

CYBER RISK LOSS MODELLING IN CYBER INSURANCE

A Thesis
presented to
the Faculty of Natural and Applied Sciences
at Notre Dame University-Louaize

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Actuarial Science

by
MARIA GERGES AL GHANDOUR

JULY 2021

© COPYRIGHT

By

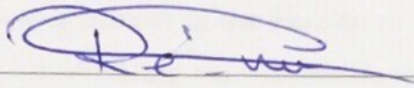
Maria Gerges Al Ghandour

2021

All Rights Reserved

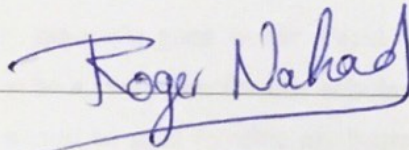
Dr. Re-Mi Hage

Associate Professor of Mathematics and Statistics, and advisor of
the Actuarial Sciences Department in the Faculty of Natural and
Applied Science at Notre Dame University Louaize



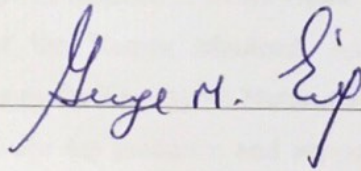
Dr. Roger Nakad,

Associate Professor of Mathematics and department Chairperson
in the Faculty of Natural and Applied Science at Notre Dame
University Louaize



Dr. George Eid

Dean of the Department of Mathematics and Statistics in the
Faculty of Natural and Applied Sciences at Notre Dame University
Louaize



Thesis Committee

Acknowledgement

First and foremost, I would be delighted to express my deep and sincere gratitude to my research supervisor, Dr. Re-Mi Hage, Associate Professor of Mathematics and Statistics, and advisor of the Actuarial Sciences Department in the Faculty of Natural and Applied Science at Notre Dame University Louaize, for providing me with invaluable guidance, help, patience, and continuous support throughout this research. Her dedication, sincerity and motivation have deeply inspired me. It was my great privilege and honor to work and study under her guidance and supervision. I am extremely grateful for her unconditional support she offered me.

I am grateful to Mr. Farid Chedid, Chairman and CEO of the Chedid Capital Holding group of companies, for giving me the opportunity to get the scholarship offered by Chedid Capital Holding to pursue my Masters in Sciences in Actuarial Sciences at Notre Dame University Louaize. My gratitude goes to Mr. Farid Chedid and his company. It was an honor and a privilege for me to be a candidate for this scholarship.

Furthermore, I would be glad to offer my heartfelt gratitude and appreciation to Dr. John Haddad and Ms. Claudia Freiji Bou Nassif, who have inspired me and have taught me the foundational knowledge and concepts that have supported this research. Their immense knowledge and plentiful experience have encouraged me in all throughout my academic research and daily life. Moreover, a special thanks to Dr. Marwan Gebran, Associate Professor in the Faculty of Natural and Applied Science at Notre Dame University, for his assistance and help.

In the memory of Dr. Ramez Maalouf, Associate Professor at the Department of Mathematics and Statistics at the Faculty of Natural and Applied Sciences, I would express my deep and sincere gratitude for his guidance and support offered to me during my Bachelor and my Master degree in Actuarial Sciences at Notre Dame University Louaize. May his soul rest in God's eternal love and peace.

My deep and sincere gratitude and thanks goes also to Dr. Roger Nakad, Associate Professor and Chairperson, and Dr. George Eid, Dean, of the Faculty of Natural and Applied Sciences at the Department of Mathematics and Statistics at Notre Dame University Louaize.

Finally, I am extremely blessed and grateful to my parents and family for their love, prayers, caring and sacrifices for educating and preparing me for my future. I thank God for all the blessings in my life

Abstract

In today's world, protecting information has become one of the most difficult tasks. Cyber security events and data breaches continue to be expensive events that affect people and businesses all around the world. A breach occurs when sensitive information is accessed. Moreover, cyber threats are constantly evolving in order to take advantage of online behavior and trends, especially when teleworking has become a necessity due to the global invasion and prevalence of the Coronavirus disease 2019 during the past two years. Therefore, the necessity for cyber insurance, which covers the liability for a cyber-breach, becomes more evident as more business activities are automated and an increasing number of computers are used to hold sensitive information. Unfortunately, research on cyber risk modeling has been fragmented and uncoordinated till date due to the lack of historical data available on cyber incidents which does not allow insurance premiums to be accurately priced, in addition to the constantly changing nature of cyber risk which makes the data easily become out-of-date. Hence, the aim of this thesis was the ratemaking of aggregate cyber loss. The VERIS dataset, one of the most extensive and publicly available datasets for global incident breaches, was used in this study. The main variables in the VERIS dataset are: type of breach, amount of a breach, timeline of the breach, Actors, Motive, Country, Variety, Assets, and Attributes.

Since the loss amounts are available in contrast to the loss frequency, we modeled, in this research, only the cyber risk severity, as a first step toward pricing cyber insurance coverage policies which require both the severity and the frequency distribution of cyber losses using the R programming language; R studio 4.0.3. First, the severity distribution was estimated using the loss distribution approach. Second, using machine learning, the Random Forest algorithm was applied to the data in order to select the most important variables that have the highest significant impact on cyber risk losses. Next, we applied the Generalized Linear Model using the most important variables selected by the Random Forest and the fitted distribution, in order to estimate the future loss amount. Last, we used the classical credibility theory to estimate the minimum number of observations required to reach 95% level of accuracy I modeling cyber risk.

Keywords: Cyber risk, Cyber security, Cyber insurance, Ratemaking, Loss Distribution Approach, Machine Learning, Random Forest, Generalized Linear Model, Classical credibility theory, R Studio.

Table of Contents

Acknowledgement	iii
Abstract	1
Table of Figures	4
Table of Tables	6
Chapter 1: Cyber Risk Insurance	9
1.1 Cyber Risk	9
1.1.1 Definition of Cyber Risk	9
1.1.2 Types of Cyber incidents and losses	9
1.2 Cyber Security	10
1.2.1 Definition of Cyber Security	10
1.2.2 Problems of Cyber Security	10
1.3 Cyber Insurance	11
1.3.1 Definition of Cyber Insurance	11
1.3.2 Contents of Cyber Insurance Policy	11
1.3.3 Types of Cyber Insurance coverage	12
1.3.4 Market challenges of Cyber insurance:	13
Chapter 2: Modeling of Cyber Risk	15
Chapter 3: Methodology	20
3.1 Fitting a distribution	20
3.1.1 Parametric methods for fitting a distribution	20
3.1.2 Goodness of fit test	21
3.1.3 Akaike information criterion (AIC)	22
3.2 Random Forest	23
3.2.1 Definition of Random forests	23
3.2.2 Selection Variable importance “VI (Xj)”	24
3.2.3 Influence of Parameters on Variable Importance	25
3.2.4 Radom Forest Algorithm	25
3.3 Generalized Linear Models	27
3.3.1 Components of Generalized Linear Models	27
3.3.2 Examples of Exponential Dispersion Models with their link function (EDM)	29
3.3.3 Generalized Linear Model Assumptions	29
3.3.4 Fisher scoring iteration	30

3.3.5	Diagnostics for Generalized Linear Models	31
3.3.6	Outliers and Influential Observations.....	32
3.3.7	Assessing the model.....	32
3.4	Credibility Theory.....	33
Chapter 4:	Modeling and prediction of Loss amount of Breaches.....	35
4.1	VERIS dataset and schema	35
4.1.1	Actors: whose actions affected the asset?.....	35
4.1.2	Actions: what actions affected the asset?.....	35
4.1.3	Assets: which assets were affected?	36
4.1.4	Attributes: how the asset was affected?	37
4.1.5	Incident timeline: The incident timeline is the timeline of events leading up to and following an incident.....	37
4.2	Descriptive Statistics	38
4.3	Fitting distributions	41
4.3.1	Introduction to Fitting distributions	41
4.3.2	Fitting distribution on all data set	42
4.3.3	Fitting distribution on the data set of each year.....	44
4.3.4	Conclusion	52
4.4	Radom Forest	54
4.4.1	Radom Forest on the overall data	54
4.4.2	Radom Forest on each year.....	56
4.5	Generalized Linear Model.....	62
4.5.1	Generalized Linear Model on the overall data.....	62
4.5.2	Generalized Linear Model on each year	68
4.6	Credibility Theory.....	79
4.6.1	Credibility theory on the overall data	80
4.6.2	Credibility theory on each year	80
Conclusion	83
References	84

Table of Figures

Figure 1: Summary of the research papers on modeling Cyber Risk	19
Figure 2: Percentage distribution for each Type of Breach	38
Figure 3: Total number of breaches for year 2013, 2014, 2015 and 2016.....	39
Figure 4: Q-Q plot of the overall impact loss amount of breaches.....	42
Figure 5: P-P plot of the overall impact loss amount of breaches.....	43
Figure 6: Probability density function of the overall Impact loss amount of a breach.....	44
Figure 7: Q-Q plot of the impact loss amount of breaches for year 2013.....	45
Figure 8: P-P plot of the impact loss amount of breaches for year 2013.....	46
Figure 9: Q-Q plot of the impact loss amount of breaches for year 2014.....	47
Figure 10: P-P plot of the impact loss amount of breaches for year 2014.....	48
Figure 11: Q-Q plot of the impact loss amount of breaches for year 2015.....	49
Figure 12: P-P plot of the impact loss amount of breaches for year 2015.....	50
Figure 13: Q-Q plot of the impact loss amount of breaches for year 2016.....	51
Figure 14: P-P plot of the impact loss amount of breaches for year 2016.....	52
Figure 15: Variable importance selection graph using the Random forests algorithm on the overall data.....	56
Figure 16: Variable importance selection graph using the Random Forest algorithm Year 2013.....	57
Figure 17: Variable importance selection graph using the Random Forest algorithm Year 2014.....	58
Figure 18: Variable importance selection graph using the Random Forest algorithm Year 2015.....	60
Figure 19: Variable importance selection graph using the Random Forest algorithm Year 2016.....	61
Figure 20: Pearson Residual Plot for Gaussian GLM applied on the overall data	66
Figure 21: Boxplots of the important variables used in fitting the Gaussian GLM to the overall data.....	67
Figure 22: Pearson Residual plot of the GLM applied on the loss amount of year 2013.....	70
Figure 23: Pearson Residual plot of the GLM applied on the loss amount of year 2014.....	73

Figure 24: Pearson Residual plot of the GLM applied on the loss amount of year 2015..... 76
Figure 25: Pearson Residual plot of the GLM applied on the loss amount of year 2016..... 79

Table of Tables

Table 1: The classical credibility y-values for the case of the normal distribution	34
Table 2: Total number and Percentage distribution for each Type of Breach	38
Table 3: Total number of breaches for each Type of Breach for year 2013, 2014, 2015 and 2016.	39
Table 4: Results of the Mean, Median and Standard Deviance of the loss amount per type of breach.....	40
Table 5: Results of the Mean, Median, Mode, Variance and Standard Deviance of the loss amount per year.....	40
Table 6: Results of the fitted distributions of the overall impact loss amount of breaches.	42
Table 7: Results of the fitted distributions of the impact loss amount of breaches for year 2013.	44
Table 8: Results of the fitted distributions of the impact loss amount of breaches for year 2014.	46
Table 9: Results of the fitted distributions of the impact loss amount of breaches for year 2015.	48
Table 10: Results of the fitted distributions of the impact loss amount of breaches for year 2016.	50
Table 11: Results of the important variable selection process using Random forests for the overall data.....	55
Table 12: Results of the important variable selection process using Random Forest for Year 2013.....	57
Table 13: Results of the important variable selection process using Random Forest for Year 2014.....	58
Table 14: Results of the important variable selection process using Random Forest for Year 2015.....	59
Table 15: Results of the important variable selection process using Random Forest for Year 2016.....	61
Table 16: Results of the important variable selection process using Random Forest for Year 2013, 2014, 2015 and 2016.....	62
Table 17: Results of the coefficient for the Gaussian GLM on the overall data	64
Table 18: Results of the Deviance for the Gaussian GLM applied on the overall data.....	64

Table 19: Results of the Minimum, Maximum, Median, First and Second Quantile for the GLM of overall data.	65
Table 20: Results of the AIC and Fisher Scoring for the Gaussian GLM applied on the overall data.	65
Table 21: Results of the MAPE and RMSE for the Gaussian GLM applied on the overall data.	65
Table 22: Results of the coefficient for the GLM applied on the loss amount of year 2014.	68
Table 23: Results of the Deviance for the GLM applied on the loss amount of year 2013.	69
Table 24: Results of the Min, Max, Median, First and Thirsd Quantile for the GLM applied on the loss amount of year 2013.	69
Table 25: Results of the AIC and Fisher Scoring for the Gaussian GLM applied on the loss amount of year 2013.	69
Table 26: Results of the MAPE and RMSE for the Gaussian GLM applied on the loss amount of year 2013.	70
Table 27: Results of the coefficient for the GLM applied on the loss amount of year 2014.	71
Table 28: Results of the Deviance for the GLM applied on the loss amount of year 2014.	72
Table 29: Results of the Min, Max, Median, First and Thirsd Quantile for the GLM applied on the loss amount of year 2014.	72
Table 30: Results of the AIC and Fisher Scoring for the Gaussian GLM applied on the loss amount of year 2014.	72
Table 31: Results of the MAPE and RMSE for the Gaussian GLM applied on the loss amount of year 2014.	73
Table 32: Results of the coefficient for the GLM applied on the loss amount of year 2015.	74
Table 33: Results of the Deviance for the GLM applied on the loss amount of year 2015.	75
Table 34: Results of the Min, Max, Median, First and Thirsd Quantile for the GLM applied on the loss amount of year 2015.	75
Table 35: Results of the AIC and Fisher Scoring for the Gaussian GLM applied on the loss amount of year 2015.	75
Table 36: Results of the MAPE and RMSE for the Gaussian GLM applied on the loss amount of year 2015.	75
Table 37: Results of the coefficient for the GLM applied on the loss amount of year 2016.	77
Table 38: Results of the Deviance for the GLM applied on the loss amount of year 2016.	78

Table 39: Results of the Min, Max, Median, First and Third Quantile for the GLM applied on the loss amount of year 2016.	78
Table 40: Results of the AIC and Fisher Scoring for the Gaussian GLM applied on the loss amount of year 2013	78
Table 41: Results of the MAPE and RMSE for the Gaussian GLM applied on the loss amount of year 2016.....	79
Table 42: Results of the Classical credibility Theory applied on the overall data.	80
Table 43: Results of the Classical credibility Theory applied for year 2013.....	80
Table 44: Results of the Classical credibility Theory applied for year 2014.....	81
Table 45: Results of the Classical credibility Theory applied for year 2015.....	81
Table 46: Results of the Classical credibility Theory applied for year 2016.....	82

Chapter 1: Cyber Risk Insurance

1.1 Cyber Risk

1.1.1 Definition of Cyber Risk

The Geneva Association (2016), the leading international insurance think tank for strategically important insurance and risk management issues, defines cyber risk as any risk resulting from the use of information and communication technology (ICT) that threatens the confidentiality, availability, or credibility of data or services. The OECD (2017) stated that the insurance industry associations have put forward a couple of definitions because there is no specific definition for Cyber risk used broadly within the insurance sector.

1.1.2 Types of Cyber incidents and losses

The OECD (2017) presents a description of the different types of cyber incidents as well as the types of losses that can happen. Concerning Cyber incidents, four broad categories are included:

(i) Data confidentiality breach:

Among the most common types of cyber accidents are incidents involving the compromise of sensitive data. The most common cause of data confidentiality accidents has historically been the release of confidential data through employee error. Also, incidents caused by malicious attacks have accounted for an increasing share of data confidentiality incidents, particularly portable device encryption that has become more prevalent.

(ii) System malfunction and issues:

There are five sub-categories of system malfunction/issue:

1. Own system malfunction
2. Own system affected by malware
3. Network communication malfunction
4. Inadvertent disruption of third-party system
5. Disruption of external digital infrastructure.

(iii) Data integrity and availability:

In this category, the classification of an incident is based on the identification of deleted, lost or encrypted data, rather than the underlying cause.

(iv) Malicious activity:

There are three sub-categories of malicious activity:

1. System misuse such as digital system misuse to transmit defamatory or embarrassing messages.
2. Targeted malicious communication such as phishing attempts to secure sensitive information.
3. Cyber fraud, cyber theft such as unauthorized transfer of financial information.

Cyber incident can potentially lead to a variety of different types of damage, including damage to tangible and intangible properties, business interruption and theft-related damages, as well as multiple forms of consumer, retailer, employee and shareholder liability.

1.2 Cyber Security

1.2.1 Definition of Cyber Security

According to Cyber Security & Infrastructure Security Agency (CISA), Cyber security is the art of protecting networks, computers and data from unapproved access or criminal usage, and the practice of ensuring that information is confidential, integral and accessible. Nowadays, in addition to communication, entertainment, transportation, and shopping, work also rely on the internet for exchange of information and etc.

1.2.2 Problems of Cyber Security

The most significant impediment to the management of cyber risk is the lack of data on cyber incidents (OECD, 2017). Cyber risk information is not publicly accessible because it is not reported by organizations with security breaches or who have been targeted. Also, the lack of a clear-cut concept of cyber risk is another impediment to the collection of cyber risk data (Martin and Jan, 2018).

1.3 Cyber Insurance

1.3.1 Definition of Cyber Insurance

According to the Association of British Insurers (2020), Cyber Insurance can protect businesses from cyber-attacks by covering losses relating to the loss of information from information technology (IT) systems and networks. Cyber insurance is one of the fastest growing lines of insurance business and it is becoming a primary component of companies risk management (Andrew et al., 2018). Any business of any size, relying on information technology (IT) infrastructure, will be under the risk of business interruption, income loss, damage management and repair, and possibly reputational damage (Association of British Insurers, 2020). Many companies buy a cyber insurance policy to protect themselves against cyber loss (Andrew et al., 2018).

1.3.2 Contents of Cyber Insurance Policy

As stated by the Federal Trade Commission (FTC), Cyber Insurance Policy must include coverage for cyber-attacks that occur anywhere in the world, such as breaches of the policy holder's network, breaches of the policy holder's data held by vendors and other third parties, incidents involving theft of personal information and terrorist acts. Also, Cyber Insurance Policy should defend the policy holder in a lawsuit or regulatory investigation, provide coverage in excess of any other applicable insurance that the policy holder's has and offer a breach hotline that is available every day of the year at all times. The OECD (2017) stated that the insurance market has a main role to play in providing greater information about the coverage available for cyber risk and which policies provide that coverage. As stated by the Association of British Insurers (2020), cyber insurance generally provides coverage for the theft or loss of "first party" and "third party".

First-party insurance covers the policy holder's business's own assets. As to the "First-party" coverage, an insurer may cover loss or damage to digital assets such as data or software programs, business interruption from network intermission, cyber extortion where third parties threaten to spoil or unleash data if money is not paid to them, customer notification expenses when there is a legal requirement to inform them of a security or privacy breach, reputational damage originating from a breach of data resulting in loss of intellectual property, theft of money or digital assets through theft of equipment or electronic theft, etc...

Third-party insurance covers the assets of others. Usually, this type of Cyber Insurance protects businesses that are responsible for a client's cyber security. "Third-party" coverage may include security and privacy breaches, and all the costs associated with them, such as investigation, defense costs and civil damages. Also, it may include multi-media liability to cover investigation, defense costs and civil damages arising from defamation. Moreover, it can include loss of third-party data, including payment of compensation to customers for denial of access, and failure of software or systems, etc...

1.3.3 Types of Cyber Insurance coverage

There are different types of cyber insurance available (Andrew et al., 2018):

- Stand-alone cyber insurance (or affirmative cyber insurance): covers a company's costs it would incur as a result of a cyber-attack. The stand-alone cyber insurance market is rapidly growing.
- Errors and Omissions (E&O) insurance: covers a company's liability to a third party. It is one of the oldest types of cyber insurance.
- Commercial property all-risk insurance: covers physical damage and business interruption resulting from a cyber-attack.
- Personal line insurance: covers a loss resulting from a cyber-attack on home computers or compensation for family members who have personal or financial data compromises. Some homeowners

1.3.4 Market challenges of Cyber insurance:

The OECD (2017) mentioned that a risk is insurable only when certain principles of insurability are met. These principles of insurability include:

- Risks must be quantifiable: the probability of a given risk, its severity and its impact in terms of harm and losses must be assessable.
- Risks must be mutual: the risk must be shared by a sufficiently large community with assets at risk.
- Risks must occur randomly: the time and the placement of an insured risk must be unpredictable.

The price at which insurance companies are willing to cover a given risk is affected by many factors including the level of uncertainty in estimating expected losses (quantifiability), the size of expected losses (economic viability) and the diversity of the pool of risks covered (limited correlation) (OECD, 2017).

In the case of cyber insurance, the most critical challenges in underwriting cyber risk are the difficulties in quantifying a newly emerging risk (quantifiability), and the potential for significant correlation across policyholders (accumulation risk).

A-Factors affecting the price of cyber insurance:

I. Quantifiability of cyber risk:

In terms of cyber risk quantification, there are three key challenges:

i. Limited availability of historical data on cyber incidents:

This lack of information is compounded by the general unwillingness of cyber incident victims to share information about these incidents and their impacts (unless required) out of concern about future reputational impacts (CRO Forum, 2014). The lack of historical data does not allow insurance premiums to be accurately priced.

ii. Changing nature of cyber risk:

A potentially more critical problem is that, even though more information is available, the data will easily become out-of-date as a result of the constantly changing nature of cyber risk (CRO Forum, 2014).

iii. Access to corporate security information:

The lack of transparency regarding security procedures and past accidents has been described by a number of insurance providers as a major barrier to underwriting coverage OECD (2017)

II. Accumulation risk:

As stated by the OECD (2017), building a broad pool of diversified risks (independent and randomly-occurring losses) helps insurers to distribute losses over a large number of policyholders and mitigates the risk for losses to be affected concurrently by a large share of the pool. All things being equal, a smaller pool, or a pool with greater reliance on the risks covered, would lead to higher premiums being required by insurers. In the case of cyber risk, the potential for losses to be associated across policyholders and across various types of coverage given to a single policyholder ('accumulation risk') is important. Also, building a diversified risk pool based on geography or even industry is also more difficult, given the reliance on the same infrastructure, software and services.

Therefore, the potential for risk accumulation across policyholders is, according to some studies, the primary reason why insurers limit the coverage available for cyber risk (OECD, 2017).

III. Reinsurance availability:

The availability of cyber risk reinsurance coverage is affected by the lack of historical experience, the changing risk environment and in particular the potential for risk accumulation OECD (2017).

Some reports have indicated that there is limited availability of reinsurance for cyber threats, that this could hinder primary insurers' ability to offer coverage, and that a devastating cyber incident could involve a government backstop OECD (2017).

B-Factors affecting the willingness-to-pay for cyber insurance coverage:

- I. Lack of awareness of potential cyber losses
- II. Misunderstandings about coverage
- III. Coverage that is not suited to the needs of policyholders

Chapter 2: Modeling of Cyber Risk

Over the past years, cyber risk has gotten a lot of attention from academics, industry, and governments. Unfortunately, scientific progress has been slow and insufficient. Industry and academic research on cyber risk modeling has been fragmented and uncoordinated till date due to the lack of historical data available on cyber incidents which does not allow insurance premiums to be accurately priced, in addition to the constantly changing nature of cyber risk which makes the data easily become out-of-date.

Böhme and Kataria (2006) systematically investigate correlations on two levels: the first level is when a cyber-event could affect several systems within a single entity (e.g. company) and the second level is when there is correlation across different entities (e.g. an insurer's portfolio of policies). Böhme and Kataria (2006) suggested the t-copula to model the correlation of extreme events. They used honeynet data from the "Leurre.com" project to empirically estimate the size of the correlation. However, they employed the t-copula purely for simulation of random variables exercises and highlighted that better data would be required in order to estimate suitable copulas and parameters empirically.

Maillart and Sornette (2010) investigated personal data breaches such as credit card, social security numbers, banking accounts, or medical files in their research. They focused on a specific criminality, which is the theft of personal information (ID thefts), using a dataset from the Open Security Foundation that contains 956 documented events reported mainly in the USA between the years 2000 and 2008. The frequency of such incidents has been faster than exponentially growing in the period from 2001 to 2006. Then, the development plateaued out by 2006. Furthermore, they found that the severity per incident has an extremely heavy-tailed distribution (Pareto index of 0.7) and the laws governing its distribution have been stable over time. Maillart and Sornette (2010) argue that these findings are representative of other types of cyber risks originating on the Internet, while personal data breaches are only one type of cyber risk.

Herath and Herath (2011) developed a cyber-insurance pricing model, where the premiums depend on the number of computers affected, the firm level dollar loss distribution, and the timing of the breach event. They illustrated a copula-based Monte Carlo simulation model for pricing cyber insurance using empirical loss distributions based on publicly available ICSA survey data. They computed the premiums for first party losses using three types of insurance policy models: basic policies, policies with a deductible and policies with a deductible and co-insurance. Herath and Herath (2011) analyzed losses due to virus incidents and found that the marginal distributions are not normal, and the risks are correlated in a non-linear fashion.

Mukhopadhyay et al. (2013) proposed a Copula-aided Bayesian Belief Network (CBBN) for cyber-vulnerability assessment (C-VA), and expected loss computation using averaged log data collected from a premier business management school in India, for a period of two years. They concentrate only on malicious events. Mukhopadhyay et al. (2013) used the normal copula in order to aggregate the number of failures (frequency) and the costs given loss (severity) in order to derive the overall loss distribution on a cyber-risk portfolio.

Building on the work of Maillart and Sornette (2010), Wheatley et al. (2015) conduct a similar analysis using an up-to-date and broader data set which is a combination between the dataset from the Open Security Foundation and the dataset from the Privacy Rights Clearing House.

Wheatley et al. (2015) discovered that the number of breaches per incident have had an even more heavy-tailed distribution since 2007 (Pareto index of 0.37 for 2015). Also, the breach size is expected to double in the next five years from an estimated 2 billion personal items to 4 billion.

Similarly as Maillart and Sornette (2010) and Wheatley et al. (2015), Edwards et al. (2015) investigate trends in data breaches between 2006 and 2015. They studied data obtained from a publicly available dataset published by the Privacy Rights Clearinghouse (PRC) and develop Bayesian Generalized Linear Models to investigate trends in data breaches. However, they find no evidence for an increasing trend in frequency (Number of data breaches) or in severity (Number of records per data breach). Edwards et al. (2015) found that the widespread intuition that the frequency and severity of data breaches are increasing, can be explained by the heavy-tailed statistical distributions underlying the dataset. They found that the severity of data breaches is best

described by the log-normal family of distribution and the frequency data breaches follows a negative binomial distribution.

Eling and Loperfido (2017) used multidimensional scaling and goodness-of-fit tests to analyze the distribution of data breach information. They studied data obtained from a publicly available dataset published by the Privacy Rights Clearinghouse (PRC). They showed that different types of data breaches need to be modeled as distinct risk categories. For severity modeling, the log-skew-normal distribution provides promising results. For frequency modeling, the Poisson distribution or the negative binomial distribution provides promising results.

Fetterman (2019) developed a quantitative model from the management perspective to facilitate understanding of how factors including human data analysis and resourcing affect the time-to-detect an incident. Understanding the relationships of variables in cyber security incidents can show how engineering managers can properly distribute and concentrate resources to improve the detection time of incidents. In his research, Fetterman (2019) used the data from the VERIS Community Database. The VERIS dataset is publicly available and contains over 7,000 records of anonymized cyber security incidents, with 136 fields per record. Fetterman (2019) used multiple forms of regression to measure the relationship between malware features, hacking techniques, and milestones of the incident response timeline as reported in the VERIS dataset.

Farkas et al., (2020) analyzed cyber claims via regression trees in order to constitute clusters of cyber incidents because Regression trees are good candidates to understand the origin of the heterogeneity, since they allow performing regression and classification simultaneously. In their research, they devoted special attention to large claims for which heavy tail distributions are fitted, since the analysis of large claims raises the problem of risk insurability, and thus the clustering technique may separate between type of events that can or cannot be covered without endangering risk pooling. In their research, Farkas et al., (2020) used the Privacy Rights Clearinghouse (PRC) database available for public download. Since the severity of cyber events is highly volatile, it seems necessary to develop a specific approach for the tail of distribution. Therefore, Generalized Pareto Distributions (GPD) naturally appears in the analysis of heavy-tailed variables.

Title	Model	Data
Models and measures for correlation in cyber-insurance (Böhme and Kataria, 2006)	T-copula to model the correlation of extreme events.	Honeynet data from Leurre.com project
Heavy-tailed distribution of cyber-risks (Maillart and Sornette, 2010)	Heavy-tailed distribution	Dataset from the Open Security Foundation
Copula-based actuarial model for pricing cyber-insurance policies (Herath and Herath, 2011)	Copula-based Monte Carlo simulation model	Publicly available ICSA survey data
Cyber-risk decision models: to insure it or not? (Mukhopadhyay et al., 2013)	Copula-aided Bayesian Belief Network (CBBN)	Data collected from a premier business management school in India
The extreme risk of personal data breaches & the erosion of privacy (Wheatley et al., 2015)	Heavy-tailed distribution	Combination between the dataset from the Open Security Foundation and the dataset from the Privacy Rights Clearing House (PRC)
Hype and heavy tails: a closer look at data breaches (Edwards et al., 2015)	Bayesian Generalized Linear Models	Dataset from the Privacy Rights Clearinghouse (PRC)
Data breaches: Goodness of fit, pricing, and risk measurement (Eling and Loperfido, 2017)	The log-skew-normal distribution for severity modeling, and the Poisson distribution or the negative binomial distribution for frequency modeling	Dataset from the Privacy Rights Clearinghouse (PRC)
Regression-Based Attack Chain Analysis and Staffing	Multiple forms of regression	VERIS Community Database.

Optimization for Cyber Threat Detection (Fetterman, 2019)		
Cyber claim analysis through Generalized Pareto Regression Trees with applications to insurance (Farkas, Lopez, & Thomas, 2020)	Generalized Pareto Regression Trees	Dataset from the Privacy Rights Clearinghouse (PRC)

Figure 1: Summary of the research papers on modeling Cyber Risk

Chapter 3: Methodology

The aim of this chapter is to explain the Statistical and Machine learning methods used to model and predict the loss amount of a cyber breach. In this study, the loss amount of a breach was fitted to a parametric distribution and goodness of fit test was used to test the hypothesis of the fitness. Credibility theory was also used in order to estimate minimum number of losses required to have a certain level of accuracy. Then, the Random Forest was used to select the most important variables affecting the loss amount of a breach. Finally, the Generalized Linear Model (GLM) of the fitted distribution was applied on the most important variable provided by the Random Forest in order to estimate and predict future loss amount due to cyber risk.

This chapter will explain in detail the following:

- Distribution fitting
- Random Forest
- Generalized linear model
- Credibility theory

3.1 Fitting a distribution

Distribution fitting is the process used to select a statistical distribution that best fits the data. Fitting a distribution is done by finding the parameters of the fitted distribution and then testing the hypothesis of the fitness using Goodness of fit test.

3.1.1 Parametric methods for fitting a distribution

There are several parametric methods for estimating the parameters of a probability distribution such as the method of moments, the maximum spacing estimation, the method of L-moments and the Maximum likelihood method. Maximum likelihood estimation is a method that is applied to estimate the parameters in a parametric distribution. The parameter values are found such that they maximize the likelihood that the process described by data that were actually observed (Broverman, 2014).

The first step in applying the maximum likelihood estimation is finding the log-likelihood function $L(\theta)$ where θ is the parameter (or parameters) to be estimated in a distribution with probability density function $f(x, \theta)$ and cumulative density function $F(x, \theta)$.

The likelihood function based on a random sample is as follow:

$$L(\theta) = \prod_{j=1}^n f(x_j, \theta) ; \text{ where the random sample is } x_1, x_2, \dots, x_n$$

The objective is to find the value of θ that maximizes $L(\theta)$. Therefore, the natural log of the likelihood function (log-likelihood), $l(\theta) = \ln L(\theta)$, is maximized and this results in the same maximum likelihood estimate of θ .

In order to maximize the log-likelihood function, the following equation is set and solved for θ :

$$\frac{d}{d\theta} \ln L(\theta) = 0$$

3.1.2 Goodness of fit test

Goodness-of-fit tests are statistical tests aiming to determine whether a set of observed values in a sample comes from a specific distribution of a population. There are multiple types of goodness-of-fit tests, such as the chi-square test and the Kolmogorov-Smirnov test used for large samples (Broverman, 2014). Thus, the hypothesis are:

H_0 : The data comes from the estimated model.

H_1 : The data does not come from the estimated model.

In this study the Kolmogorov-Smirnov (KS) statistic Test is used to measure how well the empirical distribution function of the sample agrees with the cumulative distribution function of a pre-specified theoretical distribution (Broverman, 2014). The Kolmogorov-Smirnov (KS) test is:

$$D = \max_{\text{all } x_i} |F_n(x_j) - F^*(x_j)|$$

$F_n(x_j)$ is the empirical distribution function.

$F^*(x_j)$ is the cumulative distribution function of the model based on the estimated parameter value. Thus, the Kolmogorov-Smirnov statistic measures of how "far" the empirical and estimated cumulative model distributions are from each other. If the value of the Kolmogorov-Smirnov (KS) test \mathbf{D} is greater than the critical value, the null hypothesis H_0 will be rejected.

The critical value depends on the significance level of the test and the sample size n :

Level of Significance	0.1	0.05	0.01
Critical Value	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

3.1.3 Akaike information criterion (AIC)

The Akaike information criterion (AIC) is an information criterion that is commonly used for model selection. The idea of AIC is to adjust the empirical risk to be an unbiased estimator of the true risk in a parametric model (Dunn & Smyth, 2017).

Formally, the AIC is defined in terms of the log-likelihood as:

$$AIC = -2l(\theta) + 2k$$

Where $l(\theta) = \ln L(\theta)$ is the log-likelihood evaluated at the MLE for the model under consideration and K is the number of unknown parameters.

The AIC is a powerful tool for comparing models and estimating the quality of each model. Smaller values of the AIC represent better models (Dunn & Smyth, 2017).

3.2 Random Forest

Random forests are a statistical learning method used in many fields of application and are adapted to both supervised classifications (categorical response variable) problems and regressions (continuous response variable) problems for qualitative and quantitative explanatory variables together without preprocessing (Genuer & Poggi, 2020). The general principle of random forests is to aggregate a collection of random decision trees. Since individual trees are randomly perturbed, the forest benefits from a more wide exploration of the space of all possible tree predictors, which, in practice, results in better predictive performance (Genuer & Poggi, 2020).

3.2.1 Definition of Random forests

Let $(\hat{h}(\cdot, \theta_1), \dots, \hat{h}(\cdot, \theta_q))$ be a collection of tree predictors, with $\theta_1, \dots, \theta_q$ q i.i.d. random variables. The random forest predictor \hat{h}_{RF} is obtained by aggregating this collection of random trees. The aggregation is done as follows:

- In regression: $\hat{h}_{\text{RF}}(x) = \frac{1}{q} \sum_{l=1}^q \hat{h}(x, \theta_l)$ (average of individual tree predictions).
- In classification: $\hat{h}_{\text{RF}}(x) = \arg \max_{1 < c < C} \sum_{l=1}^q 1_{\hat{h}(x, \theta_l)=c}$ (majority vote among individual tree predictions)

There are two main objective of using Random Forest (Genuer & Poggi, 2020):

1. The first main learning objective is prediction:

Random Forest can be used to construct a predictor, using the learning sample, which associates a prediction y of the response variable corresponding to any given input observation.

2. The second main objective is selection of important variable:

Random forest involves determining a subset of the input variables that are important and active in explaining the input–output relationship. It helps in constructing a hierarchy of input variables based on a quantification of the importance of the effects on the output variable. Thus, Random Forest provides a ranking of variables, from the most important to the least important.

3.2.2 Selection Variable importance “VI (X^j)”

Definition 3.2.2: Bootstrap sample:

A bootstrap sample of a learning sample L_n of size n is obtained by randomly drawing n observations from L_n with replacement, each observation (X_i, Y_i) of L_n having a probability $1/n$ of being selected in each draw (Genuer & Poggi, 2020).

Definition 3.2.3: Out Of Bag error “OOB error”

According to Genuer and Poggi (2020), the main idea of an Out Of Bag error estimator is to use observations (X_i, rY_i) that were not selected in a bootstrap sample as test data. In other word, an Out Of Bag error must be understood as out of the Bootstrap error. To predict the i th observation X_i , we only aggregate predictors built on bootstrap samples not containing (X_i, Y_i) . This provides a prediction \widehat{Y}_i for the output of the i th observation. The OOB error is then calculated as follows:

- In regression: $\frac{1}{n} \sum_{i=1}^n (\widehat{Y}_i - Y_i)^2$
- In classification: $\frac{1}{n} \sum_{i=1}^n 1_{Y_i \neq \widehat{Y}_i}$

Definition 3.2.4: Variable importance “VI (X^j)”

Let $j \in \{1, \dots, p\}$, then the importance index $VI(X^j)$ of variable X^j is calculated as following (Genuer & Poggi, 2020):

- Consider a bootstrap sample $L_n^{\Theta_1}$ of size n and the associated OOB₁ sample, that is, all observations that do not belong to $L_n^{\Theta_1}$.
- Calculate err_{OOB_1} , the error made on OOB₁ by the tree built on $L_n^{\Theta_1}$ (mean square error or misclassification rate).
- Then randomly permute the values of variable X^j in the OOB₁ sample. This gives a perturbed sample, noted $\widetilde{OOB_1^j}$.
- Finally, calculate $err_{\widetilde{OOB_1^j}}$, the error made on $\widetilde{OOB_1^j}$ by the tree built on $L_n^{\Theta_1}$.

- Repeat these operations for all bootstrap samples. The importance of the variable X^j , $VI(X^j)$, is then defined by the difference between the average error of a tree on the perturbed OOB sample and that on the OOB sample:

$$VI(X^j) = \frac{1}{q} \sum_1^q (\widetilde{err_{OOB_1^j}} - err_{OOB_1^j})$$

Hence the higher the error increase originating from the random permutations of the variable, the more important is the variable X^j .

3.2.3 Influence of Parameters on Variable Importance

The main parameters of the Random Forest function in machine learning are:

- The number of variables randomly selected at each node, which by default is \sqrt{p} in classification and $p/3$ in regression where p refers to the number of input variables.
- The number of trees in the forest, denoted as q in this chapter, which by default is 500.
- The minimum number of observations that a leaf of a tree must contain, which by default is 1 in classification and 5 in regression.

When we wish to have information on the variables, or even to select variables, we must try to adjust the Random Forest parameters by looking at their impact on Variable Importance VI. For instance, increasing the number of trees has the effect of stabilizing the Variable Importance VI, and the default value 500 seems large enough in this study (Genuer & Poggi, 2020).

3.2.4 Radom Forest Algorithm

The random forest is based on an aggregation of decision trees which are independently developed on different sample bags taken from the training set. The importance estimation of a variable is calculated as the loss of classification accuracy caused by a random permutation of variable values of cases. First, the loss of classification accuracy is computed individually for all decision trees in the forest which make use of a given feature to classify cases and then the average and standard deviation of the loss of classification accuracy are computed.

Accuracy of Radom Forest:

Let $h_1(x), h_2(x), \dots, h_k(x)$, be an ensemble of classifiers:

- The margin function is: $mg(X, Y) = av_k I(h_k(x) = Y) - \max_{j \neq y} av_k I(h_k(x) = j)$

Where $I(.)$ is the indicator function. The larger the margin, the more confidence in the classification of the Random Forest.

- The generalization error is:

$$PE^* = \text{Probability}_{(X,Y)}(mg(X, Y) < 0) = P_{(X,Y)}(mg(X, Y) < 0)$$

Where the probability of $mg(X, Y) < 0$ is over the X, Y space.

As the number of trees increases, for all Θ_1 , PE^* converges to:

$$P_{(X,Y)}(P_{(\Theta)}(X, \Theta) = Y) - \max_{j \neq y} P_{(\Theta)}((h(X, \Theta) = j) < 0)$$

Strength and correlation:

As mentioned previously, the margin function for Random Forest is:

$$mr(X, Y) = P_{(X,Y)}(P_{(\Theta)}(X, \Theta) = Y) - \max_{j \neq y} P_{(\Theta)}((h(X, \Theta) = j) < 0)$$

The strength of the set of classifiers $h(X, \Theta)$ is: $s = E_{X,Y}mr(X, Y)$

The mean value of the correlation $\bar{\rho}$ is:

$$\bar{\rho} = \frac{E_{\Theta, \Theta'}(\rho(\Theta, \Theta')sd(\Theta)sd(\Theta'))}{E_{\Theta, \Theta'}(sd(\Theta)sd(\Theta'))}$$

3.3 Generalized Linear Models

A Generalized Linear Models (GLM) is a generalized form of a linear model that is used to express the relationship between an observed response variable Y , and a number of covariates or predictor variables X . In a GLM, the response variable Y is assumed to be a member of the exponential dispersion model family (EDM), the variance is permitted to vary with the mean of the distribution and finally the additive effects of the covariates on the response variable Y is taken into consideration (Dunn & Smyth, 2017).

3.3.1 Components of Generalized Linear Models

There are two components of Generalized Linear Model:

1. Random component of the model: What probability distribution is appropriate?
2. Systematic component of the model: How are the explanatory variables related to the mean of the response μ ?

- The Random Component: Exponential Dispersion Models (EDM):

In a GLM, the response variable Y is assumed to be a member of the exponential dispersion model family (EDM). Continuous exponential dispersion model families include the normal and gamma distributions. Discrete exponential dispersion model families include the Poisson, binomial and negative binomial distributions (Dunn & Smyth, 2017).

The exponential dispersion model family (EDM), have a probability distribution function of the form:

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}$$

$$P(y, \theta, \phi) = a(y, \phi) \exp\left\{\frac{y\theta - k(\theta)}{\phi}\right\}$$

Where

- θ is called the canonical parameter related to the mean.

- $k(\theta)$ is a known function, and is called the cumulant function.
- $\phi > 0$ is the dispersion parameter or a scale parameter related to the variance.
- $a(y, \phi)$ is a normalizing function ensuring that $P(y, \theta, \phi)$ is a probability function.
- The mean μ is a known function of the canonical parameter θ
- The variance of an Exponential Family of Distribution of Y_i is a function of its mean:

$$\text{Var}(Y_i) = \phi \cdot V(\mu_i)$$

- The Systematic Component: Exponential Dispersion Models (EDM):

In GLM, in addition to assuming that the response variable Y is assumed to be a member of the exponential dispersion model family (EDM), GLM assume a specific form for the systematic component where the linear predictor:

$$\eta_i = \alpha_i + \beta_0 + \sum_j^p \beta_j X_{ji}$$

is linked to the mean μ through a link function $g(\cdot)$ so that $g(\mu) = \eta$. The link function $g(\cdot)$ is a monotonic, differentiable function relating the fitted values μ to the linear predictor η .

In a simplest way, the general equation for GLM is as following:

$$\eta_i = \hat{Y}_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Where, β_k in a GLM are coefficients or weights assigned to the predictor variables X_k .

β_0 , the intercept, is the predicted value of Y when all of the X_k equal 0.

In a more specific terms, β_k gives the predicted change in Y for a one unit change in the X_k , given that everything else constant (Dunn & Smyth, 2017).

3.3.2 Examples of Exponential Dispersion Models with their link function (EDM)

i. Normal Distribution:

The probability density function for the Normal distribution with mean μ and variance σ^2 is:

$$f(y, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y - \mu)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{y\mu - \left(\frac{\mu^2}{2}\right)}{\sigma^2} - \frac{y^2}{2\sigma^2}\right\}$$

- $\theta = \mu$ is the canonical parameter.
- $k(\theta) = \left(\frac{\mu^2}{2}\right) = \left(\frac{\theta^2}{2}\right)$ is the cumulant function.
- $\phi = \sigma^2$ is the dispersion parameter.
- $a(y, \phi) = \frac{1}{\sqrt{2\pi\phi}} \exp\left\{-\frac{y^2}{2\phi}\right\}$ is a normalizing function.

ii. Gamma Distribution:

The probability density function for the Gamma distribution is:

$$f(y, \alpha, \theta) = \frac{x^{\alpha-1}}{\theta^\alpha \Gamma(\alpha)} \exp\left\{-\frac{y}{\theta}\right\}$$

- $\theta = -\frac{1}{\mu}$ is the canonical parameter.
- $k(\theta) = \log(\mu) = -\log(-\theta)$ is the cumulant function.
- $\phi = \theta$ is the dispersion parameter.
- $a(y, \phi) = 2\left\{-\log\left(\frac{y}{\phi}\right) + \frac{y-\mu}{\mu}\right\}$ is a normalizing function.

3.3.3 Generalized Linear Model Assumptions

The GLM assumptions are as follows:

GLM1: Random Component: Each component of the response variable Y is independent and is assumed to be a member of the exponential family of distributions (EDM).

$$Y_i \sim \text{EDM}\left(\mu_i, \frac{\phi}{w_i}\right) \text{ for } i = 1, 2, \dots, n.$$

The W_i are non-negative prior weights, or credibility, that weight each observation i and are all equal to one.

GLM2: Systematic Component: The p covariates are combined to give the linear predictor η :

$$\eta_i = \alpha_i + \beta_0 + \sum_j^p \beta_j X_{ji}$$

α_i are offsets that are usually equal to zero.

GLM3: Link function: The relationship between the random and systematic components is specified via a known, monotonic, differentiable link function:

$$g(\mu_i) = \eta_i$$

$$\mu_i = g^{-1}(\eta_i)$$

Thus the GLM is GLM (EDM; Link function):

$$\left\{ \begin{array}{l} Y_i \sim \text{EDM} \left(\mu_i, \frac{\phi}{w_i} \right) \\ g(\mu_i) = \eta_i = \alpha_i + \beta_0 + \sum_j^p \beta_j X_{ji} \end{array} \right.$$

Hence the GLM is dependent of the choice of distribution from the EDM class and the choice of link function.

3.3.4 Fisher scoring iteration

The Fisher scoring algorithm provides an important method for computing and estimating the maximum likelihood estimates (MLEs) of the coefficients or weights β_j assigned to the predictor variables X_k .

The Fisher scoring algorithm computes the $\widehat{\beta}_j$ by iteratively refining the working estimates until convergence.

The working response of the Fisher scoring algorithm is defined as:

$$z_i = \eta_i + \frac{d\eta_i}{d\mu_i} (y_i - \mu_i)$$

Each iteration of the Fisher scoring algorithm is equivalent to the least squares regression of the working responses z_i on the covariates X_{ji} using the working weights W_i .

Thus, at each iteration, the z_i and W_i are updated, and the regression is repeated to obtain new coefficients $\widehat{\beta}_j$.

Moreover, since at each repetition of the Fisher scoring iteration the linear predictor η_i is updated from the working coefficients, the fitted values $\mu_i = g^{-1}(\eta_i)$ are also updated.

A high number of iterations of the Fisher scoring may be indicating that the algorithm is not converging properly. Thus, the lower the number of iterations of the Fisher scoring the better the model fits (Dunn & Smyth, 2017).

3.3.5 Diagnostics for Generalized Linear Models

Diagnostics for Generalized Linear Models are tools for detecting violations of the assumptions in a GLM, and then discussing possible solutions. The assumptions of the GLM are as the following:

- Lack of outliers.
- The correct link function is used.
- All important explanatory variables are included on the correct.
- The correct variance function $V(\mu)$ is used.
- The dispersion parameter ϕ is constant.
- The responses Y_i are independent of each other.

The main tool for diagnostic analysis is residuals. Pearson, deviance and quantil residuals. Quantile residuals are highly recommended for discrete EDMs while Pearson and deviance residuals are highly recommended for contiuous EDMs (Dunn & Smyth, 2017).

i. Pearson residual:

The Pearson residual can be used to handle the non-constant variance in EDMs by dividing out the effect of non-constant variance (Dunn & Smyth, 2017).

$$r_p = \frac{y - \hat{\mu}}{\sqrt{\frac{V(\hat{\mu})}{W}}}$$

ii. Deviance residual:

The deviance residuals can be used to check the model fit at each observation for generalized linear models. The deviance residuals r_D is defined as the signed square root of the unit deviance (Dunn & Smyth, 2017).

$$r_D = \text{sign}(y - \hat{\mu})\sqrt{\text{wd}(y, \hat{\mu})}$$

The function $\text{sign}(y - \hat{\mu})$ equals 1 if $(y - \hat{\mu}) > 0$ and -1 if $(y - \hat{\mu}) < 0$ and 0 if $x = 0$.

3.3.6 Outliers and Influential Observations

Outliers are observations incompatible with the rest of the data, and influential observations are outliers that substantially modify the fitted model when removed from the data set. Influential observations are outliers with high leverage. The measures of influence used for generalized linear models are Cook's distance D, DFFITS, DFBETAS and the covariance ratio (Dunn & Smyth, 2017).

The approximation to Cook's distance for GLM is:

$$D \approx \left(\frac{r_p}{1-h}\right)^2 \cdot \frac{h}{\phi p'}$$

3.3.7 Assessing the model

In this study, to evaluate the quality of the developed prediction model, the coefficient of variance root mean square error (CV RMSE) and the mean absolute percentage error (MAPE), were used.

$$\text{CV RMSE} = \frac{\sqrt{\sum_{m=1}^n \frac{(Y_i - \hat{Y}_i)^2}{Y_i}}}{\sum_{m=1}^n Y_i} \cdot 100\%$$

$$\text{MAPE} = \frac{1}{n} \sum_{m=1}^n \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \cdot 100\%$$

According to ASHRAE Guideline 14 criteria, for the model to be considered well calibrated, the value of the evaluation indices should not exceed:

- CV RMSE index 15%
- MAPE INDEX: 0% to 10% High Accuracy.

3.4 Credibility Theory

The credibility Theory helps actuaries in determining the risks associated with providing coverage and allows insurance firms to reduce their claims and losses exposure. Credibility theory provides tools to deal with the randomness of data that is used for predicting future events or costs (Broverman, 2014).

The basic formula for calculating credibility weighted estimates is:

$$\text{Estimate} = Z \cdot [\text{Observation}] + (1 - Z) \cdot [\text{Other Information}]$$

Where:

- Z is the credibility assigned to the observation and $0 \leq Z \leq 1$
- $(1 - Z)$ is referred to as the complement of credibility.

There are three credibility model that are used to calculate the credibility weight Z :

- The classical credibility model:
It is also referred to the limited fluctuation credibility since it attempts to limit the effect of random fluctuations on the estimates.
- The Buhlmann credibility model:
It is also referred to the least squares credibility. The goal with this approach is the minimization of the square of the error between the estimate and the true expected value of the quantity being estimated.
- Bayesian credibility theory:
The Bayes Theorem is the foundation for this analysis.

In this study, the classical credibility model is used in order to determine number of data needed in order to assign to it $P\%$ credibility. The number of observations of X needed is calculated as follows:

$$n > n_0 \cdot \frac{V(X)}{E(X)^2}$$

$$n_0 = \frac{y}{k}$$

Where:

- K is the range parameter and is often 5% (but other values are also possible, such as 2% or 1%)
- P is the probability level associated with y and is often 90% (but other values are also possible, such as 95%)

Table 1 provides an example of the y-values in the case of normal distribution

Probability level	90%	95%	98%
y Value	1.645	1.96	2.326

Table 1: The classical credibility y-values for the case of the normal distribution

Chapter 4: Modeling and prediction of Loss amount of Breaches

4.1 VERIS dataset and schema

The Vocabulary for Event Recording and Incident Sharing (VERIS) database is a publicly available database for global incident breaches and includes the following main variables:

4.1.1 Actors: whose actions affected the asset?

There can be more than one actor involved in any particular incident, and their actions can be malicious or non-malicious, intentional or unintentional, causal or contributory.

There are three primary categories of threat actors:

- a. External: external threats initiate from sources outside of the organization and its network of partners.
 - i. Motive: what motives drove the external actor(s) to act? (financial, fear, ...)
 - ii. Variety: what varieties of external actors were involved? (competitor, terrorist, ...)
 - iii. Origin (country): what are the geographic origins of the external actor(s)?
- b. Internal: internal threats are those originating from within the organization.
 - i. Motive: what motives drove the internal actor(s) to act?
 - ii. Variety: what varieties of internal actors were involved?
- c. Partner: partners include any third party sharing a business relationship with the organization.
 - i. Motive: what motives drove the partner's actor(s) to act?
 - ii. Industry: which industry best describes the services provided by the partner(s)?
 - iii. Origin (country): what is the partner's country of operation?

4.1.2 Actions: what actions affected the asset?

- a. Malware: malware is any malicious software, script, or code that is specifically designed to disrupt, damage, or gain unauthorized access to a computer system such as viruses, worms, spyware, etc.
 - i. Variety: what varieties or functions of malware were involved?
 - ii. Vector: what were the vectors or paths of infection?
 - iii. Vulnerabilities: enter any cves exploited by this malware.

- b. Hacking: according to VERIS, hacking is defined as all attempts to intentionally access or harm information assets without (or exceeding) authorization by circumventing or thwarting logical security mechanisms.
 - i. Variety: what varieties or methods of hacking were involved?
 - ii. Vector: what was the vector or path of attack?
 - iii. Vulnerabilities: enter any caves exploited through hacking.
- c. Social: social tactics employ deception, manipulation, intimidation, etc to exploit the human element, or users, of information assets.
 - i. Variety: what varieties of social tactics were involved?
 - ii. Vector: what vectors or communication channels were used?
 - iii. Target: who was the target of these social tactics?
- d. Misuse: misuse is defined as the use of entrusted organizational resources or privileges for any purpose or manner contrary to that which was intended.
 - i. Variety: what varieties of misuse were involved?
 - ii. Vector: what vectors or access methods were misused?
- e. Physical: physical actions contain threats that involve proximity, possession, or force.
 - i. Variety: what varieties of physical attacks were involved?
 - ii. Vector: how was access gained to the location(s)?
 - iii. Location: where did these physical attacks occur?
- f. Error: error broadly encompasses anything done (or left undone) incorrectly or inadvertently
 - i. Variety: what varieties of errors were involved?
 - ii. Vector: why did these errors occur?
- g. Environmental: the environmental category not only includes natural events such as earthquakes and floods, but also hazards associated with the immediate environment or infrastructure in which assets are located.
 - i. Variety: what varieties of environmental events were involved?

4.1.3 Assets: which assets were affected?

- i. Variety: what varieties (and number) of assets were compromised during this incident?

4.1.4 Attributes: how the asset was affected?

4.1.5 Incident timeline: The incident timeline is the timeline of events leading up to and following an incident.

Furthermore, there are six primary security attribute used in the VERIS database:

- Confidentiality: refers to limited observation and confession of an asset or a data.
- Possession: refers to the owner holding possession and control of an asset or a data.
- Integrity: refers to an asset or a data being complete and unmodified from the original or authorized state, content, and function.
- Authenticity: refers to the validity, conformance, correspondence to intent, and genuineness of the asset or the data.
- Availability: refers to an asset or a data being present, accessible, and ready for use when needed.
- Utility: refers to the usefulness or fitness of the asset or the data for a purpose

4.2 Descriptive Statistics

In this study, we used a sample data consisting of 276 observations from the publically available VERIS data for cyber breaches. The frequency and the percentage distribution of type of breaches are represented in Table 2 and Figure 2. The type of breach “Misuse” recorded the highest number breaches of 26%.

Type of Breach	Error	Hacking	Malware	Misuse	Physical	Social	Unknown
Number of breaches	54	67	14	73	52	14	2
Percentage of breaches	20%	24%	5%	26%	19%	5%	1%

Table 2: Total number and Percentage distribution for each Type of Breach

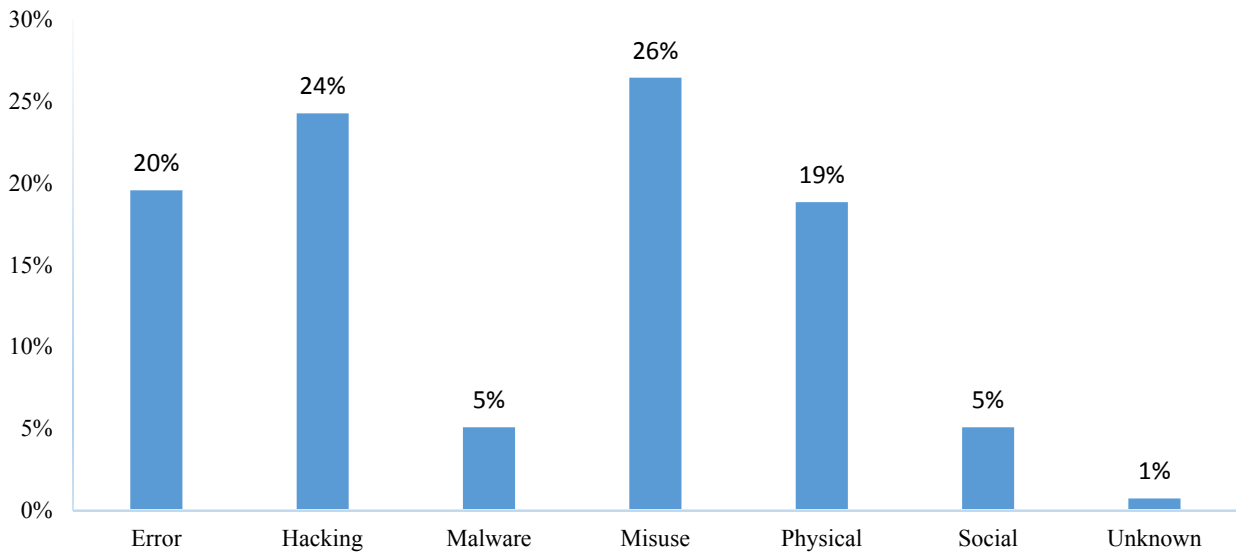


Figure 2: Percentage distribution for each Type of Breach

The frequency distribution of breach from the year 2013 till 2016 is presented in Table 3 and Figure 3. Year 2013 recorded the highest number breaches of 59 breaches.

Type of Breach	2013	2014	2015	2016
Error	8	14	6	2
Hacking	11	9	9	6
Malware	5	2	0	5
Misuse	13	8	9	6
Physical	18	5	7	6
Social	2	2	4	2
Unknown	2	0	0	0
Total	59	40	35	27

Table 3: Total number of breaches for each Type of Breach for year 2013, 2014, 2015 and 2016.

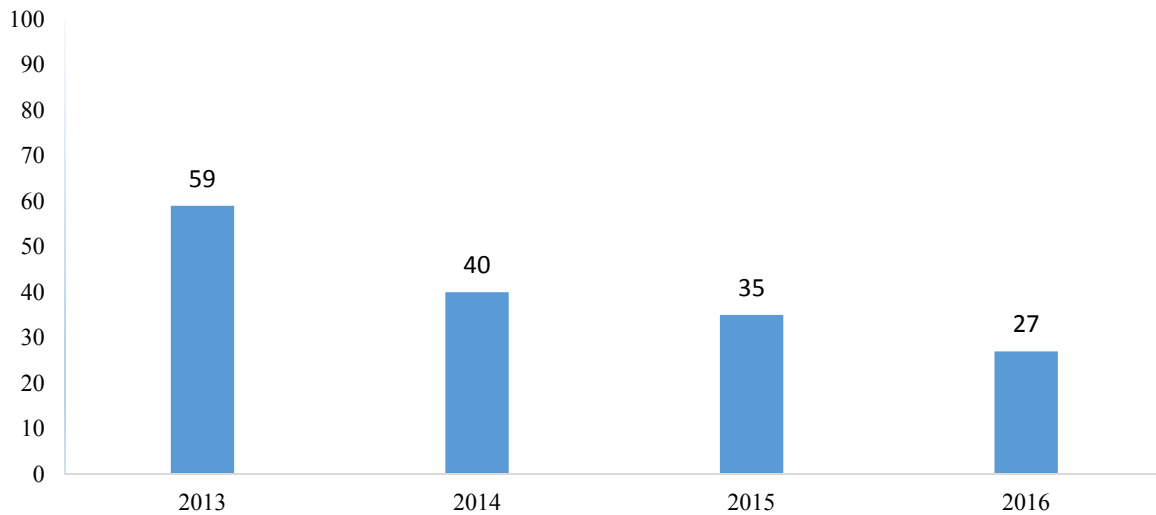


Figure 3: Total number of breaches for year 2013, 2014, 2015 and 2016

The Mean, Median and Standard Deviance of the loss amount for type of breach are presented in Table 4. The type of breach “Hacking” recorded the highest average mean loss of breaches of 13.77. The type of breach “Malware” recorded the highest standard deviation loss of 3.51.

The average mean loss of the overall data is 12.51 and the standard deviation loss is 2.05.

Type of Breach	Mean Loss	Median Loss	Standard deviation Loss
Error	11.84	12.01	2.31
Hacking	13.77	13.53	3.36
Malware	9.50	8.17	3.51
Misuse	12.53	12.21	2.94
Physical	12.07	12.43	2.84
Social	13.72	13.72	2.72
Overall data	12.51	12.44	2.05

Table 4: Results of the Mean, Median and Standard Deviance of the loss amount per type of breach.

The Mean, Median and Standard Deviance of the loss amount for each year are presented in Table 5. The average mean loss of breaches is approximately similar for all years. The year 2016 recorded the highest standard deviation loss of 4.45.

Year	Mean Loss	Median Loss	Mode Loss	Variance Loss	Sd loss
2013	12.14	12.51	11.56	10.15	3.19
2014	12.18	12.39	17.86	9.88	3.14
2015	12.48	12.21	14.77	9.05	3.01
2016	11.67	11.54	17.50	19.82	4.45

Table 5: Results of the Mean, Median, Mode, Variance and Standard Deviance of the loss amount per year.

4.3 Fitting distributions

4.3.1 Introduction to Fitting distributions

Fitting a distribution to data entails selecting an appropriate probability distribution for modeling the random variable as well as estimating the distribution's parameters. Several distributions such as normal, weibull, lognormal, and gamma distributions are used in this study.

First, their parameter estimates, estimated standard errors, log likelihood and akaike information criteria AIC are calculated. Then, the following plots will be presented:

- Q-Q plot of the empirical quantiles on the y-axis against the theoretical quantiles on the x-axis.
- P-P plot of the empirical distribution function evaluated at each data point on the y-axis against the fitted distribution function on the x-axis.

Noting that the Q-Q plot emphasizes the lack-of-fit at the distribution tails whereas the p-p plot emphasizes the lack-of-fit at the distribution center.

Finally, the P-value of the Kolmogorov-Smirnov (KS) test is calculated for each distribution. If the p-value is less than 0.05, we reject the null hypothesis H_0 . Thus, we have sufficient evidence to say that the sample data does not come from the pre-specified theoretical distribution under the null hypothesis H_0 .

This procedure is performed to fit the distribution of the loss amount as follows:

1. On all data set
2. On each year

4.3.2 Fitting distribution on all data set

Table 6 presents the fitted parameter, AIC, KS p-value of the fitting distribution procedure on all data set.

Fitted distribution	Parameter 1	Parameter 2	AIC	KS p-value	Decision
Normal distribution	12.51	3.07	1407.08	0.060547	Accept
Lognormal distribution	2.49	0.27	1437.32	0.000491	Reject
Weibull distribution	4.17	13.68	1425.67	0.00825	Reject
Gamma distribution	15.27	1.22	1417.32	0.005126	Reject

Table 6: Results of the fitted distributions of the overall impact loss amount of breaches.

The overall data of the loss amount of breaches in the VERIS dataset may fit a normal distribution since its P-value is greater than 0.05. Thus, in this study the normal distribution is chosen as the fitted distribution since it has the lowest AIC.

Figure 4 presents the Q-Q plot of the normal, lognormal, weibull and gamma distributions, on the loss amount of the breaches.

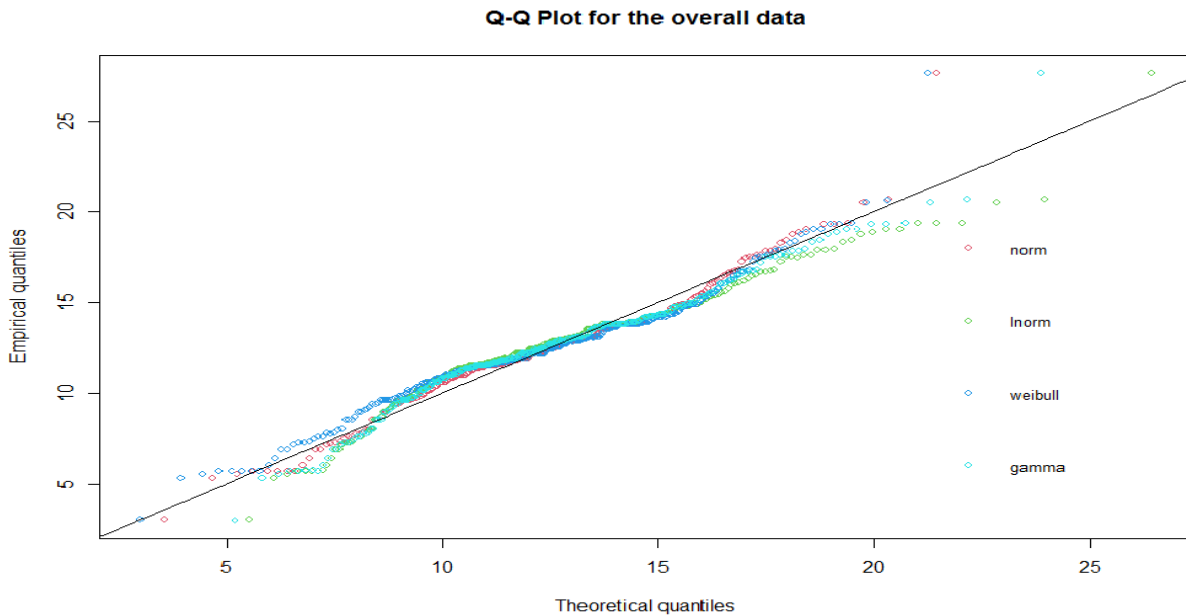


Figure 4: Q-Q plot of the overall impact loss amount of breaches.

The points of the normal distributions are the closer to the diagonal line in the q-q plot indicating that is the best fit in line with the KS test.

Figure 5 presents the P-P plot of the normal, lognormal weibull and gamma on the loss amount of the breaches

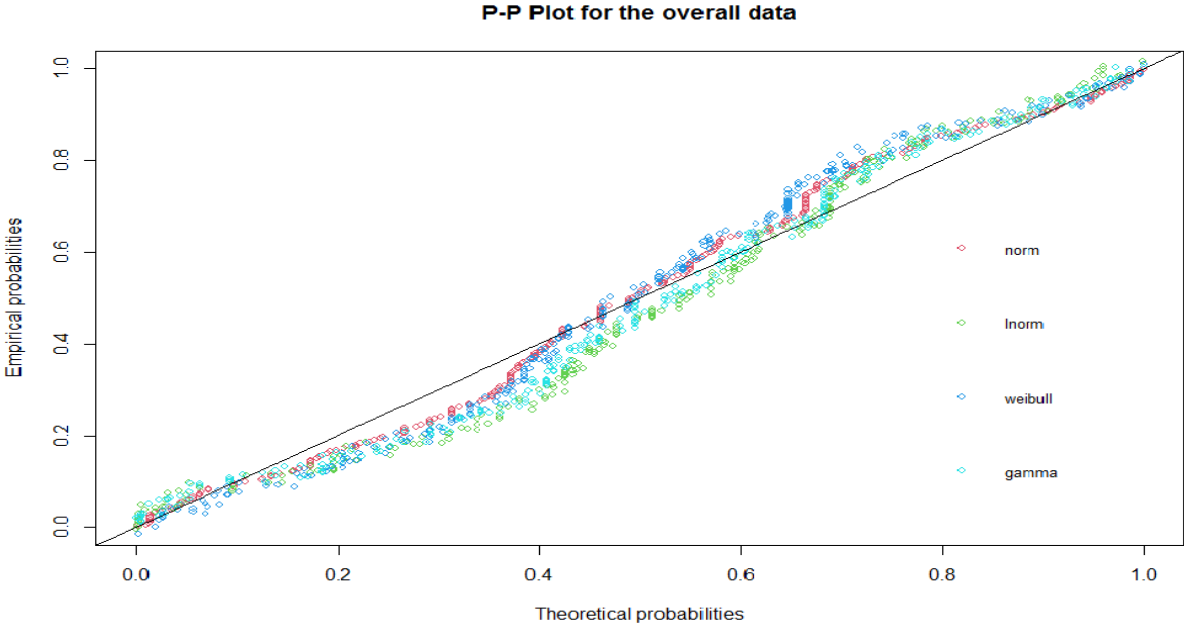


Figure 5: P-P plot of the overall impact loss amount of breaches.

The points of the normal distributions are the closer to the diagonal line in the P-P plot indicating that is the best fit in line with the KS test.

Figure 6 presents the Probability density function of the overall Impact loss amount of a breach.

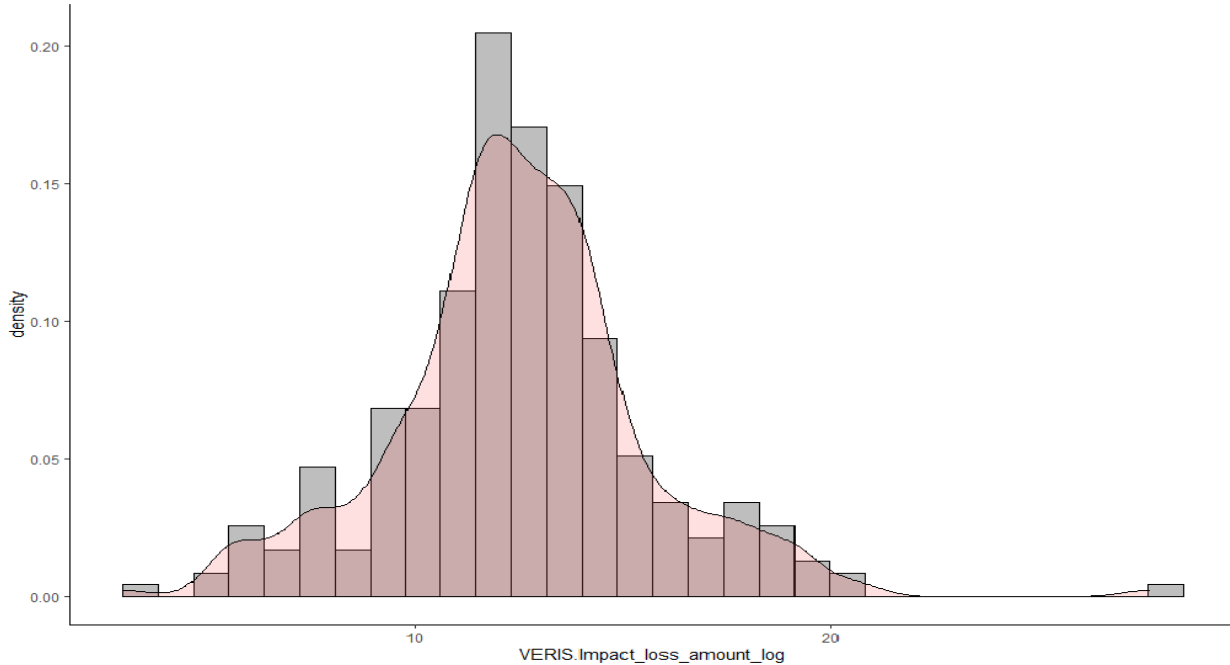


Figure 6: Probability density function of the overall Impact loss amount of a breach

The shape of the Probability density function of the overall Impact loss amount of a breach is similar to the shape of the Probability density function of the normal distribution.

4.3.3 Fitting distribution on the data set of each year

4.3.3.1 Fitting distribution on the data set of year 2013

Table 7 presents the fitted parameter, AIC, KS p-value of the fitting distribution procedure on the data set of year 2013.

Fitted distribution	Parameter 1	Parameter 2	AIC	KS p-value	Decision
Normal distribution	12.14	3.16	307.15	0.652375	Accept
Lognormal distribution	2.45	0.32	326.68	0.112922	Accept
Weibull distribution	4.45	13.30	306.04	0.662046	Accept
Gamma distribution	11.64	0.96	317.77	0.232882	Accept

Table 7: Results of the fitted distributions of the impact loss amount of breaches for year 2013.

The data of the loss amount of breaches for year 2013 in the VERIS dataset may fit a normal, lognormal, weibull or a gamma distribution since their P-values are greater than 0.05. In this study the weibull distribution is chosen as the fitted distribution since it has the lowest AIC.

Figure 7 presents the Q-Q plot of the normal, lognormal, weibull and gamma, on the loss amount of the breaches for year 2013.

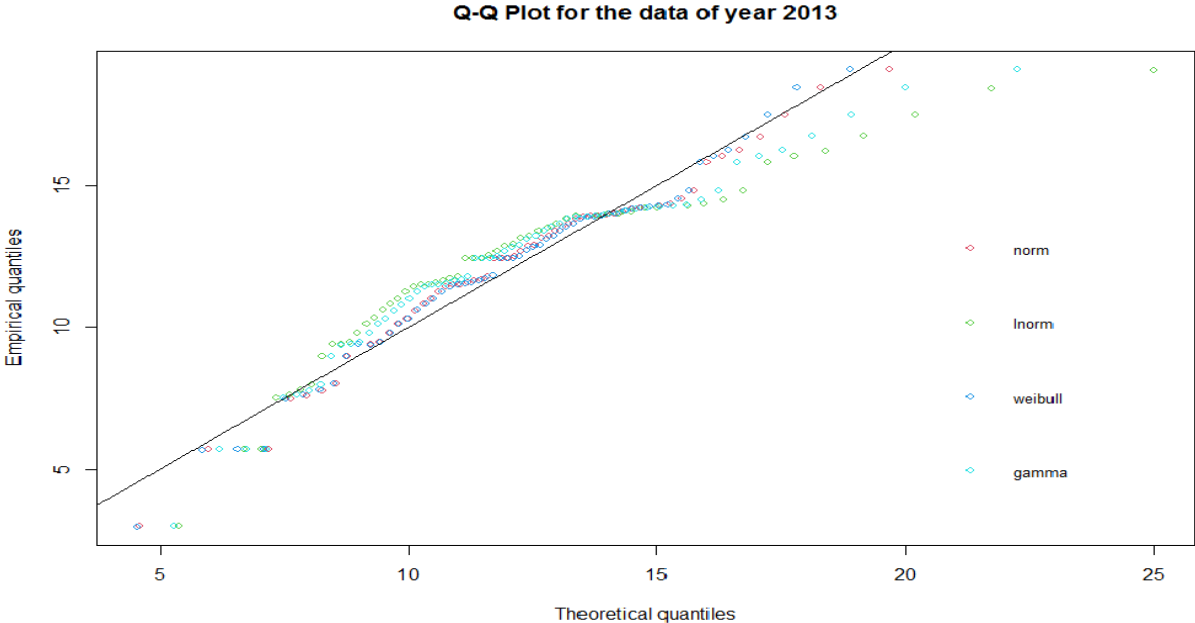


Figure 7: Q-Q plot of the impact loss amount of breaches for year 2013.

The points of the weibull distributions are the closer to the diagonal line in the Q-Q plot indicating that is the best fit in line with the KS test.

Figure 8 presents the P-P plot of the normal, lognormal weibull and gamma on the loss amount of the breaches for year 2013.

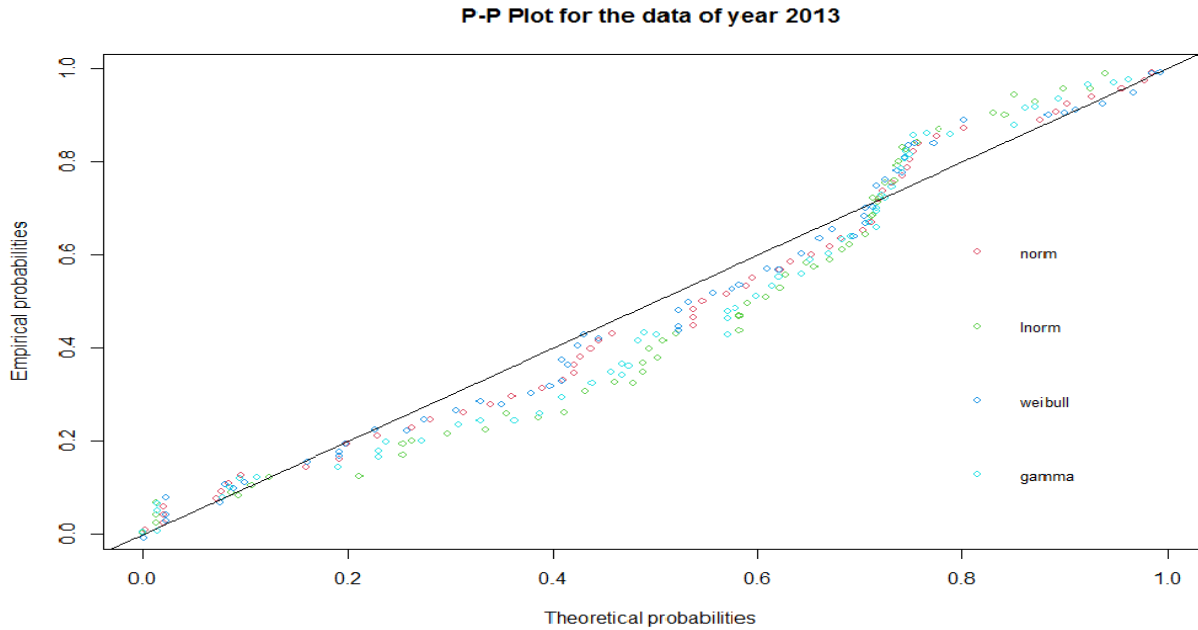


Figure 8: P-P plot of the impact loss amount of breaches for year 2013.

The points of the weibull distributions are the closer to the diagonal line in the P-P plot indicating it is the best fit in line with the KS test.

4.3.3.2 Fitting distribution on the data set of year 2014

Table 8 presents the fitted parameter, AIC, KS p-value of the fitting distribution procedure on the data set of year 2014.

Fitted distribution	Parameter 1	Parameter 2	AIC	KS p-value	Decision
Normal distribution	12.18	3.10	208.12	0.764426	Accept
Lognormal distribution	2.46	0.28	212.85	0.493206	Accept
Weibull distribution	4.27	13.36	208.85	0.602672	Accept
Gamma distribution	13.94	1.14	210.15	0.688467	Accept

Table 8: Results of the fitted distributions of the impact loss amount of breaches for year 2014.

The data of the impact loss amount of breaches for year 2014 in the VERIS dataset may fit a normal, lognormal, weibull or a gamma distribution since their P-values are greater than 0.05. In this study the normal distribution is chosen as the fitted distribution since it have the lowest AIC. Figure 9 presents the Q-Q plot of the normal, lognormal, weibull and gamma, on the loss amount of the breaches for year 2014.

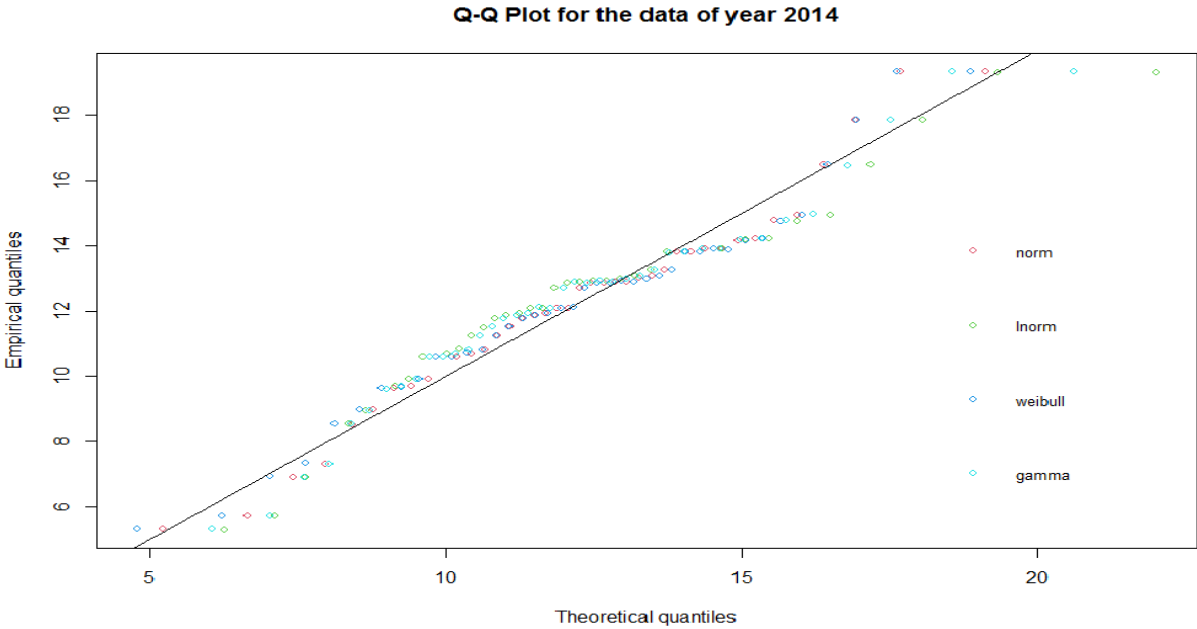


Figure 9: Q-Q plot of the impact loss amount of breaches for year 2014.

The points of the normal distribution are the closer to the diagonal line in the Q-Q plot indicating that they are the best fit in line with the KS test.

Figure 10 presents the P-P plot of the normal, lognormal weibull and gamma on the loss amount of the breaches for year 2014.

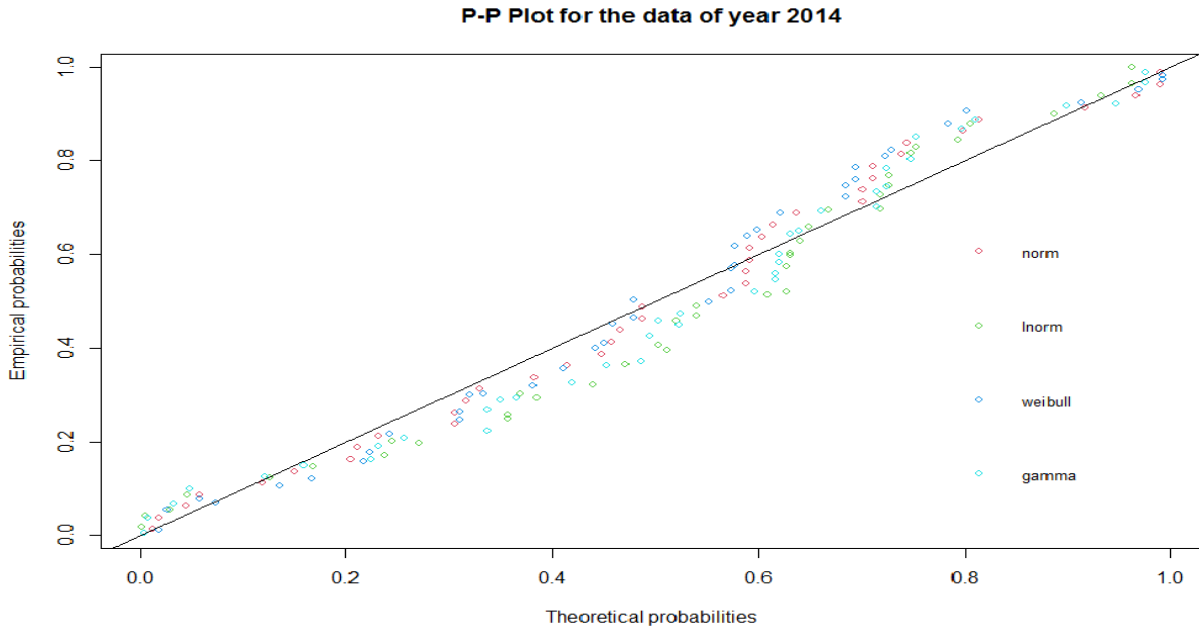


Figure 10: P-P plot of the impact loss amount of breaches for year 2014.

The points of the normal distribution are the closer to the diagonal line in the P-P plot indicating that they are the best fit in line with the KS test.

4.3.3.3 Fitting distribution on the data set of year 2015

Table 9 presents the fitted parameter, AIC, KS p-value of the fitting distribution procedure on the data set of year 2015.

Fitted distribution	Parameter 1	Parameter 2	AIC	KS p-value	Decision
Normal distribution	12.48004	2.964725	179.4006	0.799368	Accept
Lognormal distribution	2.492501	0.261067	183.7924	0.34054	Accept
Weibull distribution	4.756898	13.63097	179.2093	0.683601	Accept
Gamma distribution	15.97903	1.280195	181.5496	0.489348	Accept

Table 9: Results of the fitted distributions of the impact loss amount of breaches for year 2015.

The data of the loss amount of breaches for year 2015 in the VERIS dataset may fit a normal, lognormal, weibull or a gamma distribution since their P-values are greater than 0.05. In this study the weibull distribution are chosen as the fitted distribution since it has the lowest AIC.

Figure 11 presents the Q-Q plot of the normal, lognormal, weibull and gamma, on the loss amount of the breaches for year 2015.

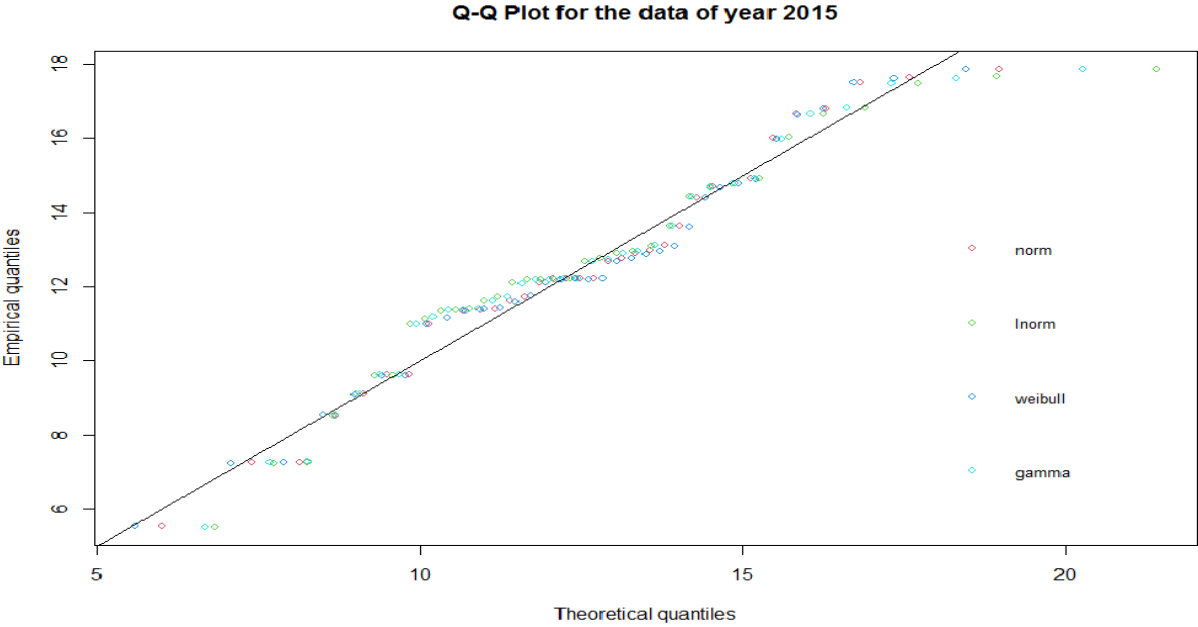


Figure 11: Q-Q plot of the impact loss amount of breaches for year 2015.

The points of the weibull distributions are the closer to the diagonal line in the Q-Q plot indicating that it is the best fit in line with the KS test.

Figure 12 presents the P-P plot of the normal, lognormal weibull and gamma on the loss amount of the breaches for year 2015.

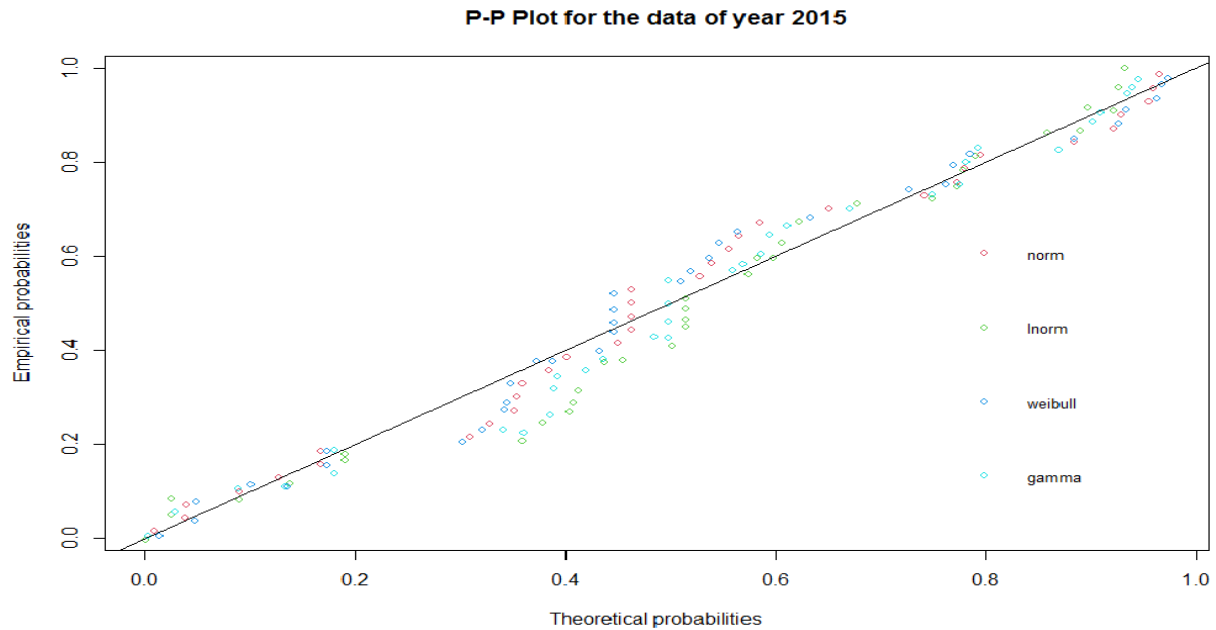


Figure 12: P-P plot of the impact loss amount of breaches for year 2015.

The points of the weibull distributions are the closer to the diagonal line in the P-P plot indicating that it is the best fit in line with the KS test.

4.3.3.4 Fitting distribution on the data set of year 2016

Table 10 presents the fitted parameter, AIC, KS p-value of the fitting distribution procedure on the data set of year 2016.

Fitted distribution	Parameter 1	Parameter 2	AIC	KS p-value	Decision
Normal distribution	11.66	4.36	160.24	0.551959	Accept
Lognormal distribution	2.39	0.34	153.08	0.672962	Accept
Weibull distribution	2.70	13.08	159.41	0.590899	Accept
Gamma distribution	8.22	0.70	154.15	0.864803	Accept

Table 10: Results of the fitted distributions of the impact loss amount of breaches for year 2016.

The data of the loss amount of breaches for year 2016 in the VERIS dataset may fit a normal, lognormal, weibull or a gamma distribution since their P-values are greater than 0.05. In this study the lognormal distribution is chosen as the fitted distribution since it has the lowest AIC.

Figure 13 presents the Q-Q plot of the normal, lognormal, weibull and gamma, on the loss amount of the breaches for year 2016.

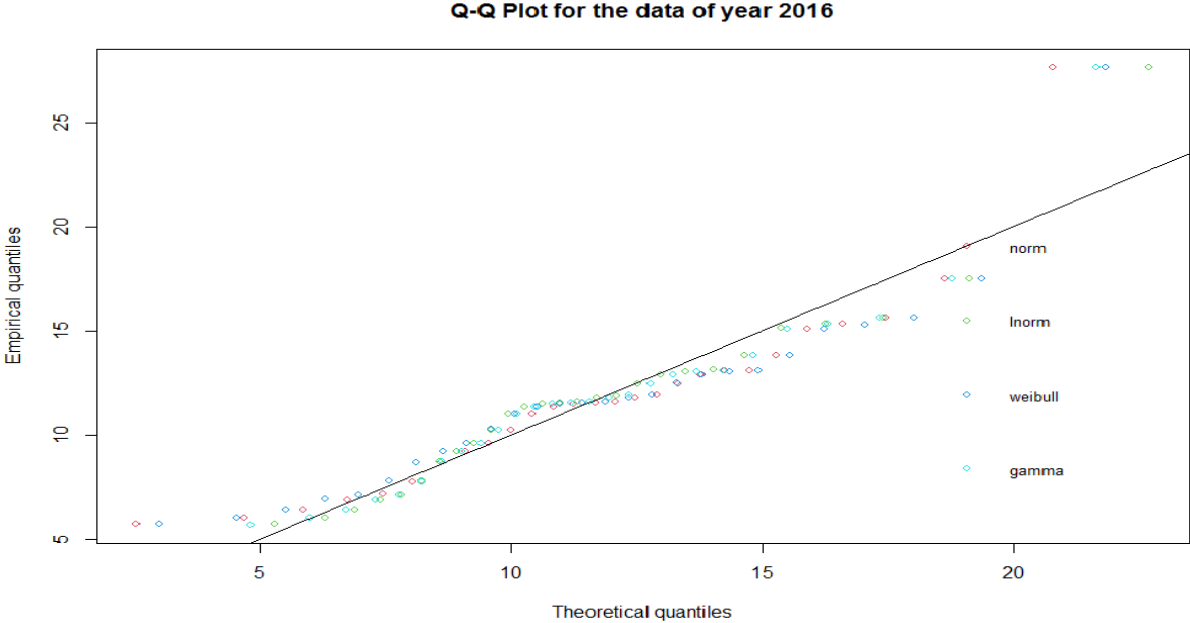


Figure 13: Q-Q plot of the impact loss amount of breaches for year 2016.

The points of the lognormal distribution are the closer to the diagonal line in the Q-Q plot indicating that it is the best fit in line with the KS test.

Figure 14 presents the P-P plot of the normal, lognormal weibull and gamma on the loss amount of the breaches for year 2016.

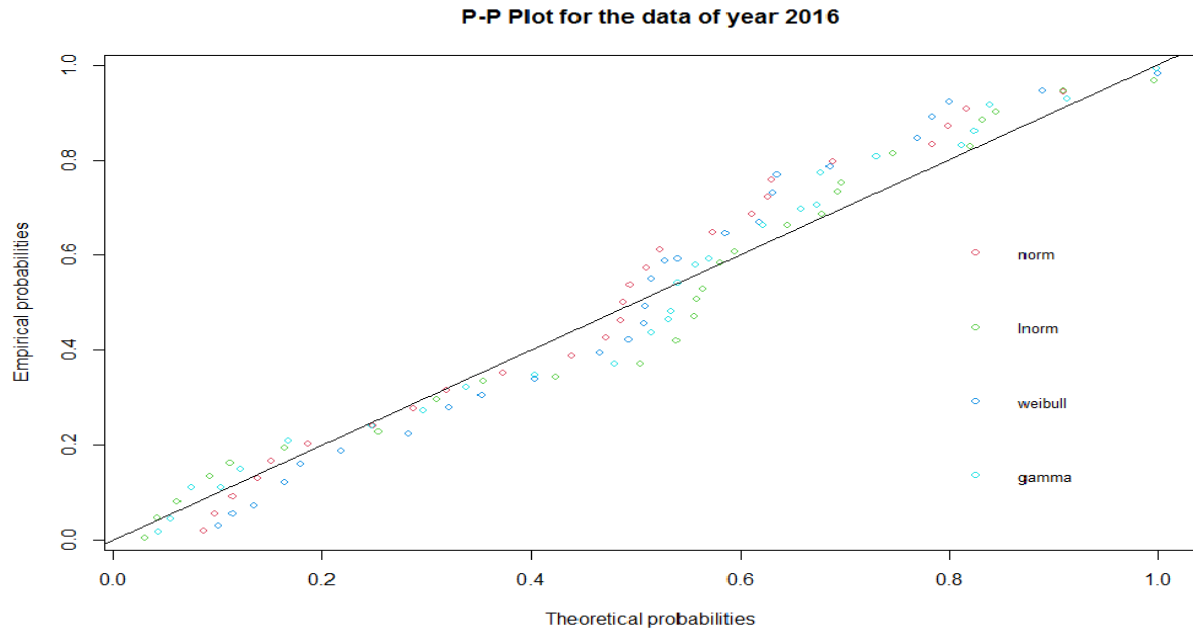


Figure 14: P-P plot of the impact loss amount of breaches for year 2016.

The points of the lognormal distributions are the closer to the diagonal line in the P-P plot indicating that it is the best fit in line with the KS test.

4.3.4 Conclusion

In section 4.1, we fitted the distribution on the overall data and yearly data for two purposes:

1. Calculate the probability density function
2. Apply the Generalized Linear model

The fitted distribution are:

- For overall data: Normal distribution
- For year 2013: Weibull distribution
- For year 2014: Normal distribution
- For year 2015: Weibull distribution
- For year 2016: Lognormal distribution

However, since the Generalized Linear model is applied only to the distribution that belong to the exponential dispersion family(EDM), since the Weibull distribution does not belong to the EDM family, thus we will apply the GLM to the distribution that belong to the EDM family and have the lowest AIC. Thus, the fitted distribution to the GLM are the following:

- For overall data: Normal distribution
- For year 2013: Normal distribution
- For year 2014: Normal distribution
- For year 2015: Normal distribution
- For year 2016: Lognormal distribution

4.4 Radom Forest

Random forests are a statistical learning method used in many fields of application and are adapted to both supervised classifications problems and regressions problems for qualitative and quantitative explanatory variables together without preprocessing. Moreover, Random forest involves determining a subset of the input variables that are important and active in explaining the input–output relationship. Thus, in this study, we used random forest for the purpose of ranking the variables, from the most important to the least important and selecting the important variables.

We applied the Random forests algorithm in this study to select and rank relevant the variables. In addition to generation of the feature ranking, the Random forests algorithm classifies features into three types: Confirmed, Tentative, Rejected.

In this study, in order to determine the most important variables associated to the Impact loss amount of breaches, Radom Forest was applied:

1. On the overall data set
2. On each year

4.4.1 Radom Forest on the overall data

The most important variables associated to the Impact loss amount of breaches in the overall VERIS data are: Action, Actor Motive, Asset Variety and Impact overall rating of breach.

The results of the important variable selection process using Random forests for the overall data in Table 11 show that, out of 12 variables, four features are confirmed, four of them are rejected and four features are marked tentative.

Attributes of all data	Mean Imp	Median Imp	Min Imp	Max Imp	Norm Hits	Decision
------------------------	----------	------------	---------	---------	-----------	----------

Action	1.939564	1.98204	-1.16234	4.17727	0.60521	Confirmed
Actor	1.486202	1.627141	-1.90815	4.015203	0.462926	Tentative
Actor Job change	-1.43487	-1.91509	-3.71758	1.364827	0	Rejected
Actor Motive	3.023557	3.135606	-2.6976	7.622461	0.797595	Confirmed
Asset Variety	2.941667	3.073702	-1.58557	6.110115	0.815631	Confirmed
Actor Country	1.530922	1.567441	-1.2412	3.295682	0.460922	Tentative
Vector	1.600348	1.676884	-1.83772	3.949662	0.472946	Tentative
Variety	0.603957	0.69382	-2.61614	2.667121	0.012024	Rejected
Actor Variety	1.471256	1.597318	-1.95171	4.862392	0.430862	Tentative
Victim country	1.190812	1.321239	-1.41697	3.383747	0.072144	Rejected
Victim Employee Count	0.119336	0.455071	-1.71535	0.915919	0.002004	Rejected
Impact overall rating	2.030967	2.106408	-1.21446	4.364237	0.635271	Confirmed

Table 11: Results of the important variable selection process using Random forests for the overall data

Table 11, the meaning of columns is as follows:

- Mean Imp: the mean of the importance.
- Median Imp: the median of the importance.
- Min Imp: the minimum of the importance.
- Max Imp: the maximum of the importance.
- Norm Hits: the number of hits normalized to number of importance source runs, where important variables is the importance measure computed over multiple iterations.

The resulting graph in Figure 15 generated using the Random forests package in the R program, indicates the level of importance of each variable on the vertical Y axis of the analyzed variables on the horizontal X axis by ranking them.

Figure 15 shows the ranking of the variables of the overall data from the least important variable to the most important variable.

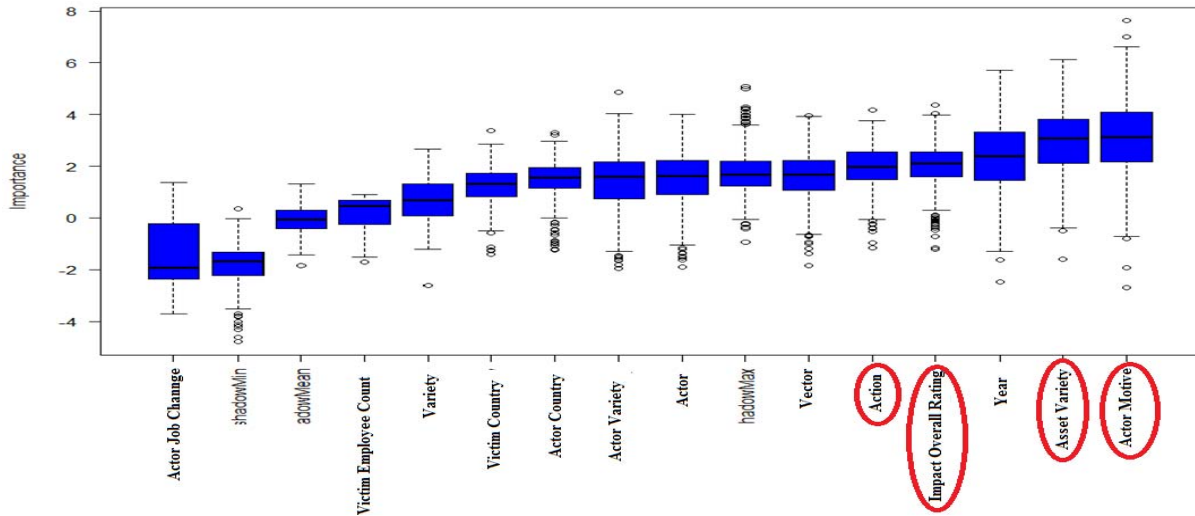


Figure 15: Variable importance selection graph using the Random forests algorithm on the overall data

4.4.2 Radom Forest on each year

4.4.2.1 Radom Forest on year 2013

The most important variables associated to the loss amount of breaches in year 2013 are: Action and Actor Country.

The results of the important variable selection process using Random Forest for Year 2013 in Table 12 shows that, out of 11 variables, two features are confirmed, eight of them are rejected and one features are marked tentative.

Attributes 2013	Mean Imp	Median Imp	Min Imp	Max Imp	Norm Hits	Decision
Action	2.772174	2.718084	-2.98117	7.894898	0.645291	Confirmed
Actor	0.26029	0.525724	-1.54013	1.922371	0.004008	Rejected
Actor Job change	0.567141	1.00189	-1.001	1.419496	0.006012	Rejected
Actor Motive	0.915077	0.870481	-0.52696	2.585596	0.004008	Rejected
Asset Variety	0.544866	0.487726	-0.63335	2.124574	0	Rejected

Actor Country	3.71067	3.750664	-1.69535	7.587299	0.771543	Confirmed
Variety	2.139153	2.261745	-3.58676	6.970542	0.531062	Tentative
Actor Variety	0.561011	0.627709	-3.58037	3.017593	0.008016	Rejected
Victim country	1.124575	1.366718	-2.53966	4.279988	0.032064	Rejected
Victim Employee Count	1.063677	0.80201	-1.57359	5.341147	0.022044	Rejected
Impact overall rating	1.241235	1.378603	-1.87723	3.911378	0.04008	Rejected

Table 12: Results of the important variable selection process using Random Forest for Year 2013.

Figure 16 shows the ranking of the variables of the year 2013 from the least important variable to the most important variable.

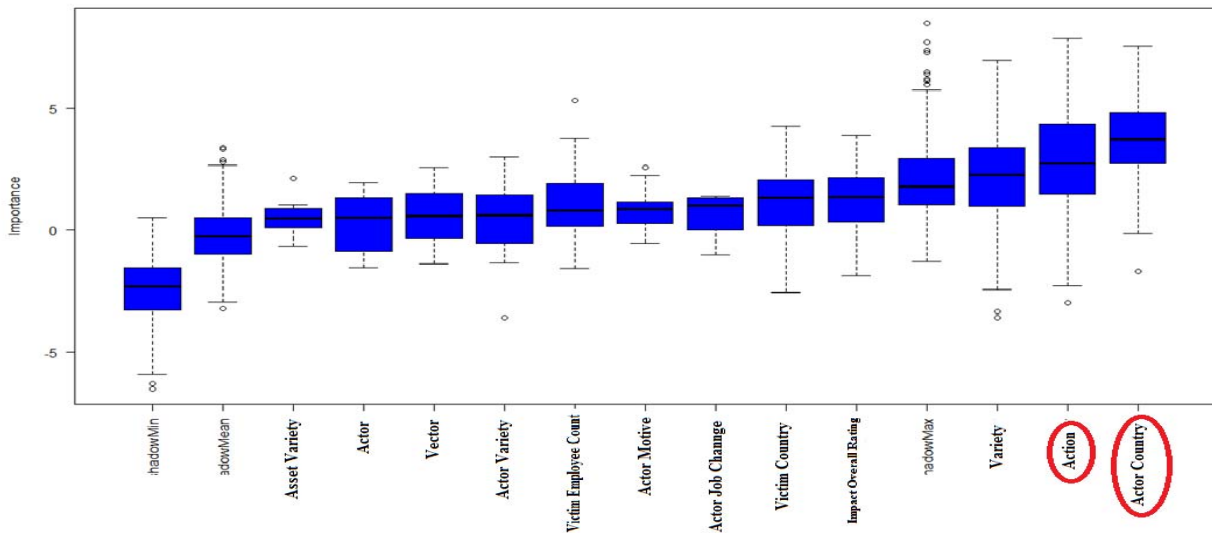


Figure 16: Variable importance selection graph using the Random Forest algorithm Year 2013.

4.4.2.2 Radom Forest on year 2014

The most important variables associated to the Impact loss amount of breaches in year 2014 are: Action, Actor, Actor Country, Actor Motive, and Actor Variety.

The results of the important variable selection process using Random Forest for Year 2014 in Table 13 show that, out of 11 variables, five features are confirmed, four of them are rejected and two features are marked tentative.

Attributes 2014	Mean Imp	Median Imp	Min Imp	Max Imp	Norm Hits	Decision
Action	5.103649	5.186105	2.008472	7.408126	0.961924	Confirmed
Actor	2.544368	2.567243	-1.51704	4.57887	0.693387	Confirmed
Actor Job change	0.051995	0	-1.04644	1.573751	0	Rejected
Actor Motive	4.090782	4.115516	-0.56373	7.161377	0.913828	Confirmed
Asset Variety	0.941943	1.011781	-1.15224	2.795975	0.014028	Rejected
Actor Country	3.346285	3.467217	-0.83229	6.172511	0.795591	Confirmed
Variety	2.222895	2.161499	-1.84683	7.415378	0.539078	Tentative
Actor Variety	3.476779	3.492707	-0.33372	6.322986	0.819639	Confirmed
Victim country	1.665043	1.701127	-1.59042	4.379066	0.43487	Tentative
Victim Employee Count	0.569568	0.425989	-0.91326	2.429987	0.002004	Rejected
Impact overall rating	0.554755	0.98537	-1.28611	2.242005	0.006012	Rejected

Table 13: Results of the important variable selection process using Random Forest for Year 2014.

Figure 17 shows the ranking of the variables of the year 2014 from the least important variable to the most important variable.

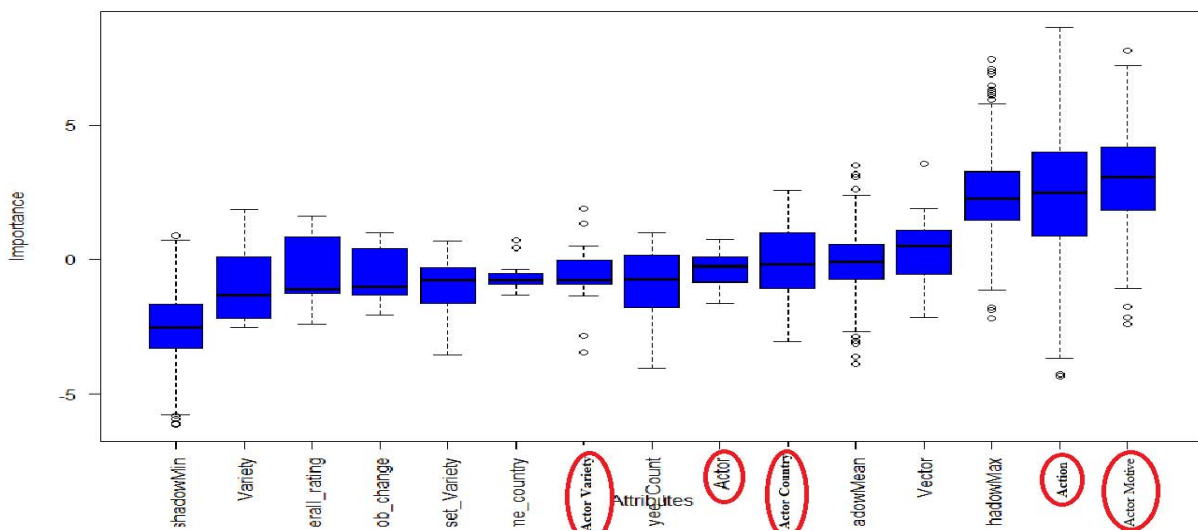


Figure 17: Variable importance selection graph using the Random Forest algorithm Year 2014.

4.4.2.3 Radom Forest on year 2015

The most important variables associated to the Impact loss amount of breaches in year 2015 are: Actor Motive.

The results of the important variable selection process using Random Forest for Year 2015 in Table 14 show that, out of 11 variables, one feature is confirmed, ten of them are rejected and one features are marked tentative.

Attributes 2015	Mean Imp	Median Imp	Min Imp	Max Imp	Norm Hits	Decision
Action	2.378927	2.479711	-4.35712	8.609157	0.52505	Tentative
Actor	-0.34847	-0.26746	-1.64255	0.749827	0	Rejected
Actor Job change	-0.55714	-1.00097	-2.04472	0.995632	0	Rejected
Actor Motive	3.067028	3.053289	-2.4134	7.763122	0.633267	Confirmed
Asset Variety	-1.03736	-0.79127	-3.56242	0.6839	0	Rejected
Actor Country	-0.01228	-0.18766	-3.07782	2.578567	0.008016	Rejected
Variety	-0.86821	-1.34774	-2.55969	1.869778	0	Rejected
Actor Variety	-0.63845	-0.77023	-3.45912	1.909605	0.002004	Rejected
Victim country	-0.60404	-0.78216	-1.33652	0.72033	0	Rejected
Victim Employee Count	-0.97049	-0.76682	-4.05697	0.972634	0	Rejected
Impact overall rating	-0.43765	-1.11981	-2.40771	1.596163	0.002004	Rejected

Table 14: Results of the important variable selection process using Random Forest for Year 2015.

Figure 18 shows the ranking of the variables of the year 2015 from the least important variable to the most important variable.

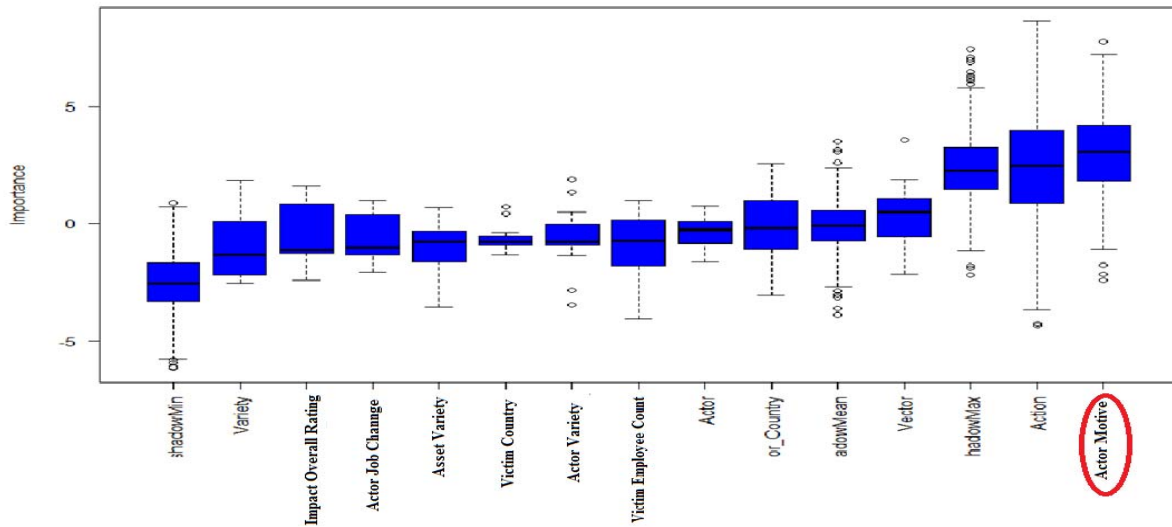


Figure 18: Variable importance selection graph using the Random Forest algorithm Year 2015.

4.4.2.4 Radom Forest on year 2016

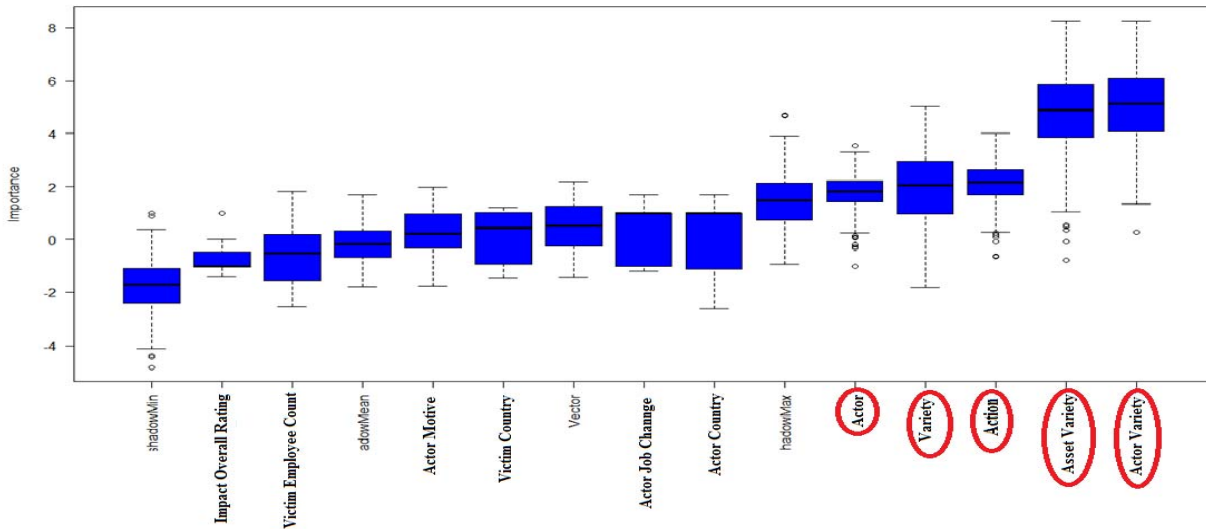
The most important variables associated to the Impact loss amount of breaches in year 2016 are: Action, Actor, Actor Variety, Asset Variety and Variety.

The results of the important variable selection process using Random Forest for Year 2015 in Table 15 show that, out of 11 variables, five features are confirmed, six of them are rejected.

Attributes 2016	Mean Imp	Median Imp	Min Imp	Max Imp	Norm Hits	Decision
Action	2.140336	2.161116	-0.62456	4.014966	0.691406	Confirmed
Actor	1.761744	1.816274	-1.00426	3.537311	0.601563	Confirmed
Actor Job change	0.18119	0.998319	-1.19765	1.6774	0	Rejected
Actor Motive	0.282092	0.225222	-1.76466	1.974021	0.003906	Rejected
Asset Variety	4.734954	4.902964	-0.7805	8.267939	0.953125	Confirmed
Actor Country	0.037695	1.001	-2.61832	1.675678	0.011719	Rejected
Variety	1.961308	2.05836	-1.82558	5.035086	0.628906	Confirmed
Actor Variety	5.034025	5.122577	0.271975	8.267894	0.976563	Confirmed

Victim country	0.0335	0.433942	-1.4349	1.212316	0	Rejected
Victim Employee Count	-0.67647	-0.51122	-2.53438	1.831339	0	Rejected
Impact overall rating	-0.61151	-1.001	-1.38539	1.000997	0	Rejected

Table 15: Results of the important variable selection process using Random Forest for Year 2016.



shows the ranking of the variables of the year 2016 from the least important variable to the most important variable.

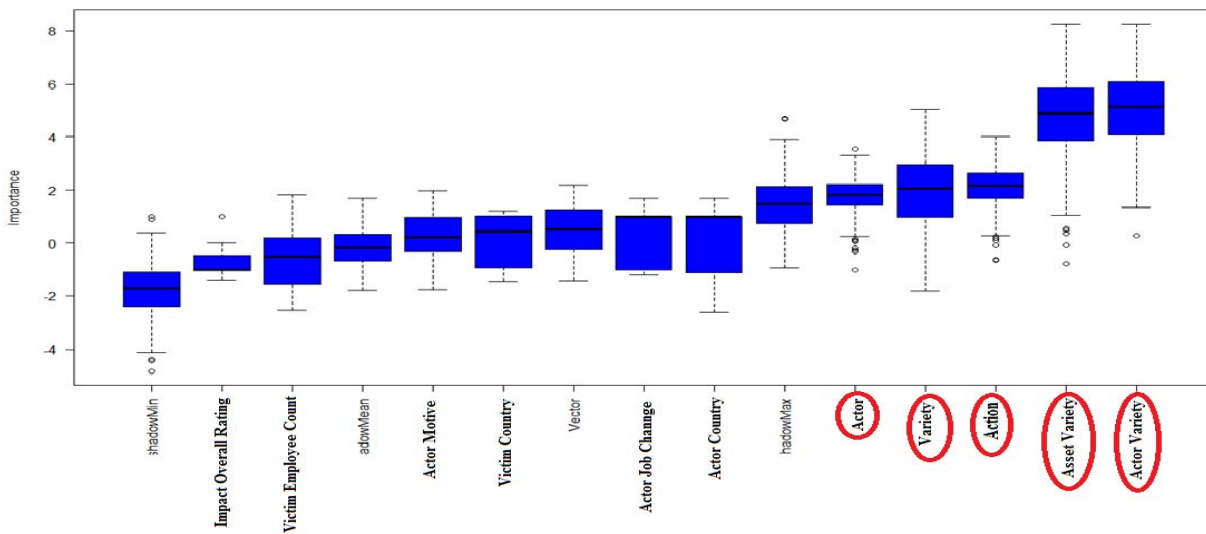


Figure 19: Variable importance selection graph using the Random Forest algorithm Year 2016.

Table 16 summarizes all the results of the important variable selection process using Radom forest for Year 2013, 2014, 2015 and 2016.

Year	Most Important Variables
2013	Action and Actor Country
2014	Action, Actor, Actor Country, Actor Motive, Actor Variety
2015	Actor Motive
2016	Action, Actor, Actor Variety, Asset Variety and Variety

Table 16: Results of the important variable selection process using Random Forest for Year 2013, 2014, 2015 and 2016

4.5 Generalized Linear Model

In this study, the Generalized Linear Model (GLM) of the fitted distribution in section 4.1 was applied on the most important variable provided by the Random Forest in section 4.2 in order to estimate and predict loss amount due to cyber risk.

4.5.1 Generalized Linear Model on the overall data

In section 4.1, the best fitted model of the overall impact loss amount is the normal distribution. In section 4.2, the most important Action, Actor Motive, Asset Variety, and the Impact overall rating. Therefore, we applied the Generalized Linear Model to the normal distribution. We excluded the outliers and influential points from the data using the cook's distance in order to get a high accuracy of the predicted model.

a. Coefficients estimates and variable significance

Table 17 provides the estimated values of the parameters in the fitted Gaussian Model and their significance level. The most significant variables are: the intercept, Action Hacking, Asset Variety Authentication,

Coefficients	Estimate	Std. Error	t value	Pr(> t)	Significance
(Intercept)	15.0629	1.70783	8.82	8.68E-16	***
Action Hacking	2.63141	0.65735	4.003	9.06E-05	***
Action Malware	0.09927	1.32888	0.075	0.94053	
Action Misuse	1.67289	0.64427	2.597	0.01018	*
Action Physical	1.18617	0.82238	1.442	0.1509	
Action Social	0.36908	3.26081	0.113	0.91001	
Action Unknown	-0.10215	2.88771	-0.035	0.97182	
Actor Motive Espionage	-0.15748	1.95903	-0.08	0.93602	
Actor Motive Financial	-3.27337	1.41753	-2.309	0.02204	*
Actor Motive Fun	-0.91409	1.84748	-0.495	0.62135	
Actor Motive Grudge	-6.27307	2.00487	-3.129	0.00204	**
Actor Motive Ideology	-1.28256	2.69588	-0.476	0.63482	
Actor Motive NA	-2.54857	2.7306	-0.933	0.35187	
Actor Motive Other	-4.70035	2.1531	-2.183	0.0303	*
Actor Motive Unknown	-2.15814	1.38189	-1.562	0.12007	
Asset Variety ATM	0.46241	1.00298	0.461	0.64532	
Asset Variety Authentication	11.83177	2.74551	4.309	2.66E-05	***
Asset Variety Call	-1.02836	2.34586	-0.438	0.66163	
Asset Variety Cashier	3.12402	4.22559	0.739	0.46066	
Asset Variety Customer	3.75105	1.28906	2.91	0.00406	**
Asset Variety Database	0.49298	0.69244	0.712	0.4774	
Asset Variety Desktop	1.42933	0.89411	1.599	0.11162	
Asset Variety Disk	-1.11924	2.71836	-0.412	0.68101	
Asset Variety Documents	0.67359	0.72109	0.934	0.35146	
Asset Variety End-user	-3.85543	1.30737	-2.949	0.0036	**
Asset Variety Executive	0.36644	2.35005	0.156	0.87626	
Asset Variety File	4.02253	2.35005	1.712	0.08864	.
Asset Variety Finance	3.63171	3.33197	1.09	0.27716	
Asset Variety Flash	-0.85537	1.73458	-0.493	0.62251	

Asset Variety Gas	1.03406	1.8095	0.571	0.56839	
Asset Variety Kiosk	-1.11149	2.41428	-0.46	0.64579	
Asset Variety Mail	0.93354	2.36768	0.394	0.69383	
Asset Variety Mainframe	5.45501	2.76366	1.974	0.0499	*
Asset Variety Partner	1.0268	2.35005	0.437	0.66268	
Asset Variety Payment	-2.32256	1.05483	-2.202	0.02892	*
Asset Variety PBX	5.09037	2.76366	1.842	0.0671	.
Asset Variety Peripheral	2.83066	2.36768	1.196	0.23341	
Asset Variety POS	0.35915	1.72547	0.208	0.83534	
Asset Variety Router	1.21393	2.35005	0.517	0.60609	
Asset Variety System	6.65362	2.35005	2.831	0.00515	**
Asset Variety Tapes	2.48582	2.36768	1.05	0.29514	
Asset Variety Unknown	0.87056	0.87487	0.995	0.32101	
Asset Variety Web	0.16332	0.71604	0.228	0.81983	
Impact overall rating Distracting	-2.5479	1.09921	-2.318	0.02155	*
Impact overall rating Insignificant	-2.16139	1.84384	-1.172	0.24262	
Impact overall rating Painful	-0.90918	1.33027	-0.683	0.49518	
Impact overall rating Unknown	-1.73759	0.85833	-2.024	0.04438	*

Table 17: Results of the coefficient for the Gaussian GLM on the overall data

b. Deviance Residuals for GLM:

The Null Deviance model in Table 18 refers to the model in which all of the terms are excluded, except the intercept. The degrees of freedom for this model are the number of data points n minus 1 if an intercept is fitted.

The Residual Deviance model in Table 18 refers to the fitted model. The degrees of freedom for this model are equal to n minus p , where p is the number of parameters including any intercept.

The dispersion parameter \emptyset is a measure of how much a sample fluctuates around its mean value and it is calculated by dividing the Residual deviance by its corresponding degrees of freedom.

In Table 18, the Residual Deviance has been reduced by 945.08 with a loss of 46 degrees of freedom when we added all the parameters to the model including the intercept.

The Residual Deviance is 940.04 on 184 degrees of freedom, thus the dispersion parameter \emptyset is equal to 5.108901.

Null deviance	1885.12 on 230 degrees of freedom
Residual deviance	940.04 on 184 degrees of freedom

Table 18: Results of the Deviance for the Gaussian GLM applied on the overall data

Table 19 illustrates the values of the Minimum, Maximum, Median, First and Third Quantile for the Gaussian GLM applied to the loss amount of the overall data.

Minimum	First Quantile	Median	Third Quantile	Maximum
-6.5424	-0.9802	0	0.9907	6.818

Table 19: Results of the Minimum, Maximum, Median, First and Second Quantile for the GLM of overall data.

Table 20 presents the value of the AIC and the Fisher Scoring for the Gaussian GLM applied on the overall data. The lower the AIC and the Fisher Scoring, the better the model fits.

AIC	1075.8
Number of Fisher Scoring iterations	2

Table 20: Results of the AIC and Fisher Scoring for the Gaussian GLM applied on the overall data.

Table 21 presents the value of the MAPE and RMSE for the Gaussian GLM applied on the overall data. The lower the MAPE and the RMSE, the higher the accuracy of the forecast of the model. The MAPE is 12.706322% which indicate a high accurate forecast of the model.

MAPE	12.7063226445
RMSE	2.819E-13

Table 21: Results of the MAPE and RMSE for the Gaussian GLM applied on the overall data.

c. Residuals Plots for GLM:

The Pearson residual plot is a graph that shows the residuals on the vertical axis and the fitted variables on the horizontal axis. If the model fits well, the residuals should show no pattern, just constant variability around zero for all values of the covariates X_i .

The Pearson residual plot in Figure 20 shows the residuals on the vertical axis and the fitted variables on the horizontal axis of the loss amount of the overall data. The more the variability of the residuals is constant around zero for all values of the covariates X_i , the better the model fits.

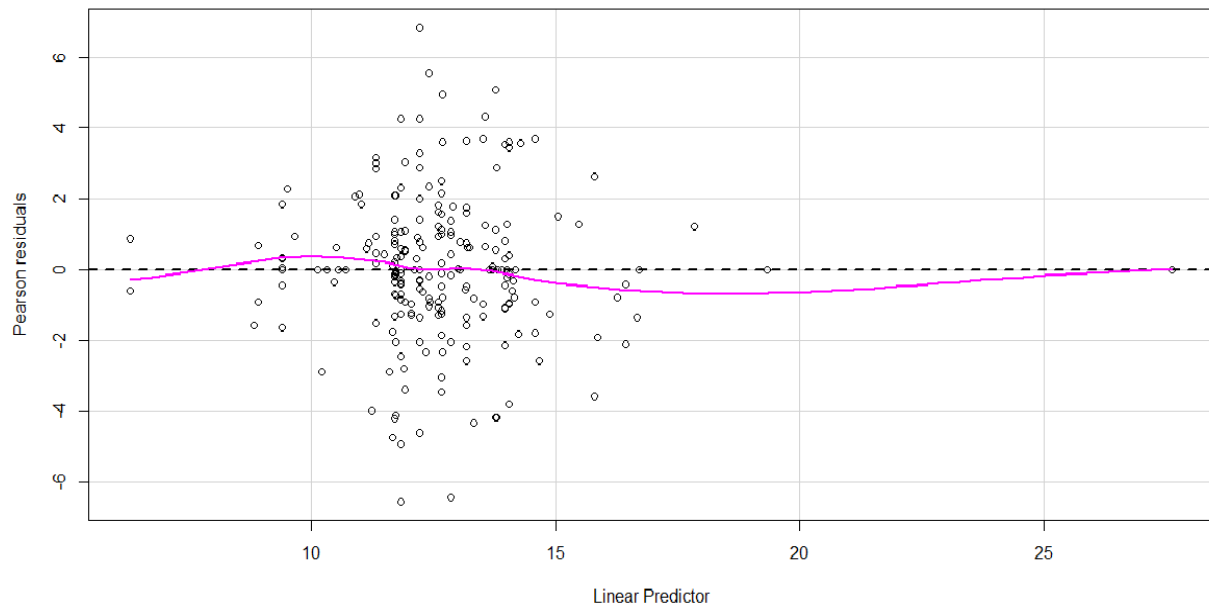
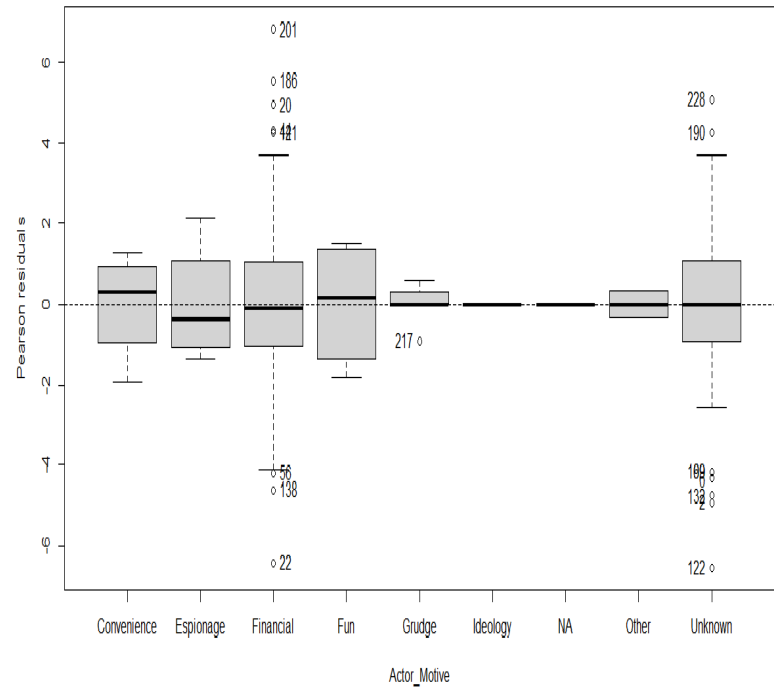
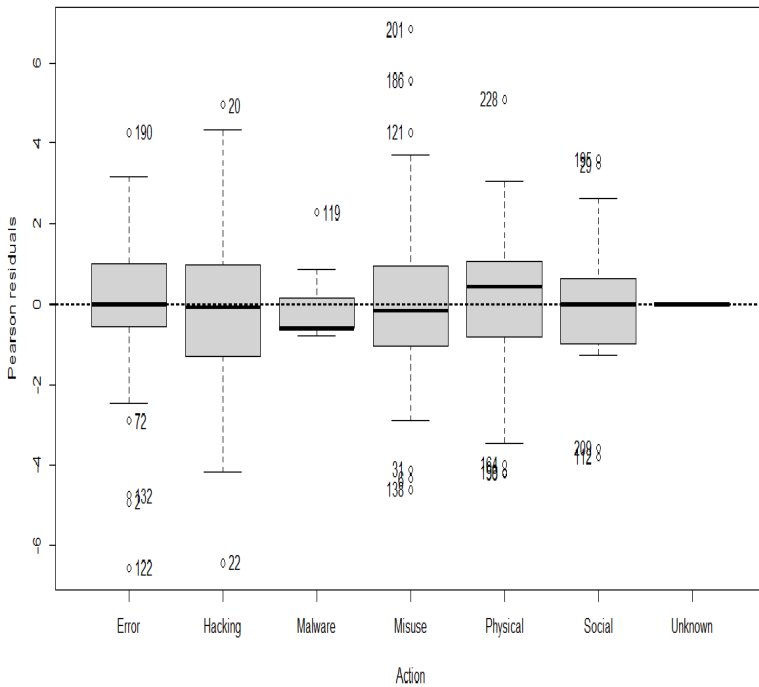


Figure 20: Pearson Residual Plot for Gaussian GLM applied on the overall data

d. Boxplots for GLM:

A boxplot is a graph that shows the Minimum, Maximum, Median, First and Third Quantile for each variable.

In the boxplots of Figure 21, outliers are denoted with an asterisk. The bottom whisker extends to the lowest value that is not an outlier and the upper whisker extends to the highest value that is not an outlier. The box represents the middle of observations with the lower end of the box at the 25th percentile (First Quartile) and the upper end of the box at the 75th percentile (Third Quartile). The line in the middle of the box represents the median.



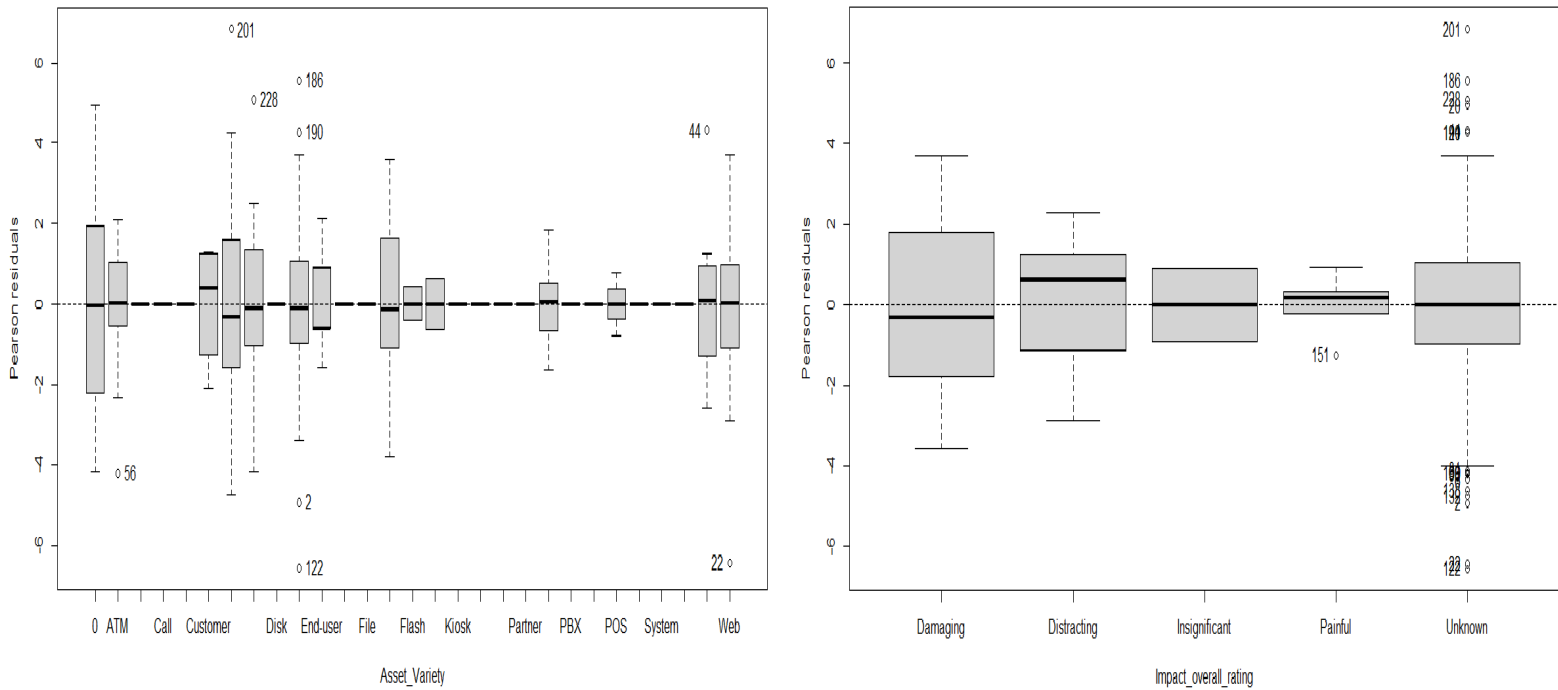


Figure 21: Boxplots of the important variables used in fitting the Gaussian GLM to the overall data

4.5.2 Generalized Linear Model on each year

4.5.2.1 Generalized Linear Model on each year 2013

In section 4.1, the best fitted model of the loss amount for year 2013 is the weibull distribution. However, the weibull distribution does not belong to the exponential dispersion family (EDM), thus the GLM cannot be applied to the weibull distribution. Therefore, in this case the GLM is applied to the Normal distribution since it has the lower AIC among the fitted distribution belonging to the exponential dispersion family (EDM) and the closer AIC to the AIC of the Weibull distribution.

In section 4.2, the most important variables selected for year 2013 are: Action and Actor Motive. Therefore, for year 2013, we applied the Generalized Linear Model to the normal distribution. We excluded the outliers and influential points from the data using the cook's distance in order to get a high accuracy of the predicted model.

a. Coefficients estimates and variable significance:

Table 22 provides the estimated values of the parameters in the fitted Gaussian Model applied on the loss amount of year 2013 and their significance level. The most significant variables are: the intercept, Actor Motive.

Coefficient	Estimate	Std Error	t value	Pr(> t)	Significance
(Intercept)	15.037	2.387	6.3	7.02E-07	***
Action Hacking	1.238	1.334	0.928	0.3609	
Action Misuse	-1.07	1.712	-0.625	0.5368	
Action Physical	-1.136	1.629	-0.697	0.4911	
Actor Motive Financial	-1.733	1.796	-0.965	0.3427	
Actor Motive Fun	5.096	2.352	2.167	0.0386	*
Actor Motive Unknown	-2.44	2.303	-1.06	0.2981	

Table 22: Results of the coefficient for the GLM applied on the loss amount of year 2014

b. Deviance Residuals for GLM:

In Table 23, the Residual Deviance has been reduced by 62.939 with a loss of 6 degrees of freedom when we added all the parameters to the model including the intercept.

The Residual Deviance is 80.219 on 29 degrees of freedom, thus the dispersion parameter \emptyset is equal to 2.76617.

Null deviance	143.158 on 35 degrees of freedom
Residual deviance	80.219 on 29 degrees of freedom

Table 23: Results of the Deviance for the GLM applied on the loss amount of year 2013

Table 24 illustrates the values of the Minimum, Maximum, Median, First and Third Quantile for the Gaussian GLM applied to the loss amount for 2013.

Minimum	First Quantile	Median	Third Quantile	Maximum
---------	----------------	--------	----------------	---------

-4.1612	-0.6547	-0.084	1.3965	2.6595
---------	---------	--------	--------	--------

Table 24: Results of the Min, Max, Median, First and Third Quantile for the GLM applied on the loss amount of year 2013.

Table 25 presents the value of the AIC and the Fisher Scoring for the Gaussian GLM applied on the loss amount of year 2013. The lower the AIC and the Fisher Scoring, the better the model fits.

AIC	147.01
Number of Fisher Scoring iterations	2

Table 25: Results of the AIC and Fisher Scoring for the Gaussian GLM applied on the loss amount of year 2013

Table 26 presents the value of the MAPE and RMSE for the Gaussian GLM applied on the loss amount of year 2013. The lower the MAPE and the RMSE, the higher the accuracy of the forecast of the model. The MAPE is 9.9555% which indicate a high accurate forecast of the model.

MAPE	9.95550339
RMSE	1.67766432

Table 26: Results of the MAPE and RMSE for the Gaussian GLM applied on the loss amount of year 2013

c. Residuals Plots for GLM:

The Pearson residual in Figure 22 shows the residuals on the vertical axis and the fitted variables on the horizontal axis of the loss amount of year 2013. The more the variability of the residuals is constant around zero for all values of the covariates X_i , the better the model fits.

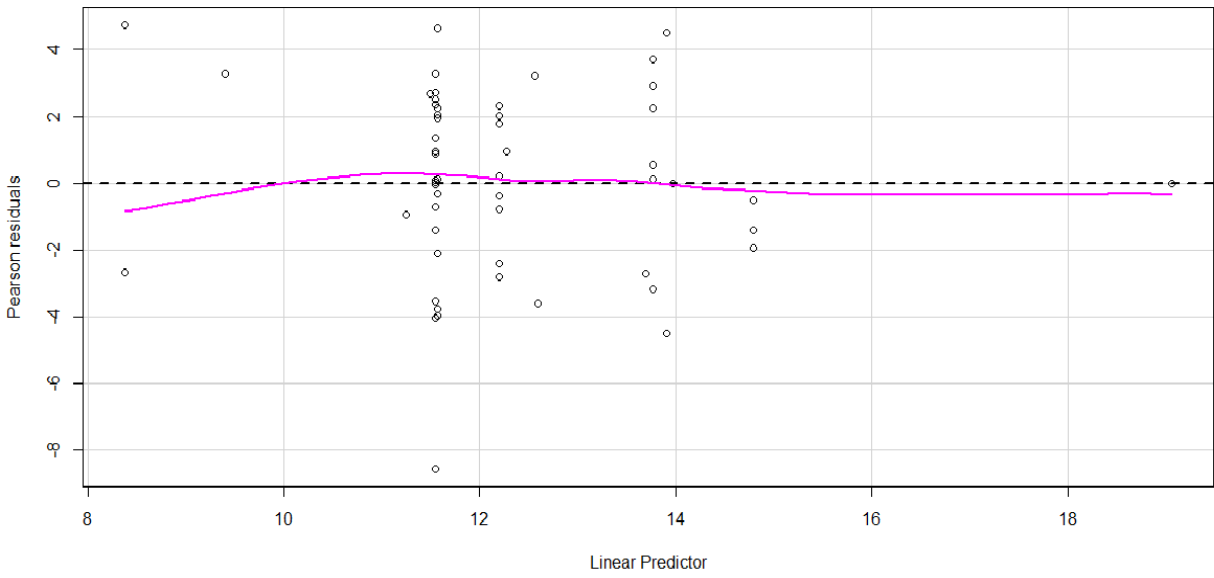


Figure 22: Pearson Residual plot of the GLM applied on the loss amount of year 2013.

4.5.2.2 Generalized Linear Model on each year 2014:

In section 4.1, the best fitted model of the loss amount for year 2014 is the Normal distribution. In section 4.2, the most important variables selected for year 2014 are: Action, Actor, Actor Country, Actor Motive, Actor Variety and Vector. Therefore, for year 2014, we applied the Generalized Linear Model to the normal distribution. We excluded the outliers and influential points from the data using the cook's distance to get a high accuracy of the predicted model.

a. Coefficients estimates and variable significance:

Table 27 provides the estimated values of the parameters in the fitted Gaussian Model applied on the loss amount of year 2014 and their significance level. The most significant variables are: the intercept, Action Hacking, Actor Motive Financial, Actor Motive Other, Action Social and Actor Motive Unknown.

Coefficient	Estimate	Std Error	t value	Pr(> t)	Significance
(Intercept)	16.959	2.779	6.102	0.000115	***
Action Hacking	4.868	1.437	3.386	0.006929	**
Action Misuse	2.141	1.429	1.498	0.164914	
Action Physical	2.073	1.599	1.296	0.224133	
Action Social	5.113	1.856	2.755	0.0203	*
Actor Internal	2.131	1.087	1.961	0.078342	.
Actor Partner	1.946	1.437	1.354	0.205622	
Actor Country CN	-2.655	2.305	-1.152	0.276214	
Actor Country RU	3.846	2.15	1.788	0.104005	
Actor Country Unknown	-1.826	1.802	-1.013	0.334894	
Actor Country US	-1.791	1.76	-1.017	0.333059	
Actor Motive Financial	-6.336	1.437	-4.408	0.001319	**
Actor Motive Other	-7.917	1.856	-4.267	0.001646	**
Actor Motive Unknown	-5.23	1.856	-2.818	0.018208	*

Table 27: Results of the coefficient for the GLM applied on the loss amount of year 2014

b. Deviance Residuals for GLM:

In Table 28, the Residual Deviance has been reduced by 112.142 with a loss of 13 degrees of freedom when we added all the parameters to the model including the intercept.

The Residual Deviance is 10.331 on 10 degrees of freedom, thus the dispersion parameter \emptyset is equal to 1.033103.

Null deviance	122.473 on 23 degrees of freedom
Residual deviance	10.331 on 10 degrees of freedom

Table 28: Results of the Deviance for the GLM applied on the loss amount of year 2014

Table 29 illustrates the values of the Minimum, Maximum, Median, First and Third Quantile for the Gaussian GLM applied to the loss amount for 2014.

Minimum	First Quantile	Median	Third Quantile	Maximum
-1.43737	-0.06939	0	0.25322	1.15129

Table 29: Results of the Min, Max, Median, First and Thirsd Quantile for the GLM applied on the loss amount of year 2014.

Table 30 presents the value of the AIC and the Fisher Scoring for the Gaussian GLM applied on the loss amount of year 2014. The lower the AIC and the Fisher Scoring, the better the model fits.

AIC	77.879
Number of Fisher Scoring iterations	2

Table 30: Results of the AIC and Fisher Scoring for the Gaussian GLM applied on the loss amount of year 2014

Table 31 presents the value of the MAPE and RMSE for the Gaussian GLM applied on the overall data. The lower the MAPE and the RMSE, the higher the accuracy of the forecast of the model. The MAPE is 3.6105% which indicate a high accurate forecast of the model.

MAPE	3.610515222
-------------	-------------

Table 31: Results of the MAPE and RMSE for the Gaussian GLM applied on the loss amount of year 2014

c. Residuals Plots for GLM:

The Pearson residual plot in Figure 23 shows the residuals on the vertical axis and the fitted variables on the horizontal axis of the loss amount of year 2014. The more the variability of the residuals is constant around zero for all values of the covariates X_i , the better the model fits.

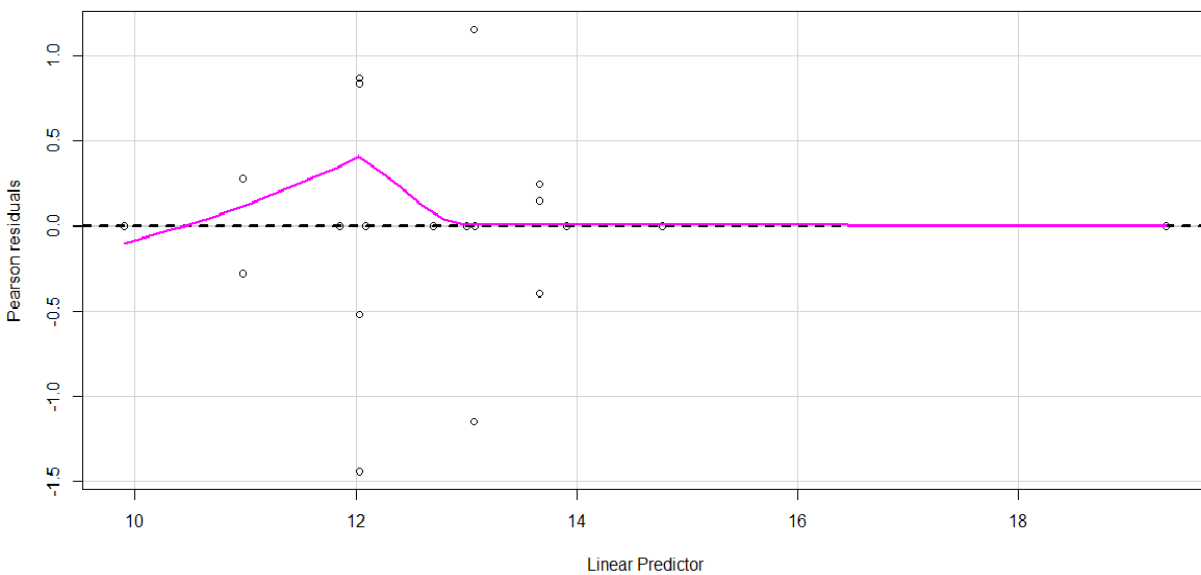


Figure 23: Pearson Residual plot of the GLM applied on the loss amount of year 2014.

4.5.2.3 Generalized Linear Model on each year 2015:

In section 4.1, the best fitted model of the loss amount for year 2015 is the Weibull distribution. However, the Weibull distribution does not belong to the exponential dispersion family (EDM), thus the GLM cannot be applied to the weibull distribution. Therefore, in this case the GLM is applied to the Normal distribution since it has the lower AIC among the fitted distribution belonging to the exponential dispersion family (EDM) and the closer AIC to the AIC of the Weibull distribution.

In section 4.2, the most important variables selected for year 2015 are: Actor Motive. Therefore, for year 2015, we applied the Generalized Linear Model to the normal distribution. We excluded the outliers and influential points from the data using the cook's distance in order to get a high accuracy of the predicted model.

a. Coefficients estimates and variable significance:

Table 32 provides the estimated values of the parameters in the fitted Gaussian Model applied on the loss amount of year 2015 and their significance level. The most significant variables are: the intercept and Actor Motive Unknown.

Coefficient	Estimate	Std Error	t value	Pr(> t)	Significance
(Intercept)	9.616	1.228	7.832	3.31E-07	***
Actor Motive Financial	2.544	1.271	2.001	0.0607	.
Actor Motive Grudge	2.12	1.736	1.221	0.2378	
Actor Motive Unknown	3.552	1.326	2.678	0.0153	*

Table 32: Results of the coefficient for the GLM applied on the loss amount of year 2015

b. Deviance Residuals for GLM applied on the loss amount of year 2015:

In Table 33, the Residual Deviance has been reduced by 12.314 with a loss of 3 degrees of freedom when we added all the parameters to the model including the intercept.

The Residual Deviance is 27.134 on 18 degrees of freedom, thus the dispersion parameter ϕ is equal to 1.50744.

Null deviance	39.448 on 21 degrees of freedom
Residual deviance	27.134 on 18 degrees of freedom

Table 33: Results of the Deviance for the GLM applied on the loss amount of year 2015

Table 34 illustrates the values of the Minimum, Maximum, Median, First and Third Quantile for the Gaussian GLM applied to the loss amount for 2015.

Minimum	First Quantile	Median	Third Quantile	Maximum
-2.5435	-0.7772	0	0.7608	2.6117

Table 34: Results of the Min, Max, Median, First and Third Quantile for the GLM applied on the loss amount of year 2015.

Table 35 presents the value of the AIC and the Fisher Scoring for the Gaussian GLM applied on the loss amount of year 2014. The lower the AIC and the Fisher Scoring, the better the model fits.

AIC	77.048
Number of Fisher Scoring iterations	2

Table 35: Results of the AIC and Fisher Scoring for the Gaussian GLM applied on the loss amount of year 2015

Table 36 presents the value of the MAPE and RMSE for the Gaussian GLM applied on the overall data. The lower the MAPE and the RMSE, the higher the accuracy of the forecast of the model. The MAPE is 6.90056% which indicate a high accurate forecast of the model.

MAPE	6.900596
RMSE	1.33E-15

Table 36: Results of the MAPE and RMSE for the Gaussian GLM applied on the loss amount of year 2015

c. Residuals Plots for GLM:

The Pearson residual plot in Figure 24 that shows the residuals on the vertical axis and the fitted variables on the horizontal axis of the loss amount of year 2015. The more the variability of the residuals is constant around zero for all values of the covariates X_i , the better the model fits.

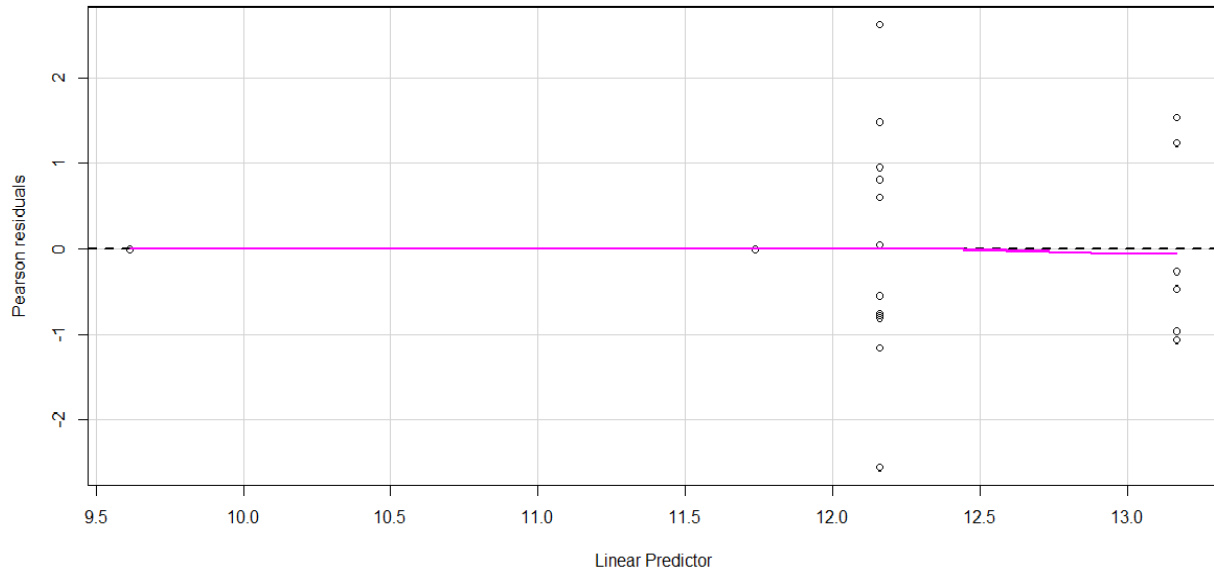


Figure 24: Pearson Residual plot of the GLM applied on the loss amount of year 2015.

4.5.2.4 Generalized Linear Model on each year 2016:

In section 4.1, the best fitted model of the loss amount for year 2016 is the lognormal distribution. In section 4.2, the most important variables selected for year 2016 are: Action, Actor, Actor Variety, Asset Variety and Variety. Therefore, for year 2016, we applied the Generalized Linear Model to the lognormal distribution. We excluded the outliers and influential points from the data using the cook's distance to get a high accuracy of the predicted model.

a. Coefficients estimates and variable significance:

Table 37 provides the estimated values of the parameters in the fitted Gaussian Model applied on the loss amount of year 2016 and their significance level. The most significant variable is: the intercept.

Coefficient	Estimate	Std Error	t value	Pr(> t)	Significance
(Intercept)	1.93551	0.69286	2.794	0.0162	*
Action Hacking	0.3491	0.62073	0.562	0.5842	
Action Malware	0.66086	0.6164	1.072	0.3047	
Action Misuse	0.6357	0.64727	0.982	0.3454	
Action Physical	0.33076	0.57637	0.574	0.5766	
Action Social	0.63121	0.61385	1.028	0.3241	
Actor Internal	-0.06591	0.35693	-0.185	0.8566	
Actor Partner	0.37975	0.64513	0.589	0.567	
Actor Variety End-user	0.06305	0.73464	0.086	0.933	
Actor Variety Espionage	-0.71247	0.6916	-1.03	0.3232	
Actor Variety Nation-state	0.7823	0.40839	1.916	0.0795	.
Actor Variety Organized crime	-0.31992	0.42228	-0.758	0.4633	
Actor Variety Other	-0.34202	0.57153	-0.598	0.5607	
Actor Variety Unaffiliated	-0.06833	0.31004	-0.22	0.8293	
Actor Variety Unknown	0.20917	0.37765	0.554	0.5898	

Table 37: Results of the coefficient for the GLM applied on the loss amount of year 2016

b. Deviance Residuals for GLM:

In Table 38, the Residual Deviance has been reduced by 389.94 with a loss of 14 degrees of freedom when we added all the parameters to the model including the intercept.

The Residual Deviance is 125.47 on 12 degrees of freedom, thus the dispersion parameter ϕ is equal to 10.4557.

Null deviance	515.41 on 26 degrees of freedom
Residual deviance	125.47 on 12 degrees of freedom

Table 38: Results of the Deviance for the GLM applied on the loss amount of year 2016

Table 39 illustrates the values of the Minimum, Maximum, Median, First and Third Quantile for the Gaussian GLM applied to the loss amount for 2016.

Minimum	First Quantile	Median	Third Quantile	Maximum
-6.1557	-0.8045	0	0.4428	6.1557

Table 39: Results of the Min, Max, Median, First and Third Quantile for the GLM applied on the loss amount of year 2016.

Table 40 presents the value of the AIC and the Fisher Scoring for the Gaussian GLM applied on the loss amount of year 2016. The lower the AIC and the Fisher Scoring, the better the model fits.

AIC	150.1
Number of Fisher Scoring iterations	5

Table 40: Results of the AIC and Fisher Scoring for the Gaussian GLM applied on the loss amount of year 2016

Table 41 presents the value of the MAPE and RMSE for the Gaussian GLM applied on the loss amount of year 2016. The lower the MAPE and the RMSE, the higher the accuracy of the forecast of the model. The MAPE is 3.6105% which indicate a high accurate forecast of the model.

MAPE	4.59791759
RMSE	9.25869186

Table 41: Results of the MAPE and RMSE for the Gaussian GLM applied on the loss amount of year 2016

c. Residuals Plots for GLM:

The Pearson residual plot is graph that shows the residuals on the vertical axis and the fitted variables on the horizontal axis of the loss amount of year 2016. The more the variability of the residuals is constant around zero for all values of the covariates X_i , the better the model fits.

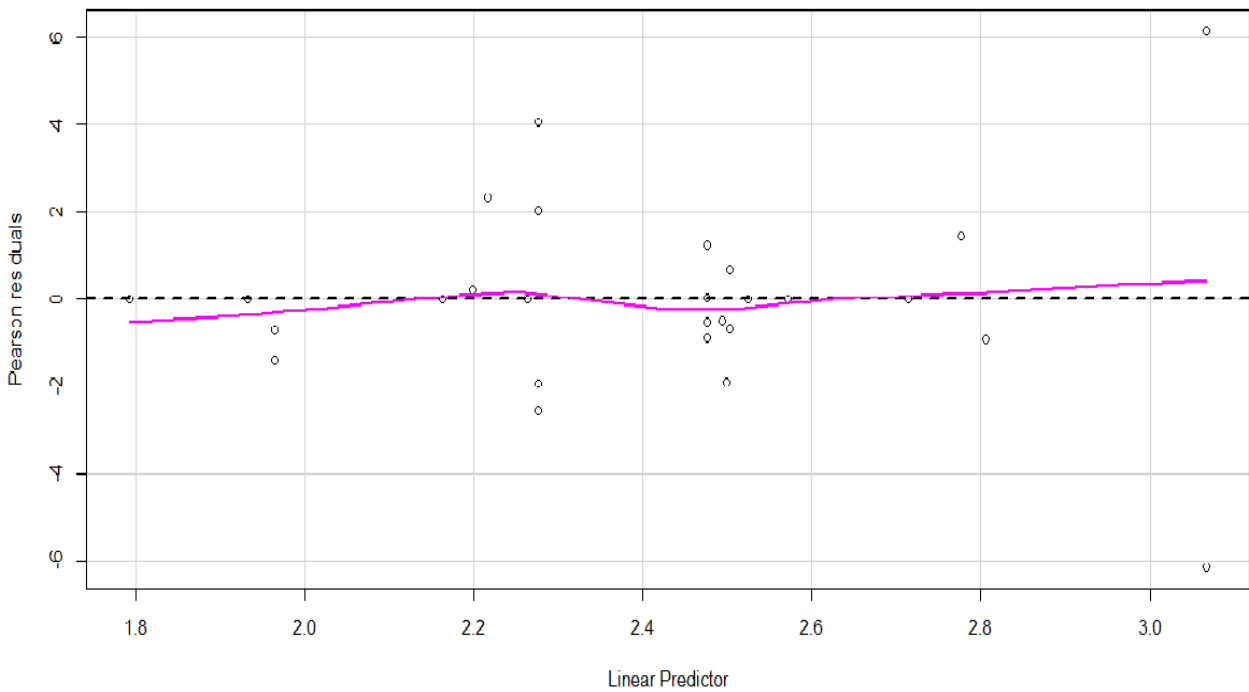


Figure 25: Pearson Residual plot of the GLM applied on the loss amount of year 2016.

4.6 Credibility Theory

In this section we applied the Classical credibility theory in order to estimate minimum number of losses required to have a certain level of accuracy.

4.6.1 Credibility theory on the overall data

Table 42 shows that the estimated minimum number of total losses required to have a 95% level of accuracy is approximately 94. Since the number of losses in the sample data of the VERIS data is 276, we can conclude that the fitted model on the overall data is accurate.

Probability level	90%	95%	98%
<i>Y</i> value	1.645	1.96	2.326
<i>k</i> value	5%	5%	5%
Number of observation needed	65.57	93.08	131.09

Table 42: Results of the Classical credibility Theory applied on the overall data.

4.6.2 Credibility theory on each year

4.6.2.1 Credibility theory on year 2013

Table 43 shows that the estimated minimum number of losses for year required to have a 95% level of accuracy is approximately 106. Since the number of losses in the sample data of the VERIS data for year 2013 is 59, we can conclude that we need a larger sample for 2013 in order to get a high level of accuracy.

Probability level	90%	95%	98%
<i>Y</i> value	1.645	1.96	2.326
<i>k</i> value	5%	5%	5%
Number of observation needed	74.53	105.81	149.01

Table 43: Results of the Classical credibility Theory applied for year 2013.

4.6.2.2 Credibility theory on year 2014

Table 44 shows that the estimated minimum number of losses for year required to have a 95% level of accuracy is approximately 103. Since the number of losses in the sample data of the VERIS data for year 2014 is 40, we can conclude that we need a larger sample for 2014 in order to get a high level of accuracy.

Probability level	90%	95%	98%
<i>Y</i> value	1.645	1.96	2.326
<i>k</i> value	5%	5%	5%
Number of observation needed	72.08	102.32	144.11

Table 44: Results of the Classical credibility Theory applied for year 2014.

4.6.2.3 Credibility theory on year 2015

Table 45 shows that the estimated minimum number of losses for year required to have a 95% level of accuracy is approximately 90. Since the number of losses in the sample data of the VERIS data for year 2015 is 35, we can conclude that we need a larger sample of data for year 2015 in order to get a high level of accuracy.

Probability level	90%	95%	98%
<i>Y</i> value	1.645	1.96	2.326
<i>k</i> value	5%	5%	5%
Number of observation needed	62.88	89.27	125.72

Table 45: Results of the Classical credibility Theory applied for year 2015.

4.6.2.4 Credibility theory on year 2016

Table 46 shows that the estimated minimum number of losses for year required to have a 95% level of accuracy is approximately 224. Since the number of losses in the sample data of the VERIS data for year 2016 is 27, we can conclude that we need a larger sample of data for year 2016 in order to get a high level of accuracy.

Probability level	90%	95%	98%
<i>Y</i> value	1.645	1.96	2.326
<i>k</i> value	5%	5%	5%
Number of observation needed	157.63	223.78	315.16

Table 46: Results of the Classical credibility Theory applied for year 2016.

Conclusion

Cyber-attacks can lead to different types of losses, such as loss of information or loss of revenue. Therefore, cyber insurance policies have become a necessity in order to protect policy holders from catastrophic losses and liabilities due to cyber breaches. This thesis makes a significant contribution to modeling cyber security insurance. The overall data of the loss amount of breaches in the VERIS dataset fit a normal distribution since its P-value of the Kolmogorov-Smirnov test is greater than 0.05 and it has the lowest AIC. Moreover, the Random Forest applied to the loss amount of breaches in the overall VERIS data indicated that the most important variables associated are: Action, Actor Motive, Asset Variety and Impact overall rating of breach. Then, the Generalized Linear Model applied to the most important variable provided by the Random Forest, lead to the conclusion that the normal distribution is a good candidate for fitting the severity distribution of the VERIS dataset with a high level of accuracy indicated by the mean absolute percentage error (MAPE) of 12.70%. Also, the classical credibility theory indicated that the fitted model on the overall data is 95% accurate. However, the classical credibility theory applied to the loss amount of each year indicated that a larger sample of data needs to be fitted in order to get a high level of accuracy. Thus, this study discovered some weaknesses and gaps in the VERIS reporting schema which may be obscuring accurate reporting of all related, detectable adversary activity linked to these attacks. Therefore, a worldwide collaboration must take place in order to contribute to a larger database that allows actuaries and statisticians to model the severity and frequency of the loss amount for future breaches with a high level of accuracy. As per the future work, it will be essential to fit the frequency based on the timeline event of a cyber risk loss using a larger and worldwide database in order to predict the aggregate loss distribution and estimate the price of a cyber-insurance policy coverage.

References

- Bejamin, E., Steven, H., & Forrest, S. (2015). Hype and Heavy tails: a closer look at data breaches. *Journal of Cybersecurity*, 2(1).
- Böhme, R., & Kataria, G. (2006). *Models and Measures for Correlation in Cyber-Insurance*.
- Breiman, L., & Schapire, R. (2001). *Random Forests*. 45, 5–32.
- Broverman, S. A. (2014). *ACTEX C Study Manual* (Fall 2014, Vol. 1). ACTEX Publications.
- Carfora, M., Mercaldo, F., Orlando, A., & Martinelli, F. (2019). Cyber risk management: an actuarial point of view Analysis of nonlinear ODE models: theory simulation and parameter estimation View project WCCI 2018 -FUZZ IEEE -Special Session on Business Process and Fuzzy Logic (BPFL) View project Cyber Risk Management: an actuarial point of view. *Article in Journal of Operational Risk*.
<https://doi.org/10.21314/JOP.2019.231>
- CISA. (2009, May 6). What is Cybersecurity? | CISA. Retrieved from us-cert.cisa.gov website:
<https://us-cert.cisa.gov/ncas/tips/ST04-001>
- Cyber insurance | Business Insurance | ABI. (2014). Retrieved from Abi.org.uk website:
<https://www.abi.org.uk/products-and-issues/choosing-the-right-insurance/business-insurance/cyber-risk-insurance/>
- Cyber insurance market challenges. (2017). *Enhancing the Role of Insurance in Cyber Risk Management*, 93–109. <https://doi.org/10.1787/9789264282148-6-en>
- Dunn, P. K., & Smyth, G. K. (2017). *Generalized Linear Models With Examples in R*. 233 Spring Street, New York, NY 10013, U.S.A.: Springer Science+Business Media, LLC.
- Eling, M., & Loperfido, N. (2017). Data breaches: Goodness of fit, pricing, and risk measurement. *Insurance: Mathematics and Economics*, 75.
- Farkas, S., Lopez, O., & Thomas, M. (2020). Cyber claim analysis through Generalized Pareto Regression Trees with applications to insurance. *HAL Archives Ouvertes*.
- Fetterman, R. (2019). *Regression-Based Attack Chain Analysis and Staffing Optimization for Cyber Threat Detection*.
- Geneva Association. (2016). *Ten Key Questions on Cyber Risk and Cyber Risk Insurance THE GENEVA ASSOCIATION*. Retrieved from website:

- https://www.genevaassociation.org/sites/default/files/research-topics-document-type/pdf_public/cyber-risk-10_key_questions.pdf
- Genuer, R., & Poggi, J.-M. (2020). *Random Forests with R*. Gewerbestrasse 11, 6330 Cham, Switzerland: Springer Nature Switzerland AG.
- Herath, H. S. B., & Herath, T. C. (2011). Copula-based actuarial model for pricing cyber-insurance policies. *Insurance Markets and Companies: Analyses and Actuarial Computations*, 2(1).
- Maillard, T., & Sornette, D. (2010). Heavy-tailed distribution of cyber risk. *The European Physical Journal B*, 75.
- Mukhopadhyay, A., Chatterjee, S., Saha, D., Mahanti, A., & Sadhukhan, S. K. (2013). Cyber-risk decision models: To insure IT or not? *Decision Support Systems*, 56, 11–26.
<https://doi.org/10.1016/j.dss.2013.04.004>
- OECD. (2017). *Enhancing the Role of Insurance in Cyber Risk Management* | READ online. Retrieved from oecd-ilibrary.org website: https://read.oecd-ilibrary.org/finance-and-investment/enhancing-the-role-of-insurance-in-cyber-risk-management_9789264282148-en#page1
- VERIS. (n.d.). VERIS The vocabulary for event recording and incident sharing. Retrieved from veriscommunity.net website: <http://veriscommunity.net/>
- Wheatly, S., Maillard, T., & Sorette, D. (2015). The extreme risk of personal data breaches and the erosion of privacy. *The European Physical Journal B*, 89(7).