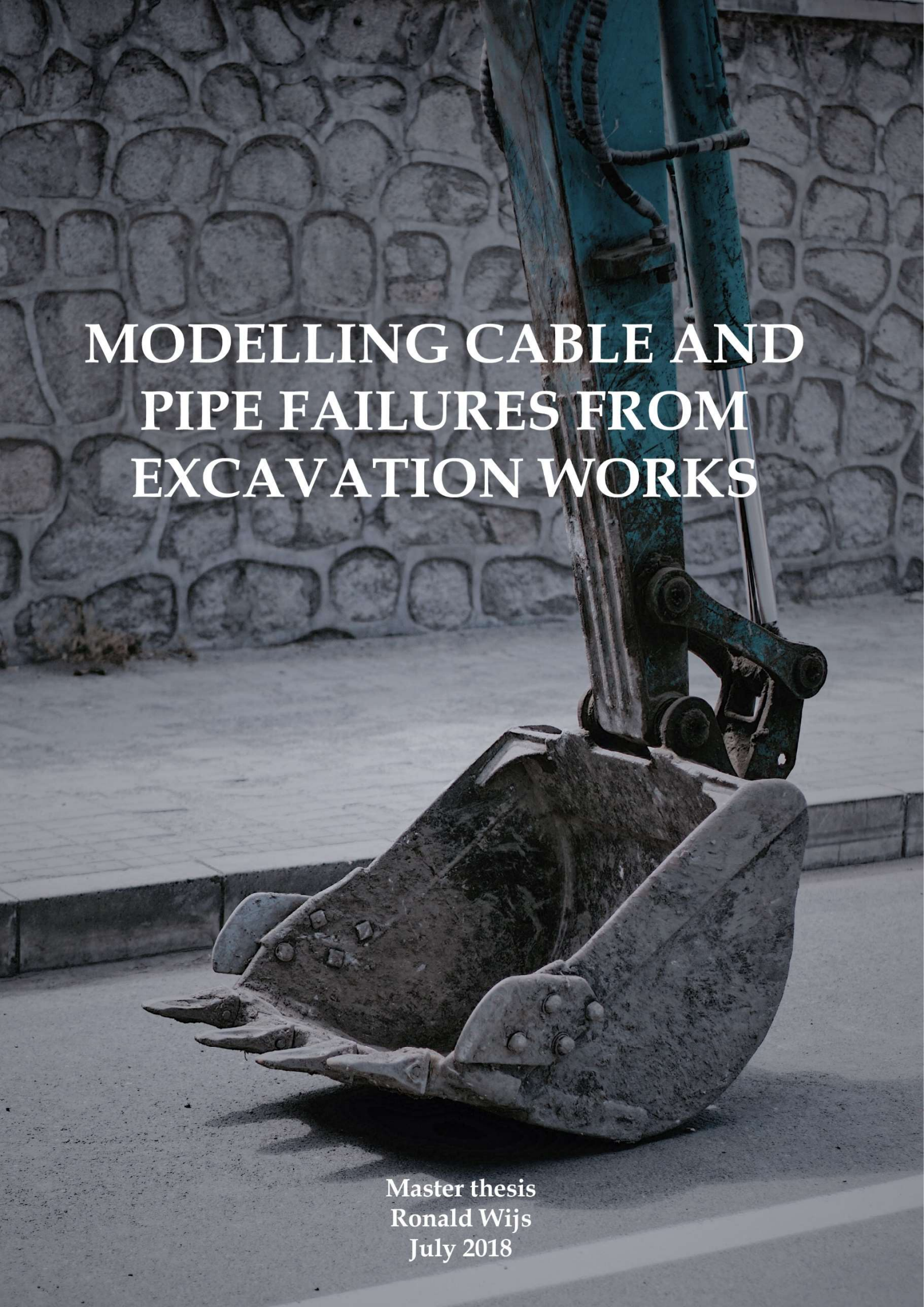


MODELLING CABLE AND PIPE FAILURES FROM EXCAVATION WORKS

Master thesis
Ronald Wijs
July 2018

COLOPHON

AUTHOR

Name	Ronald Wijs
Student number	4152964
E-mail	rjawijs@gmail.com
Telephone number	+31 6 41 87 21 30

GRADUATION COMMITTEE

Chairman	Prof.dr.ir. R. (Rogier) Wolfert Faculty Civil Engineering and Geoscience
Supervisor	Dr. M.L.C. (Mark) de Bruijne Faculty Technology, Policy and Management
Supervisor	ir. G. (George) Leontaris Faculty Civil Engineering and Geoscience
Supervisor	dr. ir. G.F. (Tina) Nane Faculty Applied Mathematics
External supervisor	ir. T. (Thomas) van Manen Evides Waterbedrijf

GRADUATION THESIS

University	Delft University of Technology
Faculty	Civil Engineering and Geosciences (CiTG) Stevinweg 1, 2628 CN Delft
Master	Construction Management and Engineering (CME)
Course	CME 2000 Graduation Thesis
Front page table	Watergate

A part of this thesis will also be published as paper for a scientific journal, probably under the title *"Modelling cable and pipe failures from excavation works"*

(This page is intentionally left blank)

"Knowledge is power only if man knows what facts not to bother with."

- Robert Staughton Lynd

(This page is intentionally left blank)

PREFACE

For the past seven months, I have been in a roller coaster while studying the possibilities of predictive modelling on cable and pipe networks. Two years ago, I was relieved that I pass my exam 'probability and statistics' because I never had to do statistics again. One year ago, I would have laughed if someone told me that I would be graduating based on a quantitative study, applying multiple programming languages during the thesis. However, during this journey at the TU and Evides water company both, statistics and the programs R, SQL and GIS piqued my interest and became daily matter. Evides taught me a lot about cable and pipe networks, network operators as company and their currently applied statistical methods. Over the last month, I have been working on a paper for a scientific journal about the statistical methods I have used. Who would have thought that two years ago?! Several people contributed largely during my journey to overcome the unknown aspects on my path within the world of probability.

Therefore, I would like to thank my graduation committee, who guided me through the past couple of months. Firstly, special thanks to my first supervisor Mark de Bruijne who contributed largely by his extensive comments on the report and the structuring of my chaotic thoughts. Your serenity and positive feedbacks during our sessions gave me self-confidence about my thesis progress. Secondly, I would like to thank Tina Nane for sharing her extensive knowledge about statistics and math. Without her, finalizing my thesis would have been impossible. Then, many thanks to the chairman of the committee; Rogier Wolfert. Your knowledge and good eye for interesting studies directed me from the kick-off meeting on. I am really grateful that you sent me in the more quantitative direction and the sharp but constructive criticism that I received during our meetings. Last but certainly not least, my last committee member from the TU, George Leontaris. Thank you for participating in my committee, you always made time for feedback sessions where you had a sharp eye for the details and equations.

Also, I would like to thank Evides water company in Rotterdam and especially my company supervisor Thomas van Manen. From the first meeting on, you provided me with a lot of knowledge about statistics, network operators and helpful tips and tricks to create a proper thesis. Your experience as former graduation student at the Civil Engineering department of the TU Delft really helped me in many ways. Evides as a company made me feel welcome from the very beginning. All nice colleagues which were always willing to answer my questions, shared their knowledge and helped me in the thought process, many thanks. Also, special thanks to Patrick van den Ende who was always willing to support me during the programming in PostGIS.

Furthermore, many thanks to Sandra Schuchmann, who was always willing to help and have a small talk. Also, thanks to my fellow (graduating) students and the CME board for providing me such a fun time at 'het afstudeerhok' and 'het hok', including many cups of coffee.

Last but not least, I would like to thank my parents, brothers, friends and family for the support during my graduating process. Your support and the daily phone calls during (or before) rush hour provided me the welcome distraction I was looking for.

Ronald Wijs
Delft, July 2018

(This page is intentionally left blank)

EXECUTIVE SUMMARY

INTRODUCTION

Cables and pipes are critical infrastructure systems (CISs) which are mostly located in the very crowded subsurface. Especially in urban areas, a typical road includes five to ten infrastructure systems which are all owned and managed by different entities. The CISs are spatially interdependent as these are highly interconnected due to the close spatial proximity. Despite the critical function of cables and pipes, over 30,000 cable and pipe failures from excavation works are reported in the Netherlands yearly. Multiple studies have been conducted to reduce the risk of excavation damage. These studies have mainly focused on the impact side. Remarkable as from an extensive cooperation between the network operators and other stakeholders, a guideline (CROW500) was formed that seeks to prevent cable and pipe damage from excavation works.

Despite the extra guideline and the close spatial proximity between cables and pipes in cities, it is still unexplored what the effect of spatial interdependencies is on the probability of failure from excavation works.

RESEARCH QUESTION AND SUB-QUESTIONS

Therefore, the objective of this thesis is to develop a model to accurately predict cable and pipe failures from excavation works, considering spatial interdependencies. The associated research question is:

"What method can predict the influence of spatial interdependencies on the probability of failure from excavation works on the cables and pipes of subsurface utility operators?"

Three sub-questions were stated to guide the research and answer the main research question. Herewith, a method to accurately predict failures from excavation works is provided by:

- Identifying the variables that are most related to cable and pipe failure from excavation works;
- Assessing to what extent the variables affect the probability of failure from excavation and how this can be predicted accurately;
- Exploring how the findings can be implemented by network operators.

SUB-QUESTION I

What variables are most related to cable and pipe failure from excavation works?

A literature review and three qualitative in-depth expert interviews served to identify the related variables that were already known. The gained knowledge was adopted during data collection in which multiple databases had to be combined, resulting in the categorical dependent variable failure or non-failure. Currently, network operators measure the network performance in failures per kilometer per year, whereby the network is assessed integrally. In this study, all locations and excavation activities are considered unique, which alters the possible modelling methods. As the modelling method, logistic regression was selected to identify what variables are related to cable and pipe failures from excavation works and to predict failure from excavation works.

By assessing the statistical significance on the one hand and application of a stepwise backward elimination procedure based on the Akaike Information Criterion on the other hand, the relevant variables were identified. The three methods together; literature review, expert interviews and logistic regression were able to answer the first sub-question. Multiple variables were identified by all three methods, but also new variables were identified by the expert interviews and logistic regression model. Newly identified were the diameter of the own asset and emergency excavation requests, which both stood out from the logistic regression model as factors with statistical significance in relation to failures from excavation works.

SUB-QUESTION II

To what extent are the identified variables affecting the probability of failure from excavation works and how can we accurately predict the probability of failure?

In logistic regression, the probability of Y (the dependent variable) occurring is predicted. In this way, a proper model can be applied to identify to what extent other variables affect the probability of failure from excavation works. Therefore, the statistical significance of both the individual variables as well as from the entire model (with the relevant variables) were tested. In order to assess the predictive performance of the logistic regression, two methods (repeated K-Fold cross validation and the Receiver Operator Curve) were used to validate the model.

However the full dataset which is used for the logistic regression is imbalanced as it contains more than 100,000 non-failures and only 180 failures. From the validation, it followed that the predictive performance of the logistic regression is poor, as the balanced accuracy was 0.50 which is like a coin flip. Therefore, three alternative (sampling) techniques for logistic regression with rare event data were tested.

First, the method of King & Zeng (2001) was applied, which uses weighting and under sampling. In this way, instead of maximizing the log-likelihood, the weighted log-likelihood is maximized. From the validation, it followed that a sample set with four times more non-failures than failures, including the weight, was able to predict with a balanced accuracy of 0.66, whereby 38% of the failures were accurately predicted and 94% of the non-failures.

The second approach to overcome the imbalanced data is the Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), in which new data for the minority (failures) is created. This is done by taking the k-nearest neighbors and selecting a random point on the line in between. On the other hand, the majority class (non-failures) is under sampled to create an almost perfectly balanced sample set. During validation, a balanced accuracy was found of 0.61, whereby 58% of the failures were accurately predicted and 63% of the non-failures. A third approach, Bayesian logistic regression had similar results as the logistic regression applied on the entire dataset, so it had no predictive power at all.

Both, SMOTE and weighting were capable to predict failures more accurate than the full data model. However, none of the models was capable to accurately predict all failures.

SUB-QUESTION III

In what way can network operators use the model to reduce failure from excavation works?

A gap analysis identified the difference between the current situation and the desired situation. Next, a SWOT-analysis has been used to assess the strong and weak points of the current situation, whereas the opportunities and threats followed from the developed model to come to the desired situation in the best possible way.

From the analysis, three possible application strategies were found by a TOWS-analysis. First, the risk-assessment that is currently done by expert judgements can be complemented by the model. Besides it can train the experts and contribute to reach more consistency during these risk-assessments. Second, the model proved once again that incorrect data has large influence on failures from excavation works. However, despite the incorrect data, the models are still capable to predict 38% and 58% of the failures, which were not averted by the expert judgements.

CONCLUSION

From this thesis it followed that various steps are necessary to come to an accurately predicting model. In the end a combination was found of methods for rare event data and a binary logistic regression model. The models that followed from the two sampling techniques, Weighting and SMOTE were capable to predict failures from excavation works with a balanced accuracy up to 0.66 and 0.61. Even though both models (weighted and SMOTE) cannot predict all failures accurately, it can increase the accuracy of the predictions which are currently done by experts. A combination of the two predicting

methods, by expert and one of the models will be able to predict cable and pipe failures from excavation works more accurate than is currently attained by experts only.

During the final stages of this thesis research, the primary focus moved to writing a scientific article as mentioned in the Colophon. Consequently, the paper contains more detail on some parts than this thesis. The paper is focused on the application of rare event data methods for network operators.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF TERMS

AC	Asbestos Cement
AIC	Akaike Information Criterion
AIR	Autonomous Inspection Robot
AMI	Asset Management Infra
AUC	Area Under the Curve
BAG	Basic Registration Addresses and Buildings
BEEL	Assessment External Effects Pipes
CIS	Critical Infrastructure System
CPH	Cox Proportional Hazard
GIS	Geographic Information System
IV	Independent Variable
KLIC	Cable and Pipe Information Center
KLO	Cable and Pipes Consultation
KWR	Watercycle Research Institute
NA	Non-Available
ROC	Receiver Operating Characteristic
SGT	Stochastic Gradient Treeboost
SMOTE	Synthetic Minority Over-sampling TEchnique
SUE	Subsurface Utility Engineering
WIBON	Law Information Exchange Surface and Subsurface Utilities
WION	Law Information Exchange Subsurface Utilities

(This page is intentionally left blank)

1 INTRODUCTION

1.1 RESEARCH BACKGROUND

An infrastructure system is a network of independent, "man-made systems and processes that function collaboratively and synergistically to produce and distribute a continuous flow of essential goods and services" (Marsh, 1997, p.3). Well-being of citizens and the economy of a nation depends on the continuous and reliable functioning of its infrastructure systems. Of those systems, the Critical Infrastructure Systems (CISs) are the ones "whose incapacity or destruction would have a debilitating impact on the defense and economic security" (Ouyang, 2014, p. 44) of nation states.

Many CISs are entirely or partially located in the subsurface, where the underground is very crowded in urban areas. A typical city road includes five to ten underground infrastructure systems, all owned and managed by different entities, mostly making decisions without any mutual coordination or information sharing (Osman, 2016). Over 1.7 million kilometers of cables and pipes are already situated in the subsurface in the Netherlands and the amount is anticipated to increase as the economy and population are expected to growth, as well as through innovation (e.g. fiberglass) (Groot, Saitua, & Visser, 2016; Rijksoverheid.nl, 2017). Each year, major investments are made in subsurface infrastructure in the Netherlands. The forecasts are that about €100 billion will be invested between 2015 and 2030 (Groot et al., 2016). The investments are made for extension and for rehabilitation of the networks. Rehabilitation in the Netherlands is defined by EN 752: "measures for restoring or upgrading the performance of existing drain and sewer systems" (Tscheikner-Gratl et al., 2016, p. 13). So, rehabilitation contains all preventive maintenance activities, concerning all aspects of the network's assets (Tscheikner-Gratl, 2015). Rehabilitation is always planned for the longer term, therefore infrastructure companies moved their focus toward pro-active approaches, using predictive analyses (Engelhardt, Skipworth, Savic, Saul, & Walters, 2000; Tscheikner-Gratl, 2015).

In contrast to rehabilitation, planning of repairs is not possible since repairs are done almost immediately after failures because cables and pipes have a vital function for a country and its citizens (Tscheikner-Gratl, 2015). Failure can be caused by excavation activities. In 2015 more than 530,000 excavation requests and 32,858 damages from excavation works were reported in the Netherlands alone (Kabel- en Leiding Overleg, 2016), which is 5.7% of all cable and pipe failures (Kabel- en Leiding Overleg, 2016). Excavation damage and third-party damage of cables and pipes refers to any damage caused by a person which is not directly associated to the network (Wei & Han, 2013, p. 2527). The direct repair costs of the excavation damages are over € 26 million per year, and the indirect costs are estimated to be €100 million per year in the Netherlands alone (Van Mill, Gooskens, Noordink, & Dunning, 2013).

1.2 PROBLEM ANALYSIS

1.2.1 REGULATIONS FOR EXCAVATION

To reduce excavation damage the Law Information Exchange Subsurface Utilities (WION) has been introduced. WION was introduced with the goal to limit danger and economic damage due to damage from excavation works (Kadaster, 2017). It obliges a 'KLIC-request' (Cable and Pipe Information Centre) before any mechanical excavation work is conducted (Kabel- en Leiding Overleg, 2015, 2016; Van Mill et al., 2013). A KLIC-request is in this thesis defined as the obligatory request that is done before mechanical excavation takes place. There are three types of KLIC-requests (Kadaster, n.d.):

- Orientating KLIC-request, which is used if the applicant is not planning to excavate within 20 days. It is even not allowed to excavate with this type of request and therefore serves other goals than excavating, such as preparation or design work.
- Regular KLIC-request, which allows the applicant to start excavating from the third day after the request until the twentieth day after the request.

- Emergency KLIC-request is for unexpected situations that could harm civilians, cause other major damage, or result in a critical supply stop. The emergency KLIC-request allows the applicant to start immediately after the request.

WION and KLIC were renewed in 2017, to modernize the data exchange methods and implement INSPIRE, the European legislation. The renewed WION has been introduced in the first quarter of 2018 as Law Information Exchange Surface and Subsurface Utilities (WIBON) (Kadaster, 2017). KLIC will become KLIC-WIN in the beginning of 2019.

As of the revision of WION and KLIC, CROW250 was also revised in 2017 into CROW500. Both, CROW250 and CROW500 have been developed by the Cable and Pipes Consultation (KLO), which is an extensive cooperation between network operators, excavation contractors and clients in the cable and pipes branch (Van Mill et al., 2013). The CROW 500 (2016) is, like CROW 250 was, a guideline that seeks to '*prevent damage on cables and pipes from excavation works, guideline carefully excavating from initiation to implementation phase*' and expands the obligations and responsibilities that are already stated by the law (WIBON and KLIC). Both, the duty of providing information for network operators as well as the duty of careful excavating by contractors and clients were specified (CROW-werkgroep, 2016). Among the specifications the early mapping of risks and the localization of cables and pipes during the design phase are included (CROW-werkgroep, 2016).

Figure 1: CROW500 is the guideline in the Netherlands for anyone who is planning to excavate (mechanically) (CROW-werkgroep, 2016)

1.2.2 RISK-ASSESSMENT

The mapping of the risk and localization of the cables and pipes should be done already in the concept phase of the project life cycle (CROW-werkgroep, 2016). The gathered data is used as a basis for the risk-assessment, which is obligatory for the initiator before any (mechanical) excavation work is conducted (CROW-werkgroep, 2016). In the risk-assessment all possible conflicts between cables and pipes are analyzed, as also other characteristics of the excavation location (e.g. soil conditions). During the risk-assessment, more detailed information can be requested from other network operators such as function, diameter and depth. Altogether, it should result in suitable control measures and agreements between parties about the required precautions, such as (temporary) relocation of cables and pipes, alternative designs design or protective measures (CROW500, 2016).

The precautions and control measures follow from theoretical (location) data from the orientation KLIC-request. As the real location of cables and pipes could deviate from the theoretical location, CROW500 also prescribes the obligatory localization of cables and pipes through trial trenching. Trial trenching comes from archaeology and is used to provide a sampling of a larger area (Palmer, 2015). For excavation works, the size of trial trenches depends on three parameters (CROW-werkgroep, 2016):

- The excavation profile is the area where the actual excavation work will be conducted. The size of the area is determined during the design phase (CROW500, 2016).
- Design and measure tolerance is the area directly around the excavation profile. The size of the tolerance area is based on the kind of work (e.g. cable construction) and the type of machine used for the work (e.g. small excavator). For a standard excavator, the tolerance is 0,50 (horizontal) meters, for all other types of excavating (e.g. drilling, piling) the tolerance should be determined per unique situation (CROW500, 2016).
- The total size of the search area follows by adding a margin for error of the data. According to WION, a cable or pipe must be located within a meter (two-sided) of its presumed virtual location, therefore CROW500 prescribes another meter trial trenching (CROW500, 2016). An example of a

search area is shown in Figure 2. Network operators determine how many trial trenches should be used to reduce the risk.

Figure 2: The excavation profile (CROW 500, 2016). For example: the light blue dashed line could be the location of the actual trial trench.

Altogether, according to the CROW 500 guideline, the area to be investigated by trial trenches must thus be at least 1.5 meters around the actual excavation location. WIBON does not prescribe anything on the accuracy of vertical location data, therefore it became mandatory by CROW500 to search an extra 20 cm in depth than the expected depth (CROW500, 2016). Multiple network owners, poor data management and unnotified relocation of utilities are reasons for inaccuracy, out-of-date and sometimes even incomplete utility location data (S. Li, Cai, & Kamat, 2015).

Several methods and innovations, such as Subsurface Utility Engineering (SUE), have been presented over the past years to improve the accuracy of the available data (Li et al., 2015; Osman & El-Diraby, 2007). SUE is an emerging engineering process that aims to accurately locate and depict cables and pipes to reduce underground excavation accidents (Jeong, Abraham, & Lew, 2004). Furthermore, for example the water distribution companies developed BEEL (Assessment External Effects Pipes), which was introduced for mains nearby important objects such as highways, public locations and flood defenses. If the impact of a mains failure nearby these important locations is high, extra control measures like protecting covers and constructive measures are taken to reduce the probability of failure. So, the risk which exists of a probability and an impact determines what kind of measures the water network operators should take. BEEL is mainly focused on the impact-side of risk as illustrated in Figure 3.

During the evaluation of the WION it was found that the obligated KLIC-request created more awareness about excavation damage for both, contractors and network operators, since they always receive the locations of other cables and pipes before excavating. This follows from the increase in both, preparation- and excavation requests, which resulted in the absolute reduction of excavation damage by 9% in 2015 (Kabel- en Leiding Overleg, 2016; Van Mill et al., 2013).

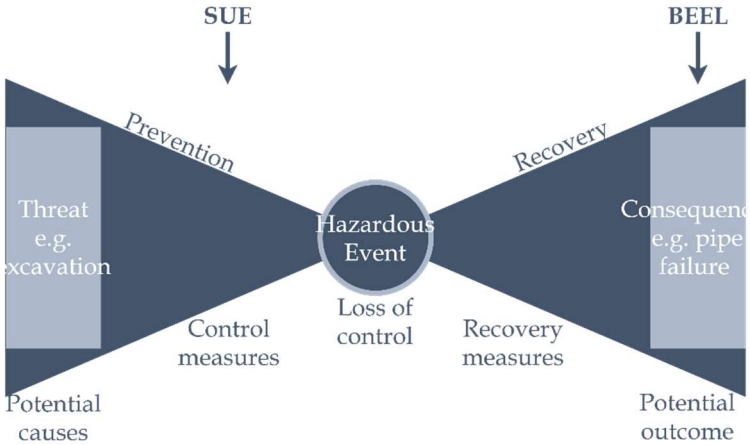


Figure 3: Subsurface Utility Engineering (SUE) and Assessment External Effects Pipes (BEEL) within a bow-tie model

KLO (2015) studied excavation damages and all kinds of components in the Netherlands. It was found that almost half of the excavation damages were related to data communication cables (Kabel- en Leiding Overleg, 2015, p. 15). Besides, half of the damages were on service connections, which are not included in the maps that excavators receive from KLIC (Kabel- en Leiding Overleg, 2015). Service connections concern all cables and pipes between the distribution networks and clients' property, both private individuals as companies. More details of excavation damages are discussed in section 1.2.4. So, multiple studies on many aspects have been conducted, resulting in measures to reduce the risk of excavation damage.

1.2.3 SPATIAL INTERDEPENDENCIES

Nevertheless, CROW 500 and KLO do not consider the probability of failure from excavation works based on distances between cables and pipes (spatial interdependencies). The minimum horizontal distance of 1.50 meters for localization of nearby cables and pipes through trial trenches, was determined without empirical data. It was only based on uncertainties of the kind of work and possible real locations that deviate from the virtual locations as described in the previous section (CROW-werkgroep, 2016). (Geo-) Spatial interdependencies are considered important for collocated infrastructures when these are considered for rehabilitation or renewal (Islam & Moselhi, 2012). However, the impact of spatial interdependencies on excavation damage is unknown.

The CISs are highly interconnected and interdependent, which indicates a bidirectional interaction (Osman, 2016; Ouyang, 2014), meaning that "the state of one infrastructure affects or is correlated according to the state of another infra-structure" (Utne, Hokstad, & Vatn, 2011). Interdependency is defined as: "A bidirectional relationship between two infrastructures through which the state of each infrastructure influences or is correlated to the state of the other. More generally, two infrastructures are interdependent when each is dependent on the other" (Rinaldi, Peerenboom, & Kelly, 2001, p. 14).

Geographical interdependency, a specific type of interdependency arises when the state of multiple or all infrastructures can be changed through one local (environmental) event (Rinaldi et al., 2001). This occurs when the infrastructure elements are in close spatial proximity, whereby events such as a pipe burst can affect the geographical interdependent infrastructure systems. As a result, it is also called spatial interdependency. Based on the physical proximity, more than two systems can be spatial interdependent (Rinaldi et al., 2001).

Many studies were already conducted on spatial (or geographical, geospatial, co-located) interdependencies between critical infrastructure systems (Hokstad, Utne, & Vatn, 2012). It followed that the concept of spatial interdependency had a similar meaning in all the examined literature. From here, this study will use the term '*spatial interdependency*' for the concept.

Just as scientists, governing bodies acknowledge the importance of (underground) infrastructure and the spatial interdependencies either. Therefore, norms and supplementary criteria were drafted. In the Netherlands, NEN 7171-1 and NPR 7171-2 (*Underground utility networks planning - part 1: Criteria*, *Part 2: Planning process)* norms are the basis for the design of the underground utility networks (Normcommissie 349 200 "Dwarsprofielen," 2009a, 2009b). Both norms are complemented by local (municipal) procedures, such as '*Handboek Leidingen Rotterdam – 2015*' (Slee & Tjan, 2015).

The stated design criteria by the norms and procedures are based on possible consequences in case of failure and set limitations to the design possibilities that can affect spatial interdependencies. All relations between the various stakeholders, norms, guidelines and regulations for all types of networks are summarized in Figure 4.

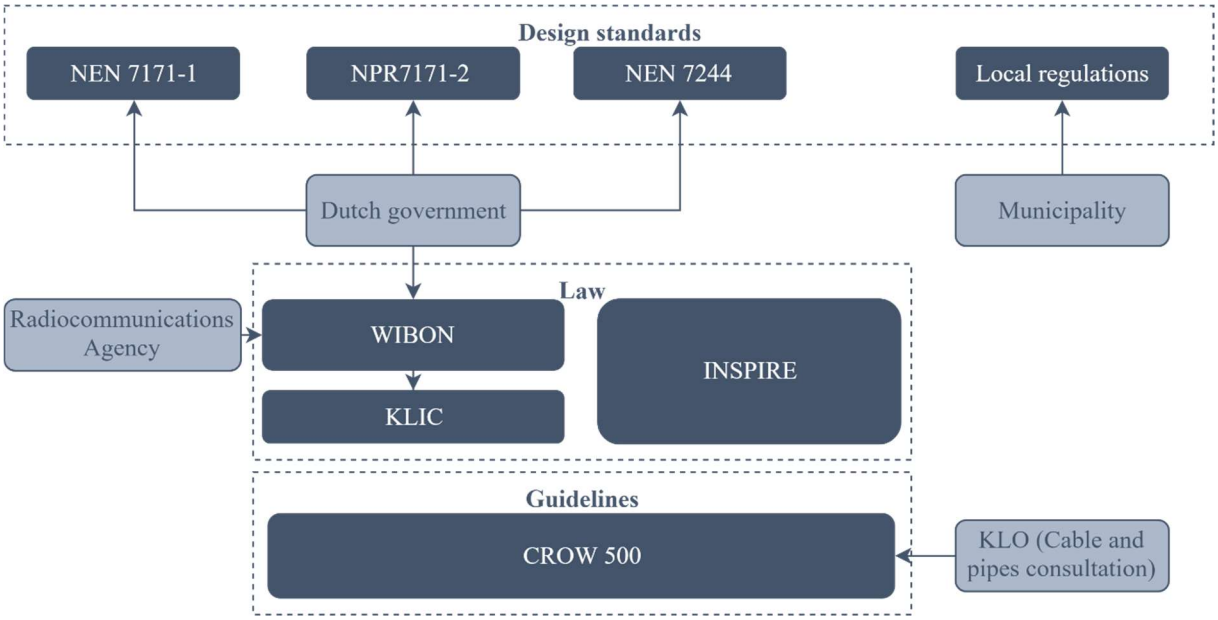


Figure 4: The various laws, regulations and design standards concerning excavation activities and cable and pipe designs in the Netherlands. The light blue boxes are the stakeholders responsible for the development and supervision of the corresponding boxes (own illustration)

1.2.4 EXCAVATION DAMAGE

Notwithstanding the many regulations, consultations and guidelines to reduce the number of failures from excavation works, excavation damage remains a big problem. Of all failures, over 18% of water distribution system (KWR, 2011) and over 12% of gas network's failure in the Netherlands (Kiwa, 2013) are caused by third parties, mostly through excavation work. Within urban areas excavation works cause even 30% of all failures. In the entire world, third party damage is a main cause for infrastructure's failure as shown in Table 1 (Wei & Han, 2013).

Table 1: Overview percentage damages caused by third parties

Country	Period	Third party induced damage (% of total failure)
USA	1987-2006	38% (gas)
EU	1970-1998	50% (gas)
China	1999-2002	40% (gas)

The Kabel- en Leiding Overleg (2015) studied excavation damage and found that 83% of third party-induced damage was caused during mechanical excavation against 14% by manual excavation. Further specification of the mechanically caused damages showed that excavators caused 54% of those. The other 46% were caused by manual excavation (16%) and other machines. From all types of work, constructing and removing cables and pipes causes most damage (59% of all damages). Despite the CROW500 guidelines, which prescribe localization of cables and pipes, more than half of all damage was caused without localization through trial trenches (53%) (Kabel- en Leiding Overleg, 2015). Some type of networks experience more problems than others as illustrated in Figure 5.

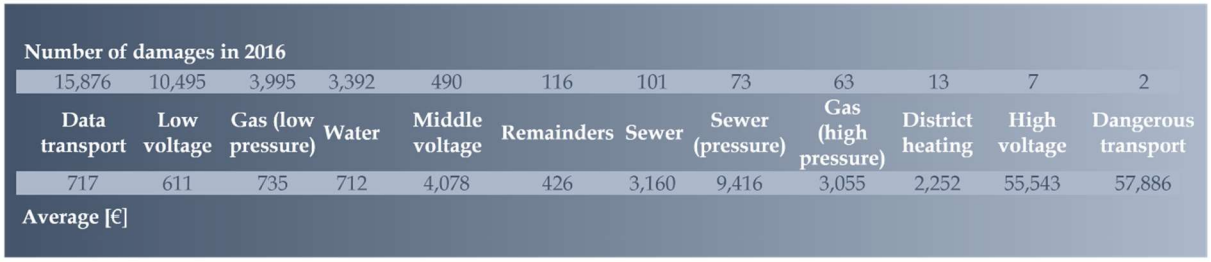


Figure 5: The number of damages from excavation works that network operators had in the Netherlands in 2016 and the average costs per damage (Agentschap Telecom, 2016)

1.3 PROBLEM STATEMENT

Even though spatial interdependencies are considered as important for collocated infrastructures, it is still unknown how spatial interdependencies affect cable and pipe failure from excavation works. The gap of knowledge is remarkable as network operators stated their own stricter guidelines, additional to the governmental norms. This indicates a clear acknowledgment of a problem around excavation damage by cable and pipe operators. Although risk is calculated by multiplying probability and impact (consequence), all reviewed literature was focused on the consequences of failure from excavation works, and none on the probability. Regardless of the close spatial proximity between cables and pipes in cities, it is still unexplored what the effect of spatial interdependencies is on the probability of failure from (a type of) excavation. As a result, network operators are not able to pro-actively prevent failures from excavation works.

1.4 STRUCTURE OF THE THESIS

This research will focus on spatial interdependencies in relation to excavation damage. The subject has been introduced in this chapter. Hereafter, the research approach will be explained, as shown in Figure 6.

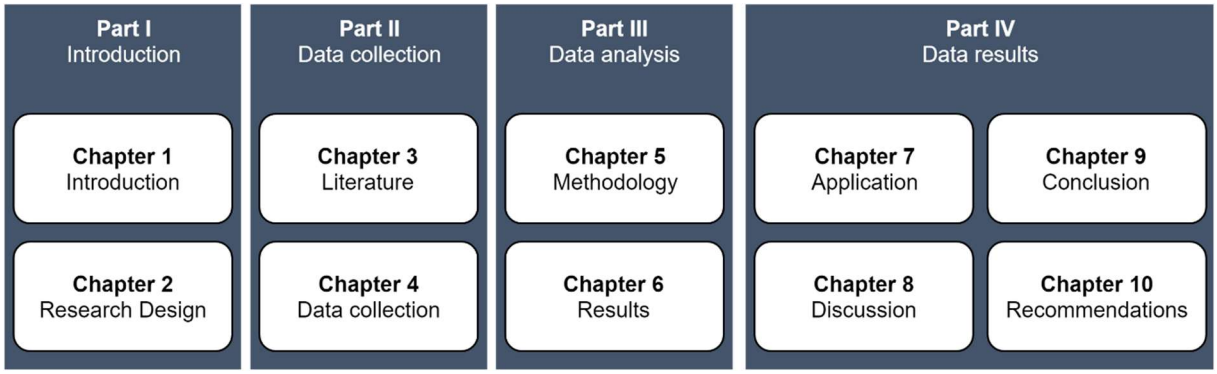


Figure 6: Overview of the thesis structure

2 RESEARCH DESIGN

2.1 OBJECTIVE

The objective of this research is to develop a model which accurately predicts cables and pipes failures from excavation works, considering spatial interdependencies. On the one hand, application of the model should predict the probability of failure from excavation works, allowing network operators to decide if risk mitigating measures are necessary.

2.2 RESEARCH QUESTIONS

The analysis, problem statement and establishment of the research objective led to the following main research question:

What method can predict the influence of spatial interdependencies on the probability of failure from excavation works on the cables and pipes of subsurface utility operators?

For the structuring of the research, and as contribution to answering the main research question, three sub-questions will guide the research:

1. *What variables are most related to cable and pipe failure from excavation works?*
2. *To what extent are the identified variables affecting the probability of failure from excavation works and how can we accurately predict the probability of failure?*
3. *In what way can network operators use the model to reduce failure from excavation works?*

Based on the relevant literature, expert knowledge and model selection, the most related variables will be identified in the first sub-question. So, what variables are most related to excavation damage? Once identified, the statistical significance of the variables in a predictive model will be tested and validated, whereby the focus is mainly on the predictive accuracy. The model will help to answer the second question. Third, it is examined how the outcome from sub-questions one and two can be implemented into network operators' systems. The knowledge of all sub-questions together will enable answering the main research question by combining all answers into one method.

2.3 SCOPE

To define what is included in this thesis, a scope is set to define the boundaries of the project. The literature and expert interviews will focus on all networks and the spatial interdependencies between those. In contrast, for the modeling in this thesis, Evides - a water distribution company in the Netherlands - is used as case. For the excavation damage on water distribution systems, the interdependencies between for example the gas pipes and sewer systems are less important. Therefore, the interdependencies between other infrastructures than between water distribution systems and others are examined to a lesser extent.

Second, the study has a specific focus on city centers and old residential areas. These areas have a high population and building density and are both used for mixed functions (such as living and shopping area). Due to upper ground developments in the past, the underground situation is often unorganized and crowded. On the one hand, because of the crowded underground, there is a large probability of failure from excavation works. On the other hand, the consequences of failure are large in the busy areas. Both, the probability and the consequence result in a high risk. The unorganized underground increases the probability for excavation damage (Vloerbergh & Beuken, 2011).

Furthermore, infrastructures are socio-technical systems, a special type of complex systems engineering (Ottens, Franssen, Kroes, & Van de Poel, 2006). The social side of the socio-technical systems refers to human involvement, such as regulations, laws, procedures and standards, where the technical refers to systems engineering (Ottens et al., 2006). For the goal of this thesis, which is to develop a predictive

method to determine the influence of spatial interdependencies, quantifiable variables are preferred since more modelling methods for numerical variables exist and data is available. On the other hand, for example, carelessness of an excavating contractor is hard to measure, whereas mutual distance between cables and pipes are numeric. The former is not quantified, whereas the latter is already from itself. To support the modelling, this thesis is mainly focused on the technical side of the socio-technical systems. Once the technical possibilities are identified, a later study could test for example the social side of the spatial interdependencies and excavation damage.

2.4 CASE: EVIDES, WATER DISTRIBUTION COMPANY

As the objective of this thesis is to develop an accurate predictive model of cable and pipe failures from excavation works, detailed data of cables and pipes and the accompanying failures are essential. Predicting these failures can be integrated in decision making processes for Evides. The model can be adapted generally by other network operators. To gain these data, a case study at a network operator is conducted. A case study is an appropriate approach to explore the possibilities of modelling cable and pipe failures from excavation.

This thesis is conducted from the point of view of Evides Water Company, the second largest drinking water company in the Netherlands. Evides has 722 fte employees, who served in 2016 156.7 million m³ safe and clean drinking water to 2.5 million consumers and business in three provinces: The South of Zuid-Holland, Zeeland and the West of Noord-Brabant in the Netherlands (Figure 7). Besides, Evides Industrial water served 95.4 million m³ to over 400 large industrial companies in the Netherlands, Belgian and Germany in 2016 (Evides, 2017).

Evides N.V. is owned by B.V. Gemeenschappelijk Bezit Evides for 50%, which are the municipalities in the South of Zuid-Holland. The other half is owned by PZEM, which is owned by the provinces Zuid-Holland, Zeeland and Noord-Brabant and 16 municipalities, mostly in Zeeland (Evides, 2017). Within Evides, the department Asset Management Infra (AMI) manages all 14,717 kilometers of pipes. In 2016 there were around 500 pipeline failures, causing an average downtime of 18.2 minutes per customer per year, from which 11.4 minutes were scheduled downtime (for replacement activities) (Evides, 2017). The average unplanned downtime which is caused by failures, is 6.8 minutes per customer per year.

To manage all the assets, AMI uses extensive databases in a Geographic Information System (GIS). The databases include information about the characteristics and locations of pipes, leakages, KLIC (excavation) request and other assets. This research is conducted with the databases possessed by Evides only, within the region named *Rijnmond area* which is highlighted in Figure 7. From this, Rotterdam is selected for the analysis as there is an extra database available (Rotterdam3D) with the cable and pipe data of all other networks.

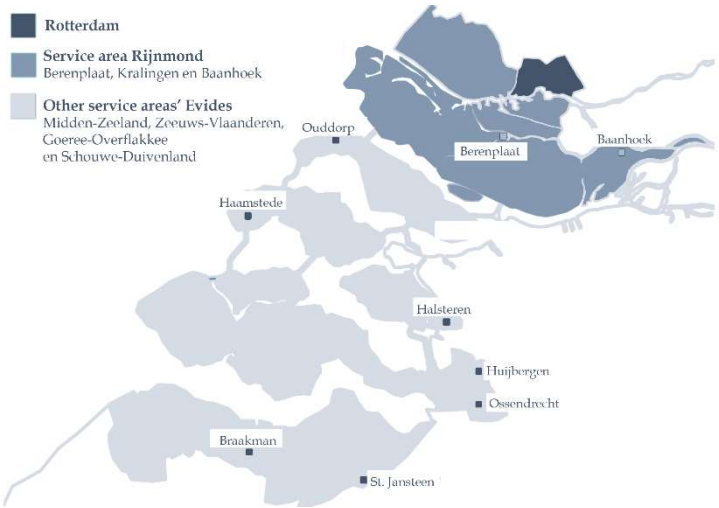


Figure 7: Evides' service area, the Rotterdam area is highlighted in dark blue (Evides, 2017)

2.5 RESEARCH STRATEGY

2.5.1 DATA COLLECTION

A research strategy has been developed to come to the defined research objective. To identify the relevant variables, three approaches are applied. First, scientific and technical literature which addresses pipeline failure and spatial interdependencies is reviewed to determine what variables were already identified in earlier studies. The literature also serves as a starting point for the data collection that will be done from Evides' databases. During the data collection, four databases are connected based on multiple criteria, such as geographical location and dates.

Second, once the data collection is completed, expert interviews are conducted to both verify the variables from literature as well as to identify other relevant variables. The interviews will be semi-structured interviews with three experts in the field of excavation damage within Evides.

The newly found variables from the expert interviews are, if possible and available, added to the dataset that is prepared for the third step in answering sub-question one. The experts, if familiar with the case, can also contribute to find representative replacements for the incorrect and incomplete data. These data irregularities could be an outcome of the data selection. The third step for answering sub-question one contains the model selection of a logistic regression model which should both support the earlier found variables as well as identify other relevant variables. As the modeling approach, logistic regression has been selected as it is an accepted way to assess the relation between a categorical dependent variable and various independent variables (Ariaratnam, El-Assaly, & Yang, 2001). The modeling is done with statistical software R, but before employing the model, multiple tests on the dataset are conducted. Multicollinearity, over- and underfitting and completeness of the data are checked before the model is developed. All steps together, literature, data collection, expert interviews and modelling should enable to answer the first sub-question.

Sub-Question I: *What variables are most related to cable and pipe failure from excavation works?*

2.5.2 DATA ANALYSIS

Next, the developed model, including all relevant variables that are available, is tested on variables individual contribution to the model fit as well as with respect to the goodness of fit of the overall model. Given that the objective of this thesis is to accurately predict cable and pipe failures from excavation works, the model is also evaluated from a predictive point of view. Therefore, a validation step is undertaken.

First, the individual contribution of each variable is tested with the Wald-statistic, which shows a variables contribution in predicting the outcome of the model. Second, the goodness of fit of the entire model is tested in three ways, by using the log-likelihood ratio, McFadden's pseudo coefficient of multiple determination (R^2) and the Akaike Information Criterion (AIC) as these are generally accepted for logistic regression. Third, the model is validated by both, determining the Area Under the ROC Curve, as well as by repeated K-fold cross validation. The former is a manual validation technique, the latter a machine learning technique.

Because cable and pipe failures are rare event data (<1%), the dependent variable is imbalanced. Imbalanced data refer to a dataset where "there are many more instances of some classes than others" (Chawla, Japkowicz, & Drive, 2004, p. 1). This particularity of the data lead to a decrease in the predictive power of the logistic regression, which has been confirmed by the analysis in this study. Therefore, three alternative sampling techniques (Bayesian logistic regression, weighting and Synthetic Minority Oversampling) were applied in an attempt to overcome this problem. The results of the methods are compared in order to test what method increases the predictability of the model to the most accurate failure predictions.

After selecting the 'best' rare event data strategy for a particular dataset, the individual contribution and goodness of fit of the model is tested again in order to check whether the performance of the model has improved in comparison with the full data model. The best predicting model based on the AUC and repeated K-fold cross validation and all the other tests contribute in answering the second sub-question.

Sub-Question II: *To what extent are the identified variables affecting the probability of failure from excavation works and how can we accurately predict the probability of failure?*

2.5.3 DATA RESULTS

Once the relevant variables are known and implemented in a predictive model, new knowledge has been gained. To test how the model can contribute to reduce cable and pipe failure from excavation works, possible implementation and application strategies for network operators are developed.

First, the difference between the current and the desired situation is analyzed by a gap-analysis, which is a generally applied technique to determine the proposed state (Marra, Biccari, Lazoi, & Corallo, 2018). What is the difference and how could the model fill this gap? The contribution of the model is determined based on a SWOT analysis, whereby the opportunities and threats follow from the developed predictive model. In a TOWS analysis, the opportunities and threats are combined with the strong and weak points of the current procedures to come up with possible application strategies. This should lead to the answer of sub-question three.

Sub-Question III: *In what way can network operators use the model to reduce failure from excavation works?*

Once the answers for all sub-questions are found, all information is combined to be able concluding and answering the main research question. Furthermore, the research limitations and recommendations are discussed.

3 LITERATURE

Literature has been reviewed to identify what variables are already known as related to failures from excavation works from previous studies. However, first some basic assumptions from this thesis are discussed. The parties that cause the actual excavation damage are the contractors of the network operators. In this report, the excavating parties are considered as the network operators (clients) instead of the actual excavating contractors. So, if the hired contractor of Stedin (gas network operator) causes damage, the gas company is held responsible. This is possible as Kabel- en Leiding Overleg (2015) found that there is no significant difference between the number of damages that these contractors in the Netherlands cause per excavation request.

Before the relevant variables are determined, one should define what is included and what is excluded. Excavation damage, or third-party damage was defined once as "any accidental damage done to the pipe as a result of activities of personnel not associated with the pipeline" (Muhlbauer, 2004, p. 3/43). The definition indicates the potential for third-party damage. On the contrary, this research will focus on the probability of pipeline failure from third-party damages, which is an important distinction. The former excludes variables such as pipe strength and characteristics of the excavation activities, whereas the latter does include the extra variables, since these variables could be the difference between damage and (registered) failure (Muhlbauer, 2004). Muhlbauer (2004) already conducted research on the former, where this thesis focusses on the latter. Therefore, the definition of excavation damage used in this research is:

An event that is registered by one network operator as failure which is caused by excavation activities of another party's personnel.

The definition excludes damage due to earth movement, natural disasters, intentional damage and damages caused by own personnel accidentally, as these are not related to spatial interdependencies. There are many possible variables that affect the probability of failure from excavation works. These variables are classified into three categories: how is failure caused, physical factors and non-physical factors.

3.1 CAUSES FOR FAILURE FROM EXCAVATION WORKS

From the definition of excavation damage, the way assets are damaged becomes important. Kabel- en Leiding Overleg (2015) did study which actions led to excavation damages in the Netherlands. In 83% of all registered third party failures, mechanical excavation was the cause against 14% manual excavation and 3% other reasons such as heavy trafficking (Kabel- en Leiding Overleg, 2015). The kind of activities that led to the failures were also analyzed as shown in Table 2.

The Watercycle Research Institute (KWR) did study the increasing excavation activities in the Netherlands. A significant relation was found between excavation areas and failure rate of water mains. In particular, a significant increase of the failure rates in excavation areas of sewers was found, and to a lesser extent also in the excavation areas' gas pipes (van Eijk & van Daal, 2013). Excavation damage on the water distribution systems was in 40% of all the excavation damages water distribution systems encounter caused by contractors of the sewer network operator. Van Eijk and van Daal (2013) found that imprudent excavation and ground settlement are major reasons for failure from excavation works, but do not elaborate on the reason why excavation on sewer systems causes damage more often.

Multiple causes for third party induced damage were found in literature, including some that are not within the scope of this study. Among the aspects that are not included, are all activities other than excavation damage, whereby other third party induced damages will be excluded from this study. For instance, heavy trafficking and leakages owing to ground movement in a certain period after excavation activities are excluded *("na-ijleffect")* (van Eijk & van Daal, 2013). Furthermore, Muhlbauer (2004) classified wildlife, seismic charges and projectiles also as third-party induced damage, which all are

categorized as factors beyond control. Last, malicious (excavation) damage (J. Li, Zhang, Han, & Wang, 2016; Wei & Han, 2013) is not included, since intentionally caused damages are not predictable considering the same aspects as unintentional damages. The probability of failure from illegal excavation (J. Li et al., 2016) without a KLIC-request is not part of the problem in this study and is therefore excluded as it is not considered problematical.

Table 2: Kind of activity that led to third party-induced failure (Kabel- en Leiding Overleg, 2015)

Kind of work	Percentage	Specification	Percentage
Constructing and removing cables and pipes	59%	Sewer	36%
		Fiber glass	35%
		Unknown	28%
		Other	1%
Paving work and road construction	9,7%		
Planting and grubbing up trees	2,6%		
Other activities (e.g. directional drilling, dredging, sheet piling, drainage, unknown, heavy trafficking)	28,7%		

3.2 PHYSICAL INFLUENCE FACTORS

SURROUNDING CHARACTERISTICS

The probability of failure will naturally rise if more excavation activities happen nearby. Generally, this is equal to the population density factor, as well as to the presence of other cables and pipes, as both lead to more activities and more frequent excavation (Muhlbauer, 2004).

Tree roots have been identified as a cause for sewer blockage and other pipe failures (Marlow, Boulaire, Beale, Grundy, & Moglia, 2011). However, this is direct blockage (failure), there is no literature elaborating on the influence of tree roots affecting the probability of excavation damage. However, Evides conducted an internal study, from which followed that the presence of trees does not influence the failure rate, and even decreases the probability of failure from excavation works. It is strongly believed that the decrease is the result of a mechanical excavation prohibition under the crown of trees (Van den Ende, 2016).

Furthermore, Riley and Wilson (2006) found that soil conditions do matter when looking at excavation damage. Higher soil density increases the soil strength, which reduces ground movements. In here, the presence of ground water should be included as it affects the soil strength (mostly) in a negative way. In general it holds that the higher the soil strength, the smaller the probability of failure from excavation works (Riley & Wilson, 2006).

OWN PIPE CHARACTERISTICS

From both, Muhlbauer's (2004) and Riley and Wilson's (2006) analyses on third party induced damage, the minimum depth of cover followed as important variable. Third party damage reduces if a pipe is located deeper, as less excavation activities will affect this pipe. Furthermore, the wall thickness of pipes is also mentioned as influencing factor on the probability of failure from excavation works, as thicker walls are proving to provide additional protection against failure from external damage (Muhlbauer, 2004). In line with the thickness of walls, also the wall material and the pipe joint systems affect the probability of failure from excavation works (Riley & Wilson, 2006).

OTHER CABLES AND PIPES CHARACTERISTICS

As already described in the previous section, it does matter on what network type the excavation activities are conducted. On the other hand, several spatial factors were found, affecting the probability of failure from excavation works. First, the horizontal distance between cables and pipes are relevant, as well as the vertical separation between cables and pipes. The larger the separation, the smaller the probability of failure from excavation work. The spatial variables also affect the probability of failure

indirectly, because more separation reduces the probability of failure from failure of one of the other cables and pipes (collateral damage) (Riley & Wilson, 2006).

Riley and Wilson (2006) also found a relation between the diameter of pipes and the probability of failure. A large cable or pipe needs more excavation, which increases the probability of failure from excavation works.

3.3 NON-PHYSICAL INFLUENCE FACTORS

Besides the environmental factors, also soft or human factors affect the probability of failure from excavation works. One of these factors is the public education about excavation work, which should create awareness about the consequences of failure from excavation works for network operators, as most failures arise unintended and due to ignorance (Muhlbauer, 2004). Even minor scratches, appearing irrelevant for the general public, could have a potential high impact on the network (Muhlbauer, 2004).

A KLO (2015) study showed that 41% of the damages was caused without KLIC-request. In 53% of all failures, the excavator did not localize the cables and pipes through trial trenches (KLO, 2015). Especially in urban areas, the ignorance of the regulations is attributed to smaller contractors as these disruption has larger consequences for them than for large contractors (Muhlbauer, 2004). Furthermore, the incentive to avoid excavation damage is low for the excavating party, since the financial consequences of damaging other networks (financial compensation) is smaller than the costs of avoiding damage on other networks.

Van Norden (2013) found that time affects the probability of failure from excavation works. Especially time constraints cause imprudent excavation. Mostly, projects with a limited budget that are relative small compared to large excavation projects, cause more damage due to imprudent excavation. Additionally, large companies often have more modern techniques for careful excavation, such as ground radars and cable locators, than small excavating companies (Van Norden, 2013).

Tscheikner-Gratl (2015), Amador and Magnuson (2011) and Van Norden (2013) all describe the influence of coordinated excavation activities as an influencing factor on the probability of failure from excavation works. According to Tscheikner-Gratl, coordination is that "instead of examining all public networks separately (all of them in an integrated way implementing all available influences), which are intertwined in our street networks, the road network is considered a container for all together and is used for prioritization" (Tscheikner-Gratl, 2015, p. 14).

3.4 DATA QUALITY

As described in section 1.2.2, the real location of cables and pipes must be within one meter of the virtual KLIC-request location. Despite this requirement, 13% of the excavation damages (2012-2014) in the Netherlands were caused in a situation where the data deviated more than the maximum of one meter (Kabel- en Leiding Overleg, 2015). Already in 2004, Muhlbauer identified incorrect data as one of the factors affecting the probability of failure from excavation works. He proposed a line locating program, to identify the exact location of cables and pipes (Muhlbauer, 2004). Trial trenching, as described in CROW500, is like the line locating program of Muhlbauer.

The correctness of the data is hard to improve. During the evaluation of WION, Van Mill et al. (2013) found that excavating parties have a lack of knowledge about the requirement to report deviating situations. Besides, reporting charges no income for the contractors, as well as that reporting deviating situations is seen as entrusting the network operators, often also contractor' s clients, with administration work. On top of that it followed from the evaluation that in case of damage caused by excavating, the excavating party is held responsible by law, even if the data are incorrect (Van Mill et al., 2013). Therefore, the incentive to improve data accuracy of one network operators is low for other

network operators. In case the contractor reports the deviating situation properly, it is still hard to revise the pipe locations as the report is not detailed enough (Van Mill et al., 2013).

3.5 FINDINGS

Multiple aspects were found in all three categories. Table 3 summarizes the findings of literature.

Table 3: The most relevant variables found from the literature study categorized in three groups

Causes	Physical influence factors	Non-physical influence factors
Mechanical excavation	Population density factor	Awareness
Excavating party	Tree roots (positive)	Ignorance of regulations
Imprudent excavation	Soil conditions	Lack of time
	Depth (vertical)	Techniques
	Wall thickness	Coordination
	Wall material	Data correctness
	Horizontal distance	
	Diameter	

4 DATA COLLECTION

As a preparation for the predictive model that has to be developed, data is collected. Furthermore, data collection serves as a starting point of view to identify the relevant data. For the model, data from multiple data bases will be collected. Expert interviews will complement the missing data and contribute to the identification of the relevant variables.

4.1 SAMPLE COLLECTION

Multiple data from different databases are combined to become useful for the analysis. In the upcoming sections, the various databases used for this study and the applied filters are explained. The most important variables for the linking of different databases are explained in the last section and the correctness and completeness of the collected data is assessed.

4.1.1 KLIC

As already explained and elaborated in detail (§ 1.2.1), all mechanical excavation activities should be notified on beforehand at the Kadaster. Between 2012 and 2014 98.4% of all excavation damages were caused despite a prior KLIC-request (Kabel- en Leiding Overleg, 2015). The study does not indicate whether these requests are related to city centers and old residential areas or not (Kabel- en Leiding Overleg, 2015). However, comparing numbers of the percentage excavation damages per municipality in the Netherlands, does show that there is no relevant difference between the largest 30 municipalities and all the others (Agentschap Telecom, 2016).

Evides stored all the KLIC requests in its network in a GIS database since June 2010, which will be called the *KLIC database*. From this date forward, Evides started using a server for automatic KLIC-request handling, which also stores the data. As the KLIC requests are stored in GIS, geographical information is available, enabling to filter them, based on location.

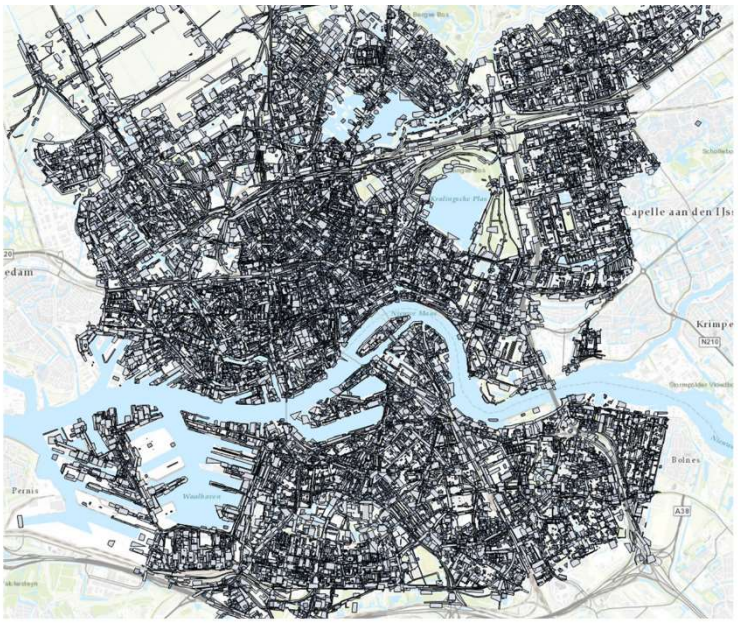
To prepare the KLIC-database for the analysis, some filters are applied on the data. First, three types of KLIC-requests are distinguished; orientation-, regular- and emergency requests. Orientation requests are only informing and do not allow parties to start excavating until a regular excavation request is done (Kadaster, n.d.), therefore orientating requests are filtered out of the main analysis. However, it should be tested whether there is a relation between KLIC orientation requests and excavation damage. In case a connection is found, orientation requests should be included in the dataset that is modelled again (section 0).

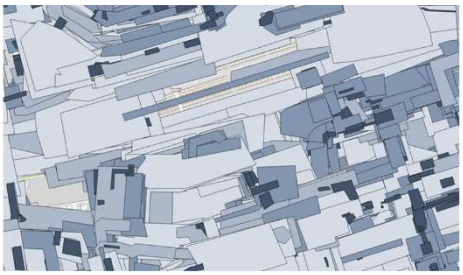
Second, the application date of the KLIC-request must be noted and should be after the database's realization date, as from this moment all KLIC-data are stored. Besides, an application date that is registered as if it is before the database's formation would indicate incorrect data.

Third, the Kadaster allows KLIC-requests up to a polygon of 500 x 500 meters. It is very likely that the size of the polygon and the number of assets located in it are related. As large polygons will contain multiple assets, it becomes hard to predict what cables or pipes are affected by the planned excavation work. Excavation activities are mostly very local. Therefore, a maximum size for the KLIC-polygon is set, where the area's type should be considered (e.g. city center or country side). The remaining polygons are categorized into four (size-)categories to ensure that small- and large polygons can be differentiated. In this way, these categories could become a variable during modelling.

The above-mentioned criteria are used to filter the data that will be used during the analysis. The KLIC-requests, indicating excavation activities, will be used as the basis for the analysis as it is the only way to identify excavation activities on a large scale. Other indicators for excavation activities are unknown. During the interviews (section 4.2) it should become clear what other available data of KLIC-requests, such as type of company and excavation type are useful for the analysis.

In Figure 8, the remaining KLIC-requests for the Evides case are shown. The figure on the right is zoomed in on the map and shows that many KLIC-requests are done for single areas. The colors indicate the four size categories that were assigned. Over 80% of the remaining KLIC-requests is smaller than 5,000 m².





	Size [m²]	Left after filtering [#]
1	0-1000	36,691
2	1000-5,000	23,975
3	5,000-10,000	8,885
4	10,000-25,000	8,901

Figure 8: Left: the remaining KLIC-requests after filtering. Right: A zoom in whereby the weights per area size have different color (see legend)

Apart from the filtering, some KLIC data is adjusted. First, when a contractor/network operator does a KLIC-request, there are 50 possibilities for the type of work, from which a maximum of six per KLIC-request can be selected. To simplify the data for the analysis the types of work are categorized into five types of work; cables and pipes, construction, landscaping and gardening, piling and drilling, remainders and unavailable. The type of work 'cable and pipe work' is specified further per network type. The further specification is also helpful to determine what network type was behind the request, as mostly only the contractor is known in the KLIC-requests. A complete overview of the grouping of the types of work can be found in appendix 12.2.

4.1.2 COMPANIES' CABLES AND PIPES

Once the KLIC-requests are filtered, the own cable or pipe data should be prepared for analysis. It is assumed that none of the network operators stores its data with the same variables in the same way. However, nowadays, it is assumed that all cable and pipe operators use a Geographical Information System (GIS) to store their assets' locations.

If available, network operators can include historical cables and pipes, as (excavation) damage could result in replacement of an asset. Historical cables and pipes refer to two types of assets, first removed assets that were taken out of the field. Second, in field assets that are out of order. Therefore, these data are still interesting for the analysis. To ensure that the analysis only has to be conducted once, it is recommended to combine the historical asset data and the current asset data if possible. Merging the two databases is done by selection of matching variables, where it is suggested to include at least date of construction, function and date of removal, as these variables are used for filtering.

Once the databases are merged, some filters are applied to it. First, service connections are removed, as these are assumed to be right-angled on the distribution cables and pipes, creating a problematic situation when mutual distances are determined later on. Furthermore, service connections are (normally) not included in KLIC-requests, the starting point of the analysis. However, in some way the service connections will be included as it will be examined whether the other cables and pipes are

located between the own asset and houses or not. Hereafter, special assets like sinker pipelines, which are pipelines located underneath waters (e.g. canal crossings and ditches), are removed as these often cause data irregularities. After filtering the 'special' assets out, the duplicate data and data with irregular dates are filtered.

Furthermore, when pipes are connected to KLIC-requests, the pipe removal date should be after the application date of KLIC to be relevant for the study. If the date is before, the possible reason of removal is certainly not the excavation activity following from that KLIC-request. To connect the data properly, correct (date-)entries are of great importance.

To conclude some technical preparations in GIS should be done, to create a workable dataset for analysis. On the one hand, cables or pipes are denoted as 'lines' in GIS, without consideration of the real asset length. Some of the lines have intercepting ends, which means that one of the lines is over the other line. If that is translated to the real world, it should mean that two pipes are on the exact same location. To avoid these double 'lines' (=cable/pipe) and to create a better entity and less chaos in the dataset the intercepting lines are merged. In this situation merge means, that two lines become one.

On the other hand, a minimum shape length is set as the 'line length' does not explain anything about an asset. Some lines referring to pipes are 300 'meters', while others are only 0.4 meters. The minimum length is set to ensure loose connections at crossings are removed.

4.1.3 FAILURES

Some aspects of failure data are relevant for the analysis. A network operator must be able to link a failure to an asset and, if related, also to a KLIC-request. First, the failure's date is important as this indicates whether it happened in a period (just) after a KLIC-request, as well as if the nearby asset was in use during failure. Second, failures are caused by many reasons. To distinguish what failures followed from excavation works, network operators need a method to classify various types of failure. Last, network operators need to have location data of failures. Figure 9 shows all registered and linked failures from excavation in Rotterdam between 2010 and 2017. A failure is defined as an event that required Evides' intervention and was recorded on their servers as a leakage (Trotter, 2017).

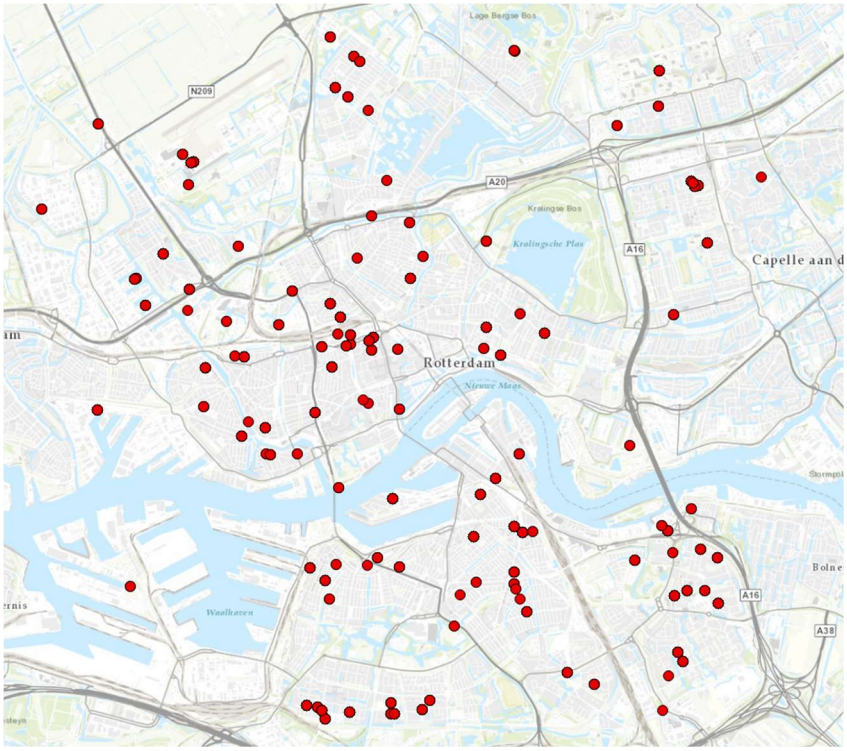


Figure 9: All registered failures from excavation works in Rotterdam that are linked to an asset and a KLIC-request

4.1.4 OTHER'S DATA

Once all internal datasets are collected and linked together, data of other networks is collected. The data from others is relevant as this study focusses on the spatial interdependencies between cables and pipes and other objects in close spatial proximity.

Identify all available data, whereby available data refers to the willingness of other parties to share the data. On one hand, this is done by contacting other network operators. On the other hand, by checking the possible existence of umbrella organizations whom possess all data. The availability of data is not self-evident, as cables and pipes data are mostly confidential, aiming to prevent malicious damage. Once available data is found, the relevant variables are identified. It is proposed to collect at least the assets' locations, diameters, types and materials.

In Evides' case, the municipality of Rotterdam developed a 3D city model to enable multiple parties to use their unique database. Besides the standard web viewer, the data is also available on request per neighborhood in multiple formats. A special map was included with a lowered ground level, which makes multiple 'hidden' city components visible, including cables and pipes, lamppost connections and tree roots. The cables and pipes are grouped by group, layer and class. Therefore, it is possible to extract the desired data from the rest of the model (that contains also upper ground data. Furthermore, the outer diameter of all cables and pipes is available, as well as some more details for a part of the data.

For this study, the locations, network type, and outer diameter of all cables and pipes in the area were received. Besides, the municipality shared the location and size of the tree roots and lampposts.

Besides, gas network operator Stedin and Evides try to coordinate their rehabilitation activities. Therefore, location data and the preferred rehabilitation moment are mutually shared. The coordination is done to prevent possible negative effects following from the other's excavation works. Within Evides, the used rule of thumb for coordination is to coordinate the activities if the Evides pipe's material is Asbestos Cement (AC) and it is located within a meter from the gas pipe. The Evides data is compared with the Rotterdam 3D data to review the data correctness in a later section.

4.1.5 HOUSES

The cables and pipes considered in this thesis are distribution and transport pipes, not the service connections. However, many failures from excavation works are on service connections. Service connections are located between the distribution pipes and buildings, providing consumers the service (e.g. gas or water). Therefore, it is relevant to know whether the other networks are located on the 'building-side' or the 'street-side' of the own asset. To implement the side of the other networks, the nearest house to a cable or pipe is determined. To do so, a buildings database is included.

In the Netherlands, the Kadaster has the buildings database which is named Basic Registration Addresses and Buildings (BAG). In this thesis, it is assumed that all network operators have access to this database, as the Kadaster provides it to all public organizations.

The BAG database contains many data and should therefore be prepared and filtered before it can be implemented in the analysis as it benefits the calculation speed. In this study, the BAG is prepared in Q-GIS where only the buildings intersecting the remaining KLIC-requests are selected. From the selection all replaced buildings are removed.

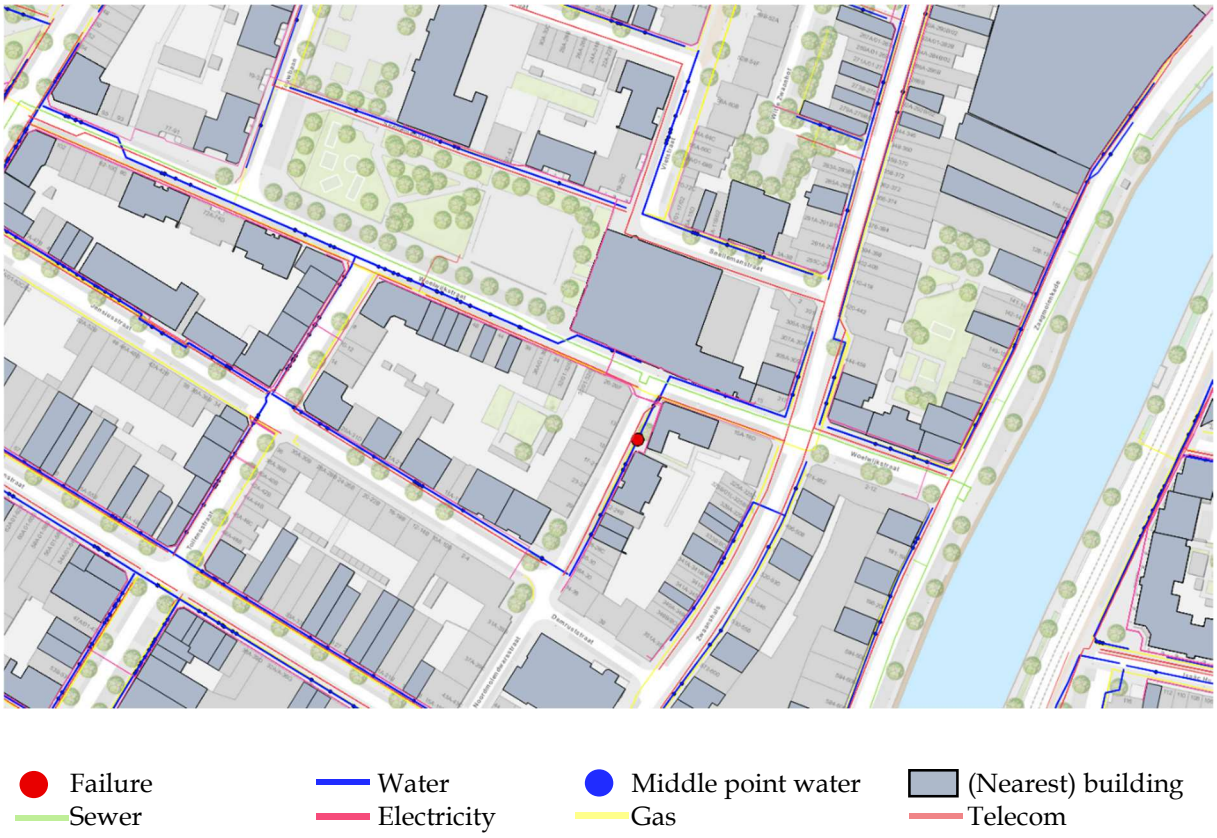


Figure 10: All network types, the middle points on the water mains and the nearest buildings including one failure. Only the KLIC polygons are left out in this figure.

4.1.6 FINDINGS

Now all individual databases are collected, and data irregularities are filtered, the databases are linked into one dataset usable for the analysis. It was found that, despite digitalization of systems, cooperation between databases does not come naturally, complicating this study, but also other analyses that network operators would like to perform. This study provides only some general guidance for the linking, as 'the best way' is very dependent on the network operators' databases. Evides' case is used to describe the linking process and the corresponding recommendations.

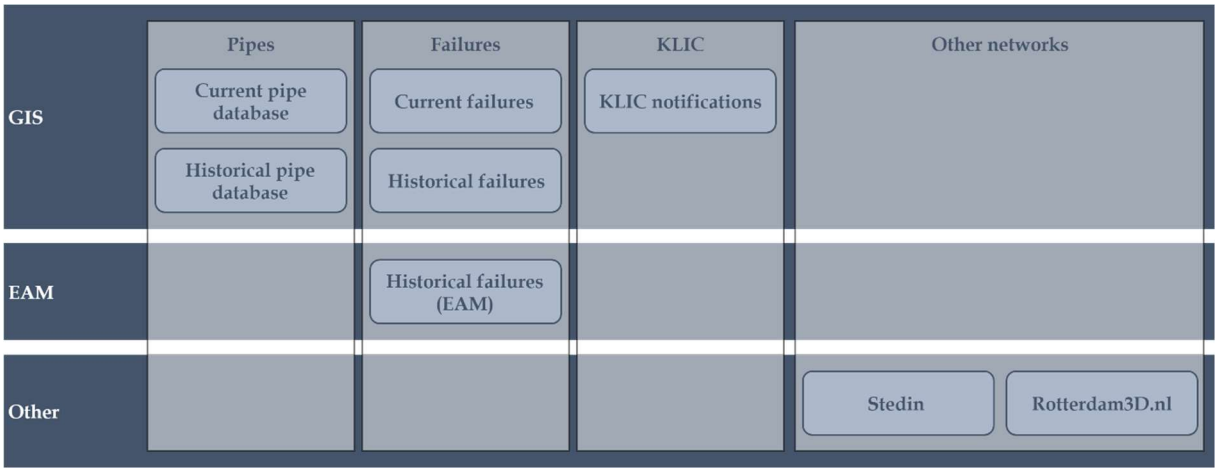


Figure 11: All used databases per data category and database type (own illustration)

As already mentioned, it is assumed that all databases contain geometry data in some way, which is the main variable during linking. First, pipes (or cables) are linked to failures. Assets are connected to the most adjacent failures within ten meters. Additionally, the asset's construction date should be before the failure date, which has to be before the asset's removal/out of use date. Connecting assets to failures succeeded for all failures.

Second, failures are connected to KLIC-requests. Where failures are "points", the KLIC-requests are polygons/areas. The first requirement for linking is that the point must be inside the polygon. Furthermore, the failure must be after the KLIC-request date, but no more than three months. From all 78,452 filtered KLIC requests between 14-06-2010 and 29-05-2018 in Rotterdam, 256 failures were connected to the KLIC-polygons. The three months period follow from the assessment of various maximum periods for connection as shown in Table 4. Besides, the three months period has been selected as an excavation activity must start within 20 days after application, but not earlier than three days after. Considering the duration of maintenance or construction work, the duration of the period could be adapted, as well as by the preferences' network operator.

Table 4: Failures related to KLIC-requests, depending on the maximum duration of the period after the request

Months after KLIC-application date	1	2	3	4	5	6
Failures from excavation works	157	210	256	289	311	332

Third, the 86,207 (distribution-, transport distribution- or transport) pipes that followed from the filtering are connected to the KLIC-requests. The connections are made based on similarities in location and date. As a result, often, multiple pipes were linked to one KLIC-request, as it is likely in a densely populated urban area such as Rotterdam, that multiple pipes are in an area when KLIC-polygons are up to 25,000 m2. Because multiple pipes (or cables) could be linked to one KLIC-polygon, the linking criteria must be considered. For example, should the assets be entirely inside the polygon, is a small intersection enough, or is a combination of both preferred. This optimal situation will differ per network operator, but they all have to consider the same aspect; on the one hand, it is preferred to model balanced data, meaning that 50% of the dataset results in failures and 50% in non-failures. On the other hand, network operators should try not to lose too much data. To reach the best situation (as many data as possible and as balanced as possible) a network operator can try several options. In this thesis eight options, following from trial and error, were considered:

1. Without any filter, the smallest interception is enough to connect an asset to a KLIC-polygon.
2. Only link the asset nearest to the middle point of the KLIC request
3. An asset should intersect the polygon with a minimum percentage of its total shape length or a minimum distance or a combination of both.
4. All assets have a virtual middle point (as explained later in this section). The asset is only connected to an KLIC-request if the middle point is located within the polygon.
5. Only the one asset with the greatest intersection length in the polygon is connected to the KLIC-request
6. Only link assets that contribute a minimum percentage to the total length within a polygon.
7. Filter the assets on their minimum shape length (7.5 meters) and set a minimum length of intersection in the polygon.
8. Filter the assets on their minimum shape length (15 meters) and set a minimum length of intersection in the polygon.

The percentage of failures and the size of the data set are plotted for each option in Figure 12. As the most balanced (or least imbalanced) dataset with as many data as possible is desired, option 8 has been selected in this thesis as it has the highest score when multiplying the two aspects (included data and failure ratio). As mentioned before, this is totally dependent on the dataset and should therefore be selected carefully. The full results of the various options are attached in appendix 12.3.

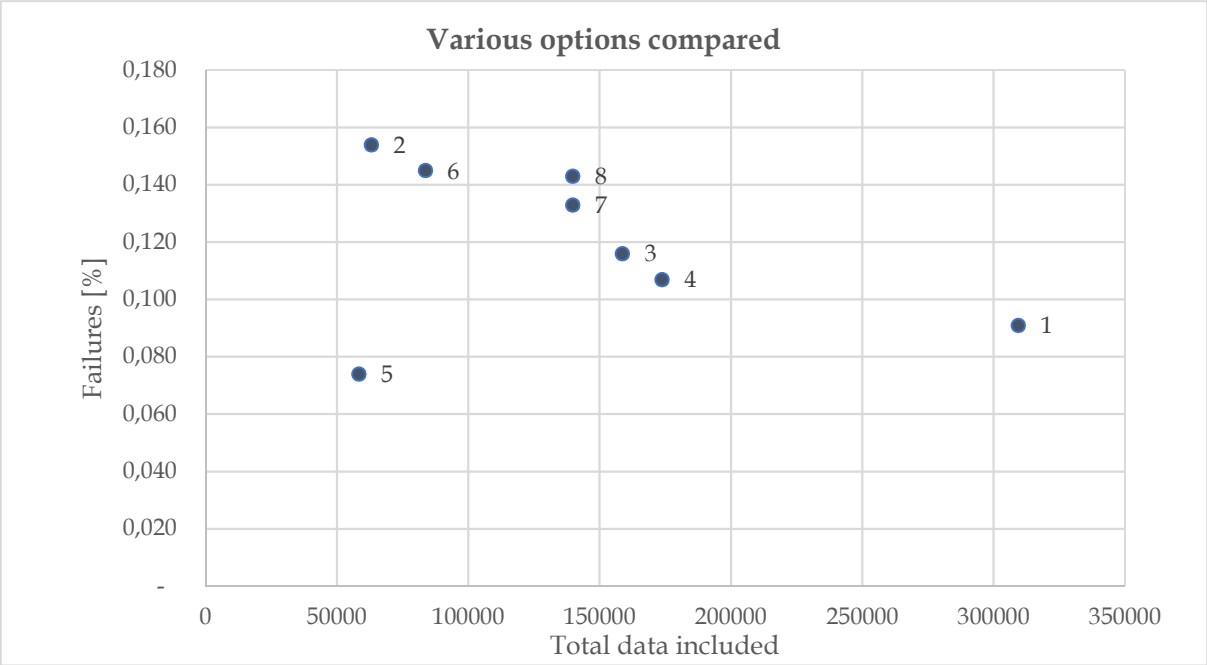


Figure 12: The results of the eight 'linking' options. Option 8 is preferred as it has the highest score when multiplying the total data included and failure ratio.

Once the connection between the own cables or pipes and the KLIC-polygons are made, the relation between the different networks is examined. To do so, the middle point of a pipe (or cable) within a KLIC-polygon is selected and a new virtual point is created. From that virtual middle point, the mutual distances to the other surrounding networks is calculated. The middle point was introduced as it was found that calculating the smallest mutual distance between networks over the entire asset length often results in a mutual distance of zero. This is because cables and pipes often cross each other whereby that point is selected as the smallest mutual distance. So, the mutual distances within a KLIC-polygon are all determined from the one middle point on the own asset.

So, to prevent misleading calculations of mutual distances, the short shape (virtual) lengths were filtered as all shapes smaller than 15.0 meters were excluded during the asset preparation already. This was done as the smaller shape lengths are mostly located at crossings where the average mutual distances are hard to determine. The mutual distance has been calculated for all networks within 10 meters from the middle point. If any further, it is considered as irrelevant when considering excavation damages, since it is not very likely that for example an excavator deviates that much (>10m) from the actual excavation location.

4.1.7 QUALITY OF THE DATASET

If possible, validation of the dataset's correctness is recommended. Especially when data is provided by other parties, the precision/correctness is unknown. When some datasets are available twice, like in the Evides case, it creates the opportunity to compare the own database with the 'foreign' database. In this thesis, Evides' data is considered as the own data and Rotterdam3D's data as the foreign data.

The deviation of foreign data against the company data can be measured on the common variables. In this study, the common variables are the (horizontal) location and diameter of the assets. It is assumed that the company data (Evides) being the network owners are original and right, and that in case of differences between the values, the foreign data are inconsistencies. As the correctness of all databases is an uncertainty, it is up to the analyzing party what dataset is assumed most reliable (true).

For the compared Evides data, it was found that 14,648 out of 19,992 compared assets (73.27%) have the exact same location. Testing the correctness of the data, results in Figure 14, which shows the decrease of the sample size if the maximum allowed deviation declines. Analyzing the data in this way, supports proper decision making by the network operators. In the Evides case, the maximum deviation from Rotterdam3D compared to the Evides data is set at 0.4 meters, which indicates that 94.5% of the data within five meters of the own assets is used for analysis. So, 94.5% of the Rotterdam3D data is considered reliable and will be included in the analysis, whereas the remaining 5.5% is excluded.

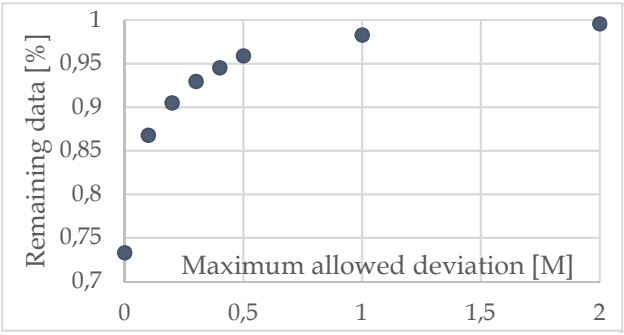


Figure 14: Remaining data after filtering the data with a deviation above threshold

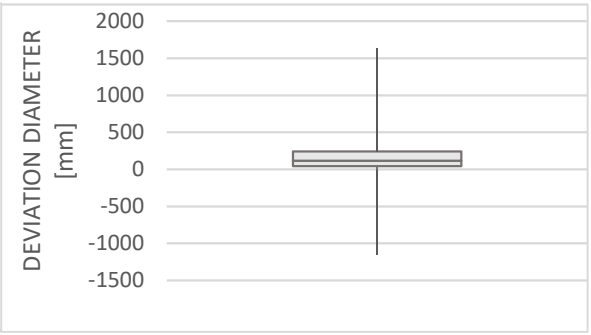


Figure 13: Distribution of the compared diameters in a boxplot

For the deviating pipes, a remarkable difference between the transport, transport-distribution and distribution pipes was found as shown in Table 5. Despite the prioritized control measures for transport pipes, the deviation from the Rotterdam3D database compared to the Evides database increases with the diameter of the pipes. The deviation of the diameter can be positive or negative, meaning that the diameter can be either larger or smaller. This fact should be used when comparing, to enable a graphical description of the quartiles by a boxplot (Figure 13). From the boxplot it becomes clear that the variation outside the upper- and lower quartile is enormous, but most of the data is quite similar. Each network operator can individually consider up to what deviation the data is reliable (enough). The average deviation of the Rotterdam3D data in the Evides case as shown in Table 5 is considered as reliable.

Table 5: Average deviation of the diameters between Evides' data and Rotterdam3D's data per type of asset (transport, transport distribution or distribution pipes)

Diameter	**Deviation [mm]**
Total [avg]	90.28
Transport pipes [avg]	457.70
Transport distribution pipes [avg]	171.58
Distribution pipes [avg]	73.27

For the modeling in this thesis, independent variable *Evi_data_quality* will refer to the correctness of the data. The variable indicates the deviation of Rotterdam3D compared to the "master" Evides database.

4.2 INTERVIEW

4.2.1 SETUP

The purpose of the interviews is to contribute to the body of knowledge that is necessary to conduct the analysis. The interviews will be qualitative in-depth, as it encourages the interviewee to describe several phenomena richly, leaving the interpretation and analysis to the researcher (DiCicco-Bloom & Crabtree, 2006).

A structure for the interviews is developed, consisting of main questions, follow up questions and probes. The main questions cover all important topics for the interview, whereas follow up questions and probes are guiding to go into more detail (Rubin & Rubin, 2005). All main questions are related to the research question of this thesis. The main questions used for interviewing were:

1. What causes excavation damage (mostly)?
2. What is the quality of the data, such as KLIC, that is available during excavation works?
3. How do spatial interdependencies affect the probability of failure from excavation works?
4. What variables affect the probability of failure from excavation works mostly?
5. According to CROW500, localization through trial trenching is obligatory over a minimum distance of at least 1,5 meters to both sides. What are the consequences of trial trenches?
6. How can knowledge about the probability of failure from excavation works be implemented, and how should it contribute to the organization?

These questions have been developed during and after the sample collection and literature study. This sequence is recommended as it contributes to develop more effective questions (DiCicco-Bloom & Crabtree, 2006). All questions are supported by multiple sub-questions, but those are partly dependent on the interviewee's answers, whereby these are not exactly prepared (Rubin & Rubin, 2005). The (semi-)structure of the interview is important for the comparability, since multiple interviews are held (Van Oel, 2017). The entire interview form, including main- and sub-questions can be found in appendix 12.3.

4.2.2 INTERVIEWEES SELECTION

The selected experts for the interviews were people within the organization who's daily work includes excavation damage somehow. In gaining initial access to the interviewees, it is most feasible if experts within the own organization are approached. If necessary, the interviewees can be asked whom else could be of interest for interviewing (Horton, Macve, & Struyven, 2004). To expose all sides of knowledge, the selected interviewees have different functions within the organization. By illuminating various sides, an 'expert' view on excavation damage from multiple asset management levels is created.

In this study, the first interviewee was a fitter, giving the infield, operational insights of failures from excavation works. What does he encounter infield during his work and is this like the expectations from the office? The interviewed fitter is employed by Evides already three years as the *district fitter* of Ommoord. Ommoord was built in the 70's, nowadays it is one of the largest districts in Rotterdam. Before Evides, the fitter did the same for the gas and electricity network operator in the area (Stedin), making him expert in several ways.

Second, an innovation and policy engineer, working at the Asset Management Infra department was interviewed. The asset manager who is primarily focused on the tactical asset management level, has according to many staff the greatest knowledge on excavation damage within the department. He has experience in the past at two other water distribution companies (Vitens and Dunea) and as researcher at the Watercycle Research Institute (KWR).

Third, a senior area manager was interviewed, who has three years' experience within Evides. Before he was employed by the municipality of Rotterdam where he did even contribute in the development of CROW500. At Evides, he oversees the entire Rotterdam harbor district.

4.2.3 RESULTS

In this thesis, the outcomes of the interviews have been analyzed in Atlas.ti, which is a powerful tool (software) for qualitative data analysis. The researcher's knowledge of the software is more important than what software is used and could therefore also be another program. In Atlas.ti, families and underlying concepts were created to quantify and clarify the results (appendix 12.3). The five families and 41 concepts were linked to interview fragments 184 times. The former are overarching themes of the concepts, which can be all topics. The following families were chosen based on the first impressions of the interviews:

- Causes, which are factors that lead to registered failure. For example, excavation damage can be caused through an imprudent contractor with an excavator.
- Data concerns all topics related to data, such as quality and completeness. So, it's not directly related to excavation damage, but it provides more insight in the data that is used.
- Variables are the concepts that could affect the probability of failure through excavation damage by a third party.
- Parties are all cable and pipe (network) operators, which are mentioned during the interviews as high-risk excavators.
- Other are the remaining factors mentioned during the interviews, such as human quality and trial trenches.

From the 41 concepts, 13 were mentioned more than 5 times. The 10 most mentioned concepts can be found in Table 7.

The model that is developed with the most correlated variables, will only use the concepts of certain families. The family *other* will not be included in it and is only used for recommendations, reflection and eventually validation of the model. This is, because the concepts in the *other* family are hard to quantify for the modelling whereas the other families are not.

A remarkable variable is the correctness of the data, which was specifically asked for during the interview, as it is an important factor for the reliability of this research. For example, correctness of the data refers to the deviation between the actual pipe location and the expected pipe location based on maps. All experts had a different opinion about the deviating distances. The area manager expects the deviation to be +-50 centimeters in general, whereas the asset manager thinks the data is 'pretty good'. The fitter thinks the virtual data is almost similar to the actual locations. Rotterdam, the area where the fitter works, is the only municipality in the Evides service area who collects all cables and pipes in a map. Furthermore, the municipality of Rotterdam is monitoring stricter if the excavation locations complies with the permitted locations than other municipalities in the area. The strict monitoring, by random checks, increases the reliability of the location data in Rotterdam, which benefits the reliability of the analysis.

Table 6: Co-occurrence of Law & Regulations concept, which gives insight in the context where the concept was used

	Horizontal position	**Human quality**	**Imprudent excavating**	**KLIC**	**Party**	**Quality Data**	**Trial Trenches**
Law & Regulations	3	1	2	1	1	3	1

Another concept that was frequently mentioned was *law and regulation*. The tendencies of the interviewees' answers varied widely on this topic. Therefore, the co-occurrences were analyzed, giving insight in the context where law and regulations were used (Table 6). The concept was mentioned three times together with the correctness of the data, three times with the horizontal distance and twice with imprudent excavating. It can be summarized as: the regulations are not followed, decreasing the correctness of the data and the horizontal distance between cables and pipes. The underlying reason, as claimed by all interviewee, is lack of time for the contractors, which results in imprudent excavation.

Some shortcomings of the procedure around KLIC-requests were found. The most reoccurring theme was the missing of service connections in KLIC-requests. All respondents agreed upon the fact that there is more excavation damage between the distribution pipes and buildings, than on the 'street side' because the service connections are located on the building side. Furthermore, when a network operator, during the design phase, determines a new possible location, issuing of the permits takes more than three days. As no excavation can be done within the first three days after the a regular KLIC-request, the new location is determined and permitted before the actual profile is determined by trial trenches. Therefore, no options are left to include the actual found profile.

According to all experts, trial trenches are a very important factor in preventing failures from excavation works. Trial trenching should be done on multiple locations in a street as data and electricity cables are flexible, which makes it possible these swerves through the streets. For water-, gas-, and sewer one trial trench should be enough as these pipes are not flexible and therefore are located parallel (and straight) to the property line. The experts had a different opinion on the minimum size of the trial trenches. The fitter believed trial trenches should be from curb to curb, covering all potential deviations. In contrary, the area manager mentioned 1,5 meters as a luxury in cities, as the cables and pipes are located very close together. The innovation and policy engineer were satisfied with the current CROW500 policy.

4.2.4 FINDINGS

First, it was found that the procedure around KLIC-requests is a problem. After the request, the contractor receives multiple maps with each map showing only one of the networks instead of one map including all networks. However, this will be solved once KLIC-WIN has been introduced. The larger problem found is the fact that half of the excavation damages are on service connections, but KLIC-requests still do not contain these connections (Kabel- en Leiding Overleg, 2015).

As already explained in section 4.1.4 (the cooperation between Evides and Stedin), network operators try to coordinate their replacement activities. According to the experts (and literature), failures from excavation works on spatial interdependent networks can be reduced by coordinating maintenance activities. For example, instead of three different moments for the maintenance activities on three different networks, all maintenance is done at the same time. This can reduce failures as most of the infrastructure systems are considered in coordination instead of focusing on a single system (Amador & Magnuson, 2011).

For the analysis, the family *causes* is related to the KLIC-requests as it all has to do with the excavation activities. The concepts that do matter according to the experts are the type of the own asset (4), the excavating party (4) and the kind of work that is executed (2).

The experts agreed that excavation activities on the sewer system cause the highest probability of failure from excavation works because it is the deepest network and has a large diameter. Furthermore, excavation work by data and telecom network operators was mentioned as having a high probability of failure. There are much more data and telecom cables in the subsurface than other types of networks (mostly, providers use their own cable network), resulting in more excavation work from the data and telecom providers. More excavation results in a higher probability of failure. Furthermore, these network operators have a lot of competition, which results in lack of time for contractors as explained in the previous section.

Some aspects from the family *variables* were mentioned very often (as shown in Table 7). Those were horizontal position, vertical position, diameter, material, and again party.

The vertical position of cables and pipes is considered important by both, literature and experts. It is important to notice that there is almost no data available on networks actual vertical position. Network operators know the depth of construction, which is based on regulations, but the real depth changes over time due to construction, landscaping and soil subsidence (S. Li et al., 2015). From the various test

points' Evides has, it is confirmed that the actual depth of cables and pipes can vary a lot from the expected depth, resulting in inaccurate data.

During the literature study and the interviews, some important aspects affecting the probability of failure from excavation works were found. In Table 7 are the findings (from literature and interviews) summarized and compared.

Table 7: Results interviews and literature study

	Family	Concept	Mentioned
1.	Variables	Horizontal position	17
2.	Variables	Vertical position	13
3.	Variables	Diameter	12
4.	Data	Quality of the data	12
5.	Variables	Material of the cable or pipe	10
6.	Other	Law & Regulations	9
7.	Other	KLIC	8
8.	Variables	What party is excavating (party)	7
9.	Parties	Sewer as damaging party	7
10.	Other	Trial trenching	7

It was found that the relevant factors that followed from the interviews were mostly similar to the relevant factors found from literature in some way. On two of the concepts found from the interviews, which are slightly expending the literature, this thesis will elaborate on further. First, it is expected that the law and regulation concept is missing in literature because it deviates per country, region or municipality. Ignorance of regulations and imprudent excavation were mentioned by literature, but from the interviews it followed that the underlying procedures are not perfect. For example, the KLIC-requests appears to be imperfect, as these do not include service connections even though 50% of the failures are caused on service connections.

Second, trial trenches were found from the interviews as very important, whereas it was not found in the scientific literature. It is obliged in the Netherlands by CROW 500, but that is (only) a guideline that followed from an extensive cooperation between network operators. However, the fact that it is included in the guideline, shows the acknowledgement of the importance of trial trenches by network operators.

5 METHODOLOGY

5.1 MODEL SELECTION

Before the collected data are analyzed, the selection of the method which is used to identify the relations between the dependent- and independent variable(s) is explained. To the best of my knowledge, almost all statistical studies on cables and pipes are conducted using failure/km/year as dependent variable, considering the network as a whole. By contrast, in this thesis the dependent variable Y is registered as failure (1) or non-failure (0), since all situations are unique and considered separately instead of all assets as a group. As a result, the dependent variable changes from numeric continuous into a categorical dichotomous type, which alters the possible modelling approaches.

Multiple statistical approaches have been reviewed and the advantages and disadvantages of every method are explained in section 5.4. In the end, binary logistic regression was selected as the method to start with. A major advantage of logistic regression is that it is generally accepted for binary outcome statistics (Hosmer, Lemeshow, & Sturdivant, 2013) and it has been shown to have good performance (Ariaratnam et al., 2001). Besides, the only studies found in literature for unique situations in the cable and pipe sector are Ariaratnam et al. (2001) and Tung (1985) whom both applied logistic regression.

LOGISTIC REGRESSION

Usually, (linear) regression models are applied to describe a linear relation between a dependent variable and the predictor (independent) variables. In multiple linear regression, the dependent variable Y is predicted by equation [1]. In this equation is b_0 the Y intercept and $b_1, \dots, b_n$ the regression coefficient of its respective independent variables $X_1, \dots, X_n$. An important assumption of linear regression is the linear relationship between the dependent variable and the independent variables, otherwise the model will not be valid.

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni} + \varepsilon_i \qquad [1]$$

where $\varepsilon_i \sim N(0,1)$ is white noise. In this study, the assumption of linearity is violated as Y is a categorical variable. To overcome this problem, logistic regression was developed in which logarithmic terms (logit) express the multiple linear regression equation (Field, 2013, p. 887). The logistic regression equation is given by equation [2], where $P(Y)$ is the probability of Y occurring and e the natural logarithms' base.

$$P(X_i = 1) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni} + \varepsilon_i)}} \qquad [2]$$

In the new equation, instead of predicting the value of dependent variable Y as for the linear regression, the probability of Y occurring is predicted. In other words, will the cable or pipe fail after the KLIC-request or not. The logistic regression method is an accepted way to assess the relation between a categorical dependent variable and various independent variables, also denoted as covariates (Ariaratnam et al., 2001). Logistic regression has been used in various applications. For example, Ariaratnam et al. (2001) applied logistic regression models to predict the likelihood of a particular cable or pipe being in a certain state (cracks versus no-cracks). They found that their application of logistic regression contributed to reduce subjectivity, as situations were assessed on probability, using historical data, rather than single numerical (condition) ratings (Ariaratnam et al., 2001).

ASSUMPTIONS OF LOGISTIC REGRESSION

Before a logistic regression model is built, some basic assumptions are tested. First, logistic regression assumes that the dependent variable follows a Bernoulli probability function having only two possible outcomes, 0 or 1, where 1 usually denotes failure and 0 non-failure with the probability

$$Y_i \sim Bernoulli(Y_i|\pi_i)$$

$$P(X = 1) = \pi_i \qquad [3]$$
$$P(X = 0) = 1 - \pi_i$$

where $\pi_i \sim N(0,1)$ is the random variable that represents the probability of failure (King & Zeng, 2001; Monroe, 2017) and is explained as in equation [2]. Furthermore, it is important that each sample is independent, so the probability without considering independent variables should be the same for every action. In this study, 1 denotes leakage and 0 denotes no leakage.

Second, the regression models assume independence of all independent variables. When some independent variables are dependent this poses some issues called multicollinearity, which should be below a certain threshold. This is, one independent variable can predict another independent variable with a certain accuracy[1]. One method to test multicollinearity is with the Variance Inflation Factor (VIF). The VIF can be tested easily by (almost) all statistical software programs. When the VIF is smaller than one, variables are not correlated, when VIF is between one and five, than the variables are moderately correlated and greater than five is highly correlated. In this thesis, VIFs lower than 2.5 are considered not to cause any problems[2]. Furthermore, independent variables can be completely separated when the individual explanatory variable predicts the outcome variable perfectly. The dataset should be tested on complete separation, especially as it often occurs in rare events data (Rainey, 2016). Complete separation arises when a dependent variable can be perfectly predicted by one variable or a combination of independent variables (Field, 2013). Rare events data are outcome variables with "dozens to thousands of times fewer ones (events, such as wars, coups, presidential vetoes, decisions of citizens to run for political office, or infections by un- common diseases) than zeros ("nonevents")" (King & Zeng, 2001, p. 138).

Third, it is known that logistic regression is affected by the proportion of 'positive' cases in the sample. Logistic regression models need many events (samples) relative to the independent variables being evaluated, especially for rare event data. For the sample size it is recommended that (Peduzzi, Concato, Kemper, Holford, & Feinstem, 1996):

$$Sample\ size = 10 * \frac{k}{p} \qquad [4]$$

where k is the number of independent variables and p the proportion of 'positive' cases.

The last test is to check if the model is overfitted. This occurs when too many (independent) variables are included in a model, resulting in a model only being predictive of the utilized dataset (Hosmer et al., 2013). A model becomes overfitted when it is trying to follow noise patterns as it follows specific points too tightly. On the other hand, underfitting also results in poor predictive power of the model,[3] which refers to the overall performance of the model to correctly classify the cases (Statsoft, 2013). The possible over/underfitting of the model makes the theoretical basis (as examined in section 3.2 and 3.3) even more important for variable selection.

The variables, relevant according to literature and interviews, are tested on *type* and *completeness*. This is done by a complete-case analysis. *Type* refers to e.g. numerical and categorical, where completeness refers to the percentage of the cases where the variable is present or available (e.g. the district heating is not present in the entire city).

PREPARE DATASET FOR MODEL FITTING

Once all tests are conducted, the dataset is modified. As the goal of the model is to determine which independent variables are most related with excavation damage, it is also important to determine what variables should be included. The variables that followed as relevant from the theoretical bases are split

[1] http://www.statisticshowto.com/

[2] Allison, P. (2012) When Can You Safely Ignore Multicollinearity? Statistical Horizons

[3] https://stats.stackexchange.com/questions/81576/

into some smaller groups of independent variables to be tested against the dependent variable. Three groups were created, two equally sized groups with both half of the variables and an all-encompassing group. The independent variables are divided randomly over the two smaller groups (e.g. odd and even variables). This step is helpful in two ways: first, it prevents overfitting the model. Second, the more variables are included in a model, the more time the computations will take. The groups that were created are shown in the next section (6.1.1).

The samples in the groups are split into a training and a test set to be able to validate the model's predictive power once it is finished. As already mentioned in findings of the data collection (section 0), only 0.2% of the dataset's dependent variables represents failure, it is not very likely that these are divided equally over the train and test group that are created. Therefore, a technique called stratified random sampling is applied (Kothari, 2004). In stratified random sampling, homogenous groups are created, called strata. From the strata random samples are taken[4]. In this study, the only stratum used is the outcome variable which is split into failure and non-failure. From both groups, around 80% is selected randomly and placed in the trainings set, where the remaining 20% is for the test set. If available, one could also use data from another (comparable) area, which would be a comparable city in this study. However, due to limited time for this thesis and data which was hard to obtain, no other data than Rotterdam are used.

DEVELOP MODEL

Once the datasets are prepared, the model is developed. Almost all statistical software will do this automatically, but the modeling theory is explained shortly. The logistic regression equation was already shown in equation [2]. The beta parameter for each variable is determined based on maximum likelihood estimation. As its name states, the beta estimates are determined by maximizing the likelihood of the data. So, based on the predictors, the model is fitted and the parameters are estimated (Field, 2013). Taking logs in combination with equation [3] simplifies the log-likelihood to:

$$ln\ L(\beta|y) = -\sum_{i=1}^{n} \ln\left(1 + e^{(1-2Y_i)X_i\beta}\right). \quad [5]$$

SELECT THE SIGNIFICANT VARIABLES

As mentioned before, only the variables that were found relevant by literature and interviews, as well as some variables that could be interesting are used. From the training models, the statistical significance of these relevant variables is tested on the one hand. The p-values that follow from the trainings model are a measure of whether the research findings are likely to have occurred by chance or not. As in most studies, a relation is considered statistically significant when the p-value is below 0.05.

On the other hand, a stepwise backward elimination procedure based on the Akaike Information Criterion (AIC) is conducted. During this procedure, which is included in most statistical software, variables are removed one by one until the AIC of the model does not improve (enough) anymore. AIC is used to establish the goodness of fit of statistical models while accounting for the simplicity of the model, i.e. the number of parameters. The AIC is calculated by

$$AIC = 2k - 2\ loglikelihood \quad [6]$$

where k is the number of variables (including intercept) and the log-likelihood measures the likelihood of the data and has been defined in equation [5].

All three kinds of methods, the stepwise backward elimination procedure, p-value and the knowledge from the theoretical background are combined and used to select the variables that will be included in the model.

[4] https://www.investopedia.com/terms/stratified_random_sampling

5.2 GOODNESS OF FIT

The variables that are relevant to the probability of failure from excavation works were selected during model selection. In order to be able to answer sub-question 2 *'To what extent are the identified variables affecting the probability of failure from excavation works and how can we accurately predict the probability of failure?'* later on, a more complex model is developed. The complex model will build on the regular model that was already established during model selection. In first instance it will follow from goodness of fit tests.

For the goodness of fit, first the individual contribution of the variables is tested. Second, the goodness of fit of the entire model is tested.

INDIVIDUAL CONTRIBUTION

When determining whether a variable is relevant for the overall logistic regression model, the Wald statistic is most important[5]. It shows the individual contribution of a variable. The more the Wald statistic is different from zero, the more it can be assumed that the variable has a greater contribution in predicting the outcome. The Wald statistic is calculated by dividing the square of the logistic regression coefficient ($\hat{\beta}$) minus the parameter of interest (β_0) by the variance of the estimates (equation [7]). The logistic regression coefficient (estimated β) of the independent variables also indicates whether the independent variable has a positive or a negative influence on the outcome.

$$W^2 = \frac{(\hat{\beta} - \beta_0)^2}{var(\hat{\beta})} \quad [7]$$

GOODNESS OF FIT

Likewise, the individual variables, the total model fit is also tested. First, the log-likelihood statistic is examined, which explains the fit of the entire model, indicating how much of the data is explained by the model. Equation [8] shows the log-likelihood-ratio, which is used to determine whether an extra variable improves the fit of the model. Equation [8] uses the deviance, which is calculated by $-2 * loglikelihood$.

$$-2\log(\frac{L_0}{L_1}) = (-2logL_0) - (-2logL_1) \quad [8]$$

Second, the coefficient of multiple determination (R^2) is tested, which is the partial correlation between the outcome variable and each of the independent variables. It provides an insight in the substantive significance of the entire model, but it was originally developed for regular linear regression models. As the 'regular' coefficient of multiple determination is missing for logistic regression models, multiple alternatives, called pseudos were developed. Therefore, cautiousness is required when interpreting R^2 pseudo's, as its trustworthiness is smaller than the regular R^2. One of the pseudos is McFadden's R squared measure, which is the default pseudo in most statistical programs and proposed by Allison since it is closely related to the linear regression R^2 definitions[6]. It uses the maximized likelihood value from a model (L_c) and the corresponding value from the null model (L_{null}) which is a model which only includes the intercept and no other predictors:

$$R^2_{McFadden} = 1 - \frac{\log(L_c)}{\log(L_{null})} \quad [9]$$

An important side-note when interpreting McFadden's R squared is that one should not expect it to be too large. According to McFadden does values between 0.2 and 0.4 already indicate an excellent model fit (McFadden, 1979).

[5] Allison, P. (2014) Another Goodness-of-Fit Test for Logistic Regression. Statistical Horizons

[6] Allison, P. (2013) What's the Best R-Squared for Logistic Regression? Statistical Horizons

As last test for the goodness of fit the Akaike Information Criterion (AIC) is tested again. The AIC-score is first used during the backward elimination procedure during model selection, but once the variables are selected, the AIC score is tested again for the model fit. AIC is used to establish the goodness of fit of statistical models while accounting for the simplicity of the model, i.e. the number of parameters.

Once the goodness of fit of the individual variables and the total model are assessed, the model will be validated.

5.3 MODEL VALIDATION

By validating the model, it is assessed how well the model can predict the outcome. As the dependent variable is categorical, validation is done by testing the capability of the model to predict the outcome accurately. Multiple validation techniques exist, such as leave-one-out, (repeated) K-Fold cross validation and some manual techniques. In this thesis, the ROC (manual validation) and repeated K-Fold cross validation are used (Kohavi, 2016).

Splitting the data into a training and a test set as mentioned in 'prepare dataset for modeling' (section 5.1) is done for the (manual) model validation. First, the Receiver Operating Characteristic (ROC) curve is used to validate the model. In a ROC curve the true positive predicted samples are plotted against the true negative predicted samples. The Area Under (the ROC) Curve (AUC) measures the predictive accuracy of the model, where it is important to realize that 0.5 would be a coin flip and 1 a perfect model.

The second method for measuring the accuracy of the model is repeated K-fold cross validation. This machine learning technique divides the data in k subsets, which mostly is 10 as also in this thesis. One of the subsets becomes the test set, all the others are for training. In this thesis this was repeated hundred times, significantly reducing the error estimation and the bias. K-fold cross validation is used to produce a confusion (error) matrix which visualizes the performance of the model. The confusion matrix helps to determine the accuracy, kappa, sensitivity and specificity of the model. An example of the confusion matrix is given in Table 8 where TP = true positive, TN = true negative, FP = false positive and FN = false negative.

Table 8: Confusion matrix from theory

		Reference model	
		No event	Event
Predicted model	No event	TN	FN
	Event	FP	TP
		$specificity = \frac{TN}{(TN + FP)}$	$sensivity = \frac{TP}{(TP + FN)}$

First, accuracy is the general measure of a classifier, which measures the overall efficiency of the model. However, with imbalanced data the majority class contributes more to the measure than the minority class and could therefore be misleading. Second, both specificity and sensitivity measure the efficiency. The former measures the accuracy of the negative cases, whereas the latter measures the accuracy of the positive predicted cases. Third, kappa determines the accuracy that follows from the difference between the model and data that is generated purely by change. Last and most important in this thesis is the balanced accuracy, which combines the specificity and sensitivity. The balanced accuracy measures the average accuracy from both, the minority and majority class. It is calculated by equation [10],

$$Balanced\ Accuracy = \frac{1}{2}(sensitivity * specificity) \quad [10]$$

whereas a high (traditional) accuracy and a low balanced accuracy indicates that the (traditional) accuracy is high because of the classifier distribution (Akosa, 2017). The difference between accuracy, kappa, sensitivity, specificity and balanced accuracy is important as these measures can be misleading

when data is highly imbalanced. Because of the imbalanced data in this thesis (section 2.5.2), the balanced accuracy is way more important for predicting failures than the other measures.

5.4 OTHER STATISTICAL MODELS

There are multiple statistical methods which can be used to model a categorical dependent variable. Four statistical methods have been considered as relevant for this study; stochastic gradient tree boost (SGT), Cox proportional hazard (CPH), 'regular' logistic regression and Bayesian logistic regression. In a literature review, several advantages and disadvantage for the three methods were found (Grzenda, 2015; Kleinbaum & Klein, 2010; Lombardo, Cama, Conoscenti, Marker, & Rotigliano, 2015).

First, CPH has the advantage that in many situations the true hazard function is unknown or complex, whereby this does not matter for CHP. Besides, CHP focuses more on the effects of the independent variables instead of on the nature of the hazard function which is not very relevant in this thesis (Harrell, 2001). However, CPH was excluded as the dependent variable is modeled as time dependent, which is not relevant in this thesis.

Second, stochastic gradient treeboost and logistic regression were compared. SGT is very efficient and easy to implement. However, in a study on landslide events, binary logistic regression was found to produce more robust models around the mean, which results in smoother and less binarized predictions of the failure probabilities (Lombardo et al., 2015).

Last, Bayesian logistic regression assumes that the model parameters are random variables. This has the advantage that it has the ability to use out the knowledge about the sample set during modelling. However, a major disadvantage is that this approach entails very large computational performance as it has a high model complexity (Grzenda, 2015). As explained earlier, the wide application of 'regular' logistic regression has the major advantage that extensive studies on model adjustments have been executed. So, this was decisive during the selection of the method. However, Bayesian logistic regression will also be tested so it can be compared with the 'regular' logistic regression model.

6 RESULTS

6.1 FULL DATA

First, the assumptions associated with logistic regression were tested for the entire dataset. Once tested, the dataset is split into groups to develop the basic model.

6.1.1 MODEL SELECTION

BERNOULLI PROBABILITY FUNCTION

The dependent variable "*leak_dummy*" has only two outcomes, failure (1) and non-failure (0), where all samples are considered to be independent. This means, the probability of failure remains the same during all trials. However, from some logical reasoning some assumptions followed. As discussed extensively during sample collection, it is possible that multiple assets cross a KLIC-polygon. In such situations, multiple samples could be affected by a similar event, leading to doubtful independencies. It is unknown what data is involved in these situations. On the contrary, it is unlikely that the actual excavation work crosses two assets, let alone all assets within the (larger) KLIC-polygons. These considerations were elaborated on in more detail in section 0 (sample collection, findings). In this thesis, due to the lack of more detailed information on multiple assets per KLIC-polygon and the considerations in this section, it will be assumed that all samples are independent. However, one of the independent variables which is included, is the total asset length in a KLIC-polygon. If this variable seems to be below the statistical significant level, it could indicate that multiple assets per KLIC-polygon does matter.

MULTICOLLINEARITY

From the Variance Inflation Factor test some situations of complete separation of several independent variables were pointed out. To overcome this problem, the gained knowledge from literature and interviews was used to take out the variables expected to be correlated, such as *type of work I* and *type of work II* that are sequential. Taking out the variables did work and an VIF table followed. To test all the multicollinearity, the remaining variables were inserted one by one, until the complete separation error appeared again. Furthermore, a detailed analysis of the dataset, specifically checking for multicollinearity was conducted and the variables that caused the complete separation were studied, whereby the background knowledge and the way of data collection were kept in mind.

It was found that the type of KLIC-request (normal vs emergency) and the type of work are completely separated. Emergency KLIC-requests do not require any specification on the type of work and are therefore always classified as '*unknown*' in the dataset while the type of work for all normal KLIC-requests was specified. As the type of work basically explains both variables, the type of KLIC-request will be excluded.

The specified type of work '*klic_type_work2*' and the responsible party which has been determined based on several variables are also completely separated. Again, the variable that is expected, from the prior knowledge, to be most informative will be included, which is the responsible party.

The final VIF table can be found in appendix 12.6. The variables that were excluded from further analyses due to complete separation or unavailability are shown in Table 9 underneath.

Table 9: The four variables that were excluded from the analysis because of complete separation or availability

Remaining cables and pipes	**Type of work**
Remainder_Distance	*Type_of_workI* (General)
Remainder_Side	*Type_of_workII* (Cables and pipe specific)

SAMPLE SIZE

The complete separation found during the multicollinearity tests indicated the need for more samples compared to the number of independent variables already. The variables that were completely separated were excluded from the analysis. This filtered and reduced the number of independent variables from 31 to 27.

In first instance, the relation between the dependent variable and 27 independent variables not being multicollinear was tested, using al 107.500 samples. The method proposed by Peduzzi et al. (1996), recommends that the sample size should be above 150,000 whereas the dataset in this thesis only contains ≈107,000 samples. As it is not possible to increase the number of samples, the number of independent variables is decreased. Using Peduzzi's equation to find k shows that 21 independent variables is the maximum for the sample size.

$$107{,}000 = 10 * \frac{k}{0{,}0002}$$

The fact that the sample size is too small is kept in mind, but no further actions are taken here. The significance test and stepwise backward elimination in a later phase (section 6.1.1.1 - 6.1.1.4) will help to determine what variables are irrelevant and can be removed from the model.

OVERFITTING

To overcome over- or underfitting problems, the theoretical basis, including the results from multicollinearity are used together. The latter filtered already some variables out, but to overcome a model from being predictive of the own dataset, some additional measures were taken.

First, the dataset is split into three groups. The odd half of the independent variables (IVs) (group 1), the even half of the IVs (group 2) and all IVs together (group 3). The odd- and even were chosen to create equally sized and random groups. On all three groups the same analyses are conducted to find potential differences during all steps. So, the actual results from the overfitting tests cannot be noticed now but should be checked later.

COMPLETENESS

Both type and availability of data are assessed, as logistic regression only includes complete samples. The results of the assessment are shown in appendix 0. It was found that some independent variables had many non-available (NA) entries. Especially remainder pipes had many NA data (+- 93% NA). The NAs can be explained by the absence of remainder pipes in large areas' Rotterdam. Furthermore, district heating can only be found in +- 30% of Rotterdam and therefore requires attention.

From the more than 107 thousand samples, less than 10% was found to be entirely complete after excluding the remainder pipes variables. Again, the simple explanation for large number of NA data is the absence of cables and pipes within the maximum distance from the measure points or irreducible responsible parties.

To overcome this problem, multiple adjustments were made to the dataset. Where most studies use the mean of a variable to insert on empty places, this thesis will need another approach. As discussed earlier, NAs are not necessarily missing, it only refers to the absence of a network type within the maximum measure distance. Therefore, imputing a variable's mean would be inappropriate for this dataset. Instead, a value not present in the dataset should be chosen to use for imputation.

First, as the mutual distance between the measure point on the Evides main and the other networks is only included within 10 meters, a value larger than 10 is selected to use for imputation. In this thesis, 12 (meters) was selected as the imputation value for missing distances. Second, an unavailable distance will automatically result in missing diameter and side. With a very few exceptions, all diameters of cables and pipes were smaller than one meter. Therefore, '1' was selected as imputation value for the missing diameters. As the cable 'side' is a categorical variable (0 and 1), the NAs will be replaced with number 2. Last, other categorical data, such as responsible party and type of work also contain NA

entries. This happens when these variables are not traceable. If so, the empty samples are imputed with 'unknown'.

SAMPLING
All three groups, which were constructed to overcome the overfitting problem, are sampled in a stratified manner. The groups were split into a training- and a test set with respectively .8 and .2 of the group data. For all groups a model based on the training set was developed. Both, the results from the backward stepwise elimination and significance as well as the difference between the results are shown in the following sections, where the results per group are shown.

6.1.1.1 GROUP I: ODD INDEPENDENT VARIABLES

First, the model selection for the group with the odd ($n = 2,4, \dots, n_i * 2$) independent variables was executed. The 13 odd independent variables were inserted into a logistic regression model, from which their p- and z-value were determined. The most significant variable is the diameter of the own assets, which has a p-value smaller than $2*10^{-16}$ and has a negative estimate, meaning that a smaller diameter increases the probability of failure. Furthermore, the unknown type of work rises the probability of failure enormously. Important to realize is the *type of work 'unknown'* refers to the emergency KLIC-request (as a result of the data collection). Moreover, the correctness of the data (deviation between Evides- and Rotterdam 3D data), diameter of the sewers, gas distance and diameter of the district heating are below the significance level and are therefore included in the model.

On the other hand, the backward elimination based on AIC ended with the same variables as were selected by the p-values, except the gas distance (p = .08). Opposite, the age of the own asset was left in the AIC model, but has a significance level of 0.26. The AIC improved from 2591 at the start, to 2580 at the end of elimination.

The intercept, the variable representing the estimated log odds baseline when all independent variables are zero, is also very large. A large negative intercept indicates that there are many more 0 outcomes than 1 outcomes, which is true for this dataset.

In Table 10 the variables with the highest statistical significance are shown. In the table, the column estimate represents the estimated β-value for each (independent) variable. Then, the Wald-statistic (z-score) which is calculated through dividing the estimate by the standard error is shown which is assumed to be normally distributed. The p-value of the normal distribution is shown in the last column. More details on the Wald-statistic will follow in the next sections.

Table 10: The odd independent variables (group 1) with a p-value below threshold (p < 0.10) and their corresponding estimates and z-values

Significance group 1	Estimate	z value	Pr(>|z|)
(Intercept)	-5,23	-9,67	0,00
klic_type_work Kabels-Leidingen	0,88	2,10	0,04
klic_type_work Unknown	2,17	4,91	0,00
evi_diameter	-0,01	-5,66	0,00
evi_data_quality	1,39	2,04	0,04
sew_diameter	-0,70	-2,70	0,01
gas_distance	-0,05	-1,77	0,08
heat_diameter	-0,76	-3,16	0,00

6.1.1.2 GROUP II: EVEN INDEPENDENT VARIABLES

Coincidentally, group 2 contains many more categorical variables than group 1. However still the same tests are conducted for this group. First the significance from the Wald statistic was tested. The variable with the lowest p-value is the side where the telecom cables are located. That is remarkable as most studies are focused on the large diameter networks. Furthermore, the mutual distance between the

sewer system and the own asset was found very significant. The side of the electricity cable, the excavating party and the length within a KLIC-polygon are also significant.

The backward stepwise elimination has more dissimilarities with the p-values below threshold than in group 1. Only five variables had a p-value considered small enough, whereas the stepwise procedure left seven variables in the model. Both the side of the district heating was left in the model based on AIC, as well as the material of the own assets. Through the backward elimination, the AIC decreased from 2640 to 2626 in the final step. The very large intercept is not surprising, as the ratio of the dependent variable is like group 1.

Table 11: The even independent variables (group 2) with a p-value below threshold (p < 0.10) and their corresponding estimates and z-values

Significance group 2	**Estimate**	**z value**	**Pr(>|z|)**
(Intercept)	-5,36	-6,34	0,00
asset_length_in_klic	0,00	-1,65	0,10
klic_partyHeat	0,92	1,90	0,06
sew_distance	-0,09	-2,71	0,01
elec_side1	0,29	1,71	0,09
elec_side2	0,66	2,01	0,04
tele_side1	-0,94	-5,38	0,00
tele_side2	-1,07	-2,92	0,00

6.1.1.3 GROUP 3: ALL INDEPENDENT VARIABLES

In group 3, where the model selection is done with all independent variables (group I and II), most of the variables correspond to the separate groups. Eleven variables were found to have a p-value below threshold ($p < .1$), just like group I and II together. The most remarkable differences found, was that the diameter of the district heating is above threshold in this group. Moreover, the mutual distance between the sewer pipes and own asset had a very low p-value in group I but is not in this group. Also, some new variables were found in group 3. First, the side where the sewer pipes are located was found important. Second, the diameter of the gas pipes has a low p-value.

The backward stepwise elimination reduced the AIC from 2531 to 2511, which is the best AIC score of all groups. Again, in this group some differences were found between the variables left based on AIC and the significant variables. Five new variables were selected based on AIC, whereas the significant sewer side was not selected by the elimination procedure. The extra variables are presented in Table 12.

Again, the intercept's estimate has a large estimate and a low p-value. However, it is smaller than in group I and II.

Table 12: All independent variables (group 3) with a p-value below threshold (p < 0.10) and their corresponding estimates and z-values

Independent variables	Estimate	z value	Pr(>|z|)
(Intercept)	-5,47	-5,18	0,00
klic_type_workUnknown	2,20	4,96	0,00
evi_diameter	-0,01	-5,95	0,00
evi_data_quality	1,60	2,34	0,02
sew_diameter	-1,21	-2,85	0,00
gas_distance	-0,07	-1,88	0,06
klic_partyHeat	0,86	1,76	0,08
sew_side2	1,53	2,23	0,03
gas_diameter	1,63	2,22	0,03
heat_side2	-1,05	-2,39	0,02
elec_side1	0,32	1,81	0,07
elec_side2	1,00	1,88	0,06
tele_side1	-0,86	-4,72	0,00
tele_side2	-1,84	-3,65	0,00

6.1.1.4 DIFFERENCES BETWEEN GROUPS

Comparing all groups, multiple similarities and differences were found as shown in Table 14. In the table each variable found relevant in group 1 and 2 is compared with the variables found relevant in the all-encompassing group 3. In the last rows, independent variables that where only selected from group 3 are shown. The last column is the number of times an independent variable was selected, where four is the maximum. As a criterion in this thesis, an independent variable should be selected in the groups at least three times, e.g. significant in group I and 0 and left in the stepwise model from group I. Nine variables met the criterion.

First, five variables were selected in all categories. The type of work, own asset's diameter, data correctness, diameter of the sewer pipes and the mutual distance between the gas network and own network. The network type behind the request (party), and side of all, telecom, electricity and district heating were selected three times. These nine independent variables are all selected based on the statistical tests as relevant variable in relation to excavation damage and will therefore be included in the model in the next section.

Table 13: Compare the results of group 1 and 2 with the results from group 3 which contains all independent variables. All variables that were selected in at least two groups by three or four of the selection criteria are included

			Group 3		
Group 1	p-value	AIC	p-value	AIC	Total
Klic_type_work Unknown	X	X	X	X	**4**
evi_diameter	X	X	X	X	**4**
evi_data_quality	X	X	X	X	**4**
sew_diameter	X	X	X	X	**4**
gas_distance	X	X	X	X	**4**
heat_diameter	X	X			2
Evi_age		X		X	2
Group 2					
asset_length_in_klic	X				1
klic_partyHeat	X		X	X	**3**
sew_distance	X			X	2
elec_side1	X		X	X	**3**
elec_side2	X		X	X	**3**
tele_side1	X		X	X	**3**
tele_side2	X		X	X	**3**
Heat_side		X	X	X	**3**
Evi_material		X		X	2
Group 3					
Evi_distance_house		X			1
Tele_distance		X			1
Elec_distance					0
Sew_side	X				1
Gas_diameter	X	X			2

When translating these results into a logistic regression model, the equation would look like equation[11]. In this equation, only the relevant variables found from the statistical tests are inserted. Based on the theoretical background some extra variables could be added, this will be done in the (sub) conclusion section. The fitted model is

$$P(Y = 1) = \frac{1}{1 + e^{-(Z)}}, \qquad [11]$$

with: $Z = b_0 + b_1\, klic_{type_{work}} + b_2 evi_{diameter} + b_3\, evi_{data_{quality}} + b_4\, sew_{diamter} + b_5\, gas_{distance} + b_6\, klic_{party} + b_7\, heat_{side} + b_8\, elec_{side} + b_9\, tele_{side} + \varepsilon$

where $\varepsilon \sim N(0,1)$.

6.1.2 GOODNESS OF FIT

In first instance, all tests were conducted on the entire dataset, so all 107,266 non-failures and all 182 failures were included (Figure 15). The results of the tests are discussed shortly.

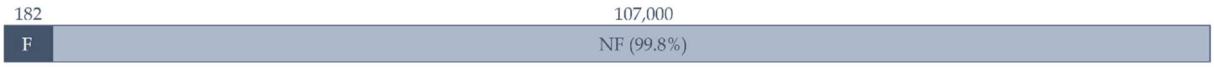


Figure 15: The sample set including all data. F indicates leakages, NF non-leakages. Only 0,169 % of the sample set represents leakages

WALD-STATISTIC

The individual contribution of the selected variables was tested based on the Wald statistic. The more the z-score differs from zero, the larger the contribution of the variable for the total model. All variables with a z-value differing more than 1 from 0 are shown in Table 14. Besides, their significance in the new model is inserted. It follows that the p-value for some variables changed compared to the group 3 model where all independent variables were included in the previous section. First, the mutual distance between the gas network and Evides' main became above the stated significance threshold level. Second, the p-value of the side of the electricity cables decreased from 0.06 to 0.01. For all other variables, the p-values remained basically equal.

Table 14: Results for the Wald-statistic and p-value of the selected independent variables

	Estimate	z value	Pr(>|z|)
(Intercept)	-5,03	-5,67	0,00
klic_type_workKabels-Leidingen	0,68	1,52	0,13
klic_type_workUnknown	2,22	5,06	0,00
evi_diameter	-0,01	-5,70	0,00
evi_data_quality	1,46	2,14	0,03
sew_diameter	-0,68	-2,54	0,01
gas_distance	-0,02	-1,28	0,20
klic_partyHeat	0,81	1,67	0,10
klic_partySewer	0,43	1,03	0,30
heat_side1	-0,61	-1,77	0,08
heat_side2	-0,70	-3,60	0,00
elec_side1	0,33	2,03	0,04
elec_side2	0,70	2,45	0,01
tele_side1	-0,90	-5,15	0,00
tele_side2	-0,93	-3,02	0,00
evi_materialPE	-1,18	-1,28	0,20
evi_materialST	0,73	1,04	0,30

LOG LIKELIHOOD RATIO

The Log Likelihood Ratio was tested to find the goodness of fit of the entire model. It was found that the Log Likelihood Ratio of the model with the selected variables, compared with a model with all variables included has a Chi-square score of 0.399, which is above the significance level ($p < 0.10$). Therefore, the null hypotheses, which states that the data are consistent with the Bernoulli distribution, is accepted[7].

COEFFICIENT OF MULTIPLE DETERMINATION

Besides the log likelihood ratio, McFadden's pseudo R^2 was calculated. It was found that the model has a partial correlation of 0.0915, which indicates a poor model fit with respect to the pseudo R^2.

[7] http://stattrek.com/chi-square-test/goodness-of-fit.aspx?Tutorial=AP

6.1.3 MODEL VALIDATION

The model also needs to be validated. Therefore, the ROC curve and the Area Under Curve (AUC) were determined. In Figure 16, the ROC curve for the developed model is given, which contains an Area Under Curve of 0.595. All scores below 0.60 are ranked as failing models, therefore this model is interpreted as not working. Nonetheless, the result is very close to 0.6.

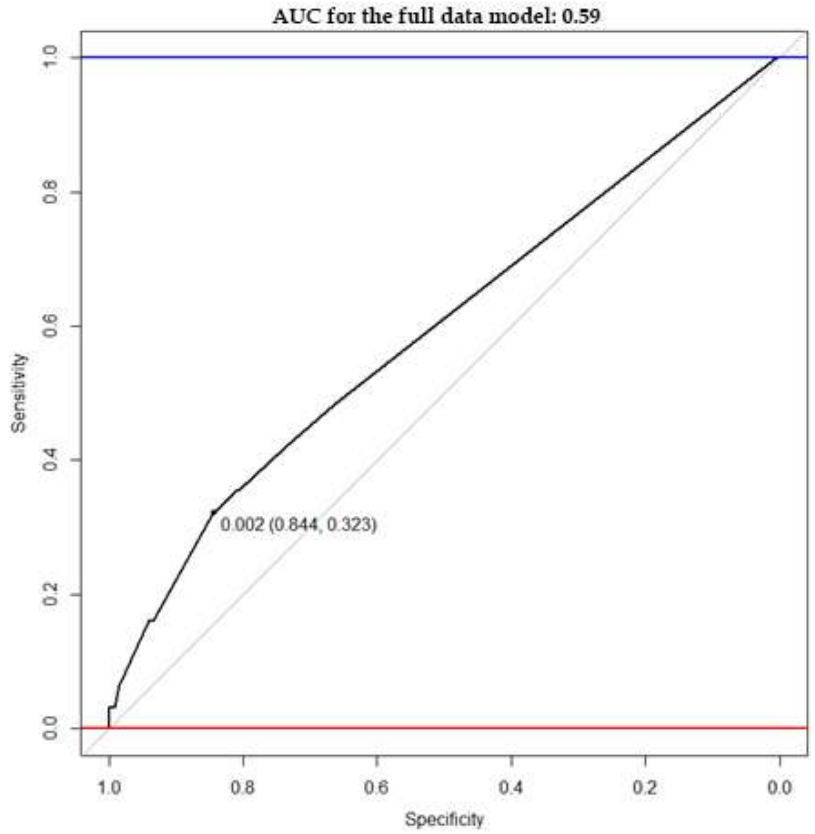


Figure 16: The ROC-curve and the corresponding Area Under the Curve for the full data model

K-Fold cross validation

During the K-Fold cross validation the confusion matrix helps to determine the failures, based on the balanced accuracy, accuracy and kappa of the model. It was found that the probability of failures was under estimated in this method (0 out 31). Because failure was never predicted, kappa is 0, and the balanced accuracy is 0.500. The balanced accuracy of the model indicates that the model has no predictive power at all. This outcome is in line with the ROC / AUC results.

Table 15: Confusion matrix of the full data model			
		Reference model	
		No event	Event
Predicted model	No event	21458	31
	Event	0	0
		$specificity = 1.0000$	$sensitivity = 0.0000$

6.2 SUB-CONCLUSION I

Before failures from excavation works can be predicted by a model with independent variables concerning spatial interdependencies and other important factors (main research question), the variables that contribute to these failures must be identified. Although many variables were found as related in earlier studies, logistic regression is useful to gain another view on the variables that are related to excavation damage. Hence, variables that could be relevant for network operators to estimate the probability of failure from other's excavation works were identified in three ways. First, a literature study identified the well-known variables. Second, expert interviews added some more relevant variables. Third, all single variables were combined into an all-encompassing model to determine the contribution and significance of all individual (independent) variables to the probability. By testing the significance and assessing a stepwise elimination based on AIC, the most related variables were found. These finding enables answering sub-question one: *What variables are most related to cable and pipe failure from excavation works?*

Multiple variables were found related to excavation damage by all methods: literature, expert interviews and logistic regression. These are included in this paragraph. First, the type of network where the excavation work is conducted for, earlier indicated as 'party'. Second, a smaller horizontal distance between cables and pipes increases the probability of failure. Third, diameters were found as related in all three methods, but the logistic regression showed some contradictions with the other methods. On one hand, logistic regression showed that some diameters have a negative estimate and others a positive estimate. On other hand, literature and interviews indicated that larger diameters increase the probability of failure. Last, if the correctness of the data reduces, the probability of failure increases.

Some variables were only implied by two of the methods. First, the process of KLIC which was mentioned in general during the interviews and was also identified by the logistic regression. The logistic regression indicates a strong increase of probability of failure in case of an emergency KLIC request. Second the material of the own assets, which were not significant in the logistic regression, were named as relevant in both literature and interviews. Third, excavation activities on cables and pipes located between the own asset and houses increase the probability of failure enormously.

Some variables found in literature and interviews were not included in the logistic regression due to unavailability, absence or other complications. On one hand, if cables and pipes are located deeper, the probability of failure from excavation works increase for the other networks, but accurate information about the vertical position is unavailable. On the other hand, whether a trial trench is made, so whether the process is followed by the rules or not is unknown and is therefore not included.

Altogether, multiple variables affect the probability of failure from excavation works. All variables that are relevant are summed in Table 16, where it is shown from what method the variable followed. In the table, the Non-Available (NA) data refers to data that was not available for the logistic regression model. From the table it follows that the pipes' wall material should be added to the logistic regression model.

Table 16: All relevant variables that were found from the literature study, expert interviews or logistic regression

	Literature	Interviews	Logistic Regression
Excavating party	X	X	X
Depth (vertical position)	X	X	NA
Wall material	X	X	
Horizontal distance between cables and pipes	X	X	X
Diameter of cables and pipes	X	X	X
Correctness of the data	X	X	X
Emergency KLIC request		X	X
Trial trenches		X	NA
Excavating on the house or street side.		X	X
Type of work	X	X	NA

It can be concluded that all relevant and available independent variables following from the three approaches, results in a logistic regression model that is like equation [12]. The fitted model is

$$P(Y) = \frac{1}{1 + e^{-(Z)}}, \qquad [12]$$

with: $Z = b_0 + b_1\, klic_{notification_{type}} + b_2 evi_{diameter} + b_3\, evi_{data_{quality}} + b_4\, sew_{diamter} + b_5\, gas_{distance} + b_6\, klic_{party} + b_7\, heat_{side} + b_8\, elec_{side} + b_9\, tele_{side} + b_{10}\, Evi_{material} + \varepsilon_i$

where $\varepsilon \sim N(0,1)$.

Further research should be conducted on the variables that seemed to be important from the literature study and expert interviews. Variables as soil type and vertical position were not included as these data were not accurate enough or unavailable.

Furthermore, from the validation of the full data logistic regression model it followed that the balanced accuracy of the model is 0.50, which has not accurate at all. Sub-question two concerns the improvement of the (balanced) accuracy in order to accurately predict the probability of failure.

6.3 RARE EVENT DATA

6.3.1 METHODOLOGY

From the previous section it followed that the full data model has no predictive power at all, i.e. the balanced accuracy was 0.50. This result from the full data model proved that the imbalanced dependent variable is not supported by generic logistic regression. Because of the second sub-question, some alternatives were tested in order to possibly increase the balanced accuracy.

First, King and Zeng (2001) studied rare events data in politics, on topics such as war, coups and uncommon disease infections. Rare event data are binary dependent variables with a positive outcome that is tens to even thousands of times smaller than the negative (0) outcome (King & Zeng, 2001). On the one hand, King and Zeng developed a widely applied method on how to correct for the underestimated event probabilities. On the other hand, a far more efficient way of data collection is proposed, as events (1) are much more informative than the non-events (0).

6.3.1.1 WEIGHTING

Both prior correction and weighting were developed to correct the model for the underestimation. Prior correction is applied on finite datasets whereas weighting can be applied on both, finite and infinite population datasets. In this thesis, excavation- requests and damages are considered infinite, as it will not stop after tomorrow and it is (almost) impossible to include all cables and pipes worldwide. Because of the infinite data, only weighting will be applied (and discussed) in this study.

Weighting is relatively simple as it uses the weighted exogenous sampling maximum-likelihood estimator. Herein, the weighted log-likelihood is maximized instead of the 'normal' log-likelihood as in 'normal' logistic regression (equation [5]). In equation [13], the weight $\boldsymbol{\omega_i}$ is bold to emphasize the difference with the regular log-likelihood equation.

$$ln\ L(\beta|y) = -\sum_{i=1}^{n} \boldsymbol{\omega_i} \ln(1 + e^{(1-2Y_i)X_i\beta}) \qquad [13]$$

With equation [14], the weights $\boldsymbol{\omega_i}$ can be determined by

$$\boldsymbol{\omega_i} = \omega_1 Y_i + \omega_o (1 - Y_i) \qquad [14]$$

where $\omega_1 = \frac{\tau}{\bar{y}}$ and $\omega_0 = \frac{(1-\tau)}{(1-\bar{y})}$, with τ as population fraction and $\bar{y}$ as the sample fraction (King & Zeng, 2001). The population fraction is calculated by the number of failures divided by all available data. On the other hand, the sample fraction is the number of included failures divided by the entire sample size.

6.3.1.2 SAMPLE SELECTION

One of the parameters used to determine the weights in the weighted log-likelihood is the sample fraction. It already implies that only a fraction of all samples is used. King and Zeng (2001) propose endogenous stratified sampling on the rare event (failure) data to ensure that these are split fairly. On the other hand, a different strategy for the majority is suggested.

If possible, it is recommended to generate an "equal shares sampling design (i.e., $\bar{y} = 0.5$)" (King & Zeng, 2001, p. 143). As equal shares are clearly not realistic for the data in this thesis, the first step is to collect all failure events' available. For non-failures this is definitely not the situation "since the marginal contribution to the explanatory variables' information content for each additional zero starts to drop as the number of zeros passes the number of ones, we will not often want to collect more than (roughly) two to five times more zeros than ones" (King & Zeng, 2001, p. 143).

Considering the sampling rule, the sample set would change as shown in Figure 18. The non-failures included in the sample set are randomly selected from all non-failures. The various ratios of non-failures compared to failures are all tested during modeling to find the best failure – non-failure ratio as recommended by King and Zeng (2001).

6.3.1.3 SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE

A second approach to overcome the imbalanced data, is synthetic minority over-sampling (SMOTE) (Chawla et al., 2002). A dataset is imbalanced if the classes are not equally represented, which is basically the same as rare event data. The smallest class is the minority, whereas the normal/largest class is the majority. In oversampling, the minority class is increased to balance the dataset. In under sampling the majority class is decreased with a similar goal as over sampling.

Chawla et al. (2002) suggest over-sampling of the minority with "synthetic" examples instead of over-sampling with replacement. The synthetic samples are generated "along the line segments joining any/all of the k minority class nearest neighbors" (Chawla et al., 2002, p. 328). The amount of over-sampling determines what neighbors from the k nearest neighbors are chosen. The difference between the sample and the nearest neighbor is multiplied by a random number between 0 and 1. In this way a random point within the correct region is selected, which enlarges the minority class (Chawla et al., 2002).

Whereas this section is about over-sampling of 1s and under sampling of 0s, one could consider the approach of King and Zeng (2001) as under-sampling and algorithm adjustment (Haixiang et al., 2017). Chawla et al. (2002) suggest to combine SMOTE and under-sampling, where the under-sampling is similar as in the previous section. The combination of both reverses the initial bias of the learner towards the majority class into the favor of the minority class. The use of both techniques could improve the classification of data (Chawla et al., 2002). Herewith this respect, classification refers to the confusion matrix and area under curve which were described in section 5.3.

A disadvantage of the SMOTE method is the incapacity to include categorical variables, because of the synthetic generating. For example, a synthetic generated datapoint could be made of material between PVC and steel, which naturally does not exist.

Figure 17: SMOTE: the five solidly filled points are the 'real' data. In between all data points, new 'synthetic' points are created on a random distance between the two 'real' points (Khurkhuriya, 2018)

6.3.2 WEIGHTING

In this section, weights are added to the samples and under sampling is applied to compensate the imbalanced dataset.

6.3.2.1 MODEL SELECTION

SAMPLE SIZE

By the proposed method of King and Zeng (2001), the sample set has been compiled. All suggested ratios that are integer numbers were tested (2, 3, 4 and 5 times). The possible ratios between failures and non-failures are shown in Figure 18. To determine what the optimal sample size is considering the AUC and balanced accuracy the goodness of fit is tested, and the four different models are validated. In the end, for each unique study it can be decided what tests are more important than others, but the considerations made in this thesis will be explained.

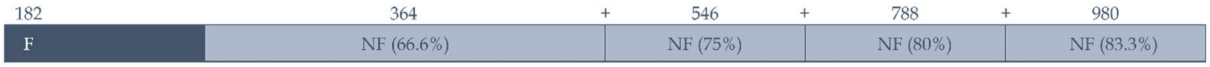


Figure 18: The sample set when adjusting the non-failure failure ratio from two to five times as suggested by King and Zeng (2001).

It is important to realize that apart from the non-failure failure ratio, the split ratio between the train and test set is also relevant. Here the training and test set contain 0.8 versus 0.2 of the entire sample set. It should be considered because the size of the test set could become very small and therefore inconsistent when tested multiple times.

WEIGHT

Weighting has been done in two ways. First, the software package *Zelig* for R, developed by King and Zeng (2001), has been used. In the package, the weighting and bias correction are performed automatically. The package is available for the statistical software Stata and R. A disadvantage of the package is that most standard statistical tests, such as the Wald-statistic, log-likelihood ratio and the coefficient of multiple determination cannot be applied to models developed within the package. Therefore, only the tests that were also included by King and Zeng (2001) could be performed.

Second, the equation to calculate the weight is known and given in equation [13]. Using this thesis' case to determine the weight gives

$$\omega_i = \frac{0.00169}{(x+1)^{-1}} Y_i + \frac{(0.998)}{(1-(x+1)^{-1})} (1 - Y_i) \quad with\ 2 < x < 5$$

where x represents the ratio zeros to ones in the sample. If x is given, the weights for the event and non-event are shown as an average of three trials in Table 17. The AUC and balanced accuracy are also included as of the second sub-question aims to accurately predict failures. Therefore the most interesting in Table 17 is a high balanced accuracy.

Table 17: The calculated weights following from the non-failure / failure ratio and the corresponding AUC and balanced accuracy

Ratio non-event / event	**Weight**		**AUC**	**Balanced accuracy**
	Event (1)	Non-event (0)		
2	0.00508	1.497	0.691	0,651
3	0.00678	1.331	0.685	0.637
4	0.00847	1.248	0.758	0.658
5	0.010	1.198	0.655	0.596

For the readability of the thesis, it is important to know that the results following from the package and the manually calculated weight are very different. Therefore, both are described and will be compared in the end.

6.3.2.2 GOODNESS OF FIT

WALD

From the test with the manually calculated weight, some very remarkable results followed. None of the variables seemed to have a p-value coming near the significance level. The unknown type of work, which refers to the emergency KLIC request has both, the lowest p-value and the highest z-score with respectively 0.772 and 0.290. The own material of the considered network has (almost) the same scores. The results are so remarkable and contradicting with all other models that it is assumed something is not right. For instance, the lowest (or best) p-value is 0.76.

Using the package to test the individual contribution of variables gave results that seemed to be more realistic. In Table 18 the p-values and z-scores of the sample set with four times more non-events than events are shown. Varying the ratio event versus non-event does adjust the values a little, but not enough to include all in this report. The p-value moves toward zero when more samples are included, whereas the z-score does the opposite.

Table 18: The p-values, estimates and z-scores of the weighted model, with four times more non-failures than failures.

	Estimate	Z value	Pr(>|z|)
Klic_party Gas	1.23	1.66	0.097
Klic_party Heat	1.51	1.84	0.066
Klic_type_work cables and pipes	0.98	1.68	0.093
Klic_type_work Unknown	2.38	3.93	8.44E-5
evi_materialHPE	640000	805.430	2.00E-16
evi_materialONB	3836000	3851000	2.00E-16
evi_diameter	-0.0096	-4.82	1.44E-06
Evi_data_quality	2.29	2.17	0.03
Heat_side1	-0.83	-1.65	0.09
heat_side2	-0.84	-2.91	0.004
elec_side1	0.48	1.97	0.049
elec_side2	0.97	2.09	0.036
tele_side1	-0.60	-2.29	0.022
Tele_side2	-1.09	-1.95	0.052

On the one hand, a major disadvantage of the rare event logistic regression model package is the absence of multiple tests, especially for goodness of fit. On the other hand, the developed model is not supported by any external goodness of fit tests. Therefore, the only model that was tested on goodness of fit was the manual weighted model from which the results only the results from the manually conducted weighting are elaborated below.

LOG-LIKELIHOOD RATIO

In line with the results from the Wald-statistic and p-value, the log likelihood ratio has remarkable values as well. After taking out the most significant variables (type of work and own diameter), the model fit increased one out of one, indicating that the variables in the model do not improve the model anything compared to the model with only an intercept.

COEFFICIENT OF MULTIPLE DETERMINATION (R^2)

Again, the unusual results found from the individual tests do already indicate that the goodness of fit will not be good either. McFaddan's Pseudo R^2 resulted in the very small 0.001429.

On the other hand, the AIC score of the (manually weighted) model is only 59.23, where the AIC of the first model from the previous section (6.1.1.3) was 40 times larger (AIC full data model = 2511). Once again, this is determined by the much smaller sample size as compared to the data for the first model.

6.3.2.3 MODEL VALIDATION

Validation could be done on both the manually weighted model as well as the automatic weighted model. Both results are included.

Remarkable enough, during validation of the manually weighted model, the AUC was 0.71. This is contradicting with the goodness of fit that was found very poor and shows that a poor model fit is not an indication of model predictive performance and vice versa. Whereas model fit indicates how much of the data is explained by the model, the predictive performance indicates the how well the model can predict the outcome.

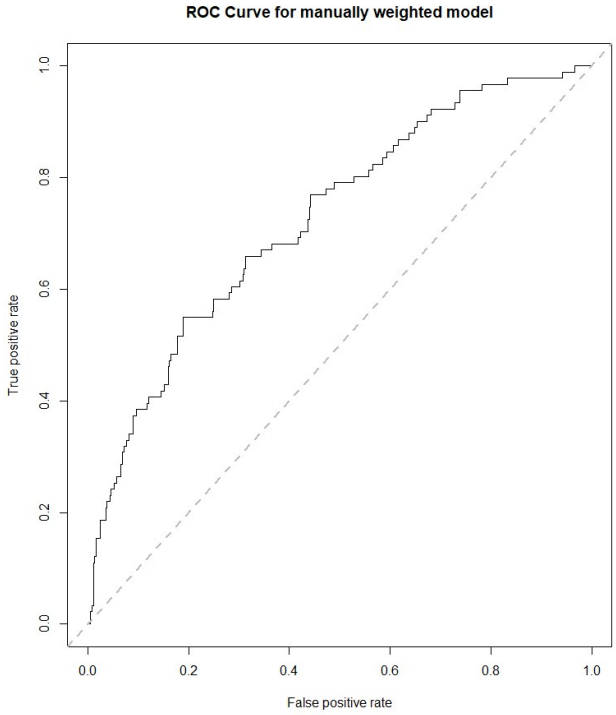


Figure 19: The ROC curve of the weighted model. The AUC is 0.71

REPEATED K-FOLD CROSS VALIDATION

Like *Results I* (with the entire dataset), the model with the manually created weight did not predict any failure. Therefore, kappa and the balanced accuracy are respectively zero and 0.50. As this is not very interesting because it does not predict anything, only the confusion matrix (Table 19) of the 'package' model is elaborated on.

Because of the weighting, the confusion matrix is affected, probably in the desired way. In Figure 20 (left), the results of the weighting by the automated software are shown. Before the weighting all samples were in the red box (100%). Through the weights, 29 percent moved from true negative to other positions since the failures are considered more important by the model. Therefore, more often failures will be predicted than without weighting. Figure 20 (right) illustrates the change from the 'old' expected values ($E(Y|X)$) into the new (blue) ones ($E(Y|X1)$).

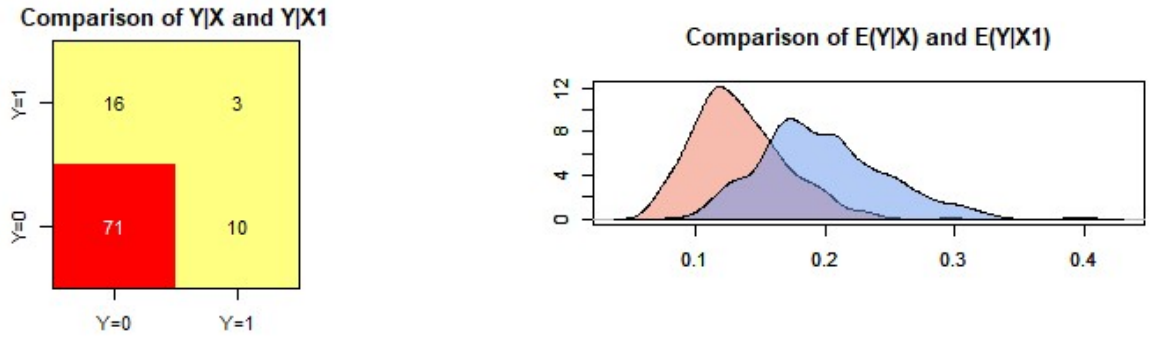


Figure 20: The result of weighting the model. Left: before the weighting 100% was in the red box. Because of the weight 29% moved from true negative to the other positions. Right: the expected value from Y given X from the unweighted model (red graph) and the weighted model (blue graph)

In the confusion matrix, the sensitivity and specificity can be found and are respectively 0.38 and 0.94. The balanced accuracy has a score of 0.66, which is the best score of all models.

Table 19: Confusion matrix of the weighted model			
		Reference model	
		No event	Event
Predicted model	No event	136	23
	Event	9	14
		$specificity = 0.9379$	$sensitivity = 0.3784$
Accuracy	0.714		
Kappa	0.368		
Balanced acc.	0.658		

6.3.3 SMOTE

6.3.3.1 MODEL SELECTION

SAMPLE SIZE

As explained in an earlier section, the dataset will be adjusted by over- and under sampling. Hereby, it is important to realize that one should be careful not to oversample too much synthetic, non-real data. Furthermore, the ratio non-event versus event should not flip over as this would be opposite to the real situation. Therefore, the non-failure / failure ratio should be at least one and this is also recommended by Chawla et al. (2002). In Table 20, the ratio of the sample set is shown for different combinations (%) of over- and under sampling. For example, when considering a 100 percent under sampling and 100 percent over sampling, one obtains a ratio of 2, meaning twice as many non-failures than failures included in the sample set.

Table 20: The non-failure / failure ratio of the sample set for different over- and under-sampling percentages

	Over sampling [%]				
Under sampling [%]	**0**	**50**	**100**	**200**	**300**
0					
50		5,67	4,00	3,00	2,67
100		3,00	2,00	1,50	1,33
150		2,00	1,33	**1,00**	1,13
200		1,50	**1,00**	0,75	0,67
250		1,22	1,27	0,60	0,54
300		**1,00**	0,67	0,50	0,45

For the various ratios the model has been tested on the ROC and corresponding AUC. The ROC and AUC depend, of course, on the sampled data set. Different samples hence provide different results. Therefore, the average AUC of five samples for every over/under sample percentage has been chosen. Considering the previous example (100% over- and under sampling), it would follow that the AUC is 0,68.

Table 21: The Area Under Curve for the various over and under-sampling percentages

	Over sampling [%]				
Under sampling [%]	**0**	**50**	**100**	**200**	**300**
0	0,58				0,65
50		0,65	0,63	0,65	0,66
100		0,62	0,68	0,68	0,66
150		0,66	0,68	**0,70**	0,69
200		0,68	**0,70**	**0,70**	0,69
250	0,58				0,65
300		0,65	0,63	0,65	0,66

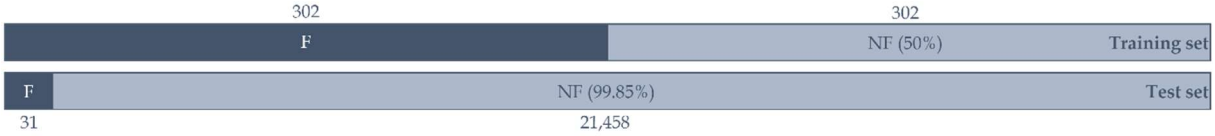


Figure 21: From the tables above, 200% under sampling and 100% over sampling were selected. From the sampling a training set with equally sized failure and non-failure follows, whereas the test set remains it original size (20% of all data)

From the table a notable increase of the Area Under the Curve follows, compared to the results when using the entire dataset. Considering the Area Under the Curve in Table 21 and the ratio between non-events and events, it has been decided to use 100% oversampling and 200% under sampling. To validate the model's performance based on the rare event sampling, we also perform a validation analysis. As usual, we choose the training set to account for 80% of the dataset, which results in 302 failures and 302 non-failures. The exceptional quality of SMOTE is that the test set remains unmodified. Only the training set is adjusted to train the model better before it needs to be validated. Figure 21 shows the sizes of the sample sets.

6.3.3.2 GOODNESS OF FIT

WALD

When testing the individual contribution, again the Wald test is conducted. Both, the z-score and the p-value are assessed. The first aspect that should be noticed is the absence of the categorical variables. This is because the Synthetic Minority Oversampling Technique cannot include categorical variables.

Table 22: The p-values, estimates and z-scores of the SMOTE model

Variable	Estimate	z value	Pr(>|z|)
(Intercept)	1.58	4.99	0.00
evi_diameter	-0.01	-4.59	0.00
evi_data_quality	2.08	2.35	0.02
gas_distance	-0.07	-2.57	0.01
gas_side	0.39	2.46	0.01
heat_side	-0.31	-2.59	0.01
elec_side	0.32	2.15	0.03
tele_side	-0.61	-3.41	0.00

The AIC score of this model (AIC=795.81) is already way better than the earlier tested alternatives. This is, however, due to the smaller number of variables. Comparing the scores shows that the AIC is more than 2.6 times smaller than in the first model.

LOG-LIKELIHOOD RATIO

Just like the AIC-score, the Log-likelihood ratio of the SMOTE model has been improved a lot compared to the original model. Where it was four times above the significance level, Chi square became very small in the SMOTE model ($6.178*10^{-11}$). This Chi square follows from the difference between all variables and selected variables (diameter of the own asset and the side of the telecom cables).

COEFFICIENT OF MULTIPLE DETERMINATION (R^2)

From McFaddan's R squared test, it follows that the total model fit is worse than in the full data model ($R^2 = 0.092$). In this model, which is over- and under sampled, R squared is only 0.069. Nonetheless, the overall change is small.

6.3.3.3 MODEL VALIDATION

ROC AND AUC

As the samples are different every time, every ROC test will have a different result. Therefore, the test has been conducted five times to have an indication of the mean. From the five trials, the difference between the largest and smallest AUCs was 0.07. In the end, the ROC curve in Figure 22 was average result of the five trials and is therefore selected as the final ROC for the SMOTE model. This results in an AUC of 0.74 when the minority is over samples a 100% and the majority is under sampled at 200%.

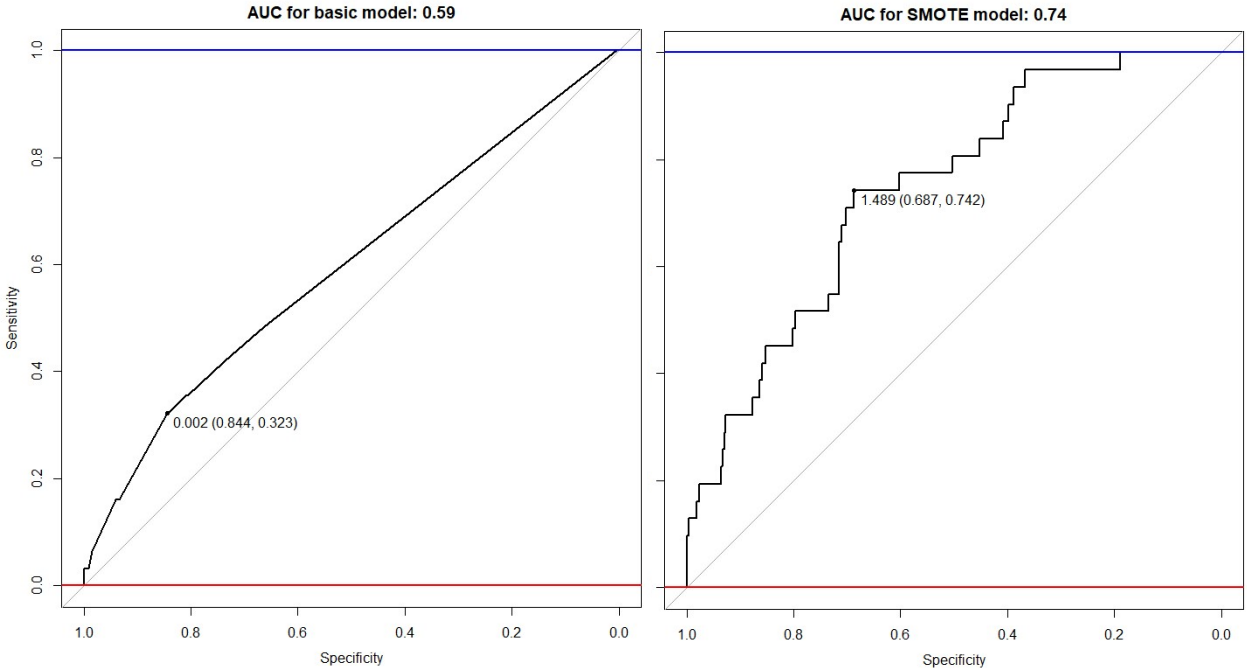


Figure 22: The ROC and AUC of the SMOTE model compared to the full data model (basic model), when the failures are oversampled 100% and the non-failures are under sampled 200%.

REPEATED K-FOLD CROSS VALIDATION

As the ROC curve already showed a major improvement for the new model, it is expected that the repeated K-fold cross validation will also show improvement. When 80% of the dataset is in the training group, this validation method indicates a balanced accuracy of 0.58 for the SMOTE model. When the test and training group have equal sizes, the model results indicate better predictable performance of 0.605. Underneath, the confusion matrix for the model with the data equally split over the training and test group is given (Table 23).

Table 23: Confusion matrix of the SMOTE model

		Reference model	
		No event	Event
Predicted model	No event	33,905	38
	Event	19,729	52
		$specificity = 0.632$	$sensitivity = 0.578$

Accuracy	0.743
Kappa	0.0019
Balanced acc.	0.605

An important note is that altering the ratio of the training/test group sizes also influences other tests. For example, the Area Under Curve is better performing (closer to 1) at the 80/20 ratio than at the 50/50 ratio of the training/test set.

6.3.4 ALTERNATIVE MODELS

It has been over thought to combine synthetic minority over sampling and weighting. However, SMOTE tries to equally balance the failure / non-failure ratio already during the sampling. Therefore, the sample set is already balanced after the synthetic minority oversampling and under sampling. On the other hand, weighting is a technique applied to compensate for imbalanced datasets, preferably on a sample set with a majority class two to five times the size of the minority (failure) class. So, the former does balance the sample set completely, where the latter is a technique to compensate for imbalanced sample sets.

Furthermore, Bayesian logistic regression has been tested on the entire dataset (as in results I). It was found that there was noteworthy difference between the results of normal logistic regression and Bayesian logistic regression. Therefore, the results are not included in this thesis.

6.4 FINDINGS

Basically, the results of five methods can be distinguished; the basic logistic regression model, the SMOTE model, the manually weighted model, the automatically weighted model and the Bayesian logistic regression model. Taking the basic model from section [5.2] (results I) as starting point creates the opportunity to compare the five models on all the tests that were conducted. Comparing the results supports decision making on what model should be used.

In Table 24, the results of all models are compared for each test that was conducted. Besides the score, also a ranking was added to create a better overview. A double plus or min indicates the best or the worst performing model on that test. A single plus or min indicates whether a model is performing better or worse than the basic model. It is important to realize that all models contained the same variables in the beginning, namely the variables found in the model selection. Using the same variables is essential to compare the models.

Looking at the P-values of the individual variables, SMOTE and the automatically weighted model perform very well, whereas the manually weighted model had strange results with very high p-values. Likewise, the other results of the manually weighted model show aberrant results. On all goodness of fit tests, it scores worst. However, it has an anomalous low AIC score. The equal scores on sensitivity and specificity is clarified by the fact that the 'basic model' does not have any predictive power either.

The missing predictive power was also found in the Bayesian model. Additionally, the other scores of the Bayesian model do also show great similarities with the basic model. Therefore, the detailed results were also left out the earlier results.

Two models perform better than the basic model on most aspects; the SMOTE model and the automatic weighted model. An earlier described disadvantage of the automatic weighting is the disability to perform goodness of fit tests on the model, whereby it becomes more complicated to compare it to other models. However, on the two good performing models also other tests were conducted which can be compared.

During validation, some differences were found. First, the sensitivity of the automatic weighted model is much better than that of the SMOTE model. In contrast, SMOTE performs much better on the specificity than the weighted model. All validation together, the balanced accuracy of the weighted model is closer to one (0.66) than the balanced accuracy of the SMOTE model (0.61).

Table 24: The five assessed alternatives compared, where the basic model is considered as the basic measure point. The models are ranked on the performance per test. Above the dotted line are standard and goodness of fit tests, underneath is validation

		SMOTE		Manually weighting		Automatic weighting		Bayesian Log.Reg.	
Test	**Basic model**	Score	Rank	Score	Rank	Score	Rank	Score	Rank
Average P-values	0.08	0.02	++	>0.77	--	0.04	+	0.28	-
Average |z-score|	|2.64|	|3.12|	++	|0.10|	--	NA		|1.85|	-
LLR (Chi squared)	0.40	6E-11	++	1	--	NA		0.37	+
Coefficient of determination	0.092	0.069	-	0.0014	--	NA		0.0093	=
AIC	2070	795	+	59	++	434	+	1950	=
AUC of ROC	0.60	0.74	++	0,71	+	0.70	+	0.72	+
Specificity	1	0.63	--	1	=	0.94	-	1	=
Sensitivity	0	0.58	++	0	=	0.38	+	0	=
Balanced accuracy	0.500	0.61	+	0.500	=	0.66	++	0.500	=

Ranking clarified:
++ Best model on the test
+ Model scoring better than basic model
= Approximately equal to basic model
- Model scoring worse than basic model
-- Worst model on the test

6.5 SUB-CONCLUSION II

The first sub-question identified the related variables affecting the cables and pipe failures from excavation works. As it was found that the full data model with the identified variables had a balanced accuracy of 0.5000 which indicates that this model cannot predict (accurate) at all, a new literature study followed to find a way to answer sub-question two: *To what extent are the identified variables affecting the probability of failure from excavation works and how can we accurately predict the probability of failure?*

Even though the full data model cannot predict accurate at all, it is very informative about the extent to which spatial interdependencies affect the probability of failure. Based on the goodness of model fit tests log-likelihood-ratio which is far above significance level (p=0.40) and the coefficient of multiple determination (R^2=0.09) it is clear that the model with the independent variables included affect the probability of failure in a large extent. Especially the contribution of the diameter of the own network, the side of the telecom cables and emergency KLIC-requests had large Wald-statistics and very low p-values whereby these three variables affect the probability to the largest extent of all identified variables.

However, it was found that the regular logistic regression model is not able to predict accurately when data is imbalanced. From literature, two effective methods were identified to deal with the rare event data for predictive modelling. First, the synthetic minority over sampling technique (SMOTE) was applied to balance the dependent variable in the trainings sample set. Second, weighting was tested to correct for the underestimating model. Both methods have shortcomings, the former is not able to handle categorical independent variables, the goodness of fit of the latter cannot be tested. Despite the deficiencies, both methods performed much better predicting (higher accuracy) than the full data model without modifications. Therefore, both methods were considered to answer the second sub-question.

First, four times more non-failures than failures was found to have the highest accuracy for the weighted and under sampled model. The sensitivity was 0.38 (failures) and the specificity 0.94 (non-failures),

resulting in a balanced accuracy of 0.66. So, the sensitivity is much better (+0.38), whereas the specificity decreases just a little (-0.06) compared to the full data model.

Second, from the sensitivity analysis the most accurate SMOTE model followed. It was able to predict 58% of the failures from excavation works (sensitivity), whereas it predicts 63% of the non-failures accurate (specificity). This results in a balance accuracy of 0.61, which is 0.11 more accurate than the full data model. Also, the regular accuracy is higher than the full data model (0.74 vs 0.59).

Concluding, both methods predict more accurate than the full data model. The SMOTE model predicts more failures accurately but also more false failures than the weighted model. Thereby, it can also be concluded that logistic regression does work to assess individual situations and therefore has a lot of potential for future developments within cable and network providers.

Whether the (balanced) accuracies are enough to implement the model or not will be discussed in the next section, but at least it is proven that the variables affect the failure rate and it is possible to predict failures accurately to a certain level.

7 APPLICATION

In the second sub-question it was found that both models, weighted (0.66) and SMOTE (0.61), predict failures more accurate than the full data model. However, as it is not 1.00 it is insufficient to accurately predict all failures. Despite the balanced accuracies, the method could still be useful for network operators. To see how the applied method from this thesis could be implemented by network operators, a gap analysis, often used in a business environment is conducted. "A gap analysis is the technique used to define the difference between the current state and the proposed state of any business and its functionalities" (Marra et al., 2018, p. 157). In this analysis, there are two questions that raise the third question. On one hand, what is the current situation (starting point)? On the other hand, what is the desired situation? The gap analysis is the process to "identify the delta between the proposed and the existing functionalities in any application" (Marra et al., 2018, p. 157). Strengths and weaknesses of the current process are highlighted to see how network operators can use the model to reduce failures from excavation works (sub-question three).

In considering strengths and weaknesses the difference between the business- and construction environment is gapped. The analysis of strengths and weaknesses in asset management is often done through a SWOT-analysis, which is an accepted method within the industry (McGrail & Roberts, 2005). A SWOT-analysis is an established method for assisting the formulation of strategy, it aims to identify the strengths and weaknesses of an organization and the opportunities and threats in the environment. Once the factors are identified, strategies will be developed which build on the strengths, eliminate the weaknesses, exploit the opportunities and counter the threats (Dyson, 2004, p. 152). Basically, the identification of strengths and weaknesses is an assessment of the current situation. The opportunities and threats can be used to fill the gap that raises from the gap analysis.

First, the current situation will be described after which the desired future situation is described according to the gap-analysis strategy. Once the desired situation is defined, the gap will be filled as good as possible by application of the SWOT-method.

7.1 CURRENT SITUATION

DATA

In 2015 about 33,000 excavation damages were registered. As shown in Table 25, about 30 large network operators serve in the Netherlands (besides sewer systems owned by the municipalities). Assuming 75% of all damages is caused on these 30 large networks, it would be around 25,000 failures. If all parties are harmed equal times, these 30 parties would all have around 800 failures from excavation per year.

All the assumptions are made as an example to illustrate the current situation. On average, network operators have access to 800 failures from excavation per year, which is only a small number compared to the total number of exaction activities (or KLIC-requests). Therefore, failures become rare event data for network operators, which hampers statistical analyses that are essential for pro-active approaches.

CROW500

As explained earlier in this thesis, several regulations and procedures are introduced to reduce the number of excavation damages on cables and pipes. The most extensive guideline was developed by the involved parties themselves and is CROW500.

Table 25: Number of registered network operators per type in the Netherlands. About 60% of the registered networks are cables/pipes serving as transport between multiple affiliates of companies (KLIC-phone, 2018)

Network type	Larger companies [#]	Comments
Water	10	
Sewer	380	All municipalities
Gas	7	
District heating	NA	
Electricity	7	
Telecom	5	
Fluids and remainders	NA	
Other (small) parties	631	
Total	1,050	*Information contained through the KLIC-phone on 18/05/2018*

CROW 500 includes tasks for every phase of the project cycle, from the initiation- to the operational phase, to avoid cables and pipes failures from excavation works. The steps that should be taken according to the guideline are (CROW-werkgroep, 2016):

- During the *initiation phase*, the initiator is responsible to provide sufficient resources (time and money) for the following project phases.
- Next, in the *concept phase* an orientation KLIC-request is done to ascertain the theoretical position and other characteristics of nearby cables and pipes. This information is the starting point for the executed risk-assessment as described in section 1.2.2. In the risk-assessment, one or multiple control measures are mapped (e.g. implement in design, move a cable or pipe, temporarily cut of a connection).
- The risk-assessment from the concept phase will be used to develop a control measurement plan during the *design phase*. To do, one of the suggested control measures from the risk-assessment is selected for every single cable or pipe. Furthermore, based on the risk-assessment locations are selected where cables and pipes should be localized through trial trenches.
- Latest, during the *work preparation phase*, the control measurement plan is developed into work instructions. The work instructions should describe what must be done to avoid damages during excavation. Together with a regular KLIC-request, the contractor should be able to excavate carefully now.

To summarize, the CROW 500 includes the early mapping of risks and the localization of cables and pipes during the design phase. The steps from the orientation KLIC-request to the work instructions are illustrated in Figure 23. In the figure the human aspect has been implemented either. All steps to the next phases are done because of human action. The responsible expert for a certain phase of the cycle must judge what (protective) measurements are desired.

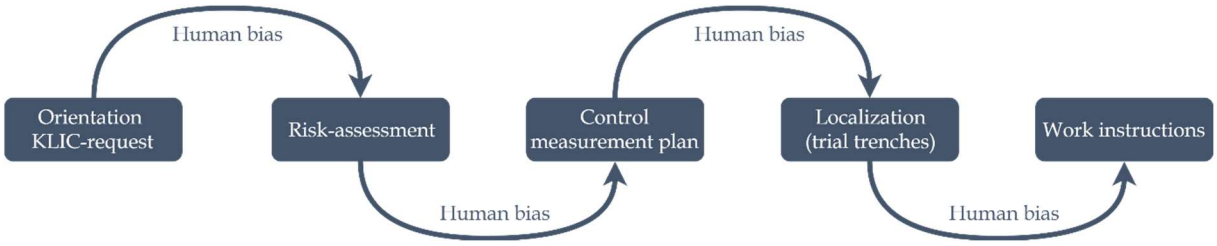


Figure 23: The CROW 500 process to prevent cables and pipes failures from excavation work as good as possible in five steps

The process of the risk-assessment as described above is for regular KLIC-requests, not for emergency KLIC-requests. This makes it plausible that the described process has been developed for rehabilitation excavation work (planned) and not for (unplanned) repairs. The dataset of the Evides case contains 5% emergency KLIC-requests, whereas this type of requests result in 24% of all failures.

WIBON AND KLIC-WIN

To reduce the failures from excavation works and to better comply to the European legislation (INSPIRE), WION was renewed (21 March 2018) and KLIC will be renewed. For the future benefit of this thesis it is assumed that KLIC-WIN has already been introduced, as this will happen within a half year. Therefore, the current application of WIBON and KLIC-WIN are considered in this section.

KLIC-WIN obliges mechanical excavators to do a KLIC-request before starting the actual work. The network operators within the concerning KLIC-polygon have to provide their data to the Kadaster where all cable and pipe data is combined into one map. The all-encompassing map is sent to the excavating party only. For the regular process, the request has to be done three to twenty days in advance of the actual work.

7.2 SWOT I

The assessment of the current situation is basically the assessment of the strengths and weaknesses of the SWOT analysis. Therefore, the strengths and weaknesses are only assessed shortly here, as it has already been described in the previous section (7.1).

7.2.1 STRENGTHS

- The by CROW500 obligated risk-assessment procedure is already very extensive.
- The WIBON and KLIC-WIN obligate a KLIC-request before excavation starts manually. Thereby, the excavator receives a map containing all the cable and pipe locations.
- Network operators possess databases already, containing detailed information on failures from excavation works.
- The KLIC-requests are (within Evides) registered and processed automatically, after which area managers check what requests have a higher failure probability. The area managers will actively go after the critical requests. So, experts assess all KLIC-requests.

7.2.2 WEAKNESSES

- There is no large data sharing program which leads to the rare event data. Rare event data are harder to analyze.
- During the process from the risk-assessment to the working instructions multiple steps are passed. All these steps are handled by in-company experts which results in a certain bias. An employee will always try to look after the interests of his company instead of other companies. Furthermore, the involvement of 4 steps increases the probability of unnecessary mistakes.
- The localization is done by infield employees, whereas the assessment of the maps is mostly done by office persons. Localization is again, based on the expert's opinion.
- A disadvantage of the current method is that only the excavating party receives the entire map. Therefore, other companies are not able to assess the full situation.
- There has never been an empirical research to study the size of the localization area. Also, spatial interdependencies are quite unfamiliar in the sector. However, the fact that the localization areas include a whole meter for the margin of data error speaks for the incorrectness of the data.
- In 20% of all KLIC-requests a cable or pipe failure occurs from the excavation works, which is obviously too much.

7.3 FUTURE SITUATION

Of course, a world without any failures from excavation works is the most desired situation. However, in the gap analysis it is more useful to state more realistic and specified goals as it becomes easier to perform the SWOT-analysis in more detail. Therefore, the desired situation has been split in similar topics as the current situation.

DATA

From the former sub-questions, it followed that data correctness influences the probability of failure from excavation works to a large extent. Therefore, it is desired that the data is very accurate (i.e. correct). In this thesis' case, more than 94.5% of all compared data (Rotterdam 3D and Evides) were located within 40 centimeters from each other. However, this is virtual data compared with virtual data. CROW500 prescribes a localization area of at least 1.50 horizontal meters, from which two third is margin for data error. In the desired situation, the data error has been reduced as far as possible. Furthermore, in the desired situations some other variables become available accurately. On one hand, the vertical position (i.e. depth) of the cables and pipes is known more accurate. On the other hand, collecting data on the positions where cables and pipes are localized through trial trenches has been started and is included in the model.

Furthermore, a shared failure database which contains all the failure data from multiple network operators should exist. The shared failure database creates the opportunity to do very large data analysis and decrease the rareness of the failure data.

CROW500

Just like the rehabilitation strategies of network operators, it is desired that the excavation strategy becomes pro-active either. This is supported by predictive analysis which also supports the reduction of the human bias as desired. The desired data collection of localization positions has contributed to develop a predicting program which tells what locations should be localized before the real excavation can start.

WIBON AND KLIC-WIN

The network operators communicate better and share data. Now, only the excavator receives the cable and pipe locations by the Kadaster. In contrast, in the future situation all network operators will receive the information as it enables them to analyze the excavation works more extensively. It enables all network operators to do a risk-assessment.

Also, it is preferable that the minimum waiting time of 3 days after a regular request is shortened. Therefore, the system should become more automated which could reduce the waiting time.

7.4 SWOT II

The developed predictive model and the newly found variables create opportunities to fill the gap to the desired situation. However, the method also has some threats that could work opposite. The results from the first two sub-questions of this thesis are considered as external input, whereby it creates the potential to use it as input for the opportunities and threats of the SWOT-analysis.

7.4.1 OPPORTUNITIES

RELEVANT VARIABLES

- The newly found variables relevant for failures from excavation work can be examined further. It can help to improve the model and find new fields of research.
- The results from sub-question one and two help network operators to move their focus to certain aspects
- With the knowledge from the results, procedures within companies can be adjusted to reduce the probability of failure from excavation works further.

MODEL

- Instead of the risk-assessment by experts only, it becomes possible to do the risk-assessment with both, the model and an expert opinion. This could result in a more consistent risk-assessment. The further application of the model can also help to improve its predictable power by either further modeling or artificial intelligence.
- Experts can be trained by the model, to improve their risk-assessment skills.
- Because the model can filter low risk excavation activities out, the workload will reduce for network operators.

METHOD

- Binary outcomes work also for cable and pipe networks to assess individual and specific situations. In this thesis logistic regression was the applied method, but for example, Cox proportional hazard which is used for time dependent variables could also be examined.
- The sampling techniques proved that it should not be limiting when data contain a rare event.

OTHERS

- Through technological innovation it becomes possible to improve the data correctness which is an important aspect considering failure from excavation. For example, the Autonomous Inspection Robot (AIR) which is under development for the water distribution sector. It includes an GPS, which can help to improve the data correctness (e.g. location) (Thienen, van Maks, & Yntema, 2016).
- The almost 33,000 excavation failures per year in the Netherlands alone would be a great source of information for the model. There is a great opportunity in data sharing among different network operators.

7.4.2 THREATS

There are also some threats that could work against the proposed method.

RELEVANT VARIABLES

- It can be that some variables are not identified because of shortcomings or unavailability of the data. These variables could fall into oblivion.

MODEL

- The developed model is fully based on Evides data and specifications. It could be that the method is only successful for Evides and that all companies are facing very different datasets and characteristics.
- The correctness of the data is uncertain and therefore the predictable power and the quality of the model could be questionable.

METHOD

- The proposed method from this thesis (binary logistic regression) is certainly not the only method suitable for binary outcomes. It could be that other methods are able to predict failures better.

OTHERS

- Network operators are not willing to share data, whereby the entire model should not work. Or KLIC is not willing to share maps with all network providers, but only with the requestor.

7.5 TOWS

In the TOWS Matrix the various factors are identified, and these are then paired with the intention of stimulating a new strategic initiative (Dyson, 2004, p. 152). It is proposed as a conceptual framework for a systematic analysis that facilitates matching the external threats and opportunities with the internal weaknesses and strengths of the organization (Weihrich, 1982). In Table 26 the results of the SWOT-analysis and the TOWS-analysis are summarized. A more elaborated explanation of the strategies that followed from the TOWS can be found below the table.

Table 26: SWOT and TOWS-analyses

	(Internal) Strengths	**(Internal) Weaknesses**
	o Extensive risk-assessment procedure o Detailed failure databases o Excavation damage is rare o Expert judgement for every single request	o No failure data sharing o Human bias and multiple steps in risk-assessment o Localization based on expert judgement o Detailed map only for excavating party o Quality of the data o 20% of all KLIC-requests lead to failure o Unavailable data (e.g. vertical location)
(External) Opportunities	**SO (Maximize)**	**WO (Compensating)**
o New variables found as relevant o Model to predict failure from excavation works o Analyses with binary outcomes o Expert training o Sampling techniques for rare event data	o Include new variables in risk-assessment o Model + expert judgement for risk-assessment o Train experts to select high probability situations o Explore possibilities with binary outcome data	o Innovation to improve data quality and collect new data (e.g. vertical location) o Reduce human bias with predictive model o Implement localization areas in the model o Use model to determine what KLIC-requests should be assessed
(External) Threats	**ST (Compensating)**	**WT (Minimize)**
o Predictive power of the model o Missing variables o Data quality used during development o Other modelling methods work better o Network operators are not willing to cooperate o Very different data within each company	o Use expert judgement and model together o Emphasize the benefits of cooperation during the procedure o Compare net. operators' failure databases	o Predictive power is worthless because of the data quality o Network operators their experts' interests

7.5.1 STRENGTHS AND OPPORTUNITIES

In this section strategies to exploit the opportunities by using the existing strengths are explained. First, instead of an extensive risk-assessment procedure based on expert judgements only, some other alternatives become available because of the predictive model. Either, the expert judgements can be supported by the model or the entire risk-assessment will be done by the model. Support the expert judgements with the model can be done in two ways. On the one hand, it could become the secondary

assessment of the risk, so it becomes one of the steps the experts consider. On the other hand, when the model has been improved it could be used to train the experts to improve their assessment skills.

Second, the model proves that it is possible to use logistic regression for network operators' rare event data. Other binary methods, such as Cox proportional hazard, can be explored on other failure databases to assess specific situations.

7.5.2 WEAKNESSESS AND OPPORTUNITIES

The opportunities are used to avoid or reduce the weaknesses. First, the smaller the margin for data error and the higher the related data correctness, the lower the probability of failure from excavation works. An improvement of the existing data correctness can be achieved through some innovative inventions. Besides, supplementary data, such as the vertical location could become available in an accurate way. For water distribution companies the data improvement could be achieved by the means of the Autonomous Inspection Robot (AIR). Second and a bit like the previous section, the model can help to reduce the human bias. This could be done by either cooperating with experts or training the experts. Third, if the data correctness improves, the localization areas could be implemented in the model. This could help to localize the best locations for trial trenches. Last, the large number of data all network operators have could be shared by innovative solutions. The best example in the nearby future is the merge of KLIC into KLIC-WIN, where most of the data will be stored in a central location. Better data facilities, correctness and availability will help to improve the model and stimulate network operators to develop new tools to improve their services.

7.5.3 STRENGHTS AND THREATS

Opposite to the previous section, here strategies are examined whereby the threats are overcome or avoided by the strengths. First, the predictable power of the model is not perfect at all and could therefore lead to very remarkable predictions. This can be overcome if experts are kept in the process and have the final judgement. So, the machine must be controlled and checked by experts as they know the details about excavation failures. Second, a campaign showing that the benefits of a model as a supplement to the already extensive risk-assessment procedure could reduce cost and risk for all network operators has to make sure no one is going to frustrate. Last, as most databases will be shared with the Kadaster for KLIC-WIN, it is possible to compare the databases on correctness and type of data to ensure there are similarities between the databases.

7.5.4 WEAKNESSES AND THREATS

Last, strategies to minimize the effect of weaknesses and overcome threats are examined. On one hand, the model's predictive power could be worthless when the data is very incorrect. Therefore, some examples from the case in this thesis will be necessary to prove that the model is also working with the incorrect data. On the other hand, the network operators should be involved in the process of implementation to ensure that the experts are also willing to join. It could be that they feel passed, but it will be sold as an additional tool.

7.6 APPLICATION FROM THE RESULTS

When the TOWS-matrix is considered and combined with the results from this thesis some possibilities for application are found. The relevant variables and to what extend the variables affect the probability of failure from excavation works are important for the network operators. Although it will deviate for all network operators, because of the dataset and the type of network they manage or possess, this section will elaborate on the variables that could help Evides reduce the number of failures from excavation works. So, what should Evides consider doing and what factors are most important during the risk-assessment.

At this moment the area managers assess the incoming KLIC-requests. Based on expert judgment the probability of failure following from the forthcoming excavation work is estimated based on several surrounding characteristics, varying per expert. This thesis identified aspects with an extensive data analysis that increase or decrease the probability of failure.

First, network operators should be more alert for the emergency KLIC-requests as these increase the probability of failure with an estimated beta of 2.22, while being way below the significance level. From a further look into the variable, it was found that emergency KLIC-requests represent only 5% of the total dataset, whereas it represents 23% of the failures.

Second, network operators (or at least Evides) should be alert for the excavation activities from telecom providers on their networks. It was found that irrespective to the side, excavation activities for telecom providers increase the probability of failure. The 'streetside' has a beta of 1.86 whereas the 'buildingside' has a beta of 1.00. This is remarkable because from the literature and expert interviews it was expected that the building side would have a larger probability as the service connections are crossed. At the crossings the probability of failure from excavation works is higher.

A hypothesis is stated that the telecom cables on the 'streetside' are located nearer to the water main than the telecom cables on the building side. This is expected since telecom providers prefer to be as close as possible to the facades for shorter service connections. However, when this hypothesis was tested, it was found that on average the telecom cables on the building side are half a meter closer to the water mains (Figure 24). Therefore, there is no clear explanation for this finding, besides the fact that excavation works for telecom providers increase the probability of failure.

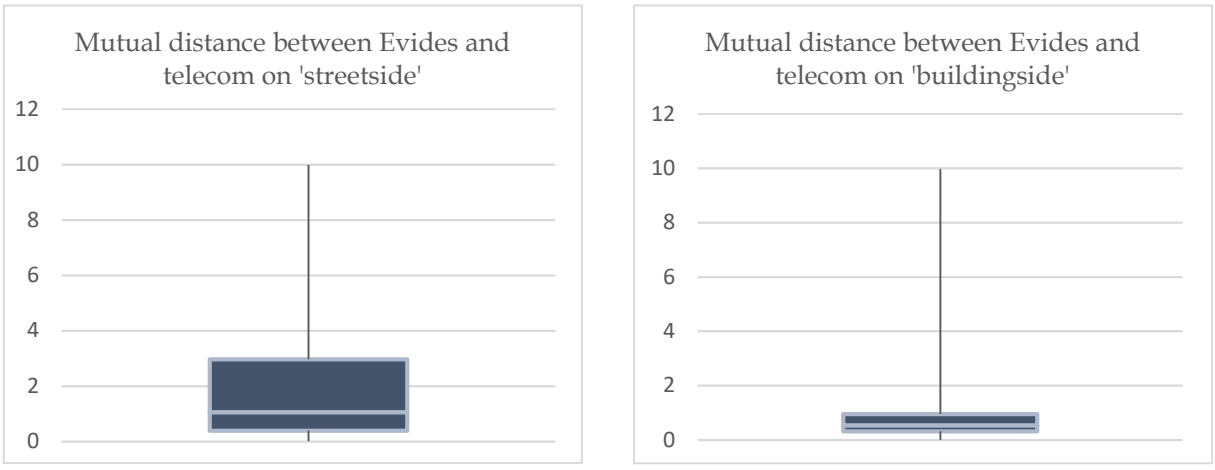


Figure 24: The distance from the telecom cable to the Evides main for the two 'sides' are compared. It was found that the telecom cables are +- 50 cm closer to the Evides main on the building sides than on the street sides.

Third, the own assets should be considered more carefully. On the one hand because of the own diameter. The smaller the diameter (β = -0.01), the larger the probability of failure. The estimate looks small, but since the Evides diameter is registered in millimeters the difference between a 20 cm and 30cm pipe is relevant. Using equation [15] and the results from Table 14 with the intercept and a 200

and 300-millimeter diameter, raised P(Y) from 0.00033 to 0.00088, which is almost the triple. Comparing a 300 and 50-millimeter diameter pipe raises P(Y) even from 0.00033 to 0.0040, which is 12 times larger. Because of the linearity, it can be stated that per 100 mm decrease of diameter, the probability of failure from excavation becomes three times larger.

$$P(Y) = \frac{1}{1 + e^{-(-5.03 - 0.01\, evi_{diameter} + \cdots + b_n X_n + \varepsilon_i)}} \quad [15]$$

On the other hand, the material of the own network is of importance. This probably follows from the material strength and stiffness, but some materials were found altering the probability positive and some negative. From Table 29 in appendix 12.8 it follows that PVC pipes (β = 14.77) increase the probability most, whereas Polyethylene, which is a relative new material had been found to decrease the probability of failure (β = -0.25). Please note that these beta values follow from a model with a very large intercept (β = -20.34) and where all variables are included.

Fourth and probably the hardest to implement for network operators is the correctness of the data. It was found that for every meter deviation of the Rotterdam3D data compared to the Evides data (which is assumed to be true) the probability of failure increases by 1,46 (per meter deviation). However, since the actual deviation with the real cable or pipe location is unknown, it could be very different and is therefore also included in the sector recommendations and discussion.

Last, but to a smaller extent, the sewer systems were found to have a negative estimate for all variables. That is, the diameter and both sides are all below the significance threshold and with negative estimates varying from -1.19 to -1.50. Even the mutual distance between the sewer- and water pipe has a (small) negative estimate. This is remarkable as sewer systems, large diameter, deeply located networks, are considered as one of the parties causing most excavation failures.

7.7 SUB-CONCLUSION III

As the model itself is not sufficient to predict all failures from excavation works accurately, other ways of application were tested. With a GAP-analysis both, the desired situation as well as the difference between that situation and the current situation were identified. The opportunities and threats that could fill the gap were identified through a SWOT-analysis. The SWOT was translated into strategies by the TOWS analysis to be able to answer sub-question three: *In what way can network operators use the model to reduce failure from excavation works?*

The model from this thesis can serve several strategies to come to the desired situation. The model can be applied to predict a certain number of failures, which are not predicted by experts now. The weighted and SMOTE models were capable to respectively predict 38% and 58% of the failures accurately. However, the former predicts 6% false positives whereas the latter predicts 37% false positives. To reduce failures from excavation works further network operators can use several strategies.

First, the model can complement the expert's risk-assessment as well as train the expert's knowledge. Complementing the expert will increase the number of predicted failures as the models itself already predicted 38% and 58% of the failures that were not predicted or prevented before. This will also train the expert's knowledge about the probability of failure from excavation works as the expert is supported in the risk-assessment. Training and complementing the expert can also help to reach more consistency during the risk-assessment procedure by reduction of the human bias. Opposite, experts can assess the outcome of the model to filter the false positive predictions.

Second, the model proved once again that incorrect data has large influence on failures from excavation works. However, despite the incorrect data, the weighted and SMOTE models are still able to accurately predict with a balanced accuracy of respectively 0.66 and 0.61. Using new techniques to improve the data correctness will result in a more accurate predicting model, which is helpful to reduce the number of failures from excavation works.

8 DISCUSSION

This thesis aimed to fill the theoretical and practical gap on the probability side of cable and pipe failures from excavation works. In this chapter both the findings are discussed as well as the limitations of the research approach.

DATA QUALITY

The reliability of the findings from the results are most dependent on the correctness of the data. CROW500 included an extra meter for the minimal localization area in all horizontal directions because of the inaccuracy of data. On top of that, the interviewed experts did not agree about the correctness of the data. One expert thought that the virtual and actual location were almost similar, whereas another expert believed that the virtual location deviates about 50 centimeters from reality. This thesis assessed the data correctness based on the deviation between two 'virtual' databases: Rotterdam3D versus Evides' asset database. From that comparison it was found that over 95% of all mains are (virtually) located within 40 centimeters from each other. However, comparing databases is not equal to comparing actual locations with virtual locations. If it is true that the real location deviates more from the virtual data than assumed, the quality of this study can move in two directions when the correctness of the location data is improved. Either, the data correctness increases, which will improve the model because the real situation is simulated better. Or the data correctness increases after which it is found that the model predicted based on randomness whereby it will become useless. However, in case of the latter, the applied method from this thesis could still be used to develop a new model based on the new data.

On top of that, this thesis has been conducted within a single area, based on one (created) dataset. As a result, validation had to be done by splitting the single database into a training and test set, whereas it would be preferred to validate the model on a completely new dataset. Furthermore, the data are separately stored in multiple databases, which are despite digitalization, hard to link whereby this and other analysis are complicated. When the various networks are linked, multiple criteria are necessary to reduce the nuisance. The linking criteria are of great importance as the most balanced dataset (failure / non-failure) is required, including as many data as possible. Despite a triangular linking process, multiple failures were lost because of the stated criteria. It is expected that more deliberate ideas could be found to improve the linking process, whereby the failure rate and the number of data increases.

Moreover, only assets within a mutual distance within 10 meters were linked to each other, whereby multiple assets were not linked, simply because these were not there, resulting in incomplete data. Consequently, empty fields are imported into the statistical software programs, that interpret the absent variables as non-available variables. A single non-available or missing variable in a sample is enough to exclude the entire sample from the analysis. Therefore, the empty fields (absent variables) were filled with virtual, non-existing data to ensure the sample is not excluded. Most studies would use the mean to replace missing variables, but since the empty fields have a meaning in this study (e.g. not within 10 meters), the mean value was not an option. Therefore, the numbers that were selected as replacement were more or less randomly selected, but in such a way that the replaced numbers are recognized (e.g. 12 meters as it is larger than 10 meters). Alternative methods should be considered to solve this 'problem'. The virtual, non-existing data as described has also been used to complement the data about the sides of other cables and pipes.

METHODOLOGY

Only one prior study on cable and pipe networks was found using logistic regression to assess sections of infrastructure systems instead of testing the system as a whole (Ariaratnam et al., 2001). The logistic regression in this thesis proves that it is an effective method to assess individual situations, such as excavation works. Probably, alternative modeling methods for binary data would also be effective. When considering a binary method, the event / non-event ratio should be kept in mind as validation proved that logistic regression on rare event data does only predict the major class. Four sampling techniques were tested and compared for the rare event data. However, the most recent method dates

back to 2002, which is already 16 years ago. Together with the (technological) innovation and digitalization over the last decades, it is expected that better sampling methods have been developed. The four applied sampling techniques are discussed later in this section.

Another important assumption due to limited time is the relation between the independent variables and the dependent variable. All independent variables in the model were assumed to have a direct relation to the dependent variable (Figure 25, simple relation). However, multiple other relations exist, which affects the way the independent variables should be implemented in the model. On the one hand, variables are moderating when they affect the relation between other variables directly. On the other hand, mediating variables explain the relationship between the independent and dependent variable by the relation it has with the two variables (Field, 2013). For example, the type of work could cause a failure. However, indirect, the horizontal distance between the work and the other possible failure location also affects if a failure will follow (Figure 25, mediation).

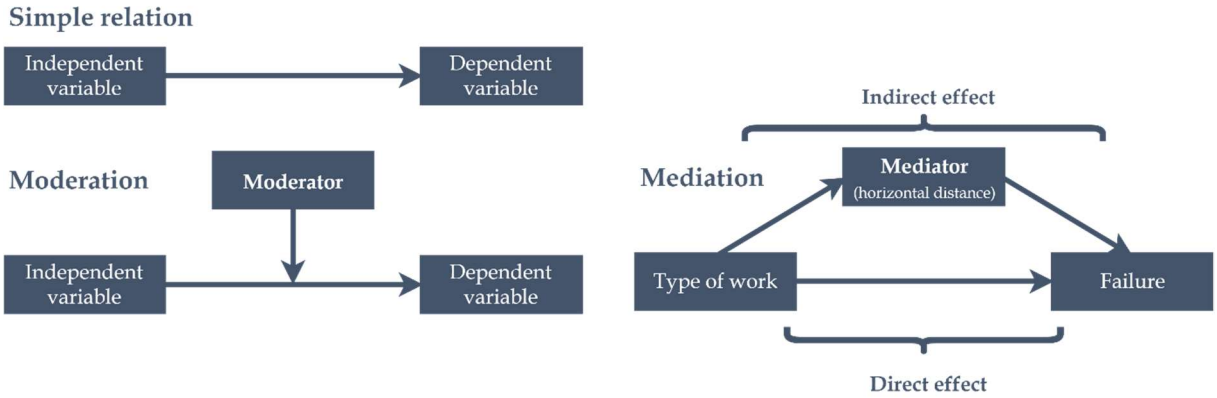


Figure 25: The possible interaction effects between variables. In this thesis the independent variables are assumed to have a direct relation to the dependent variable. For the mediating interaction effect an example of possible mediation is given (Field, 2013)

The first sampling technique that was applied was weighting, combined with under sampling from the majority class. As discussed, weighting can be done through application of the automatic *R* package or manually. The package does not support most statistical tests, such as goodness of fit tests, whereas the calculations can also be done by hand. It is preferable to do both, in order to compare the different approaches. In this thesis, the manual weighting resulted in very diverging results from the automatic weighting. Despite various trials to solve this 'problem', it did not work out whereby the results of the manual weighting can be considered as worthless. Second, a synthetic minority over sampling technique was applied. This method has the disadvantage that categorical variables are excluded as these cannot be oversampled in a synthetic way. So, as for weighting only the automatic version worked properly, no goodness of fit tests could be done, whereas with the SMOTE method the categorical variables had to be excluded. Therefore, the models from both rare event techniques were not able to fulfill all tests on all data leaving uncertainties about the model's quality.

Bayesian Logistic Regression was tested on the full data set, just like the 'full data' model. Looking back, it is doubtful whether this was the best way. In first instance, the 'full data' model was already found to have a balanced accuracy of 0.50 because of the imbalanced data. According to the reviewed literature, Bayesian logistic regression was more appropriate to handle the imbalanced (or even rare event) data. However, from the results it was found that the Bayesian logistic regression model was not able to predict better than the 'full data' model. With the current knowledge, a pre-sampling, to balance the sample set a little more would be preferred. Basically, it is expected that other modelling methods for binary dependent variables could also work when combined with a proper sampling technique and should therefore be tested. Due to the limited time, this study did not test other modeling methods than logistic regression and Bayesian logistic regression.

SECTOR

The results of this thesis can be valuable for the network operators (i.e. cable and pipe network sector). The first results indicate that certain variables are relevant when trying to predict failures from excavation works. However, more interesting are the results of the models and sampling techniques.

First, the model results indicate that binary dependent variables can be used in the cable and pipe sector. This creates the opportunity to assess specific situations in future studies, in which the dependent variable can represent any categorical variable with two outcomes. Nevertheless, it is important to realize that this study was conducted with data from Evides water distribution company. All networks have different specifications and regulations, from which characteristics follow. For example, an electricity cable has a small diameter and is flexible compared to water mains. In addition, the water main is probably located deeper and further away from buildings than the electricity cables. Therefore, the electricity cable is probably more vulnerable for excavation activities as it is smaller and crossed more often. A comprehensive study is necessary to check whether the method from this thesis could be transferred to other network types'. This could be done, by application and validation of the model for another network type, whereas the validation will show whether the model is applicable for other networks.

Second, the sampling techniques that were applied to balance the rare event data could be real interesting for network operators. Both techniques, SMOTE and Weighting were successful in increasing the balanced accuracy of the model. However, this thesis only tested the techniques on that very specific topic and did not assess other applications of the technique. Therefore, the applicability of the techniques for other issues could be questioned. Despite this uncertainty, it can be expected that it will also work for other studies of network operators. This is supposed because it is an existing technique coming from other sectors (political medical and financial sectors), where it was applied for various purposes (e.g. predicting wars or diseases). Since it has already been applied in certain sectors and for multiple purposes, it is expected that it will also be successful for various applications by network operators.

Third, it has not been studied whether the incentives and strategies of all network operators are similar. At this moment the costs of failures are lower than the costs preventing failures, whereby low costs can also be seen as an incentive for network operators. It can therefore be questioned against what price network operators are willing to reduce failures (and improve reputation).

Fourth, orientation KLIC-requests were excluded from this thesis. Due to the limited time, linking these to failures and testing whether such requests have been done prior to the regular KLIC-requests was not possible. It could be worthwhile for the sector to test whether there is a relation between orientation KLIC requests and excavation damage.

(This page is intentionally left blank)

9 CONCLUSION

The objective of this thesis was to develop a model to accurately predict cable and pipe failures from excavation works, considering spatial interdependencies. The model has to predict failures to allow network operators to take risk mitigating measures and to verify the size of the localization area by trail trenches. To meet the objective, the following main research question was answered:

What method can predict the influence of spatial interdependencies on the probability of failure from excavation works on the cables and pipes of subsurface utility operators?

First the conclusions of the sub-questions will be summarized and combined in order to answer the main research question of this thesis in the end.

9.1 SUMMARY OF THE SUB-QUESTIONS

1. *What variables are most related to cable and pipe failure from excavation works?*

Given the research objective, relevant variables that are related to cable and pipe failures from excavation works were identified first. These variables were identified via a literature review, three expert interviews, p-values and a stepwise backward elimination procedure based on the AIC score. From the literature review and expert interviews the same relevant variables followed. The p-values and stepwise backward elimination that were extracted from a binary logistic regression model complemented the list with relevant variables. The three methods together, resulted in the most related variables to cable and pipe failure from excavation works, which were all inserted in a logistic regression model.

2. *To what extent are the identified variables affecting the probability of failure from excavation works and how can we accurately predict the probability of failure?*

Regular logistic regression is not able to accurately predict probabilities of failure when data is imbalanced. To increase the accuracy of the model, rare event sampling techniques were used for over- and under sampling as well as compensations for the minority were applied.

The weighted and under sampled model predicted 38% of the failures and predicted 94% of the non-failures accurately, resulting in a balanced accuracy of 0.66. The SMOTE model predicted 58% of the failures accurately and 63% of the non-failures, resulting in a balanced accuracy of 0.61. Concluding, the weighted and under sampled model predicts more accurate looking at failures and non-failures, whereas the latter predicts more accurate when considering only the failures.

3. *In what way can network operators use the model to reduce failure from excavation works?*

From a gap-analysis together with a SWOT and TOWS analysis several strategies were found for applications of the model to reduce cable and pipe failures from excavation works. First, the models can be applied by itself to accurately predict some failures (as described above). Second, the model can complement the expert's risk-assessment (by 38% and 58%) as well as increase the consistency of the predictions by training the expert's knowledge. On the other hand experts can be used to assess the outcomes of the model.

9.2 CONCLUSION TO THE MAIN RESEARCH QUESTION

This thesis filled the theoretical and practical gap that was found on the probability side of cable and pipe failures from excavation works. The objective was to develop a method that can help network operators to accurately predict cable and pipe failures from excavation works, whereby risk mitigating measures could be taken based on the probability of failure. To meet the objective, the following question was defined:

What method can predict the influence of spatial interdependencies on the probability of failure from excavation works on the cables and pipes of subsurface utility operators?

In this thesis, a method was identified to accurately predict the cable and pipe failures from excavation works. Several steps are necessary to end with an accurately predicting model. The existing knowledge serves as starting point, in the form of a literature review and expert interviews. This basis is used for the data collection, in which data of all networks in a certain area are collected and connected, whereby KLIC-requests serve as the base and failures and non-failures as the dependent variable.

Binary logistic regression was selected as the modelling method. It was found that logistic regression is not able to handle rare event (i.e. or imbalanced) data. From literature, two effective methods were identified to deal with the rare event data. First, weighting and under sampling were tested to correct for the underestimating model. The balanced accuracy was 0.66 whereby 38% of the failures were accurately predicted and 94% of the non-failures.

Second, a synthetic minority over sampling technique (SMOTE) was applied to balance the dependent variable in the trainings sample set. The model following from this technique was able to accurately predict 58% of the failures and 63% of the non-failures (i.e. balanced accuracy 0.61).

Both methods have shortcomings; the goodness of fit of the former cannot be tested, whereas the latter is not able to handle categorical independent variables. Despite these deficiencies, both methods predicted more accurate than the basic model without modifications which was incapable to predict any failure accurately.

Even though both models (weighted and SMOTE) cannot predict all failures accurately, it can increase the accuracy of the predictions which are currently done by experts. A combination of the two predictive methods, by expert and one of the models will be able to predict cable and pipe failures from excavation works more accurate than is currently attained by experts only.

10RECOMMENDATIONS

Since this thesis applies techniques which are new for network operators and it has not been tested yet and because the study contained several limitations as discussed in the discussion section, further research is worthwhile. First recommendations for further study are enumerated from the limitations of this research, after which some recommendations for network operators follow.

10.1 RECOMMENDATIONS FOR FURTHER STUDY

First, it is recommended to have a more detailed look at improved rare event techniques. Weighting dates from 2001 and SMOTE dates from 2004. With all technical developments including artificial intelligence it is expected that new sampling techniques have been established, as were identified in "an in depth review of rare event detection from an imbalanced learning respective" (Haixiang et al., 2017, p. 220). Furthermore, the manual weighting has many potentials if it is modelled properly. Due to the limited time, the 'problem' with this model was not solved and therefore the accuracy of the model has not been tested. It is therefore recommended to test the manual weighted model, to find more plausible results.

Second, this study only focused on the probability side of the risk of failure from excavation work. As risk exists of a probability and a consequence, it would be a major improvement if possible consequences could also be included. However, as this would probably alter the dependent variable from a categorical dichotomous variable into a numerical variable, new methods should be identified for modelling with rare event data on ordinal variables.

Third, the current risk-assessment is done based on expert judgment. Whereas this thesis focused on the 'hard' statistical side, the human aspect in the risk-assessment has not been considered. If the risk-assessments will be done by both the model and the expert in parallel, it would be real interesting to study the differences and similarities in their assessment. It should be studied how the risk-assessments are currently done in practice and what decision making is based on. Since all parties have different opinions and different interest, it is expected that everyone will come up with different 'solutions'. To identify potential differences in the risk-assessment between parties, a qualitative study is necessary. If large differences are found, an automated system (based on a model) could help to create equivocality.

Last, within limited time, the balanced accuracy of 0.66 for the weighted model in this thesis has been achieved. The applied method proved to be able to accurate predict some failures as the balanced accuracy increased from 0.50 to 0.66. However, it is expected that there are multiple possibilities to improve the predictive power of the model even further by slightly moderating some details. Due to the limited time, there has not been an extensive search for the optimal performance point. It is expected that parties with more extensive resources (like time and knowledge) could increase the balanced accuracy much further while applying the same method.

In this thesis a sensitivity analysis for the dataset has been conducted (see 6.3.2.1 and 6.3.3.1). However, this thesis did not do a sensitivity analysis on the model. That is, also for the model the number of inserted variables can be varied (k). From the validation overfitting or underfitting could be found. Despite this fact, this thesis did only do goodness of fit tests, like the log likelihood ratio to test the entire model fit. It is recommended to validate different model sizes to test the sensitivity.

10.2 RECOMMENDATIONS FOR THE SECTOR

For the sector, more practical recommendations have been developed. First, it is recommended to do a study to test the actual deviation between virtual location data and the actual locations. This could be done through picking multiple sampling locations, where different situations are assessed. It is recommended as more knowledge about the data correctness can have major influence on possibilities for new studies.

Second, it is recommended to test logistic regression and other binary dependent data modelling techniques for other purposes. For example, Cox Proportional Hazard could be used to test more time related data, such as the current state of a pipe which was constructed many years ago. Also, Bayesian Networks could be tested that predicts the probability with a graphical model. The same applies to the sampling techniques that have been used. As most failures will be rare events on the many kilometers pipes, these techniques can also be used for other studies.

Third, it is recommended for network operators that the cooperation between databases comes naturally. The cooperation is essential as this thesis uses the connections between various aspects. On the other hand, it is recommended to conduct a study on the willingness of other network operators to share data and cooperate. This is helpful for studies like this thesis, as well as instructive since the network operators can benefit from each other's data and knowledge.

Fourth, study what parties are using emergency KLIC-requests above average. This could indicate unnecessary use of it (probably because one can start excavation immediately instead of waiting for three days). At this moment, the emergency KLIC-requests can be up to an area of 500x500 meters. It is recommended to reconsider if it is useful that emergency KLIC-requests, which should only be used when excavation work is so urgent that it cannot wait, should be allowed up to a polygon size of 250,000 m2. Probably, network operators know where a failure occurs and can scope to an area much smaller. Therefore, consider a standard size for the KLIC-polygon, so network operators should only point the precise location after which automatically an area of e.g. 20x20 is drawn around it.

Additionally, some shortcomings of the procedure around KLIC-requests were found. When a network operator, during the design phase, determines a new possible location, issuing of the permits takes more than three days. As no excavation can be done within the first three days after the a regular KLIC-request, the new location is determined and permitted before the actual profile is determined by trial trenches. Therefore, no options are left to include the actual found profile. Therefore, it is recommended to study the procedure around the KLIC-requests.

Finally, it is recommended to do further research on the locations of telecom cables as the model proved that it has a large effect on the probability of failure. Especially the side (street side or building side) where the cables or pipes are located seemed to be very important. It was expected that crossing the service connections, which are closer to the surface causes the high probability of failure. However, in section 7.6 it was found that this is not true. Therefore, a further study on the excavation works of telecom providers is recommended.

11 BIBLIOGRAPHY

Agentschap Telecom. (2016). *WION en schade door graven*.

Akosa, J. S. (2017). Predictive Accuracy : A Misleading Performance Measure for Highly Imbalanced Data. *SAS Global Forum*, *942*, 1–12.

Amador, L. E., & Magnuson, S. (2011). Adjacency Modeling for Coordination of Investments in Infrastructure Asset Management. *Journal of the Transportation Research Board*, *2246*, 8–15.

Ariaratnam, S. T., El-Assaly, A., & Yang, Y. (2001). Assessment of Infrastructure Inspection Needs Using Logistic Models. *Journal of Infrastructure Systems*, *7*(4), 160–165.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chawla, N. V, Japkowicz, N., & Drive, P. (2004). Editorial : Special Issue on Learning from Imbalanced Data Sets. *ACM SIGKDD Explorations Newsletter*, *6*(1), 1–6.

CROW-werkgroep. (2016). *Schade voorkomen aan kabels en leidingen. Richtlijn zorgvuldig grondroeren van initiatief- tot gebruiksfase*. Ede.

DiCicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical Education*, *40*(4), 314–321.

Dyson, R. G. (2004). Strategic development and SWOT analysis at the University of Warwick. *European Journal of Operational Research*, *152*(3), 631–640.

Engelhardt, M. O., Skipworth, P. J., Savic, D. A., Saul, A. J., & Walters, G. A. (2000). Rehabilitation strategies for water distribution networks: A literature review with a UK perspective. *Urban Water*, *2*(2), 153–170.

Evides. (2017). *Jaarverslag 2016*. Rotterdam.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (Fourth). London: SAGE Publications Ltd.

Groot, P., Saitua, R., & Visser, N. (2016). Investeren in de infrastructuur. Trends en beleidsuitdagingen.

Grzenda, W. (2015). The advantages of bayesian methods over classical methods in the context of credible intervals. *Information Systems in Management*, *4*(1), 53–63.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239.

Harrell, F. E. (2001). Cox Proportional Hazards Regression Model. In *Regression Modeling Strategies* (pp. 465–507). New York, NY: Springer.

Hokstad, P., Utne, I. B., & Vatn, J. (2012). *Risk and Interdependencies in Critical Infrastructures, A guideline for analysis*. London: Springer.

Horton, J., Macve, R., & Struyven, G. (2004). Qualitative Research: Experiences in Using Semi-Structured Interviews. In C. Humphrey & B. Lee (Eds.), *The real life guide to accounting research* (First, pp. 339–350). Elsevier Ltd.

Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied logistic regression*. *Applied logistic regression* (Third). John Wiley & Sons.

Islam, T., & Moselhi, O. (2012). Modeling Geospatial Interdependence for Integrated Municipal Infrastructure. *Journal of Infrastructure Systems*, *18*(2).

Jeong, H. S., Abraham, D. M., & Lew, J. J. (2004). Evaluation of an Emerging Market in Subsurface Utility Engineering. *Journal of Construction Engineering and Management*, *130*(2), 225–234.

Kabel- en Leiding Overleg. (2015). *Analyse graafschades 2012- 1e helft 2014*.

Kabel- en Leiding Overleg. (2016). *Factsheet graafschade voorkomen*.

Kadaster. (n.d.). Graafmelding. Retrieved January 26, 2018, from https://www.kadaster.nl/-/graafmelding

Kadaster. (2017). KLIC (WION). Retrieved November 15, 2017, from https://www.kadaster.nl/klic-wibon

Khurkhuriya, J. (2018). A-Z Machine Learning using Azure Machine Learning. Retrieved June 4, 2018, from https://www.udemy.com/machine-learning-using-azureml/?couponCode=COUPON090

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, *9*(02), 137–163.

Kiwa. (2013). *Overzicht graafschade gas in 2012*. Apeldoorn.

Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression*. (M. Gail, K. Krickeberg, J. M. Samet, A. Tsiatis, & W. Wong, Eds.) (Third). London: Springer Science + Business Media LLC.

Kohavi, R. (2016). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Learning*, (March 2001), 1137–1143.

Kothari, C. R. (2004). *Research Methodology: Methods & Techniques*. *New Age International (P) Ltd* (Third).

KWR. (2011). *Registratie geeft storingen waarde – Implementatie en evaluatie van USTORE*. BTO Nieuwegein.

Li, J., Zhang, H., Han, Y., & Wang, B. (2016). Study on failure of third-party damage for Urban gas pipeline based on fuzzy comprehensive evaluation. *PLoS ONE*, *11*(11), 1–15.

Li, S., Cai, H., & Kamat, V. (2015). Uncertainty-aware geospatial system for mapping and visualizing underground utilities. *Automation in Construction*, *53*, 105–119.

Lombardo, L., Cama, M., Conoscenti, C., Marker, M., & Rotigliano, E. (2015). Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy). *Natural Hazards*, *79*(3), 1621–1648.

Marlow, D. R., Boulaire, F., Beale, D. J., Grundy, C., & Moglia, M. (2011). Sewer Performance Reporting: Factors That Influence Blockages. *Journal of Infrastructure Systems*, *17*(1), 42–51.

Marra, M., Biccari, C. Di, Lazoi, M., & Corallo, A. (2018). A Gap Analysis Methodology for Product Lifecycle Management Assessment. *Ieee Transactions on Engineeringmanagement*, *65*(1), 155–167.

McFadden, D. (1979). Quantitative Methods for Analyzing Travel Behaviour on Individuals: Some Recent Developments. In D. A. Hensher & P. R. Stopher (Eds.), *Behavioural travel modelling*. Taylor & Francis: Taylor & Francis.

McGrail, M., & Roberts, B. (2005). Strategies in the broadband cable TV industry: The challenges for

management and technology innovation. *Info*, *7*(1), 53–65.

Monroe, W. (2017). *Bernoulli and Binomial Random Variables* (No. Lecture Notes#7). Stanford.

Muhlbauer, W. K. (2004). *Pipeline risk management manual: ideas, techniques, and resources* (Third). Oxford: Elsevier.

Normcommissie 349 200 "Dwarsprofielen." (2009a). *NEN 7171-1 (nl) Ordening van ondergrondse netten - Deel 1: Criteria*. Delft.

Normcommissie 349 200 "Dwarsprofielen." (2009b). *Npr 7171-2 Ordening van ondergrondse netten - Deel 2: Procesbeschrijving*. Delft.

Osman, H. (2016). Coordination of urban infrastructure reconstruction projects. *Structure and Infrastructure Engineering*, *12*(1), 108–121.

Osman, H., & El-Diraby, T. (2007). Implementation of subsurface utility engineering in Ontario: cases and a cost model. *Canadian Journal of Civil Engineering*, *34*(12), 1529–1541.

Ottens, M., Franssen, M., Kroes, P., & Van de Poel, I. (2006). Modelling infrastructures as socio-technical systems. *International Journal of Critical Infrastructures*, *2*(2/3), 133–145.

Ouyang, M. (2014). Review on modeling and simulation of interdependent critical infrastructure systems. *Reliability Engineering and System Safety*, *121*, 43–60.

Palmer, D. (2015). Trial Trench Evaluation. Retrieved May 30, 2018, from https://www.ucl.ac.uk/archaeologyse/our-services/excavation-and-fieldwork/trial-trench-evaluation

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstem, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*(12), 1373–1379.

Rainey, C. (2016). Dealing with separation in logistic regression models. *Political Analysis*, *24*(3), 339–355.

Rijksoverheid.nl. (2017). Graafschade aan ondergrondse leidingen en kabels. Retrieved February 19, 2018, from https://www.rijksoverheid.nl/onderwerpen/bodem-en-ondergrond/graafschade

Riley, C. L., & Wilson, M. (2006). *Pipeline Separation Design and Installation Reference Guide*. Washington.

Rinaldi, S. M. S. M., Peerenboom, J. P. J. P., & Kelly, T. K. T. K. (2001). Indentifying, Understanding, and Analyzing Critical Infrastructures Interdependencies. *IEEE Control Systems Magazine*, *01*(6), 11–25.

Rubin, H. J., & Rubin, I. S. (2005). *Qualitative interviewing (2nd ed.): The art of hearing data* (2nd ed.). Thousand Oaks, CA: SAGE Publications Ltd.

Slee, K. G. L., & Tjan, S. Y. P. Y. (2015). *Handboek Leidingen Rotterdam 2015*. Rotterdam.

Statsoft. (2013). STATISTICA Formula Guide : Logistic Regression, 1–29.

Thienen, P., van Maks, M. H., & Yntema, D. (2016). Robots inspecting drinking water pipes. *Water Matters*, *2*.

Trotter, L. (2017). *Sustainable asset management in water distribution systems*. Delft University of Technology.

Tscheikner-Gratl, F. (2015). *Integrated Approach for multi-utility rehabilitation planning of Urban*

Water Infrastructrure. Universitat Innsbruck.

Tscheikner-Gratl, F., Sitzenfrei, R., Rauch, W., & Kleidorfer, M. (2016). Integrated rehabilitation planning of urban infrastructure systems using a street section priority model. *Urban Water Journal*, *13*(1), 28–40.

Tung, Y.-K. (1985). Channel scouring potential using logistic analysis. *Journal of Hydraulic Engineering*, *111*(2), 194–205.

Utne, I. B., Hokstad, P., & Vatn, J. (2011). A method for risk modeling of interdependencies in critical infrastructures. *Reliability Engineering and System Safety*, *96*(6), 671–678.

Van den Ende, P. (2016). *Uitkomsten onderzoek "Gebruik van Bomenregister.nl t.b.v. risico analyse voor de leidingligging."* Rotterdam.

van Eijk, R., & van Daal, K. (2013). *Storingen als gevolg van graaf- werkzaamheden Ruimtelijke analyse van USTORE gegevens*. Nieuwegein.

Van Mill, B. P. A., Gooskens, B. J. F., Noordink, M., & Dunning, B. R. (2013). *Evaluatie Wion*. Den Haag.

Van Norden, P. P. (2013). *Succesfactoren bij Preventie Graafschade*. Apeldoorn.

Van Oel, C. (2017). *Case studies AR3R057*. Delft: Delft University of Technology.

Vloerbergh, I. N., & Beuken, R. H. S. (2011). *Levensduur van leidingen – Een maatschappelijk en functioneel perspectief*. Nieuwegein.

Wei, L. X., & Han, L. Y. (2013). Third-Party Damage Factors Analysis and Control Measures of Daqing-Harbin Oil Pipeline. *Applied Mechanics and Materials*, *411–414*, 2527–2532.

Weihrich, H. (1982). The TOWS matrix—A tool for situational analysis. *Long Range Planning*, *15*(2), 54–66.

12 APPENDIX

12.1 VARIABLE EXPLANATION

Variable name	Explanation
klic_id_new	A reference number created per row to trace back situations if necessary
klic_polygon_size	The size of the KLIC-request
asset_length_in_klic	The sum of all assets that are situated in a certain (KLIC) polygon
klic_notification_type	Emergency or regular KLIC-request
klic_party	For what network has the excavation works been done
klic_type_work	What sector was the work conducted for
klic_type_work2	Specification of klic_type_work if it was cable or pipe excavation work
evi_age	The number of years an Evides pipe is in use in field
evi_material	Material of the Evides pipe
evi_diameter	Diameter of the Evides pipe
evi_distance_house	Distance to the nearest building from the measure (middle) point of the Evides pipe
evi_shape_length	Total 'virtual' length. So how long is the line drawn in GIS
evi_intersection_length	Length that a certain asset intersects the KLIC-polygon
evi_data_quality	Deviation between Evides and Rotterdam3D databases
sew_distance	Smallest distance from the Evides middle point to sewer
sew_diameter	Diameter of the sewer pipe (closest to Evides)
sew_side	Is the sewer located on the building side [2] or on the 'streetside' [1] compared to the Evides pipe
gas_distance	Smallest distance from the Evides middle point to sewer
gas_diameter	Diameter of the gas pipe (closest to Evides)
gas_side	Is the gas pipe located on the building side [2] or on the 'streetside' [1] compared to the Evides pipe
heat_distance	Smallest distance from the Evides middle point to district heating
heat_diameter	Diameter of the district heating pipe (closest to Evides)
heat_side	Is the district heating located on the building side [2] or on the 'streetside' [1] compared to the Evides pipe
elec_distance	Smallest distance from the Evides middle point to electricity cables
elec_side	Is the electricity cable located on the building side [2] or on the 'streetside' [1] compared to the Evides pipe
tele_distance	Smallest distance from the Evides middle point to telecom cables
tele_side	Is the telecom cable located on the building side [2] or on the 'streetside' [1] compared to the Evides pipe
cable_distance	Smallest distance from the Evides middle point to cable (CAI)
cable_side	Is the CAI located on the building side [2] or on the 'streetside' [1] compared to the Evides pipe
remainder_distance	Smallest distance from the Evides middle point to remainder cables and pipes
remainder_side	Is the remainder cable/pipe located on the building side [2] or on the 'streetside' [1] compared to the Evides pipe
leak_dummy	Failure [1] or non-failure [0]

12.2 TYPES OF WORK

Table 27: Types of work

Kabels-leidingen	**Landscaping/gardening**	**Slaan/boren**
Gasleiding leggen	Hovenierswerkzaamheden	Aarding slaan
CAI kabel leggen	Bomen rooien/planten	Damwand/beschoeiing slaan
Huisaansluitingen maken	Stobben frezen	Diepploegen
Kabels/leidingen leggen	Vijver graven	Handholes plaatsen
Kabels/leidingen verleggen		Hekwerk plaatsen
Leggen laagspanning	**Overig**	Palen/masten plaatsen/verwijderen
Leggen middenspanning	Heien	Tanks/putten/containers in/uitgraven
Leggen hoogspanning	Baggerwerk	Persing/boring
Mantelbuis leggen	Archeologisch onderzoek	Duikers leggen enz.
Rioleringswerkzaamheden	Bodemonderzoek/sondering	
Stadsverwarming leggen	Bodemsanering	**Construction**
Telefoonkabel leggen	Proefsleuven graven	Bouwwerkzaamheden
Trafostations plaatsen	Bestratingswerk	Funderingswerk
Verwijderen kabels/leidingen	O.V.werkzaamheden	Grondwerk/bouwrijp maken
Waterleiding leggen	zie notitities	Reconstructiewerkzaamheden
Werk aan bestaande leiding	Zinker maken	Sloopwerkzaamheden
HDPE buis leggen	Drainage	Waterbouwkundige werken
	Drainage	Wegenbouw
		Woningbouw
		Woonrijp maken

12.3 LINKING OPTIONS

Option 1	Without filter	
Failures	281	281
Pipes	309355	309355
percentage	**0,091%**	

Option 2	Only link pipe nearest to middle point of polygon
Failures	97
Pipes	63000
percentage	**0,154%**

Option 3 Minimal percentage or distance for intersection

Minimal intersecting percentage	Assets left	failures left		**Minimal shape length**	Assets left	failures left	
0%	309083	271	0,088%	0	309083	271	0,088%
10%	263934	241	0,091%	3	285596	263	0,092%
20%	228978	219	0,096%	5	269156	253	0,094%
30%	202666	200	0,099%	10	204847	213	0,104%
40%	182118	185	0,102%	15	158596	184	0,116%
50%	165341	176	0,106%	20	127150	148	0,116%
60%	151571	160	0,106%	25	104262	127	0,122%
70%	139805	151	0,108%	30	87230	113	0,130%
80%	129400	140	0,108%	35	73659	102	0,138%
90%	119587	131	0,110%	40	62700	94	0,150%
100%	109531	118	0,108%	45	53384	78	0,146%

Minimal intersecting percentage	**Minimal shape length**	Percentage left	Assets left	failures left
0%	0	0,09%	309083	271
10%	3	0,09%	291043	267
20%	5	0,09%	276914	262
30%	10	0,09%	249599	237
40%	15	0,10%	226184	220
50%	20	0,10%	206002	202
60%	25	0,10%	189062	188
70%	30	0,10%	174408	179
80%	35	0,11%	161578	173
90%	40	0,11%	150042	163
100%	45	0,11%	139239	149

Minimum shape length during pipe preparation

		min shape_l = 0	min shape_l = 2.5	min shape_l = 5	min shape_l = 7.5	min shape_l = 10	min shape_l = 15	
DISL	distributie	196182	128304	115094	####	98816	86207	91%
TRDISL	Transportdistributie	11992	7954	7019	6487	6060	5371	6%
INL	Industrie	100	71	66	59	54	53	0%
PRODL	Productie	223	140	105	92	79	68	0%
RWL	Ruw water	933	874	820	769	754	712	0%
BLML	Brielse Meer Water	575	327	275	260	241	227	0%
BWL	Bron Water	357	232	210	199	189	175	0%
TRANSL	Transport	5565	3942	3327	3039	2871	2631	3%
DMWL	Demi Water	568	334	293	277	264	256	0%
		216495	**142178**	**127209**	####	**109328**	**95700**	100%

Option 4	only connect KLICs to assets middlepoints
Failures	185
Pipes	173690
percentage	**0,107%**

Option 5	Only link the pipe with the largest intersection length
Failures	43
Pipes	58267
percentage	**0,074%**

Option 6	Only link pipes that contribute a minimal percentage to the total asset length in the polygon	
	10%	20%
Failures	165	121
Pipes	132568	83610
percentage	**0,124%**	**0,145%**

Option 7	Only DISL, minimum shape 7.5		
	intersection > 0m	intersection > 7.5m	intersection > 15m
Failures	259	231	186
Pipes	257942	207808	139683
percentage	**0,100%**	**0,111%**	**0,133%**

Option 8	Only DISL, minimum shape 15		
	intersection > 0m	intersection > 7.5m	intersection > 15m
Failures	254	225	200
Pipes	215074	174655	139683
percentage	**0,118%**	**0,129%**	**0,143%**

12.4 INTERVIEW FORM

Introduction

- Wie ben ik? Studie, etc.
- Waarom ben [jij] geselecteerd?
- Wat gebeurd er met de interview data? → Anoniem, afstudeeronderzoek,
- DOEL:
  - Verband tussen variabelen van ruimtelijke afhankelijkheden en graafschade
  - Eerste setup voor de statistische analyse, correlatie.
  - Kritische blik op de reeds verzamelde data, en het aanvullen daarvan.
- **Toestemming om op te nemen**
- Introduceer [jezelf].

Onderzoeksvragen:

Welke methode is in staat om de invloed van ruimtelijke afhankelijkheden op de kans op graafschade aan kabels en leidingen te voorspellen?

1. Welke (ruimtelijke) variabelen zijn het meest te relateren aan derde schade?
2. In welke mate zijn ruimtelijke afhankelijkheden van invloed op de kans op schade door derden?
3. Hoe kunnen netwerk beheerders kennis over ruimtelijke afhankelijkheden implementeren om derden schade te voorkomen?

Graafschade: Is het een probleem en komt het veel voor?

1. **Wat is (meestal) de oorzaak van graafschade?**

- Soort werk (mechanisch/handmatig, leggen/gestuurde boring, etc.)
- Partij (Gas, Riool, Straat (gemeente), Electra, Telecommunicatie (incl. glasvezel), Stadsverwarming, openbare verlichting een verkeersregelinstallaties, afvalinzameling, industrieel/militair transport, de 'gewone burger').
- Aansluit, distributie, transport? Belang van leiding?
- Wat is hiervoor de oorzaak? (onvoorzichtig, moeilijk, **foute data)**
- Kruising/gewoon in de straat?

2. **Hoe goed is de informatie (van KLIC) die jullie/anderen hebben wanneer je gaat graven? DATA**

- Nauwkeurig?
- Hoe gebruiken jullie die data?
- Bij een probleem/melding geven jullie een probleem, oorzaak, storingscode op. Gebeurt dit op de juiste manier?
- Maken wij ook een melding als wij een schade veroorzaken?
- Wat doen wij met KLIC data? Wordt het opgeslagen en waar?
- Wordt bij een derde van de graafschades geen melding gedaan?

3. **Wederzijdse ruimtelijke afhankelijkheden**

- Wat denk je dat het is?
- Hoe beïnvloeden ruimtelijke afhankelijkheden de **kans op graafschade?**

Een tweezijdige relatie tussen 2 infrastructuren waarbij de staat van de ene de andere kan beïnvloeden. In het algemeen: twee infrastructuren zijn wederzijds afhankelijk als ze afhankelijk van elkaar zijn.

Infrastructuur systemen zijn ruimtelijk wederzijds afhankelijk als een lokale gebeurtenis beide kan beïnvloeden. Dit ontstaat doordat ze vlak bij elkaar liggen.

4. **Wat voor een dingen hebben veel invloed op de kans op graafschade?**

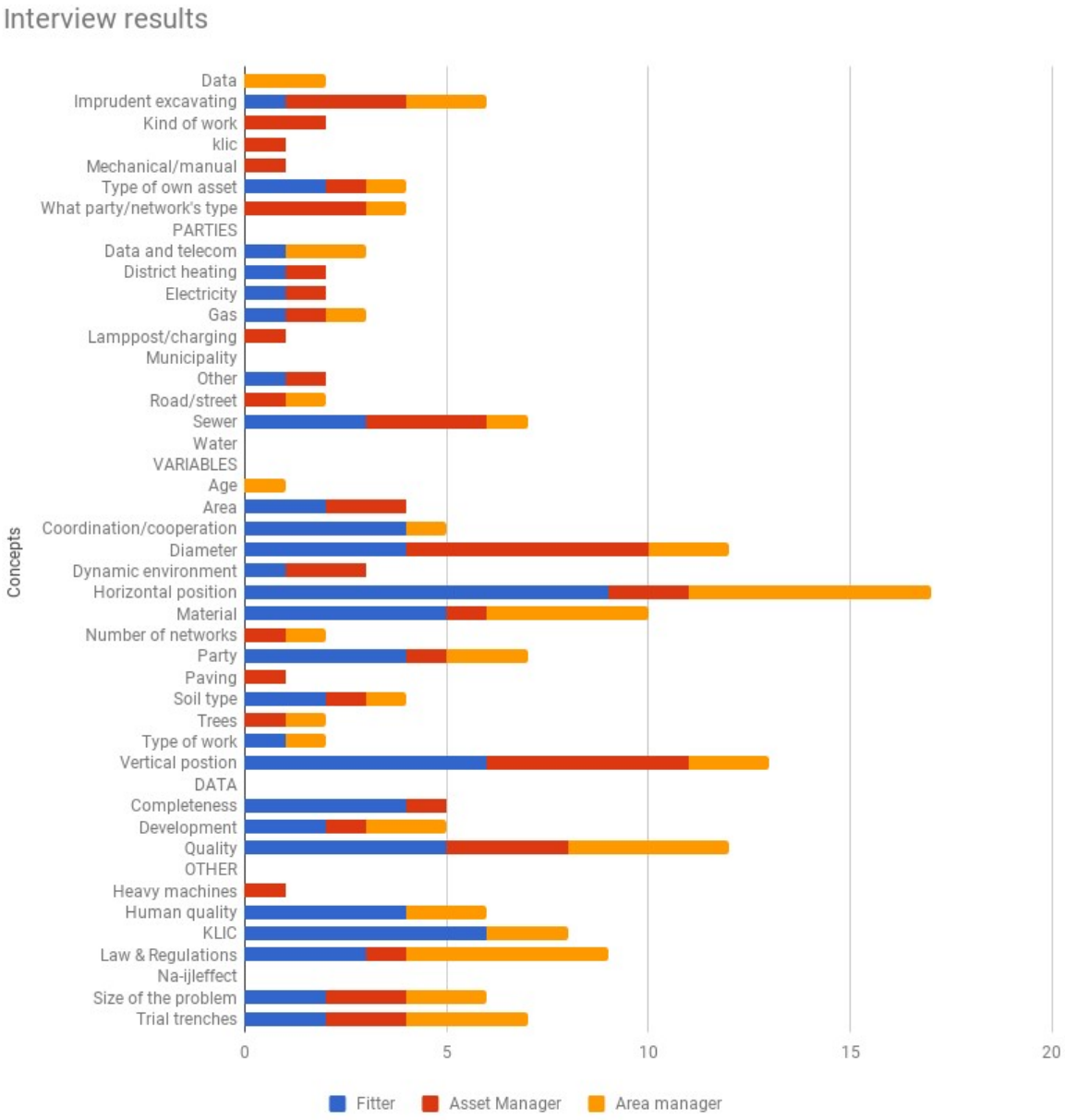
- Welke partij graaft? Waar wordt dit opgeslagen?

- De afstanden tussen de verschillende netwerken?
- Diepte? Diameter? Grondsoort? Materiaal van de waterleiding of het andere netwerk? Wie graaft, gas, riool, electra, etc.?
- Helpen de verschillende wetten en onderlinge afspraken bij het reduceren van de kans op graafschade? Welke wordt het meeste gebruikt/gehanteerd?
- Ken je **CROW 500**? Helpt het om de kans op graafschades te reduceren?
- SOORT WERK (leggen/verwijderen, damwanden, etc.)
- BOMEN
- LANTARENS
- TYPE BESTRATING
- GEBIED (stedelijk, landelijk, etc.)

5. **Volgens CROW 500 moet er bij mechanisch graven minimaal 1,50m proefsleuf worden gegraven.**
- CROW 500 zegt >1,5m. Als jij dit mag veranderen, dan meer, gelijk, of minder grote proefsleuven? Waarom?
- Weet je of de proefsleuven worden gegraven in geval van schade?
- Vroeger werkte Stedin en Evides, maar ook andere partijen meer samen. Dan werd er rechtstreeks naar het kantoor van 'de ander' gebeld wanneer er schade werd veroorzaakt. Nu niet meer, hoe komt dat denk je?
- Hoe kunnen we graafschade reduceren?

6. **Op welke manier zou kennis over de kans op graafschade en ruimtelijke afhankelijkheden Evides kunnen helpen? (Opportunity)**
- Waar in het process?
- Wie moet ermee aan de slag?
- Op welke manier helpt het?
- Hoe zou je het voor je zien?

12.5 INTERVIEW RESULTS



12.6 MULTICOLLINEARITY

Multicollinearity table

Variable	GVIF	Df	GVIF^(1/(2*Df))
klic_polygon_size	2,71	1	1,65
klic_request_type	1,44	1	1,20
evi_age	1,25	1	1,12
evi_diameter	1,17	1	1,08
evi_shape_length	1,31	1	1,15
evi_data_quality	1,03	1	1,02
sew_diameter	2,80	1	1,67
gas_distance	4,70	1	2,17
gas_side	9,27	2	1,74
heat_diameter	3,84	1	1,96
elec_distance	2,44	1	1,56
tele_distance	1,87	1	1,37
cable_distance	6,31	1	2,51
asset_length_in_klic	2,80	1	1,67
klic_party	1,54	6	1,04
evi_material	1,32	6	1,02
evi_distance_house	2,79	1	1,67
evi_intersection_length	1,34	1	1,16
sew_distance	1,50	1	1,23
sew_side	10,62	2	1,81
gas_diameter	4,80	1	2,19
heat_distance	5,84	1	2,42
heat_side	7,77	2	1,67
elec_side	6,61	2	1,60
tele_side	4,75	2	1,48
cable_side	6,76	2	1,61

12.7 TYPES OF VARIABLES

	What	Name	Included in analysis?	% available	Type	Type (statistics)	Specify
	id	id	0	100,00%			
KLIC	KLIC ID	klicnummer	0	100,00%			
	Date of request	klic_datumaanvraag	0	100,00%			
	Area size	klic_oppervlakte	1	100,00%	Num	Continuous	Interval
	Weight area size	klic_gewicht_oppervlakte	0	100,00%			
	Type of request	klic_meldingsoort	1	100,00%	Cat	Nominal	Dichotomous
Evides	Assetnumber	evi_assetnummer	0				
	Date of construction	evi_jaar_aanleg	0				
	Material	evi_materiaal	1	100,00%	Cat	Nominal	
	Diameter	evi_diameter	1	99,92%	Num	Continuous	
	Distance to nearest building	evi_distance_house	1	92,12%	Num	Continuous	Ratio
	Length in KLIC	asset_length_in_klic	1	100,00%	Num	Continuous	Ratio
	Type of asset (ansl/disl/etc)	evi_netdeel	1	100,00%	Cat	Nominal	
Sewer	Within 10 meters?	sew	1	100,00%	Cat	Dichotomous	
	House side?	sew_side	1	90,81%	Cat	Dichotomous	
	Distance to own asset	sew_distance	1	98,40%	Num	Continuous	Ratio
	Diameter	sew_diameter	1	98,40%	Num	Continuous	Ratio
Gas	Within 10 meters?	gas	1	100,00%	Cat	Dichotomous	
	House side?	gas_side	1	83,74%	Cat	Dichotomous	
	Distance to own asset	gas_distance	1	89,61%	Num	Continuous	Ratio
	Diameter	gas_diameter	1	89,61%	Num	Continuous	Ratio
District heating	Within 10 meters?	heat	1	100,00%	Cat	Dichotomous	
	House side?	heat_side	1	28,11%	Cat	Dichotomous	
	Distance to own asset	heat_distance	1	29,76%	Num	Continuous	Ratio
	Diameter	heat_diameter	1	29,76%	Num	Continuous	Ratio
Electricity	Within 10 meters?	elec	1	100,00%	Cat	Dichotomous	
	House side?	elec_side	1	91,07%	Cat	Dichotomous	
	Distance to own asset	elec_distance	1	99,55%	Num	Continuous	Ratio
Telecom	Within 10 meters?	tele	1	100,00%	Cat	Dichotomous	
	House side?	tele_side	1	90,18%	Cat	Dichotomous	
	Distance to own asset	tele_distance	1	98,54%	Num	Continuous	Ratio

CAI	Within 10 meters?	cable	1	100,00%	Cat	Dichotomous	
	House side?	cable_side	1	72,29%	Cat	Dichotomous	
	Distance to own asset	cable_distance	1	74,93%	Num	Continuous	Ratio
Other	Within 10 meters?	other	1	100,00%	Cat	Dichotomous	
	House side?	other_side	1	6,15%	Cat	Dichotomous	
	Distance to own asset	other_distance	1	7,11%	Num	Continuous	Ratio
Leakages	lek_assetnummer	lek_assetnummer	0				
	lek_id_oorzaak	lek_id_oorzaak	0				
	lek_datum_opvoeren	lek_datum_opvoeren	0				
	lek_eam_gemeld_op	lek_eam_gemeld_op	0				
	lek_storingscode	lek_storingscode	0				
	lek_probleemcode	lek_probleemcode	0				
	lek_oorzaakcode	lek_oorzaakcode	0				
	lek_outcome	lek_outcome	1	100,00%	Cat	Dichotomous	
Correctness		d_water_water	1	100,00%			

12.8 STATISTICS OF DATASET

Table 28: In the top half of the table, the percentage of certain categorical data from the entire dataset is compared to the percentage of that categorical variable within the failures. From the full data model, the estimates and p-values were also included to compare the results. In the bottom half, the numerical variables from the full data model are compared. First the average when all data is considered, then the average of the failures only.

		All data		Failures only			Full data model	
(Categorical) Variable	*What*	*Number*	*Percentage*	*Number*	*Percentage*	*Diff*	*Estimate*	*P-value*
KLIC_party	Telecom	6347	6%	7	4%	-2%		
	Sewer	**12010**	**11%**	**29**	**16%**	**5%**	**0,43**	**0,30**
	Gas	8109	8%	17	9%	2%		
	District heating	**2691**	**3%**	**12**	**7%**	**4%**	**0,81**	**0,10**
	Electricity	4985	5%	7	4%	-1%		
	Water	6235	6%	12	7%	1%		
	Unknown	**67130**	**62%**	**98**	**54%**	**-9%**		
Type of KLIC-request	Regular	101645	95%	141	77%	-17%		
	Emergency	**5862**	**5%**	**41**	**23%**	**17%**		
Evi_material	PVC	84745	79%	167	92%	13%		
	PE	**2428**	**2%**	**2**	**1%**	**-1%**	**-1,18**	**0,20**
	AC	2423	2%	3	2%	-1%		
	GGIJ	6315	6%	3	2%	-4%		
	ST	4601	4%	7	4%	0%	0,73	0,30
Sew side	0	70355	65%	129	71%	5%		
	1	21959	20%	36	20%	-1%		
	2	15193	14%	17	9%	-5%	1,53	0,03
Gas side	0	48659	45%	91	50%	5%		
	1	32003	30%	49	27%	-3%		
	2	26845	25%	42	23%	-2%		
Heat side	0	12737	12%	38	21%	9%	-0,61	0,08
	1	7505	7%	11	6%	-1%	-0,70	0,00
	2	87265	81%	133	73%	-8%	-1,05	0,02
Elec	0	47800	44%	73	40%	-4%		
	1	47312	44%	88	48%	4%	0,32	0,07
	2	12395	12%	21	12%	0%	1,00	0,06
Tele	**0**	**12647**	**12%**	**47**	**26%**	**14%**	**-0,90**	**0,00**
	1	**80296**	**75%**	**116**	**64%**	**-11%**	**-0,93**	**0,00**
	2	14564	14%	19	10%	-3%		
Cable	0	6327	6%	11	6%	0%		
	1	53045	49%	102	56%	7%		
	2	48135	45%	69	38%	-7%		
Remain	0	1315	1%	0	0%	-1%		
	1	1531	1%	1	1%	-1%		
	2	104661	97%	181	99%	2%		
Type of work	Cables and pipes	61963	58%	116	64%	6%	0,68	0,13

Construction	7946	7%	6	3%	-4%		
Landscaping/ gardening	6508	6%	4	2%	-4%		
Piling/drilling	7711	7%	6	3%	-4%		
Remainders	17517	16%	9	5%	-11%		
Unavailable	**5862**	**5%**	**41**	**23%**	**17%**	**2,22**	**0,00**

(Categorical) Variable	*Number*		*Number*	*Difference*	*Percentage*	*Estimate*	*P-value*
KLIC_polygon	7332,90		-1368,08	8700,97	-16%		
Evi_age	34,81		-0,33	35,15	-1%		
Evi_diam	**133,64**		**-28,73**	**162,37**	**-18%**	**-0,01**	**0,00**
Evi_data_quality	**0,09**		**0,01**	**0,07**	**17%**	**1,46**	**0,03**
Sew_diam	**0,27**		**-0,12**	**0,39**	**-31%**	**-0,68**	**0,01**
Gas_distance	0,19		0,00	0,19	-1%	-0,02	0,20

Table 29: The estimate, z-value and p-value of the full data model, including all (not completely separated) variables. The variables below the significance level (p ≤ 0.10) are bold.

	Estimate	z value	Pr(>|z|)
(Intercept)	-20,34	-0,03	0,98
klic_polygon_size	0,00	-0,60	0,55
klic_type_workHovenierswerkzaamheden	-0,21	-0,33	0,74
klic_type_workKabels-Leidingen	0,69	1,54	0,12
klic_type_workOverig	-0,37	-0,70	0,48
klic_type_workSlaan/boren	0,01	0,02	0,99
klic_type_workUnknown	**2,20**	**4,96**	**0,00**
evi_age	0,01	1,31	0,19
evi_diameter	**-0,01**	**-6,00**	**0,00**
evi_shape_length	0,00	0,38	0,71
evi_data_quality	**1,58**	**2,30**	**0,02**
sew_diameter	**-1,19**	**-2,81**	**0,00**
gas_distance	**-0,07**	**-1,90**	**0,06**
gas_sideBuilding	0,19	0,37	0,71
gas_sideStreet	0,38	0,81	0,42
heat_diameter	0,34	0,75	0,45
elec_distance	-0,06	-1,33	0,18
tele_distance	0,05	1,35	0,18
cable_distance	0,02	0,59	0,56
asset_length_in_klic	0,00	0,11	0,91
klic_partyElec	-0,05	-0,11	0,91
klic_partyGas	0,29	0,97	0,33
klic_partyHeat	**0,81**	**2,29**	**0,02**
klic_partySewer	**0,42**	**1,63**	**0,10**
klic_partyTele	-0,38	-0,90	0,37
klic_partyWater	0,23	0,68	0,50
evi_materialAC	14,21	0,02	0,99
evi_materialGGIJ	13,58	0,02	0,99
evi_materialHPE	-0,25	0,00	1,00
evi_materialPE	13,16	0,02	0,99
evi_materialPVC	14,77	0,02	0,99
evi_materialST	15,01	0,02	0,99
evi_distance_house	-0,01	-0,59	0,55
evi_intersection_length	0,00	0,20	0,84
sew_distance	-0,03	-0,90	0,37
sew_sideBuilding	**-1,45**	**-2,12**	**0,03**
sew_sideStreet	**-1,50**	**-2,20**	**0,03**
gas_diameter	**1,63**	**2,24**	**0,03**
heat_distance	0,01	0,28	0,78
heat_sideBuilding	0,53	0,91	0,37
heat_sideStreet	**1,05**	**2,39**	**0,02**
elec_sideBuilding	-0,65	-1,16	0,25
elec_sideStreet	**-0,96**	**-1,81**	**0,07**
tele_sideBuilding	**1,00**	**1,91**	**0,06**
tele_sideStreet	**1,86**	**3,68**	**0,00**
cable_sideBuilding	0,17	0,42	0,68
cable_sideStreet	-0,05	-0,11	0,91


TU Delft


Evides
waterbedrijf