ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Cyber Risks and Data Breaches

Master Thesis

Aline Schillig

22 October 2018

Advisors: Dr. S. Wheatley[†], Prof. M. Maathuis[‡], Dr. S. Frei, Prof. D. Sornette[†].

Chair of Entrepreneurial Risks[†] & Seminar for Statistics[‡], ETH Zürich

Abstract

In this master thesis we analyze data breaches which constitute one of the key cyber risks of today's cyber world. Motivated by previous work, in particular the work of Eling and Loperfido [14, 2017], Wheatley, Maillart and Sornette [37, 2016], Hofmann, Wheatley and Sornette [18, 2018], we analyze data breaches with at least $70k$ records lost from an insurance point of view with a new extended dataset. We use multidimensional scaling to identify severity risk classes based on the economic sector. To model the frequency we employ generalized linear models (GLMs), whereby we detect notable different scenario outcomes for the future development of the frequency of data breaches with at least $70k$ records lost. The data breach severity is analyzed with respect to various characteristics of the event, such as the size and economic sector of the affected entity as well as the type of breach medium, the mode of failure that led to the breach and whether a third party was involved in the data breach event. We estimate the severity distribution, which is best approximated by a truncated log-normal or upper-truncated Pareto distribution for various thresholds for the complete dataset. In a further step we study the reporting delay. Herefore both parametric and non-parametric methods are used to assess the development of the reporting delay over time and its relation to other variables. Furthermore, we analyze whether there have been any changes in the reporting of data breach events due to the introduction of data breach notification laws in the United States.

Acknowledgements

I would like to thank Prof. Didier Sornette, Dr. Spencer Wheatley and Dr. Stefan Frei for introducing me to this interesting topic. Furthermore, I thank Prof. Marloes Maathuis for co-supervising this master thesis. In particular I thank Dr. Spencer Wheatley, Prof. Marloes Maathuis and Dr. Stefan Frei for their help during this project. Last but not least I am very grateful for the constant support from my family and friends, whereby I especially thank Silja Haffter for proofreading my work.

Contents

1 Introduction

1.1 Cyber risks and data breaches: a definition

Within this master thesis we study data breach events, which correspond to one type of risks of the ever-changing cyber world [15]. The word ”cyber” takes its meaning from the nouns computer network and virtual reality and is used in a broad context. Generally speaking, it relates to anything that has to do with information technology systems. It includes the internet but also our dependence on electronic networks upon which our infrastructure runs. While there are many benefits coming along with the ongoing technological progress, there are also risks that need to be considered and assessed. As cyber appears in a wide range of contexts it naturally involves a vast spectrum of risks. One definition of the latter is given by the Geneva Association [15, p. 12] and defines cyber risk to be:

> ”Any risk emerging from the use of information and communication technology (ICT) that compromises the confidentiality, availability, or integrity of data or services. The impairment of operational technology (OT) eventually leads to business disruption, (critical) infrastructure break down, and physical damage to humans and properties.”

Discussing all of the risks associated with cyber exceeds the scope of this master thesis and we have therefore decided on focusing our attention on a single type of risk, namely the risk associated with data breaches. A data breach refers to the event of a loss of massive amounts of data. Hereby we can distinguish between different events by the kind and size of data that was lost, the owner, how the loss happened and if there were any negative consequences for the involved parties. The type of information included in such an event can be anything from credit card information, social security numbers, bank account information, login credentials, physical addresses up to personal health information. The size of an event is typically defined by the number of records lost, whereby a record contains any of the aforementioned types of information. In many cases the lost data corresponds to client information collected by an entity or organization, whereby this also includes online forums and governments. Knowing who is the owner of the lost data already provides a lot of information about what kind of data might be at stake. If client data was stolen from a bank, the bank’s client would generally be more worried about the privacy of their financial information than of their personal health information. Similarly, voting data of citizens are more likely to be stored by a government or political organization rather than by a financial institution. While hacking is one of the most common reasons for a loss of data [18], there are also others. In particular do not all events happen on a malicious basis and can result due to a human error or a failure of the used hard- or software. Data breach events are of importance as the stolen data is often traded on the dark web and can lead to identity and credit card fraud [37].

1.2 Aim of the thesis

The aim of this master thesis is to do a statistical analysis on data breaches. The main interest lies therefore in looking at the phenomenon from an insurance point of view and herefore to define individual risk classes, characterize both the frequency and severity distributions, evaluate the reporting delay and assess whether changes in the regulatory framework have influenced the reporting of data breaches. Of particular interest is hereby to assess the changes over time and quantify them if possible.

1.3 Previous results

There are numerous perspectives one can consider when looking at cyber risks and data breaches. In the following we limit the research review to the three topics categorization of cyber risks, related costs and previous analyses of data breach events, as they are most helpful in understanding the issue at hand.

Categorization of cyber risks
A fundamental question which has to be answered is how cyber events should be classified [15, 4]. Several methods exist, but a universal standard has yet to be established. Admittedly, this might be too much to ask for, as it greatly depends on what question one wants to answer. However, this is of immense importance as it lays the stepping stone for cyber insurance. In order to be able to ensure such risks it must be clear what they entail and what is covered. A categorization from an insurance point of view is given by the Geneva Association [15], where cyber risks are characterized by multiple dimensions. The first one differentiates between events that are caused by natural disasters or events that can be considered man-made catastrophes. The latter allows a further differentiation by considering the kind of activity (criminal, non-criminal, intentional, accidental), the type of attack (malware, insider attack, spam, denial of service, etc.) and the type of attacker (terrorist, governmental, criminal). The second dimension measures the vulnerability, whereby both the organization specific security level as well as the one from the industry or supply chain partners are considered. The third dimension distinguishes among the consequences of a cyber event, as they can for example lead to a loss of data or business interruption, which may lead to a monetary loss later on.

Costs related to data breaches
This highlights another research question which has been of high interest with regards to the pricing of cyber insurance contracts, namely what are the costs associated with cyber risks? An extensive study on the costs of data breach events has been conducted by the Ponemon institute in 2017 [22]. In this study an average amount of costs per lost record is calculated based on a global sample from 11 countries and two regional samples[1]. To get an estimate for the costs associated with a data breach event, numerous aspects were considered. Firstly, there might be a loss of costumers, which was unforeseen and not planned for, due to the data breach. Further costs stem from the detection, internal reporting and containment of the breach event. This includes for example additional assessments and audits as well as investigative tasks. Then there are costs after the data breach which stem from notifying the victims, legal expenditures and identity protection services for victims - only to name a few. While the average costs of data breaches are decreasing on a global level (158 US$ in the study from 2016 vs. 141 US$ in the study from 2017), the average costs have risen for some countries. The US is for example not only the

[1] The study considered 419 organizations from the US, the UK, Germany, Australia, France, Brazil, Japan, Italy, India, Canada, South Africa, the Middle East (including the United Arab Emirates and Saudi Arabia) and the ASEAN region (including Singapore, Indonesia, the Philippines and Malaysia). The analysis was limited to events with $1k$ up to $100k$ records lost.

country with the highest average costs per lost record (255 US$ in the 2017 study) but has also shown a significant increase from 2016 onwards. For the 419 organizations included in the 2017 study the average total cost was 3.62 million US$, while for the previous year the average total was 4 million US$. Even though the average total costs per data breach have declined, it is still a glaring additional amount of expenses to digest and further accentuates the demand for insurance products.

Previous analyses of data breaches

Previous research has already been conducted in this field and the following analysis builds upon previous work done by Eling and Loperfido [14, 2017], Wheatley, Maillart and Sornette [37, 2016], Hofmann, Wheatley and Sornette [18, 2018] and we therefore give a brief summary of their respective results.

In [14, 2017] multidimensional scaling (MDS) is used to define different risk classes with regards to frequency and severity. Hereby both the kind of attack (hack, unintended disclosure, etc.) as well as the type of organization (business, governmental, etc.) were considered and Eling and Loperfido showed that different types of data breaches should be modelled as individual risk classes. For both frequency and severity several distributions were parametrized per risk class and compared to the original dataset. While for severity the skewed log-normal distribution shows the best fit, the frequency is best modelled by a negative binomial distribution.

In [37, 2016] both frequency and severity distributions have been estimated, whereby a current maximum breach size was detected. The monthly frequency was modelled via a Poisson generalized linear model (GLM) and while the rate for events within the US remained stable, a significant increase was detected for events outside of the US. For the severity a current maximum was detected which grows sublinearly in time and is characterized by a doubly truncated Pareto distribution. Furthermore, it was found that both the frequency and severity of data breach events scale with the organization size s (here given by the market capitalization) according to $s^{0.6}$. The cumulative process is studied and the issue of the erosion of privacy is highlighted as personal information accumulates in underground markets.

In [18, 2018] the authors give a characterization of data breach events as a man-made catastrophe and show that data breaches are in particular dominated by hacking events. For the latter it was shown that both the frequency and severity have been increasing over time, whereby the half-yearly frequency counts were modelled by a log-linear negative binomial GLM and for the severity both a truncated log-normal as well as an upper-truncated pareto distribution give a suitable fit. Furthermore, challenges with regards to cyber insurance have been thoroughly discussed. The currently on the market available insurance policies have many limitations and exclusions and thus make it difficult for costumers to select a product that suits their needs. However, at the same time it is hard for insurance companies to price policies adequately as the risk is largely driven by human behavior. Further challenges with regards to insuring cyber risks were discussed from different viewpoints (insurance company, regulatory & societal, individual & firm), whereby for insurance companies the major limitations in insuring cyber risks lie in the ongoing change of the technology, the heavy-tailed nature and the violation of independence among different events.

When comparing the results from different analyses it is important to keep in mind that firstly the datasources or a combination thereof are most often not the same or cover different timespans, as the analyses have been conducted at different points of time. Secondly, most analyses only consider the upper tail of the severity distribution, whereby different truncation points have been used. Even though severe breaches account for almost all of the lost records [37], there are still numerous small events happening.

Other analyses on the topic have been conducted in [12] and [29].

1.4 Outline of the thesis

Based on the work that has been done so far, we would like to revisit some of the previously posed questions on a new dataset. Moreover, we would like to additionally look at more ways to characterize such data breach events. To this end we have built our own dataset from three publicly available sources (see section 1.5.1) and added market data as well as additional factor variables, which further classify the data breach events. As Eling and Loperfido in [14] we use multidimensional scaling (MDS) to differentiate between different risk classes in chapter two. Then we characterize both the frequency and severity distribution functions in chapters three and four. In chapter five we use an additional dataset to assess the development of the reporting delay of data breach events over the last couple of years. In chapter six we investigate whether the frequency of data breaches changed with the introduction of data breach notification laws per state within the United States (US). The results are summarized at the end of each of the corresponding subsections. A short summary thereof and comparison with previous findings of the literature is given in chapter seven, where also future questions of interest are presented.

1.4.1 General remarks

Throughout this master thesis standard mathematical notation is used, where we denote by Y, N the response variables and by X the corresponding predictor variables in a matrix. Observations are denoted by y_i and their corresponding predictor vector by x_i. Any subscripts such as t, m or q refer to the time or time interval such as months or quarters. β specifies the parameters in the statistical models and the intercept is always included if not otherwise specified. Furthermore, hypothesis tests are conducted at a 95% confidence level if not mentioned otherwise. As a general rule of thumb our estimates are shown with no more than three significant digits (most often only two) as they are only estimates of the true values and contain some amount of uncertainty. The reporting of more significant digits might cause the reader to believe that the estimates are of higher precision than they are likely to be and we thus refrain from doing so. The names of variables are printed in *italics* to make the distinction clear and for factor variables the corresponding levels are also printed in *italics*.

Throughout the master thesis several acronyms are used and in table 1.1 a list of them is given as a point of reference.

1.5 Description of the main dataset

1.5.1 Sources

For our analysis we will work with a dataset combined from the following three publicly available datasources.

1. Privacy Rights Clearinghouse (*PRC*) [8], as of 2.6.2018. Most of the events were obtained from this source. *PRC* only records events which are reported within the US[2]. The database contains events from February 2005 until May 2018.

2. Breach Level Index (*bli*) [20], as of 23.4.2018. This is the second largest source and has recorded data breach events from all over the world since 2013 until the end of 2017.

[2] This does not imply that the organization is headquartered in the US.

Table 1.1: Acronyms used throughout the master thesis and their corresponding terms.

Acronym	Term	Acronym	Term
AD	Anderson-Darling	OLS	Ordinary least squares
AIC	Akaike information criterion	other	Miscellanous sector which is a merger of the economic sectors energy, basic materials, utilities, politics and military
BIC	Bayesian information criterion	pol	Politics
bli	Breach Level Index (see section 1.5.1 for further information)	PRC	Privacy Rights Clearinghouse (see section 1.5.1 for further information)
CA	Correspondence analysis	SW	Software
edu	Education/educational	unkn	Unknwon
GLM	Generalized linear model	US	United States
gov	Government/governmental	50	Energy (economic sector)
HIBP	“Have I Been Pwnd” (see section 5.1 for further information)	51	Basic materials (economic sector)
HW	Hardware	52	Industrials (economic sector)
IiB	Information is beautiful (see section 1.5.1 for further information)	53	Consumer cyclicals (economic sector)
KS	Kolmogorov-Smirnov	54	Consumer non-cyclicals (economic sector)
MCAP	Market capitalization / market capitalized	55	Financials (economic sector)
MDS	Multidimensional scaling	56	Healthcare (economic sector)
MFACT	Multiple factor analysis	57	Technology (economic sector)
mil	Military	58	Telecommunication services (economic sector)
NPO	Not-for-profit organization	59	Utilities (economic sector)

3. Information is Beautiful (*IiB*) [21], as of 6.6.2018. This database contributes the least to our dataset, as most of the events are already contained in the other two. It also records global events and contains data breach events from 2004 onwards.

1.5.2 Dataset

Only events with at least $70k$ records lost were considered and yielded a dataset with 993 observations. The number of records lost is referred to by severity or *total records* (name of variable). For each observation a proxy *date* is available and the *location* of the headquarters of the affected entity was specified (if possible). Additional market information was included[3] in order to better understand what kind of companies were affected. The additional market data consists of the *economic sector* of the affected entity, its *market capitalization* (if available) and the *number of employees* at the proxy *date* (if available). An additional variable also distinguishes among the *type of organization* (e.g. *private, public, government*) and six factor variables add further information about the breach event. The latter provide a specification of the *medium* with which the data was lost, give information about whether *multiple firms* were involved in the same data breach, whether the breach was committed or facilitated by an *inside* or *outside* party, whether the breach happened *intentionally*, whether a *third party* was involved and specify the *mode of failure* of the data breach. The complete list and a detailed description of the variables is provided in the appendix A.1.

[3] Mostly from Thomson Reuters [13] and other publicly available resources, such as annual reports.

2 Exploratory Analysis

We start by doing an exploratory analysis of the main dataset and as in [14], we use multidimensional scaling (MDS) [23] to this end. In order to identify any subgroups within our dataset we require a dissimilarity measure that can deal both with continuous and factor variables. One possible dissimilarity measure that can handle both types of variables is Gower's similarity coefficient [17]. By using this measure no valid subgroups could be identified on the complete dataset when taking all variables into account. We have thus decided on using a simplified approach and show in this section how the frequency of data breach events of different severities is related to the economic sector. Following [14], we use multiple factor analysis (MFACT) [26] in a second step to analyze its development over time.

2.1 Multidimensional scaling

As in [14], we set up a contingency table which counts how often a sector has been victimized and thereby differentiates between the severity of the attack. Since *total records* is a continuous variable, we group it into quartiles. As some sectors (energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*)) show very few observations we merge them into a miscellaneous sector called *other*, otherwise they appear as outliers and dominate the representation.

Table 2.1: Number of events reported in the complete dataset per economic sector (industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and *other*; the economic sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*) have been merged into the miscellaneous sector *other*) differentiated by severity quartiles. The row percentages are shown in brackets and the row total in the right outmost column. Differences between the economic sectors are both visible in the row totals and the relative frequency of events per severity quartile.

	1. Quartile	2. Quartile	3. Quartile	4. Quartile	Total
52	39 (23.8%)	40 (24.4%)	46 (28.0%)	39 (23.8%)	164
53	28 (20.3%)	26 (18.8%)	46 (33.3%)	38 (27.5%)	138
54	8 (24.2%)	11 (33.3%)	7 (21.2%)	7 (21.2%)	33
55	34 (27.2%)	38 (30.4%)	31 (24.8%)	22 (17.6%)	125
56	43 (29.7%)	50 (34.5%)	35 (24.1%)	17 (11.7%)	145
57	30 (14.8%)	27 (13.3%)	54 (26.6%)	92 (45.3%)	203
58	6 (15.8%)	10 (26.3%)	7 (18.4%)	15 (39.5%)	38
edu	45 (51.7%)	26 (29.9%)	10 (11.5%)	6 (6.9%)	87
other	15 (25.0%)	20 (33.3%)	12 (20.0%)	13 (21.7%)	60

Usually, correspondence analysis is used for contingency tables. In correspondence analysis the primary goal is to reveal the dependence relationship among the row and column variables [23]. However, we are interested in identifying different risk classes and will therefore apply MDS in order to represent the dissimilarities between the economic sectors in a two or three dimensional space. For contingency tables the chi-square distance is usually used to assess the dissimilarity between rows (or columns). This was also done in [14] and we use the same distance measure here.

This allows us to express the difference in the frequency of data breach events with respect to severity quartiles for two different economic sectors. We use the classical MDS algorithm [23, p. 481] to get a two or three dimensional representation of our contingency table.

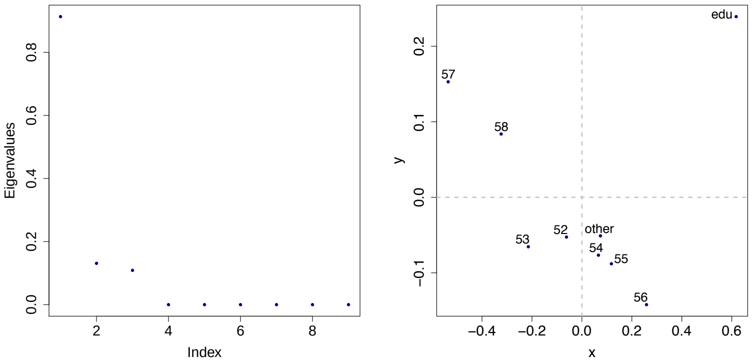


Figure 2.1: Left plot: Decay of the eigenvalues of the multidimensional scaling solution of the contingency table showing the number of events reported in the complete dataset per economic sector (industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and *other*; the economic sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*) have been merged into the miscellaneous sector *other*) differentiated by severity quartiles. Starting from the fourth largest eigenvalue all of them are equal to zero, which means we can represent the observed dissimilarities between the economic sectors exactly in a three dimensional space. Right plot: Two dimensional representation of the economic sectors given by the multidimensional scaling solution of the aforementioned contingency table. We observe in particular a separation of the education (*edu*), the technology (*57*) and the telecommunication (*58*) sectors from the other economic sectors.

In the left panel of figure 2.1 we can observe a fast decay from the first to the second eigenvalue and a plateau at the second and third largest eigenvalue. The subsequent eigenvalues are equal to zero. This indicates that we can represent the observed dissimilarities quite accurately in a two or three dimensional space. The representation in a two dimensional space is shown on the right of figure 2.1.

2.1.1 Two dimensional multidimensional scaling representation

The goodness of fit measure[1] equals 0.91 and indicates that the two dimensional MDS solution has a good fit. An interpretation of the axes is given by considering the correlations of the fitted points per axis with the columns of our contingency table.

Table 2.2: Correlations of the fitted coordinates of the two dimensional multidimensional scaling solution of the contingency table showing the number of events reported in the complete dataset per economic sector (industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and *other*; the economic sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*) have been merged into the miscellaneous sector *other*) differentiated by severity quartiles with the columns of the original contingency table. The first principal component shows a large negative correlation with the fourth severity quartile while the second principal component exhibits some negative correlation with the second and some positive correlation with the fourth severity quartile. These correlations provide an interpretation for the principal component axes and thus how the individual economic sectors differ from each other.

	1. Quartile	2. Quartile	3. Quartile	4. Quartile
Corr. x-axis	0.46	0.28	-0.46	-0.72
Corr. y-axis	0.10	-0.36	-0.16	0.24

We observe that the first axis of the MDS representation shows a high negative correlation for events with breached records in the fourth severity quartile. It is also shows a notable

[1] A goodness of fit measure is given by dividing the sum of the eigenvalues of the components used for the representation by the total sum of the eigenvalues, i.e. for the two dimensional representation this equals $(\sum_{i=1}^{2} \lambda_i)/(\sum_{i=1}^{9} \lambda_i)$ [35].

negative correlation for events with breached records in the third severity quartile and a notable positive correlation for events with breached records in the lowest severity quartile. Hence large positive x-coordinates can be associated with a high frequency of breaches in the first severity quartile *or* breaches with very few severe breaches in the third and fourth severity quartile. If sectors are positioned close to the origin along the x-axis, it indicates that they mostly suffer breaches with severity in the second quartile or across the complete spectrum of the considered breach severity[2].

If we consider again the right plot of figure 2.1, we note the following:

- The education sector (*edu*) is positioned at the right outer edge along the x-axis as it shows a high frequency of events in the lowest severity quartile and a comparably low rate of events in the fourth.
- The healthcare sector (*56*) also shows a rather low number of events for the fourth severity quartile but it has a lower x-coordinate as it suffers mostly from breaches of the second severity quartile and not of the first.
- The technology sector (*57*) is located on the far left along the x-axis, as it mostly suffers from breaches of the fourth severity quartile and from notably fewer from the first.
- For the telecommunication sector (*58*) we observe the lowest number of events in the first and the highest in the fourth severity quartile. As there is no sharp increase of the frequency from the first to the fourth severity quartile it is not positioned as far away from the origin as the technology sector (*57*).
- The consumer cyclical sector (*53*) has a similar x-coordinate as the telecommunication sector (*58*) since it has a high rate of events in the third and fourth quartile but the frequency for the first and second severity quartile are not considerably lower.

The y-axis does not show as strong correlations with the columns of the contingency table as the x-axis, but we can still observe a negative correlation with the second quartile and a positive correlation with the fourth.

If we look at figure 2.1, we see that the education sector (*edu*) shows the largest y-score as it mostly suffers from breaches of the first severity quartile and from notably fewer from the other three. Hence the large positive y-coordinate shows this discrepancy for the second severity quartile. The technology sector (*57*) also shows a large positive y-score as it mostly suffers from the fourth severity quartile and fewer from the second. The same holds true for the telecommunication sector (*58*). However in this case the y-coordinate is a bit lower as the discrepancy is not as profound as for the technology sector (*57*). The healthcare sector (*56*) shows a large negative y-score as it mostly suffers from breaches of the second severity quartile and from fewer of the fourth.

2.1.2 Three dimensional multidimensional scaling representation

We are also going to consider the three dimensional representation and therefore add the contribution of the third largest eigenvalue (see figure 2.2). We note that the financial sector (*55*) and the healthcare sector (*56*) are closest to the xy-plane. For the consumer cyclical sector (*53*), the industrial sector (*52*), the education sector (*edu*) and the technology sector (*57*) we observe positive z-coordinates. However, while the industrial (*52*) and consumer cyclical sector (*53*) are positioned relatively close to each other in the xy-plane, the education (*edu*) and technology sector (*57*) are located far apart on opposite sides.

The telecommunication sector (*58*), the consumer non-cyclical sector (*54*) and the miscellaneous sector *other* show negative z-coordinates, whereby we see that the miscellaneous

[2] Recall that our dataset only considers events with at least $70k$ records lost.

sector *other* and the consumer non-cyclical (*54*) are very close to each other along all three coordinates.

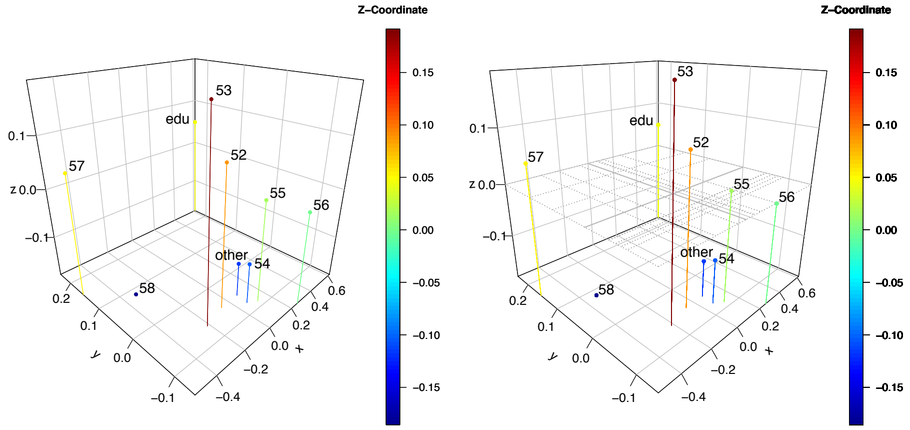


Figure 2.2: Left plot: Three dimensional representation given by the multidimensional scaling solution of the contingency table showing the number of events reported in the complete dataset per economic sector (industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and *other*; the economic sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*) have been merged into the miscellaneous sector *other*) differentiated by severity quartiles. The colors differentiate between the z-axis factor scores (see color bar on the right). Right plot: The same plot as on the left showing additionally a projection of the economic sectors onto the xy-plane. The third principal component provides in particular further differentiation between the economic sectors which are located close to each other in the xy-plane (consider for example the economic sectors industrials (*52*), consumer non-cyclicals (*54*) and *other*).

For the three dimensional representation we get a goodness of fit equal to 1, which means that we were able to find a representation in $\mathbb{R}^3$ which yields exactly the observed differences of our contingency table. Below we show the correlation of the different axis with the columns of the contingency table.

Table 2.3: Correlation of the fitted coordinates of the three dimensional multidimensional scaling solution with the columns of the contingency table showing the number of events reported in the complete dataset per economic sector industrials (industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and *other*; the economic sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*) have been merged into the miscellaneous sector *other*) differentiated by severity quartiles with the columns of the original contingency table. The first principal component shows a large negative correlation with the fourth severity quartile while the second principal component exhibits some negative correlation with the second and some positive correlation with the fourth severity quartile. The third principal component shows a large positive correlation with the first and third severity quartile. These correlations provide an interpretation for the principal component axes and thus how the individual economic sectors differ from each other.

	1. Quartile	2. Quartile	3. Quartile	4. Quartile
Corr. x-axis	0.46	0.28	-0.46	-0.72
Corr. y-axis	0.10	-0.36	-0.16	0.24
Corr. z-axis	0.71	0.54	0.77	0.46

We see that for the z-axis we get a high positive correlation for both the first and third severity quartile. Therefore, sectors with a large positive z-coordinate seem to suffer in particular from breaches of the first or third severity quartile.

Combining this interpretation with the ones from the other two axes and the contingency table 2.1 from the beginning, we note the following:

- The education sector (*edu*) seems to be very different from all other economic sectors. It mostly suffers from breaches of the first severity quartile and notably fewer from the other three.

- The technology sector (*57*) is also positioned separately from the others as it mostly experiences breaches from the fourth severity quartile and also numerous from the third.
- The telecommunication sector (*58*) is close to the technology sector (*57*) in the xy-plane, but their z-coordinates are very different. This is due to the fact that the telecommunication sector (*58*) shows a lower number of events with breach severity in the third quartile and not as clear of an increase along the severity quartiles as the technology sector (*57*) does. However, this could also be due to the lower number of observations compared to the technology sector (*57*).
- The consumer cyclical sector (*53*) has the highest z-coordinate as it shows an accumulation of breaches from the third severity quartile.
- The industrial sector (*52*) suffers breaches from all severity quartiles[3], whereby it also shows a slightly higher rate for the third severity quartile. Therefore it is positioned close to the origin in the xy-plane and shows a positive z-coordinate.
- The grouped sector *other* and the consumer non-cyclical sector (*54*) seem to be quite similar as they are closely located to each other. Both of them experience in particular breaches from the second severity quartile and fewer from all the others.
- The financial sector (*55*) and the healthcare sector (*56*) show very similar breach frequency pattern as they suffer from the complete spectrum but show an accumulation in the second quartile and a decreasing frequency for the higher quartiles. The healthcare sector (*56*) is distanced apart from the financial sector (*55*) along the y-axis as the accumulation in the second quartile is more profound.

2.1.3 Conclusion

Based on the above observations we draw the following conclusions:

1. For the frequency one can combine the miscellaneous sector *other* and the consumer non-cyclical sector (*54*) as they show a very similar frequency pattern with respect to severity quartiles.
2. Another possible risk class could be given by the technological sector (*57*) and the telecommunication sector (*58*). Even though MDS clearly separates them, if we consider the contingency table 2.1 one might wonder if this is due to the overall lower rate of events that the telecommunication sector (*58*) suffers, which can result in a less clear frequency pattern.
3. Another possible pair is given by the financial sector (*55*) and the healthcare sector (*56*). Both of them suffer from the complete spectrum with an accumulation in the second quartile, whereby this is more profound for the healthcare sector (*56*).
4. The remaining sectors (education (*edu*), industrials (*52*), consumer cyclicals (*53*)) should be considered as different risk classes since they are clearly separated in the MDS solution and show different frequency patterns in the contingency table 2.1.

2.2 Multiple factor analysis

2.2.1 Dominating subgroups

Even when taking the different timespans of the datasources into account, the number of events reported per year varies a lot (see table 2.4). Therefore we would like to identify

[3] This is not surprising if one considers that this sector offers a lot of services to other sectors and therefore operates in various fields.

years that have a similar frequency distribution with respect to breach severity (again considered in quartiles) for different economic sectors. We use multiple factor analysis (MFACT) [32] to track the evolution of the frequency distribution with respect to different severity quartiles over time. For this we have set up 13 contingency tables as the one for MDS, whereby each table summarizes the events of a given year. MFACT applies correspondence analysis (CA) to the individual contingency tables and then balances the influence of the individual tables in the overall analysis [2, 26].

Table 2.4: Number of data breaches per year for the complete dataset, built from the three datasources Privacy Rights Clearinghouse, breach level index and Information is Beautiful. The variation between the years is partially explained by the different timespans of the different datasources (Privacy Rights Clearinghouse records events with a date from 2005 onwards, breach level index for the time period 2013 until the end of 2017 and Information is Beautiful since 2004).

Year	# Events	Year	# Events	Year	# Events
2004	1	2009	27	2014	122
2005	31	2010	42	2015	107
2006	50	2011	48	2016	183
2007	49	2012	48	2017	129
2008	47	2013	90	2018	19

Recall that our observations range from 2004 until end of May 2018, which is a broad time spectrum considering the number of categories and observations we have. As an example, we show two contingency tables for 2008 and 2015 in table 2.5.

Table 2.5: Left table: Number of events reported in the complete dataset (built from the three datasources Privacy Rights Clearinghouse, breach level index and Information is Beautiful) in 2008 per economic sector (industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and *other*; the economic sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*) have been merged into the miscellaneous sector *other*) differentiated by severity quartiles. Right table: The analogous table as on the left for 2015. The observed variation in the number of events for the two years is partially due to the different timespans of the different datasources (Privacy Rights Clearinghouse records events with a date from 2005 onwards, breach level index for the time period 2013 until the end of 2017 and Information is Beautiful since 2004)).

Sector	1. Quartile	2. Quartile	3. Quartile	4. Quartile	Sector	1. Quartile	2. Quartile	3. Quartile	4. Quartile
52	2	1	2	2	52	5	1	5	3
53	4	0	0	0	53	6	1	6	3
54	0	0	0	1	54	1	4	0	0
55	1	5	2	3	55	2	1	4	1
56	6	3	1	0	56	1	5	2	5
57	0	0	0	2	57	5	4	8	15
58	0	0	0	1	58	0	4	1	3
edu	5	1	0	3	edu	2	2	1	2
other	0	0	1	1	other	2	0	1	1

As expected, the contingency tables do show some sparse entries. In 2004 we only have one event reported, and we will therefore include it into the 2005 table. Moreover, we have to keep in mind that 2018 is only partially represented, since the data was extracted at the beginning of June in 2018. We start by applying MFACT to these 13 yearly tables.

First of all, we note that the two dimensional plot in figure 2.3 shows less than 50% of the total inertia, which is an analogous measure for the total variance explained. In MFACT the first eigenvalue is between 1 and the number of tables considered. If it is close to the maximum the first dimension of the individual CAs are considered to be similar and would therefore justify a simultaneous analysis of the individual contingency tables [26]. As this is not the case, we consider several subgroups. This yields the following observations:

1. Our set of sectors can be divided into two subgroups. The first subgroup contains the economic sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*), the second consumer non-cyclicals (*54*), telecommunication services (*58*), education (*edu*) and *other* (the economic sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*) have been merged into the miscellaneous sector *other*). Considering all of them together yields

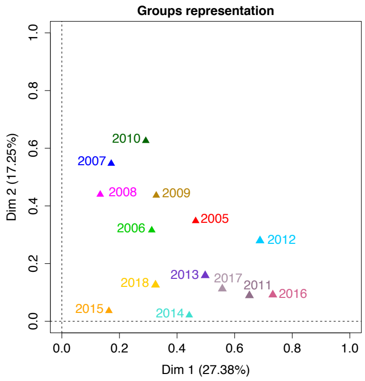


Figure 2.3: Group representation plot of the multiple factor analysis of the contingency tables showing the number of events in a year (for 2005 until 2018) per economic sector (industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and *other*; the economic sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*) have been merged into the miscellaneous sector *other*) differentiated by severity quartiles. The total inertia shown in the plot (27.38% for the first axis, 17.25% for the second) is less than 50%. As the former is an analogous measure to the total variance explained the considered two dimensional groups representation does not display most of the variance observed in the compromise table of the yearly contingency tables and thus makes a reliable comparison between different years challenging.

a representation which is dominated by the second group. If we look at the two groups we can note the following: Firstly, *54*, *58* and *other* are the sectors with the lowest rate of events and all of them are considerably lower than the ones observed in the first group. For the educational sector the overall rate of events is not as low as for *54*, *58* and *other*, but it is still notably lower than the overall rates from the first subgroup.

Table 2.6: Number of events reported in the complete dataset per economic sector (industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and *other*; the economic sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*) have been merged into the miscellaneous sector *other*). The economic sectors can be split into two groups by their total number of events, i.e. the ones with an overall total below 100 (consumer non-cyclical (*54*), telecommunication (*58*) and *other*) and above 100 (industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*).

	52	53	54	55	56	57	58	edu	other
#	164	138	33	125	145	203	38	87	60

Secondly, another similarity between the sectors consumer non-cyclicals (*54*), telecommunication services (*58*), education (*edu*) and *other* is given by the number of sparse entries in the contingency tables. These four sectors show most zero entries among the quartiles for contingency tables on a yearly basis. Even though MFACT takes the different levels of frequencies into account it is probably the combination of these two facts that leads to the dominating role of the second subgroup in the analysis with all sectors on a yearly basis.

Table 2.7: Number of sparse entries per economic sector (industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and *other*; the economic sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*) have been merged into the miscellaneous sector *other*) in the yearly contingency tables of the number of events reported in the complete dataset per economic sector (same as the aforementioned ones) differentiated by severity quartiles. In particular economic sectors with a low total number of events reported in the complete dataset show a high number of sparse entries.

	52	53	54	55	56	57	58	edu	other
# sparse entries	4	13	36	5	10	17	34	20	22

2. The contingency table for the year 2009 appears to be quite different due to its low overall number of events. Considering again table 2.4, we can see a notable decrease in comparison to the other years. The same holds true for 2018 due to its incompleteness and the time span of the different datasources (see section 1.5.1).

MFACT has a weighting mechanism in place that takes care of the different frequency levels. However, due to the observed sparsity for some sectors we have decided on analyzing the two previously mentioned subgroups separately. Furthermore, considering the low number of overall events, we have also grouped some of the years in order to reduce the number of contingency tables.

2.2.2 Sector subset industrials, consumer cyclicals, financials, healthcare and technology

In the following the sectors consumer non-cyclical (*54*), telecommunication services (*58*), education (*edu*) and *other* are excluded. For the remaining sectors the overall numbers of events per year are shown in table 2.8.

Table 2.8: Number of events reported in the complete dataset (built from the three datasources Privacy Rights Clearinghouse, breach level index and Information is Beautiful) per year from the economic sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*). The variation between the years is partially explained by the different timespans of the different datasources in which events were reported (Privacy Rights Clearinghouse records events with a date from 2005 onwards, breach level index for the time period 2013 until the end of 2017 and Information is Beautiful since 2004).

Year	# of Events	Year	# of Events	Year	# of Events
2004	1	2009	16	2014	96
2005	20	2010	35	2015	83
2006	36	2011	41	2016	149
2007	39	2012	35	2017	103
2008	34	2013	69	2018	18

We will exclude the years 2005 (incl. 2004), 2009 and 2018 as they might again distort the representation due to their low rate of events. Considering the remaining years on an individual basis gives a poor fit. A remedy for this is to group consecutive years in order to reduce the number of contingency tables. It is hereby very important to keep in mind that the following results depend on the grouping of the years.

The first overall eigenvalue is 3.29 which is close to 4, the number of tables considered, and it therefore seems appropriate to analyze the groups simultaneously. From the plots in figure 2.4 we note the following:

- In the plot on the left there is a clear separation between the years 2006-2008 and the other three groups, which indicates an evolvement over time of the frequency distribution with respect to severity quartiles for the different sectors. The groups of the more recent years are positioned more closely together. However, hereby there is also a clear distinction visible between the two groups 2010-2012, 2013-2015 and the last group 2016-2017.

- Looking at the third plot of figure 2.4, we also observe a high variation of the estimates in the factor scores for the different tables. The compromise factor scores are colored in black and especially for the sectors industrials (*52*), financials (*55*) and technology (*57*) we observe a broad range of partial factor scores around the compromise score.

  An analysis of the contributions from the different sectors to the principal components of the separate CAs (table not shown) shows that a high positive score along the first principal component is strongly negatively associated with severe breaches from the fourth severity quartile. The technology sector (*57*) is positioned far out

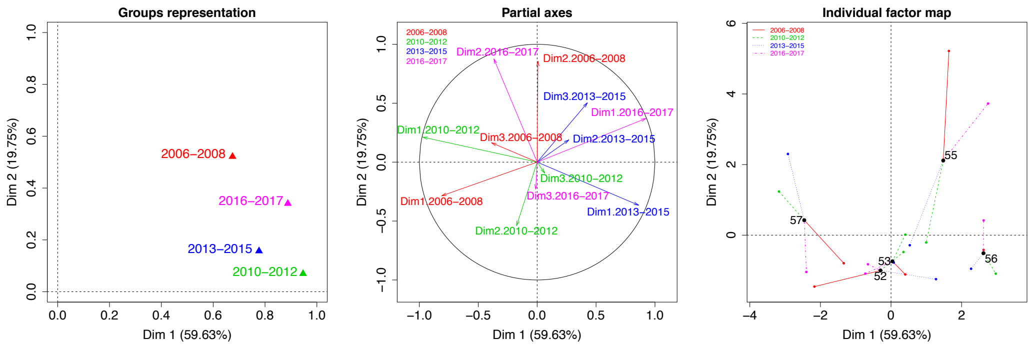


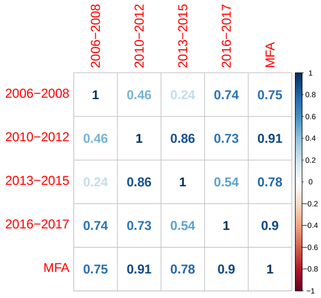
Figure 2.4: Left plot: Groups representation plot of the multiple factor analysis of the contingency tables of the yearly groups 2006-2008, 2010-2012, 2013-2015 and 2016-2017, whereby each contingency table shows the number of events reported within the years specified by the respective yearly group per the considered economic sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*) differentiated by severity quartiles. The total inertia shown in the plot is quite high (59.63% in the first dimension, 19.75% in the second). As the former is an analogous measure to the total variance explained, this two dimensional groups representation plot displays most of the variance observed in the compromise table of the contingency tables of the yearly groups. A simultaneous analysis of the different yearly groups is only appropriate to do if the first compromise eigenvalue is close to the number of tables considered. In this case this is given as the first overall eigenvalue equals 3.29 and we consider four contingency tables simultaneously. Middle plot: For the aforementioned multiple factor analysis we show the correlations of the first three principal components of the individual correspondence analyses of the yearly groups tables 2006-2008, 2010-2012, 2013-2015 and 2016-2017 with the first two compromise principal components (along the x-axis the correlation with the first compromise principal component is shown, along the y-axis the correlation with the second principal component). The length and angle of the correlation arrows of the principal components of the individual yearly groups show how much a principal component contributes to a compromise principal component. In particular we observe that all first principal components of the individual yearly groups tables are very correlated with the first compromise principal component and thus the first principal component is a good representation for all of them. The second compromise principal component mostly reflects the second principal components of the yearly groups 2006-2008 and 2016-2017. Right plot: For the aforementioned multiple factor analysis we show the partial factor scores of the considered economic sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*) of the individual correspondence analyses of the yearly groups tables 2006-2008 (red), 2010-2012 (green), 2013-2015 (blue), 2016-2017 (pink) and the respective compromise factor scores (black). In particular for the sectors industrials (*52*), financials (*55*) and technology (*57*) we observe a high variation of the partial factor scores which is an indicator for a high variation of the frequency across severity quartiles over time.

on the left which shows that it suffers in particular from these breaches. Moreover, the plot tells us that this is the case for all yearly groups. On the other side we see a notable positive compromise factor score for the financial (*55*) and healthcare (*56*) sector. Both of them seem to suffer less breaches from the fourth severity quartile. These observations are in line with what we have seen in the MDS section 2.1.

The interpretation of the second compromise axis is not as straight forward. By looking at the contribution from the individual CAs we observe that almost half of the contribution to this axis stems from 2006-2008 and roughly a third from 2016-2017 (this can also be seen in the partial axes plot). This major contribution to the second principal components in these two individual tables stem from the financial (*55*) and healthcare (*56*) sectors, whereby a large positive score is associated with many breaches from the first severity quartile or not so many from the second. Considering the individual tables we see that the financial sector (*55*) shows most of the events in either the first or second quartile, whereby the healthcare sector has always suffered mostly from events from the second severity quartile. Hence the second compromise principal component is primarily showing the discrepancy between these two sectors for the two mentioned yearly groups.

The first observation for the groups representation plot can be verified by looking at the coefficient of similarity [3] between the different contingency tables shown in figure 2.5.

We see especially low correlations between 2006-2008 and both groups 2010-2012 and 2013-2015 which matches their position in the groups representation plot. As expected, the groups 2010-2012 and 2013-2015 show a higher correlation coefficient than each of



	2006–2008	2010–2012	2013–2015	2016–2017	MFA
2006–2008	1	0.46	0.24	0.74	0.75
2010–2012	0.46	1	0.86	0.73	0.91
2013–2015	0.24	0.86	1	0.54	0.78
2016–2017	0.74	0.73	0.54	1	0.9
MFA	0.75	0.91	0.78	0.9	1

Figure 2.5: Pairwise similarity correlation coefficients of the contingency tables of the yearly groups 2006-2008, 2010-2012, 2013-2015, 2016-2017, whereby each contingency table shows the number of events reported within the years specified by the respective yearly group per the considered economic sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*) differentiated by severity quartiles, and the compromise table (MFA). We observe both large and small positive similarity correlation coefficients for various pairs of yearly groups, which indicates that some yearly groups can be considered similar (e.g. 2010-2012 and 2013-2015) but that there has as well been an evolvement over time as for example the two groups 2006-2008 and 2013-2015 show a very low similarity correlation coefficient.

them individually with the group 2016-2017. Hence the two groups 2010-2012 and 2013-2015 can be considered to be more similar than 2016-2017. What is surprising, is the low value of the correlation coefficient between 2013-2015 and 2016-2017 of 0.54 and the rather high value between 2006-2008 and 2016-2017 of 0.74.

Overall it appears that for the sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*) we have found three different similarity clusters of the frequency distribution with respect to severity quartiles. The first one consists of 2006-2008, the second of 2010-2012, 2013-2015 and the third cluster of 2016-2017.

2.2.3 Sector subset consumer non-cyclicals, telecommunication services, education and other

We proceed analogously as in the previous section for the subset of sectors consumer non-cyclical (*54*), telecommunication services (*58*), *other* and education (*edu*). The yearly number of events observed for this subset is shown in table 2.9.

Table 2.9: Number of events reported in the complete dataset (built from the three datasources Privacy Rights Clearinghouse, breach level index and Information is Beautiful) per year from the economic sectors consumer non-cyclical (*54*), telecommunication services (*58*), education (*edu*) and *other*. The economic sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*) have been merged into the miscellaneous sector *other*. The variation between the years is partially explained by the different timespans of the different datasources in which events were reported (Privacy Rights Clearinghouse records events with a date from 2005 onwards, breach level index for the time period 2013 until the end of 2017 and Information is Beautiful since 2004).

Year	# of Events	Year	# of Events	Year	# of Events
2004	0	2009	11	2014	26
2005	11	2010	7	2015	24
2006	14	2011	7	2016	34
2007	10	2012	13	2017	26
2008	13	2013	21	2018	1

In the following we only exclude the years 2004 and 2018. For the resulting 13 contingency tables there are six columns without any observations. Since the inverse of the column weights are needed in the individual analyses the current method will not work with such sparse tables. Therefore we directly consider five groups of consecutive years. The first eigenvalue of the compromise fit is 3.73, whereby the maximum is at 5. This is not very close, but we can still try to infer something from this fit.

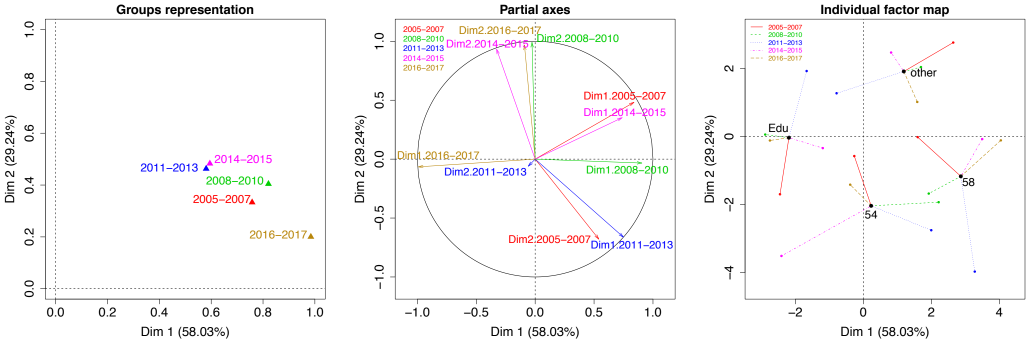
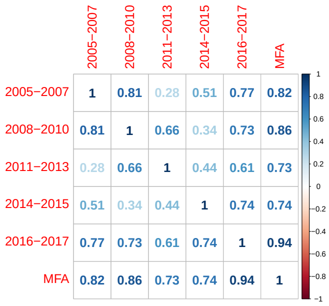


Figure 2.6: Left plot: Groups representation plot of the multiple factor analysis of the contingency tables of the yearly groups 2005-2007, 2008-2010, 2011-2013, 2014-2015 and 2016-2017, whereby each contingency table shows the number of events reported within the years specified by the respective yearly group per the considered economic sectors consumer non-cyclical (*54*), telecommunication services (*58*), education (*edu*) and *other* (the economic sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*) have been merged into the miscellaneous sector *other*) differentiated by severity quartiles. The total inertia shown in the plot is quite high (58.03% in the first dimension, 29.24% in the second). As the former is an analogous measure to the total variance explained, this two dimensional groups representation plot displays most of the variance observed in the compromise table of the contingency tables of the yearly groups. A simultaneous analysis of the different yearly groups is only appropriate to do if the first compromise eigenvalue is close to the number of tables considered. In this case this is not given as the first overall eigenvalue equals 3.73 and we consider five contingency tables simultaneously. Middle plot: For the aforementioned multiple factor analysis we show the correlations of the first three principal components of the individual correspondence analyses of the yearly groups tables 2005-2007, 2008-2010, 2011-2013, 2014-2015 and 2016-2017 with the first two compromise principal components (along the x-axis the correlation with the first compromise principal component is shown, along the y-axis the correlation with the second principal component). The length and angle of the correlation arrows of the principal components of the individual yearly groups show how much a principal component contributes to a compromise principal component. For both compromise principal components we observe yearly groups principal components which do not point into the same direction and thus make a comparison challenging. Right plot: For the aforementioned multiple factor analysis we show the partial factor scores of the considered economic sectors consumer non-cyclical (*54*), telecommunication services (*58*), education (*edu*) and *other* of the individual correspondence analyses of the yearly groups tables 2005-2007 (red), 2008-2010 (green), 2011-2013 (blue), 2014-2015 (pink) 2016-2017 (brown) and the respective compromise factor scores (black). For all considered economic sectors we observe a high variation of the partial factor scores which is an indicator for a high variation of the frequency across severity quartiles over time.

To assess the similarity between different groups from the groups representation plot in figure 2.6, we directly consider the similarity correlation coefficient between the different contingency tables shown in figure 2.7. Hereby we make the following observations:



	2005–2007	2008–2010	2011–2013	2014–2015	2016–2017	MFA
2005–2007	1	0.81	0.28	0.51	0.77	0.82
2008–2010	0.81	1	0.66	0.34	0.73	0.86
2011–2013	0.28	0.66	1	0.44	0.61	0.73
2014–2015	0.51	0.34	0.44	1	0.74	0.74
2016–2017	0.77	0.73	0.61	0.74	1	0.94
MFA	0.82	0.86	0.73	0.74	0.94	1

1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Figure 2.7: Pairwise similarity correlation coefficients of the contingency tables of the yearly groups 2005-2007, 2008-2010, 2011-2013, 2014-2015, 2016-2017, whereby each contingency table shows the number of events reported within the years specified by the respective yearly group per the considered economic sectors consumer non-cyclical (*54*), telecommunication services (*58*), education (*edu*) and *other* differentiated by severity quartiles, and the compromise table (MFA). We observe both large and small positive similarity correlation coefficients for various pairs of yearly groups, which indicates that some yearly groups can be considerd similar (e.g. 2005-2007 and 2008-2010) but that there has as well been an evolvement over time as for example the two groups 2005-2007 and 2011-2013 show a very low similarity correlation coefficient.

- What we observe from the table is that the plot is not telling us everything and might even be misleading. The group 2016-2017 is clearly separated from all other years, while it is the group that shows the highest similarity coefficients between $0.61 - 0.77$ with all other groups. While the groups 2011-2013 and 2014-2015 are close in the plot, they show a low coefficient of similarity of 0.44. On the other hand, 2005-2007 and 2008-2010 are close to each other in the plot and show a similarity coefficient of 0.81, which one would expect based on the groups representation plot.

- Also when considering the partial axes plot in figure 2.6 we see that the first dimensions of the principal component of groups 2005-2007 and 2011-2013 are likely to be too far away for a clear interpretation of the first compromise dimension. Even when looking at the contribution of the sectors to the individual first principal components there is no clear interpretation.
- It becomes evident from the third plot of figure 2.6 that for most observations the partial factor scores cover a wide range of the principal component plane and even cross compromise axes. Especially the consumer non-cyclical (*54*) and telecommunication (*58*) sector do not show many events. Therefore the number of events is too low and the tables are too sparse to observe any clear similarities or dissimilarities between the different sectors and yearly groups.

2.2.4 Conclusion

We had to simplify the individual tables in order to get a reliable group representation. In particular we had to partition the economic sectors into two sets as otherwise one of them dominates the fit due to its sparsity and overall low rate of events. Moreover, we also had to group consecutive years in order to reduce the number of contingency tables. With these restrictions in mind, we were able to identify for the first subgroup (sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*)) some similarities in the frequency distribution for the years 2010-2012 and 2013-2015. The group 2016-2017 appears to be more similar to 2006-2008 and 2010-2012 than to 2013-2015. In particular it seems like the frequency distribution of the groups 2010-2012 and 2013-2015 is rather different from the one observed in 2006-2008.

For the second subgroup (consumer non-cyclical (*54*), telecommunication services (*58*), education (*edu*) and *other*) the group representation is not reliable. Based on the similarity correlation coefficients of this subgroup one can consider the older groups 2005-2007 and 2010-2012 to be similar, but otherwise the groups should be considered separately. Interestingly the group 2016-2017 shows some degree of similarity with all other groups.

Generally we observed throughout the various fits a high variation between the partial factor scores, which is an indicator for a high variation of the frequency across severity quartiles over time.

3 Frequency

For analyzing the frequency of data breaches we start by looking at the histograms of the monthly, quarterly and half-yearly counts of the complete dataset in figure 3.1 (i.e. of data breach events with at least $70k$ items lost). For all three granularities we note two things:

- There is an apparent shift of the frequency level from 2013 onwards.
- From mid 2016 onwards there is a decline of the number of data breach events, which raises again the question of missing events and the length of a reporting delay.

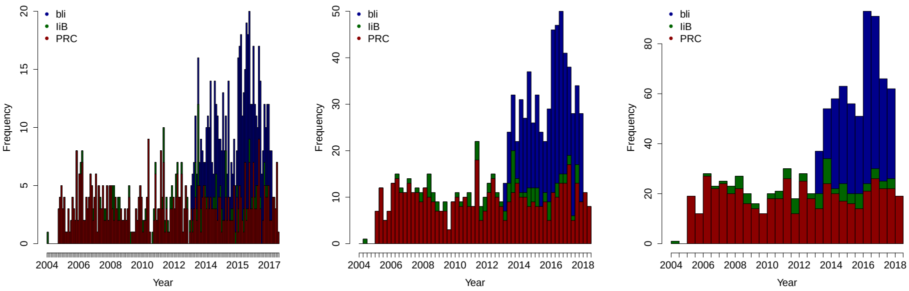


Figure 3.1: Histograms of frequency counts for monthly (left), quarterly (middle) and half-yearly buckets (right) for the complete dataset (i.e. for events with at least $70k$ records lost; the dataset is built from the three datasources Privacy Rights Clearinghouse, breach level index and Information is Beautiful). The datasources of the events are differentiated by colors: breach level index (*bli*; blue), Information is beautiful (*IiB*; green) and Privacy Rights Clearinghouse (*PRC*; red).

The observed non-stationarity is partly due to the different timespans of the databases that have been used to generate the dataset. To circumvent these artificial shifts we consider the following two subsets to model the frequency:

- Privacy Rights Clearinghouse (*PRC*) for 2005-2018,
- all events between the beginning of 2013 and the end of 2017 from all the datasources.

3.1 Frequency fit for Privacy Rights Clearinghouse

3.1.1 Time only predictor variable

We analyze the *PRC* dataset by looking both at monthly and quarterly counts, which are shown in figure 3.2. For both sets there is some strong variation visible but we cannot identify any clear pattern or systematic relationship between the counts and the *date* variable. If we consider the empirical mean and variance, we observe for the monthly counts a slight overdispersion (here we have $\hat{\mu}_m = 3.23$ and $\hat{\sigma}_m^2 = 3.68$) and a slight underdispersion for the quarterly counts ($\hat{\mu}_q = 9.63$ and $\hat{\sigma}_q^2 = 9.07$).

As the over- and underdispersion is relatively small we have decided on using a Poisson generalized linear model (GLM) with log-link function to model the counts over time (i.e.

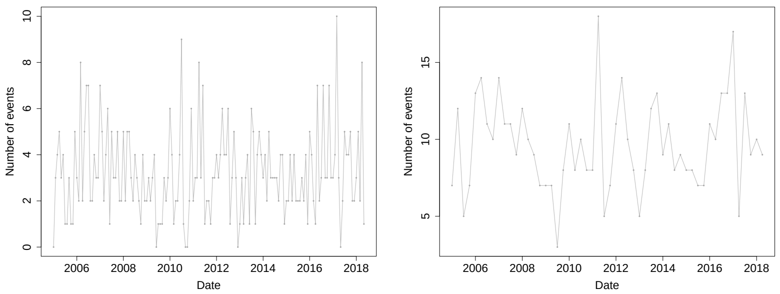


Figure 3.2: Monthly (left) and quarterly counts of events (right) for the Privacy Rights Clearinghouse datasource (considering events with at least $70k$ records lost).

we model an exponential trend in the mean over time) [31]. The residual plots and a full discussion thereof are presented in the appendix B.1.1. For both models they appear to be fine and do not show any systematic trend that remains unaccounted for in the model. In table 3.1 we show the coefficient estimates for the two fits. What clearly sticks out in both models are the non-significant p-values and the large standard errors for the *date* variable. For both models the residual deviance hardly differs from the null deviance. Moreover, for both periods the empty model is preferred according to the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) (see table 3.2).

Table 3.1: Coefficient estimates (with standard errors and z-test p-values) of the Poisson generalized linear models with log-link function for monthly and quarterly counts with *date* as predictor variable for the Privacy Rights Clearinghouse datasource (considering events with at least 70k records lost). For both models the z-test does not reject the null-hypothesis of the *date* coefficient being equal to zero at a 95% confidence level.

	Poisson model for monthly counts			Poisson model for quarterly counts		
	Estimate	Std. error	P-value	Estimate	Std. error	P-value
Intercept	9.1e-01	4.7e-01	0.056	2.1e+00	4.7e-01	8.2e-06
date	1.7e-05	3.1e-05	0.58	1.1e-05	3.1e-05	7.3e-01

Table 3.2: Residual deviance, Akaike information criterion (AIC) and Bayesian information criterion (BIC) of the Poisson generalized linear models with log-link function for monthly and quarterly counts with and without *date* as predictor variable for the Privacy Rights Clearinghouse datasource (considering events with at least 70k records lost). For both monthly and quarterly counts the model with *date* as predictor variable hardly differs from the empty model with regards to the considered goodness of fit measures.

	Residual deviance	AIC	BIC
Poisson model for monthly counts	188.4	651.3	657.4
Empty model for monthly counts	188.7	649.6	652.6
Poisson model for quarterly counts	50.1	273.9	277.9
Empty model for quarterly counts	50.2	272	274

A χ^2-test does clearly not reject the empty model in favour of the date model for both time intervals (p-values > 0.5) and we thus conclude that the empty models give the best fit as the *date* variable does not contribute in a notable way. The fits of the empty model and the model including the *date* predictor variable are shown in figure 3.3, whereby we also show the first and third quartile estimates.

3.1.2 Sector percentages as predictor variables

For the *PRC* dataset we introduce variables that specify the percentage of events from a specific sector of the total number of events reported in a period[1]. As including all of

[1] This might seem odd at first sight, as the introduced predictor variables depend on the dependent variable and thus prediction seems impossible with this model. However, the idea is to first analyze if there exists a systematic relationship between the sector percentages and the counts. If one is interested

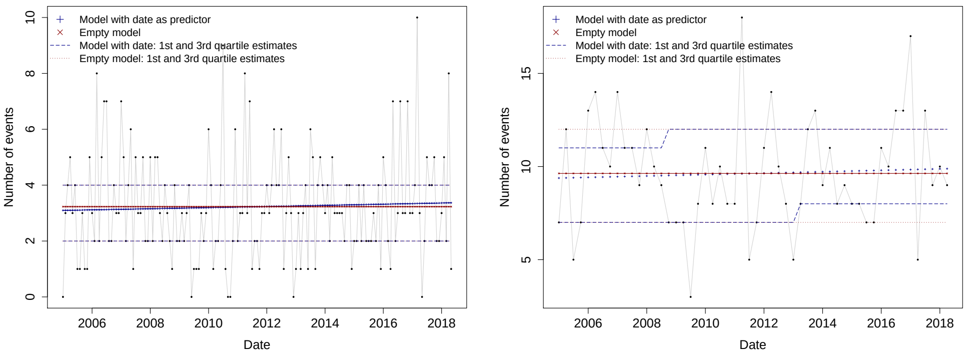


Figure 3.3: Fitted values of the Poisson generalized linear models with log-link function for monthly and quarterly counts with and without *date* as predictor variable for the Privacy Rights Clearinghouse datasource (considering events with at least 70k records lost). The model fit including the *date* as predictor variable is shown in blue and the empty model in red, estimates for the first and third quartile are also shown (dotted lines). For both the monthly and quarterly counts the fitted values of the two models are almost identical.

them would result in a linearly dependent set, we have decided on considering the sectors which belonged to the first subgroup obtained in the MFACT section 2.2.1, i.e. industrials (*52*), consumer cyclical (*53*), financials (*55*), healthcare (*56*) and technology (*57*). The introduced percentage variables for the formerly mentioned economic sectors are denoted by an *"S"* followed by the sector number.

Exploratory Analysis

The exploratory analysis of the predictors is shown in the appendix in section B.1.2. The sector percentages are not correlated and in particular for the monthly model we have a high number of zero percentages across all sectors. For the monthly dataset we also observe monthly counts for which all events originate from one sector.

Residual Analysis

We fit a Poisson GLM with log-link function to the complete dataset with the sector percentages as predictors. The complete discussion of the residual analysis is shown in appendix in section B.1.3. The residual plots are not completely satisfactory due to bent loess-smoothers and residuals which do not scatter evenly. These are partly due to the high number of zero-percentage values across the different percentage predictors, which is in particular prominent for the monthly model. The residual plots tell us that with the current model we overestimate months or quarters with a very low total and if the model predicts a high value, it generally overestimates the observations as well.

Model comparison

In table 3.3 we show the summary statistics of the two fits. For the monthly model all sector percentages contribute significantly at a 95% confidence level and their coefficients are approximately within the same range. However, for all coefficients we observe relatively large standard errors. For the quarterly counts it does not seem like the sector percentages add a lot of value as most of them are non-significant. Again we observe relatively large standard errors. Table 3.4 shows the different statistics of the two fits.

While for the monthly count model the AIC and BIC are lower in comparison to the model that only contains *date* as predictor (see table 3.2), AIC and BIC are now higher for the

in prediction, one could try to analyze the percentages and get a prediction of the latter, which could then be used to predict the counts (given that a systematic relationship between the counts and the sector percentages exists).

Table 3.3: Coefficient estimates (with standard errors and z-test p-values) of the Poisson generalized linear models with log-link function for monthly and quarterly counts with sector percentages of the respective counts as predictor variables (only for the sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*)) for the Privacy Rights Clearinghouse datasource (considering events with at least 70k records lost). For the monthly model the z-test does reject the null-hypothesis of the individual sector percentages being equal to zero at a 95% confidence level, whereby for the quarterly model for most sector percentages the null-hypothesis of the same test is not rejected (at the same confidence level).

	Poisson model for monthly counts			Poisson model for quarterly counts		
	Estimate	Std. error	P-value	Estimate	Std. error	P-value
Intercept	0.70	0.12	3.8e-03	1.7	0.25	4.2e-12
S52	0.68	0.22	2.6e-03	0.89	0.43	0.036
S53	0.58	0.23	1.1e-02	0.73	0.43	0.094
S55	0.89	0.22	5.5e-05	0.80	0.47	0.089
S56	0.48	0.19	1.3e-02	0.66	0.35	0.062
S57	0.54	0.21	1.3e-02	0.53	0.44	0.23

Table 3.4: Null deviance, residual deviance, χ^2-test p-value, Akaike information criterion (AIC) and Bayesian information criterion (BIC) of the Poisson generalized linear models with log-link function for monthly and quarterly counts with sector percentages of the respective counts as predictor variables (only for the sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*)) for the Privacy Rights Clearinghouse datasource (considering events with at least 70k records lost). Based on the χ^2-test the sector percentage model is the preferred choice for the monthly counts in comparison to the empty model while for the quarterly counts the sector percentage model does not significantly differ from the empty model at a 95% confidence level.

	Null dev.	Res. dev.	χ^2-test p-value	AIC	BIC
Monthly	188.7	167.5	0.00075	638.4	656.9
Quarterly	50.2	43.2	0.22	275	286.9

quarterly model - even in comparison to the empty model. This and the χ^2-test indicate that the empty model is a better choice for the quarterly counts (p-value: 0.22). This is supported by the fact that almost all bootstrap confidence intervals [5] for the quarterly counts model contain zero (see table 3.5).

Table 3.5: 95% bootstrap confidence interval of the Poisson generalized linear models with log-link function for monthly and quarterly counts with sector percentages of the respective counts as predictor variables (only for the sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*)) for the Privacy Rights Clearinghouse datasource (considering events with at least 70k records lost). For the monthly model zero is not included in any of the confidence intervals while for the quarterly model 0 is included in most confidence intervals of the sector percentage predictor variables.

	Intercept	S52	S53	S55	S56	S57
Monthly counts model						
lo-2.5%	0.44	0.16	0.13	0.32	0.086	0.079
estimate	0.70	0.68	0.58	0.89	0.48	0.54
up-97.5%	0.94	1.1	0.97	1.3	0.84	0.98
Quarterly counts model						
lo-2.5%	1.2	0.14	-0.18	-0.078	-0.035	-0.11
estimate	1.7	0.89	0.73	0.80	0.66	0.53
up-97.5%	2.2	1.6	1.6	1.8	1.3	1.3

Furthermore, if we do a backward stepwise selection which tries to minimize the AIC by excluding predictor variables, the best model for the quarterly counts is the empty model, which only contains the intercept. For the monthly counts the χ^2-test suggests that the sector model gives a better fit than the empty model (p-value: $7e{-}4$). An estimate for the generalization error of the monthly models can be obtained with the out-of-bootstrap sample error [5]. For the empty model we get an error of 3.688, for the model with *date* as predictor 3.762 and for the model with the sector percentages 3.896. According to the estimated generalization errors and the BIC criteria (see tables 3.2 and 3.4), the empty model is the preferred choice for the monthly counts. The three fits for the monthly counts are shown in figure 3.4.

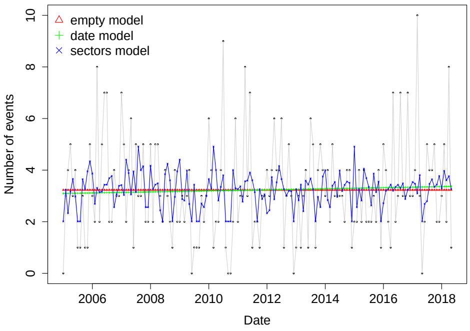


Figure 3.4: Fitted values of the Poisson generalized linear models with log-link function for the monthly counts for the Privacy Rights Clearinghouse datasource (considering events with at least 70k records lost). In red the fit of the empty model is shown, in green the fit of the model with *date* as predictor variable and in blue the fit of the model with sector percentages of the respective monthly counts as predictor variables (only for the sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*)). The latter provides a better fit but there is not enough statistical evidence for a systematic relationship between the monthly counts and the sector percentage predictor variables.

The model including the sector percentages fits the data better than the empty model, however it shows a larger generalization error and we thus do not believe the better fit was due to a systematic relationship between the counts and the sector percentages. Moreover, the residual plots for the sectors are not completely satisfactory. However, it remains an open question whether the empty model is the best one or if there are other predictor variables or a combination thereof that could yield a better fit or tell us more about the monthly frequency of events.

3.2 2013-2017 monthly count fit

In the following we consider the 2013-2017 subset for the frequency and look at monthly counts, which are shown in figure 3.5. Generally we observe an increasing trend for the monthly counts, whereby we can also make out timespans with a lower number of events, i.e. at the beginning of 2013, in the middle of 2015 and at the end of 2017.

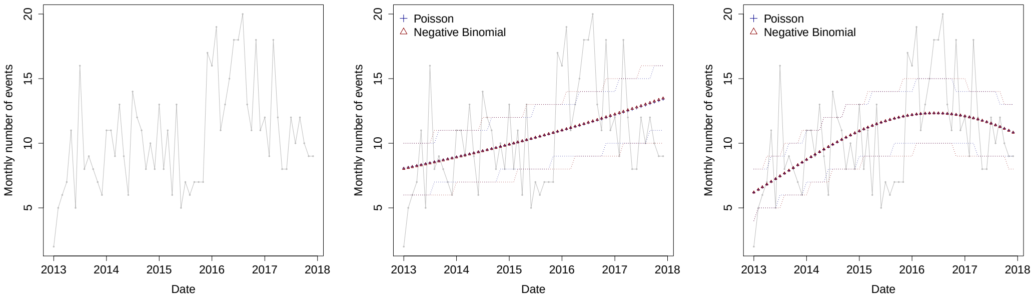


Figure 3.5: Left: Plot of the monthly frequency counts of events reported within the beginning of 2013 until the end of 2017 in the complete dataset (considering events with at least 70k records lost). Middle: The former plot on the left including the fitted values of the Poisson (blue cross) and negative binomial (red triangle) generalized linear models with log-link function and with *date* as predictor variable for the aforementioned dataset, whereby the first and third quartile estimates of both models are shown as dashed lines (blue for Poisson, red for negative binomial). Right: The analogous as the one in the middle whereby this time the Poisson and negative binomial generalized linear models take *date* as a polynomial of degree two as predictor variable. The models presented in the middle and right plot suggest two different developments: in the linear case there is an increase of the number of events over time and in the quadratic model a maximum is reached in the middle of 2016 and the number of events is decreasing afterwards.

3.2.1 Residual analysis

We start by fitting a Poisson and negative binomial GLM with log-link function and *date* as predictor variable. Doing so yields the following observations. Firstly, the Poisson and negative binomial fits are very similar (also consider table 3.6) and therefore show very similar residual plots (the residual plots are fully discussed in detail in the appendix B.1.4). The latter are not completely satisfactory as we observe slightly bent loess-smoothers. If we include the *date* variable as a polynomial of degree two, the plots improve a bit. Secondly, for both the linear and quadratic models the residuals follow mostly a stationary process but show some slight autocorrelation at a time lag of 4. (This observation is important, as we bootstrap the residuals later on.)

3.2.2 Model summary

The coefficient estimates of the four different models are shown in table 3.6. For the two linear models we have a significant increase over time, whereby for the quadratic models we observe a parabola with a negative coefficient for the quadratic term. The coefficient estimates of the Poisson and negative binomial models are very similar.

Table 3.6: Coefficient estimates (with standard errors and z-test p-values) of the Poisson and negative binomial generalized linear models with log-link function and *date* as predictor variable (top: linear in *date*, bottom: quadratic in *date*) for the monthly counts of events reported within the beginning of 2013 until the end of 2017 in the complete dataset (considering events with at least $70k$ records lost). The coefficient estimates of the Poisson and negative binomial models are almost identical (for both the linear and quadratic *date* models). The models linear in *date* suggest a significant increase over time (the *date* coefficient being equal to zero is not rejected by the z-test at a 95% confidence level) and the quadratic models suggest a parabola with a negative coefficient for the *date* coefficient of degree 2.

	Poisson			Neagtive binomial		
	Estimate	Std. error	P-value	Estimate	Std. error	P-value
Linear in date						
Intercept	2.3	0.04	<2e-16	2.3	0.046	<2e-16
date	1.2	0.31	2e-4	1.2	0.35	8.3e-4
Quadratic in date						
Intercept	2.3	0.041	<2e-16	2.3	0.044	<2e-16
date	1.3	0.33	1.1e-4	1.3	0.36	3.6e-4
date^2	-0.84	0.32	8.9e-3	-0.85	0.35	1.5e-2

The dispersion parameter θ for the negative binomial distribution specifies the relationship between the expectation and the variance via $\sigma^2_{N_t} = \mu_{N_t} + \frac{1}{\theta}\mu^2_{N_t}$. For both fits of the negative binomial model $\hat{\theta}$ shows a large standard error, whereby for the quadratic model the standard error is larger than $\hat{\theta}$.

Table 3.7: Estimated dispersion parameter $\hat{\theta}$ and their corresponding standard errors in the negative binomial generalized linear models with log-link function and *date* as predictor variable (first row: linear in *date*, second row: quadratic in *date*) for the monthly counts of events reported within the beginning of 2013 until the end of 2017 in the complete dataset (considering events with at least $70k$ records lost). The standard errors are quite large and if we test for overdispersion in the corresponding Poisson generalized linear models using the test from Cameron and Trivedi [7] the null hypothesis of no overdispersion is not rejected for the model which takes *date* as a polynomial of degree two as predictor variable (p-value 18%) and borderline rejected for the linear *date* model with a p-value of 4.6% at a 95% confidence level.

	$\hat{\theta}$	Std. error
Linear in date	35	28
Quadratic in date	58	68

We test for overdispersion in the Poisson models using the test from Cameron and Trivedi [7] at a confidence level of 95% and get a clear negative result for the model with *date* as polynomial of degree two (p-value: 18%) and a borderline positive result for overdispersion for the model which is linear in *date*, with a p-value of 4.6%.

3.2.3 Model comparison

To assess which model fits the count data best, we consider both AIC and BIC as well as the generalization error estimated via bootstrapping [5, p. 44]. As we have time-dependent data with a trend and slightly correlated errors, we cannot use the classical bootstrap method. To generate valid bootstrap samples we have used a moving block bootstrap on the Pearson residuals[2] of the respective models[3] [16]. From table 3.8 we see that the Poisson model of degree two performs best according to AIC and BIC, but the generalization error is a bit larger than for the negative binomial model of degree two. Both of them clearly perform better than their linear counterparts. Since we have seen that the overdispersion is not significant in the quadratic *date* model, we prefer the Poisson over the negative binomial fit.

Table 3.8: Null deviance, residual deviance, Akaike information criterion (AIC), Bayesian information criterion (BIC) and moving block bootstrap generalization error of the Poisson and negative binomial generalized linear models with log-link function taking *date* as a predictor variable (as a polynomial of degree 1 and 2) for the monthly frequency counts of events reported within the beginning of 2013 until the end of 2017 in the complete dataset (considering events with at least $70k$ records lost). With regards to all goodness of fit measures the models which are quadratic in *date* perform better than the models which are linear in *date*.

	Null dev.	Res. dev.	AIC	BIC	Gen. error
Poisson degree 1	91.5	77.5	329.4	333.5	14.6
Poisson degree 2	91.5	70.5	324.3	330.6	13.6
Negative binomial degree 1	70.7	59.7	329	335.3	14.5
Negative binomial degree 2	77.5	59.6	325.4	333.8	13.4

3.3 Conclusion

For the *PRC* subset we have seen a constant rate for both the monthly and quarterly counts and for the quarterly counts the empty model is the best choice. For the monthly counts the model including the sector percentages performs better than the empty model with regards to some goodness of fit criteria but not all of them. Therefore we are not convinced that this is superior to the empty model and believe that a further analysis with other predictor variables or a combination thereof might give more insights. Furthermore, using a multivariate instead of a univariate approach should also be considered, as it can deal with multiple responses (e.g. number of events per economic sector per month).

For the subset 2013-2017 the quadratic models outperform their linear counterparts with regards to all goodness of fit measures. However, the final decision on which model to choose greatly depends on what happens after the beginning of 2018. Herefore we consider three possible scenarios. In scenario number 1 the observed decrease of data breaches at the end of 2017 is merely a temporary effect and the rate of events will increase again thereafter, as it was the case in 2015. In scenario number 2 the decrease at the end of 2017 is not a temporary effect and a lower rate of events will manifest itself in the future. In this case the predictive power of the quadratic model is questionable as it was estimated on data that only contains limited information about the future development. In scenario number 3 neither a clear increase or decrease of the monthly counts will occur, which might require other modelling techniques for prediction than a GLM. Another concern that comes to mind is the effect of the reporting delay. However, since we have limited the data to events between 2013-2017 and have extracted the data at the beginning of summer

[2] The Pearson residuals are given by $r_i = (N_i - \mathbf{E}[N_i])/\mathbf{E}[N_i]^{0.5}$ and a block of length four is used.

[3] As we require the bootstrap Y_i^* observations to be integers, we have applied the same rounding as in [30]: $Y_i^* = max\big(0, \lfloor \mathbf{E}[N_i] + r_i^* \mathbf{E}[N_i]^{0.5} + 0.5 \rfloor\big)$, whereby r_i^* denotes the bootstrapped Pearson residual. Moving blocks of length 2 to 4 were considered and for all of them we reached the same conclusion.

2018, the effect should be negligible[4]. Hence without considering further information of the future development we do not believe to be able to reach a sound decision on which model should be chosen for the 2013-2017 subset. Even when more information becomes available, caution is still required in future predictions due to the ever-evolving nature of data breaches [18].

A recently published study from Risk Based Security [34, 2018] reports a decline in the number of events for the first quarter of 2018 in comparison to the first quarter of 2017. Nevertheless we cannot reach a firm conclusion as their results are based on a dataset that considers both events for which the number of records lost is unknown[5] as well as known. This highlights another limitation of the model due to the used dataset. Since we are dealing with a restricted view on the problem, the results cannot be directly generalized to the complete phenomenon and it remains unclear whether the results for events with at least $70k$ breached also hold for events with less than $70k$ items breached.

Interestingly, the models from the sections 3.1 and 3.2 tell two different stories. While for the *PRC* subset we have observed a constant rate of events and the rate remains constant even if we only consider the *PRC* events in 2013-2017. Therefore the main driver of the non-constant development in the rate of events for the 2013-2017 subset appears to originate from the other two sources, whereby most events in the period 2013-2017 stem from *bli*. Hence an additional analysis solely on the *bli* subset and a thorough comparison to the other two might reveal some further insights into the development of the rate of data breaches.

[4] In chapter 5 we analyze the reporting delay. Based on the used dataset we get an estimate for the median delay for events with at least $70k$ records lost, which equals 195 days, hence slightly above six months.

[5] In the used dataset in [34] for slightly more than 50% of the events the breach size was unknown.

4 Severity

4.1 Severity vs. single predictor variables

In the following we analyze the severity distribution of the *total records* variable. We start by comparing the *total records* variable against the other variables in the dataset. While doing so we have to keep in mind that our dataset originates from three different sources of unequal sizes, which might cause some inhomogeneities among the observations. We already know by the description of the sources that Privacy Rights Clearinghouse (PRC) focuses on events which are reported within the US. Considering the severity vs. datasource boxplot in figure 4.1, we see that this is in fact the case. The PRC subset is clearly right-skewed and shows the lowest median among all three. Breach level index (bli) is as well right-skewed but not as much as PRC and shows as well a notably higher median. On the contrary we have that Information is Beautiful (IiB) is more symmetrically distributed and therefore also shows a much higher median than the other two datasources.

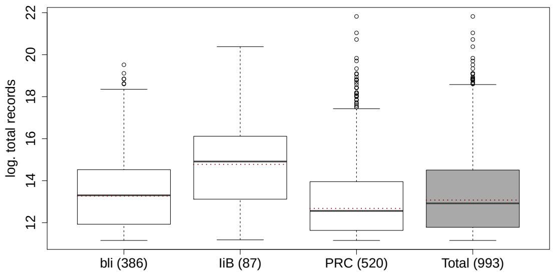


Figure 4.1: Log-transformed *total records* vs. datasource (breach level index (bli), Information is beautiful (IiB), Privacy Rights Clearinghouse (PRC)) of the complete dataset (considering events with at least $70k$ records lost), whereby the total is shown on the right (in darkgray). Enough statistical evidence is found for the groups having differently trimmed means (20% symmetric) and to follow different distribution functions (with p-values below 4.6e−4 using Welch's trimmed mean (20% symmetric) test with Windsorized variances and Welch's test on ranked data [9]).

Since we are dealing with not normally distributed observations and unequal variances among the different groups, we have used Welch's test statistic on trimmed means (20% symmetric trimming) and Windsorized variances [9] to assess whether or not there exists at least one group with a different trimmed mean from all the others at a 95% confidence level. The null hypothesis for equally trimmed means among all groups is clearly rejected with a p-value of 1.3e−14. We also used Tukey's procedure [9] to do multiple pairwise comparisons between the individual groups and for each pair the null hypothesis of equally trimmed means is rejected at a 95% confidence level with p-values below 8.6e−14. We get the same results with p-values below 4.6e−4 if we use Welch's test statistic on the ranked data to assess whether the distribution functions among the different groups are equal and if they are pairwise equal.

Therefore we need to check for differences among the three datasources in the following analysis of the severity distribution. If such differences are detected, they are mentioned in the corresponding subsections.

4.1.1 Severity vs. *date*

Below we show the logarithmized severity vs. the *date* variable, whereby we have also split the set according to the three different datasources. Fitting a truncated lognormal regression model [10] on the complete dataset shows a significant linear increase[1], which remains if we fit the model on a random half of the dataset and compare it to the fit on the other half. The same model yields a significant increase of the same magnitude over time for the *bli* subset and a borderline non-significant increase of the same magnitude for the *PRC* subset (at a confidence level of 95%). However, if we check again on two random halves, the *date* variable does not provide enough signal for a significant increase over time for any of the two subsets. While the date coefficient estimates remain of the same magnitude, a t-test does not reject the null hypothesis of them being equal to zero on both halves (p-values > 0.1 for at least one half; the *PRC* subset consists of 520 events, *bli* of 386 and *IiB* of 87). Hence we cannot determine the exact source of the observed increase, as it could be from *PRC*, *bli* or both of them.

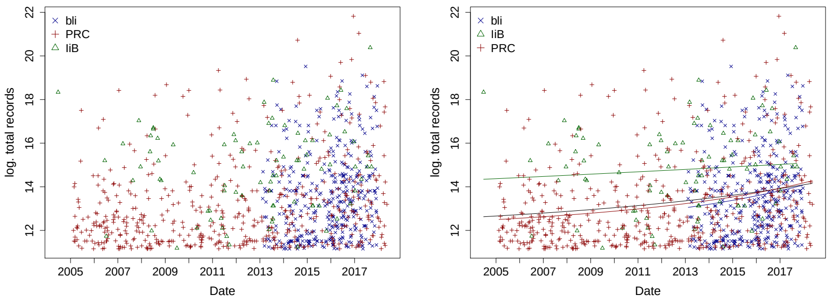


Figure 4.2: Log-transformed *total records* vs. *date* (left) and log-transformed *total records* vs. *date* including time trend estimate from the truncated regression models taking *date* as predictor variable (right) for the complete dataset (considering events with at least $70k$ records lost). The sources of the individual points have been marked: blue cross for breach level index (*bli*), red plus for Privacy Rights Clearinghouse (*PRC*) and green triangle for Information is Beautiful (*IiB*) and the same colors are used for their respective truncated regression fit. The black solid line in the right plot shows the estimated trend for the complete dataset. The datasources show different slopes over time.

4.1.2 Severity vs. *country*

The country variable is strongly dominated by the US and we therefore consider only the classes *US* and *non-US*. We can see in the boxplot in figure 4.3 that the interquartile range of the breach size is much larger for organizations headquartered outside of the US than for organizations headquartered within the US. Moreover, the median is also clearly elevated for the *non-US* group in comparison to the *US* one. While the spread of the distribution (i.e. distance between the whiskers) is much larger for the *non-US* group we do not observe any outliers (here in the sense of points outside of the whiskers), whereas there are several for the *US* group. The upper tail of the logarithmized breached records distribution is by construction right-skewed. However, we observe that the *US* group is clearly more right-skewed than the *non-US* group.

We perform the same hypothesis tests as in section 4.1 and again in both cases we can reject the null hypothesis of equally trimmed means and equal distribution functions among the

[1] The estimates of the coefficients are (standard errors are mentioned in brackets): intercept: 0.746(2.81); date: 0.0017(3.88e−4); sigma: 4.395(0.45). The t-test on the date coefficient rejects the null hypothesis of the later being equal to zero with a p-value of 6.5e−6. The *date* variable was shifted to start at 0 for the first event.

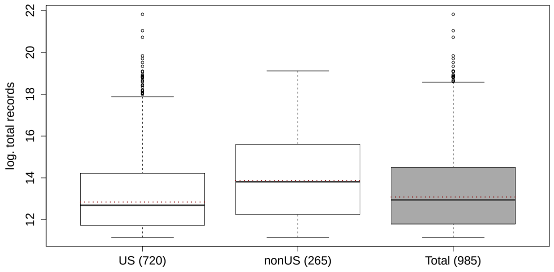


Figure 4.3: Boxplot of the log-transformed *total records* variable for the headquarter location groups *US* and *non-US* for the complete dataset (considering events with at least $70k$ records lost), the total distribution is shown on the far right (in darkgray). The 20% symmetrically trimmed mean is shown in red (dashed line) and the number of observations per group is shown in brackets next to the group label. Using Welch's trimmed mean (20% symmetric) test with Windsorized variances and Welch's test on ranked data [9] show that both their trimmed means and the distribution functions the two groups follow are unequal, (with p-values below 3.2e−8).

two groups respectively, with p-values below 3.2e−8. For this variable one might wonder if the two distributions can be considered the same above a certain threshold. This can be tested by a two-sample Kolmogorov-Smirnov test. Performing this test at a 95% confidence level shows that if the thresholds lie between 1.44 and 2.6 million, the two distributions can be considered the same (for 1.44 we have 184 observations in the *US* group and 118 in the *non-US* group, for the threshold 2.6 million 132 and 90 respectively). Afterwards the distributions differ again (p-values < .05), until we reach the last 53 observations (where 38 *US* events and 15 *non-US* events remain).

4.1.3 Severity vs. *market capitalization* and severity vs. *number of employees*

In the left plot of figure 4.4 we show the log-transformed *total records* vs. the log-transformed and inflation adjusted *market capitalization* for the 142 observations for which the latter was available. In the plot we can see two observations on the left which are located further apart from all the others. The one on the far left is a company which has lost a lot of market capitalization before the breach (Spiral Toys Inc.) and the other is an Indian company which has been in financial troubles for more than a year before the data breach happened (Aadhaar Ventures India LTD). Besides those two organizations, there is one observation at the top which catches the eye. This point represents the massive Yahoo data breach in 2016 with 3 billion records lost. All the other observations scatter evenly, whereby we have more observations in the lower half along the y-axis than in the upper half, as we are only considering the upper tail of the breached records distribution. In the plot no clear relationship between the market capitalization and the number of lost records is visible. When fitting a truncated regression model to the data [10] the *market capitalization* variable does not relate in a notable way to the *total records* variable. The same holds true if we take the datasource of the observations into account. Hence the increasing relationship between the market capitalization and data breach severity from [37] could not be confirmed within our dataset.

For the number of employees we draw the same conclusions as for the market capitalization, in particular there is no clear relationship between the *number of employees* variable and the *total records* variable visible in the right plot of figure 4.4. Also when considering the outer edges we see both very large and very small companies that suffer from a broad spectrum of data breaches.

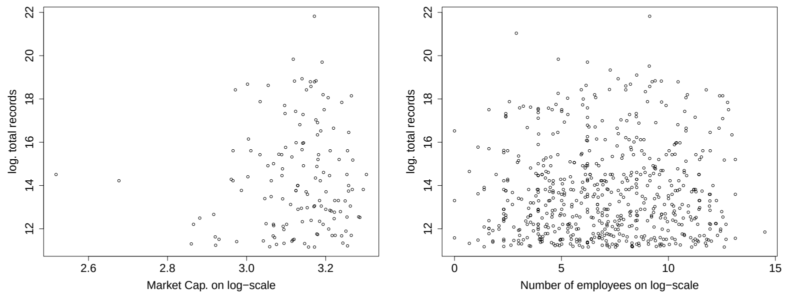


Figure 4.4: Log-transformed *total records* vs. log-transformed and inflation adjusted *market capitalization* (left) and log-transformed *total records* vs. log-transformed *number of employees* (right) for the complete dataset (considering events with at least 70k records lost). For both size variables no clear systematic relationship is visible with the number of records lost.

4.1.4 Severity vs. *economic sector*

The severity quartiles vs. the economic sector have already been discussed in section 2.1, where some differences between the sectors were found. In figure 4.5 we show the boxplot for all original economic sectors and the miscellaneous sector *other*, which consists of the sectors energy (*50*), basic materials (*51*), utilities (*59*), politics (*pol*) and military (*mil*). Notable differences both in the median level and the span of the different sectors become again evident. In particular we can verify our observations from section 2.1, as the financial (*55*) and healthcare (*56*) sector are very similar as well as the consumer non-cyclical sector (*54*) and the miscellaneous sector *other*. We observe again that the technological (*57*) sector suffers in particular from very severe breaches and the educational sector (*edu*) shows a different distribution than all the others. Moreover, we note that organizations dealing with politics and elections (*pol*) have clearly a higher median than the total median and actually show a similar severity distribution as the telecommunications (*58*) sector.

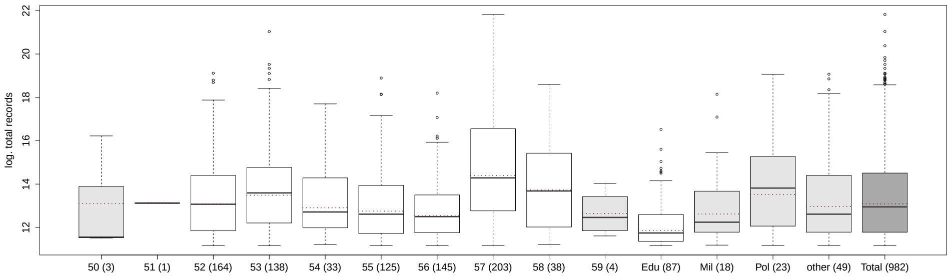


Figure 4.5: Boxplot for log-transformed *total records* vs. *economic sector* variable (taking the levels energy (*50*), basic materials (*51*), industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), utilities (*59*), education (*edu*), military (*mil*), politics (*pol*)) and the miscellaneous sector *other* (colored in lightgray and is a merger of the sectors energy (*50*), basic materials (*51*), utilities (*59*), military (*mil*) and politics (*pol*)) for the complete dataset (considering events with at least 70k records lost). On the far right the total is shown as well (in darkgray). The 20% symmetrically trimmed mean is shown in red (dashed line) and the number of observations per economic sector is shown in brackets next to the sector label. Notable differences between the economic sectors exist both in their median levels and the distribution function they follow.

Our observations are further supported by the same Welch test on the trimmed means with Windsorized variances at a 95% confidence level as in section 4.1. The null hypothesis of having equally trimmed means among the sectors is clearly rejected (p-value: 5e−20). From the multiple pairwise comparison results in table 4.1 we see that the financial (*55*) and healthcare (*56*) sector do not have significant differently trimmed means and the same holds for the consumer non-cyclical sector (*54*) and the miscellaneous sector *other*. The technological sector (*57*) has a significant differently trimmed mean than all the other sectors except for the telecommunication (*58*) sector. For the educational sector (*edu*) the null hypothesis of having equally trimmed means is rejected for each pairwise comparison with another sector.

Table 4.1: Test statistics of Tukey's multiple pairwise comparison for Welch's trimmed mean (20% symmetric) test with Windsorized variances for the *total records* variable grouped by the economic sectors (industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and the miscellaneous sector *other*, which is a merger of the sectors energy (*50*), basic materials (*51*), utilities (*59*), military (*mil*) and politics (*pol*)) based on the complete dataset (considering events with at least 70*k* records lost) at a 95% confidence level, p-values are shown in brackets. In particular the education sector has a differently trimmed mean than all the others.

	53	54	55	56	57	58	edu	other
52	-2.8(0.11)	0.78(1)	2.4(0.29)	**4.4(0.00049)**	**-8.2(0)**	-2.4(0.31)	**10.4(8.3e-14)**	0.48(1)
53		2.6(0.21)	**5.3(1.1e-05)**	**7.5(5.6e-11)**	**-5.5(2.4e-06)**	-0.88(0.99)	**13.2(4.5e-14)**	2.3(0.39)
54			0.71(1)	1.8(0.71)	**-6.4(9e-07)**	-2.6(0.22)	**5.1(0.00028)**	-0.22(1)
55				1.9(0.65)	**-10.6(9.5e-13)**	**-3.6(0.019)**	**8.1(2.1e-12)**	-0.9(0.99)
56					**-12.9(9.9e-13)**	**-4.5(0.0014)**	**7.2(3e-10)**	-1.9(0.56)
57						2.3(0.34)	**18.1(4.2e-13)**	**5.9(2.2e-06)**
58							**7.2(2.9e-07)**	2.3(0.34)
edu								**-5.2(9.4e-05)**

The results of the tests on ranked data are mostly in line with the ones from table 4.1 (table not shown). In the multiple pairwise comparison tests on the ranked data we do not get a rejection of the null hypothesis of two sectors following the same distribution function for the pairs: (industrials *(52)*, healthcare *(56)*), (consumer cyclicals *(53)*, technology *(57)*), (consumer non-cyclicals *(54)*, technology *(57)*), (financials *(55)*, telecommunication services *(58)*), (healthcare *(56)*, telecommunication services *(58)*) and (technology *(57)*, *other*) (at a 95% confidence level).

4.1.5 Severity vs. *organization type*

In figure 4.6 the boxplot for the *total records* vs. the *organization type* variable is shown. Also here we observe notable differences in the severity distribution between the various types. The median severity is the highest for publicly traded companies (group *MCAP*), whereby the latter also shows the broadest span. In sharp contrast to this is the *public* entities group which shows the lowest median and for which the third quartile is even below the overall median value. Also the not-for-profit (*NPO*) group shows a rather low median and does not spread as wide as others. Interestingly, the *private* and governmental (group *Gov*) types are similar in both median level and interquartile range.

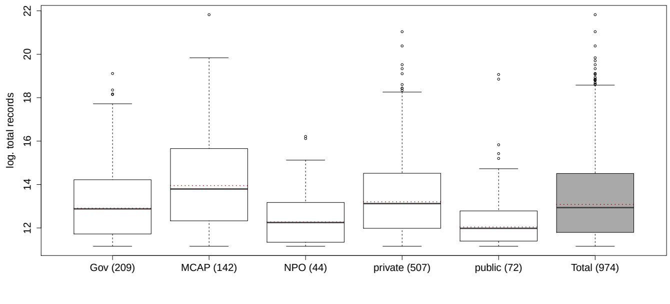


Figure 4.6: Boxplot for log-transformed *total records* vs. *organization type* variable (taking the levels government (*Gov*), market capitalized (*MCAP*), not-for-profit organizations (*NPO*), *private* and *public*) for the complete dataset (considering events with at least 70*k* records lost). On the far right the total is shown as well (in darkgray). The 20% symmetrically trimmed mean is shown in red (dashed line) and the number of observations per group is shown in brackets next to the group label. As for the economic sectors we observe notable differences between the considered groups both in their trimmed means and the distribution functions they follow.

Doing the same hypothesis tests as in section 4.1 shows that both null hypotheses of equally trimmed means and equal distribution functions among the groups are clearly rejected with p-values below 4.6e−11. The multiple pairwise comparison test shows that for the trimmed means all organization type pairs are significantly different except for *public* organizations and *NPO's* at a 95% confidence level (see table 4.2). Multiple pairwise comparison on the ranked data (see table 4.2) shows that the *MCAP* group has a distribution function that is different from all the others. The null hypothesis of a different distribution function between *government* and *private* organizations is not rejected and the *NPO* group seems to have a similar distribution function as the *government* and *public* group. However, at the same time the null hypothesis that the *government* and *public* group have the same distribution function is clearly rejected.

Table 4.2: Test statistics of Tukey's multiple pairwise comparison for Welch's trimmed mean (20% symmetric) test with Windsorized variances (top) and for Welch's test on ranked data (bottom) for the *total records* variable grouped by the *organization type* variable (taking the levels government (*Gov*), market capitalized (*MCAP*), not-for-profit organizations (*NPO*), *private* and *public*) based on the complete dataset (considering events with at least 70*k* records lost) at a 95% confidence level, p-values are shown in brackets. Based on the considered tests most groups significantly differ with regards to their *total records* trimmed means and their *total records* distribution function.

	MCAP	NPO	private	public
Welch's test on trimmed means				
Gov	**-6.3 (1.1e-08)**	**4.2 (6.5e-04)**	**-2.8 (3.8e-02)**	**6.7 (2.1e-09)**
MCAP		**9.0 (0)**	**5.0 (1.5e-05)**	**11.3 (0)**
NPO			**-6.8 (4e-08)**	1.5 (5.7e-01)
private				**10.6 (1.1e-14)**
Welch's test on ranked data				
Gov	**4.0 (9.1e-04)**	-2.7 (6.1e-02)	1.7 (4.2e-01)	**-4.1 (8.2e-04)**
MCAP		**-5.3 (1.2e-05)**	**-3.0 (2.2e-02)**	**-6.9 (1.6e-09)**
NPO			**3.9 (2.8e-03)**	-0.6 (9.7e-01)
private				**-5.6 (2e-06)**

4.1.6 Severity vs. *multiple firms*

A comparison between the severity and *multiple firms* variable is challenging as only fewer than 5% of observations are *multiple firms* events. Considering the boxplot in figure 4.7, we observe a higher median for the multiple organization events (group *TRUE*) than for the events with only one affected entity (group *FALSE*). At first sight this is in line with what one would expect, as more entities directly result in a larger exposure. However, at the same time we see that the span of the single entity events is much broader and therefore one might wonder, if there really exists a difference between the two groups. In particular we have seen in section 4.1.3 that there are also small organizations (whereby smallness is measured in number of employees), which have suffered from various sizes of breaches and that there was a very noisy relationship if any between the size of an organization and the severity of the breach. The Welch's trimmed mean (20% symmetric) test with Windsorized variances does not reject the null hypothesis of equally trimmed means at a significance level of 95% (p-value: 17%) and for the Welch test on the ranked data we get a borderline rejection of equal distribution functions with a p-value of 4.98%. Using again the two-sample Kolmogorov-Smirnov test for various thresholds shows that the null hypothesis of the two groups having the same distributions is not rejected at a 95% confidence level up to a threshold of 25.1 million, where only 5 multiple firms events remain (vs. 71 single firm events).

4.1.7 Severity vs. *insider/outsider*

The boxplot in figure 4.8 shows no notable difference between the *total records* vs. *insider* group and the *total records* vs. *outsider* group, in particular their medians and interquartile ranges are almost identical. However, what we can observe is that if it is unknown whether the breach happened because of an inside or outside party (group *unkn*), it is more likely that the breach is less severe.

Welch's trimmed mean (20% symmetric) test shows that at least one group has a different mean (p-value: 3.3e−4) and a multiple pairwise comparison confirms that the trimmed mean of the unknown group significantly differs from the other two groups with p-values

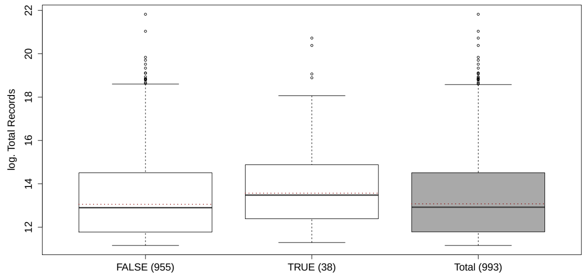


Figure 4.7: Boxplot for log-transformed *total records* vs. *multiple firms* variable (taking the level *"TRUE"* if multiple firms were involved in the same breach event and *"FALSE"* otherwise) for the complete dataset (considering events with at least 70k records lost). On the far right the total is shown as well (in darkgray). The 20% symmetrically trimmed mean is shown in red (dashed line) and the number of observations per group is shown in brackets next to the group label. A solid comparison is challenging as the two groups are of very unequal sizes.

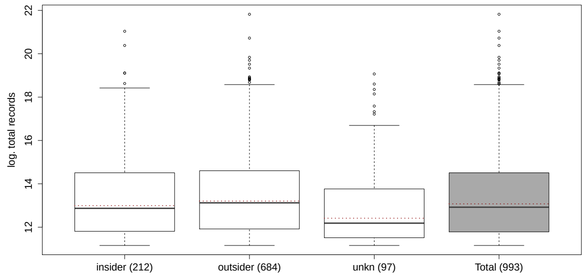


Figure 4.8: Boxplot for log-transformed *total records* vs. *insider/outsider* variable (taking the level *insider* if an inside party committed or facilitated the data breach and *outsider* if the data breach was committed or facilitated by an outside party, sometimes this is as well unknown (*unkn*)) for the complete dataset (considering events with at least 70k records lost). On the far right the total is shown as well (in darkgray). The 20% symmetrically trimmed mean is shown in red (dashed line) and the number of observations per group is shown in brackets next to the group label. The two groups *insider* and *outsider* are very similar with regards to their *total records* median levels and their *total records* distribution function. The unknown group shows a lower median and shorter span than the other two which is an indication for an information reporting problem for less severe data breaches.

below 1.02e−4 (see table 4.3). Also Welch's test on rank-ordered data rejects the null hypothesis of equal distribution functions among the groups (p-value: 0.36%). A multiple pairwise comparison rejects the null hypothesis of the unknown group and *outsider* group having the same distribution function, but not for the unknown group and the *insider* group (see table 4.3). However, one might question the latter result as the first test tells us that the trimmed means are different. Moreover, as we only consider the ranks and not the observed values the rank-based test uses less information than the trimmed mean test. Hence we conclude the unknown group to follow a different distribution function than the other two groups.

Table 4.3: Test statistics of Tukey's multiple pairwise comparison for Welch's trimmed mean (20% symmetric) test with Windsorized variances (left) and for Welch's test on ranked data (right) for the total *total records* variable grouped by the *insider/outsider* variable (taking the level *insider* if an inside party committed or facilitated the data breach and *outsider* if the data breach was committed or facilitated by an outside party, sometimes this is as well unknown (*unkn*)) for the complete dataset (considering events with at least 70k records lost) at a 95% confidence level, p-values are shown in brackets. There is no statistical evidence that the involvement of an insider or outside party has an effect on the *total records* trimmed mean or on the *total records* distribution function. The test results further support the suspicion that the reporting of information is of less quality or quantity for less severe data breaches than for larger ones.

	Welch's test on trimmed means		Welch's test on ranked data	
	outsider	unkn	outsider	unkn
insider	-2.1 (9.7e-02)	**4.2 (1e-04)**	1.2 (4.7e-01)	-2.2 (6.7e-02)
outsider		**6.8 (7.7e-10)**		**-3.4 (3.1e-03)**

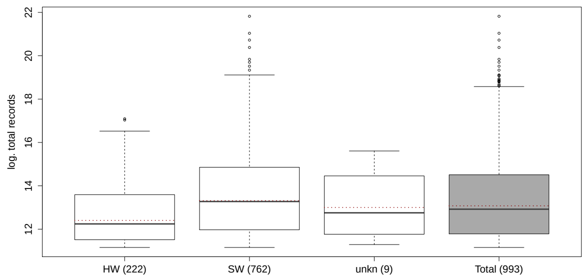


Figure 4.9: Boxplot for log-transformed *total records* vs. *medium* variable (taking the levels hardware (HW) if the data was lost through a hardware medium, software (SW) if it was lost through a software medium and unknown (*unkn*) if the information is not available) for the complete dataset (considering events with at least $70k$ records lost). On the far right the total is shown as well (in darkgray). The 20% symmetrically trimmed mean is shown in red (dashed line) and the number of observations per group is shown in brackets next to the group label. In particular for the hardware and software groups we observe notable differences with regards to the *total records* trimmed mean level and the span of the *total records* distribution function.

4.1.8 Severity vs. *medium*

In 20% of the cases the data was lost through a hardware medium (HW), whereby these losses are less severe than a breach via a software medium (SW) as visible in figure 4.9. This makes sense as with hardware media there generally exist more capacity limitations than with software media. For example, clouds can store much more data than a typical laptop and have therefore a much higher exposure. Welch's test statistic for the trimmed means and Windsorized variances clearly rejects the null hypothesis of equally trimmed means for the software and hardware group in Tukey's multiple pairwise comparison test with a p-value below 4.85e−10. We get the same result in the multiple pairwise comparison test with rank-ordered data with a p-value of 5.56e−11. For both kinds of tests there was no difference between the unknown group (*unkn*) and the other two groups detected. However, this is not surprising as it only contains very few observations.

4.1.9 Severity vs. *intentional*

While there are more than three times as many events which happen intentionally (group *yes*), there does not seem to be a difference in the breach severity of such events in comparison to unintended events (group *no*). In figure 4.10 we show the boxplot of the *intentional* variable. For unintentional and intentional data breach events the median, the interquartile range and the overall span are almost identical. For events where it remained unknown (group *unkn*) there is no clear relation to the breach severity variable visible besides showing a higher mean than the other two groups. However, this could also be due to the size. The same hypothesis tests on the trimmed means and ranked data as in the previous subsections show no statistical evidence for unequally trimmed means or different distribution functions among the three groups at a 95% confidence level with p-values above 50%.

4.1.10 Severity vs. *failure mode*

There are several modes of failure that make it possible for a data breach to occur. Considering figure 4.11, there are a few interesting things we can note. First of all, the highest median belongs to the *process* group where data breach events have occurred due to bad or non-existent security processes. Overall, these cases make up for 10% of the total. Secondly, a solid 20% of the cases are due to human error, whereby the *human* group shows a slightly lower median than the total distribution. Most events happened due to a software or hardware error (group *SW/HW*), whereby almost all of them are related to a software error. However, there is still a big proportion of events for which

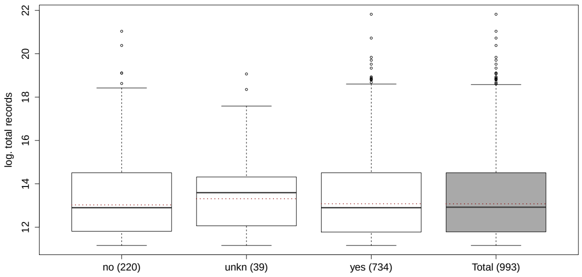


Figure 4.10: Boxplot for log-transformed *total records* vs. *intentional* variable (taking the levels *yes* if the data breach event happened on purpose and *no* if it happened unintentionally; there are as well cases for which this is unknown (*unkn*)) for the complete dataset (considering events with at least $70k$ records lost). On the far right the total is shown as well (in darkgray). The 20% symmetrically trimmed mean is shown in red (dashed line) and the number of observations per group is shown in brackets next to the group label. The intentional and unintentional groups are very similar both in their *total records* median level and their *total records distribution* function.

it is not clear what the mode of failure was (group *unkn*). Once more this group shows the lowest median and highlights again the information availability problem for less severe data breaches.

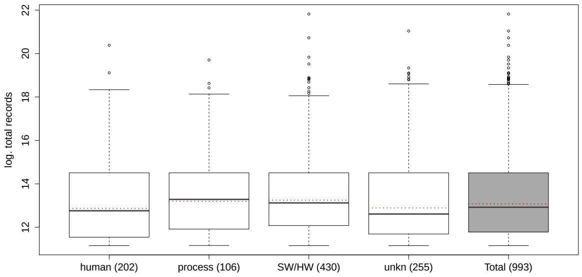


Figure 4.11: Boxplot for log-transformed *total records* vs. *failure mode* variable (taking the levels *human* if a human error let to the data breach, *process* if a process error made it possible for the data breach to happen, software/hardware (*SW/HW*) if an error in the used medium was the mode of failure and unknown (*unkn*) as this is information is not always known) for the complete dataset (considering events with at least $70k$ records lost. On the far right the total is shown as well (in darkgray). The 20% symmetrically trimmed mean is shown in red (dashed line) and the number of observations per group is shown in brackets next to the group label. Both the *total records* median levels and *total records* distribution function are very similar among all groups.

For this variable the hypothesis tests are not conclusive. While Welch's test does not reject the null hypothesis of equally trimmed means among the groups (p-value: 65.4%), Welch's test on the ranked data does reject the null hypothesis of having the same distribution function among the groups (p-value: 3.82%). However, when doing a multiple pairwise comparison on the ranked data none of the pairs reject the null hypothesis of having the same distribution function (at a 95% confidence level).

4.1.11 Severity vs. *third party*

In 15% of the cases a third party was involved in or responsible for the data breach event. What we observe from the boxplot in figure 4.12 is a higher median and interquartile range for events without a third party involvement (group *no*) than for events with (group *yes*) or events with no information (group *unkn*). Again for the unknown group we observe the lowest median and shortest span, which tells us that in particular events with few records lost belong to this category. Hence one might wonder whether events without a third party involved actually have a higher median or if this is a result of unavailable information for less severe breaches. On the other hand, for the group of observations where a third party was involved, we see an almost identical distribution as for the one of the total dataset.

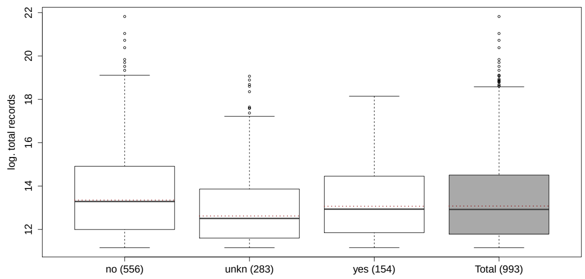


Figure 4.12: Boxplot for log-transformed *total records* vs. *third party* variable (taking the levels *no* if no third party was involved or partially responsible for the data breach to occur, *yes* if the contrary holds and unknown (*unkn*) if it is not known) for the complete dataset (considering events with at least $70k$ records lost). On the far right the total is shown as well (in darkgray). The 20% symmetrically trimmed mean is shown in red (dashed line) and the number of observations per group is shown in brackets next to the group label. Differences between the *total records* median levels and *total records* distribution functions among the three groups are visible but strong statistical evidence for the latter two being significantly different at a 95% confidence level is only present for the comparison between the no third party involved group (*no*) and the unknown group (*unkn*; see table 4.4).

Again using Welch's test statistic for the trimmed mean (20% symmetric) and Windsorized variances as well as Welch's test on ranked data, both null hypotheses of equally trimmed means and the same distribution functions among the groups are rejected with p-values below 1.12e−05. Doing a multiple pairwise comparison test using Tukey's procedure for the trimmed means rejects the null hypothesis for every pair (see table 4.4). The test on ranked data indicates that only *no* and *unkn* group follow a different distribution (see table 4.4).

Table 4.4: Test statistics of Tukey's multiple pairwise comparison for Welch's trimmed mean (20% symmetric) test with Windsorized variances (left) and for Welch's test on ranked data (right) for the *total records* variable grouped by the *third party* variable (taking the levels *no* if no third party was involved or partially responsible for the data breach to occur, *yes* if the contrary holds and unknown (*unkn*) if it is not known) for the complete dataset (considering events with at least $70k$ records lost) at a 95% confidence level, p-values are shown in brackets. Only for the comparison between the no third party involved group (*no*) and the unknown group (*unkn*) strong statistical evidence is present for the two groups having a significantly different *total records* trimmed mean and to follow different *total records* distribution functions.

	Welch's test on trimmed means		Welch's test on ranked data	
	unkn	yes	unkn	yes
no	**8.3 (0)**	**2.6 (2.9e-02)**	**-4.8 (5e-06)**	-1.7 (2.1e-01)
unkn		**-4.0 (2.2e-04)**		2.2 (8.1e-02)

4.1.12 Concluding remarks on individual predictors

From the previous subsections is becomes clear that some variables might be more important than others to describe the breach severity or still need some further investigation. Furthermore, there were some confounding effects detected due to the source of the individual observations, which has to be respected in further analyses. From the previous subsections we can summarize the following results:

- The severity of breached records above $70k$ has shown an overall increase over time, whereby we have observed different slopes for the different datasources. However, it is not entirely clear where most of this signal is coming from, as it could either be from the *bli* datasource, *PRC* or both of them.
- For the simplified country variable (with groups *US* and *non-US*) and the medium variable (with groups *HW*, *SW* and *unkn*) the groups clearly differ with regards to breach severity except for the *unkn* group. Entities headquartered outside of the

US and software media are related to more severe breaches than their counterparts, which are US-headquartered entities and hardware media respectively. Furthermore, if we test whether the two groups of the *country* variable follow the same distributions at various thresholds, there is only a limited range of threshold values for which this holds true (i.e. in between 1.44 and 2.6 million at a 95% confidence level).

- There was no statistical evidence found for a clear relationship between the *total records* variable and the size variables, whereby both market capitalization and number of employees were considered. Further possible questions to explore are whether additionally given the economic sector a relationship exists and if there exists a relationship for a higher threshold than $70k$.
- With regards to the *economic sector* variable the results are in line with what we have seen in the MDS section 2.1. Differences with respect to the total number of records breached were also observed for the more general *organization type* variable, whereby in particular the publicly traded companies group should be considered different from all the others.
- We do not believe that there was enough statistical evidence present for linking multiple firms events with a higher severity. Herefore hypothesis tests on the trimmed means (symmetric 20%), the ranked data and the empirical distributions were conducted. The null hypothesis of equally trimmed means among all groups was not rejected but the null hypothesis of the groups having the same distribution function based on ranks was rejected, whereby the corresponding p-value was close to the significance level. A two-sample Kolmogorov-Smirnov test did not reject the null hypothesis of the two groups having the same distribution function for most thresholds considered (p-value > 0.05). Furthermore, other variables have not suggested that there is a clear relationship between the size and the *total records* variable for data breach events with at least $70k$ items lost. The comparison was especially challenging due to different magnitudes of the two groups and therefore this question should again be revisited on a different dataset.
- From the severity vs. *insider/outsider* variable we can make two important observations. Firstly, breaches committed or mostly facilitated by an inside or outside party do not seem to be different with regards to breach severity, as no statistical evidence was found for the contrary (at a 95% confidence level). Secondly, in particular data breaches which are less severe belong to the group for which the information was not available. This suggests that the reporting of information is of less quality or quantity than for larger data breaches. This makes sense as large data breaches affect more people and therefore also get a broader media coverage.
- For the *failure mode* variable no clear indication with regards to the breach severity could be found. Hence this should again be reevaluated on another dataset. However, if we only consider the boxplot in figure 4.11, it looks like we are in the same situation as for the *insider/outsider* variable. Generally the groups don't differ from each other, except for the unknown group. The latter shows the lowest median and hints again at the aforementioned reporting issue for less severe data breaches.
- Further support for the reporting problem suspicion for less severe data breaches is given by the *third party* variable. Among the three groups (insider, outsider, unknown) considered, the unknown group shows again the lowest median and trimmed mean. A multiple pairwise comparison shows that for the trimmed mean test all three groups have a significantly different mean, while according to the rank-based test only the outsider group seems to have a different distribution function. The result from the first test is interesting as it suggests that if a third party was involved, the breaches are less severe and that larger breaches are more likely to be "home-made". However, this result is not supported by the second test and the

rejection is not of high magnitude. Therefore the influence of a third party involved on the severity for breaches above $70k$ items should again be considered on another dataset.

4.2 A multivariate model for *total records* based on *date* and *medium*

In the previous section we were able to identify trends and characteristics of the severity with regards to individual predictor variables. In particular we have seen that the severity is increasing over time and that for different types of breach media the distribution differs significantly at a 95% confidence level. Thus in the following we assess how the severity changes over time by medium type. Hereby we limit the analysis to the Privacy Rights Clearinghouse (*PRC*) subset, as we have observed different slopes in the increase by datasource in section 4.1.1 and because most of the data breaches which happened through a hardware medium originate from this source[2]. Moreover, for every event in the *PRC* subset it is known whether the data breach happened through a hardware medium (*HW*; 188 events) or through a software medium (*SW*; 332 events).

We consider the following three truncated gaussian regression models

- M1: *total records* $\sim$ *date*,
- M2: *total records* $\sim$ *date* and *medium*,
- M3: *total records* $\sim$ *date* and *medium* as a factor model.

In a truncated gaussian regression we estimate the conditional expectation $\mathbf{E}[y|y > u]$ based on the at u left-truncated observations y_i. The original distribution of the y_i's is hereby assumed to be gaussian [36]. The conditional expectation equals

$$\mathbf{E}[y_i|y_i > u] = x_i^T\beta + \sigma\frac{\varphi(\frac{x_i^T\beta-u}{\sigma})}{\Phi(\frac{x_i^T\beta-u}{\sigma})}, \tag{4.1}$$

whereby φ and Φ denote the standard normal probability density and cumulative distribution functions respectively. The coefficients β and the variance σ^2 of the original normal distribution $\mathcal{N}(x_i^T\beta, \sigma^2)$ are estimated via the maximum likelihood of the conditional density, which equals

$$f(y_i|y_i > u) = \frac{\varphi(\frac{y_i-x_i^T\beta}{\sigma})}{1-\Phi(\frac{u-x_i^T\beta}{\sigma})}. \tag{4.2}$$

Naturally, for a at u left-truncated random variable the expectation is higher than the one of the original gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ and its variance is smaller. The variance of a at u left-truncated $\mathcal{N}(\mu, \sigma^2)$ random variable Y equals

$$Var(Y|Y > u) = \sigma^2\big(1 - \frac{\varphi(\frac{x_i^T\beta-u}{\sigma})}{\Phi(\frac{x_i^T\beta-u}{\sigma})}(\frac{\varphi(\frac{x_i^T\beta-u}{\sigma})}{\Phi(\frac{x_i^T\beta-u}{\sigma})} - u)\big). \tag{4.3}$$

[2] To be precise, 85% of the cases which happened with a hardware medium originate from the *PRC* datasource.

4.2.1 Residual analysis

We use the truncreg package [10] to fit the truncated regression models[3]. The residual plots and a discussion thereof are included in the appendix B.2.1. In the individual residual plots we observe a separation of the residuals for the two groups. Furthermore we note that HW events are slightly overestimated in the model M1 and that the overestimation increases for higher fitted values.

4.2.2 Model comparison

In figure 4.13 we show the three different model fits. We observe an increase over time for both groups HW and SW in the models M1 and M2, whereby in the factor model M3 it is only visible for the SW group. This suggests that the observed increase of the severity over time originates from this subgroup. However, it is also evident that the HW group appears much less frequent which makes a comparison over time challenging. Considering the coefficient estimates in table 4.5 we note that for both models M1 and M2 the *date* variable is significantly contributing at a 95% confidence level, while according to model M3 the individual *date* coefficients for the HW and SW group do not differ significantly at a 95% confidence level. However, if the factor term is excluded there is sufficient statistical evidence for a different intercept for the two groups HW and SW in M2. Even when we test the model M2 on two random halves of the dataset[4], the null hypothesis of them having the same intercept is still rejected with p-values below 0.05.

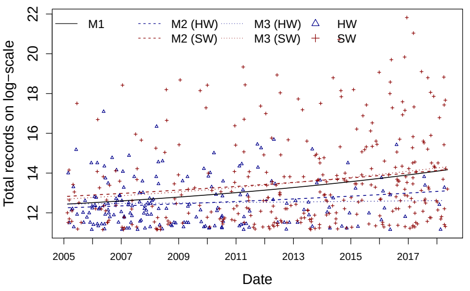


Figure 4.13: Truncated regression fits of the three models M1 (logarithmized *total records* ∼ *date*), M2 (logarithmized *total records* ∼ *date* and *medium*, whereby *medium* is a factor variable taking the two levels hardware (HW) and software (SW)) and M3 (logarithmized *total records* ∼ *date* and *medium* as a factor model) for the Privacy Rights Clearinghouse datasource (considering events with at least $70k$ records lost), whereby observations and fits of the hardware group (HW) are colored blue and observations and fits of the software group (SW) are colored red. The solid (black) line corresponds to M1, the dashed line to M2 and the dotted line to M3. Different levels with regards to the severity are clearly visible for the two groups hardware and software and in particular we observe a lower number of hardware events in more recent years.

The above observations are furthermore supported by the AIC, BIC and log-likelihood ratio test at a 95% confidence level (see table 4.6). Model M2 is clearly the preferred choice compared to model M1 based on the aforementioned measures. Even though the AIC value of M3 is slightly lower, the BIC is clearly higher in comparison to the one of M2. In particular is the hypothesis of the additional coefficient in M3 being equal to zero not rejected by the log-likelihood ratio test for the nested models M2 and M3. Hence there is not enough statistical evidence for the support of a different rate of increase over time for the two groups. However, this might be due to the lower number of observations in the HW group for more recent years.

[3] In all three models we have shifted the date variable to start at 0 for the first event in the time series. Furthermore, to run the algorithm smoothly for all three models we had to lower the threshold of $70k$ by 0.5 on the log-scale.

[4] Hereby we keep the ratio of the observations with a HW or SW medium at the same level as in the original dataset considered.

Table 4.5: Coefficient estimates (with standard errors and t-test p-values) of the three truncated regression models M1 (logarithmized *total records* ∼ *date*), M2 (logarithmized *total records* ∼ *date* and *medium*, whereby *medium* is a factor variable taking the two levels hardware (HW) and software (SW)) and M3 (logarithmized *total records* ∼ *date* and *medium* as a factor model) for the Privacy Rights Clearinghouse datasource (considering events with at least $70k$ records lost). In particular for the factor model M3 we observe large standard errors for all coefficient estimates and thus indicates that there is not enough statistical evidence for a different slope for the two groups hardware and software.

	M1			M2			M3		
	Estimate	Std. error	P-value	Estimate	Std. error	P-value	Estimate	Std. error	P-value
Intercept	7.3	1.3	5.7e-09	6.9	1.2	2.9e-08	7.9	1.3	2.9e-10
Date	0.0011	0.00023	2e-06	0.00070	0.00020	0.00052	0.00016	0.00040	0.70
MediumSW				2.5	0.69	0.00029	1.0	1.1	0.34
Date:MediumSW							0.00069	0.00046	0.13
Sigma	3.4	0.31	0	3.2	0.28	0	3.2	0.28	0

Table 4.6: Akaike information criterion (AIC) and Bayesian information criterion (BIC) values of the three truncated regression models M1 (logarithmized *total records* ∼ *date*), M2 (logarithmized *total records* ∼ *date* and *medium*, whereby *medium* is a factor variable taking the two levels hardware (HW) and software (SW)) and M3 (logarithmized *total records* ∼ *date* and *medium* as a factor model) for the Privacy Rights Clearinghouse datasource (considering events with at least $70k$ records lost). The log-likelihood ratio test statistics for the nested models (M1 vs. M2, and M2 vs. M3) is shown as well, whereby p-values are shown in brackets and the confidence level has been corrected for the family wise error rate using Bonferroni's method [35]. The latter test for the nested model comparison M2 vs. M3 does also not provide statistical support for the two groups hardware and software to have a different *date* coefficient at a 95% confidence level.

M1			M2			M3	
AIC	BIC	χ^2: M1 vs. M2	AIC	BIC	χ^2: M2 vs. M3	AIC	BIC
1'943.6	1'956.4	17.1(7.2e-05)	1'928.5	1'945.5	2.4(2.5e-01)	1'928.2	1'949.4

4.2.3 Conclusion

We conclude that there is in fact a significant increase over time (p-value < 0.001) of the severity of data breaches with at least $70k$ records lost for both medium groups. However, there is no statistical evidence for different rates of increase for the two groups. The best model M2 suggests a (massive) increase of the expected number of records lost by $e^{0.0007*365} - 1 = 0.29$ of the original gaussian distribution per year, whereby the SW breaches are more severe than the HW breaches as the former has a higher intercept. The increase of the truncated expected values of the model M2 are shown per year and for both medium types in table 4.7. The yearly change of the expected truncated records lost on the original scale starts at a different level for the two medium types, however for both of them the rate of increase has almost doubled over the last fourteen years.

Table 4.7: Expected number of *total records* lost by the truncated regression model M2 (logarithmized *total records* ∼ *date* and *medium*, whereby *medium* is a factor variable taking the two levels hardware (HW) and software (SW)) for the Privacy Rights Clearinghouse datasource (considering events with at least $70k$ records lost) for the first of January of the years 2005-2018 for both media hardware (HW) and software (SW). The estimates are shown on log-scale (LS) and on the original scale (OS), as well as a %-change per year for the original scale. For both groups the rate of increase on the original scale has almost doubled over the last fourteen years.

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
HW														
Records on LS	12.3	12.3	12.3	12.4	12.5	12.5	12.6	12.6	12.7	12.8	12.8	12.9	13.0	13.1
Records on OS	21 10^4	22 10^4	23 10^4	24 10^4	26 10^4	27 10^4	29 10^4	31 10^4	33 10^4	35 10^4	38 10^4	41 10^4	44 10^4	48 10^4
%-change		4.8%	5.1%	5.3%	5.6%	5.8%	6.1%	6.4%	6.8%	7.1%	7.5%	7.8%	8.2%	8.7%
SW														
Records on LS	12.8	12.9	13.0	13.1	13.1	13.2	13.3	13.4	13.5	13.6	13.8	13.9	14.0	14.1
Records on OS	37 10^4	40 10^4	43 10^4	47 10^4	51 10^4	56 10^4	61 10^4	68 10^4	75 10^4	84 10^4	94 10^4	106 10^4	120 10^4	137 10^4
%-change		7.7%	8.1%	8.6%	9%	9.5%	9.9%	10.4%	11%	11.5%	12.1%	12.7%	13.3%	13.9%

4.3 Density estimation

From the previous two sections we know there exist inhomogeneities among the different datasources. Evaluating everything with respect to three different datasources is quite an elaborate task. In the following we were not able to do so and thus the results based on the complete dataset have to be looked at with these reservations in mind.

4.3.1 For the complete dataset

As in [18] we fit a truncated lognormal, a Pareto and an upper-truncated Pareto distribution[5] [1] to the *total records* variable for three different thresholds ($70k$, $500k$, $1Mil$). The parameter estimates are shown in table 4.8 and the fits in figure 4.14. Considering the latter we see that the truncated lognormal and the upper-truncated Pareto yield a better fit for the tail in comparison to the Pareto distribution, as the latter is too heavy-tailed in the second half. For the Pareto distribution we have $\alpha < 1$ and hence an infinite mean. Moreover, the log-likelihood values are the highest for the truncated lognormal and upper-truncated Pareto model and almost identical for the latter two for all thresholds considered.

Table 4.8: Parameter estimates for the truncated lognormal, the Pareto and the upper-truncated Pareto distributions based on the complete dataset at various thresholds ($70k$, $500k$, $1Mil.$). The used threshold (u), the number of observations above u (n), the parameter estimates, their standard errors in brackets and the log-likelihood of the fits are given for each distribution function. For all thresholds the truncated lognormal and upper-truncated Pareto distribution give the best fit based on the log-likelihood values.

		Truncated lognormal			Pareto		Upper-truncated Pareto	
u	n	μ	σ^2	logL	α	logL	α	logL
70k	993	4.7(2.5)	24(0.37)	-15'200	0.43(0.014)	-15'210	0.40(0.015)	-15'198
500k	480	9.6(2.0)	13.8(0.3)	-8'062	0.50(0.023)	-8'070	0.45(0.025)	-8'063
1Mil.	364	8.8(3.2)	15(0.61)	-6'281	0.54(0.029)	-6'285	0.45(0.03)	-6'281

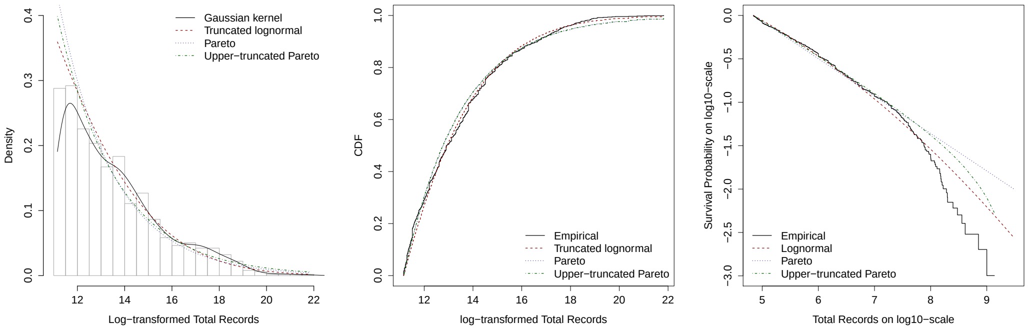


Figure 4.14: Probability density plot (left), cumulative density plot (middle) and survival probability plot on log10-scale (right) for the estimated distributions truncated lognormal (red dashed), Pareto (blue dotted) and upper-truncated Pareto (green dashdotted) for the complete dataset (i.e. with a threshold of $u = 70k$). In the density plot we used a Gaussian kernel estimate for the density of the observations (black solid line). In the other two plots we show the empirical cumulative density function (black solid line). The tail of the empirical distribution function is best approximated by the truncated lognormal distribution.

We now compare the fits by considering several goodness-of-fit measures (see table 4.9) [38]. We observe that generally the truncated lognormal and upper-truncated Pareto are valid choices. Except for the threshold $u = 500k$, the Pareto distribution is always rejected at 95% confidence level from both the Kolmogorov-Smirnov (KS) and Anderson-Darling (AS) test[6] [38]. Also with regards to AIC and BIC Pareto is the least favorable choice. For $u = 70k$ both the lognormal and upper-truncate Pareto distributions are rejected by

[5] We assume that the upper limit is known and set it to the largest observed value, which is 3×10^9.

[6] In both tests we test whether the true distribution the data originates from (denoted by F) is the

the KS test but not by the AD test, hence both of them appear to be valid choices. For $u = 500k$ only the lognormal distribution is not rejected by both hypothesis tests and is thus the preferred choice (although there is only a small difference in comparison to the upper-truncated Pareto distribution). For $u = 1Mil.$ the upper-truncated Pareto is the only distribution that is not rejected by both hypothesis tests and thus the preferred one.

Table 4.9: Goodness of fit measures (D_n = Kolmogorov-Smirnov test statistic, A_n = Anderson-Darling test statistic, Akaike information criterion (AIC) and Bayesian information criterion (BIC)) for the truncated lognormal, Pareto and upper-truncated Pareto distributions for the thresholds u = $70k$, $500k$ and $1Mil.$ based on the complete dataset (for D_n and A_n p-values are provided in brackets). With regards to the given measures either the truncated lognormal or the upper-truncated Pareto distribution provide a suitable fit.

	Truncated lognormal	Pareto	Upper-truncated Pareto
u = 70k			
D_n	0.058 (0.0024)	0.055 (0.0044)	0.046 (0.030)
A_n	2.1 (0.077)	4.7 (0.0040)	2.1 (0.077)
AIC	30'401.3	30'421.2	30'398.6
BIC	30'411.1	30'426.1	30'403.5
u = 500k			
D_n	0.051 (0.16)	0.051 (0.152)	0.033 (0.66)
A_n	1.7 (0.13)	5.5 (0.0017)	2.7 (0.041)
AIC	16'128.4	16'142.2	16'128.4
BIC	16'136.8	16'146.4	16'132.6
u = 1Mil.			
D_n	0.080 (0.019)	0.080 (0.019)	0.080 (0.019)
A_n	3.5 (0.015)	7.2 (0.00026)	2.42 (0.055)
AIC	12'566.6	12'572.9	12'564.9
BIC	12'574.4	12'576.8	12'568.8

4.3.2 For the economic sectors

From both the MDS section 2.1 as well as the previous section 4.1 it becomes clear that the economic sectors follow different severity distributions. While for the complete dataset a truncated log-normal or upper-truncated Pareto give a good fit for various thresholds, we assess in the following whether this is as well the case for the severity distribution of the individual economic sectors for the original threshold $u = 70k$. We summarize the economic sectors with a small number of observations in the miscellaneous sector *other*.

In figure 4.15 we show the different fits for the individual sectors and in table 4.10 the parameter estimates of the distributions. By considering the density fits it becomes obvious that not all distributions are a suitable fit for a specific sector and for the considered threshold. For many sectors there is a slight or even pronounced mode for the lower range of the *total records* variable, which is better captured by a truncated log-normal distribution rather than a Pareto or upper-truncated Pareto distribution. At this point it is also important to note that some of the provided *total records* have most likely been rounded to the nearest power of ten or an integer multiple thereof, as they appear more often than other random natural numbers. This explains some of the observed spikes in the histograms. A good example for this is the plot of the consumer cyclical sector (*53*), where there is a clear spike before e^{14}, which is about 1.2 million. Hence many events with records close to 1 million have probably been rounded. In fact, for the sector *53* the value 1 million appears most often with seven times, followed by 5 other values which appear three times. This is observation is important since it causes ties and we use the KS test to assess whether the observations originate from one of our assumed distributions, whereby the used test statistic D_n is given by the maximal difference of the empirical distribution $\hat{F}_n(\cdot)$ to the assumed one F_0, i.e $D_n = \sup_y |\hat{F}_n(y) - F_0(y)|$ [38].

same as our assumed distribution F_0, i.e. $H_0 : F = F_0$. Hereby we use the empirical distribution $\hat{F}_n$ as an estimate for F.

Table 4.10: Parameter estimates for the truncated lognormal, the Pareto and the upper-truncated Pareto distributions for the economic sectors (with factor levels industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and the miscellaneous sector *other*, which is a merger of the sectors energy (*50*), basic materials (*51*), utilities (*59*), military (*mil*) and politics (*pol*)) at the threshold u=70k based on the complete dataset. The number of observations per sector is given by n, the parameter estimates, their standard errors in brackets and the log-likelihood of the fits are provided for each distribution. For all economic sectors either the truncated lognormal or upper-truncated Pareto distribution give the best fit based on the log-likelihood value.

		Truncated lognormal			Pareto		Upper-truncated Pareto	
Sector	n	μ	σ^2	logL	α	logL	α	logL
52	164	12.6(0.38)	5.9(0.062)	-2’498.5	0.34(0.027)	-2’499.9	0.24(0.034)	-2’494.1
53	138	13(0.40)	6.7(0.075)	-2’173.3	0.30(0.026)	-2’177.7	0.23(0.032)	-2’173.9
54	33	12.5(0.80)	5.2(0.27)	-495.2	0.36(0.067)	-495.3	0.20(0.090)	-493.3
55	125	12.4(0.39)	4.7(0.064)	-1’851	0.38(0.034)	-1’850.7	0.29(0.042)	-1’848.4
56	145	12.4(0.22)	2.8(0.025)	-2’080.7	0.43(0.036)	-2’082	0.34(0.043)	-2’080.4
57	203	14.1(0.27)	7.4(0.042)	-3’391.4	0.24(0.017)	-3’411.2	0.15(0.023)	-3’398.7
58	38	13.4(0.57)	5.5(0.18)	-603.1	0.29(0.050)	-605.7	0.12(0.070)	-601.5
edu	87	11.5(0.40)	2.4(0.057)	-1’163.8	0.58(0.064)	-1’158.5	0.49(0.074)	-1’158.9
other	49	11.1(1.9)	12(0.83)	-754.8	0.33(0.049)	-753.7	0.22(0.062)	-751.6

With the above considerations and the goodness of fit measures presented in table 4.11, we make the following observations for the individual economic sectors.

- For the industrial sector (*52*) sector we observe a slight mode of the density at the lower spectrum of the *total records* variable. Firstly, almost a third of the observations are below 200k. Secondly, the values 100k and 200k appear most frequent (six times) and cause two spikes slightly before e^{12} and slightly afterwards, and thus lead to the observed mode in the density. Thus it is not surprising that both the Pareto and upper-truncated Pareto distributions are rejected by both the KS and AD test (p-value < 0.005). The KS test is rejected for the truncated lognormal distribution with a p-value of 3.8% whereas the AD test is clearly not. Thus the truncated lognormal distribution is a reasonable choice for the considered threshold. Presumably, the upper-truncated Pareto distribution gives a good fit for a higher threshold, as there will no longer be a mode in the density and because already now both AIC and BIC are the lowest for this distribution.

- For the consumer cyclical sector (*53*) we observe a more pronounced mode for the lower *total records* range than for the industrial sector (*52*). In particular there are three spikes visible, whereby the first one is due to multiple appearances of the high ten thousands (e.g. 70k, 80k, 95k). The second one includes values around 1 million, which in fact appears seven times and is by far the most frequent one. The third one is given by several events with 40 up to 60 millions of records lost. Again, both the Pareto and upper-truncated Pareto distributions are rejected by the KS and AD tests (p-value < 0.005). The truncated lognormal distribution is neither rejected by the KS nor the AD test (p-value > 0.1). Moreover, it scores well with regards to AIC and thus gives the best fit.

- For the consumer non-cyclical sector (*54*) we have the lowest number of observations (33). Thus there are some sparse areas in the histogram towards the upper end. For the lower range of the *total records* variable we observe again a slight mode and three spikes. There are four events in between 70k and 80k which cause the first spike, whereas the second is caused by three events above 300k and the third by three data breaches with 1.5 up to 1.7 million records lost. Again the truncated lognormal distribution gives a reasonable fit, however this time the upper-truncated Pareto distribution is a valid option as well, as it is not rejected by the KS and the AD test (p-values > 0.19) and scores best according to both AIC and BIC.

- For the financial sector (*55*) the density shows a mode at the lower range of the *total records* values. We observe two spikes, one at 100k and the other around 1

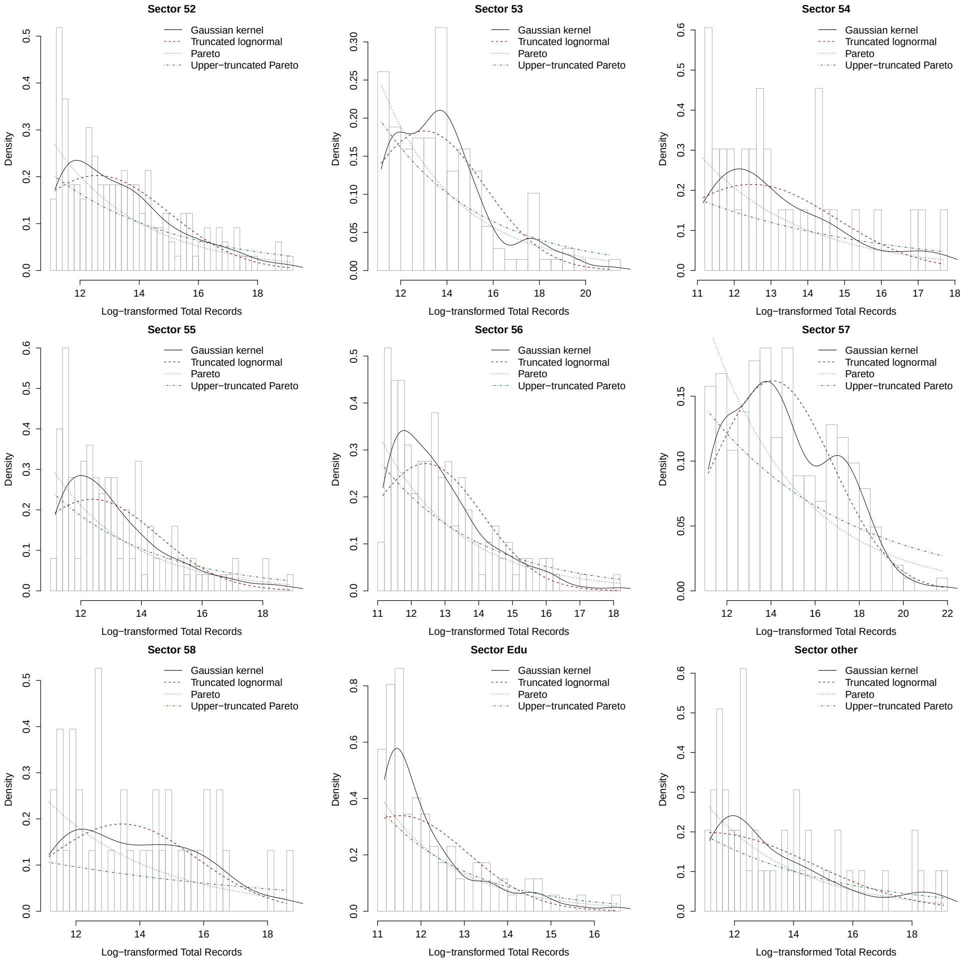


Figure 4.15: Probability density plots per economic sector (industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and the miscellaneous sector *other*, which is a merger of the sectors energy (*50*), basic materials (*51*), utilities (*59*), military (*mil*) and politics (*pol*)) including the estimated distributions lognormal truncated (red dashed), Pareto (blue dotted) and upper-truncated Pareto (green dashdotted) for the threshold $u = 70k$ based on the complete dataset. In each plot a Gaussian kernel estimate was added (black solid line). From left to right and top to bottom we show the economic sectors industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and the miscellaneous sector *other*. While for most economic sectors either the truncated lognormal or the upper-truncated Pareto distribution provide a reasonable fit, for some not a single one of the considered distributions fits well and a larger class of distribution functions should be considered.

million. The KS test rejects all distributions (p-values < 0.05), whereby the p-value of the truncated lognormal distribution is close to the significance level of 5%. The AD test rejects as well the Pareto and upper-truncated Pareto distributions. The truncated lognormal on the other hand is not rejected at a 95% confidence level, but the p-value of the test statistic is only slightly above the 5% significance level. Thus from the distributions we consider for the threshold $70k$ the truncated lognormal appears to be the most suitable choice. However, as for the industrial sector (*52*) it is very likely that the other two give a reasonable fit for a higher threshold when there is no longer a mode in the density.

- For the healthcare sector (*56*) all distributions are rejected by both the KS and AD test at a 95% confidence level. If we consider a higher confidence level, i.e. 99%, then the only distribution that becomes acceptable is the truncated lognormal. Considering the density plot in figure 4.15 we observe a profound mode for very low values of the *total records* variable which is not well-captured by the truncated lognormal. Thus for this sector and the considered threshold it is advisable to consider as well other distributions, such as the truncated Gamma distribution.

Table 4.11: Goodness of fit measures (D_n = Kolmogorov-Smirnov test statistic, A_n = Anderson-Darling test statistic, Akaike information criterion (AIC) and Bayesian information criterion (BIC)) for the truncated lognormal, Pareto and upper-truncated Pareto distributions for the individual economic sectors (industrials (*52*), consumer cyclicals (*53*), consumer non-cyclicals (*54*), financials (*55*), healthcare (*56*), technology (*57*), telecommunication services (*58*), education (*edu*) and the miscellaneous sector *other*, which is a merger of the sectors energy (*50*), basic materials (*51*), utilities (*59*), military (*mil*) and politics (*pol*)) for the thresholds u = $70k$ based on the complete dataset (for D_n and A_n p-values are provided in brackets). For most economic sectors either the truncated lognormal or the upper-truncated Pareto distribution provide a reasonable fit but for some sectors all considered distribution functions are rejected by the Anderson-Darling and the Kolmogorov-Smirnov test at a 95% confidence level and thus a larger class of distribution functions should be considered.

	Truncated lognormal	Pareto	Upper-truncated Pareto
Sector 52			
D_n	0.11 (0.039)	0.21 (6.4e-07)	0.18 (7.1e-05)
A_n	1.8 (0.12)	9.9 (1.2e-05)	5.6 (0.0014)
AIC	5'000.9	5'001.7	4'990.3
BIC	5'007.1	5'004.8	4'993.4
Sector 53			
D_n	0.089 (0.22)	0.19 (6e-05)	0.16 (0.001)
A_n	1.7 (0.14)	9.8 (1.3e-05)	6.71 (0.00046)
AIC	4'350.5	4'357.3	4'349.9
BIC	4'356.4	4'360.3	4'352.8
Sector 54			
D_n	0.13 (0.63)	0.24 (0.04)	0.18 (0.19)
A_n	0.74 (0.53)	2.3 (0.067)	1.3 (0.22)
AIC	994.3	992.5	988.6
BIC	997.3	994	990.1
Sector 55			
D_n	0.12 (0.043)	0.23 (1.5e-06)	0.20 (5.5e-05)
A_n	2.4 (0.06)	9.3 (2.6e-05)	6.6 (0.00053)
AIC	3'705.9	3'703.3	3'698.7
BIC	3'711.6	3'706.2	3'701.5
Sector 56			
D_n	0.12 (0.031)	0.26 (4.8e-09)	0.23 (4.5e-07)
A_n	2.6 (0.044)	13 (4.2e-06)	9.9 (1.3e-05)
AIC	4'165.3	4'166	4'162.8
BIC	4'171.3	4'168.9	4'165.8
Sector 57			
D_n	0.059 (0.47)	0.19 (7.8e-07)	0.12 (0.0037)
A_n	1.1 (0.32)	15 (3e-06)	7.9 (0.00013)
AIC	6'786.7	6'824.4	6'799.5
BIC	6'793.4	6'827.8	6'802.8
Sector 58			
D_n	0.10 (0.78)	0.20 (0.091)	0.14 (0.44)
A_n	0.48 (0.77)	2.7 (0.04)	1.1 (0.29)
AIC	1'210.1	1'213.3	1'205.1
BIC	1'213.4	1'214.9	1'206.7
Sector edu			
D_n	0.21 (0.00064)	0.33 (3.8e-09)	0.30 (1.2e-07)
A_n	4.1 (0.0082)	8.3 (8.5e-05)	6.9 (0.00036)
AIC	2'331.7	2'319	2'319.8
BIC	2'336.6	2'321.5	2'322.3
Sector other			
D_n	0.14 (0.26)	0.21 (0.023)	0.17 (0.11)
A_n	1.2 (0.25)	2.3 (0.064)	1.6 (0.15)
AIC	1'513.5	1'509.5	1'505.1
BIC	1'517.3	1'511.3	1'507

Again, it is very likely that for a higher threshold we get a good estimate by one of the three originally considered distributions.

- Considering the density plot of the technology sector (*57*) it becomes evident that one might need two different distributions to model the *total records* for the considered threshold as there are two clearly distinguished modes; the first one around 1 million and the second at 20 million. A compromise fit is given by the truncated lognormal, which is not rejected by any of the hypothesis tests (p-values > 0.3) and also gives the best fit according to AIC and BIC.

- For the telecommunication sector (*58*) we again have only a small amount of observations for the range considered (38). There is one spike around $300k$ records lost, a slightly higher number of events at the beginning and a sparser region for the upper range of the *total records* variable. For this sector both the truncated lognormal and upper-truncated Pareto distribution should be considered as possible distributions based on the goodness of fit measures in table 4.11. However, the density is much better captured by the truncated lognormal distribution and is thus the preferred one.

- For the education sector (*edu*) we observe a similar situation as for the healthcare sector (*56*). The density shows a clear mode for the lowest range of the *total records* variable, which is not well captured by either of the distributions. Hence it is not surprising that all of them are rejected by the KS and AD tests (p-values < 0.01). Thus for the considered threshold other distributions should be considered, which might yield a better fit.

- The miscellaneous sector (*other*) shows a slight mode in the density along the lower range of the *Total Records* variable, where we also observe two spikes. The first one around $100k$ and the second around $200k$. As the mode is only slightly visible, both the truncated log-normal and upper-truncated Pareto give a reasonable fit based on the available goodness of fit measures. Furthermore, both of them are not rejected by both the KS and AD tests (p-values > 0.1).

4.3.3 Conclusion

With regards to the distribution of the *total records* variable there are three important take-away messages. Firstly, if we consider the complete dataset at different thresholds, either the truncated lognormal or the upper-truncated Pareto distribution give a reasonable fit. Secondly, from section 4.3.2 it becomes evident that for some economic sectors other distributions might be more suitable. Thus a larger class of distributions should be considered and compared to the current ones. It is important to note that the fit of the distributions depends directly on the threshold. Therefore individual thresholds for the economic sectors should be introduced. Thirdly, for most of the economic sectors the truncated lognormal provides a reasonable fit or the best among the three distributions considered. For others the upper-truncated Pareto is as well a sensible choice. For the considered threshold the Pareto distribution was however outperformed for each sector by one of the others. This is not necessarily the case for higher thresholds.

Furthermore, in section 4.1.1 we observed an increase over time trend for the *total records* variable. In the density estimation section we did not account for this but this should be considered if one wishes to make predictions. Therefore it is a crucial next step to identify from which source(s) the time trend originate(s) from in order to be able to make the corresponding adjustments in the density estimation.

5 Reporting Delay

In the following chapter we would like to investigate if there have been any changes in the reporting duration of data breach incidents over the past couple of years. There are several reasons why a change in the reporting of such events might occur. Changes in the regulatory framework can lead to a faster reporting of data breaches as delay or non-reporting will result in heavy fines [22]. On the other hand, the ongoing coalescence of the physical and the technological world leads to an increased awareness of the risks associated to the new technology and a more responsible use.

5.1 Description of the dataset

The dataset for the analysis is taken from the website “Have I Been Pwnd” (HIBP) [19], which is run by Troy Hunt. He is a well-known cyber security expert and collects data from known breaches on his website to enable the general public to check whether or not they are affected by the breach. The dataset contains 285 events, whereby the oldest breach stems from July 2007 and the most recent from May 2018. The dataset has the following features: *title, name, domain, breach date, added date, modified date, pwn count, description, is verified, is fabricated, is sensitive, is active, is retired, is spam list, data classes*. We use this for a first analysis on the reporting delay but we believe the dataset has the following two limitations. Firstly its size, as it is rather small. Secondly, the sample might not be an accurate representation of the complete spectrum of data breaches as events are included based on a single person’s selection.

Breach date*, *added date* and *time difference

The dataset contains both a proxy date for when the breach happened (*breach date*) and when it became known to the public (*added date*). We will use the difference between these two dates as an estimate for the reporting delay. From the original 285 observations in the dataset only 223 could be used in the analysis. Some events had to be excluded as their breach date lies before the time Troy Hunt started his database at the end of 2013 [19]. If the breach date lies before that, the *added date* variable is no longer a good proxy for when the breach became known to the public and we have therefore only considered events with breach date from 2014 onwards[1].

We are interested in the reporting delay which is given by the time lag between the variables *breach date* and *added date*. Therefore we will first consider the histograms of these two variables. On the left in figure 5.1 we show the overlaid histograms for both variables for half-yearly buckets and we observe a clear time shift between the occurrence and the reporting frequency. To be more precise, we look at the reporting delay of the respective events ordered by their breach date, which is shown on the right in figure 5.1. The reporting delay has a natural upper bound, which is given by the difference between the current date and the breach date. Since this data was downloaded on the 2018-06-05, it contains all the information that was entered up to the previous day, which will be used

[1] One event had to be excluded as the added date was before the breach date, which is generally not the case.

as a reference date. We should also keep in mind that this figure might not show the complete picture as some events might have not been reported yet.

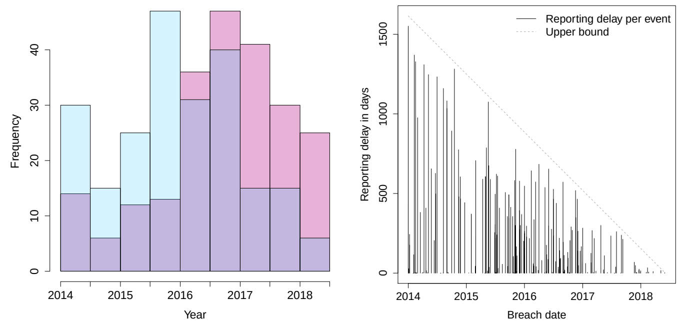


Figure 5.1: Overlaid histograms of the *breach date* (blue) and *added date* (pink) variables with half yearly buckets (left) and reporting delay in days ordered by *breach date* (right), whereby an upper bound is shown in gray with reference date 2018-06-04, for the reduced “Have I Been Pwnd” dataset, which only includes data breach events with *breach date* from 2014 onwards. Between the *breach date* and *added date* variables a clear time shift is visible.

We introduce a new variable which is called *time difference* and specifies the number of days between the *breach date* and *added date* variable. This will be our response variable and from now on we exclude the *added date* variable, as we otherwise have a dataset with linearly dependent variables. In figure 5.2 we show the histogram of the *time difference* variable, which is right-skewed[2].

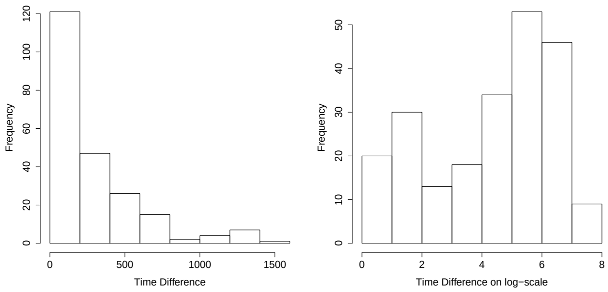


Figure 5.2: Left: Histogram of the new response variable *time difference*, which is given by the difference of the *breach date* and *added date* variable for the reduced “Have I Been Pwnd” dataset, whereby only events with *breach date* from 2014 are included. Right: The log-transformed version of the *time difference* variable, whereby the latter was shifted by 1 day for the transformation. The median of the *time difference* variable equals 139 days.

These plots motivate us to further analyze if there is statistical evidence for this visually observed time shift in the histograms and in the latter case, how we can quantify it. Before we look more closely at the other variables.

Other variables in the dataset

The five variables *is verified, is fabricated, is sensitive, is retired, is spam list* are booleans and a detailed description of them is provided in the appendix (see table A.9). The *pwn count* variable specifies the number of records lost and the *data classes* variable specifies the kind of information that was lost in the breach. The latter takes multiple values per observation and the list of the original 110 possible attributes[3] is shown in the appendix in table A.11. For the analysis we only consider attributes that appear at least ten times

[2] For the log-transformation we have shifted the *time difference* by one day, i.e. *log(time difference) = log(time difference + 1)*.

[3] Some of them refer to the same information and have thus been merged to one information attribute. The used mapping is shown in table A.10 in the appendix.

in order to get a reasonable parameter estimation. Hence the remaining variables of the dataset are:

- 5 pre-set booelans (number of “TRUE” values are shown in brackets): *is verified* (206), *is fabricated* (1), *is sensitive* (20), *is retired* (1) and *is spam list* (7),

- 1 continuous variable: *pwn count*, which is right-skewed and we thus use its log-transformed version (cf. figure 5.3),

- 13 boolean variables based on the former *data classes* variable, shown with the number of “TRUE” values in table 5.1.

Table 5.1: Information attributes of the *data classes* variable, which specify the types of information that were lost in the breach, and their corresponding frequencies in the reduced “Have I Been Pwnd” dataset, whereby only events with *breach date* from 2014 are considered. Most of the information attributes occur only seldomly.

Attribute	Frequency	Attribute	Frequency	Attribute	Frequency
Email addresses	218	Names	58	Geographic locations	21
Passwords	185	Website activity	50	Chat logs	15
Usernames	147	Phone numbers	35	Job titles	11
IP addresses	108	Physical addresses	35		
Dates of birth	62	Genders	33		

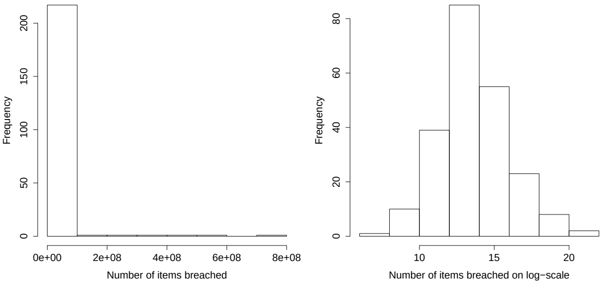


Figure 5.3: Histogram of the *pwn count* variable, which specifies the number of records lost, on the original scale (left) and log-transformed (right) for the reduced “Have I Been Pwnd” dataset, whereby only events with *breach date* from 2014 are considered. This variable is clearly right-skewed.

We check for multicollinearity among all predictor variables. To this end we look at the variance inflation factor, which measures how much of the variance of an ordinary least square (OLS) coefficient is induced by multicollinearity. If this score is above 5 we might have some multicollinearity and if it is larger than 10, the multicollinearity is assumed to be very strong [11].

Table 5.2: Variance inflation factors for the remaining predictor variables in the reduced “Have I Been Pwnd” dataset, whereby only events with *breach date* from 2014 are considered. The remaining predictor variables are *breach date*, *pwn count* (specifies the number of records lost) and the boolean information attributes variables, specifying whether a certain type of information was present in the data breach event or not, as well as the pre-defined boolean variables which provide further information about the data breach event.

BreachDate	1.1	Geographic.locations	1.5	IsSpamList	1.5	Phone.numbers	3
Chat.logs	1.3	IP.addresses	1.2	IsVerified	1.4	Physical.addresses	2.7
Dates.of.birth	1.7	IsFabricated	1.2	Job.titles	1.4	PwnCount	1.5
Email.addresses	1.2	IsRetired	1.1	Names	2.6	Usernames	1.6
Genders	1.8	IsSensitive	1.1	Passwords	1.7	Website.activity	1.2

Since all of the variance inflation factors are below 5 we do not have evidence that the different predictor variables are linearly correlated. Hence we do not exclude any variables.

5.2 Model fitting

In order to answer our question about the change of the reporting duration of data breaches over time we consider several types of models in the following subsections.

5.2.1 Generalized linear model fit

We start by fitting a generalized linear model (GLM) with identity-link function and with the log-transformed *time difference* variable as response variable to the dataset. As predictor variables we use all variables mentioned in the previous section and the *breach date* variable[4]. The diagnostic plots do not show any peculiarities for the full model and based on the summary output (not shown) we conclude that the variables *breach date*, *pwn count*, *website activity*, *geographic location* as well as *IP addresses* turn out to be the significant ones (at a confidence level of 95%). As a next step we try to reduce our model to the relevant variables in order to identify the most important ones. An all subset regression suggests to use the variables *breach date*, *pwn count*, *IP addresses*, *geographic locations* and *physical addresses*. Therefore, most of the predictor variables that were significant in the full model remain significant. The only exception is the *website activity* variable which dropped out. On the other hand, the variable *physical addresses* entered into the model. The coefficient estimates of the reduced model are shown in table 5.3.

Table 5.3: Coefficient estimates (with standard errors, t-test statistic and p-value of the t-test) of the reduced generalized linear model fit obtained by an all subset regression on the reduced "Have I Been Pwnd" dataset, whereby only events with *breach date* from 2014 are considered and the predictor variables *breach date*, *pwn count* (which specifies the number of records lost), five pre-defined booleans which provide further information about the data breach and thirteen booleans which specify the kind of information that was lost in the data breach event were available in the reduced dataset. The null deviance equals 961.13 on 222 degrees of freedom and the residual deviance 774.3 on 217 degrees of freedom, Akaike information criterion equals 924.43 and the coefficient of determination $R^2 = 19.44\%$.

	Estimate	Std. error	t value	$\Pr(>\lvert t \rvert)$
Intercept	1.5	0.8	1.9	5.9e-02
BreachDate	-0.00095	0.00032	-2.9	3.7e-03
PwnCount	0.24	0.057	4.2	3.4e-05
IP.addressesTRUE	0.99	0.26	3.9	1.5e-04
Physical.addressesTRUE	-0.97	0.35	-2.7	6.5e-03
Geographic.locationsTRUE	-1.4	0.44	-3.1	2.1e-03

Considering the estimates in table 5.3, we can make several interesting observations. Firstly, the *breach date* variable shows a negative coefficient of small magnitude. Secondly, the *pwn count* variable shows a large coefficient with a positive sign, which suggests that the larger the breach the longer the reporting delay will be. Moreover, the intercept shows a relatively large standard error. The interpretation of the information attributes is not straight forward due to the different signs of the coefficients. While the presence of the *IP addresses* has a prolonging effect, the presence of the other two shorten the delay. Additionally, the magnitude of the information attribute coefficients is largest for *geographic locations* which is only present in 10% of the observations and almost as large in magnitude as the intercept. The coefficients of the *IP addresses* and *physical addresses* variable are almost identical in size but of opposite sign. Hence if both of them are present, they almost annihilate the effect of one another on the reporting delay.

If we fit the model on a random half of the observations and compare it to the fit on the second half, the only variable that remains significant at a 95% confidence level on both fits is the *pwn count* variable and hence questions the validity of the others. This concern is further supported by the low R^2 value of roughly 20%. If we combine the information attributes into a single variable, which takes the value "TRUE" as soon as

[4] For the modelling we have shifted the *breach date* variable to start at 0 for 2014-01-01 and use as well the log-transformed *pwn count* variable.

any of them is present, it turns out that this combined information variable is clearly not significantly adding to the reporting delay. Hence the information attributes might add more noise than signal. For the latter model the *breach date* variable remains borderline non-significant at a 95% confidence level, even when fitting it on two subsets. Thus it remains questionable whether there exists a systematic relation between the reporting delay and the *breach date*.

The residual plots of the reduced model presented in table 5.3 show some peculiarities in the scale-location plot (not shown). A constant level of variance is observed for the lower part of the fitted values and a decreasing variance for the second part of the fitted values. This indicates that a GLM is not the best model choice for our data and we therefore explore other options. Hereby we use the reduced model as a starting point and check if we reach the same conclusion for the information attributes. Nevertheless we show the fit of the reduced model in figure 5.4, which exhibits a slight decreasing trend over time. One might wonder if this is the same for all quantiles of the reporting delay and we therefore fit a quantile regression to the dataset in the next section.

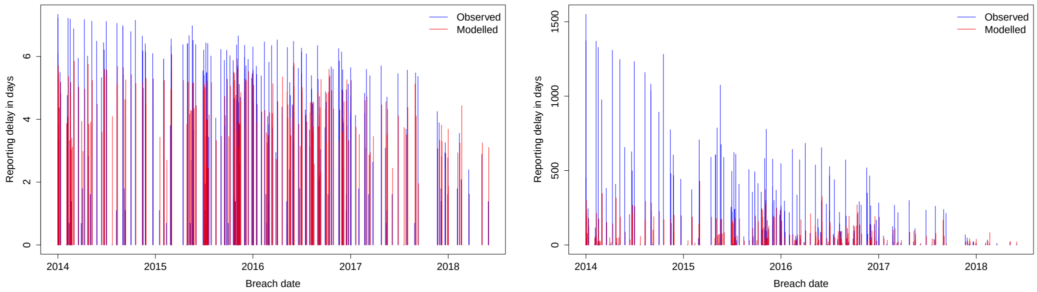


Figure 5.4: Modelled (red) vs. observed (blue) reporting delays on log scale (left) and on the original scale (right) for the reduced generalized linear model obtained by an all subset regression on the reduced "Have I Been Pwnd" dataset, whereby only events with *breach date* from 2014 are considered. The remaining predictor variables are *breach date*, *pwn count* (which specifies the number of records lost) and the three information attributes variables, *IP addresses*, *physical addresses* and *geographic locations*, which take the value *"TRUE"* when the named information attribute is present in a data breach event. A slightly decreasing trend over time is visible.

5.2.2 Quantile regression

We will perform a quantile regression [24] on the reduced model obtained in section 5.2.1. This method enables us to assess how the parameters for the different predictor variables change for different quantiles of the response. Recall that in quantile regression we assign asymmetric weights to the absolute error terms, whereby the weights depend on the quantile $\tau \in [0, 1]$. The coefficients of a τ-quantile regression solve the following minimization problem

$$\beta_\tau = \arg\min_\beta E\big[\tau|Y - X\beta|\mathcal{I}\{Y \geqslant X\beta\} + (1-\tau)|Y - X\beta|\mathcal{I}\{Y < X\beta\}\big]. \tag{5.1}$$

Therefore we get estimates of the quantile coefficients by solving

$$\hat{\beta}_\tau = \arg\min_\beta \sum_{i=1}^{N} \tau|y_i - x_i\beta|\mathcal{I}\{y_i \geqslant x_i\beta\} + (1-\tau)|y_i - x_i\beta|\mathcal{I}\{y_i < x_i\beta\}\big], \tag{5.2}$$

whereby x_i denotes the i^{th} row vector of the matrix X, which contains the values of the predictor variables for the i^{th} observation y_i.

We fit a quantile regression for all deciles. By retesting the fit on random halves of the subsets we reach the same conclusion as in section 5.2.1. In particular for the information attributes zero is mostly included in the 90% confidence interval of the quantile estimates or very close to the bounds thereof. Hence we exclude them from the model and show the quantile coefficient estimates for the intercept and the remaining two variables in figure 5.5 (Due to the exclusion of the information attributes the latter two automatically take over more importance in the model and thus show smaller p-values.)

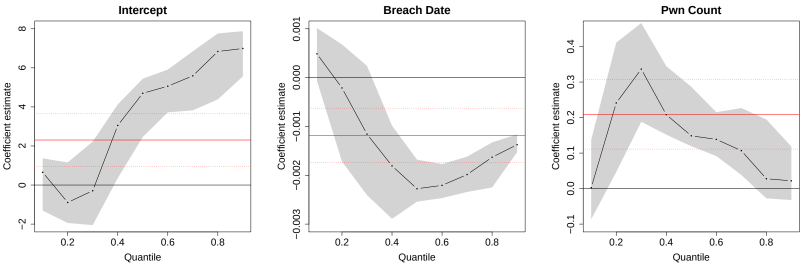


Figure 5.5: Quantile regression parameter estimates for the modelling of the log-transformed *time difference* variable with 90% confidence region (gray area) for the intercept (left), *breach date* (middle) and *pwn count*, which specifies the number of records lost, (right) per decile for the reduced "Have I Been Pwnd" dataset, whereby only events with *breach date* from 2014 are considered. The red line denotes the ordinary least square estimate and the dashed line the corresponding 90% confidence interval. The coefficients of determination R^2_τ of the quantile regressions[a] lie between $2\% - 16\%$.

[a] For quantile regression there also exists a coefficient of determination R^2_τ goodness of fit measure which is similar to the coefficient of determination R^2 measure from ordinary least squares. Contrary the latter, the R^2_τ from quantile regression focuses on a local goodness of fit for a specific quantile as it is based on a correspondingly weighted sum of the absolute residuals [25].

Based on the plots presented in figure 5.5 we note the following:

- Intercept: The intercept covers a wide range across the different deciles. While it is almost zero for the lowest three it steadily increases for the higher deciles and thus contributes more to the length of the delay for longer reporting delays.
- *Breach date*: For this coefficient we observe a change of the sign for the lowest decile. This suggests that events with very short reporting delays have actually showed an increase in the delay over the past couple of years. From the first to the second decile there is a strong decrease and a change of sign. Even though four out of eight of the following quantile estimates lie within the 90% confidence bound of the OLS regression coefficient, the OLS coefficient is underestimating the decrease over time for most of the events.
- *Pwn count*: Except for the first and the last two deciles, the coefficient estimates lie within or very close to the border of the 90% confidence interval of the OLS estimate. Moreover, the estimates of the first and the last two deciles are almost zero, which suggests that for events with very short or very long reporting delays the *pwn count* variable does not provide any differentiating information. On the contrary, the *pwn count* variable contributes more to the modelling of the delays from the second decile on, whereby it does more so for the lower deciles than for the higher ones.

In figure 5.6 we plot the reporting delay vs. the *breach date* and add the quantile regression estimates to it, as well as the GLM estimate from the model only taking *pwn count* and *breach date* as predictor variables[5]. For all models we add a loess-smoother based on the *breach date* variable. This yields a continuous non-parametric estimate based only on the *breach date* variable and is in line with our observations above from figure 5.5. For the first and the last two decile regressions the loess-smoother almost follows a straight line and the *pwn count* variable has almost no contribution. For decile regressions in which the

[5] While the *breach date* and the *pwn count* coefficients remain of roughly the same magnitude, the GLM intercept estimate increased by 50% in comparison to the reduced GLM model from section 5.2.1.

pwn count variable has a larger estimate, we observe a broader spread of the fitted values (consider for example the fitted values from the second or third decile regressions). Almost all decile regressions show a decreasing trend. The only exception to this are the regressions from the lowest two deciles. While for the lowest we can observe an increase, the reporting delay does not seem to have changed over time for the second decile. Furthermore, for the 0.3-0.6 quantile regressions we observe a stronger decrease from 2017 on towards the end. However, this could be due to the not yet reported events, i.e. events that happened recently but have a long delay and are thus still unknown. Interestingly the GLM-fit lies mostly between the 0.3-0.4 quantile regressions.

In the right plot in figure 5.6 we show the estimates from the GLM and the quantile regressions on the original scale and have again added a loess-smoother. As the observed linear decrease in the second plot of figure 5.6 was on the logarithmic scale, the decreasing effects become multiplicative on the original scale for the reporting delay. Hence we observe a strong decrease for events prior to 2016 and a less profound decrease later on.

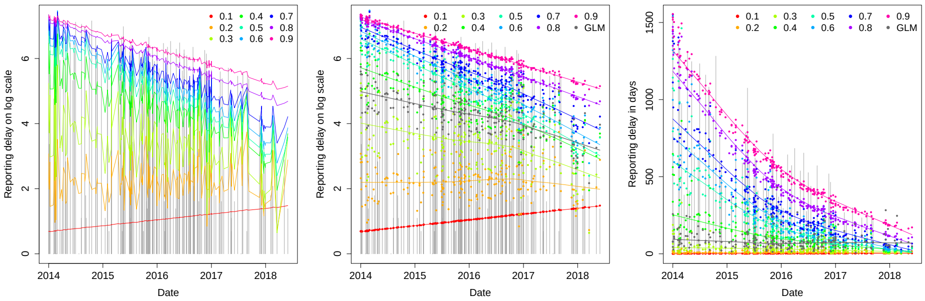


Figure 5.6: Left: Fitted reporting delay values of the quantile regression models per decile with *breach date* and *pwn count* (which specifies the number of records lost) as predictor values vs. the *breach date* variable for the reduced “Have I Been Pwnd” dataset, whereby only events with *breach date* from 2014 are considered, on log-scale. Middle: The former plot also including a loess-smoother for each quantile regression and the fitted reporting delay from the generalized linear model only containing *pwn count* and *breach date* as predictor variables, also with a loess-smoother. Right: The plot in the middle transformed back to the original scale. For the quantile regressions different rates of decrease and increase over time can be observed for different deciles.

Since we are most interested in the dependence between the *time difference* and the *breach date*, we can also fit a univariate quantile regression to a model only containing those two variables (not shown). We observe an increase in the *time difference* for the lowest decile and a decrease for all the other deciles, including the second one for which the estimate remained constant before. Since most of the coefficients change notably over different quantiles, the results from the quantile regression emphasize that a GLM is not a suitable model. Furthermore, we have no statistical evidence that the information attributes contribute in a systematic way as for several quantile coefficient estimates zero was within or close to the 90% confidence bounds. Thus we exclude them from now on. Finally, also for the quantile regressions we have observed very low R^2_τ values which question again the existence of a systematic relationship between the remaining two predictor variables and the response. In the next section we will employ a method that allows for more flexibility in the modelling of the reporting delay.

5.2.3 Generalized additive model fit

In the following we will fit a generalized additive model (GAM) to our dataset. It allows to fit non-parametric penalized regression splines to each predictor individually and hence provides the desired flexibility [5]. Below we show the results of the GAM summary statistics for the model only taking *pwn count* and *breach date* as predictor variables, whereby we have used a spline function for both of them.

Table 5.4: Coefficient estimates of the reduced generalized additive model fit taking *pwn count* (which specifies the number of records lost) and *breach date* as predictor variables based on the reduced “Have I Been Pwnd” dataset, whereby only events with *breach date* from 2014 are considered. For the predictor variables *pwn count* and *breach date* a spline function was used. In the top part the intercept estimate, its standard error, t-test statistic and p-value of the t-test are shown while in the lower part the estimated degrees of freedom (edf) of the spline functions, the reference degrees of freedom, the F-test statistic and the corresponding p-value for the *pwn count* and *breach date* variables are shown. The deviance explained by the model is 21.2%, the adjusted coefficient of determination R^2 equals 18.2%.

	Estimate	Std. error	t value	Pr(> $\|t\|$)
Intercept	4.3	0.13	34	0
	edf	Ref. df	F	p-value
s(PwnCount)	6	7.2	4.9	3.4e-05
s(BreachDate)	2	2.5	6.8	6.1e-04

We start by noting that the log-transformed *pwn count* variable is highly non-linear with an estimated degrees of freedom (edf) close to 6. Also for the *breach date* we can observe a non-linear behavior with an edf of 2, which suggests a behavior of a quadratic function. The two respective functions are shown in the first two plots in figure 5.7. We note that for both variables there exist ranges of predictor values for which the smoothing function is almost constant. For the *pwn count* variable the smoothing function takes negative values for *pwn count* below e^{13}, slightly oscillates thereafter above zero and tends toward zero for the upper range of the *pwn count* variable. This suggests that in particular for events with a low number of records lost the *pwn count* variable has a shortening effect on the reporting delay, whereas for more severe breaches it might have a prolonging effect or no influence at all. For the *breach date* the smoothing function remains constant at the beginning until a few months into 2015 and declines afterwards. Thus it has a shortening effect on the delay for more recent breaches.

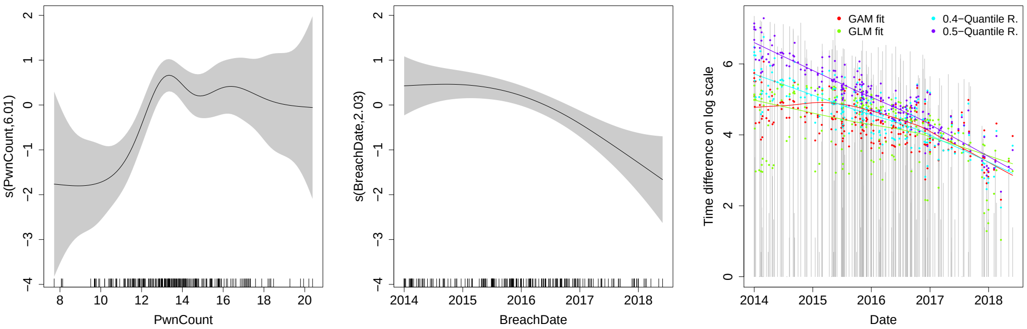


Figure 5.7: Left: Smoothing function of the *pwn count* variable (which specifies the number of records lost) in the generalized additive model fit for the log-transformed reporting delay taking *pwn count* and *breach date* as predictor variables based on the reduced “Have I Been Pwnd” dataset, whereby only events with *breach date* from 2014 are considered. The estimated degrees of freedom of the smoothing function is six which is highly non-linear. Middle: Smoothing function of the *breach date* variable from the aforementioned model, which has an estimated degrees of freedom equal to 2. Right: Fitted reporting delay values on the log-scale for the $\tau \in \{0.4, 0.5\}$ quantile regression models, the generalized linear model (GLM) and the generalized additive model (GAM), whereby all models take the *pwn count* and *breach date* as predictor variables. The generalized additive model shows a constant reporting delay for 2014 and a decline thereafter, whereby the other three models show a decline over the complete considered time horizon.

We should also note the low value of the adjusted R^2 and the deviance explained by the model (see table 5.4). This suggests that only a rather small part of the overall variance is explained by the model even though we have chosen a more flexible method this time. This supports the previous concerns about the existence of a systematic relationship between the reporting delay and the predictor variables for the given dataset.

In the right plot of figure 5.7 we show the GAM fit vs. the *breach date* variable, the τ-quantile regression estimates for $\tau \in \{0.4, 0.5\}$ and the GLM fit only based on the *pwn* *count* and *breach date* variable. We have added a loess-smoother for all fits. In comparison to the other models the GAM loess-smoother remains constant for the year 2014 and shows a similar decline as the others afterwards. While the GLM fit lies mostly between the 0.3 and 0.4 quantile regressions, besides the first year the GAM lies mostly between the 0.4 and 0.5 quantile regressions.

5.3 Conclusion

In the previous subsections we have fit three different models. First and foremost, all three of them score very low with regards to the considered goodness of fit measures and if the fit was tested on random halves, the fits were not very stable. Many of the information attributes dropped out early on and they do not appear to relate in a systematic way to the reporting delay. However, answering this question is also challenging as many of them only seldomly occur. The existence of a systematic relationship between the delay and the two continuous variables *pwn count* and *breach date* remains questionable. It is however certain that for the given dataset we cannot explain most of the variation observed in the delay solely with these two variables.

A model comparison is thus challenging as the relationship does not seem to be well captured or remains largely unaccounted for in all of the three models. With these concerns in mind, all of them have however showed a decrease of the reporting delay over time. In particular we have seen from the quantile regression model that the decrease is of different rate for the various deciles and that the shortest delays have actually shown an increase over time. Furthermore the GAM fit suggests that the variables *breach date* and *pwn count* might only contribute to the modelling of the *time difference* for a specific range of values.

Thus we do not believe to have found a sound model for the reporting delay with the given dataset. Much of the behavior of the delay remains unaccounted for and further predictor variables need to be considered. There is an indication for a decreasing trend over time, however these results have to be treated with reservations. Furthermore, in the above analysis we have used the *breach date* variable and pursued the question: if a data breach happens on a given day, how long is its delay expected to be? Naturally one can also use the *added date* variable, which then tries to answer the question: if a breach becomes known at a certain date, when is it expected to have originally occurred? In any case one needs to check for the existence of a time dependence between the delay and the considered date.

6 Data Breach Notification Laws in the United States

In the following we compare the number of data breach events before and after the introduction of data breach notification laws in the US. This is of interest as regulations have been introduced state wise and not on a national level at once. Therefore, we can analyze whether the introduction of notification regulations has had any effect on the reporting of such events.

6.1 Effective dates of notification laws in the United States

For the analysis we use the main dataset presented in section 1.5 and consider organizations which are headquartered in the US and for which the state or US territory[1] is known. Hence there are 690 observations that can be used for the analysis. To answer the posed question, we start by considering figure 6.1 which shows when notification laws became effective within the US. The covered timespan is quite broad but it is also clearly visible, that most states have introduced regulations within the beginning of 2005 and end of 2007 (38 out of the 52 considered, i.e. 73%).

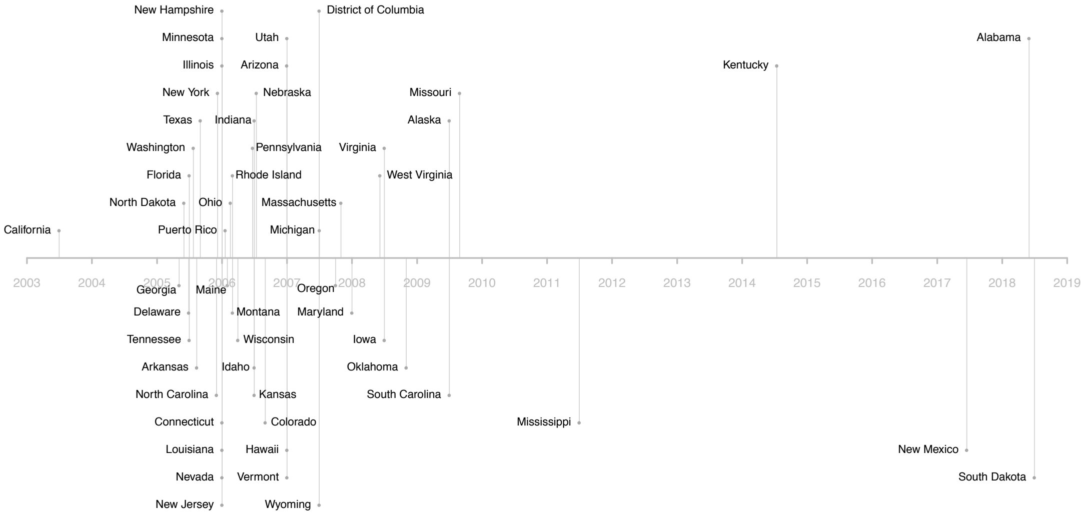


Figure 6.1: Effective date of notification regulations per state within the United States [28]. The height of the points is not meaningful and solely used for visualization purposes.

[1] Puerto Rico is not a state but one of the US territories. Besides the 50 official states, the federal district of Columbia is also included in the dataset.

6.2 Analysis

We directly exclude states with no events reported and states that have had their data breach notification laws introduced before the first reported event in the dataset from the analysis. Furthermore, we also exclude any states with no observations before or after the effective date of the regulations, as we cannot infer anything from them. This leaves us with 390 events for the set of states shown in table 6.1.

Table 6.1: States and their effective date of the first data breach notification regulations, total number of registered events and number of events before and after the effective date of notification regulations based on the complete dataset (the total is also split by datasources: Privacy Rights Clearinghouse (*PRC*), breach level index (*bli*) and Information is Beautiful (*IiB*)). Only states with a positive number of data breach events both before and after the effective date are shown.

					Split according to datasource		
State	Effective Date	# before	# after	Total	*bli*	*IiB*	*PRC*
Arizona	2006-12-31	1	6	7	2	0	5
Colorado	2006-09-01	2	9	11	1	0	10
Connecticut	2006-01-01	1	10	11	2	0	9
District of Columbia	2007-07-01	4	32	36	8	2	26
Georgia	2005-05-05	1	34	35	5	1	29
Illinois	2006-01-01	1	27	28	8	0	20
Iowa	2008-07-01	2	5	7	2	1	4
Kentucky	2014-07-15	1	5	6	2	0	4
Maine	2006-01-31	1	1	2	0	0	2
Maryland	2008-01-01	3	12	15	0	1	14
Massachusetts	2007-10-31	7	13	20	6	1	13
Michigan	2007-07-02	2	6	8	2	0	6
Minnesota	2006-01-01	2	11	13	3	0	10
Missouri	2009-08-28	2	11	13	2	0	11
Nebraska	2006-07-13	2	4	6	1	0	5
New Jersey	2006-01-01	1	15	16	6	1	9
New York	2005-12-07	4	62	66	12	5	49
North Carolina	2005-12-01	1	10	11	4	0	7
Ohio	2006-02-17	2	15	17	1	2	14
Oklahoma	2008-11-01	1	8	9	2	1	6
Oregon	2007-10-01	1	8	9	3	2	4
South Carolina	2009-07-01	1	3	4	0	0	4
Tennessee	2005-07-01	1	11	12	3	0	9
Utah	2007-01-01	1	5	6	1	0	5
Virginia	2008-07-01	4	18	22	0	3	19

Considering the above table, we see that for many states only one event is reported before the regulations became effective. This is especially the case for states which have introduced them early. In the plots in figure 6.2 we show when the individual events happened and when the corresponding regulations were introduced (red line) for the states in table 6.1. We observe the following:

- There is a high variety between the number of events per state.
- For most of the states, the frequency of data breaches does not show any clear pattern and there are also longer timespans of no data breach events, consider for example Arizona, Connecticut and Utah. For some states we also observe time periods with an accumulation of events, such as Georgia between 2016 and 2017 or New Jersey in 2014.
- The only states for which one might wonder if the rate has increased after the introduction of the new regulations are Illinois, New York, Ohio and Virginia. For Illinois there is one event in the year before the regulations became effective and four and three in the two following years respectively. At the same time we also observe larger periods of no events, i.e. 2008 until mid 2009 and mid 2010 until mid 2012. For New York we make a similar observation. While four events have been reported before the regulations became effective, ten have occurred in the same timespan afterwards. However, also for New York we find again timespans with fewer events. The same holds true for Ohio and Virginia as both of them show an accumulation of events after the introduction as well as some sparse timespans afterwards.

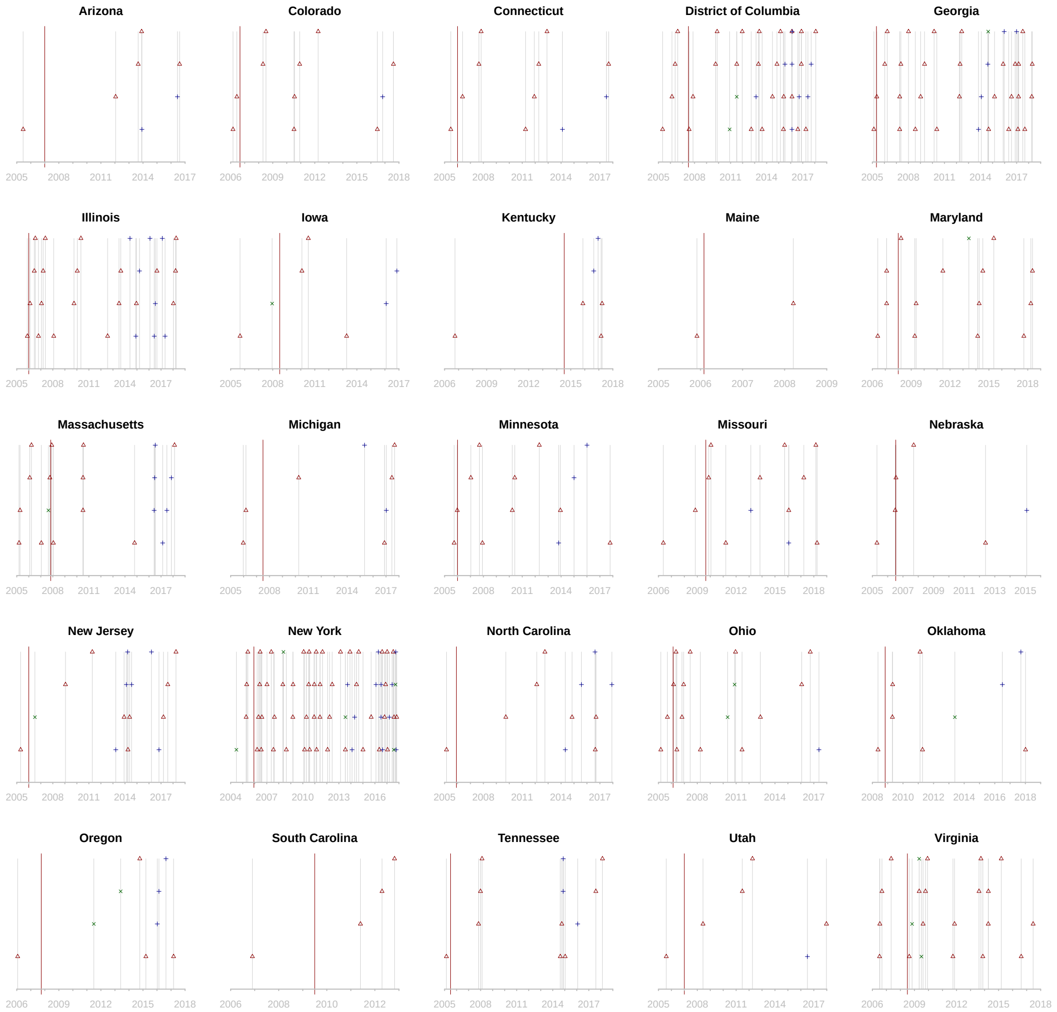


Figure 6.2: Number of events before and after the introduction of data breach notification regulations for the states (from left to right and top to bottom) Arizona, Colorado, Connecticut, District of Columbia, Georgia, Illinois, Iowa, Kentucky, Maine, Maryland, Massachusetts, Michigan, Minnesota, Missouri, Nebraska, New Jersey, New York, North Carolina, Ohio, Oklahoma, Oregon, South Carolina, Tennessee, Utah and Virginia based on the complete dataset. The red line shows when the regulations became effective and the symbols make it possible to distinguish between events that happened close in time and mark the datasource of the events: the red triangle corresponds to Privacy Rights Clearinghouse (*PRC*), the blue plus to breach level index (*bli*) and the green cross to Information is Beautiful (*IiB*). The height of the points is not meaningful and solely used for visualization purposes.

- Another state that stands out is Massachusetts. Notification laws were introduced in October 2010 but even before that many events were reported and became known to the public. From 2016 onwards we observe again more events, whereby most of them stem from the same datasource (*bli*).

6.3 Conclusion

Based on table 6.1 and figure 6.2 we have no evidence that suggests an influence of the introduction of notification laws on the number of reported events over time for the given dataset. This is primarily due to the small size of the dataset which results in few observations before the introduction of the regulations and notable sparse regions afterwards. Furthermore, we have to keep in mind that we only consider events with at least $70k$ items breached, which account for most of the records lost, but constitute only a small fraction of the total number of data breaches[2].

[2] For the dataset used in [37] it was observed that while breaches with at least $40k$ records lost account for more than 99% of the total records lost, they constitute less than 10% of the overall number of events.

7 Conclusion

Below we give a short summary of the overall results and compare them to the previous results from the literature mentioned in section 1.3. In a second part we highlight some of the key questions of interest for further analyses.

7.1 Summary and discussion of results

In chapter two we were able to identify different risk classes based on the economic sector with respect to the frequency differentiated by severity quartiles. A comparison to the results of Eling and Loperfido [14] is challenging as they identified risk classes based on the economic sector with respect to the attack type (this information was not available in the used dataset). However, in comparison to them we have obtained a much finer classification of the economic sector. The latter provides a clear basis for the definition of risk classes with regards to severity, independently of the type of attack, which is in any way hard to classify in a consistent manner.

In chapter three we analyzed the frequency of data breach events for two different subsets at a threshold of $70k$ records lost. The first subgroup consists of the PRC subset, for which a Poisson model provides the best fit for the monthly and quarterly counts and does not show an increase over time. In Eling and Loperfido [14] the same datasource was used, however they considered all data breach events for which the number of lost records was known and modelled the frequency on a daily basis, which is best approximated by a negative binomial distribution. Thus one or the other distribution is more suitable for different thresholds and time intervals considered. In the second subgroup, which consists of the events from all three sources with a date within the beginning of 2013 until the end of 2017, a mostly increasing development over time was visible. In particular is the observed trend driven by one source (bli). However, based on the given dataset we cannot predict the future development of the frequency over time.

In the first part of chapter four the severity was analyzed in detail with respect to all available variables. A truncated regression model showed a significant increase over time for the complete dataset. However, when fitting a truncated regression model on each of the three datasources, the observed increase of the complete dataset was not present for every datasource. Furthermore, from the additional factor variables it became clear that there exists an information reporting issue for data breaches which are less severe - even though we are considering a threshold of $70k$ records lost. For the same threshold we were also not able to identify any relationship between the severity and the size of the organization as it was done by Wheatley, Maillart and Sornette [37], whereby we have considered two measures for the size of the affected entity. Furthermore, we have also studied the development over time by breach medium. Already in the first section of chapter four it became evident that the two media (software and hardware) considered follow a different severity distribution. While there was enough statistical evidence for a different intercept in a truncated regression model, the increase over time of the two medium types appears to be the same (at a 95% confidence level). However, assessing this question was challenging as there were fewer events with a hardware medium for more recent years. In the third section of chapter four we estimated various distribution functions for the severity at different thresholds. Hereby the truncated lognormal and the upper-truncated Pareto distributions give a reasonable fit. In a second step we differentiated by economic sectors at the threshold $70k$ records lost and it became evident that on this granularity level both the threshold and sector are of high importance when estimating the distributions. While the truncated lognormal and upper-truncated Pareto have been suitable choices for the complete dataset at various thresholds, a larger class of distribution functions should be considered for some economic sectors (e.g. for the healthcare sector).

In chapter five the reporting delay was analyzed and there was no convincing statistical evidence found for the delay to relate in a systematic way to the breach date, that is no trend over time could be detected. In particular we have seen that most of the variation of the reporting delay remains unaccounted for with the considered variables and methods.

In chapter six we looked at the frequency of data breach events of entities headquartered in the US, per state, while taking the state wise introduction of data breach notification laws into account. Our own dataset provides only a limited view on the issue and thus it is not surprising that there were no clear changes visible.

7.2 Further questions of interest

With our analyses we detected many potential areas of research and questions within the field, whereof we would like to list here the most interesting ones.

Firstly, there is great potential to extend the analysis on the frequency to several thresholds and also apply multivariate techniques instead of remaining in the univariate setting. In particular more recent data should be included in the model in order to make a sound decision on the future development of the frequency of data breaches. Another extension of the frequency analysis is the estimation of missing events due to a reporting delay. This can for example be done by using the chain ladder method [38]. Secondly, while we get reasonable density fits for the complete dataset for the considered distributions at various thresholds, a wider class of distribution functions should be considered for a fitting on a more granular level. In particular, if a dataset with the complete spectrum of the *total records* variable is available, one could properly characterize the distribution of the small data breaches and of the large ones by identifying a suitable threshold, whereby this can as well be considered at several granularity levels for various subgroups. Thirdly, while the reporting delay was analyzed with regards to the breach date, it can as well be analyzed with respect to the date it became publicly known. Hereby a different question is pursued, namely if at a given date a data breach became known, how long has it been since the breach has happened? Furthermore, as our analysis indicated that the number of records lost has a prolonging effect on the reporting delay, a more detailed analysis with regards to various thresholds occurs to be of interest. Hereby it is recommended to consider a different dataset with more variables and which is not prone to a selection bias. Fourthly, the analysis with regards to the data breach notification laws should be revisited on a dataset without a threshold or the threshold that is specified in the respective notification regulations, if any such threshold exists.

A Description of the Datasets

A.1 Main dataset

Total records

This variable specifies the number of records breached and is available for each observation. In our dataset we have only considered events with at least $70k$ records breached, therefore the two histograms in figure A.1 show the upper tail of the severity distribution.

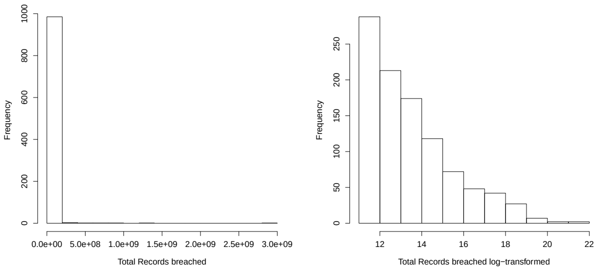


Figure A.1: Histograms of total records breached on original scale (left) and log-scale (right) based on the complete dataset (considering data breach events with at least $70k$ records lost).

The total number of records breached is clearly right-skewed.

Table A.1: Mean, standard deviation and skewness of the total records variable on original and log-scale (LS) based on the complete dataset (considering data breach events with at least $70k$ records lost).

Mean	Sd	Skewness	Mean on LS	Sd on LS	Skewness on LS
$14\ 10^6$	$115\ 10^6$	20	13	2.0	1.1

Country

There are 985 events for which it was possible to assign a headquarter location of the affected business(es). Sometimes it was not possible to do so since multiple firms were affected and they are headquartered in different countries. There are also cases, where it is not even clear which company or organization was the owner of the data. For these observations we have also not assigned a location. In table A.2 we show the different countries and their frequencies in the main dataset.

It is not surprising that the US appears most often. Data breach notification laws have been introduced state wise since 2003 and by now it is mandatory in every state [28]. As a result, the information was publicly available and made it possible to create databases of such events. From the sources that were used for our own dataset, most events originate from the *PRC* set which registers events reported within the US.

Table A.2: Number of data breach events by headquarter location based on the complete dataset (considering data breach events with at least $70k$ records lost).

Country	Freq.	Country	Freq.	Country	Freq.	Country	Freq.	Country	Freq.
US	720	Israel	8	South Africa	3	Vietnam	2	Iceland	1
United Kingdom	59	Russia	7	Chile	2	Argentina	1	Malaysia	1
Canada	26	Turkey	5	Denmark	2	Bangladesh	1	Moldova	1
Japan	24	Hong Kong	4	Iran	2	Belgium	1	Panama	1
China	17	Netherlands	4	Ireland	2	Brazil	1	Philippines	1
South Korea	15	Sweden	4	Malta	2	Bulgaria	1	Qatar	1
India	12	Italy	3	New Zealand	2	Cyprus	1	Serbia	1
France	9	Mexico	3	Pakistan	2	Czech Republic	1	Slovakia	1
Germany	9	Norway	3	Saudi Arabia	2	Finland	1	Syria	1
Australia	8	Poland	3	Spain	2	Greece	1	Taiwan	1

Location state

For the 720 affected entities headquartered in the US, it was possible for 690 of them to register the state. Again, for events with multiple entities affected which are headquartered in the US but in different states no value was entered. In table A.3 we show the frequency for each state appearing in the dataset.

Table A.3: Number of data breach events by headquarter location (state wise) within the US based on the complete dataset (considering data breach events with at least $70k$ records lost).

State	#	State	#	State	#	State	#	State	#
California	130	Indiana	19	Oklahoma	9	Utah	6	Hawaii	2
New York	66	Ohio	17	Oregon	9	Delaware	4	Idaho	2
Texas	37	New Jersey	16	Michigan	8	Louisiana	4	Maine	2
District of Columbia	36	Maryland	15	Pennsylvania	8	South Carolina	4	Montana	2
Georgia	35	Minnesota	13	Wisconsin	8	Alabama	3	New Mexico	2
Florida	31	Missouri	13	Arizona	7	Kansas	3	Vermont	2
Illinois	28	Tennessee	12	Iowa	7	New Hampshire	3	Alaska	1
Washington	23	Colorado	11	Nevada	7	North Dakota	3	West Virginia	1
Virginia	22	Connecticut	11	Kentucky	6	Puerto Rico	3		
Massachusetts	20	North Carolina	11	Nebraska	6	Arkansas	2		

It is no surprise that California and New York appear most often. A lot of companies, in particular companies which operate in the technological sector, are headquartered in California and numerous financial institutions are headquartered in New York. Furthermore, the District of Columbia is also among the top five as most federal governmental bodies are located there. For all of these three groups numerous events have been registered (see table A.8).

Date

This variable is known for every event. However, it should be seen more as a proxy date. In some cases, the actual breach date is known and might have been even entered into the database for the date. However, in other cases the date might refer to the point of time when the breach became publicly known and was reported in the media. In our dataset the first event happened on 2004-06-23 and the last on 2018-05-15. Below we show the number of reported events per year, whereby both the first and last year are considered to be incomplete.

Table A.4: Number of data breaches per year based on the complete dataset (considering data breach events with at least $70k$ records lost).

Year	# Events	Year	# Events	Year	# Events
2004	1	2009	27	2014	122
2005	31	2010	42	2015	107
2006	50	2011	48	2016	183
2007	49	2012	48	2017	129
2008	47	2013	90	2018	19

It is important to keep in mind that we have merged different databases which have been recording events over different time spans and for different locations. Therefore an increase over time can also be due to the fact that at some point of time we can observe more. In figure A.2 we show how much the different sources contribute to the complete dataset.

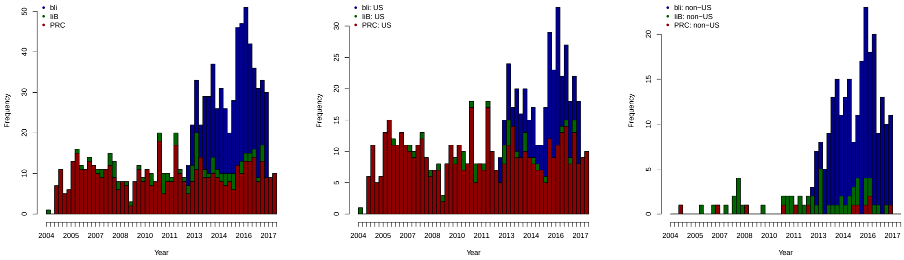


Figure A.2: Histograms of quarterly data breach events based on the complete dataset (considering data breach events with at least $70k$ records lost), color-coded for different sources (blue for breach level index (bli), green for Information is beautiful (IiB), red for Privacy Rights Clearinghouse (PRC)) and shown for all data breach events (left), data breach events with affected entities headquartered in the US (middle) and data breach events with affected entities headquartered outside of the US (right).

Market capitalization

142 entities from our dataset are publicly traded on the stock market and we extended the dataset with their market capitalization value in US dollar as of the provided date in the dataset. We acknowledge, that for a handful of events the stock price might have already been adversely affected by the breach at that point of time. However, we still believe that it can be used as a proxy for the company size in most cases. We have adjusted all market capitalizations for inflation[1] and show their values as of 1st June 2018. The histograms of the original and log-transformed market capitalization are shown in figure A.3.

Table A.5: Mean, standard deviation and skewness of the inflation adjusted *market capitalization* variable on original and log-scale (LS) based on the complete dataset (considering data breach events with at least $70k$ records lost).

Mean	Sd	Skewness	Mean on LS	Sd on LS	Skewness on LS
53 10^9	98 10^9	3.3	23	2.4	-1.1

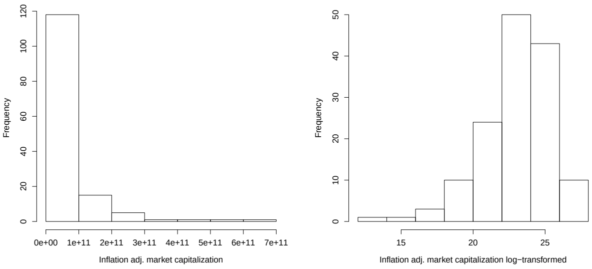


Figure A.3: Histograms of inflation adjusted *market capitalization* on original scale (left) and log-scale (right) based on the complete dataset (considering data breach events with at least $70k$ records lost).

The *market capitalization* variable is clearly right-skewed.

[1] We have used the yearly average inflation rate of the consumer price index for the US dollar [6].

Number of employees

Another indication for the size of an entity is the number of employees. For private and publicly traded companies this number is mostly available (in 550 out of 668)[2]. Overall, for 617 out of 993 observations it was possible to assign a value. Of the 376 observations without a size, 209 belong to government entities.

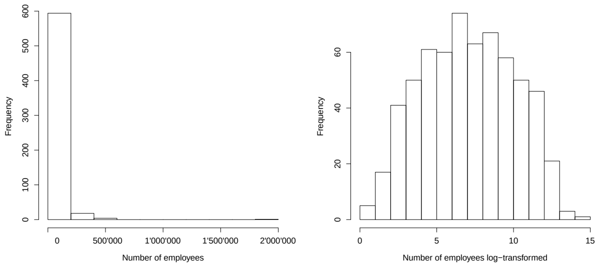


Figure A.4: Histograms of the *number of employees* variable on original scale (left) and log-scale (right) based on the complete dataset (considering data breach events with at least $70k$ records lost).

Table A.6: Mean, standard deviation and skewness of the *number of employees* variable on original and log-scale (LS) based on the complete dataset (considering data breach events with at least $70k$ records lost).

Mean	Sd	Skewness	Mean on LS	Sd on LS	Skewness on LS
$28\ 10^3$	$103\ 10^6$	12.5	7.1	3.0	0.004

This variable is also clearly right-skewed.

Economic sector

For the classification into different sectors we used the Thomson Reuters business classification scheme [33] and extended it with the three categories *mil* for military organizations, *edu* for educational institutions and *pol* for political organizations or election centers. The sectors and their corresponding frequencies are shown in table A.7.

Table A.7: Economic sectors and their frequencies based on the complete dataset (considering data breach events with at least $70k$ records lost).

Sector	Name	Freq.
50	Energy	3
51	Basic Materials	1
52	Industrials	164
53	Consumer Cyclicals	138
54	Consumer Non-Cyclicals	33
55	Financials	125
56	Healthcare	145
57	Technology	203
58	Telecommunication Services	38
59	Utilities	4
edu	Education	87
mil	Military	18
pol	Politics & Elections	23

It is important to note the broad scale of the frequency for the different sectors and that three of them show less than ten events.

[2] If available, the number of employees of the year specified by the *date* variable was taken. For publicly traded companies this is most often available in their annual reports.

Organization type

We can also look in a more general way at entities that were affected by data breaches. This variable classifies them into 5 different types of organizations, whereby we distinguish between entities belonging to a government (*Gov*), businesses that are *private*, businesses that are publicly traded (*MCAP*), not-for-profit organizations (*NPO*) and other *public* institutions that do not belong to governments, such as public schools. For 19 events it was not possible to assign a type. In table A.8 we show the contingency table of this variables split across the different economic sectors.

Table A.8: Contingency table showing both the frequency of organization type and economic sector based on the complete dataset (considering data breach events with at least 70*k* records lost).

	Gov	MCAP	NPO	private	public	NA	Total
50	2	0	0	1	0	0	3
51	0	0	0	1	0	0	1
52	92	21	1	47	0	3	164
53	4	34	2	98	0	0	138
54	7	7	0	19	0	0	33
55	24	28	7	58	1	7	125
56	32	8	25	80	0	0	145
57	6	29	4	162	0	2	203
58	0	14	0	24	0	0	38
59	1	1	0	2	0	0	4
edu	9	0	3	13	62	0	87
mil	18	0	0	0	0	0	18
pol	14	0	1	0	8	0	23
NA	0	0	1	2	1	7	11
Total	209	142	44	507	72	19	993

One should note that some of the combinations can be considered mutually exclusive, as no school district (belonging to the *edu* sector) is publicly traded on the stock exchange (*MCAP*). By looking at this table, we can clearly see some sparse regions, which can either happen because of the former mentioned point or due to the size of the dataset. Moreover, we can see that for some sectors the organization belong in more than 50% of the cases to the same organization type. For example, the organizations of the industrial sector *52* belong mostly to governments. For the sectors *53, 54, 56-59* we can see that actually over 50% of the organizations belong to the private sector, whereby the financial sector slightly misses the 50% mark. The educational sector mostly consists of public organizations, military organizations are completely part of a government and political organization mostly. Hence the organization type variable can often be considered a generalization of the Thomson Reuters' sectors.

Introduced factor variables

For our dataset we have introduced six factor variables which might give additional insights into data breach events. Below we give a short overview for these variables and mention the number of events for each level in brackets.

- *Multiple firms*: Boolean variable indicating if multiple firms were affected by the breach (holds true for 38 events).
- *Insider/outsider*: Factor variable indicating if the key person responsible for the breach was an insider (212) or an outsider (684). This is not always known (97).
- *Medium*: Specifies the medium that was involved in the breach: software (762) or hardware (222), but this is also sometimes unknown (9).
- *Intentional*: Factor variable indicating if the data was breached on purpose (734) or by accident (220), but this is also not always known (39).

- *Failure mode*: Specifies if the breach resulted due to a human error (202), an error in the process (106; for example, poor security standards are considered to be a process error) or due to the used software/hardware (430). This information is not always known (255).
- *Third party*: Specifies if a third party is to a large part or fully responsible for the breach happening (154) or not (556), but again it was not possible to determine this for all events (283).

A.2 “Have I Been Pwnd” dataset

In the following we provide further information about the dataset that was used in chapter 5. The variables *is verified, is fabricated, is sensitive, is retired, is spam list* are booleans and a description of them directly taken from the “Have I Been Pwnd” website [19] is shown in table A.9.

Table A.9: Description of the pre-defined boolean variables in the “Have I Been Pwnd” dataset directly taken from the “Have I Been Pwnd” website [19].

Variable	Description
Is verified	“Indicates that the breach is considered unverified. An unverified breach may not have been hacked from the indicated website. An unverified breach is still loaded into HIBP when there’s sufficient confidence that a significant portion of the data is legitimate.”
Is fabricated	“Indicates that the breach is considered fabricated. A fabricated breach is unlikely to have been hacked from the indicated website and usually contains a large amount of manufactured data. However, it still contains legitimate email addresses and asserts that the account owners were compromised in the alleged breach.”
Is sensitive	“Indicates if the breach is considered sensitive. HIBP enables you to discover if your account was exposed in most of the data breaches by directly searching the system. However, certain breaches are particularly sensitive in that someone’s presence in the breach may adversely impact them if others are able to find that they were a member of the site. These breaches are classed as ”sensitive” and may not be publicly searched.”
Is retired	“Indicates if the breach has been retired. After a security incident which results in the disclosure of account data, the breach may be loaded into HIBP where it then sends notifications to impacted subscribers and becomes searchable. In very rare circumstances, that breach may later be permanently remove from HIBP where it is then classed as a ”retired breach”.”
Is spam list	“Indicates if the breach is considered a spam list. This flag has no impact on any other attributes but it means that the data has not come as a result of a security compromise.”

The variable *data classes* lists the various types of information that were included in each data breach. Overall there are 110 different types of information in the original dataset. Looking more closely at these attributes, we observe that some of them refer to the same information. For example, *dates of birth* appears as an attribute, but *age*, *age groups* and *year of birth* also appear within the list and refer to the same information. In such cases the variables have been merged. The complete list of the original attributes for the *data classes* variable including their respective frequencies is shown in table A.11 and the used mapping is shown in table A.10.

Table A.10: Mapping of the *data classes* variable information attributes which refer to the same information in the “Have I Been Pwnd” dataset [19].

Original	Mapped to
Age groups	Dates of birth
Ages	Dates of birth
Email messages	Chat logs
Passport numbers	Government issued IDs
Private messages	Chat logs
Races	Ethnicities
SMS messages	Chat logs
Years of birth	Dates of birth

Table A.11: Original information attributes of the *data classes* variable and their frequencies in the “Have I Been Pwnd” dataset [19].

Information attribute	Frequency	Information attribute	Frequency
Email addresses	279	Auth tokens	1
Passwords	241	Bank account numbers	1
Usernames	192	Banking PINs	1
IP addresses	132	Beauty ratings	1
Names	78	Biometric data	1
Dates of birth	75	Buying preferences	1
Website activity	72	Car ownership statuses	1
Phone numbers	46	Career levels	1
Physical addresses	45	Cellular network names	1
Genders	43	Charitable donations	1
Geographic locations	26	Chat logs	1
Private messages	14	Credit card CVV	1
Job titles	11	Customer feedback	1
Security questions and answers	10	Customer interactions	1
Employers	8	Deceased date	1
Instant messenger identities	8	Deceased statuses	1
Spoken languages	8	Device usage tracking data	1
Government issued IDs	7	Drug habits	1
Payment histories	6	Eating habits	1
Account balances	5	Financial investments	1
Avatars	5	Financial transactions	1
Credit cards	5	Fitness levels	1
Email messages	5	Health insurance information	1
Marital statuses	5	IMEI numbers	1
Purchases	5	IMSI numbers	1
Social connections	5	MAC addresses	1
Browser user agent details	4	Net worths	1
Ethnicities	4	Nicknames	1
Home ownership statuses	4	Occupations	1
Income levels	4	Parenting plans	1
Physical attributes	4	Password hints	1
Sexual orientations	4	Payment methods	1
User website URLs	4	Personal health data	1
Education levels	3	Personal interests	1
Family members’ names	3	Political donations	1
Family structure	3	Political views	1
Historical passwords	3	Professional skills	1
Partial credit card data	3	Profile photos	1
Passport numbers	3	Purchasing habits	1
Time zones	3	Races	1
Age groups	2	Recovery email addresses	1
Credit status information	2	Religions	1
Device information	2	Reward program balances	1
Drinking habits	2	Salutations	1
Homepage URLs	2	School grades (class levels)	1
Nationalities	2	Smoking habits	1
Personal descriptions	2	SMS messages	1
Relationship statuses	2	Support tickets	1
Sexual fetishes	2	Survey results	1
Social media profiles	2	Travel habits	1
Years of birth	2	User statuses	1
Address book contacts	1	Utility bills	1
Ages	1	Vehicle details	1
Apps installed on devices	1	Work habits	1
Astrological signs	1	Years of professional experience	1

B Additional Material

B.1 Chapter 3 Frequency

B.1.1 Residual analysis of the Privacy Rights Clearinghouse time model

The residual plots for the monthly and quarterly time models of section 3.1.1 are discussed below.

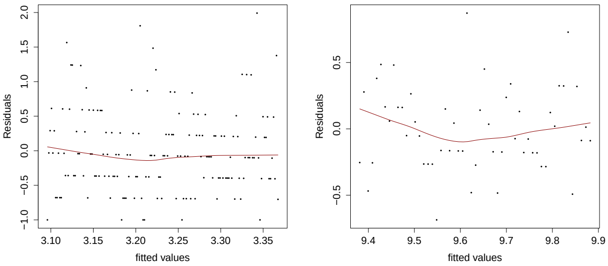


Figure B.1: Residuals vs. fitted values of the Poisson generalized linear model taking *date* as predictor variable with loess-smoother for the monthly counts (left) and quarterly counts (right) for the Privacy Rights Clearinghouse (PRC) datasource (considering events with at least $70k$ records lost). As the values are discrete horizontal lines can be observed in the residual plots.

For the monthly counts the loess smoother follows a straight line and does not show any major deviations. However, we see a clear pattern among the residuals as all of them are located on horizontal lines of different levels. This is due to the fact that our response variable N is discrete and only takes values in $\{0, \dots, 10\}$. Moreover, the range of the fitted values is quite narrow which indicates that the model is almost constant over time. Additionally there are two sparse areas visible in the upper half of the residual plot and one might wonder, if this is due to a time pattern. However, when looking at the autocorrelation of the residuals it only shows a borderline indication at a time lag of six months and we therefore consider this to be due to randomness. For the quarterly counts there is some curvature visible in the loess smoother. If we model the counts as a polynomial of degree two of the date variable the fit is only slightly improved and we therefore prefer to stick with the smaller model. For the quarterly counts the residuals spread more evenly and there is no sign of autocorrelation. We observe again some sort of levels among the residuals. However, in this case it is less pronounced as the response variable covers a broader range of values. As before, the range of the fitted values is very narrow in comparison to the range of the observations.

B.1.2 Exploratory Analysis of the sector percentage models

In the following we present the exploratory analysis of the PRC sector percentage model presented in section 3.1.2. In figure B.2 we show the histograms of the sector predictor variables, whereby the top row shows the ones for the monthly counts and the bottom row the ones for the quarterly counts. For the monthly count model zero appears most often for all sectors and we observe that there were some months for which all the affected entities belonged to the same sector. For the quarterly count model zero also appears most often except for the healthcare sector (*S56*). Moreover, we see a much more diverse picture with respect to the different sectors. Mostly the various sectors are included with a low percentage in the quarterly counts and in rare cases a single sector was predominantly present with a share above 50%. Recall that the sector percentages do not necessarily add up to one, as the total number of events also includes events from the sectors *54*, *58*, *edu* and *other*, which have not been included in the model.

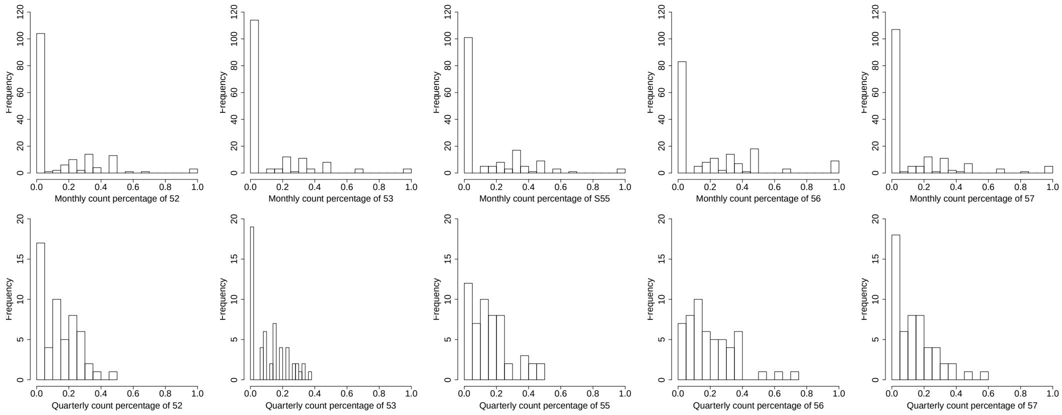


Figure B.2: Histogram for sector percentage variables (which specify the percentage of counts from an economic sector in the respective counts) for the monthly counts (top row) and quarterly counts (bottom row), whereby from left to right the sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*), technology (*57*) are shown. The x-axis shows the monthly (top row), or quarterly (bottom row) respectively, percentage value of the respective counts in $[0, 1]$.

The pairs plots of the predictor variables are shown in figure B.3.

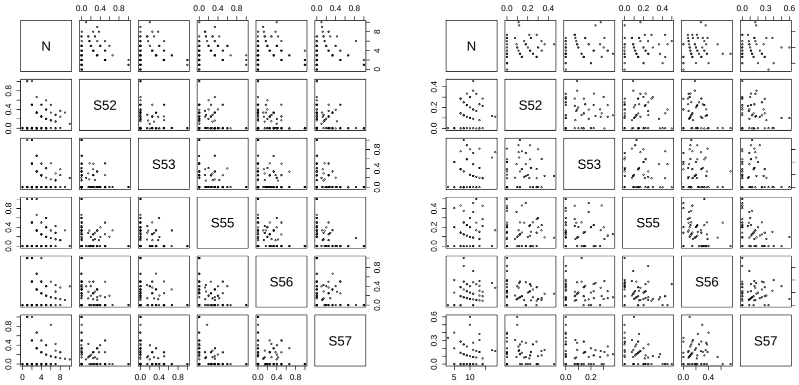


Figure B.3: Pairs plot for the count N and sector percentage variables (which specify the percentage of counts from an economic sector in the respective counts) for the economic sectors (industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*)) for the Privacy Rights Clearinghouse (*PRC*) datasource for the monthly counts (left) and quarterly counts (right).

We first start by analyzing the monthly pairs plot as it is easier and many observations can be directly transferred to the quarterly pairs plot. For the monthly counts vs. the sectors we observe several decaying lines next to each other. This pattern is a direct consequence of the monthly counts range (here $N \in \{0, \ldots, 10\}$) and the range of the number of times a sector was affected. The sectors *52*, *53* and *55* show at most three events per month which then resulted in the decaying lines $1/N$, $2/N$ or $3/N$. The same holds for the sectors *56* and *57*, whereby for *56* we have also two months in which it suffered four events and for *57* one month in which it suffered five events. The first decaying curve $1/N$ is the most clearly visible one for all sectors as all of them have suffered a breach more often once a month than twice or three times.

For the pairs plots of the sector predictor variables we also observe a pattern of diagonal lines, whereby this is clearly visible for the pairs (*S52*, *S55*), (*S52*, *S56*), (*S52*, *S57*), (*S55*, *S56*), (*S55*, *S57*) and (*S56*, *S57*). The diagonal lines at a 45 degree angle are explained by the fact that if the sector pairs both have non-negative values, in more than 50% of the cases they show the same number of events per month. Hereby this mostly consists of cases where both of them suffered from one event. The spread along the diagonal line is then given by dividing by the total number of events per month N, which varies in between zero and ten. In some cases we can even see several diagonal lines, consider for example the pair (*S52*, *S55*). This happens if the number of events per month for the sectors differ and this combination appears several times for a different total number of events per month. We conclude that even though there are some clear lines visible, the correlation among the sectors is very low.

For the quarterly counts we have a very similar picture. This time however, the range of N as well as the range for the number of events per quarter for an individual sector is larger. Considering the pairs plot of the sectors, we see that the percentage pairs spread a bit more than in the monthly dataset but in some cases we can make out the same diagonal lines. For some pairs we can observe sparse regions (e.g. consider (*S52*, *S56*)), this is partly due to the different axes and partly because the range of the sectors differ more than before as the percentages go maximally up to 40% - 80% for the individual sectors. Again we do not observe any strong correlations among the different sectors.

B.1.3 Analysis of the residual plots for the sector percentag models

We fit a Poisson GLM with log-link function to the complete dataset with the sector percentages as predictors. The residual vs. fitted value plot for both models are shown in figure B.4. For both models the curved loess-smoother function catches the eye. The loess-smoother remains curved if we consider bootstrapped versions of the models (gray lines). Comparing the fitted values to the actual observations shows that we generally overestimate months or quarters with a very low total, which leads to the curvature at the left outer edge of the range of the fitted values in the residual plots. Furthermore, if the model predicts a high value, it generally overestimates the observations. For the monthly counts the residuals are allocated again on several curved levels as N is discrete (and only takes values between zero and ten). For the quarterly model this is much less pronounced as the range of N is larger and we have fewer observations. For both models we do not detect any outliers or leverage points.

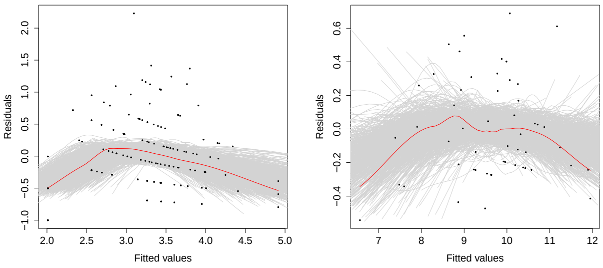


Figure B.4: Residuals vs. fitted values for the Poisson generalized linear models with sector percentages of the respective counts as predictor variables (for the economic sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*)) for the monthly counts (left) and quarterly counts (right) with loess-smoother (red line) and loess-smoother for bootstrapped versions (gray lines) based on the Privacy Rights Clearinghouse dataset (considering data breach events with at least $70k$ records lost).

Also when considering the partial residual plots in figure B.5, we observe a curved smoothing function, whereby this is again more pronounced for the monthly count model than for the quarterly count model. Resampling shows that the loess smoother is not very stable. For the partial residual plots of the monthly model this is again more pronounced as we have a much lower ratio of nonzero percentages vs. zero percentages than in the quarterly model. However, by looking at the individual points we observe that we generally underestimate very low percentages and overestimate high percentages. For the quarterly counts model we get more stable fits as we have a higher ratio of nonzero percentages vs. zero percentages and more evenly scattered residuals. For the sectors *53* and *57* we can again observe an underestimation of very low nonzero percentages which causes a bump in the smooth function. The two residual plots for which the loess-smoother follows mostly a straight line are the ones for the financial (*55*) and healthcare sector (*56*) in the quarterly count model. Considering again their histograms shown in figure B.2, we can see that in comparison to all the others these are the two sectors who are not strongly dominated by the zero percentage value. Therefore we conclude that the bent loess-smoothers in most partial residual plots are primarily due to the accumulations of zero percentages for most sectors in both models and a lower number of observations for higher percentages, which are generally overestimated in the monthly model.

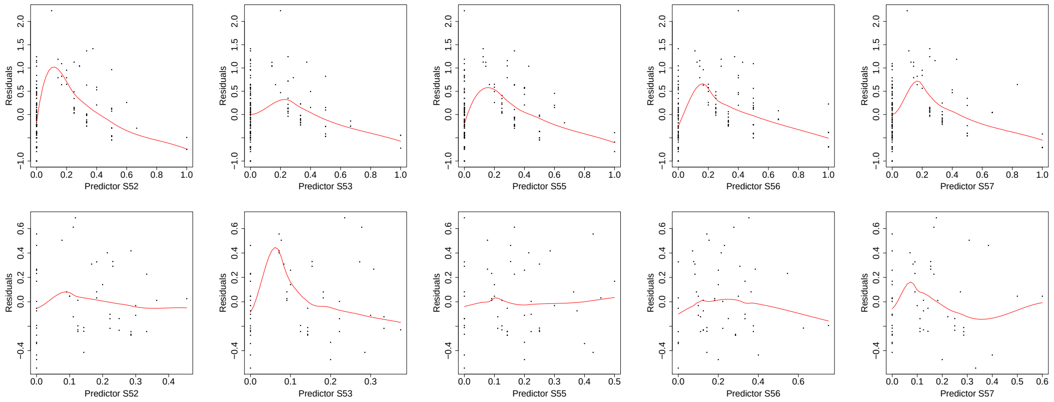


Figure B.5: Residuals vs. sector percentage variables in the Poisson generalized linear models with sector percentages of the respective counts as predictor variables (for the economic sectors industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*)) for the monthly counts (top row) and quarterly counts (bottom row) based on the Privacy Rights Clearinghouse dataset (considering data breach events with at least $70k$ records lost). A loess smoother is shown in red and from left to right the economic sector percentages are industrials (*52*), consumer cyclicals (*53*), financials (*55*), healthcare (*56*) and technology (*57*).

B.1.4 Analysis of the residual plots for the 2013-2017 models

In figure B.6 we show the residual plots of the models presented in section 3.2. For both the Poisson and negative binomial model the residual vs. fitted value plot shows a curved smoothing function for the model with *date* as a linear predictor. For the lower part of the fitted values the residuals are more or less evenly scattered whereas for the third quartile of the fitted values we generally underestimate the counts and for the fourth quartile we overestimate them. In the partial residual plot against the *date* variable we can observe the same curvature for both models. Even though the curvature is not very extreme, it could be an indication for a missing quadratic term in the model. If we include the *date* variable as a polynomial of degree two, the smoothing function no longer shows a curvature in the residuals vs. date plot, but it shows a bump for the higher fitted values in the residuals vs. fitted value plot. In particular many residuals are clustered there and show an overestimation for fitted values in between 10-12 and an underestimation for fitted values above 12. (Again these observations hold for both the Poisson and negative binomial model.)

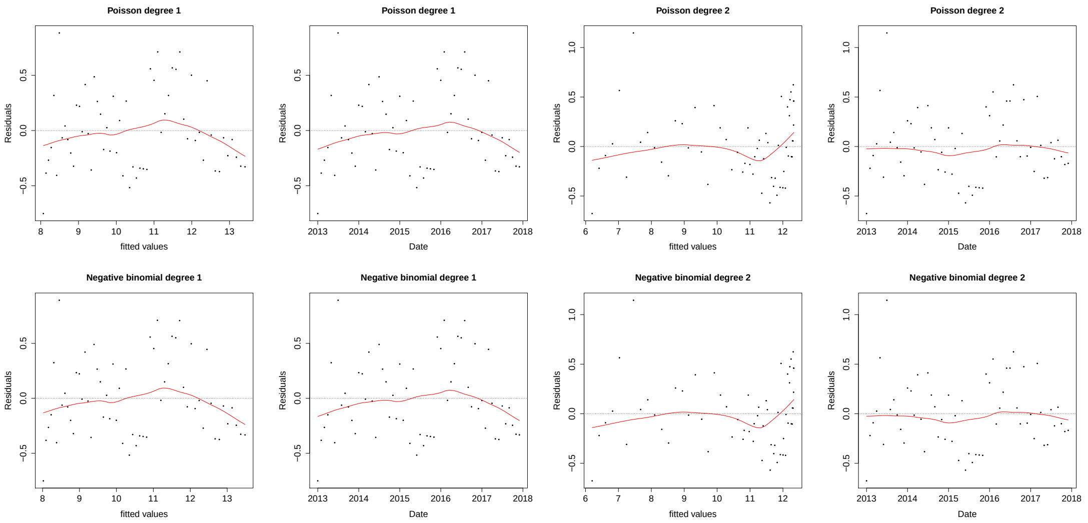


Figure B.6: Residual plots for the Poisson (top row) and negative binomial (bottom row) generalized linear models for monthly counts taking *date* as predictor variable (linear: degree 1; quadratic polynomial: degree 2), whereby on the left two plots for the linear date model are shown: i) residuals vs. fitted values and ii) residuals vs. *date* variable, and on the right the same for the quadratic *date* model. A loess-smoother is shown in each plot (red line). The models are based on the complete dataset with events reported within the beginning of 2013 and the end of 2017 with at least $70k$ records lost.

If we account for this linear or quadratic trend, we would like to know whether or not the residuals form a stationary process[1]. For this we conduct the KPSS test [27], where the null hypothesis states that the time series is stationary around a deterministic trend and the alternative states that the time series has a unit root. As we have already corrected the time series for a time trend via the fitted Poisson or negative binomial model, we conduct the test on the residuals with the null hypothesis of an intercept and no time trend. For all four models (Poisson of degree 1 and 2, negative binomial of degree 1 and 2) the test does not reject the null hypothesis[2]. If we look at the residual time series, we can see some timespans with a lower rate of events as seen in the original process, but overall the process does not seem to be non-stationary. However, the residuals show some slight autocorrelation at a time lag of 4.

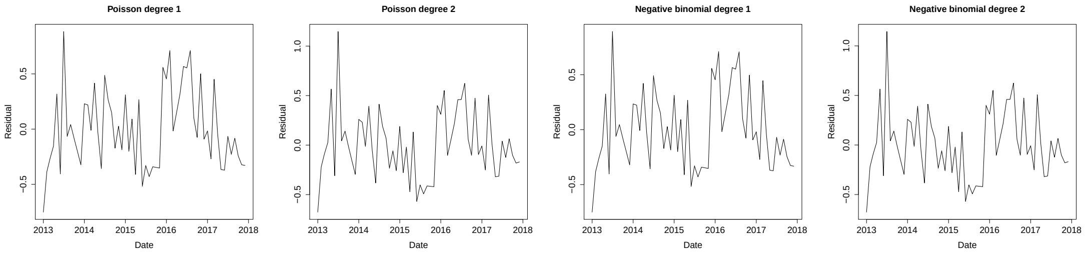


Figure B.7: Residual time series of the monthly counts Poisson generalized linear models (two on the left) and monthly counts negative binomial generalized linear models (two on the right), whereby the first and third plot show the residuals from the linear *date* models and the second and fourth from the models taking the *date* variable as a polynomial of degree two as predictor variable. The models are based on the complete dataset with events reported within the beginning of 2013 and the end of 2017 with at least $70k$ records lost. The residuals are slightly autocorrelated at a time lag of four but there was no statistical evidence found for them to be non-stationary (at a 95% confidence level using the KPSS test [27]).

[1] It is important to know the structure of the residuals as we want to bootstrap them later on.

[2] The hypothesis tests were conducted at a 95% confidence level and for all for models the results p-values were larger than 0.1.

B.2 Chapter 4 Severity

B.2.1 Residual analysis of the *date* and *medium* model

The residual plots of the three models discussed in section 4.2 are shown in figure B.8. First and foremost, it is important to keep in mind that the residuals follow a truncated distribution as well [36]. In all three models we observe a separation of the residuals from the two media groups, whereby this separation is partly visible for M1, more so for M2 and fully for M3. While the residuals of the HW group accumulate at the lower range of the fitted values, the upper range is at least dominated or fully occupied by the residuals from the SW group. Hereby we also note some differences in the spread of the residuals between the HW and SW group, as the latter covers a broader range. The negative residuals are captured within a bounded area, which is due to the lower truncation bound and the modeling of an increasing trend. Besides the three largest positive residuals of the SW group, the two largest positive residuals of the HW group and the aforementioned separation along the fitted values, the remaining residuals scatter evenly in the upper half for both groups. In all three plots the added loess smoother follows a slightly curved horizontal line. We added individual loess smoothers for both groups and in particular for the model M1 we observe that HW events are slightly overestimated and that the overestimation increases for higher fitted values.

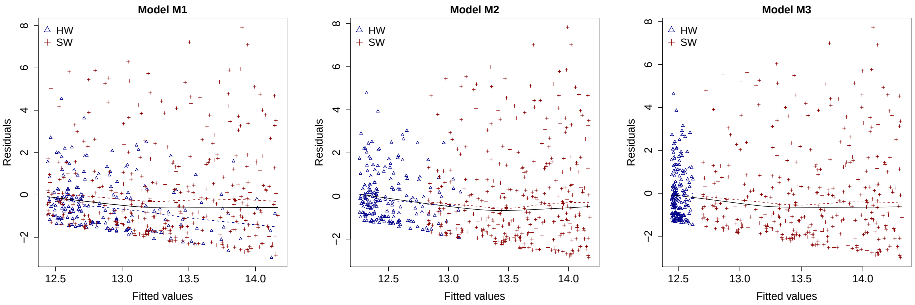


Figure B.8: Residual plots of the three truncated regression models M1 (logarithmized *total records* $\sim$ *date*; left), M2 (logarithmized *total records* $\sim$ *date* and *medium*, whereby *medium* is a factor variable taking the two levels hardware (HW) and software (SW); middle) and M3 (logarithmized *total records* $\sim$ *date* and *medium* as a factor model; right) for the Privacy Rights Clearinghouse datasource (considering events with at least $70k$ records lost), whereby residuals and loess-smoothers of the hardware group are colored blue and residuals and loess-smoothers of the software group are colored red. A loess-smoother for all the residuals is shown in black and for all three models a separation of the residuals of the two groups hardware and software is visible.

Bibliography

[1] Inmaculada B. Aban, Mark M. Meerschaert, and Anna K. Panorska. Parameter estimation for the truncated pareto distribution. *Journal of the American Statistical Association*, 101(473):270–277, 2006.

[2] Hervé Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[3] Hervé Abdi, Lynne J. Williams, and Domininique Valentin. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(2):149–179, 2013.

[4] Christian Biener, Martin Eling, and Jan Hendrik Wirfs. Insurability of Cyber Risk: An Empirical Analysis. Working Papers on Finance 1503, University of St. Gallen, School of Finance, January 2015.

[5] Peter Bühlmann and Martin Mächler. Computational statistics, 2016. Lecture notes, ETH Zürich.

[6] Bureau of Labor Statistics of the U.S. Department of Labor. *CPI-All Urban Consumers (series ID: CUUR0000SA0)*, 2018.

[7] A.Colin Cameron and Pravin K. Trivedi. Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics*, 46(3):347 – 364, 1990.

[8] Privacy Rights Clearinghouse. Chronology of data breaches, 2018. `https://www.privacyrights.org/data-breaches`, last visited 2018-06-2.

[9] Robert A. Cribbie, Rand R. Wilcox, Carmen Bewell, and H. J. Keselman. Tests for treatment group equality when data are nonnormal and heteroscedastic. *Journal of Modern Applied Statistical Methods*, 6(1):117–132, 2007.

[10] Yves Croissant and Achim Zeileis. *truncreg: Truncated Gaussian Regression Models*, 2018. R package version 0.2-5.

[11] Marcel Dettling. Applied statistical regression, 2016. Lecture notes, ETH Zürich.

[12] Benjamin Edwards, Steven Hofmeyr, and Stephanie Forrest. Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*, 2:3–14, 12 2016.

[13] Thomson Reuters Eikon, 2018. `https://eikon.thomsonreuters.com/index.html`, last visited 2018-07-15.

[14] Martin Eling and Nicola Loperfido. Data breaches: Goodness of fit, pricing, and risk measurement. *Insurance: Mathematics and Economics*, 75(C):126–136, 2017.

[15] Martin Eling, Werner Schnell, and Fabian Sommerrock. Ten key questions on cyber risk and cyber risk insurance. Report, The Geneva Association, November 2016.

[16] Leslie Godfrey. *Bootstrap Methods for Regression Models with Non-IID Errors*, pages 177–217. Palgrave Macmillan UK, London, 2009.

[17] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.

[18] Annette Hofmann, Spencer Wheatley, and Didier Sornette. Heavy-tailed data breaches & the challenge of insuring cyber risks: A man-made natural catastrophe? Working paper, Chair of Entrepreneurial Risks, ETH Zürich and School of Risk Management, Insurance and Actuarial Science, St. John's University, New York, 2018.

[19] Troy Hunt. Have I Been Pawned, 2018. `https://haveibeenpwned.com/`, last visited 2018-06-15.

[20] Breach Level Index. Data breach database, 2018. `https://breachlevelindex.com/data-breach-database`, last visited 2018-04-23.

[21] information is beautiful. Data breach database, 2018. `https://docs.google.com/spreadsheets/d/1Je-YUdnhjQJO_13r8iTeRxpU2pBKuV6RVRHoYCgiMfg/edit#gid=322165570`, last visited 2018-06-6.

[22] Ponemon Institute. 2017 cost of data breach study. Research report, Ponemon Institute LLC, sponsored by IBM Security, June 2017.

[23] Alan J. Izenman. *Modern Multivariate Statistical Techniques.* Springer, New York, NY, 2008.

[24] Roger Koenker and Kevin F. Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156, December 2001.

[25] Roger Koenker and Jose A. F. Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310, 1999.

[26] Belchin Kostov, Mónica Bécue-Bertaut, and Francois Husson. Multiple factor analysis for contingency tables in factominer package. *R Journal*, 5, 06 2013.

[27] Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159 – 178, 1992.

[28] SCHWARTZ & BALLEN LLP. State Security Breach Chart, February 2018. `http://www.schwartzandballen.com/Memos%202017/State%20Security%20Breach%20Chart%20020717.pdf`, last visited 2018-06-04.

[29] Thomas Maillart and Didier Sornette. Heavy-tailed distribution of cyber-risks. *The European Physical Journal B*, 75(3):357–364, Jun 2010.

[30] Bryan Manly. Bootstrapping with models for count data. *Journal of biopharmaceutical statistics*, 21:1164–76, 11 2011.

[31] P. McCullagh and J.A. Nelder. *Generalized linear models.* Chapman and Hall, London, 2nd edition, 1989.

[32] Jérôme Pagès and Mónica Bécue-Bertaut. Multiple factor analysis for contingency tables. In Michael Greenacre and Jorg Blasius, editors, *Multiple Correspondance Analysis and Related Methods*, chapter 13, pages 300–326. Chapman & Hall, Boca Raton, IL, 2006.

[33] Thomson Reuters. Thomson Reuters Business Classification, 2018. `https://financial.thomsonreuters.com/content/dam/openweb/documents/pdf/financial/trbc-fact-sheet.pdf`, last visited 2018-06-15.

[34] Risk Based Security, Inc. Data Breach QuickView Report: Q1 2018 Data Breach Trends. Technical report, Risk Based Security, Inc, April 2018.

[35] Fabio Sigrist. Applied multivariate statistics, 2018. Lecture notes, ETH Zürich.

[36] Michael Smithson and Edgar C. Merkle. *Generalized linear models for categorical and continuous limited dependent variables*. CRC Press, 2014.

[37] Spencer Wheatley, Thomas Maillart, and Didier Sornette. The extreme risk of personal data breaches and the erosion of privacy. *The European Physical Journal B*, 89(7):1–12, Jan 2016.

[38] Mario V. Wüthrich. Non-life insurance: Mathematics & statistics, 2017. Lecture notes, ETH Zürich.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Cyber Risks and Data Breaches

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Schillig	Aline

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zurich, 19.10.2018	Aline Schillig

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.