2019

The transition to data-driven risk based regulation on the continuity of telecommunication networks and services

IDENTIFYING THE TECHNICAL AND ORGANIZATIONAL REQUIREMENTS FOR DATA-DRIVEN RISK BASED REGULATION

STIJN SAVELKOUL

The transition to data-driven risk-based regulation on the continuity of public telecommunication networks and services

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in Engineering and Policy Analysis

Faculty of Technology, Policy and Management

by

Stijn Savelkoul

Student number: 4745795

To be defended in public on October 10th 2019

Graduation committee

Chairperson	: Dr. M.E. Warnier, Multi-Actor Systems
First Supervisor	: Dr. H.G. van der Voort, Multi-Actor Systems
Second Supervisor	: Dr. M.E. Warnier, Multi-Actor Systems
External Supervisor	: J. Wilshaus, Radiocommunication Agency Netherlands
External Supervisor	: G. Kuipers, Radiocommunication Agency Netherlands
External Supervisor	: M. Beemers, Radiocommunication Agency Netherlands

Acknowledgements

A master thesis is the final requirement for the MSc degree in Engineering and Policy analysis of the faculty of Technology, Policy and Management at the TU Delft. I have worked on this thesis for almost seven months, during the second semester of the academic year 2018-2019. I am proud to present to you my final work.

I would first like to thank my thesis supervisors Dr. H.G. van der Voort and Dr. M.E. Warnier, for their valuable feedback and guidance. They steered me in the right direction throughout the entire process of writing this thesis. Secondly, I would like to express my sincere appreciation towards all the people I consulted from the Enforcement Policy Division, Monitoring and Analysis Center, Supervisory Division Information Security and Spectrum Division Continuity of the Radiocommunication Agency Netherlands. I was offered to do an internship at the Enforcement Policy Division, where I was allowed to freely work on this research and use all their resources and expertise. Not only, did this help in writing my thesis, it also gave me a good insight of an office environment. A special thanks goes out to J. Wilshaus, G. Kuipers and M. Beemers, who were my external supervisors from the Port of Rotterdam.

I hope you will enjoy reading this, Amsterdam, October 3th, 2019 Stijn Savelkoul

Executive Summary

The continuous innovation within the telecommunications industry is compounded by changing risks for the consumer. This is the case within the Netherlands, where the use of the internet has a pervasive influence on society, necessitating robust telecommunication networks and services. The telecommunications industry is regulated by the Radiocommunication Agency Netherlands, which currently use a method of risk-based regulation to monitor relevant organizations. This is characterised by allocating resources based on the estimated level of risk that is posed by a monitored organization. While these risk analyses have developed over the years, there is no standardised system heightening the possibility for cross-departmental inconsistencies.

An overview of academic literature reveals a lack of research on the **creation and implementation of data-driven model for risk-based regulation.** Existing studies lean towards the merits of risk-based regulation, in addition to the risks associated with data-driven risk-based regulations. When this literature gap is compounded with the rising pressure for the regulator to perform quality supervision and ensure the **uninterrupted continuity of the telecommunication networks and services** in the Netherlands, it provides a suitable opportunity to explore the applicability of a data-driven solution.

The use of data-driven risk detection models have the potential to identify companies that pose higher social risks to create a more efficient process for the allocation of resources. It can illuminate the relevant risk criteria and characterise the risks that certain organizations may pose. This information supports the decision-making process of regulation, which could lead to a more cost-effective process of supervision. Furthermore, given the inconsistencies that exist between the various supervisory divisions of the Radiocommunication Agency Netherlands, a data-driven model could increase internal cohesiveness. As a result, there is significant interest in the possibility of developing **a risk detection model** that can be adapted to **the current organizational structure and** makes use of the readily available **data within the organization.**

T**he research objective** of this study is to **create and implement an exploratory data-driven model for risk-based regulation,** specifically **on the subject the continuity of the telecommunication networks and services within the Supervisory Field of Information Security** of the Radiocommunication Agency Netherlands. Thus, the primary research question that will be explored is:

How could a data-driven risk analysis for risk-based regulation support the process of risk-based regulation executed by experts?

There are four primary challenges to the creation and implementation of data-driven models: the ability to turn data into useful information, the ability to assess the quality of the data, the availability of the data within the organization and the ability to create a process with a high cost-effectiveness ratio. Additionally, there are specific issues associated with the use of data-driven models in regulatory bodies. Firstly, the use of algorithms and data metrics can be distorted when applied incorrectly. In addition, this

who create data-driven models or use the outcomes of the model are prone to interpret the data so that it will fit the pre-existing narratives The second issue is implementing the data-driven model so that it generates acceptance and can be integrated into a well-functioning working culture within the organization.

The research utilizes data that is sourced from two different departments of the Radiocommunication Agency Netherlands; survey responses from supervised organization and self-reported data regarding network incidents. The final data source consists of social media mentions naming network failures of monitored organizations.

The development of an Exploratory Model based on the available data within the organization highlighted several modelling challenges. The process of developing the Exploratory Model is preceded by an Exploratory Data Analysis which determines the quality and significance of the available data. This allows for a summary of the main characteristics of the data and supports the development of the Exploratory Model. The Exploratory Model involves clustering the data, and developing correlations between the various datasets. The execution of a K-means algorithm segments the organizations into groups that provide similar responses to the surveys. The correlations can be used for facilitating the initiation of risk profiles. The data from network incidents and failures are highly important when creating a model for data-driven risk-based regulation as they are key in defining the different risks posed by supervised organizations. The results are validated through interviews with experts of the relevant divisions in the Radiocommunication Agency Netherlands.

Two primary challenges that exist for the implementation of the model are identified; ensuring all necessary data is available for the creation of the model, and defining the owner of the model. The existence of an institutional boundary between the Supervision Division Information Security and the Spectrum Division Continuity, which hinders data sharing regarding the reports of failures that caused discontinuities of the telecommunication services and networks. This poses challenges for the implementation of a data-driven model, as it is crucial for the relevant teams to have access to all necessary data within the organization. There are two possible solutions to overcome this challenge; either dismantling the institutional boundary, or relocating the data management of reports of failures that caused discontinuities of the telecommunication services and networks from the Spectrum Division Continuity to the Supervision Division on Information Security. The second challenge is defining the most suitable owner of the model. As the model is intended to support regulation for the Supervision Division on Information Security, it is necessary for the owner to have expertise in this area. Moreover, as this will be a data-based model, knowledge on quantitative analyses is preferable. Currently, the Monitoring & Analysis Centre of Radiocommunication Agency is most suitable, due to their expertise in analysing data and working on data-driven models. Moreover, they already have access to all necessary data, which means no organizational restructuring is needed. However, there will be challenges stemming from their lack of knowledge regarding the process of supervision on the continuity of the telecommunication networks and services.

Two directions for future research are identified based on the outcomes of this research. The first step would be to further develop the Exploratory Model, which requires a higher volume of information about failures that caused discontinuities. In addition, the data currently within the databases must be frequently updated. Whilst this is certainly viable, it would be a time intensive and a bureaucratic process. In the meantime, it is possible to execute the clustering algorithm on the whole dataset of both of the surveys with the responses from around 1000 organizations and to compound this with qualitative data from experts and inspectors within the Supervision Division Information Security, to aid in the determination of the risk profiles.

Table of Contents

1. INTRODUCTION 9

1.1 The importance of telecommunication in The Netherlands 11
1.2 The role of the Radiocommunication Agency Netherlands 12
1.3 Societal and academic knowledge gap 13
1.4 Structure of the report 14

2. PROBLEM STATEMENT 15

2.1 Regulation on Telecommunication 15
2.2 Regulatory Framework 15
2.3 Exploratory Model 18

3. CASE STUDY 21

3.1 The Supervisory Field Information Security 21
3.2 Legal framework reporting obligation failures 22
3.3 The Role of Data-Driven Risk-Based Regulation 24
3.4 The Case Specific Research Question 25
3.5 The Dichotomy of The Case Study 26

4. RESEARCH APPROACH 27

4.1 Sub Research Questions 27
4.2 Methodology sub research questions 28

5. THE AVAILABLE DATA 33

5.1 Dataset outcomes survey on Information Security 33
5.2 Dataset outcomes survey on continuity 34
5.3 Data from social media 35
5.4 Dataset from Reporting Desk 36

6. EXPLORATORY DATA ANALYSIS 37

6.1 Residual 38
6.2 Re Expression (data transformation) 41
6.3 Resistant 41
6.4 Revelation (data visualization) 44
6.5 Conclusion EDA 45

7. EXPLORATORY MODEL 45

7.1 The goal of the model 45
7.2 The most important modeling challenges 46
7.3 Datasets used for clustering 46
7.4 Result K-means clustering 47
7.5 Conclusion K means clustering 53
7.6 Which questions influence the creation of the clusters the most? 54
7.7 What is the effectiveness the K Means cluster algorithm? 55
7.8 Correlations 60
7.9 The influence of the results on the process of supervision 66

8. VALIDATION 68

8.1 Validation on the datasets with the survey scores. 68
8.2 Validation on the datasets from social media 70
8.3 The data from the Reporting Desk....70

9. THE ROLE OF THE INVOLVED STAKEHOLDERS 72

9.1 Vertical vs. horizontal organizational structure 72
9.2 External stakeholders 74
9.3 Internal stakeholders....75
9.4 The organizational structure 77

10. THE IMPLEMENTATION OF THE DATA-DRIVEN MODEL....78

10.1 Data sharing within an organization with institutional borders. 78
10.2 The owner of the model within the organization 78

11. DISCUSSION....81

12. CONCLUSION....85

13. RECOMMENDATION....92

14. FUTURE RESEARCH....94

14.1 Future research for AT....94
14.2 Future scientific research....95

15. REFLECTION....96

15.1 Model reflection....96
15.2 Reflection on the organizational structure 97
15.3 Scientific relevance....97
15.4 Social relevance....98

REFERENCES: 99

APPENDIX A1 - COLUMN NAMES OF THE DATASETS ABOUT THE SURVEYS 102

A1.1 The initial column names from the datasets provided by AT 102
A1.2 The column names of the first versions of the datasets with the survey scores....111
A1.3 The column names of the second versions of the datasets with the survey scores.....117

APPENDIX A2. BAR GRAPHS DURING THE DATA VISUALIZATION 123

APPENDIX A3. THE SCATTER PLOTS AND GRAPHS FOR THE ELBOW METHOD....126

APPENDIX A4. THE ORGANIZATIONS PRESENTED IN THE CLUSTERS OF THE DATASETS AND THE MOST IMPORTANT QUESTIONS FOR THE FORMATION OF THE CLUSTERS....132

APPENDIX A5: THE AVERAGE NUMBER OF MENTIONS AND THRESHOLDS OF THE SOCIAL MEDIA DATA....144

APPENDIX A6: REPORTING DESK 147

APPENDIX A7: CONVERSATIONS WITH EXPERTS FROM AT 154

List of Figures

Figure 1: An overview of the steps taken to create the exploratory model 30
Figure 2: The formula for the Euclidean distance 31
Figure 3: an overview of the K-means clustering algorithm 31
Figure 4: The main steps of the data cleaning for the datasets with the survey scores on the subjects Information Security and Continuity 38
Figure 5: Scatterplot with the three different clusters created for the first version of the dataset on the subject Information Security 48
Figure 6: Scatterplot with the two different clusters created for the first version of the dataset on the subject Continuity 49
Figure 7: Scatterplot with the three different clusters created for the first version of the dataset on both subjects 50
Figure 8: Scatterplot with the three different clusters created for the second version of the dataset on the subject Information Security 51
Figure 9: Scatterplot with the two different clusters created for the second version of the dataset on the subject Continuity 52
Figure 10: Scatterplot with the three different clusters created for the second version of the dataset on the both subjects 53

List of Tables

Table 1: The number of times and the percentage a response was provided by the organizations after the first part of the data cleaning. 42
Table 2: The number of times and the percentage an answer was provided by the organizations after the third part of the data cleaning. 43
Table 3: The difference between the number of times and the percentage an answer was provided by the organizations after the first and third part of data cleaning. 44
Table 4: The differences between the survey scores of the clusters created for the first version of the dataset on the subject Information Security 57
Table 5: The differences between the survey scores of the clusters created for the second version of the dataset on the subject Information Security 57
Table 6: The differences between the survey scores of the clusters created for the first version of the dataset on the subject Continuity..... 58
Table 7: The differences between the survey scores of the clusters created for the second version of the dataset on the subject Continuity..... 58
Table 8: The differences between the survey scores of the clusters created for the first version of the dataset on both subjects 59
Table 9: The differences between the survey scores of the clusters created for the second version of the dataset on both subjects 59
Table 10: The total number of failures and number of failures per organization per cluster for the first version of the dataset on the subject Information Security..... 61
Table 11: The total number of failures and number of failures per organization per cluster for the first version of the dataset on the subject Information Security..... 61
Table 12: The total number of failures and number of failures per organization per cluster for the first version of the dataset on the subject Continuity 62
Table 13: The total number of failures and number of failures per organization per cluster for the first version of the dataset on the subject Continuity 62
Table 14: The total number of failures and number of failures per organization per cluster for the first version of the dataset on both subjects..... 63
Table 15: The total number of failures and number of failures per organization per cluster for the first version of the dataset on both subjects..... 63
Table 16: an overview of the advantages and the disadvantages for the possible ownership of the division involved in the process of the supervision on Information Security..... 80

List of Abbreviations

4G	The fourth generation of broadband cellular
5G	The fifth generation of broadband cellular
ACM	Authority Consumer & Market
AT	Radiocommunications Agency Netherlands
CBS	Central Statistics Office
EDA	Exploratory Data Analysis
ENISA	European Union Agency for Cybersecurity
EU	European Union
M&AC	Monitoring and Analysis Center
WIBON	Supervision Division on the law exchange of information above and below ground networks

1. Introduction

This study is conducted to identify the possibilities related to how data-driven models could improve or support the process of regulation. This is performed for the Radiocommunication Agency Netherlands (AT), which regulates telecommunication and other supervisory fields related to telecommunication in The Netherlands. A supervisory field within this organization is selected and a data-driven model is created with the available data. This subject is a test case for future research and development of data-driven models within AT. The relevance of the findings of this research could extend beyond AT, to other regulators.

Firstly, the importance of telecommunication in The Netherlands is introduced (1.1). Then, in section 1.2 the role of AT in regulating the Supervisory Field of Information Security is discussed. Following this, in section 1.3, the academic and societal knowledge gap is addressed in order to demonstrate the relevance of this research. Lastly, the structure of the report is outlined (1.4).

1.1 The importance of telecommunication in The Netherlands

A world without a high standard of telecommunication looks unthinkable nowadays, especially in developed countries such as The Netherlands. In 2018, the Central Statistics Office (CBO) in the Netherlands stated that in 2017, 98 percent of the households in The Netherlands had access to internet at home. This figure was the highest percentage for any European country, and was obtained by the CBO from data provided by Eurostat, which is responsible for providing statistical information to the institutions of the European Union (EU). For all European countries, the average percentage of people with internet access at home was 87 percent. This is 4 percentage points more than two years earlier in 2015. In addition to internet access in households, The Netherlands, alongside Sweden, scored the highest for use of internet on mobile devices. For both countries, 87 percent of the population used a mobile device with internet in 2017. This is in comparison to 55 percent in 2012 in The Netherlands, demonstrating that the number grew rapidly over the years. These figures indicate that the use of internet has become more important to individuals and businesses over the past seven years, and the trend may forecast its continuously increasing importance.

Technological development is one of the key contributors to the increasing importance of telecommunication. Developments in technology related to telecommunication evolve each year to increase the user's accessibility. Wireless internet has become stronger and more stable, and the increased range of these networks leads to an increased number of people and devices connecting to the internet. The next step for the wireless internet network in The Netherlands is the 5G network, which is the fifth generation of cellular network technology. Nowadays, most companies and households receive their internet through a fiber optic cable and create their own wireless Wi-Fi network with a router. The 5G network will make it possible to eliminate the cable and router, so companies and households can connect to the 5G network directly. This means that there will be no physical internet connection between homes and companies and the internet provider.

The fourth generation of broadband cellular network (4G) is predominantly used by mobile devices such as phones or tablets, while the 5G network will be connected to all devices that need to be connected to the internet. However, as more people would be connected to the 5G network, this increases the potential risks if there is an incident such as an outage of the 5G network.

1.2 The role of the Radiocommunication Agency Netherlands

The legislation set by the Dutch Government needs to be regulated. The Dutch regulator on telecommunication is AT regulates thirteen supervision fields in total. One of these supervisory fields is Information Security. The main goal of regulation and enforcement in this field is to provide continuity of public services and networks. The four main areas of the telecommunication law are; lawful interception, data processing and the security of data processing, accessibility of the 112 emergency number and guaranteeing the continuity of the network and/or service. The experts and inspectors within AT which specialize in Information Security observe the growing importance of the continuity of the network. Their aim is to continuously improve the supervisory process within their supervisory field.

Due to limited resources, AT cannot supervise all the telecommunication service and network providers continuously. A way to improve the supervisory field is to allocate resources to the organizations which pose the highest risks. This could be done by inspecting the organization that poses the risk. As resources are limited, AT needs to make decisions regarding where and to what extent their limited resources are used to supervise the companies involved in these subjects. The higher volume of organizations to supervise, the more difficult it is for them to allocate their resources. This leads to regulation based on the risk the organizations pose. Another reason why not all the organizations are supervised equally is because small providers , which are providers with a turnover smaller than 2 million are considered by AT as less important during the process of risk based regulation. This is due to the small impact of failures that cause a discontinuity compared to the other supervised organizations.

AT executes risk-based regulation within their supervisory fields to determine which organization pose the most risks. This is a process to systemically set regulation based on the expert's assessment of the risks of the organizations' non-compliance (Rowe,1997). These organizations are under the authority of the regulator, and could range from individuals to multinationals operating within the regulator's jurisdiction. risk-based regulation was executed for the first time around in the 1980s. The popularity of risk-based regulation has increased over the last years. Hood et al (2001) refers to this as optimizing the ‘cost benefit analysis culture’ to move away from an informal qualitative measure of the risk an organization poses, towards a more calculative and formalised approach. This will benefit regulation by optimizing the limited amount of resources that are available during the process of supervision.
The next step in risk-based regulation is to create a data-driven process based on collected quantitative data, which is called data-driven risk-based regulation. The outcomes of data-driven models could support risk-based regulation, as the implementation of indicators and data could provide the experts within the regulator more

insight on relevant developments within the supervisory fields. The use of Big Data could improve decision-making in critical development areas such as regulation (Hilbert, 2016). Data collected by the regulator could be analysed with data analysis algorithms, optimising the process of resource allocation and thereby creating a more cost-effective process. The algorithm could also define risk criteria within the datasets to extend the knowledge of the experts and inspectors on what the risk indicators are.

AT would need to create a data infrastructure that allows data scientists within the organization to create a data-driven process for risk-based regulation. This infrastructure is based on the collecting and monitoring of the data, data sharing between different teams within the organization, and the owner of the model within the organization. This research consists of two main parts that explore the possibilities for data-driven risk-based regulation within AT. The first part is the creation of an Exploratory Model that could be the first building block to a data-driven risk analysis model. The second part assesses the organizational requirements for the transition to data-driven risk-based regulation. This is executed for the subject continuity within the supervisory field of Information Security.

1.3 Societal and academic knowledge gap

The creation and implementation of an Exploratory Model for data-driven risk-based regulation could be seen as a pilot project where leads for future research may be developed. These lessons are formulated to support AT with knowledge regarding the possible influence of a data-driven model on the supervisory process within the organization, and what the organizational requirements are to create data-driven risk-based regulation. To investigate what kind of changes need to be made within AT to create a new process of regulation where data plays a central role. A data-driven model has never been created within the Supervisory Field of Information Security and only once within the whole organization. Even though the lessons from the model implemented within another supervisory field of AT can be useful to take note of, it is expected that the context of that model differs too significantly from this research's Exploratory Model for subject continuity. This is because that research was mainly aimed on the technical side of the modelling.

Scientific literature is very limited on the subject of data-driven risk-based regulation. Two main articles were found during the research on data-driven risk-based regulation. The first article explored the prospects of data-driven regulation for the Environmental Protection agency. McGarity (2008) discusses that environmental regulators rely heavily on assumption-driven models and limited empirical data. It is difficult for regulators to create data-driven regulation when there is an absence of relevant empirical data. however, there are some grounds for optimism as scientific advances could enable agencies to adopt a data-driven approach for risk-based regulation in the not-too-distant future. This approach offers an advantage in accuracy and efficiency and has the potential to offer more protection to the individuals in society. The protection would be a consequence of the improved compliance of behavior from the supervised organizations.

The second article on the digital transition in politics in the Netherlands, where people and values are central (Kool, 2018). Regulators are demonstrating a growing awareness towards the social and ethical aspects of digitization and are investigating the role of digitization within their domain. Exploring the possibilities of data-driven risk-based regulation is a consequence of the growing awareness. During the period of 2017-2018, the budget for data-driven regulation was raised, indicating that they acknowledge the future importance of data-driven regulation. The article discusses data-driven risk-based strategies for various regulators in The Netherlands, and concluded by naming the promises and downside of data-driven regulation.

The scientific literature is predominantly on the promises of risk-based regulation and there is a lack of research on the creation and implementation of data-driven model for risk-based regulation. This thesis has the research objective of creating an exploratory data-driven model for risk-based regulation on the subject continuity within the supervisory field Information Security. To investigate if data-driven regulation can support the process of risk-based regulation. This gives the following research question:

How could a data-driven risk analysis for risk-based regulation support risk-based regulation executed by experts?

1.4 Structure of the report

Firstly, the problem statement is presented (Chapter 2) by studying literature on the regulation on telecommunication, having conversations with experts within AT, different regulatory frameworks and Exploratory Models and risk detection models. Following, the case study is explained for the regulation on telecommunication in The Netherlands (Chapter 3). The findings of the literature study are used to highlight the most important findings for this specific case. Furthermore, the research approach is described for all the sub research questions and the matching research methods (Chapter 4). Subsequently, the available data is explained for the creation of the Exploratory Model (Chapter 5). Following, The Exploratory Data Analysis (EDA) is executed and the results are shown (Chapter 6). Returned datasets from the EDA are used to create the Exploratory Model (Chapter 7). In addition, the results from the Exploratory Model and the assumptions made during the process of the creation of the model are validated by expert within AT (Chapter 8). After this chapter starts the research on identifying the challenges for the implementation of the model. This sections starts with discussing the involved stakeholders (Chapter 9). Following, the challenges of the implementation are highlighted (Chapter 10) Furthermore, the outcomes and the process of the research are discussed (Chapter 11). Then the research questions are answered in the conclusion (Chapter 12). Also, recommendations are given and future research is discussed (Chapter 13 & 14). Lastly, a critical reflection is given by elaborating on the chosen methods to create the model, the organizational structure, the process and the societal and scientific contribution of the research (Chapter 15).

2. Problem Statement

Section 2.1 explores the regulation of telecommunication in The Netherlands, followed by 2.2 which highlights the regulatory framework of AT. Section 2.3 starts with the introduction of the concept of an Exploratory Model.

2.1 Regulation on Telecommunication

Citizens, companies and governments trust on the availability of telecommunication services. It has become a commodity and may almost be considered a necessity of life. Telecommunication or ICT failures can cause an enormous social and economic impact on our society. Telecommunication Acts are created to compel telecommunication service providers and network providers to take precautionary measures. These measures should improve the reliability of the continuity and availability of services and networks. According to AT, continuity consists of three main parts: a duty of care, reporting obligation on incidents, and accessibility of the emergency number (Agentschap Telecom, 2018). The precautionary measures are established to minimize the risks of a discontinuity.

Individuals are always exposed to risks within their society, knowns as social risk (Ale,2009). These individuals are used to having access to certain services and commodities provided to them by a large number of companies and the government. Some examples are electricity, clean food, clean drinking water, and telecommunication services. The government regulates all companies within their country to ensure they maintain a high quality of services and products. This is principally to protect the individuals from being exposed to risk towards their wellbeing. Therefore, the risks towards the continuity of telecommunication services and networks in this research are considered social risks, and these risks need to be minimized by a regulator.

A regulator is in most of the cases an independent and impartial body assigned by the government, which monitors and enforces the compliance of laws and regulations by organizations. AT is not an independent body, but part of the Ministry of Economic Affairs and Climate Policy. It operates within the administrative law and is part of the executing power. Traditionally, a regulator is a governmental body, however, it can also be an independent organization. The number of independent regulators are increasing over time and this can be attributed to the increasing complexity of regulation, which now necessitates specific expertise (Irion & Radu). Furthermore, the independent opinion of experts is often necessary to make individual decisions.

2.2 Regulatory Framework

Regulatory framework of the regulation on telecommunication is discussed in three different section. Starting with risk-based regulation (2.2.1), followed by data-driven risk-based regulation (2.2.2) and the final part is about the issues and challenges of r by data-driven risk-based regulation(2.2.3).

2.2.1 Risk-based regulation

AT executes risk-based regulation within all the supervised fields. It is important to understand the distinction between regulation and risk-based regulation. Governmental regulation is the determinant of the delegated legislation which is created by subject-matter experts to enforce a statutory instrument (Rowe,1997). The financial resources of a regulator are based on a budget which is specified by the government. Due the increasing number of organizations and the increasing complexities of regulation, it is becoming increasingly more difficult to determine what must be regulated, as there are not enough resources to continuously monitor all companies.

Moreover, regulation is also constrained by administrative experience and human resources, which are both short in supply (Smith & Wellenius,1999). These factors influence the cost-effectiveness of regulation in a negative way. This heightens the importance of making the right decision on what to regulate in order to maximize the use of the available resources. To improve decision making on how to appoint resources, of regulation is risk analysis introduced. Regulation based on risk analysis is called risk-based regulation. Risk analysis is a consideration which consists of three different parts: the quantification of risk, the modelling of identified risks, and how to make decisions from those models (Vose, 2008).

Therefore, it is important to maintain insight into the (social) risks that play a role within the supervisory field Information Security. A simple way to define the risks is to determine the impact of an event and to multiply this by the probability it will occur (Dietz, 2011). AT focusses on different supervisory fields, which consists of multiple subjects. As a result, different risk analysis are created for all the subjects and used to distribute resources for regulation. This results in a numerous number of different analyses. These analyses are partly based on quantitative data and mainly based on qualitative data executed by the experts and inspectors. The next step for the decision making on regulation would be implementing a data-driven structure that uses Big Data, which is data-driven risk-based regulation.

The popularity of risk-based regulation has increased over the last years. Proponents argue that as it delivers interventions in proportion to chance, the regulatory resources are used efficiently and effectively, and facilitate robust governance. There are also critics who say that the potential benefits are being mitigated by the challenges of operationalizing risk-based governance (Rothstein, 2007). This study follows the opinions of the proponents of risk-based regulation. So a robust way of governance will improve the efficiency and effectiveness of the supervision within AT. The next step for risked based regulation is data-driven risk-based regulation. This would involve the AT creating a data infrastructure that allows the data scientists within the organization to create a data-driven process for risk-based regulation. The focus of this research will focus on the organizational conditions and the creation of an Exploratory Model that could be the first building step to create a data-driven risk model for risk-based regulation.

2.2.2 Data-driven risk-based regulation

The role of Big Data in our society has become more critical over the last decade. Companies are collecting as much data as possible to sell or analyse. Moreover, the use of algorithms in processing and analysing this information has increased. These algorithms advise and sometimes decide on how the data should be interpreted and which actions should be followed post-analysis. Individuals, groups and whole societies can be affected if there are any gaps between the design and operation of the algorithms. Another change that could have an impact is the understanding of the ethical implications of the algorithm (Mittelstadt, 2016). The creation of data-driven science instead of knowledge-driven science proposes entirely different ways to make sense of the economy, society and history. It is important for the users of data-driven science to know the difference is and that they are able to distinguish those two. To improve the efficiency of the data-driven tools and models. The way research is being conducted is being reinvented by disruptive innovations such as Big Data and algorithms (Kitchin, 2014).

Implementing Big Data in risk-based regulation is the next step to create data-driven regulation. This implementation necessitates a data infrastructure that allows people within the organization to share the data with each other. In addition, it is important to collect and manage the specific data that can be used to create data-driven models. The outcomes of data-driven models could support risk-based regulation by providing the regulators more insight on relevant risk criteria and the risks that certain organizations pose. This information supports decision-making with regards to the allocation of regulation, thereby creating a more cost-effective process. In addition to improving risk analysis, data be used to detect risks before risk analysis has been initiated, which is often referred to as risk detection.

Risk detection is often used by the police to profile suspicious people or by banks that use data to detect financial risks or frauds. Banks in Turkey often perform risk detection on small and medium enterprises, which are the companies that are most vulnerable to financial risk. This is because there is a higher risk of financial distress when the owners have an insufficient background of financial knowledge. Koyuncugil & Ozgulbas (2012) created an early warning system model based on data mining that detects financial risk. Data mining is an analysis technique which was developed during the growth of large databases. The method searches for data that is of value to the database owners by finding valuable data within the mass of stale data (Hand, 2012). The chi-square automatic interaction detection algorithm is used to create a model for small, medium enterprises. This is a method of database segmentation (McCarty & Hastak 2017). These data analysis methods can be used when initiating the transition from risk-based regulation to data-driven regulation.

2.2.3 Issues and challenges of data-driven risk-based regulation

The previous parts of this chapter were focused primarily on how the use of Big Data leads to data-based knowledge to inform intelligent decision-making. On the other hand, critics of Big Data have concerns regarding privacy and human resource scarcity. The concerns surrounding privacy date back to the start of electronic database management and remains an important topic in the public debate. Compounded with concerns

regarding the privacy of the information of individuals and organizations saved in databases, is the possibility of human connection towards the process of decision-making being erased due to the use of algorithms. This could subsequently result in a loss of information during the process (Hilbert, 2016).

There are not only issues with Big Data, but also some challenges when it comes to the use of Big Data and the creation of data-driven models. Eggimann (2017) discusses four different challenges: the ability to turn data into useful information, the ability to assess the quality of the data, the availability of the data within the organization, and the ability to create a process with a high cost-effectiveness ratio. All four challenges are relevant for this research. The first two challenges are the most important for the creation of the data-driven model and the last two challenges are the most important for the implementation of the created data-driven model.

Next to the issues on the use of Big Data are four major key issues raised by the use of data-driven models in regulatory bodies (King & Brennan, 2018). The first issue is that the use of algorithms and data metrics can be distorted when they are not applied accurately. This means that a certain level of education is needed to understand the creation and execution of the models. Educating employees is resource intensive because they may have no prior experience, and the creation and execution of the models is a complex activity. The second issue is the importance of a well-functioning working culture within the organization to heighten the acceptance towards moving into more data-driven regulation.

Furthermore, people that create data-driven models or those who use the outcomes of the model are prone to interpret the data in a way that fit pre-existing narratives. This increases the importance of policy makers understanding 'non-rationalities' from individuals, as a data-driven model could be rational, but the creator and the users may not always be rational (King & Brennan, 2018). Thus, it is important to the user and policy makers to acquire reliable and relevant data.

2.3 Exploratory Model

The current risk analyses that are conducted by standard systematics has developed over the years by experts and inspectors. This systematic is mainly focused on the consideration of the experts, which is focused on the technical risks within their supervisory field (Black & Baldwin, 2010). Every expert and inspector interprets the data in their own way and based on their own systematic and experience. This could result in inconsistencies between the risk analysis from different experts, or the risk analysis could be incomplete because some vital data is not used. A data-driven model that is used by all experts could create cohesiveness between the risk analysis executed by them. This could lead to less room for error and improve the cost-effectiveness of the regulation. The first step of creating a model is called an Exploratory Model.

This research focuses on the first Exploratory Model. An Exploratory Model is an experimental, research-based systems development method (Holak & Lehmann, 1990). It is often used to develop a computer system or a product. An Exploratory Model is a

prototype of the final model, but during the development, certain scenarios are run and the optimal scenario will be picked. The risk of this model is that the results of the Exploratory Model will be less optimal than people estimate because it is a prototype. The Exploratory Model created for the Supervision Division Information Security is a prototype for a risk detection model which could improve the supervision of the continuity of the telecommunication services and networks. Historical data should be used to create a risk detection model. This model is a tool that helps the regulator to be proactive by identifying risks that exceed a certain standard and where the societal risk mitigation is relevant.

Section 2.2.3 discusses four different challenges related to the creation of a data-driven model (Eggimann, 2017). The first two challenges are significant for the creation of the Exploratory Model of the risk detection model. Firstly, whoever is creating the model needs to be able to understand the meaning of the data, and extract useful information. The second challenge is the ability to assess the quality of the data. The following two challenges are relevant for the implementation of the model. The third challenge is on the availability of the data within the organization. The data is not always managed by the same people or even the same team within the organization. Different individuals that manage the data could have different standards on the needed quality of the data. The final challenge relates to the availability of the data within the organization. Laws and regulations can limit the availability of data within an agency.

Information and data sharing can be limited by institutional borders within an organization. For an agency, information one of the most valuable resources. Pardo (2008) researched the possibilities in finding ways to bring together information from different agencies with institutional boundaries and how it could be used to integrate the findings for use in solving public problems. This led to new insights on how to bring data together between organizations with the institutional boundaries and highlighted how difficult cross-boundary information sharing is. The research concluded with three main lessons and four recommendations. The lessons and recommendations are also applicable to institutional borders within an organizations, because the borders separates two different bodies. This can be agencies or divisions within an agency.

The lessons regarding the integration of cross boundary information sharing starts with the importance of interoperability of the agencies, and is followed by the necessity of a shift in the agency culture, and concludes with the need for policymakers to hold a central role in facilitating the integration (Pardo, 2008). The recommendations were found during case studies (Pardo, 2008), which are focused on the creation of capabilities and enterprise-wide mechanisms by government executives and policy makers. It is essential that they create governance structures, resource allocation models, scalable strategies, and non-crisis capacities.

Based on the evidence presented in the previous two chapters, this research aims to investigate the following missing elements for the transition from risk-based regulation to data-driven risk-based regulation:

- Identify the modeling challenges of creating a data-driven model to make the transition from risk-based regulation to data-driven risk-based regulation.
- Identify the challenges of the implementation of a data-driven model within an agency.

3. Case study

The potential for data-driven risk-based regulation to support the regulation process was presented in Chapter one, and compounded with the identification of the knowledge gap and the formulation of the research question. Chapter two provides more information about risk-based regulation and the promises and challenges of data-driven risk-based regulation. The problem statement is used for a specific case within AT to determine the modeling challenges of creating a data-driven model for a subject within the supervisory field and to identify the challenges of implementing that model. This chapter begins with 3.1 providing information about the Supervisory Field of Information Security and the subject continuity, followed by the challenges the inspectors experience within the supervision process (3.2), and section 3.3 introduces the available data within the supervision on Information Security.

3.1 The Supervisory Field Information Security

The main question from the Enforcement Policy Division of AT was to investigate whether it is possible to create a risk detection model. A risk detection model is a possible model of data-driven risk-based regulation. In addition to the interest in creating a risk detection model, there is also interest in investigating the possibility of creating a data-driven model that supports supervision. AT wants obtain information on how they should set up the process of collecting, structuring and analyzing data to create a data-driven model for the executed risk analysis. Another desired finding of the research is to compare the outcomes of the model with the outcomes of the risk analysis executed by the experts within AT in order to determine the influence of a data-driven risk detection model on the cost-effectiveness of regulation. Members of the Enforcement Policy Division within the AT requested these areas be researched. Therefore, the supervision field Information Security was selected as a case study within the case study of AT.

The Supervision Division Information Security executes risk-based regulation. This is a process to systemically set regulatory goals based on the expert's assessment of the risks of the organizations' that are regulatory subject of regulation, being non-compliant. These organizations are those under supervision by the regulator. The range of organizations that are subjects of regulation in the regulatory field of Information Security ranges from individuals to multinationals operating within the regulator's jurisdiction. Risk analyses are executed in order to assess the risks within the different supervisory fields. Based on the risks that are detected, priorities are set and based upon the selected priorities the supervisory resources are divided.

Providers of public electronic communications networks and services do have to comply with the regulations with regards to the obligation to technical and organizational measures in order to guarantee the continuity of their networks and services (the so called duty of care) and the obligation to report incidents to AT. AT is given the power to supervise the compliance of the providers with the obligations mentioned. Moreover, providers of public electronic communications services an networks do have the obligation to register with the Authority Consumer & Market (ACM). AT uses the ACM

Registration Register data in order to s the organizations that are supervisory targets for the Duty of care obligation. 2003 organizations that provide public electronic communications networks and services are registered in the ACM Registration in 2019 whereof around 1000 organizations a year are relevant for supervision on Information Security. Around fifty percent of the organizations registered to the ACM Registration are post services

The supervisory goal of the Supervision Division Information Security is to supervise that providers of telecommunications networks and services take their responsibility in reducing the vulnerability of the continuity of the telecommunication services and networks by taking the right mitigating measures. The public interest that is being protected by the continuity obligation of telecommunications providers is the reduction of vulnerabilities of their networks and services in order to protect the interest of the user of the services and the networks. Multiple stakeholders are involved in the process of regulation on the continuity of the networks and services. The first stakeholders are AT regulated organizations, which consists of two types of organizations; network providers which provide network and the telecommunication services providers which provide only the telecommunication services. AT regulates these organizations based on the national policy from the Ministry of Economic Affairs and Climate, which is in part dictated by the policy set by the European Union.

Risk-based regulation is executed within the AT for the Supervision Division Information Security. The goal for the Supervision Division Information Security is to reduce the vulnerability of the continuity of the telecommunication services and networks and thereby protect the interest of the user of the services and the networks. All organizations within the Supervisory Field of Information Security are regulated on the continuity of the telecommunication services or network they provide. Risk-based regulation within the Supervision Division Information Security is information driven. Risk is defined by the execution of a risk analysis, which is an assessment of quantitative and qualitative data. The main source of data for the Supervision Division Information Security is a standard survey which covers all four main subjects and is sent to all supervised organizations. These subjects are authorized wiretapping, security of telecommunication data, network and service continuity, and maintenance of privacy This is done with a frequency of a maximum of three times per ten years, and a minimum of once per ten years. All organizations receive the survey when they register to the ACM. The quantitative data used during the regulation process are the outcomes of this and other surveys.

3.2 Legal framework reporting obligation failures

The supervised organizations within the Supervisory Field of Information Security are obligated to report failures that caused interruptions of the continuity of their services or networks. This is defined by legislation set by the national government of The Netherlands and consists of three main parts (Agentschap Telecom, 2019). The first part is the reporting obligation. Providers of public electronic communication networks and services are obligated to report failures that caused a breach of security and integrity, and substantially interrupted the continuity of their networks or services. The reporting is provided to the Minister of Economic Affairs and Climate. In the event of a disruption of

continuity, it is sufficient for the network operator to report the relevant breach of security or loss of integrity. The service providers who use the network and are also affected by the continuity disturbance do not have to report the disruption of the continuity. The legislation demands that the providers are obliged to appoint an official who is responsible for taking care of the report and who acts as the first point of contact for the reporting center in the event of a security breach or loss of integrity.

The Decree (Tempelman,2012) on continuity of public electronic communications law and services stipulates that the notification must at least contain the following:

1. The starting time of the breach of security or loss of integrity.
2. The nature and extent of the breach of security or the loss of integrity.
3. The affected network and the affected services caused by the breach of security or the loss of integrity.
4. A forecast of the recovery time.

The Minister must report the information regarding the breaches of security and losses of integrity to European Union Agency for Cybersecurity (ENISA) and the European Commission. That is why a variety of information is collected about the breaches of security and losses of integrity. On the basis of the Continuity of Electronic Communication Networks and Services Decree, providers are also obligated to provide the Minister of Economic Affairs and Climate with a report on a breach of security or a loss of integrity, within four weeks after the termination of the breach with the following information:

1. When the breach of security or the loss of integrity ended
2. What measures have been taken to end the breach of security or the loss of integrity
3. What measures have been taken to prevent a recurrence of the breach of security or the loss of integrity

The obligation to report does not exist with regard to every breach of security or loss of integrity, but only if continuity is '**interrupted to a significant degree**' as a result. The implementation practice will have to show how this criterion is fulfilled. According to the law, the minister can determine a lower limit of infringements or losses that must in any case be reported. The Continuity of Public Electronic Communication Networks and Services Decree states that, in line with the 'Technical Guidelines' drawn up for this purpose by ENISA, the nature and extent of the infringement or loss and the possible consequences thereof are always central. The purpose of the duty to report is to limit the intended adverse infringements for users and interconnected network as much as possible.

The second part of the legislation is the information obligation. In order to be able to assess the security and integrity of their public electronic communication networks and services, the providers involved must provide the minister with all information required for this purpose, at his request. The minister must inform the ENISA and the NRAs in the other Member States about each report. In addition, each year the minister must report

to the European Commission and ENISA on the notifications received and the measures taken in this regard.

The final part of the legislation is the disclosure of the reports. The minister may disclose a breach of security or a loss of integrity or have it made public by the provider. Objections and appeals are possible against a decision to do so. The disclosure of any confidential information relating to the network or services involved may, however, be omitted from the disclosure.

3.3 The Role of Data-Driven Risk-Based Regulation

The complexity of regulating providers of public electronic communications networks and services is constantly evolving in synergy with the risks that change over time based on a large number of variables. For example, one of these variables is technological development. It has a significant influence on the risks within the supervisory field Information Security. As prefaced in section 1.1, technology related to telecommunication changes each year and the users may not always be aware of the risks related to these changes. Wireless internet has become stronger and more stable, which has led to more people being connected to the wireless internet. This means that the continuity of the telecommunication services and networks will become more critical in the future. A relevant example of this is the development of virtualizing networks. This is the process of combining network functionalities, software network resources, and the functions of hardware into a single software-based entity. Every software developer with the right skills can generate virtual networks on their servers and sell them to telecommunication users. These two changes have an impact on the regulation of the continuity of the telecommunication services and networks.

Such examples of technological development make it difficult for AT to effectively regulate telecommunication. The experts and inspectors of Information Security need to execute their supervision task and keep up with the new technologies related to Information Security. As these developments are becoming more complex, keeping up with such changes is more resource intensive, and therefore, costly. In conjunction with the increasing complexity of regulation it will be harder to supervise the same percentage of organizations in the future due to the increase of the number of organizations. As technological development has made it possible for individuals to create virtual networks and other services that could be provided to customers, the number of organizations that need to be supervised also increases. As a result, it will become more difficult for the Supervision Division to allocate resources.

Thus, the increase in organizations and evolving complexity of the regulated services and networks makes it harder for AT to execute risk-based regulation. This leads to a more difficult process of decision making with regards to the division of resources. As a result, AT is searching for new methods to improve the cost-efficiency of the regulation on Information Security. The importance of data is recognized by AT and they sought to investigate if there are possibilities to use the data within their organization to create a data-driven risk detection model. In addition to the possibilities with the available data,

there is also interest in what improvements could be made to the data quality and the process of collecting and sharing data within the organization.

3.4 The Case Specific Research Question

The Supervision Division Information Security are interested in the possibilities of a data-driven model to support the process of regulation. There is an additional interest in obtaining knowledge about the possible uses of the data they collect, by investigating the quality of the data, the interpretation of the data, and what the challenges are for creating a data-driven model with the available data. Another area of interest that is comparing the similarities between two different factors. The first factor is the comparison between the important variables within the model and the important variables according to the experts, and the second factor is comparing the outcome of the models and the outcome of the risk analysis executed by the inspectors.

A total of three separate bodies within AT are responsible for collecting and managing four different sources of data that pivot upon subject continuity. The three bodies are: the Supervision Division Information Security; the Monitoring and Analysis Centre (M&AC), which specialises in managing and monitoring data; and the Spectrum Division Continuity, which is the executive body of the regulation on Information Security. The first two sources of data are collected by the Supervision Division on Information Security. This data is comprised of two different surveys; the first survey assesses all the main subjects of Information Security, and the second survey assesses subject continuity. Both surveys collect information from all supervised organizations and are used during the executing of the risk analysis. Whilst the Supervision Division Information Security collects the data, the management and analysis of this data is handled by experts in the M&AC. They ensure that the database with the data from the surveys remains up to date for active organizations. The third source of data is both collected and managed by the M&AC and only includes information from some organizations. This data is sourced from social media mentions regarding the continuity failures of telecommunication networks and services of certain companies. The final source of data is collected and managed by the Spectrum Division Continuity, and originates from the regulated organizations reporting incidents that caused malfunctions and discontinuities.

The data from the Spectrum Division Continuity is normally not available and shared with the Supervision Division on Information Security. The Exploratory Model uses data from both teams, and the outcomes of the model will be used by the Supervision Division Information Security. This means that there is an institutional border for information sharing, which is called a Chinese Wall (Brewer, 1998). All the require data for the creation of the Exploratory Model needs to be available in order to create a valid model. A Chinese Wall is usually implemented within an agency, by the national government, in order to keep certain divisions within the agency separated from each other. The case of regulating Information Security consists of two different parts; the Supervision Division and the Spectrum Division. To keep these two sections separated, a Chinese Wall can be set between the two divisions. The reason for this separation is that inspectors from the supervision division could create a bias based on certain information provided by the Spectrum Division. This institutional border results in a heightened importance regarding

who the owner of the model is, as it is important that the relevant data is accessible to the owner. In this case is the institutional border not set by the national government but by AT, which makes it an institutional border instead of a Chinese Wall.

Based on the specifics of the case found during the case study, the following **research question** and **problem statement** is formulated.

How could a data-driven risk analysis for risk-based regulation on the subject continuity, based on outcomes of surveys and data from malfunctions and failures of the regulated organizations on the continuity of the networks and services, support the risk-based regulation executed by experts within the Supervision Division Information Security?

3.5 The Dichotomy of The Case Study

The problem statement consists of two different elements; the modeling component and the implementation component. There are four challenges for the transition from risk-based regulation to data-driven risk-based regulation and consist of: The ability to turn data into useful information; the ability to assess the quality of the data; the data being available within the organization; and finally, the ability to create a process with a high cost-effectiveness ratio (Eggiman, 2017). The modeling component of the problem statement, which is to identify the modeling challenges of creating a data-driven model, is in line with the first two challenges. This is because, for this case, it will be crucial to turn the data related to the subject continuity into useful information and thereby indicate the meaning of the data and the value of the data. This is important for the validation and verification of the model. The second challenge relates to the quality of the data and what the indication could be for the validation of the model based on the quality of the model. In conjunction to the two challenges from Eggiman (2017) there is a third case specific challenge for the creation of the model, which concerns the indication of the modelling technique that suits the available data.

The final two challenges from the study from Eggiman (2017) are in line with the identification of the challenges of the implementation of a data-driven model on the subject continuity within AT. Data on the subject continuity is found within different divisions of AT discussed in section 3.3. It is important that all the data for the subject continuity is available in order to be able to create and implement the model. The last challenge centres upon the acceptance of the model within the organization, and ensuring that all the experts within the different teams within AT are on the same page. The main challenge with regards to the acceptance of the model will be to determine who the owner will be, and to decide the relation between the teams and the model. A high level of acceptance will create a high cost-effectiveness ratio for the regulation on the continuity of the telecommunication services and networks.

4. Research approach

The previous chapter defines the problem statement to provide an overview of the addressed missing elements within this research. The aim of this chapter is to explain the research approach. This begins with an explanation of the research approach and objective (4.1) to specify the deliverable of this research and how it will be reached. In section 4.2 the sub research questions are discussed, followed by the used research methods in section 4.3. Finally, this chapter concludes with a discussion on the required data which is collected(4.4).

4.1 Sub Research Questions

This section elaborates the **sub-questions** of the leading research question and the research methods conducted per sub-question.

1. *What is the quality of the available quantitative data within the supervisory field Information Security? What is the significance of the data?*

The available data within the Supervisory Field of Information Security on the subject continuity will be analyzed and explored. This is executed to gain a better understanding of the data and the quality of the data and to collect information to decide what the best modelling technique is for these datasets. On the other hand it is possible to gain information on how the stakeholders within AT collect, process, and manage the data, based on the quality and the significance of the data. During the data analysis, it is also decided if the initial datasets from the available data sources are used for the creation of the model or a subset of the data. This research question is in line with the first two challenges stemming from the study of Eggimann (2017), which explores how to assess the quality of the data and turn data into useful information.

2. *What is the preferred modeling technique to create a model by taking the quantitative data into consideration? Is it possible to create the model with the available data?*

Find an algorithm or combine several algorithms to develop a modeling method that could support the process of risk-based regulation executed by the Supervision Division Information Security on the subject continuity. Identify why this method is suitable to be developed into a risk detection model and why it is data-driven. Furthermore it is important to understand the role of the different datasets on the outcome of the modeling method.

3. *How to validate and verificate the created model?*

Several assumptions are made regarding the meaning of the data, which is one possible method to turn data into useful information. In addition, there certain decisions to be made on the data to improve the quality of the data. These assumptions and decisions are made during the second and the third sub question and need to be verified by experts within AT

who manage the used datasets. This results in the formation of final datasets that will be used for the modelling method.

Along with the verification, the model must also be validated. The validation is done by experts within the Supervision Division Information Security specialized on the continuity of the telecommunication network and services and by experts from different supervisory fields. Experts within the Supervision Division Information Security validate the outcome of the model. Validation by experts from different supervisory fields will be done to investigate if the modeling method is applicable within different datasets in order to test how robust the model is.

4. *Whom are the stakeholders involved in the regulation on the subject continuity within the Supervisory Field of Information Security and what is their role?*

After the creation and validation of the Exploratory Model will the focus of the research shift towards the implementation of the Exploratory Model. The first step in understanding the organizational structure is identifying the involved stakeholders. The key stakeholders will be identified, and their role and interests towards the regulation of information security will be indicated. In addition, it is important to understand which data the stakeholders within AT collect and what the role of that data is during the execution of risk-based regulation.

5. *What are the organizational requirements to implement the created model?*

It is important to determine if it is possible to create a data infrastructure that enables the use of a data-driven model within the organization. To understand what the organizational requirements are for the implementation of the data-driven Exploratory Model. In addition the data structure is the acceptance, interest, and ownership towards the model from the different stakeholders within AT. This means that it is important to define the organizational requirements and challenge for the implementation of a data-driven risk detection model. This sub question provides the answer for the case study to the final two challenges posed by Eggimann (2017), which probes whether the data within the organization is available for the creation of the model , and if it is possible to create a process within the organization that enables the use of the model to create a high cost-effectiveness ratio.

4.2 Methodology sub research questions

Four main research methods are used in this study: a literature review, interviewing experts, Exploratory Data Analysis, and Exploratory Modelling. Supplementing these main research methods is theory about actor strategy models, a clustering algorithm, and theory on calculation correlation. These methods are used to support the main research questions in order to provide an answer to the sub research questions.

4.2.1 Methodology for the first sub question

Exploratory Data Analysis

An Exploratory Data Analysis (EDA) is executed in order to define the quality and significance of the available data within the organization on the subject continuity of the telecommunication services and network providers. The steps of the EDA are based on the article by Yu (2017). He argues that EDA is a strategy of data analysis that emphasizes maintaining an open mind to alternative possibilities, and that EDA consists out of four basic elements. This idea was based on the outline for EDA set by Velleman and Hoaglin (1981). The outline consists out of four basic elements called the four Rs: residual, reexpression, resistant, revelation. Reexpression could be referred to as data transformation and revelation could be referred to as data visualisation. The residual analysis is seen as data = fit + residual or data = fit + error by EDA. This part is also referred to as data cleaning. The residual of the error are the values that deviate from the expected value and the fit are the expected values. The model's adequacy can be assessed by investigating and examining the residuals. After reducing the residuals, the data could be reexpressed to improve the interpretability of the data. This could be a transformation based on a mathematical function or a transformation in the kind of data. The next step is executing some parametric tests on the data to obtain more information. The final step is visualizing the data or outcomes of the resistant.

4.2.2 Methodology for the second sub question

Exploratory Model

Exploratory Modelling and Analysis (EMA) supports the understanding of uncertain and complex systems (Kwakkel & Pruyt, 2013). This method allows the researcher to experiment with different datasets as input in order to discern the influence of the datasets on the outcomes of interest. The determination of different risk criteria and risk profiles for the risk detection model is based on different input data. The creation of the Exploratory Model supports the evaluation of the conceptual uncertainties and the sensitivity of outcomes to the different input data. An Exploratory Model is an experimental, research-based systems development method (Holak & Lehmann, 1990). The Exploratory Model created for the supervision on the continuity of the telecommunication services and network is a prototype for the final risk detection model. This study will develop a method that can be used to create a model. The creation method consists out of two different phases, starting with clustering and followed by finding correlations.

An overview of the Exploratory Model is shown in Figure 1. The creation of the model consists out of two different parts. Starting with the formation of clusters based on the survey scores from the supervised organizations. The survey scores are used as input data and a cluster algorithm called K-means clustering is executed for on these datasets. This algorithm created K clusters for the dataset, but how efficient was the cluster algorithm on dataset to the creation of cluster. To define this is the average difference calculated per cluster between the scores from all the organizations within a cluster. A lower difference means that the answers from the organizations within the cluster are

more similar to each other. The assumptions is made that when the average difference is smaller the effectiveness of the algorithm is higher.

The second part of the Exploratory Model is defining the different risk profiles. One cluster is equal to one risk profile but the specifics of the profile needs to be defined. These specifics are calculated based on the correlations between the organizations within the clusters and the data from social media and the Reporting Desk. Which are two data sources about information of failures.

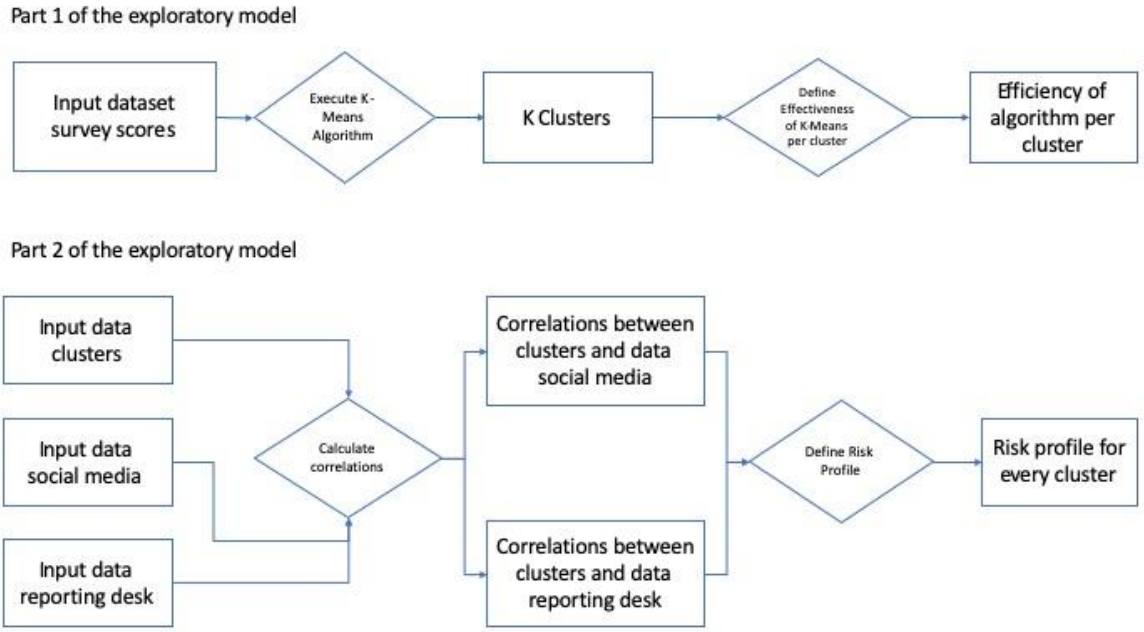


Figure 1: An overview of the steps taken to create the exploratory model

Clustering

Clustering analysis makes quantitative comparisons of multiple characteristics to discover if multiple objects within a set will fall into different subsets (Anil,2009). It is used to develop an algorithm that discovers clusters within the set of regulated organizations. Clustering the organizations will allow a more in depth analysis on the organizations within the clusters. The composition of the clusters illuminates some properties of the organizations. The similarities will not be taken into consideration for the research method in the second phase. Detecting groups in the data can help regulation to establish the appropriate strategies to regulate the organizations under supervision.

Data Clustering has been used for three main purposes:

- **Underlying structure**: to gain insight into data and find notable features.
- **Natural classification**: to identify the degree of similarity among the objects
- **Compression**: to summarize and organize the data

K-means clusters

The modelling method K-means clustering is used as the clustering algorithm. It discovers the natural grouping of a set of objects or points (Hartigan & Wong,1979). The K-means algorithm finds K groups based on a measurement method that finds similarities between objects within a group. The measurement method used is the Euclidean distance and the formula of this distance is shown in Figure 2. Objects from different groups have a low number of similarities and objects in the same group have multiple similarities. The total number of objects is equal to n. Therefore, n different objects are split into K different clusters. Cluster analysis can be used in any discipline that has a set of objects and wants to analyse multivariate data.

$$Distance[(x,y),(a,b)] = \sqrt{(x-a)^2 + (x-b)^2}$$

Figure 2: The formula for the Euclidean distance

So the input values of the model are the value of K and n different points. The first step is that K different centroids are placed at random locations in the defined space. For all n points is the distance calculated from the point to the different centroids. The points are assigned to the centroid where the distance between the point and the centroid is the smallest. The second step of the algorithm is to calculate for all the centroids the new positions of the centroid. This position is the mean of all the points assigned to that cluster. These are the two steps of the algorithm to create the clusters and are iterated till the optimal value of the centroids in found. The steps of the K-means clustering algorithm can be found in Figure 3.

For this algorithm, multiple values for K are possible, which stand for the number of clusters that are created by the algorithm. After obtaining the data, two important things need to be done before executing the algorithm. The first part is transposing the multivariate data into points and the second is to define the optimal value of K. Finding the optimal value of K can be achieved by a method called the elbow method.

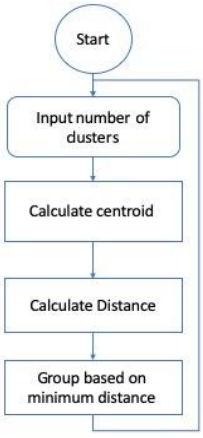


Figure 3: an overview of the K-means clustering algorithm

Elbow Method
The Elbow method sets the square error against different values of K. In other words, the percentage of variance explained as a function of the number of clusters (Bholowalia & Kumar,2014). When K increases, the average number of objects will decrease per cluster. This means that the average distortion will decrease when K increases, resulting in fewer objects closer to the centroid. The elbow point is the point where the distortion declines the most. At this point, the marginal gain of creating an extra cluster drops drastically. The idea is that adding another cluster doesn't provide much better modelling of the data. The value of K where the elbow point is located will be the K selected for the K-mean clustering algorithm. The range of K for the Elbow method is selected from 1 till 9 for all the datasets. The outcomes of the elbow method for the different datasets are demonstrated in the results section of the K-means in Chapter 7.

Correlation
Correlation is a method that can be used to measure two variables (X and Y) for each case, and to quantify the relation between these variables (Zhou, 2011). No assumption is made on whether any of the variables depend on other variables within the dataset. It describes only the association between them and it does not study the relationship between variables, which is done in regression. For this study, the X variables are the formed clusters and the Y variables consist of information about failure that occurred in the past for the telecommunication service and network providers within the cluster. This is calculated to study the frequency at which failures occur, and when a failure occurs, what the properties are for these failures.

4.2.3 Methodology for the third sub questions
The validation and verification of the model is executed through expert interview. For the validation process, the outcomes of the model and the risk criteria shown to the expert within the supervisory field Information Security. They are asked for an assessment on whether they see similarities in the results of the exploratory modeling and the results of the risk analysis that they execute. The interviews for the verification are not focused on the creation of the model, but on the assumptions and decisions made regarding the data. This allows for the investigation of whether different assumptions and decisions need to be formulated in the future for the creation of the risk detection model.

4.2.4 Methodology for the fourth sub question
Identifying the role of the involved stakeholders in the regulation on the subject continuity within the Supervisory Field of Information Security is achieved through literature review of internal documents within AT, interviews with experts on the subject, and participatory observation. This provides information about two different groups of stakeholders, which are the external and internal stakeholders. The external stakeholders are governmental bodies, regulated organizations, and AT as an agency involved in the regulation of Information Security with a focus on the continuity of the telecommunication services and network. This results in a better understanding of their roles in the process of regulation. The internal stakeholders are the different divisions within AT that are involved during the

execution of risk-based regulation on the subject continuity, and points to which divisions are involved in data-driven risk-based regulation within AT.

4.2.5 Methodology for the fifth sub question

The methods used for the final sub question are expert interviews and a review of internal documents from AT. These documents concern the legislation and regulation on the continuity of telecommunication networks and services and investigate whether there are any legal restrictions on the implementation of the model. The interviews are conducted with experts within AT on the data sharing and ownership of the model, allowing for an assessment of whether the current organizational structure allows the model to support the regulation process. These research methods will provide the organizational requirements of the implementation of the model.

5. The available data

The aim of this chapter is to explain the different datasets that are used for Exploratory Data Analysis. Four different datasets from three different sources within AT are used to create the model. A description of each dataset is provided, along with an explanation of the relevant internal and external stakeholders. Section 5.1 explains the dataset comprised of answers to a survey on the subject Information Security. Following this, the survey on subject continuity is discussed (5.2). Section 5.3 overviews data from the report desk, which consists of data regarding reported failures of the continuity of telecommunication networks and services by regulated organizations. Lastly, the collected data from social media concerning failures on the continuity of telecommunication networks and services by the regulated organizations is discussed(5.4).

5.1 Dataset outcomes survey on Information Security

The first data source is the database of the Supervision division on Information Security, which is managed by the M&AC. This data is composed of information from all the providers of public electronic communications networks and services that are listed to ACM. The first dataset from this source is data from a survey on the subject of Information Security. Questions within this survey are related to all the main subjects of Information Security (Agentschap Telecom, 2018). Not all the questions are answered by the supervised organizations. The Supervision Division Information Security sends a survey to all the new organizations registered to the ACM. The survey consists of 28 questions, whereof the first four questions are general questions about the organization and 24 questions assess the main subjects of Information Security. These 24 questions comprise of the following subjects and frequencies:

- Six questions on authorized wiretapping
- Ten questions on keeping privacy
- One question on the accessibility of the 112 emergency number
- Seven questions on continuity

5.2 Dataset outcomes survey on continuity

The second dataset are outcomes of a separate survey. This survey focuses on the subject continuity and consists of 29 multiple choice questions (Agentschap Telecom, 2013). The title of the survey is called The Supervision of Duty of Care and The Obligation to Report on Continuity Incidents, the Baseline Measurement. The purpose of the survey is to assess if the telecommunication services and network providers comply with laws and regulations. Not all the questions are answered by the supervised organizations. New regulation was set in 2012 by the European Union (EU) (Agentschap Telecom, 2012). The EU introduced two new laws to stimulate the trust of business and society in electronic communication, which are the duty of care of the continuity of the telecommunication services and networks. The survey was introduced following the implementation of these laws.

Duty of care continuity
Providers of public electronic communication networks and public electronic communication services are obliged to take appropriate technical and organizational measures to control the risks of the safety and integrity of networks and / or services. In the event of disruptions or power outages, there is a legal obligation for providers of public telephone services to take all necessary measures. The purpose of this is to guarantee the continuity and availability of networks and / or services as much as possible. This concerns breaches of security and a loss of integrity of the continuity of the network and / or service. It does not concern breaches of personal data (privacy).

Reporting obligation continuity
Providers are obliged to report an incident to AT in the event of breaches of security and / or a (partial) loss of integrity.

Goal of the survey
AT mapped out the level of compliance with the duty of care and reporting continuity with the survey of the baseline measurement. This measurement serves to make the current level of compliance visible. By comparing the baseline measurement and follow up measurement, the improvements in compliance can be made visible. In addition, the outcome of this investigation serves as input for the risk analysis executed by the experts within the Supervision Division Information Security, especially for future inspections and audits. The outcomes can also be used to determine the areas for improvement that must be tackled in order to get the market moving and improve compliance.

This report provides insight into the achievement of the policy objectives of the duty of care and reporting continuity at the time of entry of the legislation and regulations. This regulation is new, so it is important that feedback measuring the levels of compliance are provided to policymakers. The outcome of the baseline measurement provides insight into possible areas of significance in the implementation of the duty of care and reporting continuity.

Finally, AT returns the result of the baseline measurement to the market, which includes the providers participated in the survey. On the basis of the anonymous outcome of the

baseline measurement, providers can see to what extent the market complies with current legislation and regulations and what the areas for improvement are for the regulated organizations. This also makes it clear to the market where the emphasis is on supervision.

The regulated providers
The duty of care and reporting continuity applies to all providers in the Netherlands. Three target groups emerge from the target group analysis of AT Netherlands:

- Network providers
- Service providers with their own network
- Service providers without their own network.

Providers are obliged to be included in the register. However, providers who have failed to do so also fall within the scope of the duty of care and reporting continuity.

5.3 Data from social media

The final data source is data from social media. This data is scraped from the social media of 38 of the 2003 regulated organizations. An expert within the M&AC division within AT advised use of the website of “all storingen” to identify these organizations. This is a website that tracks failures from not only telecommunication providers and network providers, but other service providers related to these providers in The Netherlands. As such, gaming or emailing platforms. The website utilises social media to track failures of different systems, however, it is not possible to download the data of the failures This website is used so that it is guaranteed that there is information on social media about failures for these organizations, as it is not possible to manually check for all 2003 organizations whether there is data available on failures on social media. As a result, not all the organizations from the ACM list are checked for mentions on social media regarding a failure.

AT uses software to collect data about failures mentioned on social media in order to track the number of failures per organization. The software Coosto is used by AT to collect data from social media, regarding failures. A query can be entered, and Coosto can then measure how many times this query is presented within the text of a social media post per day. The software counts the post both when the query words consecutively follow each other in the text, in addition to when they are separated from each other within a block of text. All the queries begin with “storing”, the Dutch word for failure, and are then followed by the name of the organization. The collected data spans from the first of January of 2012 and ends at the first of May 2019. It shows the number of times the query was mentioned per day in a social media post. AT has set a threshold for the number of times an incident is mentioned on social media for 20 different organizations. These thresholds are selected by experts based on their subjective intuition, with no established selection process.

5.4 Dataset from Reporting Desk

The second source is data from the Reporting Desk of the Spectrum Division Continuity. The law indicates that reporting a failure is voluntarily and organizations are only required to report a failure when they think it is significant. The data encompasses the failures and the following information is viewable in database from the Reporting Desk.

- Provider
  - Labeled
- Year
- Did it hit the alarm number?
  - Yes or no
- Affected service
  - E.g. fixed telephony, mobile telephony, SMS or Internet
- Number of affected customers
- Impact
  - National, regional or international
- The duration of the failure
  - In Hours
- The cause
  - E.g. nature, people, attack, hardware and extern organization
- Hours immediately

6. Exploratory Data Analysis

Before the exploratory model is created an Exploratory Data Analysis (EDA) is executed on the data. As extrapolated in Section 4.3.2, EDA is used to determine the quality and significance of the available data. This allows for a summary of the main characteristics of the data and to investigate what kind of information the data can provide before the modelling is started. The execution of the EDA resulted in 6 different versions of the initial datasets. These were two versions of the dataset with the outcomes of the survey on the subject Information Security, and two versions of the dataset with the outcomes of the survey on the subject continuity. On top of these four datasets are two additional datasets created by taking the organizations that are present in both datasets and linking the data from both datasets for these organizations. The reason why two different versions of the datasets are used as input for the exploratory modeling is to investigate the influence of the changes made to the datasets on the outcomes of the exploratory modelling.

The first version of the datasets after the EDA is created from the initial dataset provided by AT, which is comprised of the data discussed in the previous chapter . Some information from the initial datasets are removed due to unusability. The reasons for the data being unusable are; the data is not related to the subject, or a question is answered by only two or less organizations. The decisions made based on those two restrictions is considered the first part of the data cleaning. The second version of all the datasets incorporates the two new conditions in order to make a reduced version of the original datasets. This is followed by the second component of data cleaning. Here, all the questions that were answered the same by more than 90 percent of the organizations that replied to the survey are removed by AT . This condition is created to investigate what the influence is of the questions that were answered the same by the organizations and to investigate which questions will be removed when this condition is set. The datasets that resulted after the second part of the data cleaning are not used as input for the exploratory model. The last and third component of the data cleaning is based on the condition that all data from questions that were answered by less than 50 percent of the organizations are removed from the dataset. This condition is set to investigate what the influence is of removing unanswered question on the execution of the k means algorithm, and to investigate which questions will be removed. The datasets returned after the third step of data cleaning are considered the second versions of the datasets.

The EDA consists out of four basic elements starting with the residual in section 6.1,which is highlights the minimization of the unexpected values within the datasets of the survey scores. Followed by the re expression of the data also known as data transformation (6.2). Section 6.3 shows the resistance which are the outcomes of several parametric tests executed on the datasets with the survey scores. Finally, the data is visualized in the revelation section of the EDA (6.4)

6.1 Residual

The residual are the values in the dataset that deviate from the expected value. These values should be minimized to improve the performance of the algorithms executed on the data. Therefore, multiple decisions are made regarding the data before the algorithm is executed. Another word for minimizing the residuals is data cleaning. Data cleaning can executed in multiple different ways to investigate if it leads to a difference in the final outcome. That is why the residual of the dataset is minimized on two different occasions for all the datasets. Both times, different restrictions are applied to clean the data, and when these conditions are met certain actions are executed upon the dataset.

Two different actions are used to minimize the residual of the dataset. The first action is removing columns based on the answers provided by the responders and the second action is dividing the column into multiple columns as some organizations responded with multiple answers to a single question. Splitting the columns is performed to avoid a large number of possible answers for some of the questions. Figure 4 shows the different parts of the data cleaning and the returned datasets. The first part of data cleaning is based on information collected during conversations with experts from AT. For this part, no conditions are set. The data that is not relevant for the creation of Exploratory Model is removed from the dataset during the first part of data cleaning and the returned datasets after the first part of the data cleaning are called the first version of the datasets. These datasets are used for the first part of the K-means clustering during the creation of the exploratory model. The second and third part of the cleaning data are based on additional conditions. The returned dataset after the second part is not used for the exploratory modelling, however, the returned dataset after the third part is used for the exploratory modelling. The reason for multiple elements of data cleaning is to investigate if the answers from the organizations to the survey will be more similar to each other after the different parts of data cleaning. More similar answers will lead to higher effectiveness of the K means algorithm.

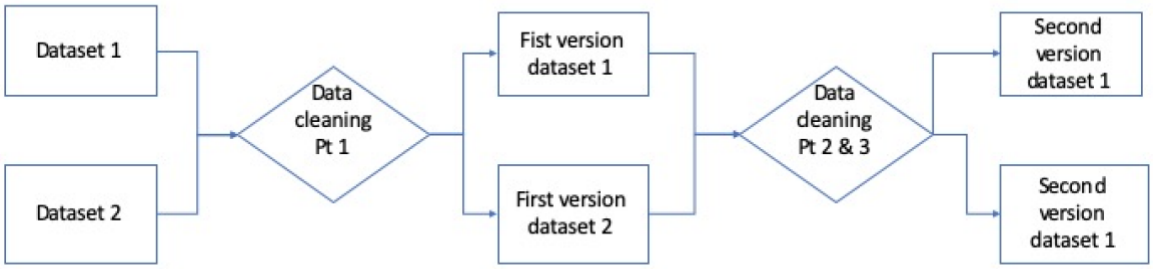


Figure 4: The main steps of the data cleaning for the datasets with the survey scores on the subjects Information Security and Continuity

6.1.1 Data cleaning part 1

The first discussed action is the elimination of specific columns. Every column represents one question from the survey or other information related to the organization. Some of the columns are removed because the data is not relevant for the creation of the exploratory model. Another reason for removing a question is because the question is unanswered by almost all of the organizations. The threshold here is set at 2 and this is defined with a visual test rather than a parametric test. This means that the columns are removed when less than ten percent of the organizations answered the question. Thus, the first part of the data cleaning was focused on removing the columns with column names that are uninteresting to the research, and questions that were unanswered.

Survey with questions on the subject Information Security

The dataset with the information about the survey on the subject Information Security is cleaned first. The initial version of the dataset consisted of 91 columns. In this dataset, every column represents information about organizations that answered the survey. A single cell in the data frame represents a question or a multiple choice answer of a question answered by an organization. Some of the questions were answered with multiple possible answers. In that case, the question split into multiple columns. These columns are now seen as separate questions.

The columns with the answers to question 12 and 22 kept there structure in the dataset with a column representing a single multiple choice answer. These questions kept their structure because most organizations answered multiple possible answers to these questions. Take for example, question number 12. This question consists of ten different answers in total so this means that there would be ten different answers possible if one organization supplies only one services. Most of the organizations provide or supply more than one service. So the question is split to keep the dataset organized and easy to understand. Another reason why the question is split is to create an easier way to cluster the answer to these questions. After the first part of the data cleaning, multiple columns were erased and this resulted into a new dataset with 41 columns in total. The initial and the new column names of the dataset of can be found in the appendix.

Survey with questions on the subject continuity

The second dataset is the data from the survey on the subject continuity. The initial dataset consisted of 41 columns and the resulting dataset after the first step of data cleaning consisted of 33 questions. The columns are removed based on judgement if the data in those columns are relevant to the research. All the initial column names and the column names after the first step of data cleaning can be viewed in the appendix. None of the columns or questions were split into multiple columns for this dataset.

6.1.2 Data cleaning part 2

The datasets that resulted from the first step of data cleaning are used as the starting point for the second part of data cleaning. Only the action of removing columns is used during this part. A condition is set to define which columns should be removed. This condition is a threshold on the percent of organizations with the same response for one question. This threshold is set at 90 percent for both datasets. So if a column within the dataset comprises of 90 percent of the same answers, the column is erased from the dataset. Not answering a question is also seen as an answer, so when 90 percent did not answer the question, the question is removed. The threshold is set at 90 percent because the datasets consists of 31 and 33 respondents, respectively. So 90 percent or higher indicates that three or less organizations provided different responses to the rest.

This condition is applied because columns with the same value in all the cells do not have a large influence on cluster forming done by the K Means algorithm. Thus, the idea is that the effectiveness of the K Means algorithm will increase when columns are removed with almost the same values. Due to the minimal changes, the datasets returned after this part of the data cleaning are not used as input for the K-means algorithm. These datasets are used as the datasets for the third part of the data cleaning.

Survey with questions on the subject Information Security

Within this survey, two questions are answered exactly the same and nine other questions are answered the same for more than 90 percent. The threshold is set at 90 percent so eleven columns are removed from the dataset, resulting in a dataset with 31 columns and 30 rows.

Survey with questions on the subject continuity

There are two questions within this survey that are answered exactly the same and one other question is answered the same for more than 90 percent of organizations. These three columns are removed from the dataset, resulting in a dataset with 30 columns and 31 rows.

6.1.3 Data cleaning part 3

A new condition is set for the final part of the data cleaning to investigate the influence of unanswered questions on the effectiveness of the K-Mean algorithm. The new threshold is set on the percentage of organizations that gave no answer to a question. This threshold is set at a percentage of 50. So all the questions that are unanswered by more than 50 percent of the organizations are removed from the datasets. This is done to reduce the residual which are the unexpected values. The unexpected values in this case are the number of columns with a high percentage of organizations that did not answer the question. To create a dataset with more similarities between the answers given by the organizations.

Survey with questions on the subject Information Security

There are eleven questions within the dataset on the subject Information Security that are not answered by more than 50 percent of the organizations. All these eleven questions are removed from the dataset. This resulted in a dataset with only 20 columns left.

Survey with questions on the subject continuity

Two questions within the dataset regarding the questions on the subject continuity were not answered by more than 50 percent of organizations. After these two columns were removed, this resulted in a dataset with 28 columns.

The returned datasets after the third part of data cleaning are referred to as the second version of the datasets during the creation of the exploratory model. The set of column names of these datasets can be found in the appendix next to the names of the columns that were removed between the first and final parts of data cleaning. Two additional datasets are created for the organizations that answered both surveys. The organizations that only answered one of the two surveys are not included in these datasets. These two datasets are created after the first part of the data cleaning, which is the first version of the datasets and after the third part of the data cleaning, which is the second version of the datasets.

6.2 Re Expression (data transformation)

It is key to know what kind of data is needed for the creation of the exploratory model. The data that is needed to execute the K-means cluster algorithm is numerical data. The data of the survey in the datasets created during the residual is ordinal data. The data is transformed from ordinal data to numerical data in order to run the cluster algorithm. The data of the surveys are answers of multiple choice questions given by the supervised organizations. The answers are written out in the initial dataset, which is seen in computer programming as a string. In other words, a sequence of characters. Thus, as the K-means algorithm which is used to create clusters can only process numerical data, the outcome of the surveys are changed into numerical values in order to run the K-means algorithm. This is executed in the following way; the first possible answer of the question has been replaced with the value ‘1’, the second answer with the ‘2 and so on. When a question is not answered, an extra value is added as a possible answer. This is the value ‘5’ and this value is used within all the transposed datasets.

6.3 Resistant

Some parametric tests are executed on the datasets to obtain more information about the data during the resistant. Most parametric tests such as the mean are hard to apply to this dataset as it is ordinal data transformed into numerical data. That is why only two values are calculated for both versions of all the datasets. The first value is the frequency of an answer in the dataset, and the second value is the percentage of an answers being present in the dataset in the dataset. The assumption here is that all the answers for the multiple choice questions are the same. This is not always true, but for many questions is. In fact, there is only one answer that is the same for all questions, which is answer number 5. This means that the question is unanswered

6.3.1 Resistant of the first version of the datasets

The number of time and the percentage of the answers is calculated for the datasets after the first part of the data cleaning. The results are shown in Table 1. There are three interesting observations that can be made by examining the table. The first is the difference of unanswered questions between the survey about the subject Information Security and the survey about the subject continuity. The difference between the two surveys is more than 20 percent. This means that the survey about the subject Information Security has 20 percent more unanswered questions. The second observation is the low frequency of answers with the value '3' or '4' in the dataset, especially within the dataset about Information Security. This is because only three of the 41 questions have a third possible answer and none of the questions have a fourth possible answer. The final observation regards the difference in percentage for the answers with the value '1' and '2' between the dataset about the subject Information Security and the dataset about the subject continuity. The first possible answer is "Ja" which means "Yes" in Dutch, and this answers accounts for 90 percent of the questions in the dataset about subject continuity and is 95 percent of the questions in the dataset about Information Security the first possible answer to the questions. The second possible answer is "Nee" which means "No" in Dutch and this answer for 90 percent of the questions for the continuity dataset and 93 percent for the questions of the survey on the subject Information Security. The table indicates a significant difference of percentage between these answers. Whereas the first answer is responded to by 22 percent more in the dataset about continuity, the second answer is responded to by less than 12 percent in the dataset about continuity.

Version 1	**Both Surveys**		**Continuity**		**Information Security**	
	Count	Percentage (%)	Count	Percentage (%)	Count	Percentage (%)
Answer 1	607	35,65	476	49,58	349	27,27
Answer 2	609	36,78	293	30,52	541	42,27
Answer 3	84	5,07	100	10,42	2	0,16
Answer 4	37	2,23	46	4,79	0	0
Answer 5	319	19,26	45	4,69	388	26,41
Total	1656	100	960	100	1280	100

Table 1: The number of times and the percentage a response was provided by the organizations after the first part of the data cleaning.

6.3.2 Resistant for the second version of the datasets

The same values are calculated after the third part of data cleaning. These results and the differences between the data in the datasets after the first part of the data cleaning and the third part of the data cleaning are shown in table 2. All the differences are negative because data could only be removed from the dataset between the first and the third part of the data cleaning. The difference in percentage can be positive or negative because it a comparison to the total number of answers, which will naturally decrease when removing data from the dataset.

Three observations were made after the first step of data cleaning. The new values for the second part of the dataset are used to investigate if the observation are still holding up for the new datasets. The first observation is the high percentage of unanswered questions within the dataset about Information Security compared to the dataset about continuity. This difference is reduced to 5 percent after the second part of the data cleaning, which is significantly lower compared to the 20 percent after the first part of the data cleaning. The second observation was about the low number of answers three and four. This observation still holds up for the new datasets, but the fifth answer can be added alongside these answers.

The final observation was on the difference in the first and second answer between the dataset about the survey on continuity and the survey on Information Security. This observation still holds up as well. The first answer is responded more frequently to the survey about continuity and the second answer is responded more frequently to the survey about Information Security. A new observation is that after the third part of the data cleaning 150 answers are removed from the dataset about continuity and 672 from the dataset about Information Security.

Version 2	**Both Surveys**		**Continuity**		**Information Security**	
	Count	Percentage (%)	Count	Percentage (%)	Count	Percentage (%)
Answer 1	458	43,29	383	47,28	235	38,65
Answer 2	439	41,49	266	32,84	332	54,61
Answer 3	83	7,84	100	12,35	1	0,16
Answer 4	37	3,5	46	5,68	0	0
Answer 5	41	3,88	15	1,85	40	6,58
Total	1058	100	810	100	608	100

Table 2: The number of times and the percentage an answer was provided by the organizations after the third part of the data cleaning.

Table 3 shows the differences of the number of times an answer is present in the dataset and the change in percentages. The biggest changes are the decrease in the percentage of the fifth answer for the dataset on the subject Information Security and the dataset of the organizations that replied to both surveys. The creation of the Exploratory Model will clarify what the influence is of the decrease of the number of unanswered questions on the creation of the clusters.

Part 1- Part 3	**Both Surveys**		**Continuity**		**Information Security**	
	Difference	Percentage (%)	Difference	Percentage (%)	Difference	Percentage (%)
Answer 1	-149	7,64	-93	-2,3	-114	11,38
Answer 2	-170	4,71	-27	2,32	-209	12,335
Answer 3	-1	2,77	0	1,93	-1	0
Answer 4	0	1,27	0	0,89	0	0
Answer 5	-278	-15,38	-30	-2,84	-348	-19,83
Total	598	0	150	0	672	0

Table 3: The difference between the number of times and the percentage an answer was provided by the organizations after the first and third part of data cleaning.

6.4 Revelation (data visualization)

Three different observations were made during the resistant part of the EDA and only one of the observations was different for the datasets after the first part of data cleaning and the third part of data cleaning. This was the difference in the percentage of unanswered questions which decreased drastic for the dataset with the survey scores on the subject Information Security. This is demonstrated by comparing Figure A2.1 and Figure A2.2 from the second chapter of the appendix. The unanswered questions in the datasets have the value ‘5’. Figure A2.3 and Figure A2.4 show the percentage of all the answers in the datasets of the survey on the subject Continuity. The comparison of the figures show that the number of unanswered questions decreased but not as drastically as for the dataset on the subject Information Security. Figure A2.5 and Figure A2.6 show the percentage of answers in the datasets on both subjects. For this dataset is viewable that the number of unanswered questions decreased significantly. This is caused by the decrease of the unanswered question in both datasets. All the bar graphs demonstrate the other observations made during the resistant.

6.5 Conclusion EDA

The primary difference between the first version of the datasets and the second version of the datasets is the decrease of the unanswered question within the dataset on the subject Information Security and the dataset with the organizations that responded to both surveys. Furthermore, the first and second answers are the most present within the datasets and the percentage of these answers per dataset increase in the second version of the dataset on the subject Information Security and the dataset with the organizations that responded to both surveys. For the dataset on the subject continuity, there are minor changes after the final two steps of the data cleaning. The expectation is that the clusters will be relatively similar for both datasets on the subject continuity. Two different versions of all the datasets are generated to facilitate the creation of the exploratory model.

7. Risk detection model

This chapter discusses the creation of the Exploratory Model as the first stepping stone for the creation of the Risk Detection model. Section 7.1 highlights the goal of the model towards the case study. Following, the most important challenges on the creation of the Exploratory Model are discussed (7.2). The creation of the Exploratory Model consist of two steps. The first step is the creation of the cluster. Section 7.3 provides information on the used dataset for the creation of the cluster. In addition, for the questions within the datasets with the survey scores is the importance defined(7.4). Following, the effectiveness of the K-means cluster algorithm is defined for the dataset and the clusters within the different datasets. The final step of the creation of the exploratory model is the calculation of different correlations between the clusters and the dataset from Social Media and the Reporting Desk(7.8). To conclude is the influence of the results of the Exploratory Model on the process of supervision discussed.

7.1 The goal of the model

The model supports two different levels within AT with information about data-driven risk-based regulation. The first is on the level of the experts of Information Security. These experts make decisions based on a number of values within the dataset and their own qualitative assessments. Quantifying these qualitative assessments can lead observers to assume that all risk-based systems are purely quantitative, whereas in practice, their character can vary (Black & Baldwin, 2010). Only the quantitative data is taken into consideration for this model. This is data obtained by the experts over the years. It is supposed to be a tool for the experts to support their decision making on risked based regulation by providing them with additional information about the risks the organizations pose. This information can be used to improve the effectiveness of the regulation, where effectiveness is measured as regulation commensurate to the risks that organizations pose.

The second level within AT is on the strategic level for the whole organization. By creating the model, the possibilities and promises of creating data-driven models to support risk-based regulation can be exhibited to managers within AT who determine the strategy of

the regulator. It can provide them with new insights on future developments, and how to react to the developments by harnessing data-driven decision making.

7.2 The most important modeling challenges

Certain assumptions had to be made during the development of the model. These assumptions were related to which data should be used in the analysis. The first assumption pivoted upon the relation between organizations within the supervisory field. Multiple regulated organizations are daughter companies of other regulated organizations. For the creation of the model, these organizations are seen as independent and taken as separate entities. The second assumption is made on the elimination of questions that have less than three responses in the dataset. These questions have been removed from the dataset during the first part of the data cleaning. The challenge on what to do with the other questions that are partly unanswered follows from this assumption. The third assumption made is on the questions were multiple answers were given by the supervised organizations. These questions are split into different columns. The final assumption is on the unanswered questions, which is tackled by adding the value '5' as extra possible answer to all the questions. extra answer not answered is added to the possible answers.

The survey with the questions seeking general information of the organizations comprises of multiple subjects. There is a difference in the importance of these subjects to the subject continuity. During the creation of the model, however, all the subjects are treated as equally important, as only the numerical trends are found within the data. The question is whether this is realistic towards the risks related to the subject continuity.

7.3 Datasets used for clustering

The K-means clustering algorithm is executed on the six different datasets that were created during the EDA. The first four datasets are the outcomes after the first and third part of the EDA that considered the two initial datasets provided by AT. The first version of the datasets are the returned datasets after the first part of the data cleaning an the second version of the datasets are the returned datasets after the third part of the data cleaning. Note that not all the organizations conducted both of the surveys. As such, the last two datasets contain information on the organizations that conducted both surveys. This means that the organizations that only participated in one of the surveys are left out of these datasets.

The input data are outcomes of the surveys, which are the possible answers for each question. The K-means algorithm can only process numerical data, but the answers of the organizations are written out in the dataset. Which means that during the re-expression part of the EDA the outcome of the surveys had to be changed into numerical values in other to run the K-means algorithm. This is process is discussed in the data transformation part of the EDA. The datasets with the numerical data is used for the K-means clustering.

7.4 Result K-means clustering

The figures
The previous section explained how steps in the methodology of the algorithm are executed on the different datasets and the three different figures that are created for all the datasets. These figures are used to support the results of the algorithm for all the datasets. The first figure is a scatter plot of the calculated X and Y values from all the objects in the dataset. The X and Y values are based on the survey scores of the organization. The second figure is the outcome of the elbow method, which is used to define the optimal number of clusters for the dataset. All the scatterplots and the graphs of the elbow method can be found in the third chapter of the Appendix. The final figure is the same scatter plot, but with the K clusters, where every cluster has their own colour. All the scatterplots with the created clusters can be found in the results section of the K-mean clustering.

A point in the scatterplots
Objects in the datasets are represented by dots in the scatterplots. In these datasets, an object is the organizations that replied to at least of the surveys. Thus, a point in the scatter plot represents an organization that participated in the questionnaire. All these points are placed in a space, and this space is defined by the X and Y coordinates of all the organizations surveyed. The centroids are placed and the clusters are created in this space. The number of formed clusters is equal to the K number of placed centroids.

Cluster
The goal of the exploratory modeling is to create a risk detection model. Different risk profiles are created based on the available data in order to create a risk detection model. A cluster in the results represents a risk profile and is based on the survey data from the past. The organizations that answered the questions the most similar to each other are placed in the same cluster. A certain trend could be present among answers of the organizations. When new organizations fill in the surveys, they will be placed into one of the clusters based on the similarity of their answers and the average of all the answers of the organizations within the clusters. The number of clusters is equal to the number of centroids that should be placed into the space, which is the value for K.

The results of the K-means clustering for the different datasets are discussed in the following order; The results of K-means clustering on the datasets after the first part of the data cleaning is followed by discussion of the results of K-means clustering on the datasets after the third part of data cleaning. The K-means clustering is executed for six different datasets in total.

7.4.1 Results K-means clustering for the first version of the datasets

Data from the survey on general information about Information Security
The first dataset that is analyzed is the data from the survey on general information about Information Security. This dataset consists of the survey scores collected by AT on the subject Information Security for 32 organizations. Figure A3.1 shows the scatterplot of the X and Y coordinates of all the organizations within the dataset. It is directly visible that the

points are mainly spread out on the left side of the figure. There are a number of points visible that are close to each other, but it is hard to develop clusters directly based on the plot. To define how many clusters should be formed, the value of K is calculated. This is done by the elbow method, and the outcome of this method in shown in figure A3.2. The marginal gain of creating an extra cluster changes drastically when K equals three. Thus, for this dataset, the optimal value for K is three.

The last figure is Figure 5 which shows the three clusters created for the dataset. All the points in the scatterplot have a colour that corresponds to the cluster that the organizations were placed in. The colours are light green for cluster 1, purple for cluster 2 and grey for cluster 3. The first cluster consists of 11 points, the second cluster,17 points, and the third cluster comprises 4 points.

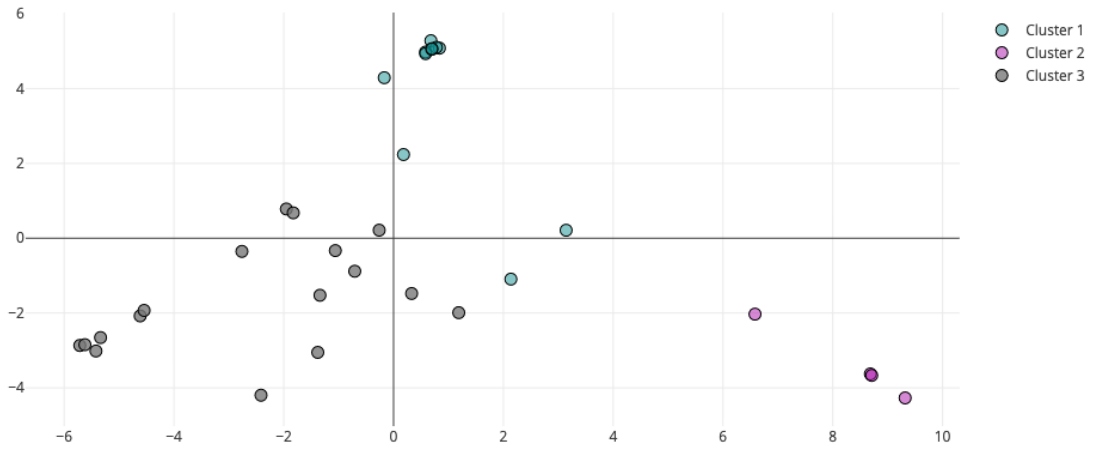


Figure 5: Scatterplot with the three different clusters created for the first version of the dataset on the subject Information Security

Data from the survey on the subject continuity

The second dataset consists of data from the survey on the subject continuity. This dataset is developed from the data of 30 different organizations. Figure A3.3 shows the scatterplot of the data. It is clear that the points are segmented primarily into two sections. The first part are the points with the negative x value, and the second are the points with a positive y value. To define how many clusters should be formed, the value of K is calculated. This is done by the elbow method and the outcome of this method in shown in Figure A3.4. The marginal gain of creating an extra cluster changes drastically when K is equal to two. Thus, for this dataset, the optimal value of K is two. The final figure is figure 6 which shows both cluster. The points on the scatterplot have a colour that corresponds to the cluster that they were placed into. The colours are light green for cluster 1, purple for cluster 2. Both clusters consists of 15 points.

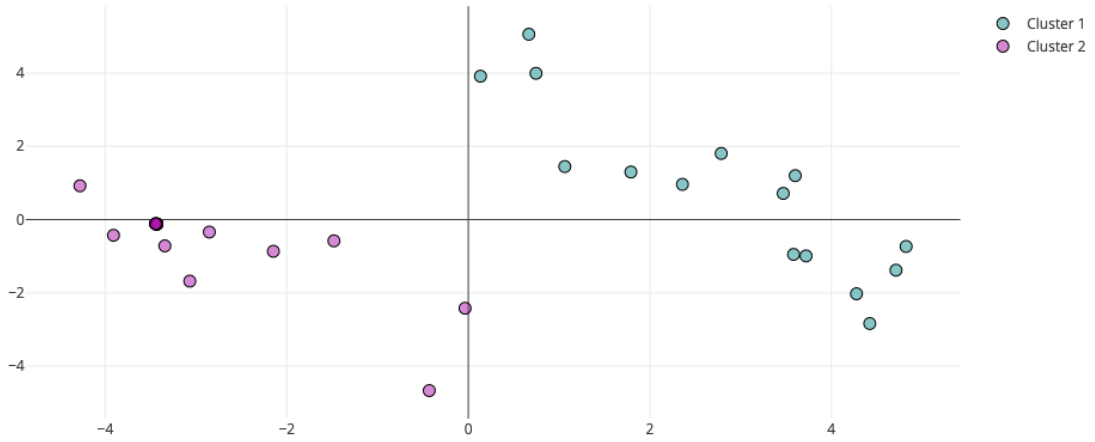


Figure 6: Scatterplot with the two different clusters created for the first version of the dataset on the subject Continuity

Data from the organizations that responded to both surveys

The third dataset that is analyzed is comprised of data from organizations that participated in both of the previous surveys. This accounts for 23 organizations in total, and Figure A3.5 shows the scatterplot of the data. It is visible that there is a group of points near each other between minus two and minus four. The remainder of the points are spread out. To define how many clusters should be formed, the value of K is calculated. This is done by the elbow method and the outcome of this method is shown in figure A3.6. The marginal gain of creating an extra cluster changes drastically when K equals three. Thus, for this dataset, the optimal value for K is three. The last figure is Figure 7 which shows all three clusters. The points on the scatterplot have a colour that corresponds to the cluster that they were placed into. The colours are light green for cluster 1, purple for cluster 2 and grey for cluster 3. The first cluster consists of 11 points, the second cluster has 10 points, and the third cluster has 2 points. These two points are outliers on the left upper side.

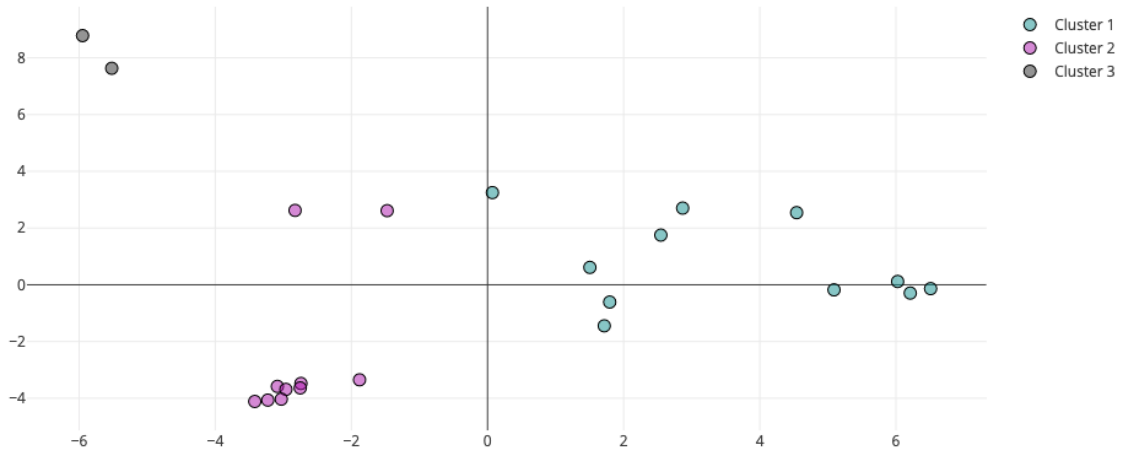


Figure 7: Scatterplot with the three different clusters created for the first version of the dataset on both subjects

The number of clusters created for the data from the survey on the subject continuity is equal to two, and for the other two datasets, the number of clusters is equal to three. These values are relatively low due to the number of organizations within the datasets, which is between 23 to 32 organizations per dataset. The first scatterplot shows the most distribution between the points for the dataset with three created clusters. For this dataset, the third cluster the most spread cluster, as shown in Figure, and consists out of the most data points. The data points of the other clusters are located closer to each other. The third dataset is the other dataset with three created clusters and for this dataset the data points in the clusters are less dispersed. The data points for the third cluster are closer to each other but only consist of two data points. Two outliers are visible for the second cluster and the data points of the first cluster are more spread away from each other. The final dataset is the second discussed dataset, which consists of two clusters. All the points on the right side of the Y axis create the first cluster and all the points on the left side of the Y axis create the second cluster. The points within the clusters are spread within the different sides of the Y-axis.

7.4.2 Results K-means clustering for the second version of the datasets

The K-means algorithm is executed on a different version of the same initial datasets as described in section 7.4.1, which are the second version of the datasets after the third part of data cleaning. The scatterplots and the graphs from the elbow methods can be found in the Appendix A3. The scatterplot with the created clusters are discussed and shown in this chapter.

Data from the survey on general information about Information Security
The first dataset is the data of the survey on the subject Information Security. For this dataset, K is equal to three, which is the same as the value of K for the first version of this dataset. Figure 8 shows the scatterplot with the created clusters. The first cluster is the

group of points on the lower left side of the plot and they are colored green . The data points of this cluster are located relatively close to each other. The second cluster is the points on the right side of the plot and they are more dispersed from each other. This cluster is marked with the colour purple. The final cluster is the points on the upper left side of the plot, which are marked with the colour grey. A change that could be made is adding another cluster that will split the second cluster in an upper right cluster and a lower right cluster. This would result in a K-means clustering where K is equal to four.

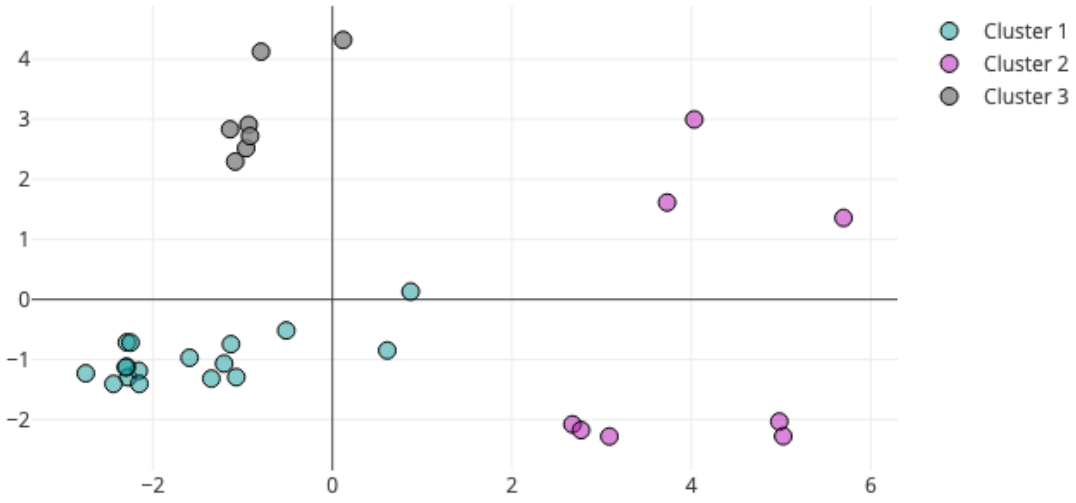


Figure 8: Scatterplot with the three different clusters created for the second version of the dataset on the subject Information Security

Data from the survey about continuity

For the dataset of the survey about continuity is it hard to define the elbow point. It is difficult to determine if the marginal gain of creating an extra cluster changes more when K is equal to two or equal to four. As a result, the K-means algorithm is run for both K is equal to two and K is equal to four to see what the differences are for both values of K. The graph of the elbow method and the graph of the scatterplot that includes the clusters for K is equal to four can be found in the appendix. The value of K is equal to two is taken for this dataset. This is because the K means clustering algorithm did not create four groups of points that could be seen as clusters, in contrast to the execution of the K-means algorithm for K is equal to two, which created two clusters. The first cluster is the group of points on the left side of the plot marked in green, and the second cluster is the group of points on the right side of the plot marked with a purple colour.

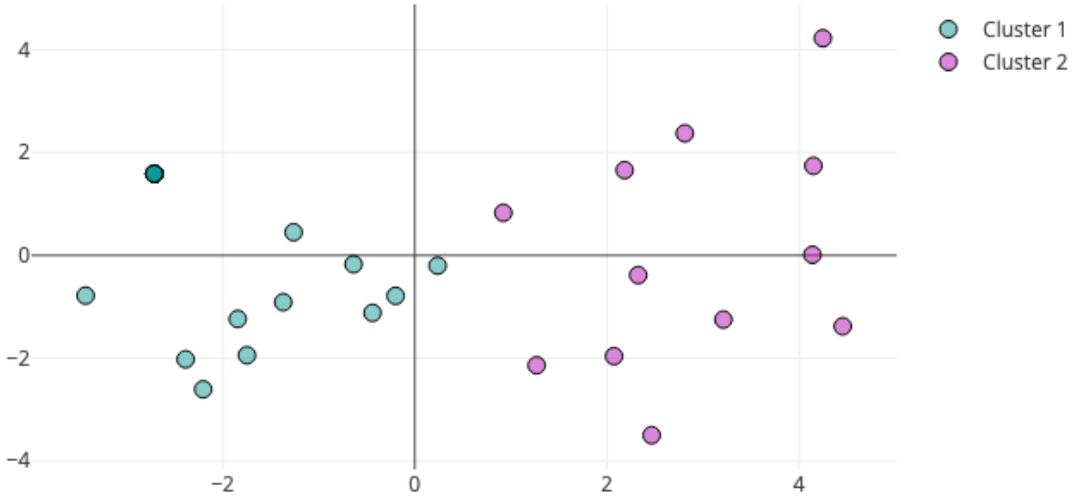


Figure 9: Scatterplot with the two different clusters created for the second version of the dataset on the subject Continuity

Data from the organizations that responded to both surveys

The final dataset consists out of data from the organizations that responded to both surveys. This resulted in the creation of three different clusters. The first cluster is located on the left side of the plot with the positive Y values and marked with a green colour. The second cluster is the purple points, which are primarily on the right side of the plot, along with a single outlier in the middle upper part of the graph. The final cluster is the grey points in the middle lower part of the plot.

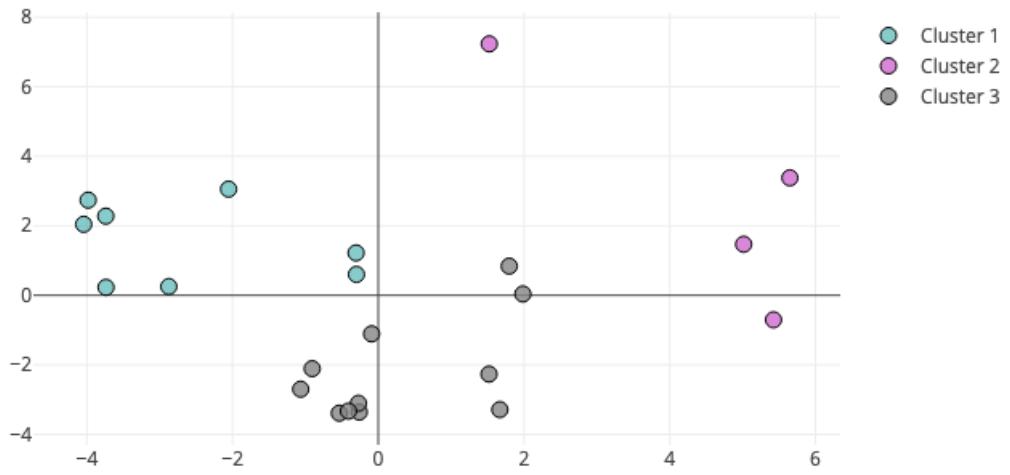


Figure 10: Scatterplot with the three different clusters created for the second version of the dataset on the both subjects

7.5 Conclusion K means clustering

The K-means clustering algorithm is run for two different versions of all the datasets. This did not lead to a different number of clusters for the datasets, but the average distance between the data points change in the resulting clusters. Differences are viewable when the scatterplots of the different versions are compared with each other. The first dataset about the survey on the subject Information Security demonstrates the largest visible difference. The data points in the clusters created for the second version of this dataset are closer located to each other than the cluster for the first version of the dataset. This means that the answers provided by the organizations within the second datasets are more similar to each other that the answers given in the first dataset.

The clusters created for the dataset on the subject continuity are similar to each other. For these datasets, two clusters are created and the clusters for the second version of this dataset are further dispersed than in the first dataset. The change of the dataset did not create clusters with organizations that answered more similar as for the change for the first dataset on the subject Information Security.

For the final dataset, three clusters are created for both versions and for both datasets the data points are located close to each other, excluding some outliers. The creation of the different versions of the dataset did not lead to a visible improvement in the distance between the data points of the clusters created in the scatterplots.

The percentage of organizations from the datasets used for the creation of the cluster covers between 1.15 and 1.6 percent of the organizations registered to ACM, which is a significantly low representation. This is due to the low number of organization data that is sourced from social media or the Reporting Desk for failures. The K-means clustering method is a machine learning technique which is normally used for datasets with a large sample size. The dataset used for the clustering algorithm have a small sample size of organizations. The influence of a small sample size on the K-means clustering algorithm is that the outcomes of the created clusters are less reliable and less accurate than when the algorithm is used for datasets with large sample sizes.

In the next section, the properties of the clusters are explained. The properties are the questions that have the biggest influence on the forming of the clusters. This is followed by a section defining the average distance between the answers of the organizations within the clusters. This is calculated to investigate what the influence is of the changes made between the first version of the datasets and the second version of the datasets.

7.6 Which questions influence the creation of the clusters the most?

The previous section discusses the results of the cluster creation after the execution of the K-means algorithm on the datasets. All the formed clusters are marked in the scatterplots. The next step is to investigate which organizations are placed into which cluster and what the properties of these clusters are. These properties are the columns that have the highest influence on the cluster forming, which correspond to the questions of the surveys. For all the datasets, a table with the information about which organizations are included in which cluster is shown This is followed by a table with the questions that have the largest influence on the creation of the clusters. These questions have the most influence on the cluster creation are the criteria that separate the supervised organizations based on their survey scores. The important towards the process of supervisions is compared with the opinion of experts during the validation. The tables with the cluster and the corresponding organizations and the tables with the questions with the largest influence on the cluster formation are shown in Appendix A4.

The determination of which questions have the largest influence on the cluster forming is based on the comparison of the averages for every column value of every cluster from a single dataset. When the difference between the value of the averages is greater than 0,75 for a minimum of two clusters within the dataset, the assumption is made that these questions have a high influence on the formation of the clusters. This is done starting with the two datasets on the subject Information Security, followed by the dataset with the information about the subject continuity, and the last two datasets contain the data from the organizations that responded to both surveys.

The organizations within the clusters for the datasets about Information Security are shown in Table A4.1. Part 1 stands for the first version of the dataset about Information Security and Part 3 stands for the second version of the same dataset after the final part of data cleaning. To develop a better understanding of the characteristics of the clusters, Table A4.2 is created for the first version of the dataset on the subject Information

Security. This table includes the name of the columns that have the most influence on the creation of the clusters. Cluster three is the smallest cluster for the first dataset and consists of four organizations, followed by the first cluster with twelve organizations, and cluster two is the biggest with sixteen organizations. The questions that most influence the creation of the clusters the can be found in Table A4.2. The third cluster of the second version of the dataset consists of eight organizations, the second of nine, and the first of thirteen. The questions with the highest influence on the cluster forming for the second version of the dataset can be found in Table A4.3.

The second dataset consists of the outcomes of the survey on the subject continuity. The cluster algorithm created two clusters for both version of this dataset. Organization placed in the clusters for both datasets are shown in Table A4.4. The clusters created by the algorithm for the first version of the dataset consists of fifteen organizations for both. The most important questions for the creation of these two clusters, which are the risk criteria, can be found in Table A4.5. For the second version of the dataset on the subject continuity, the number of organizations within the cluster not equal to each other. Cluster one comprises of twelve organizations and cluster two of eighteen. In Table A4.6 the questions with the highest influence on the creation of the clusters can be found.

The final two datasets are a combination of the first and second version of both datasets. Cluster one of the first version of this dataset consists of four organizations, cluster three of eight, and cluster two of eleven. The organizations per cluster can be found in Table A4.7. Table A4.8 comprises of the questions that have the highest influence on the creation of the clusters. The second dataset consists of; cluster two comprises of four organizations, cluster one comprises of eight organizations, and cluster three of eleven. Table A4.7 shows the organizations per cluster for the second version of the dataset on both subjects. The questions with the highest influence on the creation of these cluster can be found in Table A4.9.

7.7 What is the effectiveness the K Means cluster algorithm?

In this section, the effectiveness of the cluster algorithm is calculated, because the number of clusters for every dataset is the same after the different parts of the EDA. The effectiveness of the algorithm is based on the difference between the rows in a dataset, which are the differences between the answers given by the organizations. A row represents an organization and a column a question. The data is changed to numerical data for the creation of the cluster, so when the difference between two rows is low, the assumption is made that the answers given by the organizations are similar to each other.

Four different values are calculated for every cluster. The first value is the total difference between all the answers of all the organizations within the dataset. All the organizations are compared with each other one time. This gives the total difference for every cluster. The second value is the difference per organization, which is the total difference divided by the number of organizations within the clusters. The clusters of a single dataset are easy to compare with the difference per organization. The third value is the difference per question, which is the total difference divided by the number of questions. This value provides information regarding the difference in responses to the question per cluster.

This is insightful to compare the clusters of a single dataset and the clusters created by the dataset with the same subject. The final value is the difference per organization per question. This value is calculated to compare the datasets of a subject after the different parts of the EDA. This is calculated because there is a difference in questions between the datasets after the first and the final part of the EDA. This value is calculated to compare different datasets of the same subject.

Two comparisons are made for all the formed clusters. The first comparison is between the created clusters within a single dataset to investigate which clusters have the lowest average distance between all the rows within the dataset. The second comparison is between the clusters created for the first and second version of a single dataset. This is performed to investigate what the effect is of the algorithm on the changes made to the dataset between the first and second version of the dataset. All four values are calculated for the datasets about Information Security, continuity and for the organizations that responded to both surveys.

The first comparison is on the effectiveness within the single datasets on the subject Information Security. The total difference is not important for this comparison because the number of organizations within the cluster differs. On the contrary, the values of the differences per organization and the differences per question are important to compare the clusters within the dataset. The values of the differences of the clusters created for the first version of the dataset on the subject Information Security can be found in Table 4. This table indicates that the difference per organization is the smallest for the first cluster followed by the third cluster, and the second cluster has the largest difference per organization. The cluster algorithm has the most influence on the creation of the clusters for the cluster with the smallest difference, because the answers given by the organizations are the most similar to each other. Cluster number two has the largest difference per organization and the largest difference per question. For the creation of this cluster was the K-means algorithm least effective. The striking thing is that the first cluster has the least difference per organization, but the second highest difference per question. This is because the first cluster consists of more organizations than the third.

The differences of the second version of the dataset on the subject Information Security is shown in Table 5. For this dataset, the order of the effectiveness of the clusters is the same as the first version of this datase, however, the clusters consists of different organizations. The difference per organization is still the smallest for the first cluster, followed by the third, and the second cluster has the largest difference. Looking at the difference per question, the average is still the lowest for the third cluster, followed by the first cluster, and the second cluster has the largest difference per question. By defining the effectiveness of the cluster algorithm within these datasets, the difference per organization is the most important value. This means that for both datasets, the first cluster has the lowest difference and the second cluster had the highest difference.

In addition to comparing the cluster within the single datasets is the comparison made between the clusters of both datasets. Two different tables are created for two different versions of the dataset on the subject Information Security. The two different datasets are

based on different steps of data cleaning during the EDA. The last two parts of the resistant within the EDA led to 22 questions being removed from the dataset. In order to determine the influence of the data cleaning on the significance of the clusters, the difference per organization per question is calculated to compare the two different datasets with each other. This is computed because the number of questions differs between each of the datasets. Comparing the difference per organization from two datasets with a different number of questions makes the comparison insufficient. The tables demonstrate that the differences per organization per question are lower for all the clusters after the final part of the data cleaning, which means that those extra steps of data cleaning improved the effectiveness of the K-means algorithm on the dataset with the survey scores on the subject Information Security.

Version 1	**Information Security**			
	Total difference	per organization	per question	per organization per question
Cluster 1	1116	93	27,9	2,33
Cluster 2	3034	189,63	75,85	4,74
Cluster 3	952	119	13,22	1,65

Table 4: The differences between the survey scores of the clusters created for the first version of the dataset on the subject Information Security

Version 2	**Information Security**			
	Total difference	per organization	per question	per organization per question
Cluster 1	489	81,5	10,63	1,77
Cluster 2	1462	132,91	31,78	1,89
Cluster 3	660	94,29	14,35	2,05

Table 5: The differences between the survey scores of the clusters created for the second version of the dataset on the subject Information Security

After the determination of effectiveness of the cluster algorithm on the datasets for the subject Information Security, the effectiveness of the cluster algorithm on the datasets for the subject Continuity defines. The differences for the first version of the dataset is shown in Table 6. Two clusters are created for this dataset, where the second cluster has an approximately fifty percent higher difference for the total difference, difference per organization, and difference per question. This indicates that effectiveness of the cluster algorithm of the first cluster is higher than the second cluster.

In Chapter 6, section 6.3 explored the resistant component of the EDA, and demonstrated that the second and third part of the EDA did not have a significant effect on the dataset on the subject continuity. Only six questions were removed from the dataset. This leads to small differences in effectiveness of the cluster algorithm between the cluster within one of the datasets on the subject continuity and small differences when the significance of both datasets are compared with each other. Within the dataset, the values of the differences for the first cluster are around 50 percent higher than the values of the second cluster. This is dichotomous to the first dataset. Thus, cluster one from the first dataset corresponds with cluster two of the second dataset and cluster two of the first dataset corresponds with cluster one of the second dataset. The influence of removing those six questions from the dataset increases the difference per question and the difference per organization per question. As a result, when measuring the difference per organization per question, the second dataset performs worse than the first dataset. The calculated differences for the second version of the dataset on the subject continuity can be found in Table 7.

Version 1	**Continuity**			
	Total difference	per organization	per question	per organization per question
Cluster 1	1636	109,07	51,13	3,41
Cluster 2	2436	162,4	76,13	5,01

Table 6: The differences between the survey scores of the clusters created for the first version of the dataset on the subject Continuity

Version 2	**Continuity**			
	Total difference	per organization	per question	per organization per question
Cluster 1	2442	161,47	83,52	5,57
Cluster 2	1610	107	55,52	3,7

Table 7: The differences between the survey scores of the clusters created for the second version of the dataset on the subject Continuity

The final two datasets consist of data from the organizations that participated in both surveys. Data from the subjects of Information Security and continuity were merged together into one dataset for the organizations that responded to both surveys. The second and third part of the data cleaning removed much of the data from the survey on Information Security. The question to be explored is the influence of these changes on the significance of the clusters created for both datasets, and which clusters have the smallest difference within the datasets. The second component will be answered first by looking at the first version of the dataset. Cluster one had the lowest difference per organization, followed by cluster three, and cluster two had the biggest difference per

organization. This means that effectiveness of the cluster algorithm the highest is for cluster one and also has the lowest value for difference per question. Cluster three has the second lowest value, and cluster two has the highest value for the difference per question.

For the second version of the dataset, the second cluster is the most significant, followed by the first and third clusters. Comparing the difference per organization per question of both datasets with each other by calculation provides the average of the three clusters for both datasets. The resulting averages are 2,12 for the first dataset, and 2,13 for the second dataset. This means that the difference per organization per question increased by one hundredth, however, as this is a minute different, it is considered as having no change.

Part 1	**All Data**			
	Total difference	per organization	per question	per organization per question
Cluster 1	322	80,5	4,47	1,12
Cluster 2	2832	257,45	39,33	3,58
Cluster 3	952	119	13,22	1,65

Table 8: The differences between the survey scores of the clusters created for the first version of the dataset on both subjects

Part 3	**All Data**			
	Total difference	per organization	per question	per organization per question
Cluster 1	920	115	20	2,5
Cluster 2	190	47,5	4,13	1,03
Cluster 3	1446	131,45	31,43	2,86

Table 9: The differences between the survey scores of the clusters created for the second version of the dataset on both subjects

Conclusion effectiveness of the cluster algorithm

The major change applied during the EDA was the reduction of the unanswered questions between the first and the second version of the dataset on the subject Information Security and the dataset with the data from the organizations that responded to both surveys. It is important to understand the impact of these changes on the effectiveness of the cluster algorithm for both versions of these datasets. The impact on the effectiveness of the cluster algorithm on the second version of dataset compared with the first version of the dataset on the subject Information Security is high. The average difference per

organization per cluster decreased from 134 to 103. This is not the average difference per organization but the average difference per cluster. For the dataset with the information of the organizations that responded to both surveys, the changes made result in only a minor visible increase between the first and second version of the dataset. For the average distance per cluster, the first version of the dataset provides an average of 2,12 and the second version of the dataset provides an average of 2,13.

To develop a better understanding of the influence of the changes made during the second and third part of the data cleaning on the effectiveness of the cluster algorithm, the difference per organization is calculated for all the datasets, starting with the dataset on the subject Information Security. The average difference per organization for the first version of dataset is equal to 159 and the average difference per organizations for the second version of the data is equal to 82. This means that the effectiveness of the K Means cluster algorithm improved because the difference decreased by almost 50 percent. The second datasets are the two versions of the dataset on the subject continuity. The difference per organization for the first version of the dataset is equal to 136, and equal to 135 for the second version of the dataset. The difference per organization for all the organizations within both datasets decrease only by one point. This means that the effectiveness improved to a small degree for the datasets on the subject continuity. The final dataset consists of the data from the organizations that responded to both surveys. The difference per organization for the first version of the dataset is equal to 179, and equal to 111 for the second dataset. These figures indicate that the effectiveness of the cluster algorithm improved after the finals steps of the data cleaning.

7.8 Correlations

7.8.1 Social media

The first results of the correlations are the correlations between the created clusters and the social media data. AT collects data about failures that cause interruptions of the continuity of the telecommunication services and networks via social media. This data is comprised of number of posts per day that mention a failure by an organization monitored by AT. A threshold is set for the number of social media mentions, and once the threshold is exceeded, it is assumed that there was a failure on that day for that organization. Next to the 13 organizations where AT defined the thresholds consists the datasets of 26 of other organizations without a threshold defined by AT. In Appendix A5 are the thresholds and the average number of mentions of failures for the specific organizations shown. Thresholds defined by AT can be found in Table A5.1 The new thresholds are based on the threshold that was already set by AT. The average of the number of mentions on social media per day are calculated for all the organizations and the value of the thresholds for the organizations where of a threshold was missing are defined by comparing the average number of mentions. The new thresholds are shown in Table A5.2 and the average number of mentions per day are shown in Table A5.3.

The correlations of the social media derived data is first calculated for the datasets about the subject Information Security. Table 8 shows the total failures per cluster for the first version of the dataset on the subject Information Security. This table shows that the total

number of failures and the failures per organization is the highest for cluster number one. Cluster number two has the second highest total clusters, but the highest number of failures per organization. The third cluster has the lowest number of total failures, but the second highest number of failures per organization. The results for the second version of the dataset are shown in Table 9. For the second dataset, the first cluster had the highest number of total failures and highest number of failures per organization. This is followed by the second cluster which has the second highest number of total failures and failures per organization. The final and third cluster has the lowest number for both the total failures and the failures per organization.

The biggest difference between the results for both datasets is that for the first dataset, the cluster with the lowest total failures has the second highest number of failures per organization, and the second dataset, the cluster with the lowest total failures has the lowest number of failures per organization. In addition, the cluster with the second highest number of total failures in the first dataset has the lowest number of failures per organizations, and the cluster with the second highest number of total failures for the second dataset has the second highest number of failures per organization.

Version 1	**Information Security**			
	Total failures	Failures per organization	Failures per year	Failures per organization per year
Cluster 1	1134	94,5	154,71	12,89
Cluster 2	185	11,56	25,24	1,58
Cluster 3	164	41	22,37	5,59

Table 10: The total number of failures and number of failures per organization per cluster for the first version of the dataset on the subject Information Security

Version 2	**Information Security**			
	Total failures	Failures per organization	Failures per year	Failures per organization per year
Cluster 1	1153	67,82	157,3	9,25
Cluster 2	230	28,75	31,38	3,92
Cluster 3	100	14,29	13,64	1,94

Table 11: The total number of failures and number of failures per organization per cluster for the first version of the dataset on the subject Information Security

The next datasets are on the subject continuity, which consist of two different clusters. The total number of failures and the failures per organization per cluster can be found for the first version in Table 10 and the second version Table 11. For the first dataset, the first cluster has the highest number of total failures and the highest number of failures per organization. The second cluster has the lowest number of total failures and the lowest number of failures per organization. There is a significant difference in the failures per organization for both clusters, with a difference of 78 percent between both datasets. For the second dataset on the subject continuity, the first cluster has the lowest number of total failures and the highest number of failures per organization, and the second cluster had the highest number of total failures and the lowest number of failures per organization. In comparison with the first dataset where the difference in failure per organization is equal to 78 percent, the difference between the failure per organization for the second dataset equals 1,3 percent.

Version 1	**Continuity**			
	Total failures	Failures per organization	Failures per year	Failures per organization per year
Cluster 1	1290	86	175,99	11,73
Cluster 2	278	18,53	37,93	2,53

Table 12: The total number of failures and number of failures per organization per cluster for the first version of the dataset on the subject Continuity

Version 2	**Continuity**			
	Total failures	Failures per organization	Failures per year	Failures per organization per year
Cluster 1	632	52,67	86,22	7,19
Cluster 2	936	52	127,69	7,09

Table 13: The total number of failures and number of failures per organization per cluster for the first version of the dataset on the subject Continuity

The final two datasets consists of the data of the organizations that responded to both surveys. The calculated number of total failures and failures per organization for the datasets can be found for the first version in Table 12 and the second version in Table 13. The clustering algorithm created three different clusters for both datasets. The third cluster of the first dataset has the highest number of total failures within the dataset and the highest number of failures per organization. The first cluster has the second highest number of total failures and failures per organization, and the third cluster has the lowest number of total failures and failures per organization. For the second dataset, the second cluster has the lowest number of total failures and failures per organization, followed by

the third cluster with the second highest number of total failures and failures per organization. Finally, the first cluster has the highest value for both variables. The biggest difference between the results for both datasets is that the organizations within the third cluster for the first dataset are accountable for 77 percent of all the failures, and the cluster with the most failures is cluster number one for the second dataset, accounting for 43 percent of the failures.

Version 1	**All Data**			
	Total failures	Failures per organization	Failures per year	Failures per organization per year
Cluster 1	185	46,25	25,24	6,31
Cluster 2	136	12,36	17,75	1,68
Cluster 3	1086	135,75	148,16	18,52

Table 14: The total number of failures and number of failures per organization per cluster for the first version of the dataset on both subjects

Version 2	**All Data**			
	Total failures	Failures per organization	Failures per year	Failures per organization per year
Cluster 1	611	76,38	83,36	10,42
Cluster 2	188	47	25,65	6,41
Cluster 3	608	55,27	82,95	7,54

Table 15: The total number of failures and number of failures per organization per cluster for the first version of the dataset on both subjects

7.8.2 Reporting Desk data

The second dataset that is used to calculate correlations is the dataset from the Reporting Desk. The tables about the organizations in the dataset of the Reporting Desk, the selected cluster and the calculated correlation can be found in Appendix A6. This dataset consists of information about the incidents involving the supervised organizations that led to failures that had a negative impact on the continuity of the provided network or services. There are sixteen organizations within this dataset whereby ten of these organization are represented within the datasets with the surveys, the organizations are shown in Table A6.1. A total of 32 organizations is included in the datasets for the survey on the subject Information Security and continuity, and 24 for the dataset with the data of the organizations that responded to both surveys. Thus, the organizations within the dataset covers around one third of the organizations for all datasets.

The number and the percentage of organizations within the created clusters that are represented within the dataset from the Reporting Desk are shown in Table A6.2 and Table A6.3. This number is defined to investigate if it is possible to calculate the correlation between the data of the Reporting Desk and the clusters. The number of organizations and the percentage needs to be high enough within the clusters to be able to calculate a correlation that is significant enough to be representative for the whole cluster. This is not the case for most clusters due to a low coverage ratio between the cluster and the data from the Reporting Desk. As a result, the correlations are only calculated for one cluster to provide an example of what the influence could be of reliable correlations. This is executed for the third cluster of the first dataset with the data about the organizations that responded to both surveys, because the number of organizations is equal to six and the coverage is equal to 75 percent.

Six organizations are represented within the dataset with the information about the surveys and the dataset from the Reporting Desk. The cluster consists of eight organizations in total. To calculate the correlations, two organizations which are not represented within the dataset of the Reporting Desk are removed from the cluster. This led to a new cluster of six organizations, and these organizations are listed in Table A6.4. These six organizations are responsible for 315 of 431 incidents reports, which accounts for 72 percent of the reported incidents between 2015 and 2019. To develop further insight regarding which organizations cause the highest number of incidents, the percentages of the failures per organization are calculated and also shown in Table A6.4.

The number of failures per year between 2015 and 2019 is calculated, and this provides a better understanding on how the failures are distributed over the years and on the quality of data each year. Distribution and quality is analysed in yearly increments as only the year is known for the reportings of the incidents, as opposed to the specific date of the reported incident. This resulted in the number of incidents per year which are shown in Table A6.5. What stands out immediately is that the number of incidents during the year of 2015 and 2019 are clearly lower than the number of incidents during the other years. However, this research is carried out during the year of 2019, and it is unclear what months of data is provided by AT for the year 2019. As such, this is assumed to be the reason for the low number of incidents in 2019. The reason of the low number of incidents is difficult to determine. The assumption is made that the organizations reported fewer incidents in 2015 due to the fact that reporting incidents if voluntary.

The next step is calculating the characteristics of the incidents for the organizations within the selected cluster. There are seven different characteristics within the dataset of the Reporting Desk. This begins with the ‘type of services’ that are compromised due to the incidents. There are seventeen types of reported services that experiences a failure, and can include a combination of multiple services. The number of times a service was compromised, and the rounded percentage are shown in Table A6.6. Five ‘types of services’ have a percentage of 10 percent or higher and these services are email, mobile internet, mobile telephony, broadcasting distribution, and fixed telephony. In addition to the services with a higher percentage of incidents, is it notable that the emergency number was compromised twice during the selected time period, as this is an important

service and their disruption comes with high risks. The calculated percentages provide information about the probability of the kind of service that could be compromised when a failure occurs.

The second characteristic is ‘the number of affected customers’ that were compromised by the incident. This number of customers is divided into three different groups based on the data. The first group is the incidents that reported to have a number of affected customers that is equal to minus one, the second group are the incidents for which the number of affected customers is unknown, and the final group are the incidents with a number of affected customers that is higher than zero. Table A6.7 shows the number of incidents and the percentage per group. For the first group of incidents, it is assumed that no customers were impacted. The average number of affected customers for the final group is equal to 409.626 customers with a standard deviation of 1.118.904. The spread of the number of affected customers is [21,5 ; 300.000].

The third variable is ‘the range’ of the incidents. Results of the correlation between the cluster and the range of the incidents can be found in Table A6.8. There are five different options for the range within the data: (N)ational, (R)egional, S, X and unknown. Most of the incidents have a regional range which is the case for 63 percent of the incidents, followed by a national range for 27 percent of the incidents.

The fourth characteristic is the question ‘if the emergency number got hit’ during the failure. The number of times these answers occurred and the percentage are shown in Table A6.9. This is a variable that is different from the emergency number represented as a service. So the assumption is made that the service that got hit indirectly affected the emergency number during the incident. Four different answers to the question if the emergency number was affected are possible. The first answer is yes, the second is no, the third is does not apply and the last answer is unknown.

The fifth variable is ‘the cause’ of the incidents. Thirteen different causes are noted for all the incidents related to the organizations within the cluster. Third parties cause the most incidents at 33 percent, followed by hardware failures and hardware or software failures which are together responsible for 39 percent of the failures. For nine percent of the failures, the cause of the incident is unknown. Other notable causes are human at seven percent, and power failure at third party for five percent. The correlations between the cause of the incidents and the cluster can be found in Table A6.10

The sixth variable is the ‘hours immediately’, refers to the time between the start of the incident and when the incident is reported to AT. The correlations between the hours immediately and the cluster can be found in Table A6.11. This value is lower than zero for one incident, equal to zero for one incident greater than zero for 309 incidents and unknown for four incidents. So for 98 percent of the incidents, the hours are immediately known. The average number of hours for these incidents is equal to 166, 29 with a standard deviation of 1006. In addition is the spread for this average [0,35;8936,37]. The data on the hours immediately has a couple outliers, so the average is calculated again and the fours highest and lowest values are removed from the dataset. This resulted in

an average of 53,44 hours with a standard deviation of 110. The spread is equal to [0,77;882,73]

The final characteristic of the incidents within the dataset from the Reporting Desk is the 'duration of the incidents' in hours. The correlations between the duration of the incident and the cluster can be found in Table A6.12. The duration is defined from the starting time of the incident till the time that the failure is revolved. One incident within the dataset has a duration equal to zero, followed by 309 incidents where the duration of the incident is greater than zero, and the duration is unknown for five incidents. The average duration of the incidents is equal to 207 with a standard deviation of 3489. In addition, he spread for this average is [0,02;61344,73]. The data on the hours immediately has a couple outliers, so the average is calculated again and the three highest and lowest values are removed from the dataset. This resulted in an average of 6,1 hours with a standard deviation of 8,88. The spread is equal to [0,07;61,68]

Conclusion correlations Reporting Desk data
Multiple correlations and other values are calculated for the picked cluster. This paragraph provides a summary of the most useful information found for the organizations within the cluster. Starting with the two organizations that are together responsible for more than 50 percent of the failure, which are organizations V and Z. Followed by information of the data from the Reporting Desk that the failures of all the organizations came for 75 percent from the years 2016, 2017 and 2018. Next to the years is found that the services that got hit the most by failures are email, mobile internet, mobile telephony, broadcasting distribution and fixed telephony. The percentage of these services being hit by a failure is 10 or higher. The average number of affected customers is equal to 409.626 customers for the failures from which the number of affected customers is known. For 23 percent of the failures is the number of affected customer unknown. The range of the incidents is for 27 percent of the failures national and 63 percent of the failures regional. Next to the range of the failure is also noted if the emergency number is hit due to the failure, which is true for 29 percent of the failures. The two main causes of the failures are caused by a hardware failure or due to a third party. The final two characteristics of the failures are two different durations of time, which are the hours immediately and the duration of the failure. The average value of hours immediately is 166 and the average value for the duration of the failure is equal to 8.65 hours.

7.9 The influence of the results on the process of supervision

The method for the creation of the Exploratory Model consists of two different components; the clustering part, and the calculation of the correlations between the clusters and the data from social media as well as the data from the Reporting Desk. These generated two different sets of results beginning with the clusters and followed by the correlations. It is important to understand how these results improve the process of supervision. The goal of supervision is to improve the compliance behavior of the organizations under supervision. In this case, this is specific to the supervision on the continuity of the telecommunication networks and services. The results of the clustering algorithm could provide insights to the experts within the Supervision division regarding which organizations within the dataset with the survey scores responded with a similar

set of answer to each other. This allows the experts to execute the risk analysis not for a single organizations but for a subset for the dataset. Therefore, the resulted clusters could be directly used to support the process of supervision. The clusters created during this study could be used, however, it would be more precise to execute the cluster algorithm again for the whole initial databases with the survey scores. In addition, the expert could select certain questions from the survey and create clusters based on these questions. This could be utilised when an expert deems that specific questions from the survey are more important to define the risk of the organization than others. Additionally, the use of qualitative data of the expert could change how the cluster algorithm is executed. By executing the risk analysis on groups of organizations, it could it be easier for the expert to process more information from the survey scores in a shorter time span.

The results of the correlations are harder to directly apply to the process on supervision. This is because of the low coverage ratio between the organizations in the clusters and the organizations within the dataset of the Reporting Desk. It is not possible to create reliable risk profiles due to the lack of data for the organizations in the subsets of the clusters. Nevertheless, the results of correlations between one of the cluster for the dataset with the data from the organizations that responded to both surveys and the Reporting Desk are an example on what the possibilities are if there would be enough data to calculate correlations for all the created clusters. It could supply the inspector with information about failures that caused interruptions of the continuity of the telecommunication networks and services. Take as an example the cause of the failure, which is calculated for all the organizations within the cluster. A specific cause could score high for the organizations within the cluster, which could support the inspector during the inspections. The inspector could point out during the inspection that the inspected organizations should take action to reduce the number of failures caused by the highest scoring cause. Additionally, the results of the correlation between the social media data and the created clusters could be directly used by the experts of the Supervision Division Information Security. Nevertheless, the of the data from social media is only available for a small part of the initial datasets. The low number of data points for the creation of the clusters could mean that the survey scores of the organizations within the same groups are not as similar as the K-mean algorithm suggests. A way to define the reliability of the cluster is to look at the average difference between the organizations within the clusters, which can be found in Section 7.7, or to look at the scatterplots in Section 7.4.

8. Validation

The goal of this chapter is to test the assumptions made during the creation of the Exploratory Model are valid. This is done by conducting interviews with experts on the subject Information Security. Section 8.1 provides the validation of the assumptions made on the datasets with the survey scores, followed by the validation of the thresholds created of the data from social media (8.2), and section 8.3 highlights the validation of the assumptions made on the data from the Reporting Desk. Two interviews were conducted during the process of validation. The first interview is conduct with a senior inspector from the supervision division on Information Security for the validation of the assumptions made on the datasets with the survey scores and the thresholds created for the data from social media. Followed by an interview with an expert from the Spectrum Division Continuity on the assumption made for the data from the Reporting Desk. The experts are picked based on their expertise and knowledge about the supervision of the continuity of the public telecommunications services and networks and the datasets. An overview of all the consulted experts is shown is Appendix A7. The experts consulted for the validation are also shown in this table.

8.1 Validation on the datasets with the survey scores.

The validation of the assumption made for the dataset with the survey scores consists of four different parts. **The first step** is selecting the appropriate organizations from the datasets with the survey scores that corresponds with the names of the organizations that are mentioned in combination with a failure on social media. The website from “alle storingen” is used to pick the organizations that are mentioned on social media. A decision had to be made on which organizations names within the datasets with the survey scores match with the names of the organizations on the website from “alle storingen”. The database with the survey scores includes some organizations that have a variation of very similar names. This could be the same organization that replied to the survey multiple times, or different organizations with almost the same name. A possible reason for organizations having very similar names could be that some organizations are daughter companies of the original mother company. In these cases, the annual turnover of each organization was compared, and the organization with the highest annual turnover was assumed to be the mother company, and therefore, selected for analysis. When the annual turnover of the organizations is equal, the organizations with the most similar names to the organization’s name on “all storingen” was selected.

The exact names of some organizations mentioned on “alle storingen” are presented multiple times within the dataset. This can be the case when an organization replies to a survey multiple times. When this is the case, the latest survey score was used for the creation of the clusters.

Two questions were asked during the interview:

1. *Are the most suitable organizations from the database of surveys that correspond with the organizations from "alle storingen" selected?*
2. *By selecting the latest version of the survey scores for the organizations that replied to the same survey multiple times, are the most suitable survey scores for the organizations selected?*

Expert H indicated that the most suitable organizations that correspond to the organizations from "alle storingen" are selected from the database. This is because the data from the mother company is accurately selected, rather than that of the daughter companies. Furthermore, selecting to use the latest version of the survey scores is supported as the correct decision. This is because it provides the most recent data regarding the status of the organization on compliance behaviour.

The **second step** discusses the elimination of columns from the initial dataset on the subjects Information Security and continuity to create the first version of the dataset during the first part of the data cleaning. In essence, this first part of data cleaning removes certain questions and other information, which was deemed as not useful information of the creation of the exploratory model. This decision is based on a personal opinion that is created through conversations with supervisors and inspectors.

The question to the experts was:
For the risk analysis that they normally execute, is the removed information considered important to define the risk of an organization?

Expert H supports the decisions made by commenting that the removed information from the database in not usable or critical for defining the risks that the organization poses. When looking at the datasets for the first version of the dataset on the subject Information Security and continuity, the expert expressed the opinion that more questions can be removed from the dataset. For the dataset on the subject Information Security, the second question could be removed and, for the dataset on the subject continuity the second, third, fourth, sixth and thirty third be removed. All these questions are considered as irrelevant by the expert from the Supervisory Division Information Security.

The **third step** is on the blanks within the dataset with the survey responses. A lot of the answers within the datasets are blank and the assumption is made that the organizations did not reply to these questions. Section 6.3 of the EDA indicates that these blanks were mostly present in the initial dataset on the subject Information Security. So the final assumption made is that all blank values represented unanswered questions.

The question to the expert:
Do all the blanks within the dataset classify as unanswered?

Expert H responded that this may not be the case for all the blanks within the datasets. Some of the blanks represent that a certain question was not applicable for some

organizations. One of the options within some questions of the surveys is actually ‘does not apply’. For example, an organization may answer a question within the dataset which generates instructions to skip the following question. The skipped question within the dataset is not marked as ‘does not apply’ even though one of the optional answers to such a question includes does not apply, and some organizations choose to select this option. As a result, a blank within the dataset could be unanswered by an organization or does not apply.

8.2 Validation on the datasets from social media

The data from social media consists of the amount of social media posts per day with a mention of a failure by an organization. Thresholds were set by AT for some of the organization, but not all the organizations. According to AT, a failure occurs when the threshold for the organization is exceeded. The thresholds which are not set by AT are based on the comparison of the average number of mentions per day between 2012 and 2019 for the organizations with the thresholds and without the thresholds. Organizations with a similar average number of mentions were given the same value for the threshold. Therefore, the assumption made is that all the organizations were active during the time period of 2012 to 2019.

The question asked to the expert:
Are the defined values of the thresholds for the organizations reliable and realistic?

Expert H responded with stating that the values of the thresholds are realistic and well defined. He would not change any of the value. Remarkable is that all the lists of the new values scores low compared with the other thresholds. Only the threshold of organization JJ scores high.

8.3 The data from the Reporting Desk

For the data from the Reporting Desk, three different assumptions made. **The first assumption** is on the number of incidents during the year 2015. The dataset consists of data from the years 2015 to 2019. The number of reported incidents for the year 2015 is around 75 percent lower than the number of reported incidents for the following three years. Reporting incidents is completely voluntary for the organizations, so the assumption was made that during the year 2015, the urge to report was lower than during the following three years.

The question to the expert of the Spectrum Division Continuity:
Can the assumption be made that the urge to report was lower in 2015 compared with the period from 2016 to 2019?

The response of expert I was that in 2015, only the number of reports was considered, whereas from 2016 the service interruptions were counted separately. A report of an incident can contain multiple service interruptions. The table supplied for these interruptions contains the interruptions per service for the years from 2016 to 2019. For the year 2015, it is only the reports which may contain information regarding multiple

service interruption. Thus, due to differences in the reporting method between 2015 and the other years, it is difficult to cannot retroactively define the number of interruptions for 2015.

The **second assumption** regards the affected customers. If the number of affected customers is equal to the value minus one, than it is assumed that no customers have been affected because of the failure.

The question to the expert of the Spectrum Division Continuity:
Can the assumption be made that the value of minus one in the dataset means that no customer is affected by the failure?

Expert I responded that minus means unknown, similar to blank values and the value unknown in the dataset.

The **final assumption** is on the question in the dataset of whether the emergency number is affected by the failure. The assumption is made that this question regarding whether the emergency number is affected refers to an indirect consequence that arises from the malfunction of another service Because the main type of service that is hit by a malfunction can also be the emergency number. So the assumption is made that when the type of service hit is the emergency number than it only concerns the emergency number and the column with the question: if the emergency number is affected is it a result of a malfunction of another service.

The question to the expert of the Spectrum Division Continuity:

Can the assumption be made that the question on if the emergency number is affected is an indirect consequence that arises from a malfunction that primary hits another service*?*

Expert I responded that this is right assumption for the meaning of the data within the column with the question: if the emergency number is affected?

9. The role of the involved stakeholders

Chapter 9 is the start of second part of the research which consider the implementation of the data-driven risk-based model. This part highlights the final two challenges found in the article from Eggimann (2017) which consider the availability of the data with the organization and the ability to create a process with a high cost-effectiveness ratio. The initial research approach was aimed on the first two challenges of Eggimann (2017). Due to quality and quantity of the data available is the research shifted to the implementation of the model.

The aim of this chapter is to explain the role of the stakeholders involved in the regulation of the continuity of the telecommunication services and networks provided by organizations within The Netherlands. There are two levels of abstraction for the stakeholders. The first level of abstraction is the level of AT as a whole, in other words, on organizational level. The stakeholder analysis starts in section 9.1 with some extra literature research which was need to gain information about organizational structures. These consist out of governmental bodies and organizations involved in the regulation process. These stakeholders are called the external stakeholders and they can be found in section 9.2. The other level of abstraction for the stakeholder are the stakeholders within AT on the subject continuity. In section 9.3, the internal stakeholders are elaborated. Lastly, the organizational structure within AT is explained in section 9.4.

9.1 Vertical vs. horizontal organizational structure

There are two different kinds of structures possible within an organization. The first is a vertical structure, which is often chosen for businesses with a large number of employees. The typical structure can be likened to a pyramid, with executives at the top, mid-level managers in the middle, and employees at the bottom (Quain, 2019). The executives make the major decisions, and communicate those to the mid-level managers who are responsible for implementing the given decisions and communicating the changes to the employees. The employees must then adapt their daily tasks to align with the goals of the executives. This structure is also often referred to as a top-down organizational structure. It is hard for the employees to make suggestions that reach the top level of the pyramid because of strict protocols in the chain-of-command.

The second kind of organizational structure is for companies that aim to develop a culture that harnesses the creativity and imagination of the employees, and empowers them to perform their tasks without micromanagement(Quain,2019). This structure provides employees the autonomy to make their own decisions. It is also typical to have only a few managers, and to remove the barriers between the employees and any manager. The main characteristics of the horizontal structure within an organization are an emphasis on teams, and the exchange of ideas and collaboration.

Masahiko Aoki (1986) studied the role of both structures within organizations. He made a comparison of industrial organizations in the U.S. and Japan by looking at hierarchical control and horizontal coordination. Hierarchical control corresponds with the vertical

organizational structure and horizontal coordination with the horizontal structure. Decisions that response to a high degree of irregularity and that these decisions are involved with a high degree of uncertainty are decision placed under hierarchical control. These decisions are often referred to as strategic decisions and function to provide the overall framework for production. On the other hand, efficiency for the organization with the vertical structure is often obtained by on-the-spot knowledge and rapid problem solving through practical, hands-on learning.

There are also some issues related to the vertical structure. A poor quality of information could lead to managers being slow and imprecise towards the recognition of problems that need to be tackled. Furthermore, excluding subordinates in the decision-making could create a lack of motivation for the subordinates to report problems. In addition, it could it counteract the implementation of hierarchical orders due to a lack of motivation. Thus, the implementation of hierarchical structure could create costs of monitoring due to the bounded rationality of the supervisors, and costs of implementation stemming from a lack of incentive for subordinates..

Horizontal coordination corresponds with the horizontal structure within an organization and is emphasized after the framework for production is laid down. The structure can be split in horizontal information exchanges and semiautonomous coordination of operations by relevant subordinates. Aoki (1986) concluded during his case study that horizontal coordination within an organization facilitates a smooth production flow because there is no intervention by a supervisor. In addition, the efficiency of the firm with the horizontal structure is attained by rational technocratic control. On the other hand, subunits are able to coordinate their decisions with the other subunits, which is consistent with the purpose of the organizational structure, however, is limited by their lack of understanding of the finer mechanisms of all the operations within the firm. A way to combat this issue is to learn by doing, but this is costly in terms of time. Furthermore, the lack of centralization of information on events that affect different subunits could have a negative influence on the problem solving capability of the subunits.

Nunan (2012) studied the organizational challenges of policy integration for governments with a vertical structure and a horizontal structure. The paper concludes that a vertical approach indicates political commitment, which could lead to resources and support from the government, but may also leave it vulnerable to the whims of the government, threatening durability. Contrastingly, the horizontal approach could improve the chance of an effective implementation due to the integrated technical and analytical expertise, however, resources and mandates are needed to execute this approach effectively. Due to these merits and drawbacks, Nunan (2012) concludes that the best approach is a hybrid model where the horizontal approach is supported by a top-down commitment.

9.2 External stakeholders on the supervision of Information Security

Radiocommunication Agency Netherlands (AT)
AT is a Dutch regulator that supervises compliance with laws and regulations. It regulates the government policy on subjects related to wireless and wired telecommunications. In total, thirteen supervision fields are regulated by AT. The continuity of telecommunication and networks is a subject within the supervisory field Information Security. It grants licenses for the use of frequencies for telecommunication services such as the (mobile) internet and telephony.

The Dutch government sets laws and regulations on telecommunication services, within the framework provided by the European directives. It consists of requirements to be able to offer public telephony, internet access and (wireless) networks. The four main subjects of the telecommunication law regarding Information Security is authorised wiretapping, security of telecommunication data, continuity and keeping privacy. The European Union has set a legal framework for telecommunication that AT needs to regulate in their national laws.

Ministry of Economic Affairs and Climate
The national policy within The Netherlands is set by the Dutch government. The Ministry of Economic Affairs and Climate is one of the ministries within the Dutch government that sets the laws and regulations for the whole nation. The legislation set by the Ministry of Economic Affairs and Climate Policy on telecommunication is mainly based on the guidelines set by the European Union. The main responsibility of the ministry is to ensure that the citizens feel secure about the services provided by companies in The Netherlands. They are responsible for directing the legislation on the companies that provide services to citizens and other businesses. The primary objective of the Ministry of Economic Affairs and Climate is to create a sustainable and entrepreneurial Netherlands; a climate neutral society and a strong, open economy.

Telecommunications providers
A telecommunications service provider has traditionally provided telephone and similar services. This kind of communication provider includes local exchange carriers and mobile wireless communication companies. It excludes internet service providers, cable television companies, satellite TV, and managed service providers. They provide mainly telephone and communication services related to telephony. Services are provided to businesses and individuals. In the past, most telecommunication providers were government owned, but recently there are predominantly private players in The Netherlands.

Network providers
A network service provider is a business or organization that sells bandwidth or network access. This could be done by providing direct internet backbone access to internet service providers or access to its network access points. For such a reason, network service providers could also be referred to as internet providers. Network service providers may consist of telecommunications companies, data carriers, wireless

communications providers, Internet service providers, and cable television operators offering high-speed Internet access.

9.3 Internal stakeholders on the supervision of Information Security

Monitoring and Analysis Center
Within AT there is a team called Monitoring and Analysis Center (M&AC). This team consists of twelve people who manage and monitor all the data collected within the organization. This is data from all the thirteen supervisory fields. Some examples are data from social media, the use of frequencies for telecommunication networks, and inspection results. Data management is an administrative process that includes acquiring, validating, storing, protecting, and processing the data collected within the organization. This is done to ensure the accessibility, reliability, and timeliness of the data for its users within the organization.

The data that they monitor is data from the frequency spectrum. There are two different ways they collect this data. The first is data collected from radio-receivers of thirteen different static measurement points throughout the country. The second method of collecting data is through mobile measurement equipment, which is installed in the cars used by inspectors. These inspectors travel through The Netherlands to visit organizations which they need to inspect, and during their commute, the measurement equipment installed in their cars collects data on existing frequencies.

Two different forms of analysis are conducted by M&AC. The first is standard analyses done by a business analysis tool called Watson, which is a software package from IBM. It is a cognitive system that guides end users in analyzing data. The end users are the experts and inspectors who make use of the analysed data. The analysis in Watson are Business Intelligence based, which means it is focused on collecting and analyzing data from different systems to make it usable for inspectors, experts and management. These analyses are called standard analyses because they are continuously conducted by standardised analysing processes of Watson when new data is stored in the databases.

The second kind of analysis is custom analyses, ordered by experts. All the analyses conducted by M&AC are executed by analysts that are commissioned by the people outside of M&AC. All members from the Supervision Divisions can submit a request for an analysis at M&AC, which will be executed if possible. These analysis are not done through Watson, but via other software packages or programming languages

Enforcement Policy Division on Information Security
The Enforcement Policy Division on Information Security is responsible for the policy of the supervision on Information Security. Enforcement policy on supervision is about how could the compliance behavior of the supervised organizations being increased. There are two different kinds of supervision for the Enforcement policy division. The first one is called comply supervision, which it not used for the supervision on Information Security. The second kind of supervision is ‘quality supervision’, which is used of the supervision on Information Security. The enforcement police team wants to sustain “quality supervision” for the supervision field Information Security, especially on the supervision

of continuity of the telecommunication network and services. This kind of supervision aims to create a system that tackles socials problem related to Information Security. According to the experts from the Enforcement Policy Division on Information Security is the reason for ‘quality supervision’ to support the creation of a well-organized and reliable telecommunication infrastructure. In addition, the experts acknowledge the importance of the data from the Reporting Desk. Their vision is that reports of failures could be used as a learning curve to improve the created system, which is the basis of ‘quality supervision’.

Supervision Division Information Security

There are multiple Supervision Divisions within AT. One of these is the Supervision Division Information Security. This team includes inspectors on the subject Information Security. One of the goals of the Supervision Division Information Security is to increase the compliance behaviour of the organization within their supervisory field. As explained in Chapter 3, AT uses risk-based regulation to distribute their resources. This is executed by the experts to determine the distribution of the inspectors over the organizations. The experts and inspectors of the Supervision Division Information Security partly make use of the collected data, and the other component of the decision making is subjective. They also assess themselves what the importance of the collected data is .

Only a limited set of criteria are used for risk analysis, and these criteria are not calculated based on all the collected data. The risk analysis takes into account a subset values from the databases, and if organizations demonstrate values that are deemed to pose a higher risk, they become subject to inspection. Surveys distributed to all regulated organizations are only reviewed on a subset of the questions asked, not on the complete set of questions. AT is interested to learn whether a model would substantiate their personal judgement, or provide new insights .

Spectrum Continuity

This team on Continuity is situated elsewhere within AT, outside the Supervision Division. The Reporting Desk where the telecommunication providers and network providers need to report certain incidents is part of this team. It sits within the executive branch of legislation created by the national government and develops protocols towards the laws created. One of the protocols is the obligation of organizations to report incidents on the continuity of their services. One challenge is that reporting is voluntarily, based upon an assessment of whether a failure is significant enough to be reported. This results in some cases where there is no official record regarding an incident of discontinuity, as the organization did not deem it significant enough to report.

The Supervision Division Information Security is independent from the Spectrum Division Continuity to guarantee an unbiased Supervision Division. The Spectrum Division Continuity normally does not share any data with the Supervision Division Information Security on subjects such as discontinuity incidents. They call this the institutional border between the Supervision Division Information Security and the Spectrum Division Continuity on the continuity of the telecommunication networks and services. The purpose of this organizational divide is to encourage the monitored organizations to be more willing in cooperating and reporting incidents. If the Spectrum Division Continuity needs to obtain

more information about a certain organization, they can engage the Supervision Division Information Security to obtain that information. The Spectrum Division Continuity also holds the power to execute reactive research. An example of this is a report about blackouts and the impact of the blackouts on certain parts of the society such as telecommunication services

9.4 The organizational structure of AT

A regulator has extensive information within their organization. This information is referred to as data, and different kinds of data exist. The most important distinction is the difference between qualitative and quantitative data. Qualitative data is often unstructured and more difficult to analyze. Nevertheless, experts and inspectors utilize this kind of data to execute risk-based regulation. The new challenge that a regulator is facing is the collection and management of quantitative data and how to incorporate this data into risk-based regulation.

A data-driven model could support the risk analysis of the experts and in the future, potentially substitute the manual risk analysis executed by experts. Implementing the created model into the risk-based regulation process within the supervisory field will be crucial for the cost-effectiveness of the model. The stakeholder analysis revealed two primary challenges within the organization. The first challenge is the institutional border between the Spectrum Division Continuity and the Supervision Division Information Security. This boundary for information sharing mainly revolved around the subject continuity of the telecommunication networks and services. Creating a model based on data from two different sources; the surveys and continuity failures, brings the two databases together. This conflicts with the currently existing institutional border and national policy of the government, and is thus the most significant challenge in implementing the model within AT.

The second challenge is developing agreement with regards to the implementation of the model between the three different teams within AT. The user of the results of the model are the inspectors within the supervisory field of Information Security, and they will use the collected information to support their decision making in risk-based regulation. The manager and monitoring team of the model is the M&AC within AT. They will run the model with the new data provided by the Supervision Division Information Security. They also have the task of maintaining and updating the model. Updating the model can be done by expanding the analysis to create the model. The model is not a self-learning model, which means it does not update every time new data is used as input. Furthermore, the Spectrum Division Continuity needs to be willing to share information about the failures. This is difficult due to the institutional border, and there are concerns that the regulated organizations will become hesitant to voluntarily report their failures to AT.

10. The Implementation of the data-driven model

This chapter aims to elaborate the two main challenges of the implementation of the model within AT. An expert from all the internal actors is interviewed about their opinion on the implementation of the model. Two main challenges are discussed during the interviews. The first challenge for the implementation of the model is the availability of the data within the organization which is discussed in section 10.1. The second challenge which is discussed in section 10.2 is about which internal actor could have the ownership of the model within the organization. It is key to find the owner for which the efficiency of the model will be the highest in order to improve the cost effectiveness of the model.

10.1 Data sharing within an organization with institutional borders.

Three different sources of data are used to create the Exploratory Model all these data sources are managed by three different internal stakeholders. One of these stakeholders could be the owner of the model, but all the data needed for the Exploratory Model must be available for the internal stakeholder that is the owner of the model. The first stakeholder is the Spectrum Division Continuity on continuity of the telecommunication networks and services. They manage the data collected by the Reporting Desk, which is situated within the Spectrum Division Continuity. For this stakeholder, are all the datasets available. Following this, the Supervision Division Information Security is the owner of the dataset with the information from the different surveys. For this stakeholder, the data from social media is available, however, they do not have access to the data from the Reporting Desk. This is due to the institutional border between the Reporting Desk and the Supervision Division Information Security. Finally, the M&AC is the owner of the data from social media, and for them, all the data is available. So is the Supervision Division were to be the owner of the model, the institutional border between the Spectrum Division Continuity and the Supervision Division would pose an issue. Solutions for this will be discussed in Section 10.2.

10.2 The owner of the model within the organization

Four different internal actors are identified for the creation and the implementation of the model during the stakeholder analysis in section 6.2. One of these actors needs to be the owner of the model in order to implement it. There are two main challenges to overcome in order to be the most suitable owner of the model. The first challenge for the ownership is related to the institutional borders discussed in 10.1 and pivots upon the availability of the data for the owner of the model. This leads to the key question of whether all the data which is needed to create the model is available for the owner of the model. The second challenge for ownership is that the right expertise needs to be available within the internal actor to monitor, maintain, and execute the model. An expert within the internal actor needs to be appointed as the person ultimately responsible for the model. All four internal stakeholders are considered as possible owners of the model. Table 13 provides an overview with the advantages and the disadvantages of the possible owners of the model.

Two of the four internal actors are considered as the least suitable candidates for the ownership of the model, and are the Spectrum Division Continuity and the Enforcement

Policy Division of the supervision on Information Security. The main reason why these internal actors are not considered as suitable owners of the model is that they are not directly involved with the supervision of the telecommunication network and service providers. The model is made to support the process of supervision on Information Security and both teams are considered too distanced from this process. They are both indirectly involved and although they may have the capacity to overcome one or both challenges mentioned in the previous section, it would not be fitting for either of them to own the model.

This then leaves the other two internal actors, the Supervision Division Information Security, and the M&AC, to be the possible owner of the model. The first option is the Supervision Division Information Security. This team is directly involved with the supervision on the subject Information Security. The model is created to support risk-based regulation executed by the experts within the Supervision Division Information Security. So it is necessary to determine whether it is possible for the Supervision Division to meet the requirements to be a suitable owner for the model. The first requirement arises from the availability of the data that is needed to manage the model. Data from two of the three sources is available to the Supervision Division Information Security, and these are the datasets from the surveys they manage. The datasets from the reports on the discontinuities of the services and network, however, is not available for the Supervision Division. This data can be made available to the Supervision Division in three different ways. The first one is to remove the institutional border between the Spectrum Division Continuity and the Supervision Division. This makes the data available to the Supervision Division while the Spectrum Division continues to be the owner of the dataset. The second option is to move the ownership of the dataset including the tasks related to managing the dataset from the Spectrum Division to the Supervision Division. This makes the data available for the Supervision Division and could lead to a more cost effective process of supervision. The final option is to aggregate the data in a way that it is untraceable for the expert of the Supervision division Information Security. The downside of aggregating the data is that information will be lost and the value of the data towards the created Exploratory Model will decrease.

The second requirement demands an expert on data-driven models within the Supervision Division. This is needed to improve and use the model to support the Supervision Division. A person with the knowledge of the model within the Supervision Division could improve the efficiency of the model on the process of supervision because there is a direct relation between the expert of the model and the process of supervision. This expert would only work on data-driven models for the subject Information Security, which will lead to an improved version of the Exploratory Model and the possibility to create other data-driven model that will support the process of regulation.

The second suitable situation is for the owner of the model to be the M&AC. This team consists of multiple experts with expertise on data analytics and the creation of data-driven models. Thus, they are the experts on data managing and analysing within AT, which directly overcomes the second challenge of needing a specialised operator for the model. The first requirement can also be met by this internal stakeholder because all the

datasets are available for them. The institutional border is not set between the Spectrum Division Continuity and the M&AC. The only drawback of assigning the M&AC ownership of the model is their distance from the supervision on Information Security which manages and analyse the data for all the supervisory field. As a result, the M&AC has a lower level of expertise for the subject Information Security.

The opinion of the interviewed stakeholders is that the final actor would be the most suitable option in terms of the current organizational structure within AT. This is largely in part due to the lack of an institutional border between the Spectrum Division Continuity and the M&AC, which facilitates data sharing between the two teams. The other reason is the lack of knowledge within the Supervision Division regarding data-driven models. Hiring an expert on data-driven models for the Supervision Division is not possible for AT at this moment due to the associated costs. M&AC has four people working on managing and analysing data for thirteen different supervisory field at this moment, making it unrealistic to hire just one expert for the Supervision Division on Information Security.

Owner	**Advantages**	**Disadvantages**
M&AC	Data availability	Not directly involved with the supervision on Information Security
	Knowledge about data analysis or data-driven models	
Supervision Division Information Security	Directly involved with the supervision on Information Security	Institutional border for data sharing with Spectrum Division Continuity
		No knowledge about data analysis or data-driven models
		The use of aggregated data
Spectrum Division Continuity		Not directly involved with the supervision on Information Security
		Institutional border for data sharing with Spectrum Division Continuity
Enforcement Policy Division		Not directly involved with the supervision on Information Security
		No knowledge about data analysis or data-driven models

Table 16: an overview of the advantages and the disadvantages for the possible ownership of the division involved in the process of the supervision on Information Security

11. Discussion

This chapter discusses the interviews conducted on the subjects validation and implementation. Furthermore, the reason of why a verification of the Exploratory Model is impossible is highlighted. Four interviews are conducted on the implementation of the model; two on the validation of the model, and one new interview is conducted with an expert of the supervisory with the highest development of data-driven model to support their process of supervision. The first discussed interviews are the interview on the implementation of the model. Experts from all the internal stakeholders were interviewed on this subject and they all shared a common opinion. This was that the M&AC would be the most suitable owner of model when taking the current organizational structure within AT into consideration. This opinion was formed due to the possibility of sharing of data between the Spectrum Division and M&AC as there is no institutional border between these two team. The other two sources of data are also available to the M&AC. In addition, the team at the M&AC possess the expertise to manage, monitor and execute the model on the available data.

The other viable option of the ownership of the model is the Supervision division on Information Security. All the interviewed experts from the internal stakeholders share the same opinion on them having the ownership of the model. This internal stakeholder lacks knowledge about data-driven models. Hiring an expert on data-driven models for one specific Supervision division is not possible for AT at this moment. The costs are too high for the organization to make that decisions. M&AC has four experts working on managing and analysing data for thirteen different supervisory field at this moment. This makes it unrealistic to hire one expert for the Supervision division on Information Security, but it could be an option in the future. The second requirements that should be met is the availability of the data. For this subject, the interview with the Enforcement Policy Division on Information Security was insightful.

Two different options of making the data available to the Supervision Division Information Security were discussed, starting with removing the institutional border between the Spectrum Division Continuity and Supervision Division Information Security. This is something that was already expressed in 2015 by an expert within the Enforcement Policy Division on Information Security, where the opinion of the expert was that data sharing between the Spectrum Division and the Supervision division could lead to a higher compliance of behaviour. The data from the Reporting Desk is seen as data that could be used as a learning curve by the experts of the Supervisory Division Information Security. The institutional border between the two internal stakeholders is set by AT and not by the Ministry of Economic Affairs and Climate Policy in consultation with the supervised organizations. This was established because the supervised organizations were hesitant that the data of self-reported failures would be used against them. The importance of the supervision on the continuity of the telecommunication networks and services has increased immensely since the decision to share less data between the Spectrum Division and the Supervision Division. Thus, the data sharing between the stakeholders should be revised. The institutional boundary could only be removed by the Spectrum Division Continuity.

The second option discussed for making the data available to the Supervision Division Information Security was to relocate the Reporting Desk from the Spectrum Division to the Supervision Division Information Security. The decision to assign the Reporting Desk to the Spectrum Division was made fifteen years ago. In contrast, newly created Reporting Desk such as the one created for failures that caused discontinuity of energy supplied by power plant is created three years ago. The Reporting Desk for this area was placed in the Supervision Division and not the Spectrum Division with the reason that the data from the incidents could improve the process of regulation. Putting the Reporting Desk outside the Supervision Division could be outdated due to the importance of the supervision on the continuity of the telecommunication networks and services. However, it is structurally difficult to move the Reporting Desk, which could only be done by the director of AT. This could also be demotivating for the Spectrum Division as work that they take ownership of, would be taken away from them.

The other subject discussed with the Enforcement Policy Division Information Security are the legal instruments and other instruments to increase the volume of the database from the Reporting Desk. A legal instrument to increase the volume of the database could be imposing sanctions to the supervised organizations that do not report their failures. This is hard to implement, however, as the current laws are unclear about when failures should be reported. In addition, there is no specific rule for when organizations should report their failures. The Spectrum Division decided this together with the supervised organizations and individually for every organization. This makes it hard to obtain information from all the supervised organizations because the thresholds need to be defined for all the organizations separately. The other factor of the law that makes it hard to obtain information about failures is that when a supervised organizations uses the services or networks from another company than they do not need to report the failure if the failure is caused by the organizations that they use the services or networks from. Fifty percent of the clients of this organization could be hit by this failure, which, for the larger company, may only represent two percent of all the people that they serve. This means that the failure could be significant enough to report for the smaller company, but does not get reported because the failure is caused by the company that they use the services or networks from. Furthermore, many larger organizations do not report the incident because the failure is not considered significant enough for them to report. This leads to the failure never being reported to the Reporting Desk.

The first interview was conducted with an expert from the supervisory field Information Security. The assumptions made for the dataset with the survey scores and the defined thresholds for the data on social media were discussed during the interview. According to the expert, there is one issue with one of the three assumptions made, which regards the assumption on the blanks within the dataset. These blanks are considered as questions that were unanswered by the supervised organizations. This is not the case because it could also mean that the question did not apply for that specific organization. For future use of the dataset for the creation of a model or other analyses, it is important that the data is saved separately for unanswered questions and questions that are not applicable for an organization. This is important for maximizing the value of the data for

the organizations. In addition, thresholds are created for the data for social media and validated by the same expert. The defined thresholds are realistic, but relatively low. Low thresholds mean that the average number of mentions for a failure for a specific organization is also low. A low threshold makes it difficult to determine a failure for an organization because the word failure and the name of the organization needs to be in a social media post together. This could also be a positive tweet instead or a negative tweet, so the low threshold makes it harder to define if a failure actually happened.

The second interview conducted for the validation was conducted with an expert from the Reporting Desk on the assumptions made for the data from the report desk. The first assumption made for this dataset was that the urge to report lower was in 2015 than in the period from 2016 to 2019. This is not the case because in 2015, only the number of reports was considered as an failure, despite the fact that a report from a supervised organizations can contain multiple service interruptions. From 2016 not only were the number of reports saved, but the service interruptions were counted separately. This calls into question the reliability of the data from the Reporting Desk that is reported before 2016. In addition, the assumption that there are no affected customers when the value for the affected customers is minus one is false. Therefore is it important that the database manager from the Reporting Desk saves the data the right way with the dataset. In order to maximize the value of the data.

The final interview is an interview with an expert from the Supervision Division on the law exchange of information above and below ground networks (WIBON), also known as the excavation contract. This law obliges diggers to report any "mechanical grounding", such as digging, digging in saints, dredging and laying pipes. Cable and pipeline managers must have all their (underground) cables and pipelines digitally available within specified accuracy during the act of mechanical grounding. Since October 1, 2008, it is mandatory to submit an excavation report to the Land Registry Office for every "mechanical earthwork". This supervisory field gradually adapted data-driven risk-based regulation over the past ten years. The main subject of the interview was to develop a better understanding on which factors had the most influence during the transition from risk-based regulation to data-driven risk-based regulation and how this relates to the supervisory field Information Security.

Three different developments ensured the transition to data-driven risk-based regulation over the last ten years. The first development was the change in the national policy set by the Ministry of Economic Affairs and Climate Policy. It became an obligation for the excavators to report an excavation to the Land Registry Office. This increased the volume of data collected by the Land Registry Office to a volume high enough for the creation of a data-driven model. The second development is the higher standards for the reportation of excavation damage caused by the supervised organizations. Changes in the national policy allowed the WIBON to use certain legal instruments to set these standards. Excavation damages need to be reported to the Land Registry Office, so these new standards increased the volume of the database from the Land Registry Office with data about failures.

Data about the supervised organizations, the excavations and the excavation damages is all available in high volume to the WIBON. There is no institutional border between the WIBON and the Land Registry Office, which creates full transparency between the supervised organizations and the Supervision Division. The final development is that an expert of the M&AC took this data and created a data-model that supports the supervisory process. The use of this model within the WIBON created more cost efficient process of supervision. It enables the Supervision Division to appoint resources of supervision to the organizations that pose the highest risk-based on the outcomes of the data-driven model.

The fourth sub question of this study is on the validation and the verification of the Exploratory model. Chapter 8 discusses the validation of the Exploratory Model, but the verification had not been executed. This is because it is impossible to verify the created Exploratory Model and the results of the calculated correlations. It is impossible due to the lack of data from the sources; the Report Desk and social media. The low number of organizations in these datasets leads to a lower number of organizations in the datasets with the survey scores. Low number of organization in the dataset with the surveys scores lead to small clusters. For these cluster is only data from social media available for all the organizations and not from the reporting desk. The low number of organizations in the datasets that are used to create the clusters causes that the outcome of the verification of the exploratory model is not substantial.

12. Conclusion

1. *What is the quality of the available quantitative data within the supervisory field information safety? Is the data usable for the creation of a data-driven model?*

Data regarding the subject the Continuity of the services and networks provided by providers of public electronic communications networks and services are sourced from three different areas within the organization. All these data sources are owned by different internal stakeholders. The first data source are the outcomes of the surveys, which are owned by the Supervision Division Information Security. This data source consists of two different datasets with the outcomes of two different surveys. The first survey is on the subject Information Security and comprises of questions about all the four main subjects of Information Safety. Continuity is one of these subjects, so a part of the survey is about continuity. The other survey is on the subject Continuity and only assesses subject Continuity. The dataset on the subject Information Security consists of data from 1327 organizations and the dataset on the subject continuity consists of data of 457 organizations.

The EDA is executed for both datasets with the survey scores. This begins with a minimization of the residual, which are the values that deviate from the expected values. This initial process is often referred to as data cleaning, and consists of an additional two steps; high percentage of questions that were unanswered by the supervised organizations found. During the third part of the data cleaning were all the questions that are unanswered by 50 percent or more of the organizations removed from the dataset. The purpose of the data cleaning is to improve the efficiency of the algorithm that will be executed on the datasets. Six different dataset are created from the 2 initial datasets of the survey scores. All six the datasets can be used for the creation of the exploratory model.

The second data source is data from social media mentions regarding failures. This data is scraped from social media for 38 of the 1000 regulated organizations because there is only data on social media for these organizations. The dataset consists of the number of social media posts per day that a failure is mentioned per organization, between January 2012 and ending in May 2019. Experts from the supervision team on information safety defined a threshold for the amount of social media posts an organization can be mentioned in before a failure is considered to have occurred. The thresholds are specifically and independently set for all the organizations, so the value of the threshold varies. An EDA is not executed for this dataset because it only consists of the number of social media posts per day and the thresholds differ for each organization. Even if an EDA is performed with this data, only a small amount of information could be returned, and would not be of any interest for the creation of the Exploratory Model.

The final dataset is the data from the Reporting Desk, which is part of the Spectrum Division Continuity. This dataset consists of data regarding self-reported failures from 16 different organizations, and data of 12 of the 16 organizations are also present in the other two datasets. The data is derived from information about failures that occurred in

the years between 2015 till 2019, and includes information about the duration of the failure, the causes, and other properties of failures that cause interruption of telecommunication networks and services. Due to the low number of organizations represented in this dataset, it is more difficult to use this dataset for the creation of the Exploratory Model. In addition, the data currently within the database must be frequently updated.

2. *What is the preferred modeling technique to create a model when taking the quantitative data into consideration? Is it possible to create the model with the available data?*

An Exploratory Model is created with the available data, which is a prototype of the final model. A modelling method that is applicable to all the available data is created. The first method that is used for the creation of the data-driven model is the K-means clustering algorithm, which is used to cluster the organizations based on their survey scores. The cluster algorithm is executed on the two different versions for all the datasets with survey scores. The effectiveness of the cluster algorithm increased for the second version of the dataset on the subject Information Security and the dataset with the data from the organizations that responded to both surveys. In addition to the effectiveness of the K-means algorithm, the influence of the questions of the surveys is investigated to identify which questions the most influence have on the creation of the clusters. These questions are the risk criteria and can be found in Appendix A4. These risk criteria are presented to the experts for validation during an interview. The outcome of this will be explained during the conclusion of the next sub question.

After the creation of the clusters, different correlations between the created clusters and the data from social media and the Reporting Desk are calculated. The data from social media can be used for all the created clusters because there is data available from social media for all the organizations in the datasets with survey responses. There is a bigger difference between the value of the failures per organization per cluster for the first version of the dataset on the subject Information Security that the second version of the dataset. The decrease of unanswered questions within the dataset and the removal of questions that received similar responses resulted in a smaller difference for the failures per organization for the created clusters. This is also the case for the other two datasets and for the dataset on the subject continuity became value for the number of failures per organizations almost equal to each other for both datasets.

Correlations are also created between the survey scores and the data from the Reporting Desk. This is done for only one cluster because of the low number of organizations represented in the dataset from the Reporting Desk. The selected cluster is the cluster with the highest percentage of organizations present in the both the cluster and the data from the Reporting Desk. Six of the eight organizations in the third cluster on the second version of the dataset are also part of the dataset from the Reporting Desk. The correlations are calculated for these six organizations to provide an example of what kind of information could result from calculating the correlations between the survey scores and the data from the Reporting Desk. The results demonstrate that organizations V and

Z are over 50 percent responsible for all the failures caused by the six organizations. Furthermore, 75 percent of failures occurred between the years 2016 and 2018. The service that is compromised the most often is fixed telephony, and third parties and hardware failure cause the most incidents. The other calculated correlations can be found in section 9.2.1

Thus, the preferred modeling technique for these datasets is K-means clustering followed by the calculation of the correlations with the data from social media and the Reporting Desk. The quality of the datasets with the survey scores is sufficient enough to create the model, contrary to the data from social media and the data from the Reporting Desk, which was lacking. For 1960 organizations, the survey scores are stored in the database of the Supervision Division Information Security. The volume and the quality of this data is appropriate for the creation of the model. The quality of the data from social media is also satisfactory, however, the volume of the data is too low for the creation of the model. For the data from the Reporting Desk, the quality is sufficient enough to calculate correlations, however, the volume is even lower than the data from social media, which means that this dataset is even less suitable for the creation of a data-driven model.

3. *How can the created model be validated and verificated?*

A variety of assumptions are applied to the datasets from the Supervision Division Information Security that consist of the survey scores and the dataset from the Reporting Desk about the reportings of the failures during the EDA and the creation of the exploratory model. These assumptions influence the interpretation of the data. Seven assumptions were made, and three are incorrect. Two of the three assumptions that are wrong are due to an incorrect interpretation of a certain value within the dataset. Therefore, it is important that the meaning of the collected data is clear for anyone who may not be familiar with the dataset. A value in the dataset cannot be interpreted in two different ways. It is important that the user of the dataset can translate the data to valuable information. The other incorrect assumption concerned the discrepancies between the reports from 2015, and the reports from 2016 to 2019. The reports from 2015 consider only full reports, however, a report can consist out of multiple interruptions of services. From 2016 to 2019, all the interruptions within a report were saved separately. This means that the information expressed in the dataset for 2015, and the data for the years following 2015 are different. Therefore, it is hard to draw conclusive information from the dataset for all the years. This is because it is critical that all rows and columns within a dataset have the same meaning, so that the value and use of the dataset is to maximized.

4. *Whom are the stakeholders involved in the regulation on the subject continuity within the Supervisory Field of Information Security and what is their role?*

The stakeholders involved in the supervision on the subject continuity within the Supervisory Field of Information Security are divided into two different kinds of stakeholders; the external and internal stakeholders. The external stakeholders can be grouped as the organizations, and the internal stakeholders are the different stakeholders

within the organizations, more specifically, the teams within AT involved in the process of supervision.

External stakeholders
The external stakeholders can then be divided into a further two divisions, which are the organizations involved in the process of supervision and the supervised organizations. There are two different organizations involved in the supervision on the continuity of the telecommunication services and networks. The first organization is the Ministry of Economic Affairs and Climate Policy. They set the legislation on telecommunication which are primarily based on the guidelines provided by the European Union. The second organization is AT, the Dutch regulator on telecommunication and multiple other industries. They supervise compliance with laws and regulations on the subject continuity of the telecommunication networks and services.

The second division of organizations are the supervised organizations. This division consists of two different kind of organizations, including the telecommunication providers and the network providers. Telecommunication providers provide fixed telephony, mobile wireless communication services, and other similar services. A network provider is a business or organizations that sells bandwidth or network access, often also referred to as internet providers. They provide the direct internet backbone to internet service providers or provide access to internet access points.

Internal stakeholders
There are four different kind of internal stakeholders within AT. The first internal stakeholder is the Supervision Division on Information Security. This team is directly involved in the supervision on the continuity of the networks and services provided by providers of public electronic communications networks and services. The experts within this team send out survey to the supervised organizations to obtain knowledge about how the organizations compliance with the law. Distribution of the resources for the supervision of these organizations is based on the outcomes of a risk analysis conducted by the experts. Scored from the survey are used as input data for the risk analysis. The risk analysis is based on the input data and the gut feeling of the experts and inspectors.

The second internal stakeholder is the Enforcement Policy Division on Information Security. Enforcement policy on supervision sets policy to improve the compliance behavior of the supervised organizations. The kind of supervision advocated by the Enforcement Policy Division on Information Security is ‘quality supervision’, which aims to create a system that tackles socials problem related to Information Security. According to the experts of this Policy Divisions is this kind of supervision highly effective to improve the telecommunication infrastructure in The Netherlands. It creates an reliable and well-organized telecommunication system. Furthermore, the data from the reporting desk is considered and valuable to the process of supervision of the continuity of the telecommunication services by the policy-makers.

The third internal stakeholder is the Spectrum Division Continuity. They are the executing part of the legislation created by the national government and responsible for the creation

of protocols based on the national policy. One of these protocols is the obligation of to report incidents which caused discontinuities of the services and networks provided by the supervised organizations. The law defines that the organizations need to report an incident when the organizations think that the impact of the incident is significant enough to be needed to do so. Data from the reports of the incidents that cause interruption of the continuity is not shared with the Supervision Division o due to an institutional boundary between the Spectrum Division and the Supervision Division, which is called a institutional border. The reason for this institutional border is to increase the willingness to report of the organizations because the law says that the organizations only need to report an incident if the organization thinks that the failure is significant enough.

The final internal stakeholder is the Monitoring and Analysis Center, their role is to monitor and manage all the data within the organization. This team had the most expertise and knowledge about analyzing data and creating data-driven models. They manage and collect the data from social media on incidents that caused failures. A failure is occurs according to them when the number social media posts per day that mentions a failure from a specific organizations exceeds the threshold that is set for that specific organizations. This means that the threshold differs per organizations.

The Exploratory Data Analysis is executed for both datasets with the survey scores. This begins with a minimization of the residual, which are the values that deviate from the expected values. This initial process is often referred to as data cleaning, and consists of an additional two steps; high percentage of questions that were unanswered by the supervised organizations found. During the third part of the data cleaning were all the questions that are unanswered by 50 percent or more of the organizations removed from the dataset. The purpose of the data cleaning is to improve the efficiency of the algorithm that will be executed on the datasets. Six different dataset are created from the 2 initial datasets of the survey scores. All six the datasets can be used for the creation of the exploratory model.

The second data source is data from social media mentions regarding failures. This data is scraped from social media for 38 of the 2003 regulated organizations because there is only data on social media for these organizations. The dataset consists of the number of social media posts per day that a failure is mentioned per organization, between January 2012 and ending in May 2019. Experts from the Supervision Division Information Security defined a threshold for the amount of social media posts an organization can be mentioned in before a failure is considered to have occurred. The thresholds are specifically and independently set for all the organizations, so the value of the threshold varies. An EDA is not executed for this dataset because it only consists of the number of social media posts per day and the thresholds differ for each organization. Even if an EDA is performed with this data, only a small amount of information could be returned, and would not be of significant use for the creation of the Exploratory Model.

The final dataset is the data from the Reporting Desk, which is part of the Spectrum Division Continuity. This dataset consists of data regarding self-reported failures from 16 different organizations, and data of 12 of the 16 organizations are also present in the

other two datasets. The data is derived from information about failures that occurred in the years between 2015 till 2019, and includes information about the duration of the failure, the causes, and other properties of failures that cause interruption of telecommunication networks and services. Due to the low number of organizations represented in this dataset, it is more difficult to use this dataset for the creation of the Exploratory Model.

5. What are the organizational requirements to implement the created model?

There are two principal requirements to implement the model within the organization. The first is the availability of the data for the creation of the model. Data from three different sources is used to create the model. All the data needs to be available for the creation of the model. The model is created to support the supervision on the continuity of the telecommunications network and services, but there is an institutional boundary between the Spectrum Division Continuity and the Supervision Division Information Security, which makes the data from the Reporting Desk inaccessible to the Supervision Division Information Security. Thus, a key component to implementing the model is that all the data is available to the owner of the model, which is the second requirement of an efficient implementation.

There are four different internal stakeholders that could be the possible owner of the model within the organization. The Spectrum Division Continuity and the Enforcement Policy Division on Information Security are the least suitable candidates for the ownership of the model. Both of these internal stakeholders are indirectly involved in the supervision on Information Security. The model is created to support the supervision process and they are too far distanced from this process, making them the least suitable owners of the model. A better candidate would be the Supervision Division Information Security. They are the main actor for the supervision of the model, however two requirements need to meet in order for them to be the owner of the model.

The first requirement is the availability of the data for the Supervision Division. For now, the data from the Reporting Desk is not available to them due to the institutional border within the organization. There are two different options to make the data available to the Supervision Division. The first option is to remove the institutional border and make the data available to the Supervision Division within the organization. The second option is to relocate the Reporting Desk from the Spectrum Division to the Supervision Division Information Security. These changes could only be applied and executed by the Ministry of Economic Affairs and Climate Policy. The other requirement for the Supervision Division is to add an expert on data-driven models to their team. This means somebody within the internal actor would be appointed as the person ultimately responsible for monitoring, maintaining and executing the model. It would be essential for this person to be suitable for this role.

The final internal stakeholder is the M&AC. This stakeholder is the most suitable option to be the owner of the model as they are already equipped with expertise on data analysis and data-driven models, and all the data sources are available to this actor. The only drawback of them being the owner of the model is the indirect connection with the process

of supervision on the continuity of the telecommunication network and services. This gives the experts within the M&AC a lower level of expertise on the subject Information Security. During interviews with experts of all four stakeholders, the overwhelming opinion that the M&AC would be the best option to be the owner of the model emerged.

13. Recommendation

The model is created to form better insights on the effects and possibilities for the transition from risk-based regulation to data-driven risk-based regulation. The goal of the creation of the model is to inform AT on two different levels within the organization; AT as an organization, and in the specific division of supervisory Information Security. For AT, the recommendations are on the implementation of the model, whereas for the supervisory Information Security, the recommendations focus on the subject Information Security. These recommendations suggest improvements for the exploratory model.

- Overcome the issue of the data from the Reporting Desk not being available for the Supervision Division Information Security. This can be done in two ways; removing the institutional boundary between the Spectrum Division and the Supervision Division, or moving the Reporting Desk from the Spectrum Division to the Supervision Division.
- Find a way to aggregate the data from the Reporting Desk in a way that the organizations are unrecognizable for the Supervision Division Information Security, in order for them to use the data for the creation of the model.
- Encouraging better self-reporting of incidents on the discontinuity of the network and services from organizations, as only 16 of the 1000 organizations reported incidents during the last 5 years.
- The frequency of the surveys needs to be higher. Over the last ten years, surveys have been distributed between one to the supervised organizations. This results in an outdated database. A higher frequency will lead to a more up to date database.
- The survey should be sent digitally to the supervised organizations to reduce the administrative burdens and allow for the frequency of surveys to be increased. This will make it easier for the supervised organizations to reply to the survey and for the Supervision Division Information Security to cope with all the responses.
- When an organization reply a specific answer to a question they could be requested to skip certain questions. This makes the skipped questions not applicable for the organization which is one of the possible multiple choice answers for the question. The experts within the of Supervision Division Information Security register this as unanswered. This makes that there is no difference between the not applicable questions and the unanswered questions. Some information for the creation of the clusters got lost because of these answers having the same value.
- The structure of the dataset of the survey should be uniform. This would make it easier to understand the dataset when viewing the dataset for the first time, and easier to use for the creation of a data-driven model.
- Multiple organizations are daughter companies of other organizations or sell services or networks from other organizations. Although a failure could occur for them, a different organization is listed as responsible for this failure. Data should be collected about the source of the failures and all organizations affected by the failure.

- The social media monitoring threshold needs to defined more precisely for all the providers of public electronic communications networks and services, and not only the organizations mentioned on the website of “ alle storingen”. This will increase the volume of the dataset on the failures mentioned on social media.
- Increase the volume of data in the dataset of the Reporting Desk. AT should consider setting stricter standards for the organizations to be obligated to report a failure. This is the only way to obtain more data from failures. This should be done digitally for the same reasons as the surveys

14. Future research

The future research consist of two primary components, starting with the possible future research for AT in section 14.1, followed by the future scientific research in section 14.2. The possible future research for AT highlights the options for a full data-driven model and a semi data-driven model.

14.1 Future research for AT

Full data-driven model
There are two different ways to follow up the research of the exploratory model. The first one is to proceed the creation of a data-driven model, which can be divided into two different options of further research for data-driven models within AT. The first option is to further explore how data-driven risk detection models could support the supervisory field Information Security. This would demand changes in legislation that need to be implemented at the level of the national policy. These changes will take time, as will the collection of data following any changes. Thus, further research on how data-driven models within the Supervisory Field of Information Security will support the process of supervision will be a time intensive process.

This gives room for the second option which is to research how this method is applicable to other supervisory fields with a higher availability of data. This gives AT a better understanding on how the model works, if it is applicable to other supervisory fields, and if the outcomes of the model support the process of regulation. This will further develop a better understanding of what the value is of clustering organizations with similar properties and to investigate if clustering could improve the cost effectiveness of the process of regulation. On the other hand, there is more time to research the value of calculating correlations for the created clusters.

Semi data-driven
The second way for future research is to create a semi data-driven model on the subject of the continuity of the telecommunication networks and services for the supervisory field Information Security. In this case, a semi data-driven model would only create the clusters with the data from the surveys and creating the risk profiles with the knowledge from the experts. Data of the survey on the subject Information Security is available for 1960 organizations and for the survey on the subject continuity for 1250 organizations, which means that the volume is high enough to execute the K-means clustering algorithm. This algorithm will generate clusters based on the survey scores. The clusters can then be utilised as a risk profile for the risk detection model and are based on data from the Reporting Desk. For the semi data-driven model, the risk profiles are based on a combination of qualitative data and quantitative data.

The experts of the Supervision Division Information Security will look at the organizations within the dataset and the average scores of the questions. The average scores are the answers that are provided the most frequently within a cluster for a question. Risk profiles are created for all the clusters based on the outcomes of the risk analysis of the experts.

During the current process of regulation, the risk analysis executed individually per organization. In the semi data-driven model, the risk analysis can be executed for all the organization within a specific cluster.

14.2 Future scientific research

Researchers can further research two dichotomous elements of this study. The first is on the technical creation of a data-driven risk detection model for the supervision of telecommunication services and network providers. The supervised organizations are interrelated with each other due to the mother and daughter companies relationship and the use of each other's services or networks. A risk detection model creates risk profiles for the organizations based on historical data. During this study, it is assumed that all the organizations that reply to at least one of the surveys are independent from each other, which is not the case for many of the supervised organizations. The future research for the creation of model should be focused on the creation of risk profiles that accounts for all the relations between the supervised organizations. This would result in more specific risk profiles.

The other element to be further researched are the organizational challenges. Researchers could further explore the role of an institutional boundary between the Reporting Desk and supervisions teams within agencies. Furthermore, research on the role of reports on interruptions of services where the continuity is important is vital for the creation of data-driven models that support the supervision of the continuity of these services. Research could also be carried out on the importance of data with information on incidents, reports and other information that is reported to Reporting Desk for all agencies. This will better determine the role of reporting desks when it comes to data-driven risk-based regulation, and the importance of the dataset from the reporting desks.

15. Reflection

The goal of this chapter is to reflect upon the research conducted on the subject of how data-driven models could improve or support the process of regulation. The reflection consists of four different parts. Section 15.1 provides a reflection on the exploratory model, followed by the reflection on the implementation of the model within the organization (15.2). Section 15.3 highlights the scientific relevance of the study and section 15.4 discusses the social relevance of the study.

15.1 Model reflection

The reflection on the model consists of two different parts; the reflection on the data and a reflection on the modeling method. The data can be reflected upon by looking at the quality of the data and the quantity of the data. The quality of the dataset with the survey scores of the subject continuity is high. A high percentage of the questions of this were answered. This was found during the EDA in Chapter 6. The quality of the dataset with the survey scores of the subject Information Security is lower than the quality of the datasets of the subject continuity. The lower quality is caused by the large amount of blanks within the dataset, which are considered as unanswered during the EDA and the creation of the exploratory model. During the process of validation, it emerged that a blank could mean that the question is unanswered by the organization or that the questions is not applicable for the organization. Based on the assumption that the blanks meant that a question is unanswered, in an attempt to improve the quality of the data during the EDA, questions with a high percentage of blanks were removed. The quality of the data can improved by recording the difference between unanswered and nonapplicable questions within the dataset. In addition, the datasets with the survey scores can be influenced due self-reporting bias. This can be caused by people filling in different answers as they attend to do. The quality of both datasets is high enough to use for the creation of the clusters.

It is hard to define the quality of the data from social media because it tracks the number of times per day a failure is mentioned in a social media post for a specific organization. This software creates the highest quality of data based on the query that is given by the user of the software. For the calculation is the quality of the data high enough to define the average number of failures per year of the organizations within the cluster. The quality of the data from the Reporting Desk is high enough to define the correlations between the clusters and the data of the characteristics of the failures. Nevertheless, the data from the Reporting Desk could be improved by reducing the number of unknown values within the dataset. Reducing these values will improve the precision of the defined risk profile.

The quantity of the data with the survey scores in the initial datasets is high enough to create clusters with the K-means cluster algorithm. On the other hand, the quantity of the data from social media and the Reporting Desk is too low to define the risk profiles. The percentage of datasets with the survey scores and within either or both datasets with the data from social media and the Reporting Desk is lower than three percent. This makes it impossible to create meaningful risk profiles for all the supervised organizations.

Furthermore, the data from social media only monitors organizations that provide network and services to the public sector. The telecommunication providers that provide services and networks to the private sector are not taken into consideration. The initial approach of the research was to create a model for all providers of telecommunication services and networks. Due to the lack of data of the providers to the private sector is this model only created for public providers of telecommunication services and networks.

The second part of the reflection of the model is on the modeling method. During the EDA, the questions with a majority of blanks were removed from the dataset with the survey scores because the assumption was made that these questions were unanswered. The purpose of removing unanswered questions was to improve the quality of the data, however it may have led to a loss of information because many of the the questions were unanswered by a high percentage of the organizations. It could be the case that the organizations with a high percentage of unanswered questions have worse compliance behavior and that more inspections are needed for these organizations. This outcome will be less likely to occur if questions with a majority of blanks were removed.

15.2 Reflection on the organizational structure

During the study on how to implement the model within the organization, it was found that there is an institutional border between the Spectrum Division Continuity and the Supervision division on Information Security. This border makes it difficult for the model to be implemented into the organization, which leads to the question of whether data-driven risk-based regulation would be the best solution for increasing the cost-effectiveness of the process of regulation on the subject Information Security. AT wanted to explore the possibilities of data-driven risk-based regulation for the supervisory field Information Security, however, due to the institutional border it could be better for AT to explore other kinds of risk-based regulation or an alternate model where the institutional border is not problematic. Another option could be to create a model that supports the process of supervision where the data from the Reporting Desk will not be needed. This could lead to the transition from risk-based regulation to data-driven risk-based regulation.

15.3 Scientific relevance

The scientific literature lacks in research on the creation and implementation of data-driven model for risk-based regulation (Eggimann,2017). The literature predominantly focuses on the promises of risk-based regulation, and issues associated with data-driven risk-based regulation. This study had the research objective to create and implement an exploratory data-driven model for risk-based regulation on the subject of the continuity of the telecommunication network and services within the supervisory field Information Security, to investigate whether a data-driven model can support the process of risk-based regulation. For the creation of a model for data-driven risk-based regulation, it was found that the data from incidents and failures are important for the supervision of the continuity of systems. The information of interruptions of the systems is key to defining the risk or risk profiles of the supervised organizations. Therefore, it is important to understand what kind of data is needed to create a data-driven model.

In addition, the role of the freedom to share the data within the organization is important for the creation and implementation of the model. Institutional borders between bodies or divisions within a regulator could hinder the creation of data-driven models. Furthermore, it is important that people with the adequate expertise are being present in the organization and especially in the division that is ultimately responsible for the model. An understanding of what is needed is important to run, manage and monitor the model. Therefore, the organizational structure is key for the model to be used the most efficient way.

15.4 Social relevance

Improving the cost-effectiveness of the used resources during the process of regulation for the Supervisory Field of Information Security leads to improving the compliance behaviour of the supervised organizations. The organizations on the supervisory of the continuity of the telecommunications network and services are telecommunications providers. The role of telecommunications is increasing in the private and public sector. More systems are relying on telecommunication so the importance of the continuity of the telecommunication is increasing. The created method is a tool that can be used by the supervisors on the continuity of the telecommunication services and networks provided by the supervised organizations. The risk of failures that cause discontinuities will be reduced when the compliance behaviour of the organization increases, which means the continuity of the telecommunication services and networks will increase for both the public and the private sector.

References:

Agentschap Telecom (2017,December). Biedt u openbare telefonie, internettoegang en/of een netwerk aan?

Agentschap Telecom (2013). Toezicht zorg- en meldplicht continuïteit, de 0-meting.

Agentschap Telecom (2018, October). Vragenlijst nieuwe aanbieders van openbare telecommunicatie netwerken en services.

Agentschap Telecom (2019). Wijzigingen van de Telecommunicatiewet ter implementatie van de herziene telecommunicatierichtlijnen.

Ale, B., & Ale, B. J. M. (2009). *Risk: an introduction: the concepts of risk, danger and chance*. London: Routledge.

Aoki, M. (1986). Horizontal vs. Vertical Information Structure of the Firm. *The American Economic Review, 76*(5), 971-983. Retrieved from http://www.jstor.org/stable/1816463

Black, J., & Baldwin, R. (2010). Really Responsive Risk-Based Regulation. *Law & Policy,32*(2), 181-213. doi:10.1111/j.1467-9930.2010.00318.x

D.D.C. Brewer and M.J. Nash (IEEE Symposium on Security and Privacy, pp. 206-14, 1988)

Dietz, S. (2010). High impact, low probability? An empirical analysis of risk in the economics of climate change. *High Impact, Low Probability? An Empirical Analysis of Risk in the Economics of Climate Change, 108(3), 519–541. doi: 10.1007/s10584-010-9993-4*

Eggimann, S., Mutzner, L., Wani, O., Schneider, M., Spuhler, D., Moy de Vitry, M., Beutler, P.,Maurer, M. (2017). The potential of knowing more: A review of data- driven urban water management. *Environmental Science and Technology*. 51(5), 2538-2553

Haines, F. (2012). *The paradox of regulation: What regulation can achieve and what it cannot*. Cheltenham: Edward Elgar.

Hand, D. J., Mannila, H., & Smyth, P. (2012). *Principles of data mining*. New Delhi: PHI Learning Private Limited.

Hartigan, J., & Wong, M. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100-108. doi:10.2307/2346830

Hood, C., Rothstein, H., & Baldwin, R. (2001-08-23). The Government of Risk: Understanding Risk Regulation Regimes. : Oxford University Press. Retrieved 4 Aug. 2019, from https://www.oxfordscholarship.com/view/10.1093/0199243638.001.0001/acprof-9780199243631.

Hilbert, Martin. (2013). Big Data for Development: From Information- to Knowledge Societies. SSRN Electronic Journal. 10.2139/ssrn.2205145.

Holak, S. L., & Lehmann, D. R. (1990). Purchase Intentions and the Dimensions of Innovation: An Exploratory Model. *Journal of Product Innovation Management, 7(1), 59–73. doi: 10.1111/1540-5885.710059*

Irion, K., & Radu, R. (2013). Delegation to independent regulatory authorities in the media sector: A paradigm shift through the lens of regulatory theory. Retrieved from https://www.ivir.nl/publicaties/download/Radu_2013.pdf

King, R., & Brennan, J. (2018, February). Data-driven risk-based quality regulation. Retrieved from https://www.qaa.ac.uk/docs/qaa/about-us/data-driven-quality-assessment-final.pdf?sfvrsn=916ff681_8

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society. https://doi.org/10.1177/2053951714528481

Koyuncugil, A. S., & Ozgulbas, N. (2012). Financial early warning system model and data mining application for risk detection. *Expert Systems with Applications,39*(6), 6238-6253. doi:10.1016/j.eswa.2011.12.021

McCarty, J., & Hastak, M. (2007, January 16). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. Retrieved from https://www.sciencedirect.com/science/article/pii/S0148296306002323

Mittelstadt, Brent & Allo, Patrick & Taddeo, Mariarosaria & Wachter, Sandra & Floridi, Luciano. (2016). The Ethics of Algorithms: Mapping the Debate. Big Data & Society. In press. 10.1177/2053951716679679.

Nunan, F. , Campbell, A. and Foster, E. (2012), ENVIRONMENTAL MAINSTREAMING: THE ORGANISATIONAL CHALLENGES OF POLICY INTEGRATION. Public Admin. Dev., 32: 262-277. doi:10.1002/pad.1624

Pham KTM, Zhou Q, Kurasawa Y, Li Z. (2011, November 15) A coiled-coil- and C2-domain-containing protein is required for FAZ assembly and cell morphology in Trypanosoma brucei J Cell Sci. 2019 Jul 15; 132(14). Epub 2019 Jul 15.

Purnima, B., & Kumar, A. (2014, November 9). BK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.735.7337&rep=rep1&type=pdf

Quain, Sampson. (2019, February 01). The Definitions of Horizontal and Vertical Organizations. Small Business - Chron.com. Retrieved from http://smallbusiness.chron.com/definitions-horizontal-vertical-organizations-23483.html

Rowe, W. (1977, March). Governmental Regulation of Societal Risks. Retrieved from https://heinonline.org/HOL/LandingPage?handle=hein.journals/gwlr45÷=40&id=&page=

Rothstein, Henry & Irving, Phil & Walden, Terry & Yearsley, Roger. (2007). The risks of risk-based regulation: Insights from the environmental policy domain. Environment international. 32. 1056-65. 10.1016/j.envint.2006.06.008.

Smith, P. & Wellenius, B. (1999, March 31). Strategies for Successful Telecommunications Regulation in Weak Governance Environments. Retrieved from http://siteresources.worldbank.org/INTRANETTRADE/Resources/BP2telcm.pdf
Vose, D. (2008). *Risk analysis: A quantitative guide*. Chichester: Wiley.

Tempelman, J. A. (2012, June 5). Meldplicht continuïteitsonderbrekingen. Retrieved from https://www.navigator.nl/document/idb738da8fdc58431f805b55ed6595e059?cpid=WKNL-LTR-Nav2&cip=hybrid

Velleman, P. F., and D. C. Hoaglin. 1981. Applications, basics, and computing of exploratory data analysis. Boston: Duxbury.

Yu, C. H. (2017). Exploratory Data Analysis. *Oxford Bibliographies Online Datasets. doi: 10.1093/obo/9780199828340-0200*

Tempelman, J. A. (2012, June 5). Meldplicht continuïteitsonderbrekingen. Retrieved from https://www.navigator.nl/document/idb738da8fdc58431f805b55ed6595e059?cpid=WKNL-LTR-Nav2&cip=hybrid

Appendix A1 - column names of the datasets about the surveys

Chapter one of the appendix consists out of the tables with the column names of all the used datasets with data of the survey scores. These datasets are the information from the initial dataset and the returned data during the EDA. Starting with section A1.1 that provides the initialing column names of dataset with the survey scores. Following, the column names are shown from the first version of the datasets. Finally, the column names of the second version from the datasets are shown. The column names for both the datasets with the survey scores on both subject are shown separately, because the dataset with the merged data has the column names from both datasets.

A1.1 The initial column names from the datasets provided by AT

Relatienummer AT
Naam AT
Dossier Nummer
Lijst ID
Lijst omschrijving
Datum inspectie
Aantal vragen
Aantal antwoorden
01) Naam bedrijf ?
02) Vestigingsadres: straat; huisnummer; postcode; plaats (alleen bij nieuw invoeren ;) ?
03) Relatienummer ?
04) Dossiernummer Kamer van Koophandel ?

05) ID nummer inspectie ?
06) Datum inspectie ?
07 a) Uitgevoerd door: C.J. Roomer?
07 b) Uitgevoerd door: W. Herder?
07 c) Uitgevoerd door: E. de Geus?
07 d) Uitgevoerd door: D.M.S. Jansen?
07 e) Uitgevoerd door: J.K.M. Beulen?
07 f) Uitgevoerd door: J.P.A. Hoeven?
07 g) Uitgevoerd door: M. Beemer?
07 h) Uitgevoerd door: Oud medewerker?
08) Soort inspectie?
09) ACM geregistreerd?
10) Wat is de jaaromzet telecom gerelateerd?
11 a) Wie zijn de klanten? Zakelijke eindgebruikers?
11 b) Wie zijn de klanten? Particuliere eindgebruikers?
11 c) Wie zijn de klanten? Andere netwerk en/of dienst aanbieder?
11 d) Wie zijn de klanten? Niet beantwoord?
11 e) Wie zijn de klanten? Onbekend?
12 a) Wat wordt aangeboden? Vaste internettoegang?

12 b) Wat wordt aangeboden? Overige (beschrijven bij bijzonderheden)?
12 c) Wat wordt aangeboden? Dienst/netwerk voor verspreiden van programma's?
12 d) Wat wordt aangeboden? Mobiele telefonie?
12 e) Wat wordt aangeboden? Mobiel netwerk?
12 f) Wat wordt aangeboden? Mobiele internettoegang?
12 g) Wat wordt aangeboden? Vast netwerk?
12 h) Wat wordt aangeboden? Passieve (glasvezel) infrastructuur?
12 i) Wat wordt aangeboden? Vaste telefonie (waaronder CS/CPS/VOIP)?
12 j) Wat wordt aangeboden? E-mail?
12 k) Wat wordt aangeboden? Niet beantwoord?
13) Valt de geinspecteerd binnen het toezichtsdomein?
14) Worden alle diensten/netwerken aangeboden middels eigen apparatuur of netwerk ?
15) Zijn er (wholesale) ketenpartners waaraan een dienst en/of netwerk worden aangeboden ?
16) Biedt uw organisatie enkel passieve (glasvezel) infrastructuur aan? ?
17) 5. Worden de gegevens aangeleverd op het CIOT informatiesysteem ?
18) 6. Is de uitbesteding vastgelegd in een schriftelijke overeenkomst ?
19) 7. Is (zijn alle) dienst(en) en/of netwerk(en) aftapbaar ?

20) 8. Wie voert het proces over aftappen uit ?
21) 9. Is de uitbesteding vastgelegd in een schriftelijke overeenkomst ?
22) 10. Heeft de organisatie een onderhoudscontract afgesloten bij eigen apparatuur of wordt een contract bij uitbesteding periodiek geevalueerd ?
23) 11. Verwerkt de geinspecteerde verkeers en/of locatiegegevens langer dan noodzakelijk is voor het overbrengen van de communicatie ?
24) 12. Indien verkeersgegevens worden verwerkt voor facturatie Is de abonnee/gebruiker in kennis gesteld welke verkeersgegevens worden verwerkt en met welke duur ten behoeven van de facturatie ?
25) 13. Indien verkeers- en/of locatiegegevens worden verwerkt ten behoeve van marktonderzoek en/of verkoopactiviteiten en/of toegevoegde waarde diensten wordt vooraf aan de klant toestemming gevraagd waarbij de klant wordt geïnformeerd over welke verkeers- en/of locatiegegevens worden verwerkt en de duur hiervan? ?
26) 14. Kan de abonnee of gebruiker te allen tijde de toestemming intrekken en is hij hierover geinformeerd ?
27) 15. Is er een procesbeschrijving van het anonimiseren dan wel het verwijderen van verkeers en/of locatiegegevens binnen de gestelde termijn ?
28) 16. Is uw organisatie ISO/NEN 27001 gecertificeerd ?
29) 17. Is er een beveiligingsplan ?
30) 18. Wordt het beveiligingsplan periodiek geëvalueerd? ?
31 a) 19. Wordt er periodiek ge-audit op beveiliging Ja, extern?
31 b) 19. Wordt er periodiek ge-audit op beveiliging Nee?
31 c) 19. Wordt er periodiek ge-audit op beveiliging Ja, intern?
31 d) 19. Wordt er periodiek ge-audit op beveiliging Niet beantwoord?

32) 20. Is een verklaring omtrent gedrag verkregen voor de betreffende medewerkers? ?
33) I. Er is een fucntionaris, belast met het toezciht op de uitvoering en naleving van de beveiligingsmaatregelen? ?
34) II Er wordt voldaan aan beveiliginseisen ten aanzien van personeel? ?
35) III Er wordt voldaan aan fysieke beveiliging en beveiliging van de omgeving ?
36) IV er wordt voldaan aan de eisen met betrekking tot beheer en bedieningsprocessen. ?
37) V Er wordt voldaan aan de toegangsbeveiliging geautomatiseerde systemen. ?
38) VI Er wordt voldaan aan ontwikkeling, onderhoud en reparatie van geautomatiseerde informatiesystemen. ?
39) 21. Is er een procesbeschrijving van de invulling van een ononderbroken toegang to het alarmnummer, waarin in ieder geval aandacht wordt besteed aan congestie, electriciteitsuitval of technische storing? ?
40 a) 22. Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Ja, passende maatregelen?
40 b) 22. Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Ja, noodzakelijke maatregelen?
40 c) 22. Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Nee, geen noodzakelijke maatregelen genomen?
40 d) 22. Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Nee, geen passende maatregelen genomen?
40 e) 22. Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Niet beantwoord?

41) 23. Zijn continuïteitsafspraken gemaakt tussen de geïnspecteerde en zijn ketenpartners (bijvoorbeeld over up- en down-time) in een schriftelijke overeenkomst vastgelegd? ?
42) 24. Levert u telecomdiensten en/of netwerken aan bedrijven/instellingen die binnen de vitale infrastructuur vallen? ?
43) 25. Is er een procesbeschrijving van de invulling van de zorgplicht continuïteit, waarin in ieder geval aandacht wordt besteed aan de adressering van gesignaleerde risico's door middel van passende technische en organisatorische maatregelen? ?
44) 26. Worden de risico's op de veiligheid en integratie van het netwerk en/of de dienst periodiek geïnventariseerd, beoordeeld en geëvalueerd? ?
45) 27. Beschikt uw organisatie over een vastgesteld continuiteitsplan? ?
46) 28. Wordt het continuiteitsplan periodiek herzien? ?
47 a) Welke processen zijn aanwezig t.a.v. management van de dienstverlening? Availability management?
47 b) Welke processen zijn aanwezig t.a.v. management van de dienstverlening? Change management?
47 c) Welke processen zijn aanwezig t.a.v. management van de dienstverlening? Incident management?
47 d) Welke processen zijn aanwezig t.a.v. management van de dienstverlening? Problem management?
47 e) Welke processen zijn aanwezig t.a.v. management van de dienstverlening? Niet beantwoord?
47 f) Welke processen zijn aanwezig t.a.v. management van de dienstverlening? Onbekend?
48) Is het proces BC en ICT weerbaarheid beschreven? ?
49) Is governance ingericht? ?
50) Einde inspectie formulier ?

Table A.1: The column names of the initial dataset on the survey about Information Security provided by AT

0010 Naam geinspecteerde
0020 Vestigingsadres; straat, huisnummer, postcode, plaatsnaam
0030 Postadres; straat/postbus, nummer, postcode, woonplaats
0040 Relatienummer
0050 Contactperso(o)n(en); naam; functie; tel nr; email adres
0060 Dossiernummer Kamer van Koophandel
0070 ID nummer inspectie
0100 Datum inspectie
0110 Uitgevoerd door
0120 Soort inspectie
0130 OPTA-registratie nummer
0140 o.e.c.d.
0150 o.e.c.n.
0159 Wat is de jaaromzet
0160 Valt de geinspecteerde binnen het toezichts domein
1000 Beschikt uw organisatie over een continuiteitsplan
1100 Wanneer is de verwachting dat er wel beschikt wordt over een continuiteitsplan
2000 Zijn er richtlijnen en doelstellingen afgegeven mbt continuiteit en beschikbaardheid
2100 Wordt er op directieniveau gerapporteerd waar het de geleverde prestatie van de organisatie betreft mbt de continuiteit en de beschikbaarheid van uw telecom diensten en/of netwerken

2200 Wordt er regelmatig op directieniveau gerapporteerd waar het de geleverde prestatie van de leveranciers van ICT als TI betreft mbt de continuiteit en de beschikbaarheid van diensten
2300 Zijn majeure incidenten welke zicht vanaf 6-2012 voorgedaan hebben opgenomen in de bijlage van het continuiteitsplan inclusief oorzaak en verbetermaatregelen ter voorkoming van herhaling en is uw directie op de hoogte gebracht
2400 Op welke niveau wordt/is het continuiteitsplan vastgesteld
3000 Is in het continuiteitsplan beschreven hoe het proces incident management is ingericht
3100 Is bij een majeure calamiteit de maximum toelaatbare hersteltijd na een onderbreking van ICT diensten (incl. telefonie) vastgelegd en beschreven in het continuiteitsplan
3200 Worden incidenten geregistreerd en worden de impact, urgentie en hersteltijd aangegeven
3300 Is er prioritering bij het oplossen van incidenten op basis van maatschappelijke en/of economische impact
4000 Is in het continuiteitsplan beschreven hoe het proces problem management is ingericht
4100 Is in het continuiteitsplan beschreven hoe het proces change management is ingericht
4200 Is in het continuiteitsplan omschreven welk deel van de infrastructuur op korte termijn aan vervanging toe is
4300 Zijn er additionele passende technische en organisatorische maatregelen genomen om de continuiteit over het te vervangen stuk infrastructuur te garanderen
4400 Is er een verhoogde paraatheid/alertheid ingevoerd voor het te vervangen stuk infrastructuur
4500 Bij uitbesteding bent u verantwoordelijk dat de derden de zorgplicht contiuiteit naleven. Zijn deze verplichtingen vastgelegd en verwijst u daarnaar in het continuiteitsplan
4600 Is er een actueel overzicht van huidige leveranciers en/of service partners op het gebied van ICT en TI, welke een relatie hebben met de continuiteit van uw diensten en/of netwerken, opgenomen in het continuiteitsplan

5000 Zijn in het continuiteitsplan de risico's beschreven die de continuiteit van de diensten of netwerken bedreigen
5100 Staan in het continuiteitsplan de risico's beschreven, de kans van het optreden van het risico en de impact bij het optreden van het risico omschreven
5200 Staan in het continuiteitsplan de maatregelen beschreven welke u neemt om deze risico's te adresseren
5300 Is er in het continuiteitsplan, bij het beschrijven van de risico's, rekening gehouden met grootschalige uitval van elektriciteit en/of ICT
5400 Is er bij een grootschalige uitval van elektriciteit en/of ICT de dan in werking treden kritieke processen en activiteiten in het continuiteitsplan beschreven m.b.t. het waarborgen van de continuiteit van de netwerken en diensten
6000 Worden er telecom diensten en/of netwerken geleverd aan bedrijven/instellingen welke binnen de definitie vitale infrastructuur vallen en is uw directie hiervan op de hoogte
6300 Heeft u inloggegevens aangevraagd en ontvangen? Dit zijn de inloggegevens die nodig zijn om incidenten te melden bij het loket Meldplicht
9999 Einde formulier

Table A.2: The column names of the initial dataset on the survey about continuity provided by AT

A1.2 The column names of the first versions of the datasets with the survey scores

1	Naam AT
2	ACM geregistreerd?
3	Wat is de jaaromzet telecom gerelateerd?
4	Zakelijke eindgebruikers?
5	Particuliere eindgebruikers?
6	Andere netwerk en/of dienst aanbieder?
7	Klanten zijn onbekend
8	Vaste internettoegang?
9	Overige diensten?
10	Dienst/netwerk voor verspreiden van programma's?
11	Mobiele telefonie?
12	Mobiel netwerk?
13	Mobiele internettoegang?
14	Vast netwerk?
15	Passieve (glasvezel) infrastructuur?

16	Vaste telefonie (waaronder CS/CPS/VOIP)?
17	Geen service?
18	Valt de geinspecteerd binnen het toezichtsdomein?
19	Worden alle diensten/netwerken aangeboden middels eigen apparatuur of netwerk ?
20	Zijn er (wholesale) ketenpartners waaraan een dienst en/of netwerk worden aangeboden ?
21	Worden de gegevens aangeleverd op het CIOT informatiesysteem ?
22	Is de uitbesteding vastgelegd in een schriftelijke overeenkomst ?
23	Is (zijn alle) dienst(en) en/of netwerk(en) aftapbaar ?
24	Wie voert het proces over aftappen uit ?
25	Is de uitbesteding vastgelegd in een schriftelijke overeenkomst ?
26	Heeft de organisatie een onderhoudscontract afgesloten bij eigen apparatuur of wordt een contract bij uitbesteding periodiek geevalueerd ?
27	Indien verkeersgegevens worden verwerkt voor facturatie Is de abonnee/gebruiker in kennis gesteld welke verkeersgegevens worden verwerkt en met welke duur ten behoeven van de facturatie ?
28	Kan de abonnee of gebruiker te allen tijde de toestemming intrekken en is hij hierover geinformeerd ?

29	Is er een beveiligingsplan ?
30	Is een verklaring omtrent gedrag verkregen voor de betreffende medewerkers? ?
31	Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Ja, passende maatregelen?
32	Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Ja, noodzakelijke maatregelen?
33	Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Nee, geen noodzakelijke maatregelen genomen?
34	Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Nee, geen passende maatregelen genomen?
35	Zijn continuïteitsafspraken gemaakt tussen de geïnspecteerde en zijn ketenpartners (bijvoorbeeld over up- en down-time) in een schriftelijke overeenkomst vastgelegd?
36	Levert u telecomdiensten en/of netwerken aan bedrijven/instellingen die binnen de vitale infrastructuur vallen?
37	Is er een procesbeschrijving van de invulling van de zorgplicht continuïteit, waarin in ieder geval aandacht wordt besteed aan de adressering van gesignaleerde risico's door middel van passende technische en organisatorische maatregelen?

38	Worden de risico's op de veiligheid en integratie van het netwerk en/of de dienst periodiek geïnventariseerd, beoordeeld en geëvalueerd?
39	Beschikt uw organisatie over een vastgesteld continuiteitsplan?
40	Wordt het continuiteitsplan periodiek herzien?

Table A.3: The column names of the first version of the dataset on the subject Information Security

1	Naam
2	OPTA geregistreed
3	o.e.c.d.
4	o.e.c.n.
5	Wat is de jaaromzet
6	Valt de geinspecteerde binnen het toezichts domein
7	Beschikt uw organisatie over een continuiteitsplan
8	Wanneer is de verwachting dat er wel beschikt wordt over een continuiteitsplan
9	Zijn er richtlijnen en doelstellingen afgegeven mbt continuiteit en beschikbaardheid
10	Wordt er op directieniveau gerapporteerd waar het de geleverde prestatie van de organisatie betreft mbt de continuiteit en de beschikbaarheid van uw telecom diensten en/of netwerken

11	Wordt er regelmatig op directieniveau gerapporteerd waar het de geleverde prestatie van de leveranciers van ICT als TI betreft mbt de continuiteit en de beschikbaarheid van diensten
12	Zijn majeure incidenten welke zicht vanaf 6-2012 voorgedaan hebben opgenomen in de bijlage van het continuiteitsplan inclusief oorzaak en verbetermaatregelen ter voorkoming van herhaling en is uw directie op de hoogte gebracht
13	Op welke niveau wordt/is het continuiteitsplan vastgesteld
14	Is in het continuiteitsplan beschreven hoe het proces incident management is ingericht
15	Is bij een majeure calamiteit de maximum toelaatbare hersteltijd na een onderbreking van ICT diensten (incl. telefonie) vastgelegd en beschreven in het continuiteitsplan
16	Worden incidenten geregistreerd en worden de impact, urgentie en hersteltijd aangegeven
17	Is er prioritering bij het oplossen van incidenten op basis van maatschappelijke en/of economische impact
18	Is in het continuiteitsplan beschreven hoe het proces problem management is ingericht
19	Is in het continuiteitsplan beschreven hoe het proces change management is ingericht
20	Is in het continuiteitsplan omschreven welk deel van de infrastructuur op korte termijn aan vervanging toe is
21	Zijn er additionele passende technische en organisatorische maatregelen genomen om de continuiteit over het te vervangen stuk infrastructuur te garanderen
22	Is er een verhoogde paraatheid/alertheid ingevoerd voor het te vervangen stuk infrastructuur
23	Bij uitbesteding bent u verantwoordelijk dat de derden de zorgplicht contiuiteit naleven. Zijn deze verplichtingen vastgelegd en verwijst u daarnaar in het continuiteitsplan

24	Is er een actueel overzicht van huidige leveranciers en/of service partners op het gebied van ICT en TI, welke een relatie hebben met de continuiteit van uw diensten en/of netwerken, opgenomen in het continuiteitsplan
25	Zijn in het continuiteitsplan de risico's beschreven die de continuiteit van de diensten of netwerken bedreigen
26	Staan in het continuiteitsplan de risico's beschreven, de kans van het optreden van het risico en de impact bij het optreden van het risico omschreven
27	Staan in het continuiteitsplan de maatregelen beschreven welke u neemt om deze risico's te adresseren
28	Is er in het continuiteitsplan, bij het beschrijven van de risico's, rekening gehouden met grootschalige uitval van elektriciteit en/of ICT
29	Is er bij een grootschalige uitval van elektriciteit en/of ICT de dan in werking treden kritieke processen en activiteiten in het continuiteitsplan beschreven m.b.t. het waarborgen van de continuiteit van de netwerken en diensten
30	Worden er telecom diensten en/of netwerken geleverd aan bedrijven/instellingen welke binnen de definitie vitale infrastructuur vallen en is uw directie hiervan op de hoogte
31	Is er beschreven welke additionele maatregelen zijn genomen, om de continuiteit te kunnen garanderen, voor wat betreft de levering van telecom diensten en/of netwerken aan bedrijven en/of instellingen die binnen de definitie vitale infrastructuur vallen
32	Is er een actueel overzicht van de klanten (bedrijven en/of instellingen) die binnen de defenitie vitale infrastructuur vallen en die u telecommunicatie diensten en/of netwerken levert, opgenomen in het continuiteitsplan

Table A.4: The column names of the first version of dataset on the subject Continuity

A1.3 The column names of the second versions of the datasets with the survey scores.

1	Naam AT
3	Wat is de jaaromzet telecom gerelateerd?
4	Zakelijke eindgebruikers?
5	Particuliere eindgebruikers?
6	Andere netwerk en/of dienst aanbieder?
8	Vaste internettoegang?
9	Overige diensten?
11	Mobiele telefonie?
12	Mobiel netwerk?
13	Mobiele internettoegang?
14	Vast netwerk?
16	Vaste telefonie (waaronder CS/CPS/VOIP)?
17	Geen service?

19	Worden alle diensten/netwerken aangeboden middels eigen apparatuur of netwerk ?
20	Zijn er (wholesale) ketenpartners waaraan een dienst en/of netwerk worden aangeboden ?
21	Worden de gegevens aangeleverd op het CIOT informatiesysteem ?
24	Wie voert het proces over aftappen uit ?
25	Is de uitbesteding vastgelegd in een schriftelijke overeenkomst ?
32	Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Ja, noodzakelijke maatregelen?

Table A.5: The column names of the second version of the dataset on the subject Information Security

2	ACM geregistreerd?
7	Klanten zijn onbekend
10	Dienst/netwerk voor verspreiden van programma's?
15	Passieve (glasvezel) infrastructuur?
18	Valt de geinspecteerd binnen het toezichtsdomein?
22	Is de uitbesteding vastgelegd in een schriftelijke overeenkomst ?
23	Is (zijn alle) dienst(en) en/of netwerk(en) aftapbaar ?

26	Heeft de organisatie een onderhoudscontract afgesloten bij eigen apparatuur of wordt een contract bij uitbesteding periodiek geevalueerd ?
27	Indien verkeersgegevens worden verwerkt voor facturatie Is de abonnee/gebruiker in kennis gesteld welke verkeersgegevens worden verwerkt en met welke duur ten behoeven van de facturatie ?
28	Kan de abonnee of gebruiker te allen tijde de toestemming intrekken en is hij hierover geinformeerd ?
29	Is er een beveiligingsplan ?
30	Is een verklaring omtrent gedrag verkregen voor de betreffende medewerkers? ?
31	Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Ja, passende maatregelen?
33	Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Nee, geen noodzakelijke maatregelen genomen?
34	Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Nee, geen passende maatregelen genomen?
35	Zijn continuïteitsafspraken gemaakt tussen de geïnspecteerde en zijn ketenpartners (bijvoorbeeld over up- en down-time) in een schriftelijke overeenkomst vastgelegd?
36	Levert u telecomdiensten en/of netwerken aan bedrijven/instellingen die binnen de vitale infrastructuur vallen?

37	Is er een procesbeschrijving van de invulling van de zorgplicht continuïteit, waarin in ieder geval aandacht wordt besteed aan de adressering van gesignaleerde risico's door middel van passende technische en organisatorische maatregelen?
38	Worden de risico's op de veiligheid en integratie van het netwerk en/of de dienst periodiek geïnventariseerd, beoordeeld en geëvalueerd?
39	Beschikt uw organisatie over een vastgesteld continuiteitsplan?
40	Wordt het continuiteitsplan periodiek herzien?

Table A.6: The column names of the removed columns of the dataset of the survey on Information Security between the first and the second version.

1	Naam
3	o.e.c.d.
5	Wat is de jaaromzet
7	Beschikt uw organisatie over een continuiteitsplan
8	Wanneer is de verwachting dat er wel beschikt wordt over een continuiteitsplan
9	Zijn er richtlijnen en doelstellingen afgegeven mbt continuiteit en beschikbaardheid
10	Wordt er op directieniveau gerapporteerd waar het de geleverde prestatie van de organisatie betreft mbt de continuiteit en de beschikbaarheid van uw telecom diensten en/of netwerken

11	Wordt er regelmatig op directieniveau gerapporteerd waar het de geleverde prestatie van de leveranciers van ICT als TI betreft mbt de continuiteit en de beschikbaarheid van diensten
12	Zijn majeure incidenten welke zicht vanaf 6-2012 voorgedaan hebben opgenomen in de bijlage van het continuiteitsplan inclusief oorzaak en verbetermaatregelen ter voorkoming van herhaling en is uw directie op de hoogte gebracht
13	Op welke niveau wordt/is het continuiteitsplan vastgesteld
14	Is in het continuiteitsplan beschreven hoe het proces incident management is ingericht
15	Is bij een majeure calamiteit de maximum toelaatbare hersteltijd na een onderbreking van ICT diensten (incl. telefonie) vastgelegd en beschreven in het continuiteitsplan
16	Worden incidenten geregistreerd en worden de impact, urgentie en hersteltijd aangegeven
17	Is er prioritering bij het oplossen van incidenten op basis van maatschappelijke en/of economische impact
18	Is in het continuiteitsplan beschreven hoe het proces problem management is ingericht
19	Is in het continuiteitsplan beschreven hoe het proces change management is ingericht
20	Is in het continuiteitsplan omschreven welk deel van de infrastructuur op korte termijn aan vervanging toe is
21	Zijn er additionele passende technische en organisatorische maatregelen genomen om de continuiteit over het te vervangen stuk infrastructuur te garanderen

22	Is er een verhoogde paraatheid/alertheid ingevoerd voor het te vervangen stuk infrastructuur
23	Bij uitbesteding bent u verantwoordelijk dat de derden de zorgplicht contiuiteit naleven. Zijn deze verplichtingen vastgelegd en verwijst u daarnaar in het continuiteitsplan
24	Is er een actueel overzicht van huidige leveranciers en/of service partners op het gebied van ICT en TI, welke een relatie hebben met de continuiteit van uw diensten en/of netwerken, opgenomen in het continuiteitsplan
25	Zijn in het continuiteitsplan de risico's beschreven die de continuiteit van de diensten of netwerken bedreigen
26	Staan in het continuiteitsplan de risico's beschreven, de kans van het optreden van het risico en de impact bij het optreden van het risico omschreven
27	Staan in het continuiteitsplan de maatregelen beschreven welke u neemt om deze risico's te adresseren
28	Is er in het continuiteitsplan, bij het beschrijven van de risico's, rekening gehouden met grootschalige uitval van elektriciteit en/of ICT
29	Is er bij een grootschalige uitval van elektriciteit en/of ICT de dan in werking treden kritieke processen en activiteiten in het continuiteitsplan beschreven m.b.t. het waarborgen van de continuiteit van de netwerken en diensten
30	Worden er telecom diensten en/of netwerken geleverd aan bedrijven/instellingen welke binnen de definitie vitale infrastructuur vallen en is uw directie hiervan op de hoogte
33	Heeft u inloggegevens aangevraagd en ontvangen? Dit zijn de inloggegevens die nodig zijn om incidenten te melden bij het loket Meldplicht

Table A.7: The column names of the second version of the dataset on the subject Continuity

Appendix A2. Bar graphs during the data visualization

Bar graph are created for all the values in the datasets with the survey scores. Six different datasets are creates during the EDA and all the answers in these datasets are visualized in six different bar graphs.

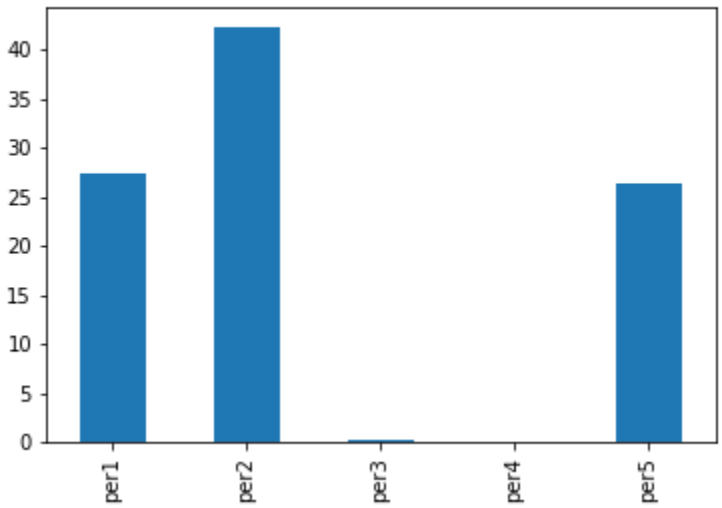


Figure A2.1: Bar graph of the percentage of every multiple choice answers in the first version of the dataset on the subject Information Security

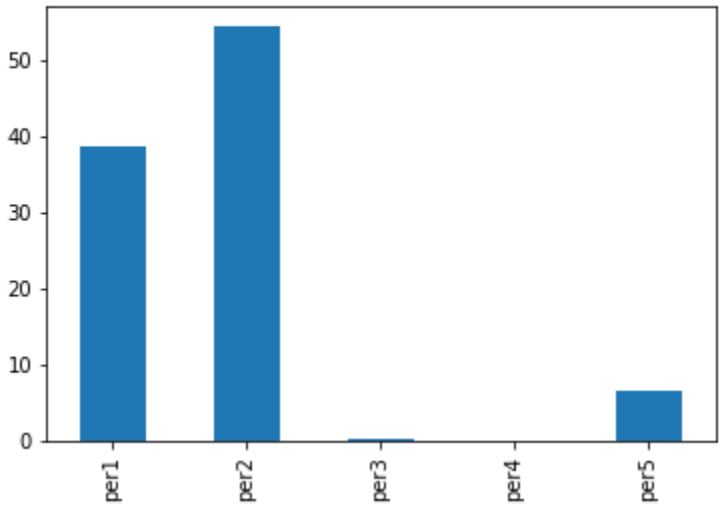


Figure A2.2: Bar graph of the percentage of every multiple choice answers in the second version of the dataset on the subject Information Security

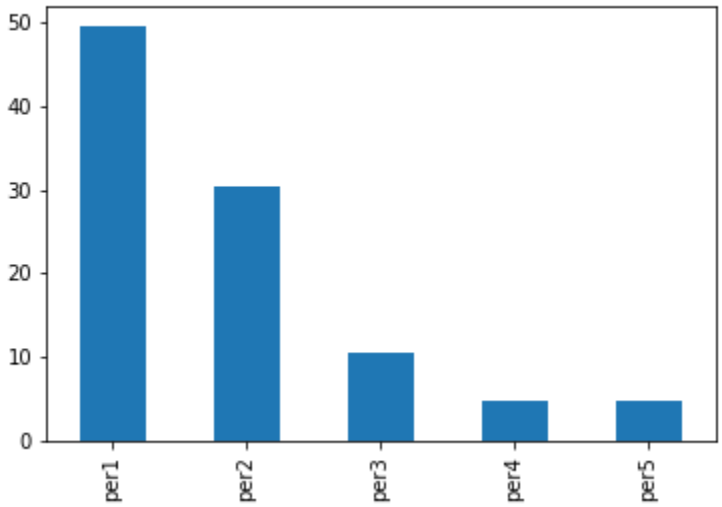


Figure A2.3: Bar graph of the percentage of every multiple choice answers in the first version of the dataset on the subject Continuity

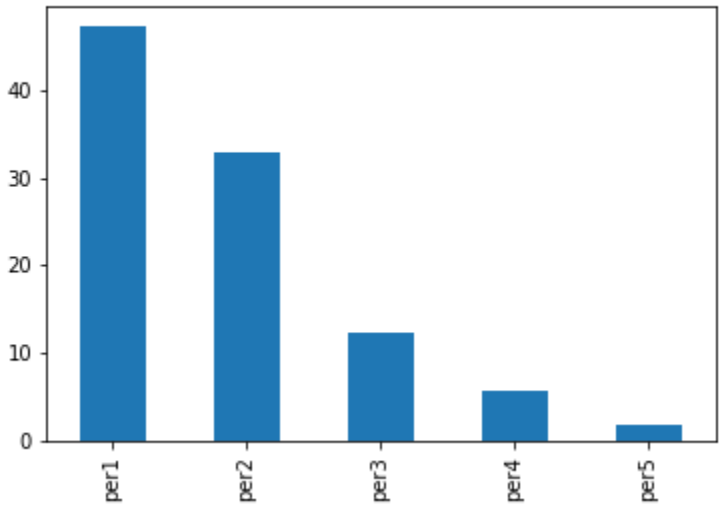


Figure A2.4: Bar graph of the percentage of every multiple choice answers in the second version of the dataset on the subject Continuity

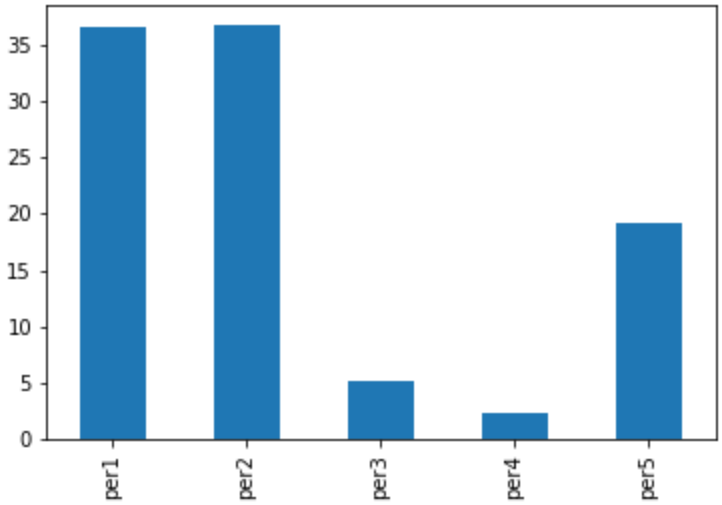


Figure A2.5: Bar graph of the percentage of every multiple choice answers in the first version of the dataset on both subjects

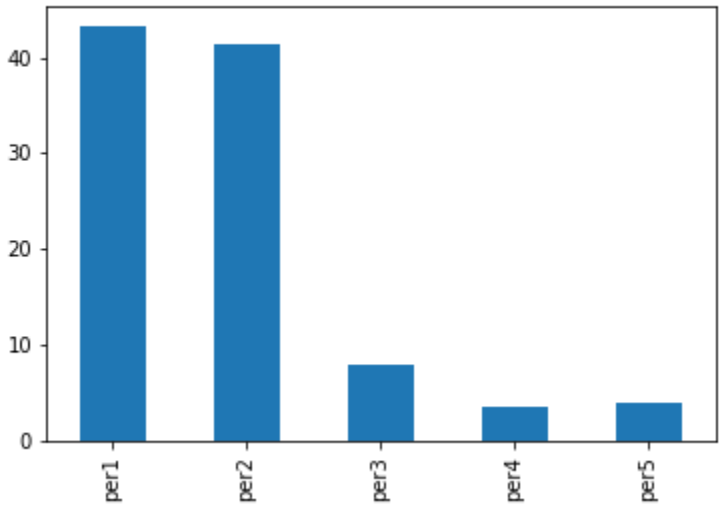


Figure A2.6: Bar graph of the percentage of every multiple choice answers in the second version of the dataset on both subjects

Appendix A3. The scatter plots and graphs for the elbow method

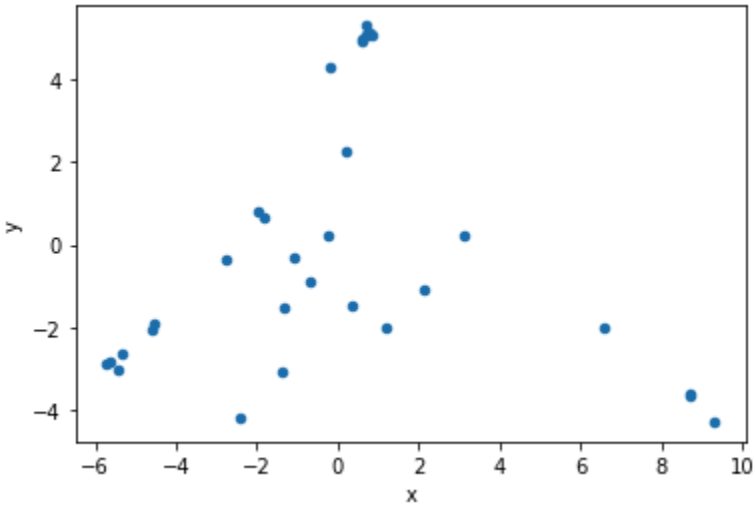


Figure A3.1. Scatterplot data for the first version of the dataset on the subject Information Security

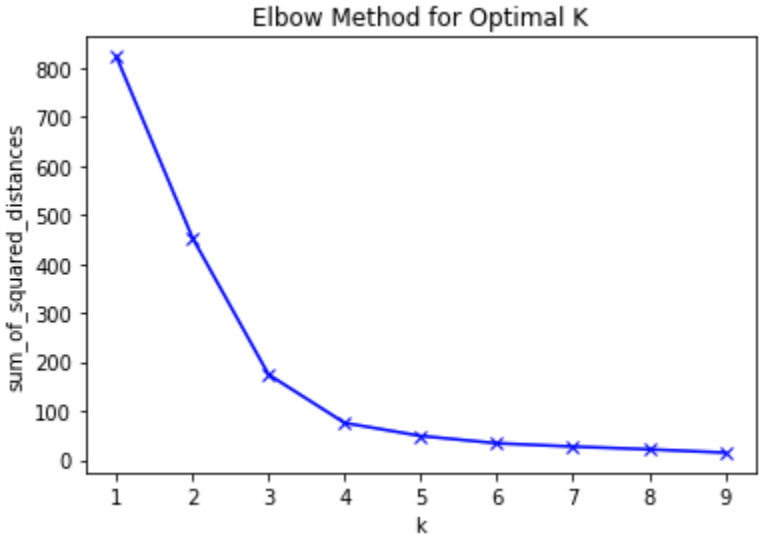


Figure A3.2. Result of the elbow method of the first version of the dataset on the subject Information Security

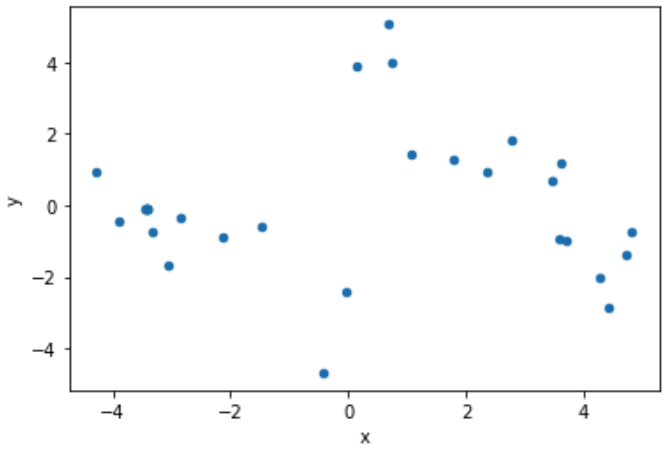
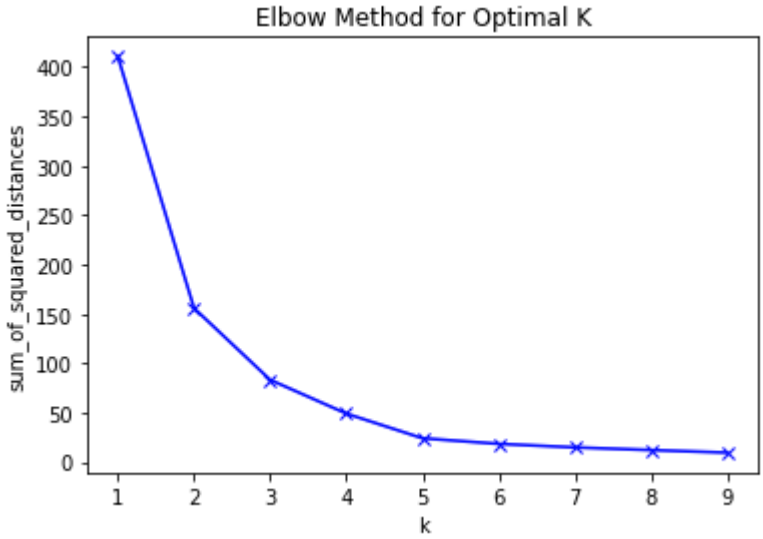


Figure A3.3. Scatterplot data for the first version of the dataset on the subject Continuity



.

Figure A3.4. Result of the elbow method of the first version of the dataset on the subject Continuity

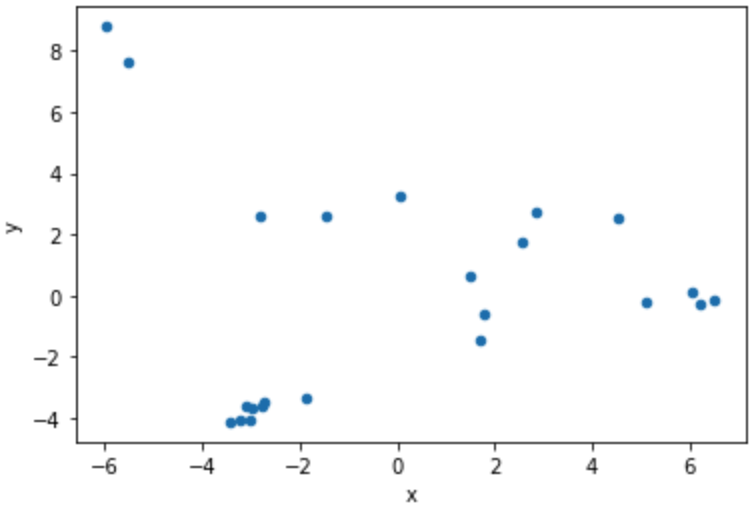


Figure A3.5: Scatterplot data for the first version of the dataset on both subjects

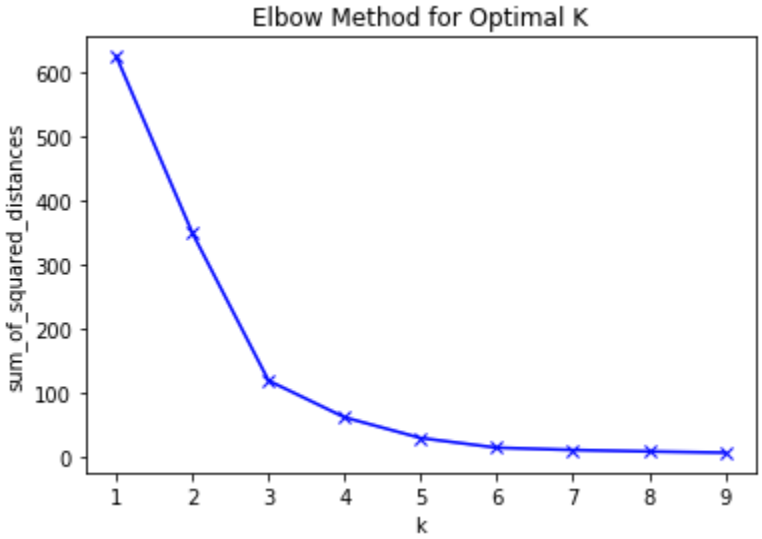


Figure A3.6. Result of the elbow method of the second version of the dataset on both subject

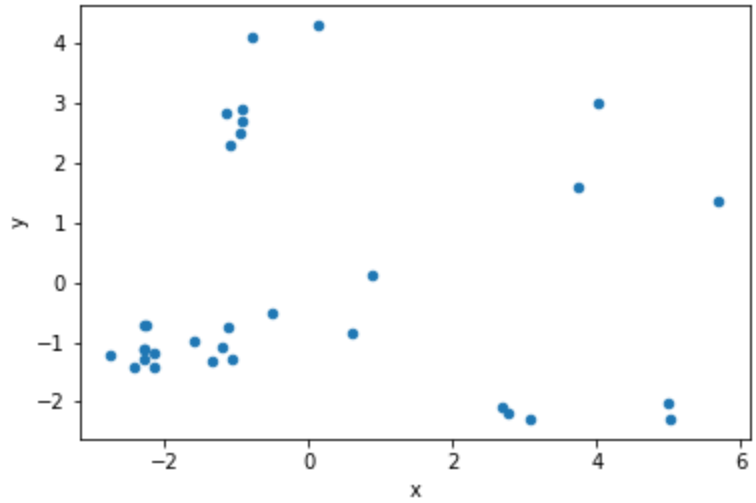


Figure A3.7. Scatterplot data for the second version of the dataset on the subject Information Security

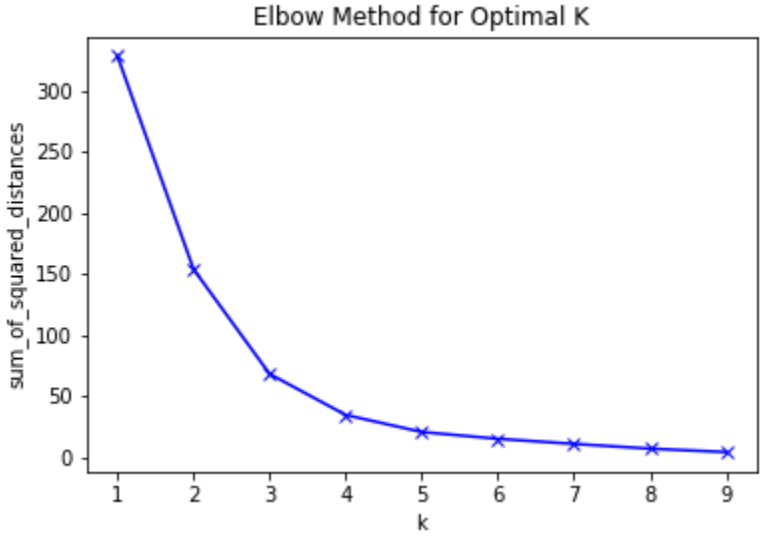


Figure A3.8. Result of the elbow method of the second version of the dataset on the subject Information Security

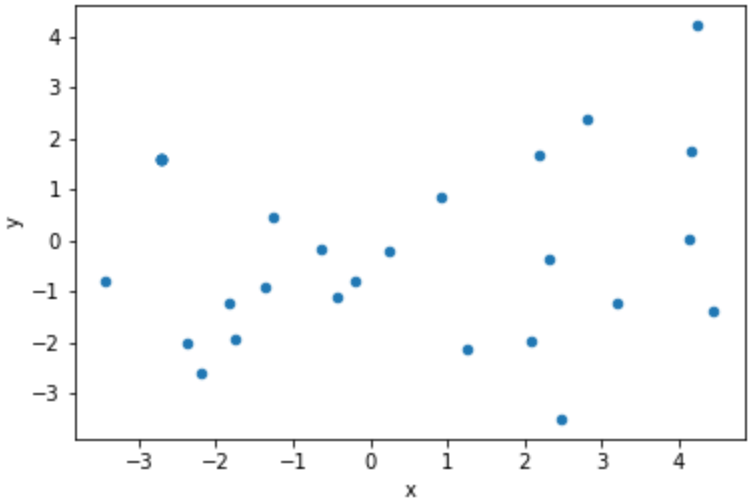


Figure A3.9. Scatterplot data for the second version of the dataset on the subject Continuity

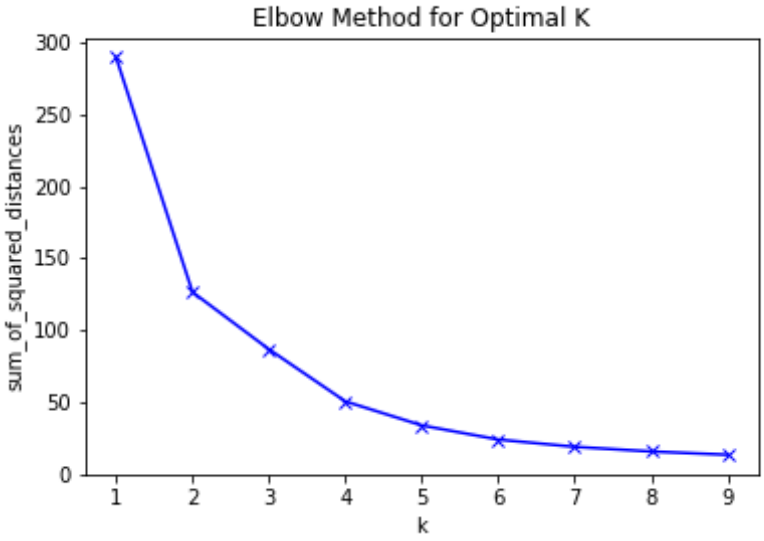


Figure A3.10. Result of the elbow method of the second version of the dataset on the subject Continuity

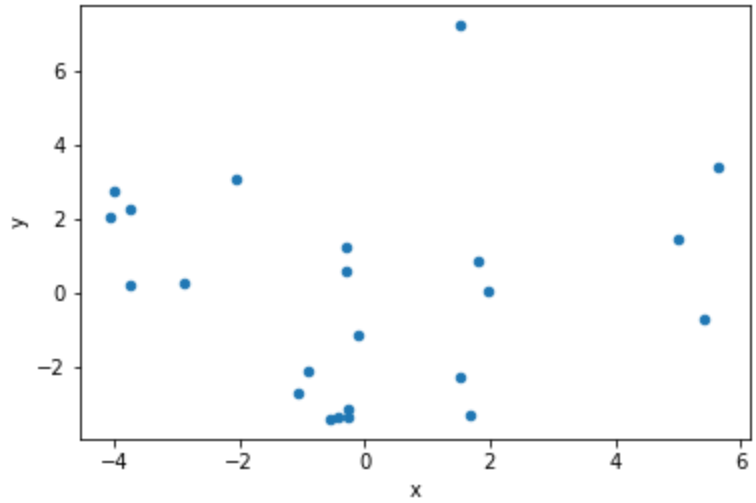


Figure A3.11: Scatterplot data for the second version of the dataset on both subjects

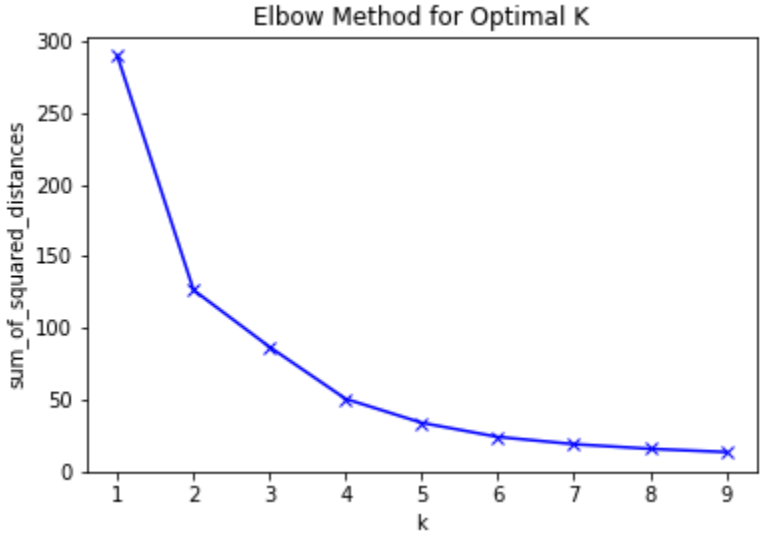


Figure A3.12. Result of the elbow method of the second version of the dataset on both subject

Appendix A4. The organizations presented in the clusters of the datasets and the most important questions for the formation of the clusters

Information Security	**Version 1**			**Information Security**	**Version 2**	
Cluster 1	Cluster 2	Cluster 3		Cluster 1	Cluster 2	Cluster 3
A	FF	C		B	A	FF
B	E	P		G	C	E
D	F	L		I	D	F
S	G	R		K	P	H
U	H			M	L	J
V	I			O	R	N
X	J			Q	CC	T
Y	K			S	DD	
Z	M			U		
AA	N			V		
BB	O			W		
EE	Q			X		
	T			Y		

	W			Z		
	CC			AA		
	DD			BB		
				EE		

Table A4.1: The organizations within the clusters for the datasets on Information Security.

Nummer	Vraag
3	Wat is de jaaromzet telecom gerelateerd?
5	Particuliere eindgebruikers?
16	Vaste telefonie (waaronder CS/CPS/VOIP)?
19	Valt de geïnspecteerde binnen het toezichtsdomein?
21	Zijn er (wholesale) ketenpartners waaraan een dienst en/of netwerk worden aangeboden ?
22	Worden de gegevens aangeleverd op het CIOT informatiesysteem ?
24	Is (zijn alle) dienst(en) en/of netwerk(en) aftapbaar ?
25	Wie voert het proces over aftappen uit ?
26	Is de uitbesteding vastgelegd in een schriftelijke overeenkomst ?
27	Heeft de organisatie een onderhoudscontract afgesloten bij eigen apparatuur of wordt een contract bij uitbesteding periodiek geëvalueerd ?

28	Indien verkeersgegevens worden verwerkt voor facturatie Is de abonnee/gebruiker in kennis gesteld welke verkeersgegevens worden verwerkt en met welke duur ten behoeve van de facturatie ?
29	Kan de abonnee of gebruiker te allen tijde de toestemming intrekken en is hij hierover geïnformeerd ?
30	Is er een beveiligingsplan ?
31	Is een verklaring omtrent gedrag verkregen voor de betreffende medewerkers? ?
32	Heeft de geïnspecteerde passende en, indien van toepassing, noodzakelijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Ja, passende maatregelen?
36	Zijn continuïteit afspraken gemaakt tussen de geïnspecteerde en zijn ketenpartners (bijvoorbeeld over up- en down-time) in een schriftelijke overeenkomst vastgelegd?
37	Levert u telecomdiensten en/of netwerken aan bedrijven/instellingen die binnen de vitale infrastructuur vallen?
38	Is er een procesbeschrijving van de invulling van de zorgplicht continuïteit, waarin in ieder geval aandacht wordt besteed aan de adressering van gesignaleerde risico's door middel van passende technische en organisatorische maatregelen?
39	Worden de risico's op de veiligheid en integratie van het netwerk en/of de dienst periodiek geïnventariseerd, beoordeeld en geëvalueerd?
40	Beschikt uw organisatie over een vastgesteld continuïteitsplan?

41	Wordt het continuïteitsplan periodiek herzien?

Table A4.2: The most significant questions for the formation of the clusters for the dataset on Information Security after the first part of data cleaning.

3	Wat is de jaaromzet telecom gerelateerd?
8	Vaste internettoegang?
16	Vaste telefonie (waaronder CS/CPS/VOIP)?
19	Valt de geïnspecteerde binnen het toezichtsdomein?
20	Worden alle diensten/netwerken aangeboden middels eigen apparatuur of netwerk ?
21	Zijn er (wholesale) ketenpartners waaraan een dienst en/of netwerk worden aangeboden ?
24	Is (zijn alle) dienst(en) en/of netwerk(en) aftapbaar ?
25	Wie voert het proces over aftappen uit ?

Table A4.3: The most significant questions for the formation of the clusters for the dataset on Information Security after the third part of data cleaning.

Continuity	**Version 1**		**Continuity**	**Version 2**
Cluster 1	Cluster 2		Cluster 1	Cluster 2
D	A		A	C
E	B		B	FF
L	C		E	D
N	FF		G	J

P	G		GG	L
Q	GG		K	N
T	J		O	HH
V	K		II	P
Y	HH		S	Q
JJ	O		W	T
LL	II		Y	V
AA	S		KK	JJ
BB	W			Z
CC	Z			LL
EE	KK			AA
				BB
				CC
				EE

Table A4.4: The organizations within the clusters for both versions of the dataset on continuity.

5	Wat is de jaaromzet
7	Beschikt uw organisatie over een continuïteitsplan

8	Wanneer is de verwachting dat er wel beschikt wordt over een continuïteitsplan
9	Zijn er richtlijnen en doelstellingen afgegeven mbt continuïteit en beschikbaarheid
10	Wordt er op directieniveau gerapporteerd waar het de geleverde prestatie van de organisatie betreft mbt de continuïteit en de beschikbaarheid van uw telecom diensten en/of netwerken
15	Is bij een majeure calamiteit de maximum toelaatbare hersteltijd na een onderbreking van ICT diensten (incl. telefonie) vastgelegd en beschreven in het continuïteitsplan
16	Worden incidenten geregistreerd en worden de impact, urgentie en hersteltijd aangegeven
30	Worden er telecom diensten en/of netwerken geleverd aan bedrijven/instellingen welke binnen de definitie vitale infrastructuur vallen en is uw directie hiervan op de hoogte
31	Is er beschreven welke additionele maatregelen zijn genomen, om de continuiteit te kunnen garanderen, voor wat betreft de levering van telecom diensten en/of netwerken aan bedrijven en/of instellingen die binnen de definitie vitale infrastructuur vallen
32	Is er een actueel overzicht van de klanten (bedrijven en/of instellingen) die binnen de definitie vitale infrastructuur vallen en die u telecommunicatie diensten en/of netwerken levert, opgenomen in het continuïteitsplan

Table A4.5: The most important questions for the formation of the clusters for the first version of the dataset on Continuity.

5	Wat is de jaaromzet
8	Wanneer is de verwachting dat er wel beschikt wordt over een continuiteitsplan

9	Zijn er richtlijnen en doelstellingen afgegeven mbt continuiteit en beschikbaardheid
10	Wordt er op directieniveau gerapporteerd waar het de geleverde prestatie van de organisatie betreft mbt de continuiteit en de beschikbaarheid van uw telecom diensten en/of netwerken
12	Zijn majeure incidenten welke zicht vanaf 6-2012 voorgedaan hebben opgenomen in de bijlage van het continuiteitsplan inclusief oorzaak en verbetermaatregelen ter voorkoming van herhaling en is uw directie op de hoogte gebracht
15	Is bij een majeure calamiteit de maximum toelaatbare hersteltijd na een onderbreking van ICT diensten (incl. telefonie) vastgelegd en beschreven in het continuiteitsplan
16	Worden incidenten geregistreerd en worden de impact, urgentie en hersteltijd aangegeven
17	Is er prioritering bij het oplossen van incidenten op basis van maatschappelijke en/of economische impact
23	Bij uitbesteding bent u verantwoordelijk dat de derden de zorgplicht contiuiteit naleven. Zijn deze verplichtingen vastgelegd en verwijst u daarnaar in het continuiteitsplan
24	Is er een actueel overzicht van huidige leveranciers en/of service partners op het gebied van ICT en TI, welke een relatie hebben met de continuiteit van uw diensten en/of netwerken, opgenomen in het continuiteitsplan
25	Zijn in het continuiteitsplan de risico's beschreven die de continuiteit van de diensten of netwerken bedreigen
26	Staan in het continuiteitsplan de risico's beschreven, de kans van het optreden van het risico en de impact bij het optreden van het risico omschreven

27	Staan in het continuiteitsplan de maatregelen beschreven welke u neemt om deze risico's te adresseren
29	Is er bij een grootschalige uitval van elektriciteit en/of ICT de dan in werking treden kritieke processen en activiteiten in het continuiteitsplan beschreven m.b.t. het waarborgen van de continuiteit van de netwerken en diensten

Table A4.6: The most important questions for the formation of the clusters for the second version of the dataset on Continuity.

Both surveys	**Part 1**			**Both surveys**	**Part 3**	
Cluster 1	Cluster 2	Cluster 3		Cluster 1	Cluster 2	Cluster 3
A	FF	B		B	A	C
C	E	S		E	D	FF
D	G	V		G	L	J
L	J	Y		K	CC	N
	K	Z		O		Q
	N	AA		S		T
	O	BB		W		V
	Q	EE		Y		Z
	T					AA
	W					BB
	CC					EE

Table A4.7: The organizations within the clusters for the datasets with the organizations that responded to both surveys.

5	Valt de geinspecteerde binnen het toezichts domein
10	Wordt er regelmatig op directieniveau gerapporteerd waar het de geleverde prestatie van de leveranciers van ICT als TI betreft mbt de continuiteit en de beschikbaarheid van diensten
35	Zakelijke eindgebruikers?
46	Passieve (glasvezel) infrastructuur?
51	Worden alle diensten/netwerken aangeboden middels eigen apparatuur of netwerk ?
52	Zijn er (wholesale) ketenpartners waaraan een dienst en/of netwerk worden aangeboden ?
53	Worden de gegevens aangeleverd op het CIOT informatiesysteem ?
54	Is de uitbesteding vastgelegd in een schriftelijke overeenkomst ?
56	Wie voert het proces over aftappen uit ?
57	Is de uitbesteding vastgelegd in een schriftelijke overeenkomst ?
58	Heeft de organisatie een onderhoudscontract afgesloten bij eigen apparatuur of wordt een contract bij uitbesteding periodiek geevalueerd ?
59	Indien verkeersgegevens worden verwerkt voor facturatie Is de abonnee/gebruiker in kennis gesteld welke verkeersgegevens worden verwerkt en met welke duur ten behoeven van de facturatie ?

60	Kan de abonnee of gebruiker te allen tijde de toestemming intrekken en is hij hierover geinformeerd ?
61	Is er een beveiligingsplan ?
63	Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Ja, passende maatregelen?
64	Heeft de geinspecteerde passende en, indien van toepassing, noodzakeijke maatregelen getroffen om de risico's voor de veiligheid en de integriteit van het netwerk en/of dienst te beheersen? Ja, noodzakelijke maatregelen?
68	Levert u telecomdiensten en/of netwerken aan bedrijven/instellingen die binnen de vitale infrastructuur vallen?
69	Is er een procesbeschrijving van de invulling van de zorgplicht continuïteit, waarin in ieder geval aandacht wordt besteed aan de adressering van gesignaleerde risico's door middel van passende technische en organisatorische maatregelen?
70	Worden de risico's op de veiligheid en integratie van het netwerk en/of de dienst periodiek geïnventariseerd, beoordeeld en geëvalueerd?
71	Beschikt uw organisatie over een vastgesteld continuiteitsplan?
72	Wordt het continuiteitsplan periodiek herzien?

Table A4.8: The most important questions for the formation of the clusters for the first version of the dataset on both subjects.

8	Vaste internettoegang?

19	Valt de geinspecteerd binnen het toezichtsdomein?
20	Worden alle diensten/netwerken aangeboden middels eigen apparatuur of netwerk ?
21	Zijn er (wholesale) ketenpartners waaraan een dienst en/of netwerk worden aangeboden ?
24	Is (zijn alle) dienst(en) en/of netwerk(en) aftapbaar ?
25	Wie voert het proces over aftappen uit ?
39	Zijn er richtlijnen en doelstellingen afgegeven mbt continuiteit en beschikbaardheid
40	Wordt er op directieniveau gerapporteerd waar het de geleverde prestatie van de organisatie betreft mbt de continuiteit en de beschikbaarheid van uw telecom diensten en/of netwerken
42	Zijn majeure incidenten welke zicht vanaf 6-2012 voorgedaan hebben opgenomen in de bijlage van het continuiteitsplan inclusief oorzaak en verbetermaatregelen ter voorkoming van herhaling en is uw directie op de hoogte gebracht
45	Is bij een majeure calamiteit de maximum toelaatbare hersteltijd na een onderbreking van ICT diensten (incl. telefonie) vastgelegd en beschreven in het continuiteitsplan
46	Worden incidenten geregistreerd en worden de impact, urgentie en hersteltijd aangegeven
47	Is er prioritering bij het oplossen van incidenten op basis van maatschappelijke en/of economische impact
58	Is er in het continuiteitsplan, bij het beschrijven van de risico's, rekening gehouden met grootschalige uitval van elektriciteit en/of ICT

59	Is er bij een grootschalige uitval van elektriciteit en/of ICT de dan in werking treden kritieke processen en activiteiten in het continuiteitsplan beschreven m.b.t. het waarborgen van de continuiteit van de netwerken en diensten

Table A4.9: The most important questions for the formation of the clusters for the second version of the dataset on both subjects.

Appendix A5: the average number of mentions and thresholds of the social media data

Organization	Threshold
AA	35
B	50
BB	40
D	5
DD	12
EE	100
G	7
JJ	25
L	80
P	2
V	5
Y	20
Z	15

A5.1: The defined thresholds by AT

Organization	**Threshold**
A	2
C	2
CC	5
E	5
F	2
FF	5
GG	2
H	2
HH	5
II	2
J	2
JJ	25
K	2

KK	2
L	2
LL	2
M	2
N	5
O	5
Q	7
R	2
S	2
T	2
U	5
W	2
X	2

A5.2: The thresholds defined by comparing the average number of mentions per day

Organization	Average #mentions
A	0,039597
B	7,52783
C	0,026896
D	0,320882
E	0,104968
F	0,014195
G	0,524468
H	0,035487
I	0,012327
J	0,056406
K	0,019425
L	42,834143
M	0,054539
N	0,366829
HH	0,20508
O	0,220396
P	0,025775
Q	0,519238

R	0,011207
S	0,069107
T	0,029511
U	0,11991
V	0,624953
W	0,035487
X	0,06313
Y	5,328726
JJ	10,519612
Z	10,291744
KK	0,048188
LL	0,022787
AA	28,96638
BB	12,339933
CC	0,131864
DD	1,655958
EE	36,449384
FF	0,330968

A5.3: The average number of mentions per day per organization

Appendix A6: Reporting Desk

This chapter of the appendix consists of information about the cluster picked for the calculation of the correlations between the organizations in the cluster and the data from the Reporting Desk. In addition are the correlations shown between the data about the failures and the cluster.

Organization
B
D
L
P
Q
V
Y
Z
AA
EE

Table A6.1: All the organizations that are represented in the dataset of the Reporting Desk.

Survey	**Version dataset**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Information Security	1	7	1	2

Information Security	2	7	3	0
Continuity	1	8	2	-
Continuity	2	2	8	-
Both	1	2	1	6
Both	2	2	2	5

Table A6.2: The number of organizations within the cluster that is represented within the dataset from the Reporting Desk.

Survey	**Version dataset**	**Cluster 1** (%)	**Cluster 2** (%)	**Cluster 3** (%)
Information Security	1	58	6	50
Information Security	2	41	38	0
Continuity	1	53	13	-
Continuity	2	17	44	-
Both	1	50	9	75
Both	2	25	50	45

Table A6.3: The percentage of organizations within the cluster that is represented within the dataset from the Reporting Desk.

Organization	Percentage (%)

AA	10
B	20
EE	10
V	28
Y	7
Z	26

Table A6.4: The organizations within the third cluster of the first data of the data about the organizations that responded to both surveys and the corresponding percentages of responsibility for reported incidents.

Year	Incidents	Percentage (%)
2015	22	7
2016	83	26
2017	92	29
2018	96	30
2019	22	7

Table A6.5: the number of incidents and the percentage per year from 2015 to 2019

Service	Incidents	Percentage (%)
Emergency number: 1-1-2	2	1
Email	33	10

Internet	2	1
Internet access	21	7
Internet Access, Fixed telephony, E-mail	1	0
Mobile	2	1
Mobile internet	41	13
Mobile, short message	1	0
Mobile telecom services, short message services	1	0
Mobile telecommunication services from other providers	1	0
Mobile Telephony	54	17
Broadcasting distribution	34	11
SMS	18	6
Fixed internet	23	7
Fixed Telephony	69	22
Fixed telephony, internet access , e-mail	6	2
Unknown	6	2

Table A.6.6: correlations between the cluster and the type of service that got affected by the failure.

Number of customers	Incidents	Percentage (%)
Minus 1	50	16
Unknown	21	7
Higher than 0	244	77

Table A.6.7: correlations between the cluster and the number of affected costumers

Range	Incidents	Percentage (%)
N	84	27
R	198	63
S	4	1
X	7	2
Unknown	22	7

Table A.6.8: correlations between the cluster and the range of the incident

1-1-2 affected?	Incidents	Percentage (%)
Yes	91	29
No	43	14
Does not apply	118	37
Unknown	63	20

Table A.6.9: correlations between the cluster and if the emergency number was affected due to the failure

Cause	Incidents	Percentage (%)
Attack	2	1
Third party	104	33
Fire at third party	2	1
Power failure at third party	17	5
Fiberglass breakage	2	1
Hardware failure	94	30
Hardware or software failure	29	9
Local power supply failure	1	0
Human	23	7
Nature	6	2
Power failure	2	1
Unknown	29	9
Others	4	1

Table A.6.10: correlations between the cluster and the cause of the incident

Hours Immediately	Incidents	Percentage (%)
Lower than 0	1	0
Equal to 0	1	0
Greater than 0	309	98
unknown	4	1

Table A.6.11: correlations between the cluster and the hours immediately

Duration of incident	Incidents	Percentage (%)
Equal to 0	1	0
Greater than 0	309	98
Unknown	5	2

Table A.6.12: correlations between the cluster and the duration of the incident

Appendix A7: Conversations with experts from AT

Multiple conversations and interviews were conducted with expert from different divisions from AT. The most important information is used in the report. This table provides an overview of the experts of the divisions within AT provided information that supported the research. The chapters for which the information is used are shown in the table together with the label of the expert and the division the expert is working for within the organization.

Expert	**Division**	**Chapters**
A	Enforcement Policy	1 to 3
B	Enforcement Policy	1 to 3, 9
C	Enforcement Policy	1 to 3, 5, 10 to 15
D	Enforcement Policy	1 to 3, 5, 10 to 15
E	M&AC	3, 5
F	M&AC	3, 5
G	M&AC	3, 5, 9 to 13
H	Supervisory Information Security	1 to 3, 5, 10 to 15
I	Spectrum Continuity	3, 5, 7 to 13
J	Spectrum Continuity	3, 5, 7 to 13
K	Supervisory WIBON	11

Table A7.1: An overview of the consulted experts within the organization