Pipe failure modelling for water distribution networks using boosted decision trees

Winkler, Daniel; Haltmeier, Markus; Kleidorfer, Manfred; Rauch, Wolfgang; Tscheikner-Gratl, Franz

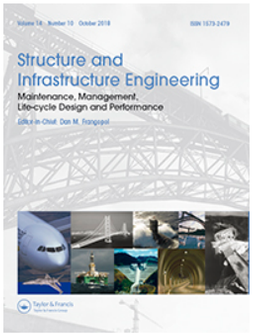
DOI
10.1080/15732479.2018.1443145

Publication date
2018

Document Version
Final published version

Published in
Structure and Infrastructure Engineering

Citation (APA)
Winkler, D., Haltmeier, M., Kleidorfer, M., Rauch, W., & Tscheikner-Gratl, F. (2018). Pipe failure modelling for water distribution networks using boosted decision trees. *Structure and Infrastructure Engineering*, *14*(10), 1402-1411. https://doi.org/10.1080/15732479.2018.1443145



Taylor & Francis
Taylor & Francis Group

Structure and Infrastructure Engineering
Maintenance, Management, Life-Cycle Design and Performance

ISSN: 1573-2479 (Print) 1744-8980 (Online) Journal homepage: http://www.tandfonline.com/loi/nsie20

Pipe failure modelling for water distribution networks using boosted decision trees

Daniel Winkler, Markus Haltmeier, Manfred Kleidorfer, Wolfgang Rauch & Franz Tscheikner-Gratl

To cite this article: Daniel Winkler, Markus Haltmeier, Manfred Kleidorfer, Wolfgang Rauch & Franz Tscheikner-Gratl (2018) Pipe failure modelling for water distribution networks using boosted decision trees, Structure and Infrastructure Engineering, 14:10, 1402-1411, DOI: 10.1080/15732479.2018.1443145

To link to this article: https://doi.org/10.1080/15732479.2018.1443145

Published online: 27 Feb 2018.

STRUCTURE AND INFRASTRUCTURE ENGINEERING, 2018
VOL. 14, NO. 10, 1402–1411
https://doi.org/10.1080/15732479.2018.1443145

Taylor & Francis
Taylor & Francis Group

Pipe failure modelling for water distribution networks using boosted decision trees

Daniel Winkler[a], Markus Haltmeier[b], Manfred Kleidorfer[a], Wolfgang Rauch[a] and Franz Tscheikner-Gratl[c]

[a]Unit of Environmental Engineering, University of Innsbruck, Innsbruck, Austria; [b]Department of Mathematics, University of Innsbruck, Innsbruck, Austria; [c]Sanitary Engineering, Delft University of Technology, Delft, The Netherlands

ABSTRACT

Pipe failure modelling is an important tool for strategic rehabilitation planning of urban water distribution infrastructure. Rehabilitation predictions are mostly based on existing network data and historical failure records, both of varying quality. This paper presents a framework for the extraction and processing of such data to use it for training of decision tree-based machine learning methods. The performance of trained models for predicting pipe failures is evaluated for simple as well as more advanced, ensemble-based, decision tree methods. Bootstrap aggregation and boosting techniques are used to improve the accuracy of the models. The models are trained on 50% of the available data and their performance is evaluated using confusion matrices and receiver operating characteristic curves. While all models show very good performance, the boosted decision tree approach using random undersampling turns out to have the best performance and thus is applied to a real world case study. The applicability of decision tree methods for practical rehabilitation planning is demonstrated for the pipe network of a medium sized city.

ARTICLE HISTORY
Received 4 July 2017
Revised 20 November 2017
Accepted 24 November 2017

KEYWORDS
Deterioration; rehabilitation; decision support systems; water supply; statistical models; environmental engineering

1. Introduction

Deterioration models predicting pipe failure play a major role in planning and decision support processes for water distribution system asset management, helping to prioritise system rehabilitation actions (Martins, Leitão, & Amado, 2013). The ability to make a prediction about the remaining service life of a technical asset provides valuable information for optimal prioritisation of maintenance, rehabilitation or replacement of assets (Syachrani, Jeong, & Chung, 2013). Solving the problem of forecasting and predicting the future state of an asset implicitly or explicitly implies a theoretical model of the complex process of pipe deterioration (Puz & Radic, 2011). An extensive amount of factors (Salehi, Jalili Ghazizadeh, & Tabesh, 2017) affect this process, which makes the prediction when a pipe will fail a difficult task (Ana & Bauwens, 2010).

The physical mechanisms that lead to pipe breakage are very complex and thus not fully graspable by existing physical models (Kleiner & Rajani, 2001). At the moment, these models treat only a small amount of influencing factors at a time, consider only a limited description of the physical deterioration processes or are applicable only for a certain kind of pipe material or failure type (Sorge, 2006). Wilson, Filion, and Moore (2017) provide an extensive overview of existing physical models. The main limitation for application of these models is their extensive need for network, condition and environmental context data. Accumulation of these data is only justifiable for large water mains with costly consequence of failure (Kleiner & Rajani, 2001). While the ideal, complete and open available data-set, the so-called 'transparent infrastructure' (Tscheikner-Gratl, 2016), seldom exists, the lack of available data in the necessary quality exacerbates this situation.

Given the difficulties of applying deterministic physical models and obtaining accurate results, statistical models have been developed (Ana & Bauwens, 2010). They are used to quantify the structural deterioration of water distribution pipes based on analysing various levels of historical data (Shahata & Zayed, 2012). Scheidegger, Leitão, and Scholten (2015) provide a good overview of the statistical models used, (Kleiner & Rajani, 2001; Martins et al., 2013; Osman & Bainbridge, 2011; Tscheikner-Gratl, 2016) compare the strengths, weaknesses and limitations of those statistical models. Most of the models use different strategies to handle scarce data situations (Scholten, Scheidegger, Reichert, & Maurer, 2013), so even for limited data availability deterioration models can give valuable information, when the user acknowledges its limitations. Still data issues are a recurring nuisance throughout the statistical modelling process. Tscheikner-Gratl, Sitzenfrei, Rauch, and Kleidorfer (2016) provide a good overview on these issues (e.g. data inconsistency or gaps in data) together with overall recommendations to overcome or at least minimise their occurrence.

Another modelling category are artificial intelligence models (e.g. genetic algorithms (Nicklow et al., 2010), neural networks (Tran, Ng, & Perera, 2007) or neurofuzzy systems (Christodoulou & Deligianni, 2010)). These are purely data driven approaches

CONTACT Daniel Winkler daniel.winkler@uibk.ac.at

that enable solving of complex problems without the necessity of detailed explicitly known model assumptions. Therefore, a high amount of data and computational resources are necessary while the model itself stays a 'black box' (Ana & Bauwens, 2010).

In order to overcome the limitations of existing approaches, this paper aims to implement a new approach for water distribution pipe deterioration modelling – the family of decision tree learning methods. The underlying model, intuition, assumptions and trade-offs behind each of the methods are more transparent to the user than in other artificial intelligence models (James, Witten, Hastie, & Tibshirani, 2013). Decision tree learning defines a family of methods in the context of supervised learning (Kotsiantis, 2013). The core idea is to design a recursive partitioning of the training data based on the provided labels. This approach allows to model complex relationships between the individual features of the data, while at the same time the model can easily be interpreted (Quinlan, 1986).

Decision trees have been successfully applied for regression and classification tasks in various fields such as medicine, biology, astronomy or business (Rokach & Maimon, 2014). Despite the above benefits, in its pure form, decision tree learning methods are rarely used in the field of pipe deterioration modelling. Jilong, Ronghe, Junhui, Liang, and Chaohong (2014) applied a decision tree algorithm with a depth of three to predict water supply network faults, including valve damage, faucet damage, pipeline losses, water tank damage and bursting pipes without distinguishing between these damages. Furthermore, they only used 20 fault points without validation which gives the whole approach limited significance. There exist several approaches for sewer networks, but these are only partly comparable since the factors affecting pipe failures in water networks are different from the factors in the sewers. Rokstad and Ugarelli (2015) compared random forest algorithms with statistical deterioration models for sewers and found that random forests are not suitable to estimate condition states. Syachrani et al. (2013) employed a decision tree-based deterioration model for sewer pipes to predict the 'real' age of their pipes, using prior clustering to get slimmer decision trees. Harvey and McBean (2014) apply random forests to predict the structural condition of sanitary sewer pipes. Santos, Amado, Coelho, and Leitão (2017) used the random forest algorithm to predict pipe blockage in sewers. However, random forests constitute only one possibility of ensemble methods for decision tree learning and, moreover, this work is the first to use such methods for modelling pipe failure in water distribution networks.

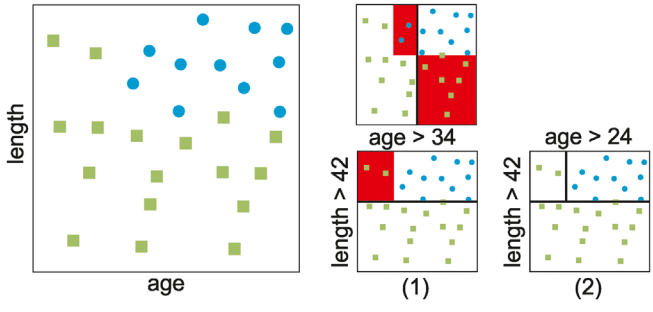


Figure 1. (left) visualises an example predictor space with observations of two classes (blue circle and green square). (1) provides two possible separations of the predictor space. (2) chooses the better classification and applies another separation on the subspaces.

Therefore, this paper benchmarks decision trees and statistically advanced extensions thereof and discusses the individual strengths and the overall performance for an application in pipe deterioration modelling using a water distribution network as case study. For water distribution networks in general *only* the occurrence of pipe bursts and the replacement of pipes are recorded due to the fact that visual inspection in water distribution networks is seldom applicable. This ambiguity in information adds an uncertainty on the exact state of the network, making it a challenge to use the available data to its full extent (Mounce et al., 2017).

This manuscript discusses the current state of the art in decision tree learning algorithms. Special attention is paid to the accurate pre-processing and interpretation of the data, which originates from the historical record of a water distribution network in a medium sized Austrian city. The performance is determined by training the models on one half of the approximately 40,000 pipes in the data-set and testing it on the disjoint other half. The results are evaluated with regard to a practical application in pipe rehabilitation. Using this criterion, the best performing method is selected (in this case boosted decision trees with random undersampling) to predict the current and future states of the pipe network, which can be used to assist tactical rehabilitation planning.

2. Methods

2.1. Decision trees and extensions

2.1.1. Decision trees

Decision trees describe a class of methods to cope with model classification and regression problems in machine learning (James et al., 2013). For the application on pipe deterioration modelling decision trees are employed to detect pipes where failure is imminent.

A major advantage of decision trees is the simplicity and computational efficiency of the method, both in terms of creating the tree as well as applying it to decision-making (Breiman, Friedman, Stone, & Olshen, 1984). Apart from the simple concept, the approach has further interesting advantages for this application. Firstly, the corresponding algorithms are easy to understand and the resulting trees can directly be visualised and interpreted, which allows to immediately perceive and highlight the most influential deterioration factors. This inherent property of the method is used to investigate the trained models, and to compare it to the statistical significant deterioration factors determined with other approaches in literature to provide plausibility to the modelling results. Secondly, decision trees are very suitable for modelling problems with complex relationships between the features and outputs such that they often outperform classical approaches (James et al., 2013). This intrinsic property does not require data augmentation with artificial features that mathematically represent relationships between single features, and can also be used for increasing the complexity of the trained model (Mitchell, 1997).

An example application of a decision tree is provided in Figure 1, which shows a predictor space with observations of two classes

(blue circle and green square). The observations are quantified according to age and length. Step (1) compares two possibilities to separate the predictor space with a rule. On top a rule tries to separate the classes according to an age based rule, nine observations are misclassified. On the bottom the length-based rule misclassifies only two observations. Thus as first rule for the decision tree the length-based rule is chosen. In step (2) the two resulting regions have to be segregated based on the previous decisions. The lower region is already perfectly classified thus no rule is added to the tree. The upper region is split according to age, note that the criterion is different from the one in the previous step. The resulting splitting rules form the final decision tree.

The main concept of decision trees is the stratification of the predictor space into a finite number of subregions. This stratification is expressed as splitting rules, which are hierarchically combined into a tree. The tree construction follows a top-down, greedy approach denoted as recursive binary splitting. Top-down indicates that the starting point (the top of the tree) is the undivided predictor space, where all observations belong to a single region (see Figure 1 left). Thereafter, the method recursively divides the predictor space corresponding to the previous split into two additional regions with every split that is performed (see Figure 1(1)). The greedy nature is due to the creation of the splitting rules, where at every time step the algorithm chooses the best split for this particular decision, ignoring splits that might be better to the overall performance. The recursion stops when the underlying region contains samples that are homogenously classified or a prescribed depth is reached (see Figure 1(2)). For the case of pipe deterioration, the predictor space is the record of all pipes in the system. An example for a binary splitting rule to stratify this space is to test for the type of material, in particular concrete or otherwise. Each of the resulting two regions are then split with an individual splitting rule that separates the region best into failure and non-failure. This process is applied recursively until an exit condition is met. The resulting tree of rules constitutes the decision tree for the prediction model.

The Gini diversity index (GDI) is used as basis for the splitting criterion (James et al., 2013), which expresses the impurity of the node according to:

$$\mathrm{GDI} = 1 - \sum_i p^2(i) \tag{1}$$

where the sum is taken over the available classes i, and $p(i)$ is the observed fraction of predictions with class i in the given node. Thus, a node with a single class has a GDI of 0, whereas for diverse nodes the GDI tend towards 1. The best predictor is chosen by selecting the smallest GDI after the split (Breiman et al., 1984). For the binary classification employed in this pipe deterioration model there are exactly two classes, which means the lower the GDI the better it separates failure from non-failure observations. Weighing the GDI with the node probability results in the node risk, which is used to estimate the importance of the final predictors (MathWorks, 2016).

2.1.2. Bagging

A major disadvantage of plain decision trees is the high variance of the classifier (Hastie, Tibshirani, & Friedman, 2009). To overcome this issue, bootstrap aggregation, in short bagging, is applied, which can be used for reducing the variance in various prediction methods (Breiman, 1996). In the context of decision trees this approach can significantly improve the prediction accuracy.

The basic principle of the method applies the fact that for a set of n independent observations $Z_1, \ldots, Z_n$ with variance σ^2, the variance of its mean $\overline{Z}$ is σ^2/n. Ideally this approach is used to first create independent classifiers $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$ from B separate training sets, which are averaged using:

$$\hat{f}_{\mathrm{avg}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x) \tag{2}$$

Generally, and also in the case of this study, there is no access available to multiple training sets. In such a situation, separate training sets can be created from a single set of observations using bootstrapping. Bootstrapping generates B different training sets by repeatedly taking samples from a single training set, which are used to calculate predictors $\hat{f}^{*b}(x)$. Averaging the predictions is called bootstrap aggregation and defined as (James et al., 2013):

$$\hat{f}_{\mathrm{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \tag{3}$$

Random forests describe an approach based on the principle of bagging that can further improve the accuracy of the decision tree classifier (Breiman, 2001). Similar to bagging a number of trees are grown, however, in the process of growing some additional randomness is introduced to lower the correlation of the individual bagged trees. To gain this property the original decision tree algorithm is altered so that it only allows to choose from a random subset of m predictors at every split. The size of these sets is provided as hyperparameter to the algorithm and is often chosen to be $m = \sqrt{p}$, where p is the overall number of predictors. The set of legitimate predictors is determined randomly for every split based on the size m.

2.1.3. Boosting

Boosting is a conceptually similar method to bagging in the sense that it improves the performance of a predictor by combining multiple classifiers (Hastie et al., 2009). However, the underlying principles are fundamentally different (Freund & Schapire, 1999). The basic concept of boosting is illustrated based on the first boosting algorithm AdaBoost.M1 (Freund & Schapire, 1997), hereafter referred to as AdaBoost.

A weak classifier is a classifier whose error rate is only slightly better than random guessing (Freund, Schapire, et al., 1996). Boosting creates a strong classifier from a list of weak classifiers by training each classifier on a slightly modified version of the dataset. The resulting sequence of classifiers $G_m(x), m = 1, 2, \ldots, M$, is associated with a sequence of weights α_m. The combination of all classifiers to a weighted majority vote results in the strong classifier:

$$G(x) = sign\left(\sum_{m=1}^{M}\alpha_m G_m(x)\right) \tag{4}$$

The classifier weights $\alpha_1, \alpha_2, \ldots, \alpha_M$ are updated during the iterative training algorithm and used to weigh classifiers with a

lower error rate higher than others. Another important aspect of AdaBoost is the additional weighing of the individual observations $(x_i, y_i), i = 1, \ldots, N$ using weights $w_1, w_2, \ldots, w_N$. Those weights are initialised to $w_i = 1/N$, implying that the first classifier is trained as a standard decision tree. After each iteration, the training samples are reweighed, so that misclassified samples have their weights increased; whereas weights of correctly classified samples become decreased (Hastie et al., 2009).

2.2. Data

As explained in the previous section, decision tree learning is based on the statistical evaluation of existing data. When applying such algorithms to real world problems several issues regarding the provided data have to be considered. Those comprise not only apparent properties like the layout and format but also intrinsic properties like data distribution. The necessary pre-processing steps to cope with such issues are described in the sequence.

2.2.1. Data curation

In a first step, the data is pre-processed by removing features that do not have any technical relevance for the model, which is denoted as curation. Apart from pipe enumeration identifiers, no geographical features (e.g. street names or coordinates) are used. Therefore, no spatial interpretation is performed (e.g. using the street name to correlate close pipes with each other or to certain districts of the city). The remaining features are separated into numerical (e.g. age) and categorical features (e.g. material), which require different pre-processing strategies. Since the classification efforts are limited to the decision tree approach and its extensions, the numerical values do not have to be normalised prior training. Categorical features like material or type are transformed from a single feature into a set of Boolean features. This process transforms a categorical feature χ with n possible values into n Boolean features $\chi_k, k = 1, \ldots, n$, where only the feature with the matching value is set to true. Due to the use of MATLAB (MathWorks, 2016) as tool for machine learning, this explicit modification is left to the software by marking categorical features as such.

In the case of pipe material, the data undergoes another pre-processing step. The provided data classifies the pipes as nine different materials, such as cast iron or polypropylene. This classification is improved by considering the fact that some materials changed their properties and thus deterioration patterns significantly due to changes in manufacturing (Roscher, 2000). As such, a finer categorisation of pipe material is used according to Table 1.

For example, according to Table 1 the data for pipes made of CI is separated into (a) pipes built between 1900 and 1930 (CI 1st generation) and (b) pipes built between 1930 and 1970 (CI 2nd generation) (Roscher, 2000). From a classification perspective, these two data-sets represent two completely different materials, such that individual deterioration patterns are inferred. This categorisation is applied for the other materials (DI, ST, PE) listed in Table 1 accordingly.

Table 1. Timetable for pipe materials that changed their deterioration patterns due to different manufacturing processes as classified by Roscher (2000).

Material	Interval boundaries in years			
CI	1900	1930	1970	
DI	1950	1980	2000	
ST	1900	1940	1980	2000
PE	1950	1975	1995	

2.2.2. Failures

The training of the models is based on the data representing individual pipes in the pipe network. Furthermore, the data needs to be augmented with records on damages and repair measures on the pipes, i.e. it needs to be evident when a pipe failed and whether it has been repaired or replaced. Such information can be available as relational database or simply in form of a flat list containing the pipe failures. Regardless of the data storage option, to use the data for training the structure is transformed to a matrix form X expressing individual pipes as rows and features as columns. Consequently, the entry X_{ij} represents the value of feature j for pipe i.

To allow the model to predict a hypothesis, it is necessary to provide an expected output value (failure or no failure) for the training. Considering the current state of the system there are no pipe failures such that the vector $\mathbf{y}$ representing the expected values consists only of zeros indicating no failure. Thus, in the system $(X, \mathbf{y})$ each row $\mathbf{x}_p$ representing a pipe with index p has an expected output value $y_p = 0$. Obviously, the data needs to be extended with recorded pipe failures in order to be able to learn pipe deterioration patterns from the data. This is done by concatenating all recorded pipe failures to the data-set $(X, \mathbf{y})$ such that each added row $(\mathbf{x}_f, y_f)$ has an expected output value $y_f = 1$. Since the network connection represented by a pipe may have failed multiple times in the past, there may be several entries $(\mathbf{x}_{f1}, 1), \ldots, (\mathbf{x}_{fn}, 1)$ for a single connection, each representing an individual pipe failure occurrence.

The input vector $\mathbf{x}_f$ is created by duplicating the pipe feature vector $\boldsymbol{x}_p$ and adjusting some features as follows. Table 2 provides an overview over all features that are used for training. After a pipe failure there are two options, either the pipe is repaired or the pipe is replaced. In both cases the geographical related information does not need to be changed as it is representing the pipe as connection in the network. For the case of the physical features, only if the pipe has been replaced the properties of $\mathbf{x}_f$ like material or diameter has to be changed to match the preceding pipe. Furthermore, all existing data are complemented with the current number of damages (i.e. total failures at pipe location) and damages since replacement (i.e. failures since pipe installation or replacement). Clearly, these values differ only for pipes that have been replaced since the initial installation. This data entries are created by chronologically adding pipe failures $\mathbf{x}_f$ for each pipe $\mathbf{x}_p$ and continuously incrementing both values by 1, starting from 0. If a pipe replacement occurs in this process the counter for damages since replacement has to be reset to 0. To model the age influence on the deterioration the installation date is replaced with either the age at pipe failure or the current age.

2.2.3. Skewed data

If the data-set contains an unbalanced number of samples for the individual classes, the data are called to be skewed (Seiffert, Khoshgoftaar, Van Hulse, & Napolitano, 2010). For example, a provided data-set might be skewed with a ratio of $\approx 1/10$ of failure class to intact samples class. Such a property is problematic

Table 2. Feature vector ***x*** of the data-set with the abbreviations used in the results. The type indicates if the feature is categorical (C) or numerical (N).

	Name	Type	Description
	Failure	C	Indicating pipe failure or not (*y*)
physical	Age	N	The age in years
	Type	C	The type of the pipe, classified as DP (distribution pipe), HC (house connection) and HY (hydrant pipe)
	Diameter	N	The diameter in millimetres
	Pressure	N	The nominal pressure in bar
	Length	N	The length of the pipe in metres
	Material	C	The material of the pipe section, categorised as described in Table 1
geographically derived	HC_Str	N	Number of house connections in the same street section
	HY_Str	N	The number of hydrants in the same street section
	Valves_Tot	N	The number of valves on the pipe
	Valves_St	N	The number of valves in the same street section
historically derived	Failure_Tot	N	The number of total damages recorded on the pipe
	Failure_New	N	The number damages recorded since the pipe has been replaced. If the pipe has never been replaced it coincides with Failure_Tot

for training as a simple classifier $G(x) = 0$, predicting always no failure, would be correct in 90% of the cases. To avoid this discrepancy, the classes for training have to be more evenly distributed, which can be achieved by sampling a subset of the predominant class.

2.2.4. Data subsampling

For the reduction of training data, two approaches known as simple random sampling (SRS) and stratified sampling (Cochran, 2007) are used. SRS is the simplest form of probabilistic sampling where n units out of the N observations in the data-set are selected (Hastie et al., 2009). The n observations are drawn randomly unit by unit with equal chance and at most once. Stratified sampling allows to improve the sampling regarding certain aspects. For a population of N units that is divided into L characteristic subpopulations $N_1, N_2, \ldots, N_L$ such that $N = \sum_{h=1}^{L} N_h$, stratified sampling provides means to represent each subpopulation, called stratum, in the selected sample. If the set of selected observations in each stratum is chosen randomly, the method is called stratified random sampling (Cochran, 2007). Stratified random sampling with proportional allocation is performed, which means that the condition:

$$\frac{n_h}{n} = \frac{N_h}{N} = W_h \tag{5}$$

needs to be fulfilled, where n is the set of sampled observations, n_h is the set of sampled observations in stratum h and W_h is the fraction of the hth stratum. This type of stratified sampling is known as proportionate stratified random sampling (PSRS). In the case of this paper, the pipe material is used as the stratification condition because each type of material has a specific deterioration pattern (Ahmadi, Cherqui, Aubin, & Le Gauffre, 2015). This choice influences the distribution of the subsampled data such that all materials are represented in the training set but is independent from the actual learning process. The importance of the pipe material for the classifier is thus entirely determined by the learning algorithm and not prescribed by this choice.

PSRS is performed on the training data used for the decision tree, random forest and AdaBoost classifiers. This paper furthermore investigates RUSBoost, which has been specifically designed as a variation of AdaBoost that employs random undersampling (RUS) on the data (Seiffert et al., 2010). Due to the fact that the sampling is embedded in the method, the RUSBoost classifier is trained on the entire training data-set, which is an advantage over the other methods that are only trained on a sampled subset. In case of the manual subsampling discussed above only a small percentage of the non-failure class are leveraged to gain a training class ratio of 50:50. Using the above example of a ratio $1/10$ would mean that only 10% of the dominant class would be used for training. RUSBoost improves this drawback by individually undersampling the entire data-set for every weak classifier, such that a larger fraction of the majority class is used for training.

2.3. Case study

The case study, on which the described machine learning approach is applied, is a medium sized city (app. 95,000 inhabitants) in Austria with an overall network length of 851 km with 17,268 house connections (32% of the network length). The failure recordings started already 1983 but the time series recording has gaps (Tscheikner-Gratl, Sitzenfrei, Hammerer, Rauch, & Kleidorfer, 2014). The original network data was of mediocre quality and therefore is enhanced with the help of a data reconstruction method (Tscheikner-Gratl et al., 2016) and divided into street sections to simplify processing. The reconstructed data-set consists of approximately 39,637 pipes with 20 documented properties, including material and length. The available data contains 3743 documented failures, which represent a fraction of 8.63% of all observations. Thus, the data are skewed with a fraction lower 1/10 of failure to intact samples.

The data distribution of the most important network features is visualised in Figure 2. A graphical representation of the pipe pressure distribution has been omitted since 99% of the pipes are recorded with a pressure of .5 MPa. Sixty-nine per cent of the pipes are house connections and 25% distribution pipes. According to this distribution, the data-set contains a high percentage of pipes with small diameter, 72% of the pipes have a diameter less than 50 mm, which corresponds with the high amount of house connections. The provided data contains nine different pipe materials, which are distributed on the network as: 3.46% asbestos cement (AC), 6.92% cast iron (CI), 7.05% ductile iron (DI), .01% glass reinforced plastic (GRP), 1.70% polypropylene (PP), 51.83% polyethylene (PE), 8.43% polyvinyl chloride (PVC), .06% lead (Pb) and 20.54% steel (ST). These materials have been refined according to Table 1 such that 17 different materials are used as input to the models.

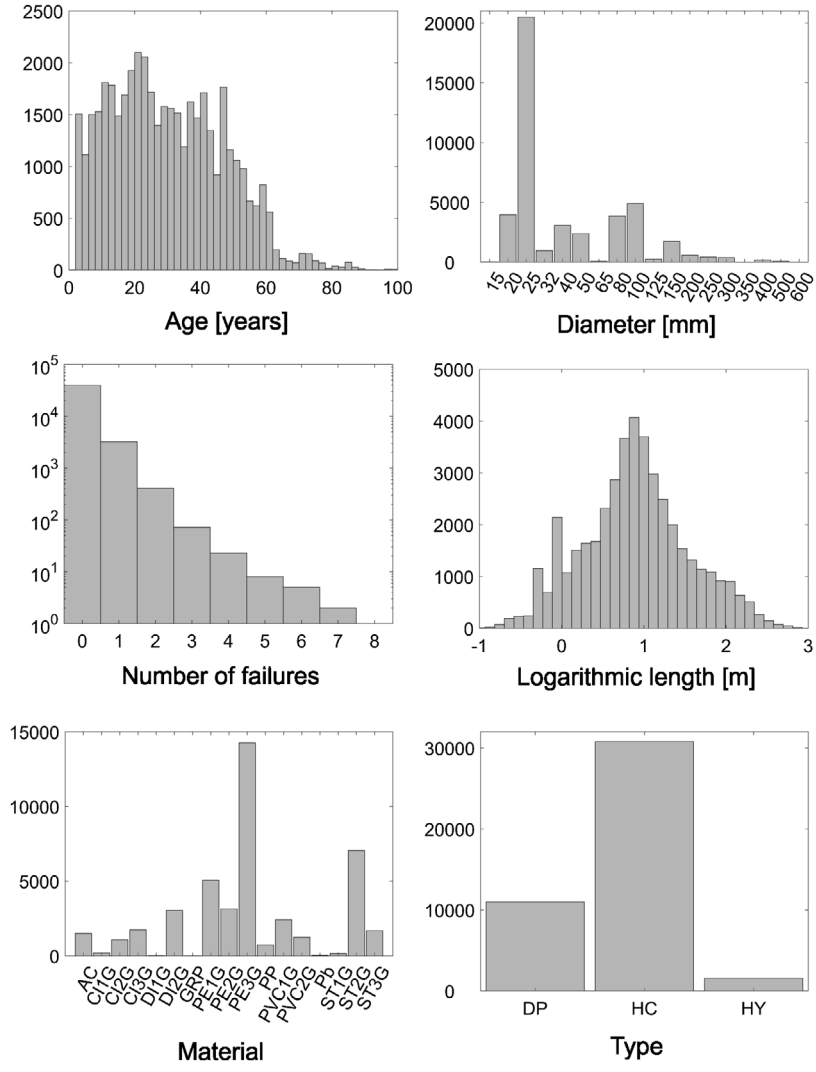


Figure 2. Histograms of the pipe network data. Material appendices 1G, 2G and 3G indicate the sub-classification of materials according to Table 1. The material abbreviations denote asbestos cement (AC), cast iron (CI), ductile iron (DI), glass reinforced plastic (GRP), high impact polypropylene (HIT), polyethylene (PE), polyvinyl chloride (PVC), lead (Pb) and steel (ST). The type abbreviations denote distribution pipe (DP), house connection (HC) and hydrant (HY).

Applying the described pre-processing steps transforms the used data to training and test data for the machine learning algorithms. Note that decision tree learning does not require mathematical processing steps like feature scaling and normalisation, which are necessary, for example, for linear methods. Furthermore, the mathematical combination of features or unary manipulations like exponentiation is not necessary for tree based approaches (Breiman et al., 1984). The 12 features that are selected from the original data-set to train the model for predicting pipe failures in the current system are listed in Table 2.

The learning and testing is conducted with the MathWorks MATLAB Statistics and Machine Learning Toolbox (MathWorks, 2016). Pre-processing of the data is performed using Python in combination with the Python Data Analysis Library (McKinney & Team, 2015).

3. Results

All results for the performance evaluation are created with a 50% holdout rate of test data from the entire data-set resulting in an equal partitioning of training and test data, i.e. the model is trained on one half of the data and validated against the other disjoint half. Increasing the ratio of training to test data did not result in significant improvement of the results. Subsampling of the data are performed only on the training data. Performance evaluation for all methods is executed on the skewed test data, which represents the practical application of the classifier that has to be applied on the entire skewed data-set to model pipe deterioration.

3.1. Classification performance

The performance in terms of predictions is measured by estimating the accuracy, confusion matrix and receiver operating characteristic (ROC) curve. Accuracy is calculated as the fraction of correct predictions to total predictions. The confusion matrix provides more insight by explicitly categorising the predictions according to actual class and predicted class. Each column in the matrix represents instances in a predicted class while each row represents instances in an actual class. Thus, the predictions are separated into true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predictions. Dividing

these metrics by the number of actual observations results in the respective rate, e.g. TP/(TP + FN) results in the true positive rate (TPR).

Table 3 summarises the confusion matrices for the trained models evaluated on the test data. An interesting characteristic of these results is that the TPR of the methods trained on the stratified data is approximately 10% higher than for RUSBoost that is trained on the entire training data. In contrast, RUSBoost has a much lower false positive rate (FPR), which is reflected in an overall accuracy of .96. Due to the skewed nature of the data, the FPR has a much higher contribution on the accuracy measure than the TPR, such that the accuracies of the other methods are significantly lower at .87 (AdaBoost), .89 (Random forest) and .83 (Decision tree).

For classification problems with ensemble methods each DT of the ensemble votes for a specific class, the overall class prediction is then based on a majority vote. The results from the confusion matrix are thus strictly distinguished at a threshold of .5. Apart from the label, predictions comprise scores that describe the probability that the observation belongs to a certain class. This information allows to vary the threshold at which a pipe is classified as damaged, thus trading a decreased false positive rate (FPR) for a lower true positive rate (TPR) and vice versa. As an example, if FPR of 1 is accepted then also a TPR of 1 is trivially achieved by simply classifying all pipes as broken. On the other hand, if only a very low FPR of .01 is accepted the classification model will have a relatively low TPR, containing only pipes where the probability is high enough. This relationship is visualised by the so-called ROC curve, which is created by altering the discrimination threshold and plotting the TPR against the FPR as function of thereof (Fawcett, 2006). The ROC is rated as good when the curve is above the 45° line which represents random guessing, perfect classification is graphically interpreted by the union of two lines corresponding to FPR = 1 and TPR = 1, respectively.

As shown in Figure 3 the ROC curves for the three ensemble methods are quite close with the best characteristic for the RUSBoost method. The legend of the Figure furthermore gives information on the area under the curve (AUC), which is a quantity in the range $0 \leq \text{AUC} \leq 1$ that integrates over the respective ROC functions. As argued in (Hosmer & Lemeshow, 2000), a model that achieves an area under the ROC above .8 is excellent and an AUC higher than .9 is outstanding. This indicates that decision tree learning is well suited for deterioration modelling, as all ensemble methods perform with AUC higher than .9.

Table 3. Confusion matrices for the evaluated methods showing rates in per cent. RUSBoost is significantly different from the other methods by having a lower true positive rate and a significantly lower false positive rate.

		predicted				*predicted*	
RUSBoost		Yes	No	*Random Forest*		Yes	No
actual	Yes	70.19	29.81	*actual*	Yes	80.62	19.38
	No	1.17	98.93		No	9.32	90.68
		predicted				*predicted*	
AdaBoost		Yes	No	*Decision Tree*		Yes	No
actual	Yes	80.75	19.25	*actual*	Yes	79.14	20.86
	No	10.10	89.90		No	16.51	83.49

The predictor importance of the classifier is calculated as summation over the risk changes due to splits of a specific feature. For ensemble methods, this value is accumulated over all weak classifiers. For each feature, an importance value is calculated where high values indicate high relevance for the classification process. The data from Figure 4 meet the practically known relevance factors for deterioration modelling. The most important features for all tested methods are material, age and length. Debón, Carrión, Cabrera, and Solano (2010), Lei and Sægrov (1998) and Tscheikner-Gratl (2016) found in their works to be material, length and diameter to be significant factors, while Giustolisi, Laucelli, and Savic (2006) also chose these pipe features among all available information to model the occurrence of water main bursts.

It is important to note that due to the data dependency of the approach the predictor importance is representative only for this case study and not for pipe deterioration in general. An illustrative example is the influence of pipe pressure. While several point out that pressure is among the important factors studies (Friedl et al., 2012; Ghorbanian, Karney, & Guo, 2016; Salehi et al., 2017), according to Figure 4 it is the least significant of all properties for all tested models. This is explained by the data availability of the case study which did not allow a hydraulic model of existing pipe pressure in the network but only the usage of the nominal pressure of the pipes as a proxy for the pipe material quality, where 99% of the pipes are documented with a nominal pressure of .5 MPa.

Obviously, this data-set is not representative to determine the effect of pressure in pipe failure prediction, which is reflected in the results accordingly. The diversity of pipe diameters in the data-set is high enough to be used as relevant criterion, however, the importance for the DT models is on par with artificial meta properties like the amount of valves and house connections in the same street section. For the case of a single decision tree, Figure 4 shows that the pipe diameter is more important to the model than those properties. This could be explained by the implementation of a categorisation between house connections and distribution pipes, which to a certain degree also are a division between higher and lower diameters. Here again, it is important that this observation is true for the given training data-set and not for deterioration modelling in general.

3.2. Practical considerations

To use any of the trained classifiers for creating a rehabilitation strategy requires to use it on the original data. For this purpose, two steps have to be performed. Firstly, the classifier is trained on the entire data-set with a test holdout of 0, which exploits the information of the entire database. To minimise the risk of overfitting, the classification error is compared to cross-validated classifiers that are trained on the same data. For this purpose, k-fold cross validation partitions the data randomly into k equal sized subsamples. In addition, k classifiers are trained individually using one subsample as test data and the other $k-1$ subsamples as training data. The cumulative error of the prediction serves as estimate for the accuracy of the classifier (Hastie et al., 2009). Two cross-validated classifiers are trained with $k = 2$ and $k = 5$ and compared to the model that is trained on the full data-set. All three models perform very well with a classification

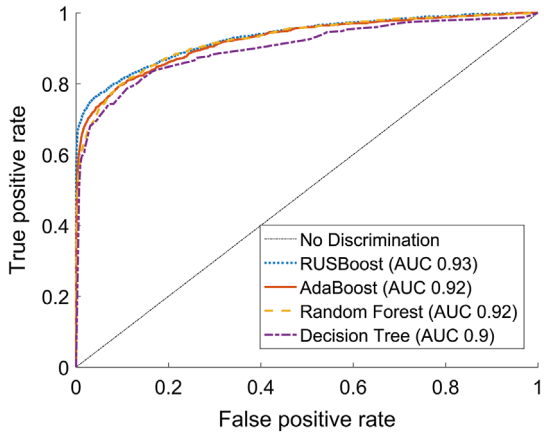


Figure 3. ROC curve for failure pipes, predicted vs. actual response on test set. RUSBoost shows the best performance as it is closes to the ideal classification, a horizontal line with true positive rate 1 for all values of false positive rate. This is also reflected by the highest area under curve (AUC) value of .93.

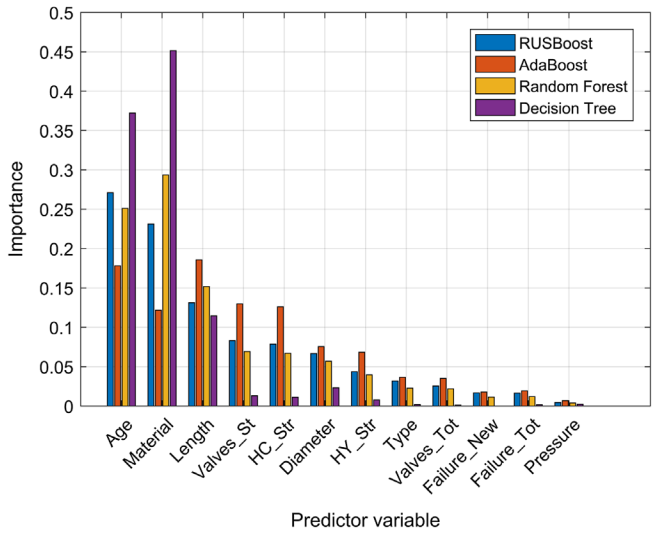


Figure 4. Comparison of predictor importance estimation. The predictors are sorted according to the importance for the first classifier (RUSBoost).

error of .039 for the final model. The cross-validated classifiers perform slightly worse with an error of .040, the standard deviation of the loss ratio of full model to k-fold is .032 for $k = 2$ and .027 for $k = 5$.

The final classifier is used hereafter to predict the failure pipes on a database that has not been augmented with pipe failures, which is basically a registry of all pipes in the current network

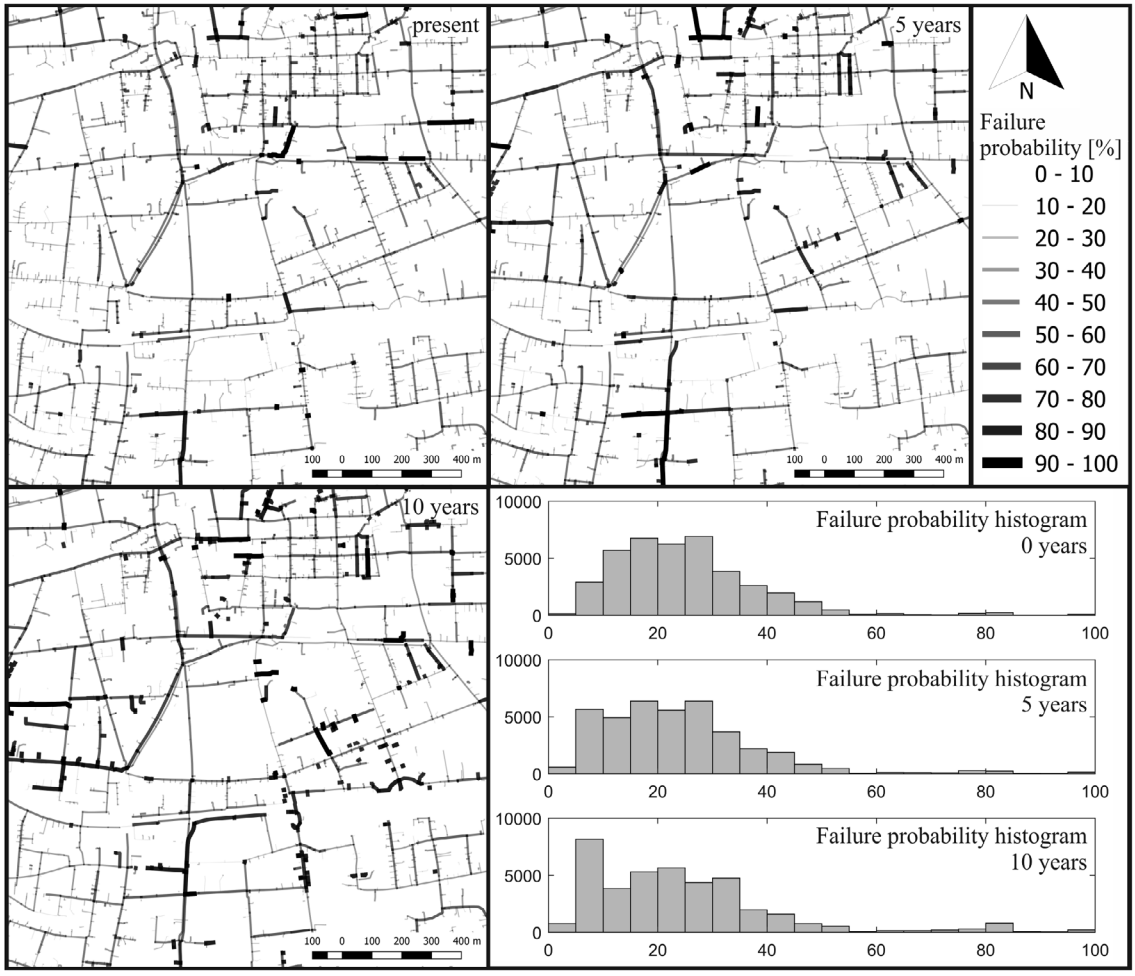


Figure 5. Visualisation of the pipe deterioration prediction for parts of the entire network for the current condition and future in five year steps. The pipe failure probability is visualised by colour intensity. The histograms show the failure probability distribution for the entire network.

that matches the input feature requirement of the trained model. Executing the model on the data determines the pipes that are, according to the model, in a failure state. For this reason, the selection of the model requires careful deliberation of the results from the previous section. According to the confusion matrices, RUSBoost is preferable since it has a very low FPR, which is important for real world rehabilitation to reduce the cost of replacing pipes before the end of their technical service life. The lower TPR compared to other methods won't affect the practical rehabilitation since generally not all detected pipes will be replaced immediately, thus more conservative model matches practically feasible strategies.

As discussed for the ROC curves, the model underlying the binary classifier predicts a class probability. A priority list for rehabilitation measures can thus be created based on the probabilities of belonging to the failure class. Mapping the probability to the geographic location of the pipe allows to create a map of the network with the failure state attached. Figure 5(top left) visualises a part of the city network with the current failure probability colour coded. This allows real world rehabilitation management to prioritise and in consequence inspect and repair clusters of high probability failure pipes.

Alegre and Coelho (2012) propose to predict the network condition in intervals of 5 years to create a tactical plan for rehabilitation. This is modelled by incrementing the pipe age feature by 5 and 10 years. In Figure 5 the deterioration pattern of the pipe network for the same section of the city are shown, furthermore a histogram of the failure probability is included for the entire network for the current state and the predictions in 5 and 10 years. Increasing the pipe age of the system grows the predicted amount of pipes in a failure state. The histograms show that the fraction of pipes with a higher probability increases for future predictions. Combination of the failure probability with geographical information allows visualise the information spatially.

An important observation of the predictions is that the system does not deteriorate monotonically, i.e. the condition of the pipe can get better with higher age. One reason for this effect is that increasing the pipe age creates observations that are outside the domain of the training data, thus the model needs to extrapolate for predictions. Since the most conservative model is chosen, it may predict low probabilities in for such cases. However, although it seems unintuitive, decreasing failure probability is actually observed in reality. This can be explained due to damages that occur during the pipe installation, such that a high failure probability exists in the initial lifetime, as well as the survival bias of older pipes, meaning that the oldest surviving pipes in the data base are in good condition because the ones in bad conditions are already replaced (Sægrov, 2005). This bias could lead to overestimation for the condition of very old pipes so for applying the model in prioritisation the setting of thresholds would be advisable.

4. Conclusions

In this paper, the novel approach of using decision tree learning methods to model water distribution pipe deterioration is proposed. The very good performance of the method (prediction accuracy of .96 and AUC of .93) shows that it can be seen as a good alternative to conventional statistical deterioration models. As a proof of its efficiency, the model was applied in a medium size case study. The pipe network database and failure recordings are transformed into a format that is suitable for machine learning. The problem of skewed data distribution of failure and non-failure observations is handled, and bagging and boosting are applied to overcome the high variance of standard decision tree classifiers.

The performance evaluation of the classifiers using a holdout of 50% for test data reveals outstanding results when applying the performance classification of Hosmer and Lemeshow (2000). Boosted decision trees using random undersampling is found to be the best performing classifier, which is used for the creation of a tactical rehabilitation plan where the model is employed to predict the pipe network state in 5 and 10 years. A further novelty is the inclusion of house connections into the approach, which is still seldom done, but is one of the weak points of a network in terms of failure occurrence.

Future work will include the application and evaluation of the model to different data-sets. Interesting measures are the performance of the approach on these data-sets, and the performance of trained models on different data-sets. A sensitivity analysis with respect to the data distribution will provide information on the generalisation ability of the method. Furthermore, measures will be tested to reduce the influence of the survival bias.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Daniel Winkler http://orcid.org/0000-0003-0131-1559
Markus Haltmeier http://orcid.org/0000-0001-5715-0331
Manfred Kleidorfer http://orcid.org/0000-0002-4001-1711
Wolfgang Rauch http://orcid.org/0000-0002-6462-2832
Franz Tscheikner-Gratl http://orcid.org/0000-0002-2545-6683

References

Ahmadi, M., Cherqui, F., Aubin, J.-B., & Le Gauffre, P. (2015). Sewer asset management: Impact of sample size and its characteristics on the calibration outcomes of a decision-making multivariate model. *Urban Water Journal, 13*(1), 41–56. doi:10.1080/1573062X.2015.1011668

Alegre, H., & Coelho, S.T. (2012). Infrastructure asset management of urban water systems. In A. Ostfeld (Ed.), *Water supply system analysis – Selected topics* (pp. 49–74). Rijeka, Croatia: InTech. doi:10.5772/52377

Ana, E.V., & Bauwens, W. (2010). Modeling the structural deterioration of urban drainage pipes: The state-of-the-art in statistical methods. *Urban Water Journal, 7*(1), 47–59. doi:10.1080/15730620903447597

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140. doi:10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. doi:10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C.J., & Olshen, R.A. (1984). *Classification and regression trees*. Boca Raton, FL: CRC Press.

Christodoulou, S., & Deligianni, A. (2010). A neurofuzzy decision framework for the management of water distribution networks. *Water Resources Management, 24*(1), 139–156. doi:10.1007/s11269-009-9441-2

Cochran, W.G. (2007). *Sampling techniques*. New York City, NY: Wiley.

Debón, A., Carrión, A., Cabrera, E., & Solano, H. (2010). Comparing risk of failure models in water supply networks using ROC curves. *Reliability Engineering and System Safety*, *95*, 43–48. doi:10.1016/j.ress.2009.07.004

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861–874. doi:10.1016/j.patrec.2005.10.010

Freund, Y., & Schapire, R. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence, 14*, 771–780.

Freund, Y., & Schapire, R.E., et al. (1996). Experiments with a new boosting algorithm. In Proceedings of the thirteenth international conference on machine learning (Vol. 96, pp. 148–156), Bari, Italy.

Freund, Y., & Schapire, R.E.R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139. doi:10.1006/jcss.1997.1504

Friedl, F., Möderl, M., Rauch, W., Liu, Q., Schrotter, S., & Fuchs-Hanusch, D. (2012). Failure propagation for large-diameter transmission water mains using dynamic failure risk index. *World Environmental and Water Resources Congress, 2012*, 3082–3095. doi:10.1061/9780784412312.310

Ghorbanian, V., Karney, B., & Guo, Y. (2016). Pressure standards in water distribution systems: Reflection on current practice with consideration of some unresolved issues. *Journal of Water Resources Planning and Management, 142*(8), 04016023.

Giustolisi, O., Laucelli, D., & Savic, D. (2006). Development of rehabilitation plans for water mains replacement considering risk and cost-benefit assessment. *Civil Engineering and Environmental Systems, 23*(3), 175–190. doi:10.1080/10286600600789375

Harvey, R.R., & McBean, E.A. (2014). Predicting the structural condition of individual sanitary sewer pipes with random forests. *Canadian Journal of Civil Engineering, 41*(4), 294–303. doi:10.1139/cjce-2013-0431

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York, NY: Springer. doi:10.1007/b94608

Hosmer, D.W., & Lemeshow, S. (2000). *Applied logistic regression*. Hoboken, NJ: Wiley. doi:10.1002/0471722146

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York, NY: Springer. doi:10.1007/978-1-4614-7138-7

Jilong, S., Ronghe, W., Junhui, P., Liang, C., & Chaohong, X. (2014). *Decision tree classification model in water supply network*. CUNY Academic Works. Retrieved from http://academicworks.cuny.edu/cc_conf_hic/65

Kleiner, Y., & Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: Statistical models. *Urban Water, 3*(3), 131–150. doi:10.1016/S1462-0758(01)00033-4

Kotsiantis, S.B. (2013, April 29). Decision trees: A recent overview. *Artificial Intelligence Review, 39*, 261–283. Netherlands: Springer. doi:10.1007/s10462-011-9272-4

Lei, J., & Sægrov, S. (1998). Statistical approach for describing failures and lifetimes of water mains. *Water Science and Technology, 38*(6), 209–217. doi:10.1016/S0273-1223(98)00582-4

Martins, A., Leitão, J.P., & Amado, C. (2013). Comparative study of three stochastic models for prediction of pipe failures in water supply systems. *Journal of Infrastructure Systems, 19*(4), 442–450. doi:10.1061/(ASCE)IS.1943-555X.0000154

MathWorks (2016). *Statistics and machine learning Toolbox 2016b*. Natick, MA: The MathWorks Inc.

McKinney, W., & Team, P.D. (2015). *Pandas – Powerful Python data analysis toolkit*. Pandas –Powerful Python Data Analysis Toolkit, 1625.

Mitchell, T. (1997). *Machine learning*. McGraw Hill Series in Computer Science. Retrieved from https://profs.info.uaic.ro/~ciortuz/SLIDES/ml0.pdf

Mounce, S.R., Ellis, K., Edwards, J.M., Speight, V.L., Jakomis, N., & Boxall, J.B. (2017). Ensemble decision tree models using rusboost for estimating risk of iron failure in drinking water distribution systems. *Water Resources Management, 31*(5), 1575–1589. doi:10.1007/s11269-017-1595-8

Nicklow, J., Reed, P., Savic, D., Dessalegne, T., Harrell, L., Chan-Hilton, A., Karamouz, M., Minsker, B., Ostfeld, A., Singh, A., Zechman, E. (2010). State of the art for genetic algorithms and beyond in water resources planning and management. *Journal of Water Resources Planning and Management, 136*(4), 412–432. doi:10.1061/(ASCE)WR.1943-5452.0000053

Osman, H., & Bainbridge, K. (2011). Comparison of statistical deterioration models for water distribution networks. *Journal of Performance of Constructed Facilities, 25*(3), 259–266. doi:10.1061/(ASCE)CF.1943-5509

Puz, G., & Radic, J. (2011). Life-cycle performance model based on homogeneous Markov processes. *Structure and Infrastructure Engineering, 7*(4), 285–296. doi:10.1080/15732470802532943

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106. doi:10.1023/A:1022643204877

Rokach, L., & Maimom, O. (2014). *Data mining with decision trees: Theory and applications* (2nd ed.). Singapore: World Scientific Pte. Ltd. doi:10.1007/978-0-387-09823-4

Rokstad, M.M., & Ugarelli, R.M. (2015). Evaluating the role of deterioration models for condition assessment of sewers. *Journal of Hydroinformatics, 17*(5), 789–804.

Roscher, H. (2000). *Zustandsbewertung städtischer Wasserrohrleitungen zur Vorbereitung der Rehabilitation* [Condition assessment of urban water distribution pipes to precede rehabilitation]. ROHRBAU-Kongress 2000 in Weimar. FITR - Forschungsinstitut für Tief- und Rohrleitungsbau Weimar e.V, Ed, Weimar, Germany.

Sægrov, S. (2005). *Care-W computer aided rehabilitation of water networks, 1843390914*. London: IWA.

Salehi, S., Jalili Ghazizadeh, M., & Tabesh, M. (2017). A comprehensive criteria-based multi-attribute decision-making model for rehabilitation of water distribution systems. *Structure and Infrastructure Engineering, 6*, 1–23. doi:10.1080/15732479.2017.1359633

Santos, P., Amado, C., Coelho, S.T., & Leitão, J.P. (2017). Stochastic data mining tools for pipe blockage failure prediction. *Urban Water Journal, 14*(4), 343–353. doi:10.1080/1573062X.2016.1148178

Scheidegger, A., Leitão, J.P., & Scholten, L. (2015). Statistical failure models for water distribution pipes – A review from a unified perspective. *Water Research, 83*, 237–247. doi:10.1016/j.watres.2015.06.027

Scholten, L., Scheidegger, A., Reichert, P., & Maurer, M. (2013). Combining expert knowledge and local data for improved service life modeling of water supply networks. *Environmental Modelling & Software, 42*, 1–16. doi:10.1016/j.envsoft.2012.11.013

Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, 40*(1), 185–197. doi:10.1109/TSMCA.2009.2029559

Shahata, K., & Zayed, T. (2012). Data acquisition and analysis for water main rehabilitation techniques. *Structure and Infrastructure Engineering, 8*(11), 1054–1066. doi:10.1080/15732479.2010.502179

Sorge, H.-C. (2006). *Technische Zustandsbewertung metallischer Wasserversorgungsleitungen als Beitrag zur Rehabilitationsplanung* [Technical condition assessment of metallic water distribution pipes as contribution to rehabilitation planning]. Weimar, Germany: Bauhaus-Universität Weimar.

Syachrani, S., Jeong, H.S., & Chung, C.S. (2013). Decision tree-based deterioration model for buried wastewater pipelines. *Journal of Performance of Constructed Facilities, 27*(5), 633–645. doi:10.1061/(ASCE)CF.1943-5509.0000349

Tran, D.H., Ng, A.W.M., & Perera, B.J.C. (2007). Neural networks deterioration models for serviceability condition of buried stormwater pipes. *Engineering Applications of Artificial Intelligence, 20*(8), 1144–1151. doi:10.1016/j.engappai.2007.02.005

Tscheikner-Gratl, F. (2016). *Integrated approach for multi-utility rehabilitation planning of urban water infrastructure*. PhD Thesis, Innsbruck University Press.

Tscheikner-Gratl, F., Sitzenfrei, R., Hammerer, M., Rauch, W., & Kleidorfer, M. (2014). Prioritization of rehabilitation areas for urban water infrastructure. A case study. *Procedia Engineering, 89*, 811–816. doi:10.1016/j.proeng.2014.11.511

Tscheikner-Gratl, F., Sitzenfrei, R., Rauch, W., & Kleidorfer, M. (2016). Enhancement of limited water supply network data for deterioration modelling and determination of rehabilitation rate. *Structure and Infrastructure Engineering, 12*(3), 366–380. doi:10.1080/15732479.2015.1017730

Wilson, D., Filion, Y., & Moore, I. (2017). State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. *Urban Water Journal, 14*(2), 173–184. doi:10.1080/1573062X.2015.1080848