other parameters (including dispersion parameter) of the distribution of the response variable as linear and/or nonlinear and/or additive non-parametric functions of explanatory variables.

Another feature that differs this model from linear models (LM), generalized linear models (GLM) and generalized additive models(GAM) is that the assumed distribution of the response variable can belong to any parametric distribution, not just to the exponential dispersion family.

The formulation presented here was defined by Rigby and Stasinopoulos (2005) and assumes that, for $i = 1, 2, \ldots, n$, independent observations Y_i have probability density function $f_Y(y_i \mid \mu_i, \sigma_i, \upsilon_i, \tau_i)$ conditional on up to four distribution parameters, each of which can be a function of the explanatory variables. The first two population distribution parameters μ_i and σ_i are usually characterized as location and scale parameters, and υ_i and τ_i are usually characterized as shape parameters, for example, skewness and kurtosis, respectively.

Response variable observations $Y_1, Y_2, \ldots, Y_n$ are independent with $Y_i \sim D(\mu_i, \sigma_i, \upsilon_i, \tau_i)$ for $i = 1, \ldots, n$, where D is any distribution with (up to) four distribution parameters. For $k = 1, 2, 3, 4$, let $g_k(.)$ be a known monotonic link function relating a distribution parameter to a predictor η_k, where

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} s_{jk}(x_{jk}) \tag{5.1}$$

where X_k, with $k = 1, 2, 3, 4$, is a known design matrix, $\beta_k = (\beta_{k1}, \ldots, \beta_{kJ\prime_k})^T$ is a parameter vector of length J_k, s_{kj} is a smooth non-parametric function of variable X_{kj} and the x_{kj}'s are vectors of length n, for $k = 1, 2, 3, 4$ and $j = 1, \ldots, J_k$. A GAMLSS model allows the modelling of the parameters of the distribution as linear, that is, $X_k\beta_k$ or smooth term functions $s_{kj}(x_{kj})$ (k is the parameter index).

Next subsection shows the results of the application of the GAMLSS.

5.1 Modelling results for frequency and severity

Considering information of the SAS database with complete records, i.e. with availability of all the explanatory variables under consideration, we developed a GAMLSS model with an application to insurance ratemaking. There were 680 policies that met our criteria. This subsection describes the modelling results of the best fitted distributions/GAMLSS models that have been applied to model claim frequency and severity.

The a priori rating variables we employ are: the type of industry (T), the size of the company (S) and the geographic region (R). The variable T consists of two categories, those in: T_1 = Financial services and T_2 = Non-financial. The variable S consists of three categories, those of size: S_1 = Small, S_2 = Medium and S_3 = Large. The variable R consists of four categories, those in: R_1 = Asia, R_2 = Europe, R_3 = North America and R_4 = Other.

Table 5.1 presents the best fitted distribution/GAMLSS model for approximating the number of claims. The Zero Modified Logarithmic distribution (ZALG) GAMLSS model was chosen considering the ML estimators of the parameters associated with the condition of significance for all combinations of the covariates. We compared the fit of the GAMLSS models for the observed claim frequencies in the database employing AIC.

Table 5.1: Result of the fitted ZALG GAMLSS model

Variable	Coeff. β	Std.Error	t value	Pr($> \|t\|$)
Intercept	-0.7087	0.3555	-1.994	0.046582*
T				
T_1	0.7972	0.1623	4.913	1.13e-06***
S				
$S_1 + S_2$	-0.9349	0.1567	-5.968	3.88e-09***
R				
R_3	1.0340	0.2973	3.478	0.000538***
$R_2 + R_4$	0.4879	0.2854	1.709	0.087832 .
AIC	1078.05			

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

Table 5.1 presents the best fitted distribution/GAMLSS model for approximating the severity of claims. The Gamma (GA) GAMLSS model was chosen considering the ML estimators of the parameters associated with the condition of significance for all combinations of the covariates and the lowest AIC.

Table 5.2: Result of the fitted GA GAMLSS model

Variable	Coeff. β	Std.Error	t value	Pr($> \|t\|$)
Intercept	4.6303	0.2508	18.465	$<$ 2e-16***
T				
T_1	-0.7906	0.1505	-5.253	2.01e-07***
S				
$S_1 + S_2$	-0.4399	0.1698	-2.590	0.00980**
R				
R_3	-0.4934	0.2024	-2.438	0.01502*
$R_2 + R_4$	-0.5929	0.2177	-2.724	0.00663**
AIC	4764.63			

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

When we compare the results obtained here by the GAMLSS models for the frequency and severity of cyber risk claims with the traditional GLM modeling (Table 5.3), which requires a probability distribution that is a member of the exponential family of distributions for the response variable, we verify a superior fit to those obtained via GLM. The Akaike criterion together with the fact that in the GLM models not all explanatory variables were statistically significant, even when regrouping them in all possible ways, indicate that the GAMLSS models appear to be more appropriate. Table 5.4 compares the AIC values of the two approaches.

Table 5.3: Result of the fitted GLM models

	Frequency				Severity			
Variable	Coeff. β	Std.Error	t value	Pr($> \lvert t \rvert$)	Coeff. β	Std.Error	t value	Pr($> \lvert t \rvert$)
Intercept	1.41601	0.18654	7.591	1.06e-13***	75.715	14.338	5.281	1.74e-07***
T								
T_1	0.30412	0.11489	2.647	0.00831**	-28.997	8.831	-3.284	0.00108**
S								
$S_1 + S_2$	-0.51587	0.12707	-4.060	5.49e-05***	-10.913	9.767	-1.117	0.26424
R								
R_3	0.36581	0.15360	2.382	0.01752*	-18.563	11.806	-1.572	0.11635
$R_2 + R_4$	0.08671	0.16090	0.539	0.59011	-22.723	12.367	-1.837	0.06660 .
	AIC: 2390.40				AIC: 8295.50			

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

Table 5.4: GAMLSS and GLM models comparison

Model	Frequency AIC	Severity AIC
GAMLSS	1078.05	4764.63
GLM	2390.40	8295.50

5.2 Calculation of the A Priori Premiums

Table 5.5 contains a detailed description of the estimated coefficients for the adjusted GAMLSS, including the tariff relativities associated with each of the risk levels of the variables considered, relativities that express in which direction and in what intensity the statistical premium should be increased or smoothed.

The relativities were obtained using the inverse exponential function, considering that the link function used in the adjustment of the GAMLSS was logarithmic. Relativities are of fundamental importance within the tariff analysis as they allow the measurement of the risk of the other classes of a certain tariff variable in relation to the reference base class, see (Ohlsson and Johansson, 2010, 15-38). This measure aims to indicate the chance or the marginal effect of the risk observed in relation to the dependent variable when there are variations or changes in the behavior of the realizations of one of the independent variables.

Table 5.5: Coefficients, β, and relativities, $\exp(\beta)$, for estimated GAMLSS

		Frequency (ZALG)		Severity (GA)		A priori premium	
Risk factor	Level	β	$\exp(\beta)$	β	$\exp(\beta)$	β	$\exp(\beta)$
Intercept	-	-0.7087	0.4923	4.6303	102.5448	3.9216	50.4812
Industry	2	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
	1	0.7972	2.2193	-0.7906	0.4536	0.0066	1.0066
Size	3	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
	1+2	-0.9349	0.3926	-0.4399	0.6441	-1.3748	0.2529
Region	1	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
	3	1.0340	2.8123	-0.4934	0.6105	0.5406	1.7170
	2+4	0.4879	1.6289	-0.5929	0.5527	-0.1050	0.9003

For a better understanding of the analyzes below, it is important to highlight that a reference level was determined for each class of risk. Level 2 for the risk factor industry, level 3 for the risk factor size and level 1 for the risk factor region were chosen to be the reference base, so the values of relativities for these classes are equal to 1.

Regarding the estimated frequency relativities, it can be observed in the fourth column of Table 5.5 that, for example, the expected average number of claims per policy is higher for companies at tariff level 1 of the industry type variable, when compared to companies at level 2 of the same variable. When analyzing relativities, policies of risk level 1, in relation to those of base risk level 2, of the type of industry variable, is 2.2193. Thus, it is estimated that the average number of claims to be observed for level 1 is approximately 2.2193 times the number observed for level 2, or that the average number of claims observed for level 1 is, approximately 121.93% higher than the number observed for level 2. Similarly, the same analysis about relativity can be extended to the variable size and region.

In relation to the severity-adjusted GAMLSS, sixth column of Table 5.5, we observed that the average severity expected for level 1 companies of the industry type variable is lower than that expected for level 2 companies. Thus, it is estimated that the average severity of claims to be observed for level 1 is approximately 0.4536 times the severity observed for level 2, or that the severity of claims observed for level 1 is approximately 54.64% lower than the average severity observed for level 2.

The relativity of risk level 1 and 2 policies in relation to base risk level 3 policies, in the size variable, is 0.6441. Thus, it is estimated that the average severity of claims to be observed for levels 1 and 2 is approximately 0.6441 times the severity observed for level 3, or that the severity of claims observed for levels 1 and 2 is approximately 35.59% lower than the average severity observed for level 3. Analogously, the same analysis about relativity can be extended to the region variable.

It is also observed in the eighth column of Table 5.5, in relation to the estimated model for the premium, that the relativity for the risk level 1 policies, in relation to the base risk level 2, of the type of industry variable, is 1.0066. This implies that the insurance rate to be paid by level 1 policyholders, of the aforementioned variable, will be equivalent to approximately 1.0066 times the premium paid by level 2 policyholders, or, that the premium paid by level 1 policyholders will suffer a increase of approximately 0.66% in relation to the premium paid by level 2 policyholders.

The relativity of risk levels 1 and 2, in relation to the base risk level 3, of the size variable, is 0.2529. This implies that the premium to be paid by levels 1 and 2 policyholders, of the aforementioned variable, will be equivalent to approximately 0.2529 times the premium paid by level 3 policyholders, or even that the premium paid by levels 1 and 2 policyholders will be decreased by approximately 74.71% when compared to the premium paid by level 3 policyholders.

Finally, the relativity of risk levels 2 and 4, in relation to that of base risk level 1, of the region variable, is 0.9003. This implies that the premium to be paid by levels 2 and 4 policyholders must be equal to 0.9003 times the amount paid by level 1 policyholders, or that the premium paid by levels 2 and 4 policyholders will be reduced by approximately 9.97% in relation to the premium paid by level 1 policyholders.

Based on the use of the net premium calculation principle, we analyzed the premium for each of the 12 different risk classes, which are determined by the relevant a priori characteristics. To calculate the premium of any insured, given their individual risk profile, so that, if $N \sim ZALG(\mu;\sigma)$, and $X \sim GA(\alpha;\beta)$, one has that, without loss of generality, the pure premium for a given policy i can be calculated by:

$$P_{R_i} = E[S_i] = E[N_i]E[X_i] \tag{5.2}$$

Applying the transformation with the logarithmic link function, we have:

$$ln(P_{R_i}) = ln(E[S_i]) = \beta_0 + \beta_{11}X_{11} + \beta_{12}X_{12} + \beta_{21}X_{21} + \cdots + \beta_{34}X_{34} \tag{5.3}$$

where β_0 represents the intercept, the β_{ij} represent the estimated parameters for the observable GAMLSS.

Applying a transformation through the inverse exponential function, the expected value of the premium is expressed by:

$$e^{ln(P_{R_i})} = e^{ln(E[S_i])} = e^{\beta_0+\beta_{11}X_{11}+\beta_{12}X_{12}+\beta_{21}X_{21}+\cdots+\beta_{34}X_{34}} \tag{5.4}$$

Considering hypothetically an insured with a risk profile categorized by being a financial company, small or medium and located in North America and considering the estimated coefficients in the Table 5.5, seventh column, the pure premium would total (in million US$) as follows

$$P_{R_i} = e^{3.9216+0.0066-1.3748+0.5406} = 22.07. \tag{5.5}$$

The group with the lowest expected rate are those small and medium companies in non-financial services located in Europe and Other, with a rate of 11.49 (in million US$). On the other hand, the group with the highest expected rate are those large companies in financial services located in North America, with a rate of 87.25 (in million US$).

6 Remarks and conclusions

The insurance ratemaking process can have several approaches that can sometimes generate considerable differences in terms of tariffs. Our intention in this work was to carry out a comparison between the results generated by the Loss Distribution Approach (LDA) and by the Generalized Additive Models of Location, Scale and Shape (GAMLSS) in the treatment of cyber risk data.

In both approaches, frequency and severity of claims were treated separately. For the LDA, covariates were not considered, thus, the individual characteristics of each company were disregarded, generating a single tariff for the portfolio, 26.77 (in million US$). Regarding the adjustment of the distributions for frequency and severity, the best were the generalized Poisson distribution for the frequency and the Weibull distribution for the severity.

GAMLSS allows considering not only the mean but also other parameters of the frequency and severity distributions in the systematic part of the model, which allows for a more efficient fit of the data. The GAMLSS model generated results for 12 risk classes resulting from the combination of the considered covariates, these: type of industry (two possibilities), geographic region (four possibilities) and company size (three possibilities). The tariff values were between 11.49 and 87.25 (in million US$).

Additionally, it was possible to test individual and embedded hypotheses regarding the estimated parameters for the frequency and severity models of claims, analyze the influence and contribution of tariff factors to the analyzed models, identify and interpret the influence and impact of the parameters of the models on the dependent random variables.

In addition, we observed that, among the probability distributions tested in the fit of the models, the ones that provided a better fit to the data were the GAMLSS with ZIPF probability distribution and logarithmic link function for the frequency and Generalized Inverse Gaussian probability distribution and logarithmic link function for severity. Distributions other than those adjusted for the LDA.

Our detailed analysis of the frequency and severity of cyber risk considering two ways of approaching the ratemaking process for this type of risk showed how much the inclusion of covariates can increase the financial need to be charged as well as how much the premium value changes depending on of the risk class. According to our calculations, insurance premiums can become expensive, which could generate disinterest on the part of both insurers in accepting such a risk and policyholders due to the high cost.

Acknowledgements

The author was partially supported by the Project CEMAPRE/REM - UIDB/05069/2020 - financed by FCT/MCTES through national funds.

References

Barracchini, C., Addessi, M. E., *et al.* (2014). Cyber risk and insurance coverage: An actuarial multistate approach. *Review of Economics Finance*, 4:57–69.

Biener, C., Eling, M., and Wirfs, J. H. (2015). Insurability of cyber risk: An empirical analysis. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 40(1):131–158.

Böhme, R. and Schwartz, G. (2010). Modeling cyber-insurance: Towards a unifying framework. In *WEIS*.

Boyd, J. (1998). Global approximations to the principal real-valued branch of the lambert w-function. *Applied Mathematics Letters*, 11(6):27–31.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304.

Buuren, S. V. and Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, 20(8):1259–1277.

Carfora, M. F., Martinelli, F., Mercaldo, F., and Orlando, A. (2019). Cyber risk management: An actuarial point of view. *Journal of Operational Risk*, 14(4):77–103.

CRO (2016). Forum concept paper on a proposed categorisation methodology for cyber risk. `https://www.thecroforum.org/wp-content/uploads/2016/06/ZRH-16-09033-P1_CRO_Forum_Cyber-Risk_web-2.pdf`.

Eling, M. and Wirfs, J. H. (2016). Cyber risk is different. *Working Papers on Risk Management and Insurance*, (160).

Emmanuel, C. S., Osagie, A. M., and Obini, N. N. (2021). Mean parameter modeling for an automobile insurance portfolio using generalized additive models for location scale and shape (gamlss). *Science World Journal*, 16(3):287–290.

Ganegoda, A. and Evans, J. (2013). A scaling model for severity of operational losses using generalized additive models for location scale and shape (gamlss). *Annals of Actuarial Science*, 7(1):61–100.

Gleeson, S. (2010). International regulation of banking: Basel II: Capital and risk requirements. *OUP Catalogue.*

Gupta, R. C. and Huang, J. (2014). Analysis of survival data by a weibull-generalized poisson distribution. *Journal of Applied Statistics*, 41(7):1548–1564.

Hambuckers, J., Kneib, T., Langrock, R., and Silbersdorff, A. (2018). A markov-switching generalized additive model for compound poisson processes, with applications to operational loss models. *Quantitative Finance*, 18(10):1679–1698.

Karam, E. (2014). Measuring and managing operational risk in the insurance and banking sectors.

Malavasi, M., Peters, G. W., Shevchenko, P. V., Trück, S., Jang, J., and Sofronov, G. (2022). Cyber risk frequency, severity and insurance viability. *Insurance: Mathematics and Economics.*

Mukhopadhyay, A., Chatterjee, S., Saha, D., Mahanti, A., and Sadhukhan, S. K. (2013). Cyber-risk decision models: To insure it or not? *Decision Support Systems*, 56:11–26.

Ohlsson, E. and Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*, volume 174. Springer.

Peng, C., Xu, M., Xu, S., and Hu, T. (2018). Modeling multivariate cybersecurity risks. *Journal of Applied Statistics*, 45(15):2718–2740.

Peters, G., Shevchenko, P. V., and Cohen, R. (2018). Understanding cyber-risk and cyber-insurance. *Macquarie University Faculty of Business & Economics Research Paper.*

Pitt, D., Trück, S., van den Honert, R., and Wong, W. W. (2020). Modeling risks from natural hazards with generalized additive models for location scale and shape. *Journal of Environmental Management*, 275:111075.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Romanosky, S. (2016). Examining the costs and causes of cyber incidents. *Journal of Cybersecurity*, 2(2):121–135.

Romanosky, S., Ablon, L., Kuehn, A., and Jones, T. (2019). Content analysis of cyber insurance policies: How do carriers price cyber risk? *Journal of Cybersecurity*, 5(1):tyz002.

Stasinopoulos, D. M. and Rigby, R. A. (2008). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23:1–46.

Tzougas, G., Vrontos, S. D., and Fragos, N. (2015). Optimal bonus-malus systems using generalized additive models for location, scale and shape. *Scale and Shape (April 31, 2015)*.

Vazquez-Leal, H., Sandoval-Hernandez, M., Garcia-Gervacio, J., Herrera-May, A., and Filobello-Nino, U. (2019). Psem approximations for both branches of lambert function with applications. *Discrete Dynamics in Nature and Society*, 2019.

Yang, Z. and Lui, J. C. (2014). Security adoption and influence of cyber-insurance markets in heterogeneous networks. *Performance Evaluation*, 74:1–17.

Zou, Y., Lord, D., and Zhang, Y. (2011). Analyzing highly dispersed crash data using the sichel generalized additive models for location, scale and shape.