Improving Subsurface Asset Failure Predictions for Utility Operators
A Unique Case Study on Cable and Pipe Failures Resulting from Excavation Work

Wijs, R. J.A.; Nane, G. F.; Leontaris, G.; Van Manen, T. R.W.; Wolfert, A. R.M.

DOI
10.1061/AJRUA6.0001063

Publication date
2020

Document Version
Accepted author manuscript

Published in
ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering

Citation (APA)
Wijs, R. J. A., Nane, G. F., Leontaris, G., Van Manen, T. R. W., & Wolfert, A. R. M. (2020). Improving Subsurface Asset Failure Predictions for Utility Operators: A Unique Case Study on Cable and Pipe Failures Resulting from Excavation Work. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, *6*(2), Article 05020002. https://doi.org/10.1061/AJRUA6.0001063

Improving subsurface assets failure predictions for utility operators

A unique case study on cable and pipe failures from excavation works

R.J.A. Wijs[1], G.F. Nane[2] , G. Leontaris[3], T.R.W. van Manen[4] & A.R.M. Wolfert[5]

1. MSc. student, Faculty of Civil Engineering and Geosciences, Delft University of Technology Stevinweg 1, 2628 CN, Delft, The Netherlands, rjawijs@gmail.com,
2. Assistant Professor/Dr.Ir., Department of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mourik Broekmanweg 6, 2628 XE, Delft, The Netherlands; g.f.Nane@tudelft.nl,
3. PhD student/ Ir., Faculty of Civil Engineering and Geosciences, Delft University of Technology Stevinweg 1, 2628 CN, Delft, The Netherlands, g.leontaris@tudelft.nl,
4. Reliability Engineer/Ir., Department of Asset Management, Evides, Rotterdam, Schaardijk 150, 3063 NH, Rotterdam, The Netherlands, Thomas.manen@evides.nl,
5. Professor/ Prof.Dr.Ir , Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628 CN, Delft The Netherlands, r.wolfert@tudelft.nl

ABSTRACT

Utility operators have to rely on predictive analyses regarding the availability of their subsurface assets which highly depend on damages by the increasing amount of excavation works. However, straightforward use of standard statistical techniques, such as logistic regression or Bayesian logistic, does not allow accurate predictions of these rare events. Therefore, in this paper, alternative approaches are investigated. These approaches involve weighting the likelihood as well as over- and under-sampling the data. It was found that these data methods can improve the accuracy of predicting the rare failure events substantially. More specifically, an application based on real data of a Dutch water utility operator showed that: under sampling and weighting improved the balanced accuracy varying between 0.61 and 0.66, whereas the proposed methods resulted in failures predictions between 38% and 58% of the validation dataset. Hence, the proposed methods will enable utility operators to arrive at more accurate forecasts enhancing their asset operation decision making.

Word count abstract: 155

KEYWORDS

Rare event data, Logistic regression, cables and pipe networks, synthetic minority oversampling, weighted sampling, network operator, excavation works, predictive maintenance.

INTRODUCTION

Uncertainty quantification and risk analysis are of paramount importance in all engineering sectors, therefore also in the subsurface utility sector. It is crucial to understand and account for the stochastic nature of underlying processes in the cable and pipe sector, in order to enable enhanced decision making, for example. Furthermore, subsurface utility companies moved their focus towards more pro-active approaches in risk analysis, by using predictive analyses. Engelhardt et al. (2000) and Tscheikner-Gratl (2016), for example, focused on predicting the deterioration state of cables or pipes before rehabilitation is planned. Likewise, Scholten et al. (2013) combined two models, a rehabilitation and pipe failure model in order to predict the long-term performance of rehabilitation strategies for water mains. It should be noted that rehabilitation in the Netherlands is defined by EN 752 as follows: “measures for restoring or upgrading the performance of existing drain and sewer systems” (Tscheikner-Gratl et al. 2016).

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, 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 et al. 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 contains all preventive maintenance activities, concerning all aspects of the network’s assets (Tscheikner-Gratl 2016). Rehabilitation is always planned for the longer term, therefore infrastructure companies moved their focus toward pro-active approaches, using predictive analyses (Engelhardt et al. 2000; Tscheikner-Gratl 2016).

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. This is remarkable because, based on an extensive cooperation between the network operators and other stakeholders, a binding guideline (CROW500) was formed that seeks to prevent cable and pipe damage from excavation works.

In contrast to rehabilitation, planning of repairs is not possible because the failures are unplanned and repairs are often executed almost immediately after failures since cables and pipes have a vital function for a country and its citizens (Tscheikner-Gratl 2016). 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 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 and Han 2013). 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 et al. 2013). 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. This paper aims to address this gap.

Failures or damages are modelled as dichotomous events, where failure or damage is denoted by one and zero denotes non-failure (non-damage). Logistic regression (LR) is, in this setting, often selected as the modeling approach, i.e., Hosmer et al. (2013), Kleinbaum and Klein (2010). Logistic regression accounts for the influence of the so-called independent variables on the probability of a given event, i.e., the probability of failure, and it has been shown to have good performance in general (Ariaratnam et al. 2001). The failure or damage is regarded as the dependent variable.

Predicting the probability of failures is widely applied in the engineering sector. In contrast, in the subsurface utility sector a scarce number of applications appear to have used logistic regression. For example, logistic regression has been applied to relate scouring potential in a channel to certain independent variables in a study conducted by water resource engineers to enable developing a risk-based design (Tung 1985). Furthermore, the likelihood that a particular infrastructure system (sewer) is in a deficient state was predicted by logistic regression in a setting to demonstrate that the use of logistic regression enables decision makers to prioritize what sewer sections should be inspected (Ariaratnam et al. 2001).

The data used in this case study have been provided by Evides Water Company, the second largest water distribution company in the Netherlands, located in Rotterdam. The data have revealed that there were 181 water main failures as compared to 107,500 non-failures, as registered by Evides from 2010 until 2017 in the municipality of Rotterdam. The data on cable and pipe failures from excavation works are therefore very imbalanced. The failures are regarded as a minority, whereas the non-failures as a majority of the data. This phenomenon is often referred to as rare event data or imbalanced data. In practice, numerous engineering sectors, as well as research fields deal with data where the events of interest (failures or damages) are scarce and therefore make the data imbalanced. An extensive list of application domains has been provided by Haixiang et al. (2017). It is noteworthy that none of these reviewed studies have been applied in the subsurface utility sector.

Modelling rare event data has been proven to pose significant challenges to standard statistical techniques. In particular, predicting rare events proves to be a challenging endeavor, since standard methods, such as logistic or Bayesian logistic regression fail to accurately predict rare events (Haixiang et al. 2017). Predicting rare events is challenging due to several reasons. Firstly, general accepted performance metrics, such as accuracy and precision induce bias toward the majority class. Secondly, models treat rare events as noise occasionally, and consider them exceptional patterns in the data space and reversely, noise can be incorrectly regarded as minority patterns. A detailed discussion about the challenges posed by the rare event data can be found in Haixiang et al. (2017).

Numerous approaches have been proposed over the years to adequately model rare event data. The strategies involve resampling techniques, such as over- and under-sampling methods, as well as hybrid methods. Oversampling methods create new minority samples. One of the best known methods is the synthetic minority over-sampling technique (SMOTE), developed by Chawla et al. (2002). Under-sampling methods discard majority (non-event) samples. The simplest method involves random elimination and has been proposed by Tahir et al. (2009). Hybrid methods entail a combination of over- and under-sampling methods. These approaches are usually referred to as data level methods. Other approaches focused on adapting the techniques or algorithms for the imbalanced data. King and Zeng (2001) have proposed logistic regression for rare event data via the maximization of a weighted log-likelihood function. Other methods have been developed for imbalanced data, for example decision trees and neural networks, which are collectively referred to as classification algorithms for imbalanced learning (Haixiang et al. 2017). An exhaustive review of methods is provided in Haixiang et al. (2017).

This study will unveil the challenges of applying standard logistic regression and Bayesian logistic regression to rare event data in the subsurface utility sector. To the authors' best knowledge, logistic regression for rare event data has not been applied in the subsurface utility sector so far. This paper aims to fill this gap in modelling and predicting failures. Moreover, the paper aims to provide guidelines of employing logistic regression with rare event data. Both data and algorithm approaches which accommodate the imbalanced data are considered. The methods are evaluated with respect to standard measures, such as area Under the Receiver Operating Characteristic (ROC) Curve (AUC) and balanced accuracy. Furthermore, since the aim of the study is to predict damages resulting from excavation works, the prediction performance is evaluated on a validation dataset.

The remainder of this paper is structured as follows. Further details on the study design and data collection process are presented. The methodology introduced the modelling approaches and discusses the assumptions employed by the methods. Afterwards, the performance of the various rare event data approaches is compared. Lastly, the concluding section provides the summary, discusses the results and recommends future research.

Study design

Case Study Area

All subsurface utility operators control Critical infrastructure systems (CISs), which indicates that the network's "incapacity or destruction would have a debilitating impact on the defense and economic security of a nations state" (Ouyang 2014, p. 44). One measure to prevent failures are mandatory excavation requests from which risk assessments follow to analyze conflicts between cables and pipes. In 2015 more than 530,000 excavation requests, from which 32,500 failures from excavation works followed were reported in the Netherlands alone (Kabel- en Leiding Overleg 2016), resulting in € 26 million direct and € 100 million indirect damage.

This research has been conducted within the Evides Water Company, the second largest water distribution company in the Netherlands, serving safe and clean drinking water to 2.5 million consumers and businesses in three provinces. Evides only had around 500 pipeline failures in 2016, causing an average unplanned downtime of 6.8 minutes per customer (i.e., household) per year. This research focuses on the municipality of Rotterdam within Evides' Rijnmond area. This is, first of all, due to the availability of other cable and pipe data. Moreover, this is because city centers and old residential areas have a high population and building density, which result in a larger probability of failure from excavation works (Vloerbergh and Beuken 2011).

Data resources and processing

Many aspects were considered in the data collection process. The study mainly focuses on spatial interdependencies, as these are regarded as important for collocated infrastructures when these are considered for rehabilitation or renewal (Islam and Moselhi 2012). Cable and pipe networks are spatial interdependent, since the state of one network can affect the state of another network by a bidirectional relation (Rinaldi et al. 2001; Utne et al. 2011). From an extensive literature review and three expert interviews within Evides, a list of important variables concerning spatial interdependencies has been considered for data preparation and further analysis. The list is included in Table 1. The variables include information about the horizontal position, diameter and wall material. These variables were collected from different data resources, which are described in the following subsections. A commonality between the databases is that these all use Geographical Information System (GIS), whereby location data is available. This enabled linking the various databases to each other.

Excavation data

Each data entry is obtained from an excavation request, which is mandatory by the Kadaster in the Netherlands before any mechanical excavation activity is started (Kadaster, n.d.). An excavation request contains information such as the location, the type of work, the contractor and the client. Three types of requests are distinguished, that is, 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. Furthermore, the Kadaster allows KLIC-requests (Cable and Pipe Information Center) up to a polygon of 500 x 500 meters. For clarification, it should be noted that a KLIC-request is defined as the obligatory request that is done before mechanical excavation takes place. 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 (25,000 m^2) for the KLIC-polygon is set. Figure 1 depicts the KLIC-requests for this study case.

Evides pipes

All network operators possess databases including assets, such as cables or pipes, and so does Evides. Firstly, service connections are removed from the dataset as these are assumed to be rightangled on the distribution cables and pipes, creating a problematic situation when mutual distances between various network types are determined later on. Service connections concern all cables and pipes between the distribution networks and clients' property, both private individuals and companies. Furthermore, the cables or pipes are visualized as 'lines' within GIS, whereby line length can vary from up to 300 'meters' to only a few centimeters. A minimum length of 15 meters has been chosen is set to ensure loose connections at for example crossings are removed.

Other cables and pipes

Data from other network operators are of importance as this study focused on spatial interdependencies between cables and pipes. The municipality of Rotterdam made available a 3D city model to enable multiple parties to use their unique database, including cables and pipes. The availability of data is not self-evident, as cables and pipes data are mostly confidential, aiming to prevent malicious damage. For the analysis the foreign assets' locations, the type of the network and the associated diameter were collected.

Buildings

Furthermore, the nearest buildings were linked to ensure whether the other networks were crossing the service connections. Service connections are relevant as failures often occur on smaller crossing connections. The Kadaster possesses such a database called Basic Registration and Buildings (BAG), which includes all building locations in the Netherlands.

Failures

In this study, the variable of interest, or the dependent variable, of each sample entry is registered as the failure (one) or non-failure (zero) of an Evides pipe due to a third party. Failures are stored in an Evides database. To identify failures from excavation works, network operators need a method to classify various types of failure, as well as the failure date which indicates whether the failure was in a certain period after the excavation request.

Data processing

Each individual data source has been cleaned already prior to the processing of all the databases into a suitable dataset for the study. During the processing, data were filtered if it could not be connected to the other databases.

Data Integration

The most important variables used for linking are the geometry data, possessed by all used databases. failures were linked to the nearest networks within 10 meters. Linking the assets and failures succeeded for all failures. 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.

Second, failures are connected to excavation requests. Where failures are "points", the excavation requests are polygons, whereby a point must be inside the polygon for linking. Furthermore, the failure must have occurred after the excavation request date, but no more than 3 months after. An excavation activity must start within 20 days after application, but not earlier than 3 days after. Considering the duration of maintenance or construction work, the duration of the period may be adapted. The 3-month period follows from an assessment of various maximum periods for connection. Considering the duration of maintenance or construction work, the duration of the period could be adapted. In this way, 256 failures out of the total of 500 excavation failures were connected to an excavation polygon.

Third, all items that followed from the prior linking of assets and failures were connected to excavation requests. The connections are made based on similarities in location and date. As a result, often, multiple pipes were linked to one excavation request, as it is likely in a densely populated urban area such as Rotterdam, that multiple pipes are in an area when excavation polygons are up to 25,000 m^2.

Because multiple pipes (or cables) could be linked to one KLIC-polygon, the criteria for linking 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. On the other hand, network operators should try not to lose too much data.

Once previous links are succeeded, the relation between the different networks is examined. Therefore, a virtual point on the middle point of each Evides pipe within an excavation polygon is created. From that virtual middle point, the mutual distances to the other surrounding networks and buildings is calculated. To prevent misleading calculations of mutual distances, the short “lines” 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.

In this way, 107,500 entries were collected from which only 181 resulted in a failure. Less than 10% of all data was found to be entirely complete, which is explained by the maximum distance that has been set for linking. In other words, only 10% of all streets in the sample contain all assessed networks. Because LR only includes complete samples, empty entries have been imputed. Even though a common approach is to use the average of the available observations for missing data, this study requires a differentf approach. As discussed earlier, not availables (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. Therefore, mutual distance NAs were imputed by 12, whereas 10 meters was the maximum connection range and NA diameters were replaced with 1 (meter). 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. When that happens, the empty samples are labeled 'unknown'.

Note on the case study

The way in which data have been collected is worthwhile discussing, since it has a large influence on the sample set and therefore on the analysis and results. Firstly, there are various manners in which multiple databases can be linked, as all kinds of criteria for the linking can be used, such as linking all intersecting pipes or only the one pipe with the largest intersection and everything in between. This research aims to retain as many unique situations, while considering the percentage of failures within the sample set which resulted in the selected linking method. Secondly, some data were unavailable, for example the vertical position of the cables and pipes, which is very relevant according to literature (e.g., Riley & Wilson, 2006) and experts. Lastly, the validity of the data is questionable, whereby the actual locations are sometimes not corresponding to the data's location. This was also confirmed when the foreign location data were compared to Evides' own data, from which it was found that more than 5% of the compared data deviated more than 0.4m from the comparable data points in the other data source. Less than 75% had the same location data.

PROPOSED METHODOLOGY

This study aims to employ logistic regression in order to predict failures from excavation works. Since logistic regression is not able to cope with rare event data, several approaches have been considered. To overcome the class-imbalance problem, data level and algorithm level techniques can be used (Chawla et al. 2004). The data level technique prepares the data by rebalancing the data before the modelling is done. Examples of re-sampling techniques are under-, over- and hybrid sampling (Chawla et al. 2002 2004; He and Garcia 2009; Xiong and Zuo 2018). At the algorithm level, the logistic regression has been adapted via a weighted log-likelihood function (King and Zeng 2001). In general, at the algorithm level, the costs of misclassifying the classes, i.e. cost sensitive learning, allocates high cost for the rare event by adding a weight, to improve the learning ability of the classifiers (Chawla et al. 2004; He and Garcia 2009; King and Zeng 2001; Xiong and Zuo 2018).

In this study, three distinct approaches were used to model and predict cable and pipe failures from excavation works These approaches have been validated and their predictive performance has been compared in order to determine the best approach for the data at hand. Moreover, characteristics of the data at hand have been emphasized in order to provide guidelines for the cable and pipe sector, as well as other sectors within the construction or maintenance industry.

The implementation and analysis for this study have been done using programming language R.

Theoretical background

Logistic regression

As already described in the introduction section, logistic regression is generally accepted for binary outcome statistics (Hosmer et al. 2013) and has been already applied for network operators (Ariaratnam et al. 2001; Tung 1985). Logistic regression assumes that the dependent variable follows a Bernoulli distribution having only two possible outcomes, 0 or 1, where 1 usually denotes failure and 0 non-failure with the probability

$$Y_i \sim Bernoulli(Y_i|\pi_i) \tag{1}$$

$$P(Y_i = 1) = \pi_i \tag{2}$$

$$P(Y_i = 0) = 1 - \pi_i, \tag{3}$$

for $i = 1, \ldots, n$ observations and where

$$\pi_i = \frac{1}{1 + e^{-X_i\beta}}, \tag{4}$$

where X_i denotes the vector of independent variables, for each observation i and β denotes the vector of parameters. Then $P(Y_i|\pi_i) = {\pi_i}^{Y_i}(1-\pi_i)^{Y_i}$ is the random variable that represents the probability of failure (King and Zeng 2001; Monroe 2017). The parameters are estimated by maximum likelihood, where the log-likelihood function simplifies

$$\ln L(\beta|y) = \sum_{Y_i=1} \ln(\pi_i) + \sum_{Y_i=0} \ln(1-\pi_i)$$
$$= -\sum_{i=1}^{n} \ln\left(1 + e^{(1-2Y_i)X_i\beta}\right). \tag{5}$$

The influence of a number of independent variables on the dependent variable is depicted via a logit transformation. Therefore the model does not require a linear relationship between the independent variables and the dependent variable, as in the linear regression models. It assumes, nonetheless, linearity of independent variables and the log odds. Moreover, the residuals do not need to be normally distributed. The observations are however assumed to be independent. Furthermore, the independent variables should not exhibit multicollinearity. Multicollinearity entails that one independent variable can predict another independent variable with a certain accuracy (Hosmer et al. 2013; Xiong and Zuo 2018).

As mentioned in the introduction section, logistic regression does not perform well with rare event data. Results will be nevertheless provided, for comparison reasons in the results section.

Weighting and under sampling

The first proposed rare event data approach is by employing weighting, as well as under-sampling. This approach addresses therefore the rare event issue both at the data level and at the algorithm level. This method has been developed for rare event data in political science, related social science and public health research, and have been proposed by King and Zeng (2001). A major advantage of the weighting approach is that it is relatively simple to employ. At the algorithm level, instead of maximizing the standard log-likelihood function, as in the regular logistic regression, a weighted log-likelihood function is maximized as in equation 6. Then

$$ln\ L(\beta|y) = -\sum_{i=1}^{n} \boldsymbol{\omega_i} \ln(1 + e^{(1-2Y_i)X_i\beta}) \tag{6}$$

With equation 1, the weights $\boldsymbol{\omega_i}$ can be determined by

$$\boldsymbol{\omega_i} = \omega_1 Y_i + \omega_o (1 - Y_i), \tag{7}$$

where $\omega_1 = \frac{\tau}{\bar{y}}$ and $\omega_0 = \frac{(1-\tau)}{(1-\bar{y})}$, and τ is the population fraction and $\bar{y}$ as the sample fraction (King and 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.

At data level, it is proposed to include two to five times more zeros than ones, “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” (King and Zeng 2001, p. 143). This weighting method has been applied in multiple studies. Similar to King and Zeng (2001), Maalouf et al. (2018) found that weighting has a higher discriminative performance than regular logistic regression. The former predicted wars for political purposes, whereas the latter predicted network intrusions for military networks. Within GIS-based (Geographic Information System) applications, Xiong and Zuo (2018) used the proposed under sampling and prior correction (which is very similar to weighting) to map prospective mineral locations (King and Zeng, 2001). The method has been implemented in the R package `ReLogit`. A disadvantage of the available package for statistical software `R` is that it does not allow for any goodness of fit tests of the models.

SMOTE

The second approach for rare event data is the Synthetic Minority Oversampling Technique (SMOTE), which has been proposed by Chawla et al. (2002). SMOTE addresses the rare event issue at data level. 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 required number of over-sampling determines how many neighbors from the *k* nearest neighbors are randomly chosen. The new samples are generated by taking one vector under consideration and its nearest neighbor, whereby a random point along the line segment between the two points is selected. In this way, a random point within the correct region is selected, which enlarges the minority class, whereby it becomes more general in the sample set (Chawla et al. 2002; He and Garcia 2009). A combination of both, over- and under sampling is recommended, as it 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).

SMOTE has proven to be successful in various applications, such as for mammography, diabetes and oil slicks (Chawla et al. 2002) and because of its success, it has been further improved over the years. For example Borderline-SMOTE, whereby the over sampling is conducted between the borderline minority class samples instead of all minority samples (Han et al. 2005) has been developed. Another example is SMOTE and Tomek, which cleans data by applying Tomek links to the over sampled training set, whereby also majority class examples are removed that form Tomek links (Batista et al. 2004). However, this study applied the basic version of SMOTE. A disadvantage of the SMOTE method is the incapacity to include categorical independent variables, since the synthetic generated data is different than the variable's categories. Nonetheless, SMOTE has been generalized to handle both continuous and categorical data. The algorithm is called SMOTE-NC, Synthetic Minority Over-sampling Technique-Nominal Continuous (Chawla et al. 2002).

Bayesian Logistic Regression

Lastly, Bayesian logistic regression (BLR) was tested. Firstly, the standard Bayesian logistic regression was employed for the entire dataset. Afterwards, Bayesian logistic regression was combined with under sampling. Bayesian logistic regression entails a Bayesian approach to the multivariate logistic regression model. That is, it starts with a prior distribution on the logistic regression parameters. The posterior distribution is then obtained by multiplying the prior with the likelihood.

Bayesian logistic regression naturally compensates for rare event data by adjusting the estimates toward the null hypothesis to reduce the bias in rare event data. If no common pattern is detected within subgroups, Bayesian logistic regression will perform little partial averaging across issues

(DuMouchel 2012). BLR has been applied for rare event data before to assess clinical safety data, such as the occurrence of a specific adverse event and other safety related issues (DuMouchel 2012). A major disadvantage is that this approach entails a very large computational performance as it has a high model complexity (Grzenda 2015). Nonetheless, the results of this study show the limitation of the Bayesian logistic regression and points out the need to consider methods for rare event data, similarly to the logistic regression.

Methodology approach for the study case

One of the assumptions implied by the logistic regression is that the independent variables should not show multicollinearity. If the independent variables are correlated, this poses the issue of multicollinearity, which can be easily tested with the Variance Inflation Factor (VIF). Along with multicollinearity, the dataset is checked 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). Thirdly, in logistic regression it is recommended for the sample size to satisfy the relation

$$Sample\ size = 10 \times \frac{k}{p} \quad (8)$$

where *k* is the number of independent variables and *p* the proportion of 'positive' cases (Peduzzi et al. 1996). The outcome of the sample size is a rule of thumb, which is kept in mind without any further action.

The model selection is a step in the analysis which will help to determine what variables are irrelevant and can be removed, in order to also overcome a too small sample size. Model selection will be done based on goodness of fit test and by employing a stepwise backward elimination procedure based on Akaike Information Criterion (AIC). The goodness of fit of the statistical model is considered, while accounting for the simplicity of the model. Model selection is of importance to prevent the model from being overfitted or underfitted. The former occurs when the model tries to follow noise patterns whereas the latter occurs when the model is not capable to follow the data points tightly enough.

The performance of the model is evaluated firstly using the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), which is a traditionally accepted performance metric in logistic regression. AUC assesses the performance between true positive (sensitivity) and false positive (specificity) error rates (Lee 2000; Swets 1988).

Given the objective to predict rare events on cable and pipe networks, the model is also evaluated from a predictive point of view rather than from a fitting perspective. Therefore, a validation step is undertaken by considering a validation set along with a training set. The training set is used to fit the model, which is afterwards used to make predictions for the variable of interest in the test set. The model predictions can subsequently be compared with the values of the variable of interest in the test set. A standard approach in the validation analysis is to use a k-fold cross validation, which uses k-1 folds for training and the remaining fold for validation (Han et al. 2005; Rodríguez et al. 2010). When k=5, this translates to using 80% of the data for training and 20% of data for testing. The k-fold cross validation typically makes use of randomly selected training and test sets and the procedure can be repeated numerous times. The prediction error can then be averaged over all the training sets to account for the predictive power of the statistical model. Finally, stratified random sampling needs to be applied, in order to ensure that the rare data are equally split over the training set and the validation set.

The output of the validation step is a confusion matrix, which is used to determine the accuracy, kappa, sensitivity and specificity of the model. Cohen's kappa denotes a measure of agreement. Sensitivity accounts for the proportion of the observed failures that were predicted as failures. Specificity denotes the proportion of the observed non-failures that were predicted as non-failures. The sensitivity and specificity determine the balanced accuracy

$$Balanced\ Accuracy\ = \frac{1}{2}(sensitivity \times specificity) \quad (9)$$

The balanced accuracy measures the average accuracy from both the minority and majority class. A high standard accuracy and a low balanced accuracy indicates that the standard accuracy is high because of the classifier distribution (Akosa 2017). Lastly, the sensitivity of both the data and the model is tested. The former depends on the sample size, therefore the performance of the model for samples of different sizes is investigated. Moreover, the sensitivity of the model explores how the performance of the model is affected by the number of independent variables.

RESULTS

The models following from the proposed rare event techniques, that is the weighting, SMOTE, as well as Bayesian logistic regression are compared on various aspects with respect to a standard logistic regression model. The standard model was also used to test the basic assumptions, as well as for the model selection.

Logistic regression

The original dataset that was identified from the literature review and from interviews accounted for 27 independent variables (Table 1), which include 107,000 non-failures and 181 failures. Employing the logistic regression model for the statistical analysis of the original dataset would require almost 160,000 samples according to Peduzzi et al. (1996). Therefore, backward elimination based on Akaike's Information Criterion (AIC) was applied to select the variables that were considered statistically significant. In the end, ten significant variables were left in the model (Table 2), which agreed with the proposed sample size of Peduzzi et al. (1996). The basic model has been tested comparing a model including all independent variables and a model with the 10 significant variables. From the Log Likelihood Ratio, which indicates how much of the data is explained by the model, a Chi-square score of 0.40 followed, which is above the significance level ($p < 0.10$) whereby the null hypothesis is accepted (Table 2). The mutual dependence of the variables, called multicollinearity was tested by the Generalized Variance Inflation Factor (GVIF), whereby all variables with a GVIF larger than 2.5 were removed.

An overall model performance of the logistic regression resulted in an AUC of 0.60, which is regarded as a poor performance and as failing model (Tape, n.d.). Afterwards, the validity of the model was tested by repeated K-fold cross validation for various test train group ratios. It was found that no failure was predicted at all, resulting in a balanced accuracy of 0.50 and specificity of 1.00 for both models, the all-encompassing model and the model with only 10 significant variables included. This finding is similar to the conclusion of Akosa (2017), also for imbalanced data. To improve the balanced accuracy and hence the model's predictive performance, the rare event techniques introduced in the proposed methodology section, are considered.

Weighting and under sampling

By employing the sampling strategy of King and Zeng (2001), a new sample dataset has been constructed. Different ratios of non-event/event have been considered and the results have been compared. For example a ratio non-event/event of 2 means that there are twice as many non-events (zeros) than events (ones or failures). All suggested ratios that are integer numbers were tested (2, 3, 4 and 5 times) and the results are presented in Table 3.

The results are obtained by performing a validation step, where the size of the training set was approximately 80% of the entire original dataset. It can be concluded that the best ratio, which is based on the balanced accuracy resulted from dataset where the ratio non-event/event was four. This represents the data sensitivity. The selected ratio also results in a sample set of 905 samples from which only 182 are selected for the test set. In the test set 37 failures are included (20%). Table 3 also includes the weights used in maximizing the weighted log-likelihood function.

Because of the weighting, the confusion matrix is affected in the desired way. Through the weights, 29 percent moved from true negative to other positions since the (rare) failures are considered more important by the model, as shown in Figure 2. Therefore, failures will be predicted more frequently with weighting rather than without weighting, which increases the sensitivity of the model.

The validation analysis confirmed that the weighted model predicts failures more accurately than the standard logistic regression model. The specificity was 0.94 and the sensitivity was 0.38, meaning that 38% of the failures were accurately predicted. The specificity and sensitivity result in a balanced accuracy of 0.66 and the AUC, following from the ROC was 0.71. In order to investigate whether the model selection for the standard logistic regression has influenced the results, different models, with different sets of independent variables were considered. No noteworthy differences were found when models with different included variables were considered.

SMOTE

With SMOTE, the dataset will be adjusted by over- and under sampling before the method (presented in the subsection methodology approach for the study case) is employed. Hereby, it is important to realize that the ratio non-failure versus failure should not flip over as this would be opposite to the real situation. Therefore, the non-failure versus failure ratio should be at least one and this is also recommended by Chawla et al. (2002). In Table 4, 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 are included in the sample set. The sample sets that were balanced perfectly (1.00) are bold.

For the various ratios, the resulting AUC of the model has been computed. The AUC metric depends, 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 5 covers all the resulting AUC values for all possible combinations of under- and over- sampling. The smallest AUC values is 0.58, whereas the largest AUC values is 0.72. This is attained when the minority class is 200% oversampled, whereas the majority class is under-sampled 250%.

Without 'flipping' the dataset's balance and considering the AUC, 200% under sampling and 100% over sampling were selected for the modelling, resulting in an equally balanced training set of 604 samples. To validate the model's performance based on the rare event sampling, a validation analysis was also performed. Whereas the training set is balanced, the exceptional quality of SMOTE is that the validation set reflects the real situation with more than 21,000 non-failures and only 31 failures included (0.15%).

From the validation analysis, an AUC of 0.74 was found. The K-fold cross validation gave a specificity of 0.63 and a sensitivity of 0.58, meaning that 52 failures out of 90 were accurate predicted. Together, the balanced accuracy of the SMOTE model is 0.58.

Bayesian Logistic Regression

Furthermore, Bayesian logistic regression (BLR) has been tested on the entire dataset, whereby all 107,500 non-failure observations were included. It was found that there was no noteworthy difference between the results of standard logistic regression and Bayesian logistic regression on the predictive performance. This means that the balanced accuracy was also 0.50, whereas the sensitivity was zero.

As a consequence of the low predictive performance, the BLR model was tested on a smaller sample set, similar to the weighted model as this did also increase the predictive performance of the standard logistic regression model. Once this more balanced sample set of the weighted model is used (4:1 non-failure/failure ratio) for the BLR model, the predictive accuracy increases. The K-fold cross validation step resulted in an increased balanced accuracy of 0.60 and a sensitivity of 0.24.

Models comparison

Considering logistic regression as the first statistical approach enables the comparison of the four models with respect to the standard performance measures, such as AUC, specificity, sensitivity and balanced accuracy. Comparing these results supports decision making on what model should be used for predicting failures resulting from excavation works. It is important to realize that all models included the same independent variables, namely the 10 variables found through the model selection. Using the same variables is essential to compare the models.

Table 6 contains these results for all the employed methods. Firstly, with respect to the P-values of the individual variables, the SMOTE and weighted model perform very well, with values equal to 0.02 and 0.04 respectively. A disadvantage of the R package for weighting is the disability to perform goodness of fit tests on the model, whereby it becomes more complicated to compare it to other models.

As this study aimed to accurately predict cable and pipe failures from excavation works, the validating tests are considered most important. The standard logistic regression model, as well as the Bayesian logistic regression model were found to have a balanced accuracy of 0.50, indicating no predictive accuracy at all for failure. Therefore, the SMOTE, the weighted and under sampled BLR models, which perform better than the other two standard models on most aspects are compared.

The SMOTE model was able to accurately predict most failures with a sensitivity of 0.58. Conversely, it has the worst specificity, with 0.63, meaning 37% of all non-failures are predicted as failures. The weighted model under sampled to a 4:1 ratio has a good specificity whereas it predicts 94% of the non-failures correctly. However, this model predicts failures less accurate than the SMOTE model as the sensitivity is 0.38. Lastly, the under samples BLR model has the best specificity (0.97) but the worst sensitivity (0.28).

When looking at the 'overall' score, the balanced accuracy, the models score quite similar within a range from 0.60 to 0.66. Based on a subsurface utility operator's requirements, the most preferred model can be selected. If preventive measures for a subsurface utility operator are relatively simple and cheap and the cost of failure is large, then the SMOTE model is recommended. On the other hand, when precautionary actions are expensive and complicated it is recommended to use the under sampled BLR model. Therefore none of the models is pointed out as the 'best' model, under any circumstance.

CONCLUSION

Over the past years, network operators have moved their focus towards pro-active approaches. Despite the initiative, they were not able to accurately predict excavation failures for unique situations because these failures are rare events. For other sectors, techniques to handle rare event data were already developed and applied. Therefore, rare event data techniques are proposed to network operators in order to enhance the predictive power of the logistic regression models, that are used to predict excavation failures. To overcome the class-imbalance problem, rare event approaches at data and algorithm level have been tested.

The proposed method has been applied in a test case concerning predictive modelling for cable and pipe failures from excavation works in Evides, a water distribution company in The Netherlands. At data level, it was found that the application of SMOTE did increase the balanced accuracy of the model by 0.11 as compared to a model based on the initial data. At the algorithm level, combined with under-sampling, weighting was tested and found to improve the balanced accuracy to 0.66. The under sampled BLR model has a balanced accuracy of 0.62.

It should be mentioned that the applied techniques which handle rare event data (weighting and SMOTE) have been developed in 2001 and 2002. More advanced techniques have been developed over the past years which could improve the predictive power of logistic regression models even further. An exhaustive overview of all (recent) rare event data techniques has been published by Haixiang et al. (2017). However, the application of the methods in this case study demonstrates the potentials of logistic regression modelling with rare event approaches.

Employing LR revealed interesting insights into the effect of spatial interdependencies on the probability of failure due to excavation works. Two variables were found to influence the probability of failure from excavation works the most. Firstly, emergency KLIC-requests influence the probability of failure the most. However, it is not startling that immediate repairs increase the probability of failure more than planned maintenance, since the latter enables one to prepare for ease. Secondly, the distance to telecom cables, especially on the building side, also increases the probability of failure considerably. With this respect, it is expected that crossing service connections which are closer to the surface cause the increased probability of failure.

Another interesting yet expected finding of this study is the statistical insignificance of the age of pipes, which is found in many studies concerning interdependent critical infrastructures (e.g., Atef and Moselhi 2014; Hokstad et al. 2012) to be a statistically significant variable for failure prediction. Nonetheless, for our case study, it is somewhat to be expected that pipes' age is not expected to be of significant influence for failures due to excavation works, since most mechanical equipment is powerful and will cause damage regardless the pipe's age.

Finally, this case study also entail a number of limitations. First of all, despite the novelty of methods in the setting of network operators, the employed sampling techniques are fairly standard. More advanced, recent, techniques might improve the predictive performance of the methods; as mentioned beforehand, a good overview of the most recent developments is included in Haixiang et al. (2017).

Furthermore, this study reveals that parties are using emergency KLIC-requests above average. An emergency KLIC-request should, in principle, only be used when excavation work is so urgent that it cannot wait. This could indicate unnecessary use of the requests, which probably occurs because one can start excavation immediately instead of waiting for three days. Currently, emergency KLIC-requests can be used in areas of up to 250,000 m^2 meters. The authors recommend that the issue of whether emergency KLIC-requests that apply to polygons with areas of up to 250,000 m^2 be revisited to determine whether they serve an useful purpose. Network operators can probably determine, within a much smaller area, where a failure has occurred. Therefore, it would be advisable to 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 meters is drawn around it.

Furthermore, it is recommended to further study the effect of altering the outcome from failure or non-failure into a numerical value and the implementation of possible consequences. In this way the outcome indicates the ‘size’ of the probability, whereas it is clear obvious that, e.g., 0.75 indicates a larger probability than 0.51. In the current study, both examples are indicated similarly, namely as failure. Moreover, if possible consequences would be also accounted for, a complete overview of the overall risk analysis would emerge.

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 is expected that crossing the service connections, which are closer to the surface causes the high probability of failure. Adjusting the distance from telecom cables to houses could prevent a lot of failures.

DATA AVAILABILITY

All data and models are proprietary or confidential in nature. All statistical code used during this study is available from the corresponding author.

ACKNOWLEGEMENTS

The authors would like to thank Evides water company for providing the dataset and their contribution during preparation of the dataset used in this study. The contribution of the municipality of Rotterdam that provided the data (Rotterdam3D) is deeply appreciated.

REFERENCES

Akosa, J. (2017, April). Predictive accuracy: a misleading performance measure for highly imbalanced data. In *Proceedings of the SAS Global Forum* (pp. 2-5).

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.

Atef, A., & Moselhi, O. (2014). Modeling spatial and functional interdependencies of civil infrastructure networks. In *Pipelines 2014: From Underground to the Forefront of Innovation and Sustainability* (pp. 1558-1567).

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter - Special Issue on Learning from Imbalanced Datasets*, *6*(1), 20–29.

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.

DuMouchel, W. (2012). Multivariate Bayesian Logistic Regression for Analysis of Clinical Study Safety Issues. *Statistical Science*, *27*(3), 319–339.

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. J. M., Saitua, R., & Visser, N. (2016). *Investeren in de infrastructuur: trends en beleidsuitdagingen*. Eib, Economisch Instituut voor de Bouw.

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

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.

Han, H., Wang, W., & Mao, B. (2005). Borderline-SMOTE : A New Over-Sampling Method in Imbalanced Data Sets Learning. In D. Huang, X. Zhang, & G. Huang (Eds.), *Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science* (p. Notes in Computer Science, Vol 3644). Berlin, Heidelberg: Springer.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Hokstad, P., Utne, I. B., & Vatn, J. (2012). *Risk and interdependencies in critical infrastructures*. Springer London.

Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied logistic regression* (Third). New York: Wiley.

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

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

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

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

Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression: A Self-Learning Text* (Third). New York: Springer.

Lee, S. S. (2000). Noisy replication in skewed binary classification. *Computational Statistics and Data Analysis*, *34*(2), 165–191.

Maalouf, M., Homouz, D., & Trafalis, T. B. (2018). Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods. *Computational Intelligence*, *34*(1), 161–174.

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

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

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

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*. Olympia, WA: Washington State Dept. of Ecology.

Rinaldi, S. M., Peerenboom, J. P., & Kelly, T. K. (2001). Identifying, understanding, and analyzing critical infrastructure interdependencies. *IEEE control systems magazine*, *21*(6), 11-25.

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, *32*(3), 569-575.

Scholten, L., Scheidegger, A., Reichert, P., & Mauer, M. (2013). Strategic rehabilitation planning of piped water networks using multi-criteria decision analysis. *Water Research*, *49*, 124–143.

Swets, J. A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science*, *240*(4857), 1285–1293.

Tahir, M. A., Kittler, J., Mikolajczyk, K., & Yan, F. (2009, June). A multiple expert approach to the class imbalance problem using inverse random under sampling. In *International Workshop on Multiple Classifier Systems* (pp. 82-91). Springer, Berlin, Heidelberg.

Tape, T. G. (n.d.). Plotting and Intrepretating an ROC Curve. Retrieved August 1, 2018, from http://gim.unmc.edu/dxtests/ROC2.htm

Tscheikner-Gratl, F. (2016). *Integrated Approach for Multi-Utility Rehabilitation Planning of Urban Water Infrastructure: Focus on Small and Medium Sized Municipalities*. innsbruck university press.

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 Mill, B. P. A., Gooskens, B. J. F., Noordink, M., & Dunning, B. R. (2013). Evaluatie Wion. Den Haag. Publisher: Kwink Groep.

Vloerbergh, I. N., & Beuken, R. H. S. (2011). *Levensduur van leidingen*. Nieuwegein. Publisher: BTO 2011.057.

Xiong, Y., & Zuo, R. (2018). GIS-based rare events logistic regression for mineral prospectivity mapping. *Computers and Geosciences*, *111*(September 2017), 18–25.

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

TABLES

Table 1: Coefficient estimates, z-values and p-values of the full data model, including all (not completely separated) variables.

Name of variable	Category	β coef.	z value	Pr(>|z|)
(Intercept)		-20,34	-0,03	0,98
Size of KLIC polygon		0,00	-0,60	0,55
Type of KLIC request (everything else than *Emergency* is a regular request)	Gardening	-0,21	-0,33	0,74
	Cables and pipes	0,69	1,54	0,12
	Other	-0,37	-0,70	0,48
	Piling/drilling	0,01	0,02	0,99
	Emergency	**2,20**	**4,96**	**0,00**
Age Evides pipe		0,01	1,31	0,19
Diameter of the (own) pipe		**-0,01**	**-6,00**	**0,00**
Shape length Evides pipe (virtual length)		0,00	0,38	0,71
Difference between the two databases		**1,58**	**2,30**	**0,02**
Diameter of the sewer pipes		**-1,19**	**-2,81**	**0,00**
Distance to gas pipes		**-0,07**	**-1,90**	**0,06**
Gas side	Building	0,19	0,37	0,71
	Street	0,38	0,81	0,42
Diameter district heating		0,34	0,75	0,45
Distance to electricity cables		-0,06	-1,33	0,18
Distance to telecom cables		0,05	1,35	0,18
Distance to cable cables		0,02	0,59	0,56

Total length of Evides pipes in KLIC polygon		0,00	0,11	0,91
KLIC requested by	Electricity	-0,05	-0,11	0,91
	Gas	0,29	0,97	0,33
	District heating	**0,81**	**2,29**	**0,02**
	Sewer	**0,42**	**1,63**	**0,10**
	Telecom	-0,38	-0,90	0,37
	Water	0,23	0,68	0,50
Material Evides pipe	AC	14,21	0,02	0,99
	GGIJ	13,58	0,02	0,99
	HPE	-0,25	0,00	1,00
	PE	13,16	0,02	0,99
	PVC	14,77	0,02	0,99
	ST	15,01	0,02	0,99
Distance to buildings		-0,01	-0,59	0,55
Intersection length of Evides pipe in KLIC polygon		0,00	0,20	0,84
Distance to sewers		-0,03	-0,90	0,37
Sewer side	**Building**	**-1,45**	**-2,12**	**0,03**
	Street	**-1,50**	**-2,20**	**0,03**
Diameter gas pipe		**1,63**	**2,24**	**0,03**
Distance to district heating		0,01	0,28	0,78
District heating side	Building	0,53	0,91	0,37
	Street	**1,05**	**2,39**	**0,02**
Electricity side	Building	-0,65	-1,16	0,25

	Street	**-0,96**	**-1,81**	**0,07**
Telecom side	**Building**	**1,00**	**1,91**	**0,06**
	Street	**1,86**	**3,68**	**0,00**
Cable side	Building	0,17	0,42	0,68
	Street	-0,05	-0,11	0,91

Note 1: The variables below the significance level ($p \leq 0.10$) are in bold.
Note 2: AC (asbestos cement), GGIJ (grey cast iron), HPE (Hard polyethylene), PE (polyethylene), PVC (polyvinyl-chloride), ST (steel).

Table 2. Ten variables selected for inclusion in models with corresponding P-values and GVIF as followed from the basic model.

Name of variable	**Category**	**β coef.**	**Pr(>|z|)**	**GVIF^(1/(2*Df))**
Type of KLIC-request	Regular	0,68	0,13	1,20
	Emergency	2,22	0,00	1,20
Diameter of the (own) pipe		-0,01	0,00	1,08
Difference between the two databases		1,46	0,03	1,02
Diameter of the sewer pipes		-0,68	0,01	1,67
Distance to gas pipes		-0,02	0,20	2,17
Excavation work on type	District heating	0,81	0,10	1,04
	Sewer system	0,43	0,30	1,04
District heating side	Building	-0,61	0,08	1,67
	Street	-0,70	0,00	1,67
Electricity side	Building	0,33	0,04	1,60
	Street	0,70	0,01	1,60
Telecom side	Building	-0,90	0,00	1,48
	Street	-0,93	0,00	1,48
Material	Polyethylene	-1,18	0,20	1,02
	Steel	0,73	0,30	1,02

Note: The values of the categorical variables shown in the table are the most extreme values.

Table 3: Weights following from non-failure/failure ratio and corresponding AUC and balanced accuracy.

Ratio non-event / event	Weight		AUC	Balanced accuracy
	Event (1)	Non-event (0)		
2	0.005	1.50	0.69	0,65
3	0.006	1.33	0.69	0.64
4	0.008	1.25	0.76	0.66
5	0.010	1.20	0.66	0.60

Table 4. Non-failure/failure ratio of sample set for different over- and under-sampling percentages.

Note: The sample sets that are perfectly balanced are in bold.

	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

Table 5. Area Under Curve (AUC) for 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		0,63	0,68	**0,72**	0,67
300	0,64	0,64	0,67	0,67	0,69

Note: The sample sets that are balanced are in bold.

Table 6. The five assessed alternatives compared. Above the dotted line are standard and goodness of fit tests, underneath is validation.

Model / test	**Full data**	SMOTE	Weighted	BLR	Under sampled BLR
Average P-values	0.08	0.02	0.04	0.28	0.42
Average |z-score|	|2.64|	|3.12|	NA	|1.85|	|1.16|
LLR (Chi squared)	0.40	6E-11	NA	0.37	0.41
Coefficient determination	0.09	0.07	NA	0.01	0.14
AIC	2070	795	434	1950	656
AUC of ROC	0.60	0.74	0.70	0.72	0.74
Specificity	1	0.63	0.94	1	0.97
Sensitivity	0	0.58	0.38	0	0.28
Balanced accuracy	0.50	0.61	0.66	0.50	0.62

Double-spaced list of figure captions

1. **Figure 1.** The KLIC-requests for the Evides study case in the Rotterdam area. Adapted from Evides (2017).
2. **Figure 2.** Result of weighting the model; (a): because of weighting, 29% of the predictions moved away from true negative. (b): expected value of the weighted model is larger.

Double-spaced list of table captions

1. **Table 2:** The estimate, z-value and p-value of the full data model, including all (not completely separated) variables. The variables below the significance level ($p \leq 0.10$) are bold.
2. **Table 3.** The ten variables that were selected to be included in the models with the corresponding P-values and GVIF as followed from the basic model. The values of the categorical variables shown in the table are the most extreme values.
3. **Table 3.** The weights following from the non-failure/failure ratio and the corresponding AUC and balanced accuracy.
4. **Table 4.** The non-failure/failure ratio of the sample set for different over- and under-sampling percentages.
5. **Table 5.** The Area Under Curve (AUC) for the various over and under-sampling percentages.
6. **Table 6.** The five assessed alternatives compared. Above the dotted line are standard and goodness of fit tests, underneath is validation.