Reliability Engineering and System Safety 241 (2024) 109611

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Evaluating the generalizability and transferability of water distribution deterioration models

Shamsuddin Daulat [a,*], Marius Møller Rokstad [a], Stian Bruaset [b], Jeroen Langeveld [c], Franz Tscheikner-Gratl [a]

[a] Department of Civil and Environmental Engineering, Norwegian University of Science and Technology (NTNU)
[b] SINTEF, Trondheim, Norway
[c] Faculty of Civil Engineering and Geosciences, TU Delft

ARTICLE INFO

Keywords:
Data preprocessing
Random survival forests
Survival functions
Variable importance
Uncertainties

ABSTRACT

Small utilities often lack the required amount of data to train machine learning-based models to predict pipe failures, and hence are unable to harness the possibilities and predictive power of machine learning. This study evaluates the generalizability and transferability of a machine learning model to see if small utilities can benefit from the data and models of other utilities. Using nine Norwegian utilities' datasets, we trained nine global models (by merging multiple datasets) and nine local models (by utilizing each utility's dataset) using random survival forest. Several pre-processing techniques including addressing left-truncated break data and break data scarcity are also presented. The global models and three of the local models were tested to predict the pipe failure of the utilities which were not included in their training datasets. The results indicate that the global models can predict other utilities with sufficient accuracy while local models have some limitations. However, if a representative utility with a sufficiently large (and information rich) dataset is selected, its model can predict the other utility's pipe breaks as accurate as the global models. Furthermore, survival curves for defined cohorts as proxies for uncertainty, and variable importance show that pipes with and without previous breaks behave extremely different. With the understanding of models' generalizability and transferability, small utilities can benefit from the data and models of other utilities.

1. Introduction

Aging water distribution networks, increasing break rates, and limited budgets are putting the utilities under increasing pressure to manage and, in consequence, rehabilitate their networks. In Norway [64], around 50,340 km of water supply pipes exist, with a new pipe installation rate of 1.4% in 2021 compared to a renewal rate of 0.68% per year. In the Netherlands, the renewal rate of water supply pipes is around 1%. Renewal rates in the UK were less than 0.6% annually [39], while in the US, they increased from 0.5% in 2015 to at least 1% (up to 4.8% based on the utility) by 2019 [3]. By a rough estimate, a 2% renewal rate is deemed suitable for OECD countries [48]. Proper maintenance of such critical infrastructure systems is important for the economy, environment, and public health [5,44,50,53,54,61]. To facilitate proper maintenance, utilities need to assess the assets' condition to elicit the maintenance and rehabilitation needs and consequently plan actions. As urban water infrastructure assets are, in contrast to other infrastructures such as roads, underground and difficult to access for inspection, the use of models to predict their conditions or their failure is more common than for easily accessible infrastructures. Therefore, deterioration models of water pipes are commonly used in research and also in practice to assess the reliability of pipes and hence to estimate the rehabilitation needs and aid in developing risk-based rehabilitation strategies.

Despite the importance of proper maintenance and rehabilitation of water pipe networks, many utilities have barely started applying asset management principles and can therefore not be expected to have gathered sufficient failure/condition data already [80]. This issue is especially prevalent among small utilities, which often struggle with limited resources, expertise, and data availability [23,25,26,75]. This lack of resources can lead to difficulties in addressing the increasing frequency of pipe failures in aging water distribution networks, which can ultimately compromise the quality of the water supply. Despite the urgent need for solutions, there are relatively few studies available that

* Corresponding author.
E-mail address: shamsuddin.daulat@ntnu.no (S. Daulat).

https://doi.org/10.1016/j.ress.2023.109611
Received 16 November 2022; Received in revised form 31 July 2023; Accepted 28 August 2023
Available online 3 September 2023
0951-8320/

address the specific challenges that small utilities face in this regard. Wood and Lence [76] proposed an approach for effective asset management in small and medium utilities, which involves identifying critical data and selecting suitable models to predict pipe breaks. The study identified material, diameter, age, length, and break data as crucial variables. To optimize asset management, the authors recommended gathering data on these variables through future data acquisition programs. Haider et al. [25] developed a performance management model for internal activities of small to medium sized utilities. Francisque et al. [23] proposed a life cycle cost approach to help small and medium utilities prioritize water pipe interventions. Winkler et al. [73] used low quality data, enhanced by a distance-based reconstruction algorithm, to train a Random Forest (RF) model for one utility. In the Netherlands, multiple drinking water companies pooled their water pipe and failure data in a shared database called USTORE to use it for failure modeling. It was reported to contribute to more reliable failure analysis [6].

Chen et al. [13] utilized data from other utilities to augment the limited break data of a target utility, with the aim of improving the performance of machine learning (ML) models built using the target utility's data. However, the study found that this approach did not consistently improve the target utility's performance. It is worth noting that the study did not explore the possibility of using models built on the data from other utilities (either individually or in combination) to make predictions on the target utility's pipe deterioration and evaluate their performance. Data of water distribution networks are usually classified as sensitive, thus restricting it from being circulated among the utilities. While models can be circulated among the utilities as they do not reveal sensitive information about the utilities from which they were trained. The question of whether a utility can use another utility's (or a group of utilities') model to predict its own pipe breaks remains open. To the authors' knowledge, no study has evaluated whether a utility can utilize other utilities' (individually or collectively trained) pipe break models and what accuracy could be expected of those models. This research gap highlights the need for further investigation into the feasibility of sharing failure models between utilities.

This study aims to fill this research gap by evaluating the transferability and generalizability of models between utilities. The term transferability in this study is defined as "the ability of single-utility trained models to produce accurate predictions for utilities that were not part of the training and testing". Similarly, generalizability is defined as "the ability of multi-utility trained models to produce accurate predictions for utilities that were not part of the training and testing". The former definition aligns with the definition provided in Wiesenfeld et al. [71] and the latter definition aligns with the one provided in Miner et al. [45]. This study specifically addresses the issue of inadequate data in small utilities where there may not be enough data to train a machine learning-based water pipe break model. To achieve this goal, this study evaluates the accuracy of both global (multi-utility trained) and local (single-utility trained) models in predicting water pipe break of other utilities which were not part of the training process. The study combines data from multiple utilities to train global models for generalizability and uses individual utility data to train local models for transferability. This approach has not been previously studied in literature. This work is particularly helpful for small utilities with limited data and expertise. By utilizing pretrained models, these small utilities are relieved of the burden of acquiring sufficiently large datasets and extensive expertise, while still achieving accurate predictions.

This study utilizes Random Survival Forest (RSF), a ML technique that is not commonly used in pipe deterioration modeling, to train the models. The trained models predict the timespan until the occurrence of the next pipe failure. Moreover, separate survival curves for different utilities with groups of selected explanatory variables are presented to provide insights into the possible uncertainties in the deterioration modelling process. They can also highlight the differences in survival probabilities for different pipe groups, and the decisive factors in shaping them. The findings from the curves are further supported by analyzing the explanatory variables using permutation-based variable importance. The developed approach can provide important insights into the factors that contribute to the failure of water pipes, which can help utilities better allocate resources and prioritize repairs. The permutation-based variable importance analysis also provides an understanding of the relative importance of the explanatory variables in predicting water pipe failure.

The remainder of the paper is structured as follows: Section 2 provides a background on different deterioration models for pipes, as well as the various factors that influence pipe break prediction. Section 3 describes the data and the method. Section 44 describes the results and discusses the prediction performances of the models in a comparative way. It also presents and discusses the survival curves, the uncertainties, and the importance of explanatory variables. Section 55 provides conclusions.

2. Pipe deterioration models and the influencing factors

The currently applied deterioration models are generally classified as physical, statistical, and machine learning (ML) models [20]. Among these model classes, physical models are closest to emulating the reality of different individual deterioration processes. However, collecting the data required for these models is usually difficult and not cost effective, except for critical and large-size transmission pipes [37]. The use of statistical models has traditionally been the most common approach for pipe deterioration modeling mainly due to its cost-effectiveness [5]. A large variety of statistical models have been developed throughout the years and applied for water distribution networks [15,56,58,60,69] and they were reviewed by several articles with different perspectives (e.g., by explaining the mechanisms or comparing the models) and details (e.g., individual or classes of models) [18,37,47,57,65,72]. Such models use historical failure data to recognize patterns and then extrapolate those patterns into the future [72]. These models, however, have a limited capability of identifying the complex relationships that may exist between explanatory and response variables. As an alternative, ML models can help to identify these complex relationships with relative ease given that enough and reliable data is provided. Several ML techniques have been applied in pipe failure modeling in recent years [19]. Some of the most applied techniques by researchers in pipe failure modeling are artificial neural networks, logistic regression, genetic algorithms, and decision tree-based models [9,16,28,66,73]. One advantage of many ML models is their ability to rank variable importance, while statistical models provide more insights into the variables, hence, complementing each other [39].

The complementary relationship between statistical and machine learning models is one reason why RSF was selected for this study, as RSF was developed for fulfilling those needs. RSF, developed by Ishwaran et al. [31], combines RF with statistical survival modeling, providing time-dependent survival probabilities for individual water pipes. The latter is key for tactical planning of water pipe rehabilitation programs. Additionally, pipes with similar characteristics can be clustered into groups, and the probability of failure for each group can be presented for strategic rehabilitation planning. A further advantage of RSF is that it utilizes the power of ensemble ML and considers right-censorship (i.e., when the event of interest has not yet occurred [36]), whose lack of consideration is a drawback of RF models. In a study by Snider and McBean [63], the RSF model was shown to outperform the Weibull proportional hazard survival model as well as the RF model for water pipe failure prediction. Laakso et al. [40] applied the RSF model for sewer pipes' life span prediction and found that it performs slightly better than the Weibull model but neither offered excellent results. Notwithstanding its advantages and suitability, RSF is a novel and scarcely applied technique in water and wastewater asset management [63]. However, in a study by Almheiri et al. [2] on failure modeling of water pipes, RSF's performance was shown to be inferior to survival support vector machine and an artificial neural network method (advanced meta-learning). As the literature does not consistently demonstrate the superiority of any method, we opted to use RSF in this study. It is important to note that this study uses RSF as a tool, not as the objective.

Although ML techniques outperform statistical methods in most cases [9,21,39,40,63], their performance relies not only on quality but also on a considerable quantity of data – something that smaller utilities rarely possess.

The same cannot be said for literature on influencing factors on the deterioration processes, and the models applied to mimic it. Variable importance analysis is conducted as part of modeling by many studies [39,55,63,73]. The explanatory variables derived are divided into three groups: operational, environmental, and pipe-intrinsic factors. These factors influence pipe failure occurrence, both simultaneously and in interaction with each other. These interactions make it difficult to value the individual influence of a single variable on pipe failures. Nonetheless, there is a need to identify the most influential factors in order to prioritize data collection on these factors and emphasize the data quality requirements. Barton et al. [4] reviewed the factors affecting pipe failure without discussing the degree of importance of these factors. Studies that have evaluated the importance of factors have got varying results. The most used variables in the literature are pipe-related factors such as diameter, length and material [59]. Winkler et al. [73] used four different decision tree-based models and concluded that age and material type are the most important factors. Konstantinou and Stoianov [39] used several statistical and ML models and concluded that maximum pressure and pressure range are the two most important variables. Jara-Arriagada and Stoianov [32] examined the effect of pressure control on a water distribution network and quantified its impact. They found that lowering the mean pressure for a group of asbestos cement and cast-iron pipes can result in pipe breaks reduction by 18% to 30%. In another study [14] where pipe age was set as an output variable for an artificial neural network model, the number of previous breaks, length, material and diameter were reported the most important variables, respectively. Yamijala et al. [79] stressed on pipe length as a key factor in predicting the number of breaks, as the length is directly correlated with the surface area exposed to hazards. Chen and Guikema [12] used spatial pattern analysis to group pipe breaks into clusters and identified areas with unusually high break rates. Iannacone and Gardoni [29] quantified the impact of earthquakes on water pipelines and the need for repair rates using a physics-based model. In addition to modelling, Fares and Zayed [22] collected municipal experts' opinion in Canada about the factors affecting pipe failure. The conclusion was that pipe age, material and failure rate were the most important factors, respectively.

3. Material and methods

This section describes the scenario building and modelling process applied in this study. It describes the used data and the (anonymized) utilities, network composition and size (3.1). Based on this, data preprocessing (3.23.2) is performed and the issues of quality and quantity are addressed. The applied model and its performance metrics are described in (3.33.3). Finally, the method of evaluation of the most influential factors is described in (3.43.4). As a naming convention, the authors chose to refer to models trained on the data of one utility as local models and for models trained on multiple utilities as global models.

3.1. Data description

Water distribution pipe data of nine utilities of Norway are used in this study. The network data initially consists of approximately 139,000 pipes. After pre-preprocessing and cleaning, around 108,000 pipes with a total length of 7,314 km remained for modeling. Pipe break data was available for most utilities from the 1970s to 2015. However, in utility 7, pipe break data had been recorded mysteriously but systematically since 1900 (as shown in Fig. 1a). After consulting with local experts, it was concluded that it was highly unlikely for Norway to have had systematic and reliable break records before 1970. Therefore, any break records prior to 1970 were deemed unreliable and excluded from further analysis. The total number of reliable breaks included in the study was 18,511.

The construction periods of the pipes are representative of the construction practices of their time. Pipes from the period before World War II are produced generally in better quality than pipes constructed in the post-World War II era (until the 1960s). This is because pipelines were mostly constructed by hand before the war, and with machines after the war. The first decades with machine-constructed pipelines were characterized by rapid urban expansion, which translated to rapid construction of pipelines. Pipes from this period perform the worst in terms of failure rates that we experience today. After that period, and especially after 1980, the quality of the construction work was highly improved. Pipe material use over time is similar to most of the other networks in developed nations, with some local variations. It started with cast iron in 1850s until the second half of 20th century, then ductile iron became popular (from the 1960s) followed then by PVC and PE pipes (from the 1970s). Exceptions were found mainly in utility 4 and to a lesser extent in utility 8, where a rapid transition from cast iron to PVC, without the use of much ductile iron pipes in between (Fig. 1b), can be observed. There is a trend for many Norwegian utilities that they have been, and are still, focusing on using predominantly ductile iron, or predominantly using PE/PVC. This is due to professional preference and local conditions. Historical and current failure rates in the utilities are correlated to historical pipe production standards [8]. There was for example a change in the production standard for ductile iron pipes around 1980s, when most producers started to add internal corrosion protection in their pipes to reduce internal corrosion pitting problems. This had a positive effect on the failure rates of ductile iron pipes. A similar effect for PE/PVC pipes was observed when a change in the production standard occurred around 1980s.

3.2. Data pre-processing

As for every modelling endeavor, quality control of the used data is of the utmost importance. As the data for the global model is the combination of multiple utilities' datasets, explanatory variables for this model must be shared among all the utilities. If one utility misses the recording of a variable, e.g., interior protection type, then the interior protection type of all the other utilities is dropped. Otherwise, the global dataset will have missing data and the treatment of missing data for the RSF model will increase the uncertainties to a level that dropping the variable can make it less uncertain. As a result, fewer explanatory variables for training the global model are available than what can be used for each local model individually (a complete list of available variable data in each utility is shown in Appendix). The variables shared among all the utilities in our datasets are material type, length, diameter, number of previous breaks, and number of other pipes connected to the target pipe's manholes at the start and at the end (these last two variables are derived by GIS analysis). Pipes with less than one meter in length are not considered in this study and removed from further analysis. More explanatory variables' data were available for some local models. For improved comparability between global and local models, the local models are also trained with the same explanatory variables as used in the training of the global models. The consequence of this is that the performance of local models is in general a bit underestimated in this study, although internal tests showed that the effect was minimal.

After selecting the explanatory variables, they were checked for anomalies. Visualization of the data (e.g., Fig. 1a and b) is helpful and an effective way of identifying anomalies such as missing data, outliers and implausible values [55]. The anomalous values of material type were grouped into a single category labeled as "unknown". Missing or anomalies of diameters were replaced with the median diameter of the

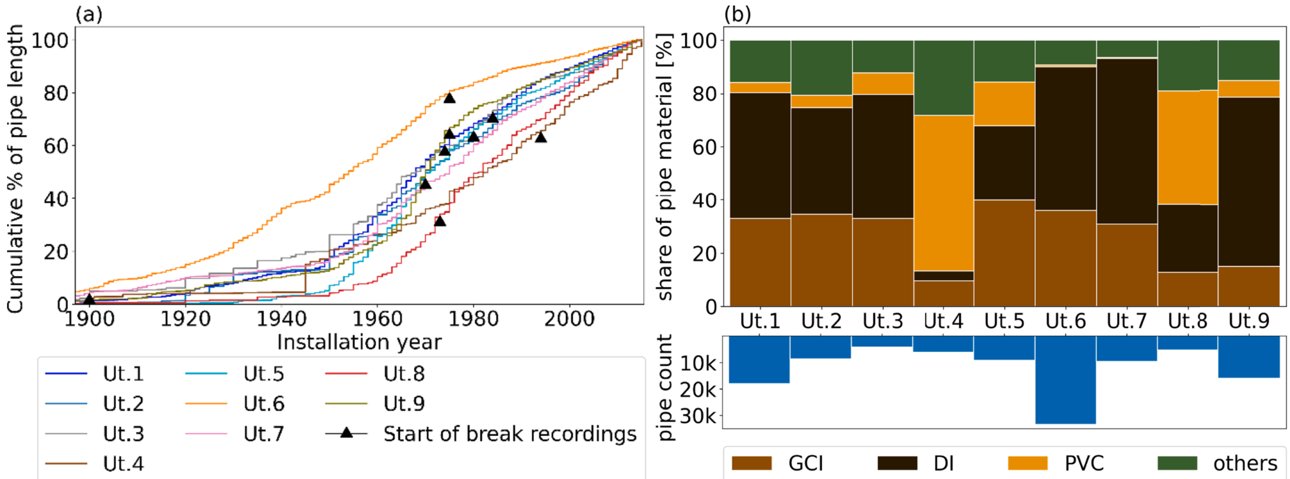


Fig. 1. (a) Cumulative% of pipe length over the years for each utility and the start of break recordings; (b) Pipe material distribution and count of pipes in each utility.

network, and pipe length was taken from the GIS mapping of the networks. One should note that there are advanced techniques for imputing missing data, which can outperform the simpler methods we used here for diameter and material type [34]. However, we chose not to focus on data imputation in this study. Instead, we aimed to look at the generalizability and transferability of models since many utilities struggle with limited data availability and lack of expertise. Thus, we decided to work with the available data, which is representative of the situation that many utilities face. It is important to note that filling in the missing values of the most important variables, such as age, with estimations can induce more uncertainties and lower the prediction performance of the models. Therefore, in the case of anomalies in age, the sample was excluded from the study due to its significance as a response variable in RSF model. Approximately 3000 pipes were removed during this stage, primarily due to the exclusion of short pipes (those of only a few centimeters in length) and pipes with anomalous age (construction year) data. More than just missing values and outliers in the dataset, there are other quality issues in a typical water pipe dataset that need to be addressed in order to increase the performance. The following paragraphs identify and address these issues.

3.2.1. *Replaced pipes*

Unlike most other typical datasets, which seldom include information on replaced pipes [58], the used Norwegian water network datasets have records of replaced pipes with the time of replacement starting in the 1980s. However, the reason for replacing the pipes is usually not recorded (e.g., if the replacement was due to a break or because of adaptation to changes in demand or environment). For our deterioration modeling, only a binary state (fail or not fail) at the time of replacement is sufficient. As the state of the replaced pipes at the time of replacement was not known, we evaluated three scenarios to see which scenario gives the highest model performance: (1) the replaced pipes did not fail at the time of replacement; (2) the replaced pipes did fail at the time of replacement; and (3) remove these pipes from modeling. None of the three scenarios showed significant improvement or decline in model performance. Hence, we continued with the first scenario for the rest of the analysis as it was the convenient choice.

3.2.2. *Dealing with left truncated break data*

Left-truncated break data refers to the unknown break history of those pipes which were installed before the start of break records. The datasets of the utilities show that the earliest break records started in the 1970s. So, the break records for the pipes installed before 1970s are left-truncated. The time and number of previous breaks of those pipes are unknown. As a result, the first recorded breaks in pipes installed before 1970s might not be the real first break which eventually can lead to incorrect results [78]. Instead of discarding all the pipes installed before 1970, which comprises 48% of all the data by count, the following workaround has been applied:

Pipes installed from 1945 and after are considered for modeling because break records for pipes installed before 1945 are not reliable given the start of break records in 1970s. The first recorded breaks for the pipes installed before 1970s are assumed to be their first break in the raw dataset. A simple descriptive statistical analysis on pipes installed after the 1970s shows that on average it takes 30 to 40 years for the first breaks to occur. Thus, the first pipe breaks for pipes installed after 1940s are captured since 1970s and after. The majority of pipes installed before the 1940s have most probably already had some breaks before the recording period. Thus, break records of pipes installed before 1940s contain more unreliable break records than reliable break records. 1945 was taken as a starting point, as not much infrastructure construction work was undertaken before 1945. In utility 4, break records started in 1994. Hence, instead of 1945, 1960 was taken as the cutoff year. By excluding the pipes installed before 1945 (1960 for utility 4), approximately 27,000 (19% of the total) were discarded, of which 25,000 were mainly composed of grey cast iron (GCI) pipes. This corresponds to approximately 42% of all GCI pipes being excluded from the combined datasets.

In the results section, the performance of the models is presented under two scenarios: all pipes and pipes installed from 1945 and onwards. The analysis indicated that the models performed better in the second scenario, which included pipes installed from 1945 onwards.

In our study, we utilized a simple yet effective method to handle left-truncated break data. Although advanced techniques have been developed, such as machine learning imputation techniques [77] and the Yule process extension used in Le Gat [41] that incorporated the left-truncation issue, we chose to adopt a simpler approach that was sufficient for our purpose. Our focus was to compare local and global models and examine the generalizability and transferability of break models between utilities. It was not our aim to create the perfect model, but rather to use a pragmatic and feasible approach that is accessible to many utilities.

3.2.3. *Previous pipe breaks*

For each recorded break in a pipe, we give new IDs to the pipe (and hence, virtually new samples) for two reasons. One, the number of broken pipes is usually by far fewer than the number of unbroken pipes, and even in some cases the number of broken pipes is not sufficient to be used for modelling. Two, we want to use the number of previous breaks as an explanatory variable for the model as it is bound to have an impact on decreasing the time to next break [67]. Each real pipe having an n number of previous breaks is transformed into $n+1$ virtual pipes. As illustrated in Fig. 2, a pipe with 3 previous breaks is transformed into 4 virtual pipes, starting from 0 previous breaks up to 3 previous breaks. Their time to the next break event is also changed accordingly. In this way the number of break records is expanded without compromising the quality or reliability of data. This technique has been successfully applied in the work of Winkler et al. [73].

Available break records do not distinguish between the breaks that happened due to external events such as earthquakes, floods, digging and construction errors, etc., and the breaks due to gradual deterioration. This is a typical problem in break records of datasets for water pipes [57]. Had the break records been differentiated between the gradual deterioration and external events, only gradually deteriorated pipes would have been selected for deterioration modeling, while the ones for external events could be used for (inter-) dependency assessments [17]. Pipes breaking due to external events cannot be considered as gradually deteriorated pipes. They are probably one of the sources of noise in the results. The share of pipes breaking due to external factor/third party is 1 out of 6 in the Netherlands' USTORE database.

Table 1 shows the variables and their ranges after data pre-processing. These variables were used as explanatory and response variables of the models used in this study. The RSF model requires two types of response variables, age, and status (1=fail, 0=no fail) at this age. We have provided the pipe age at the time of break, and status=1 if the pipe has failed at this age. Otherwise, the pipe age at the time of censoring was provided with its status=0, meaning that the break of interest has not occurred at this age.

3.3. Model description

RSF [31] is an ensemble machine learning method and an extension of the RF method [7] specifically tailored for survival modeling. In other words, RSF is a combination of RF and survival models, such as the Kaplan-Meier estimator [35] or the Nelson-Aalen estimator [1,46]. This combination allows RSF to harness the strengths of both approaches and create a powerful tool for survival modeling. It has the capability to handle censored data and predict survival outcomes. Compared to other survival analysis techniques, the RSF algorithm has several advantages. It is a non-parametric method and can handle high-dimensional data, making it a powerful tool for analyzing complex datasets. Additionally, the RSF algorithm provides measures of variable importance, which can help identify the most influential predictors of survival. This feature is especially useful when the goal is to understand the underlying mechanisms driving the survival outcomes. RSF was applied in water pipe deterioration modeling by Snider and McBean [62,63] Snider and McBean [62] describe the RSF in simpler terms. The RSF algorithm consists of survival trees, which are similar to decision trees used in RF. However, instead of using impurity measures like the Gini index or information gain as the criteria for splitting nodes, the maximum survival difference is used as the split criterion. This ensures that the nodes are split in a way that maximizes the difference in survival probabilities between the two child nodes. The log-rank test statistic is utilized to assess and optimize the differences between survival curves. The log-rank test statistic [43] is formulated as:

$$G(s) = \sum_{i=1}^{n} I\{X_{ij} \le y\} \left(\delta_i - S(t_i)\right) \quad \text{(Eq. 1)}$$

where X is the input variable j for individual i; y is the split criterion for input variable j; I is the indicator function (1 if X is less than y and 0 if not); δ_i is the censor indicator; and $S(t_i)$ is the survival curve.

Moreover, the RSF algorithm estimates the survival function at the terminal nodes of each tree using the Kaplan-Meier estimator (Eq. (2)) for survival probability or the Nelson-Aalen estimator (Eq. (3)) for the cumulative hazard function (CHF), both of which utilize time-to-event data.

$$\widehat{S}(t) = \prod_{t_i \le t}\left(1 - \frac{d_i}{n_i}\right) \quad \text{(Eq. 2)}$$

$$\widehat{H}(t) = \sum_{t_i \le t}\frac{d_i}{n_i} \quad \text{(Eq. 3)}$$

where t_i is the time when at least one event occurred, d_i is the number of events that occured at time t_i, and n_i is the pipes known to have survived up to time t_i.

Finally, to obtain the CHF of each individual pipe, the RSF algorithm averages the CHFs of all the trees in the ensemble. The CHF, denoted by $\hat{H}(t)$, represents the accumulated risk of experiencing an event up to time t. It can be converted to the survival probability, $\hat{S}(t)$, which is the probability of surviving beyond time t, as follows:

$$\widehat{H}(t) = -ln(\widehat{S}(t)) \quad \text{(Eq. 4)}$$

To better present the survival probabilities for groups of pipes with similar characteristics, we utilized the model's ability to provide

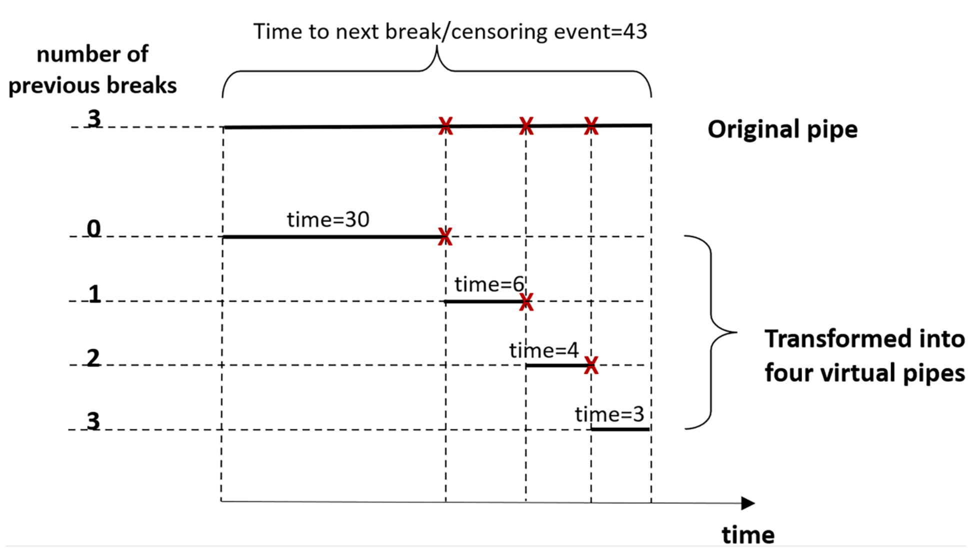


Fig. 2. Illustration of transforming a real pipe into virtual pipes to increase the number of break records for modeling purposes.

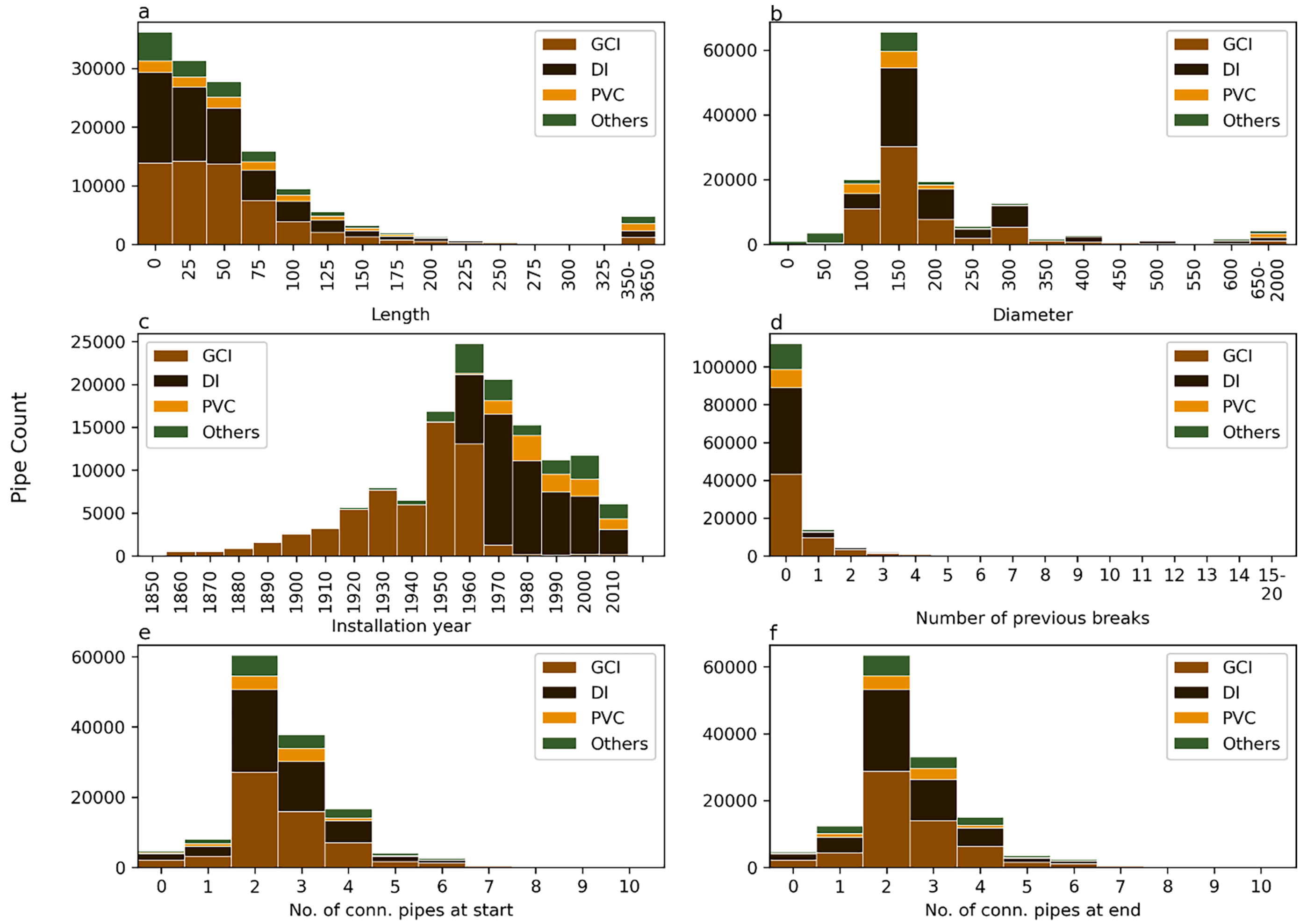


Fig. 3. Distribution of the common variables, all utilities' data combined.

survival probabilities for individual pipes and grouped them based on their material and the number of previous failures. This allowed us to present the survival probabilities for cohorts of pipes, which provides valuable information about the expected lifespan of pipes with similar characteristics.

To provide a comprehensive picture of the average and the uncertainties involved in cohort survival probabilities, several measures, including the median, 75th percentile, and 95th percentile were used. The 75th and 95th percentile survival probabilities provide information on the upper limits of the cohort's survival distribution and can help assess the range of potential outcomes.

By presenting survival probabilities for cohorts of pipes with similar characteristics, one can gain insights into the factors that influence the lifespan of pipes and make informed decisions about maintenance and replacement schedules. Moreover, by using multiple measures to describe cohort survival probabilities, we can better account for the uncertainties involved and make more accurate predictions about the lifespan of the pipes.

The concordance index (C-index), introduced by Harrell et al. [27], is the standard performance metric used to evaluate the predictive accuracy of survival analysis [42] such as RSF. The C-index compares pairs of pipes between each other. It rewards the model if the pipe that failed first had a higher predicted risk of failure. Risk of failure in RSF modeling is the area under the CHF curve. It is important to note that not all possible pairs are comparable. The following steps illustrate which pairs are comparable and how the C-index algorithm rewards the model:

- Generate all possible pairs of pipes using the test data.
- Exclude pairs where both pipes have a censored failure event. Exclude pairs where the pipe with shorter survival time is censored and exclude pairs with identical survival times unless at least one of the pipes has failed. The remaining pairs are comparable and are evaluated.
- For each comparable pair, determine if the pipe with shorter survival time has a worse predicted outcome. If the predicted outcomes are tied, count 0.5 for that pair. For pairs with identical survival times and both have failed, count 1 if the predicted outcomes are tied, otherwise count 0.5. For pairs with identical survival times where one pipe failed and the other is censored, count 1 if the failed pipe has a worse predicted outcome, otherwise count 0.5.
- Calculate the sum of these counts over all comparable pairs and call this sum Concordance.
- Calculate the C-index by dividing the sum Concordance by the total number of comparable pairs. The resulting C-index represents the probability that the model correctly predicts which case has a worse outcome.

The C-index depends on the censoring distribution and is shown to be not a sufficiently descriptive performance index for data that comprises highly right-censored events [68]. As right-censorship dominates in water distribution pipe networks (i.e., majority of pipes do not have any breaks recorded), it is recommendable to use a more suitable performance metric than the C-index alone. Uno et al. [68] presented a modified version of C-index, called C-index-inverse probability of

Table 1
Explanatory and response variables for the random survival forest models.

Explanatory-variables	Variable type	Variable range	Description
Material type	Categorical	DI, GCI, PVC, PE, PEL, PEH, PE50, PE80, PE100, UCI, GST, ST, Cu, C, Unknown,	DI=ductile iron, GCI=grey cast iron, PVC=polyvinyl chloride, PE=Polyethylene, (PEL, PEH, PE50, PE80, PE100, PE100K are PE's sub-variants based on density and/or production standard), UCI=unspecified cast iron, GS=galvanized steel, ST=steel, Cu=copper, C=Concrete
Length	Numerical	1–3627	Length of pipe segments [m]
Diameter	Numerical	20–2000	Pipe diameter [mm]
Number of previous breaks	Numerical	0–20	Count of breaks recorded on individual pipes
Number of pipes connected at starting manhole	Numerical	0–10	Count of other pipes connected to the manhole at the start of the pipe of concern
Number of pipes connected at ending manhole	Numerical	0–10	Count of other pipes connected to the manhole at the end of the pipe of concern
Response-variables			
Survival/break age	Numerical	0–70 (years)* 0–165 (years)**	Age of pipe at the time of break/ censoring
Status	Binary	0,1	0=not failed, 1=failed

* For pipes installed between 1945 and 2015,.
** For all pipes (1850-50]

The distribution of the common variables for all utilities' data combined between 1850 and 2015 is illustrated in Fig. 3.

censoring weights (C-index-ipcw) to address this shortcoming. In essence, the C-index-ipcw weighs the contribution of each observation based on the probability of it being censored, thereby adjusting for the bias introduced by the censoring mechanism. This approach allows for more accurate evaluation of the model's ability to predict the time-to-event outcome, especially in situations where there are a large proportion of censored observations. As literature for water and sewer systems asset management used the C-index [40,63], we did not only evaluate our models' results with the C-index-ipcw but also evaluated with the C-index to enable comparison of our results with literature. However, our final evaluation and emphasis is on the C-index-ipcw.

The C-index and C-index IPCW both have a range of 0 to 1, with higher values indicating better model performance. A score of 1 indicates perfect model performance, while a score of 0.5 suggests that the model is no better than random guessing. A score below 0.5 indicates that the model performs worse than random guessing.

Using the RSF algorithms written in Scikit-survival Python package [51], we developed a total of 18 models, including nine global models and nine local models, in a Python platform. To ensure reliable and robust results, using Scikit-learn [49], we utilized a train-test split of 80–20 percent for the local models, and each local model was trained 15 times, with each run predicting its respective utility's test dataset, with a randomly selected train-test dataset. For the global models, each model was trained with data from eight utilities excluding the ninth utility, which was then predicted 15 times, each time with 80% of its randomly selected pipes.

We chose these percentages to balance the need for a sufficient amount of data for training while ensuring that there was enough data for testing to provide a reliable estimate of model performance. The train-test split shuffles the data randomly and splits it into two sets based on the specified ratio. This approach assumes that the data points are independent and identically distributed (i.i.d.), and there is no inherent structure or pattern in the data (temporally or spatially) that needs to be preserved in the split. Additionally, we conducted 15 model simulations and predictions, as additional simulations did not yield significantly different results.

Moreover, we investigated the transferability of the local models by selecting the three largest utilities as reference utilities. These reference utilities were then used to predict the failure of pipes in the remaining utilities, allowing us to compare the transferability of local models with the generalizability of global models. The prediction process for reference utilities was the same as for global models, where 80% of the pipes were randomly selected for 15 iterations.

We used the default hyperparameters provided by Scikit-survival and did not perform hyperparameter tuning. Given that our main objective was to compare the results between the different models, we believe that the default hyperparameters provide a fair basis for comparison. However, a sensitivity analysis showed that the following hyperparameters could be tuned to improve the performance of the models: number of trees in the forest (n_estimators), number of features for the best split (max_features), and minimum number of samples at a leaf node (min_samples_leaf). Although hyperparameter tuning could potentially improve the performance of the models, it is not the objective of this study to try and get an optimal numerical fit, but rather to provide a proof of concept for generalizability and transferability.

3.4. Assessment of variable importance

Permutation based variable importance analysis using the algorithms in Scikit-survival [51] was carried out to pinpoint which variables contribute the most to the model prediction and if there is a common predictive variable among all the utilities. The method is theoretically based on removing individual explanatory variables and retraining the model to see how much the performance is reduced in the absence of that variable. Then the importance of the variable is calculated from the difference in the C-index (as the built-in function of RSF is based only on C-index) with and without that variable. This way can be computationally expensive for such a large amount of data due to retraining of the model for each variable, and hence, impractical. Practically, the values of the variable of interest in the test dataset are replaced by random values and the model is then tested. This way the random values of the variable of interest are assumed to have not contributed to the C-index. The importance of the variable of interest is then calculated from the difference in the C-index with actual and with random values of the variable. Sometimes the random values can accidentally be similar to the original values. To avoid such cases, each random filling of each target variable is iterated 15 times and the average change in performance is recorded.

4. Results and discussion

The results show the applicability of the selected models. In the following sections we will discuss the prediction performance as a measure of model generalizability and transferability, the importance of the different variables and the insights that the usage of the RSF model and its survival curves can give us into the uncertainty of deterioration models, that is otherwise seldom addressed.

4.1. Prediction performance of the local, global and the reference models

The model results are presented in boxplots to show the spread and average performance (in terms of C-index and C-index-ipcw) of the 15 iterations. Fig. 4 shows the performance of nine global models (in green) in predicting the local utilities, as well as the performance of nine local models (in black) in predicting parts of their own network, in terms of C-index and C-index-ipcw. Fig. 4a and b show the models' performances

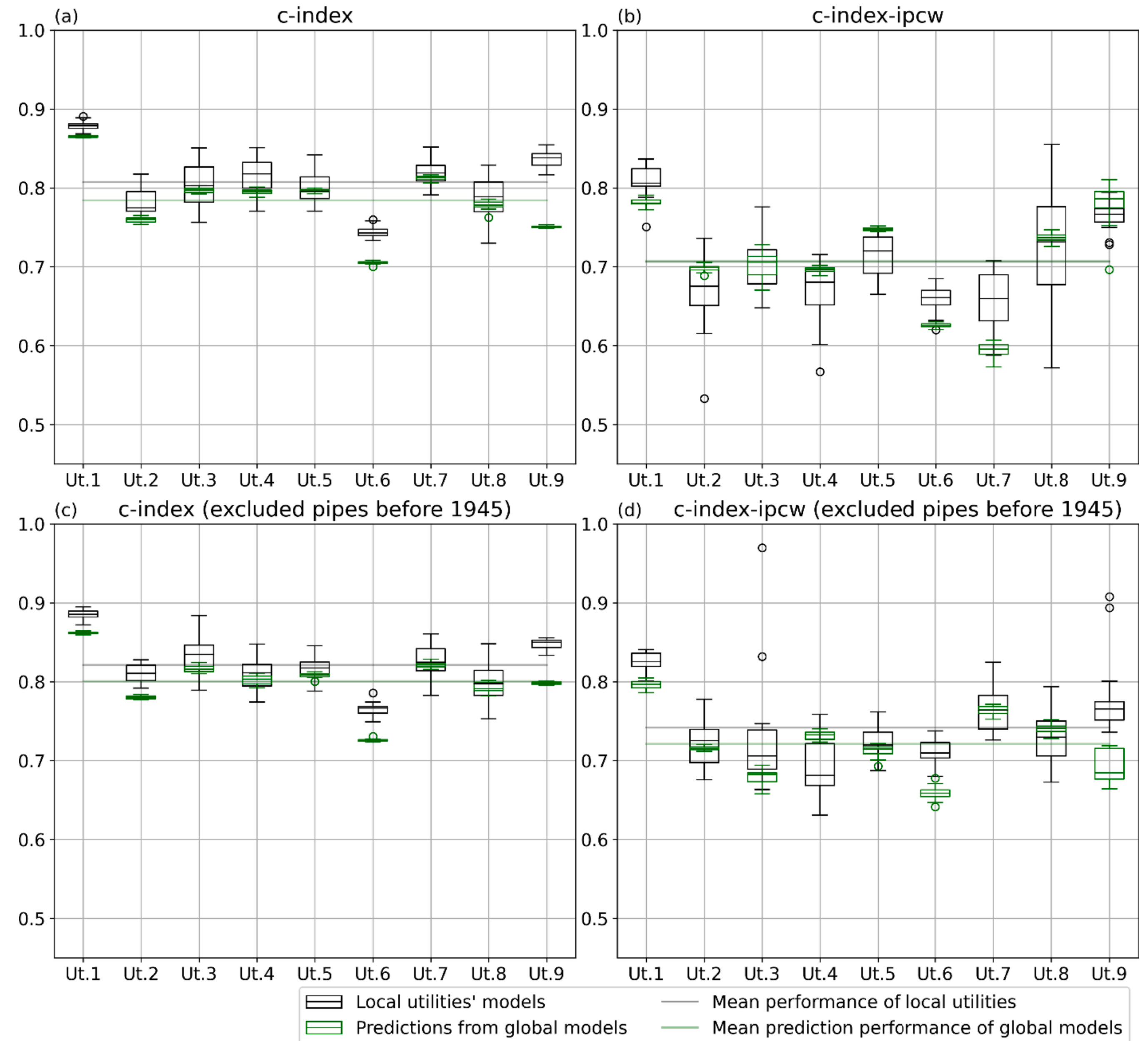


Fig. 4. In (a) and (b), prediction performance of global models compared to the local utilities (no filtration of pipes by installation year). In (c) and (d), prediction performance of global models compared to the local utilities (for pipes installed from 1945.

with input data including all the years of pipe installations, while Fig. 4c and d show the models' performances with input data including pipes installed only from 1945 and after. The figures show that excluding pipes installed before 1945 generally improved the performance. Thereby strengthening the assumptions made earlier. The results shown from this point onwards are based on the data of pipes installed from 1945 and after.

Comparing the performance of local and global models, Fig. 4c and d shows that locally and globally trained models have an overall average C-index of 0.82 and 0.8, and C-index-ipcw of 0.74 and 0.72, respectively. The global models' performances are close to local models and well beyond 0.5 (which is the threshold for a model to be not performing better than a random guess). The results clearly show the generalizability of a globally trained model for predicting local utilities. It would be interesting to check the performance of the global models for utilities in other countries with similar and different climate, geography, and historical network evolution to test the limits of such a transferability.

It is also interesting to assess if the local models are transferrable between themselves and if they deliver comparable results to the global models, i.e., if a local utility's model can predict another local utility's pipe breaks. To evaluate this, the three largest utilities (1, 6, and 9) were chosen as reference utilities. Models were trained for each reference utility using their own datasets and predictions on the other 8 utilities were made. Fig. 5 shows the prediction performance of the reference utilities in terms of the C-index-ipcw. Blue boxes show the performance of reference utilities when they predict parts of their own pipes. Their C-index-ipcw values on average are 0.83, 0.72, and 0.78 for the utilities 1, 6, and 9, respectively. Locally trained models stayed exactly the same (as nothing is changed here) with an overall average C-index-ipcw value of 0.74 while the respective index for predicting local utilities from the reference utilities (1, 6, and 9) are 0.69, 0.72, and 0.69, respectively. These results indicate that the performance of our 3 reference utilities in predicting the other utilities is way better than a random guess (i.e., 0.5).

Among the reference utilities, utility 1 and 9 are among the best

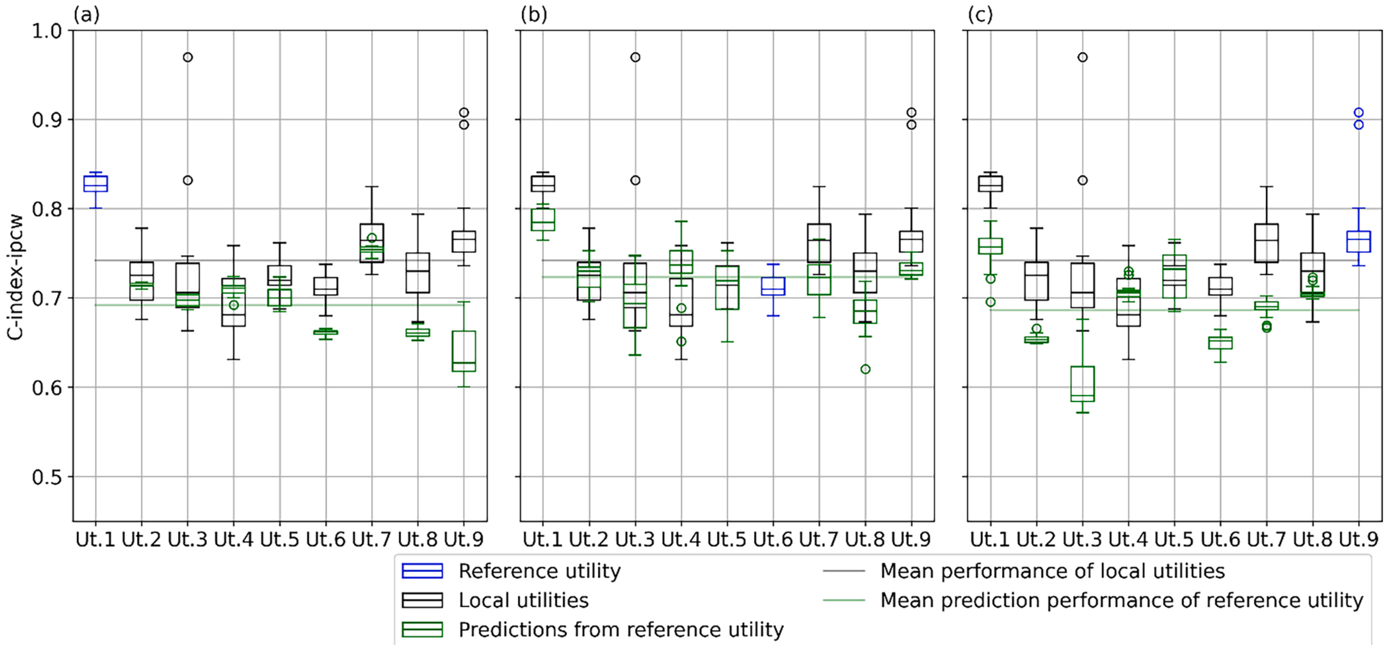


Fig. 5. Model performance in terms of C-index-ipcw of three reference utilities predicting other utilities compared to the other local utilities predicting their own pipes (for pipes installed from 1945 and after).

performances when predicting parts of their own networks, while their average performance to predict other utilities is weaker than for utility 6. On the contrary, utility 6 has the weakest performance for its own pipes but strongest performance for predicting other utilities. The reason is most likely that utility 6 is the largest among the 9 utilities (includes 30% of all pipes by total count). Utility 6 is also in a relative geographical vicinity location to most of the other utilities, sharing similar climate, demographics, and construction practices to some extent, while utility 1 and 9 are more remote from most of the other utilities. The weaker model performance of utilities 1 and 9 are even more noticeable when looking into the prediction of individual utilities (see predicted result of utility 9 by reference utility 1 in Fig. 5a and predicted results of utilities 1, 2, 3, and 7 by reference utility 9 in Fig. 5c). These results indicate that to achieve better transferability it is more important for the reference utilities to possess a large enough database with a good representativity for the other utilities than to be better performing for themselves.

Comparing the performances of the global models (Fig. 4d) with reference utilities' models (Fig. 5), utility 6′s performance (in Fig. 5b) is as high as the global models (both have an average C-index-ipcw of 0.72). However, utility 1 and 9 perform weaker than the global models. In practice if a utility manager is not sure which utility can better represent that utility, it is safer to use a global model than to use the model of a randomly selected utility. As shown in Fig. 4d, the global models predict each utility satisfactorily, while predictions from reference utilities (Fig. 5) depend on their size and representativeness.

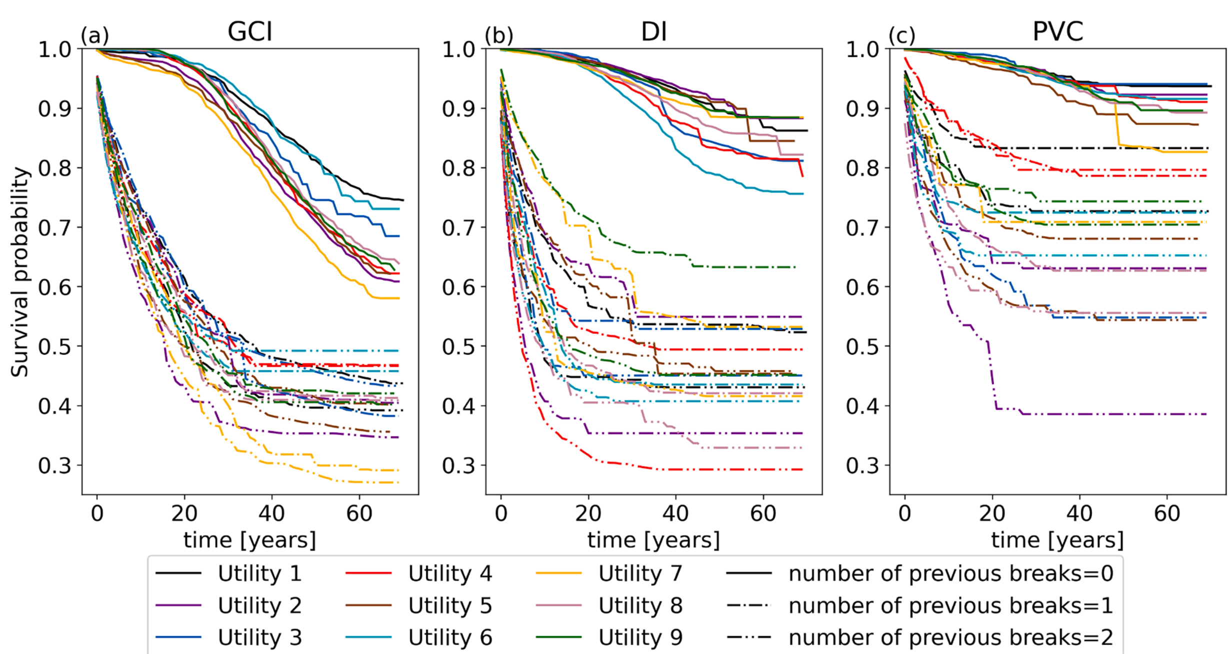


Fig. 6. Predicted survival curves for each utility using the global models, distinguished by utility, number of breaks and materials.

4.2. Survival curves

Based on the high performance of global models in predicting local utilities, survival curves for local utilities were predicted using global models shown in Fig. 6. Each survival curve represents the median survival curve of a utility for a cohort of pipes with the same material and same number of previous breaks (0,1, or 2). The three most common material types (grey cast iron, ductile iron, and polyvinyl chloride) were chosen for the analysis. The figure indicates that there is no clear survival (or time to next break) difference among the utilities for the same material type and previous breaks. However, survival times are significantly shorter for pipes with previous breaks compared to those without, which is consistent with Scheidegger et al. [57] argument that repairing pipes after a break reduces their structural strength permanently. The figure also shows that there is a small difference between survival probabilities of pipes with 1 and 2 previous breaks. This supports the findings of Tscheikner-Gratl et al. [67], that the difference in time to next failure decreases after the first break. Pipes with 2 previous breaks are a bit more prone to failure than the pipes with 1 previous break. Survival difference in different material types do exist (see the differences in Fig. 6 from a to c), however, their differences do not make the curves as diverse as the previous breaks do. Moreover, the differences in survival curves for different material types indicate that survival has improved from GCI to DI and to PVC, showing the evolution in the reliability of materials throughout time.

It is important to note that relying on the survival curves has limits. Once the survival curves start changing its slope towards horizontal direction it means less and less break data is available and that is the point where the curve starts being less reliable. Once the slope is completely horizontal no more break data is available, and no reliable conclusions can be drawn anymore. In Fig. 6a, for instance, the threshold of becoming less reliable on curves starts from around 55 years for the pipes with zero previous breaks and 20 years for the pipes with 1 or 2 previous breaks. The survival curves in Carrión et al. [10] grey cast iron pipes also start getting flat after around 50 to 55 years. The study did not further distinguish the pipes by the number of previous breaks. The flatting effect was also seen in Laakso et al. [40] model where they used RSF for lifespan prediction of sewer pipes.

The survival curves, like every modelling effort, come with uncertainties which are seldom recognized and addressed in literature. In Fig. 6 each curve was the median of a group of pipes, as is often done for cohort survival models that group by materials. To have a glimpse of the

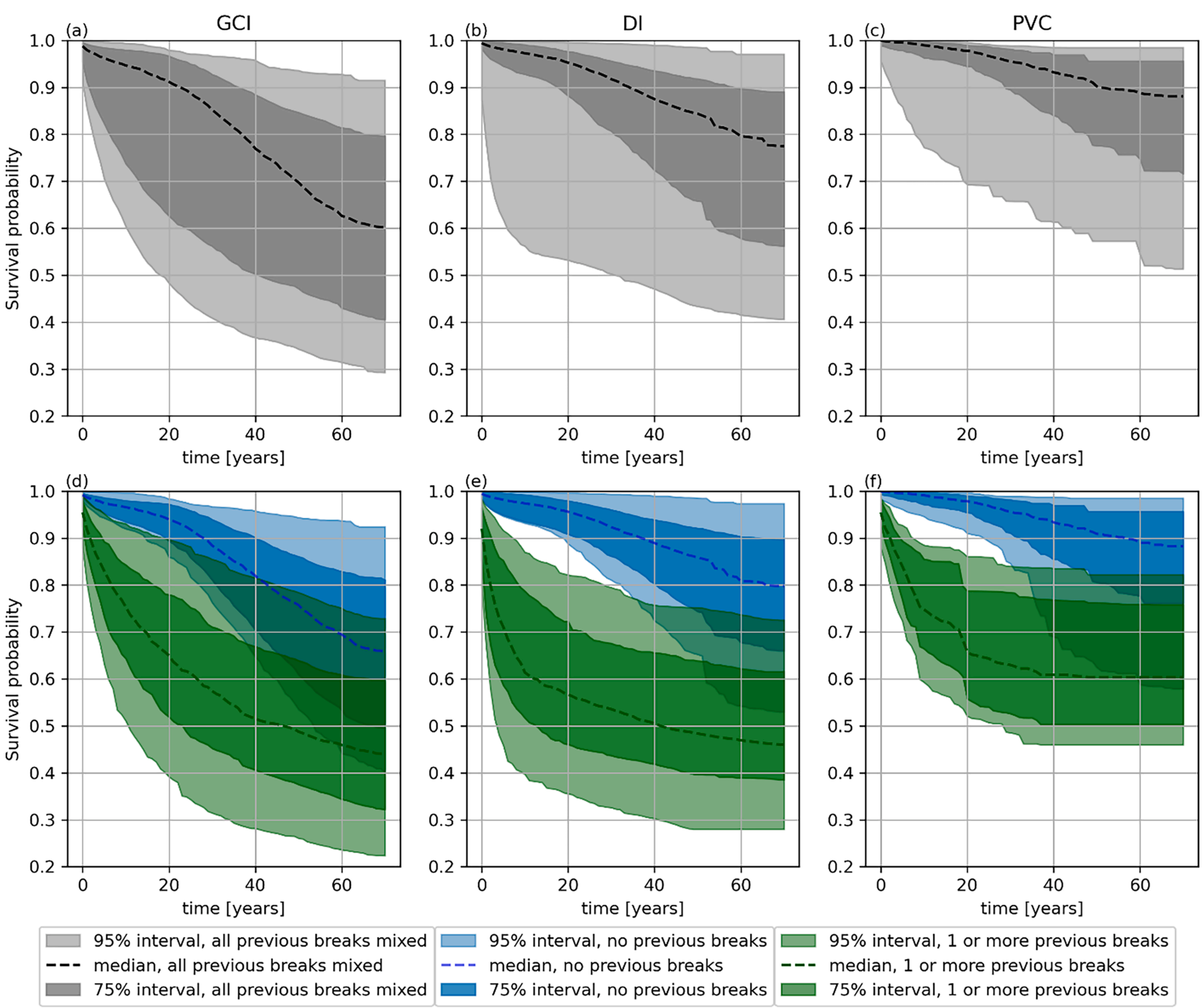


Fig. 7. The variation of survival curves as a proxy for uncertainties and pipe grouping based on the important variables.

uncertainties involved in the survival curves, Fig. 7 is drawn with 75% and 95% interval of survival curves in addition to the median curves. This figure further illustrates how uncertainties can be reduced by creating smaller and more homogeneous groups. In Fig. 7a, b, and c the pipes are distinguished by material alone, while in Fig. 7d, e, and f, they are further subdivided by the presence or absence of previous breaks. This way, uncertainties are significantly reduced, however the common practice of constructing cohorts based solely on material groups is not sufficient for our cases. Uncertainties can be further reduced by subdividing groups even more by length (e.g., length≤5 and length>5), material production standards, diameter and so on. However, in small utilities enough data might not be available to subdivide groups. In fact, groups should be as homogeneous as possible to minimize uncertainties, while still being large enough to maintain the statistical significance of findings [38]. In the case of limited data, pipe grouping should start with the variables that make groups as distinct as possible. Those variables are the most important variables that need to be identified through variable importance analysis.

The survival curves provide utilities with valuable information for both tactical and strategic planning as the results can be used both on a pipe and network level. Tactical planning involves short to medium-term planning, typically ranging up to 5 years, and focuses on identifying the most critical pipes that require rehabilitation or replacement. The survival curves provide utilities with a way to predict which pipes are likely to fail in the near future, allowing them to prioritize rehabilitation efforts based on the severity of each pipe's condition. One can combine the probability of failure with the consequences of failure to take risk-based decisions.

Strategic planning, on the other hand, is a long-term planning process that takes into account the overall condition of the network and budget constraints. The group survival probability predictions can be used for strategic rehabilitation planning by giving an overview of the time it takes for certain pipe groups to deteriorate and hence need to be replaced. Summing rehabilitation needs over the long term can provide an overview of the future of the water network condition and budget needs.

4.3. Analysis of variable importance

The results of permutation-based variable importance are depicted in Fig. 8. As also supported by the survival curves (Fig. 6 and Fig. 7), the number of previous breaks is the most influential variable among all the utilities followed by length and material type (i.e., GCI, DI, and PVC). A Pearson correlation analysis of the combined datasets (Appendix) also shows relatively high correlation of the number of previous breaks, material type (mainly grey cast iron) and length with the response variables (survival/break age, and status). These findings are aligned with the findings in Christodoulou et al. [14]. Findings in other literature can be different than the findings in this study. For instance, in Konstantinou and Stoianov [39], where they did not consider previous

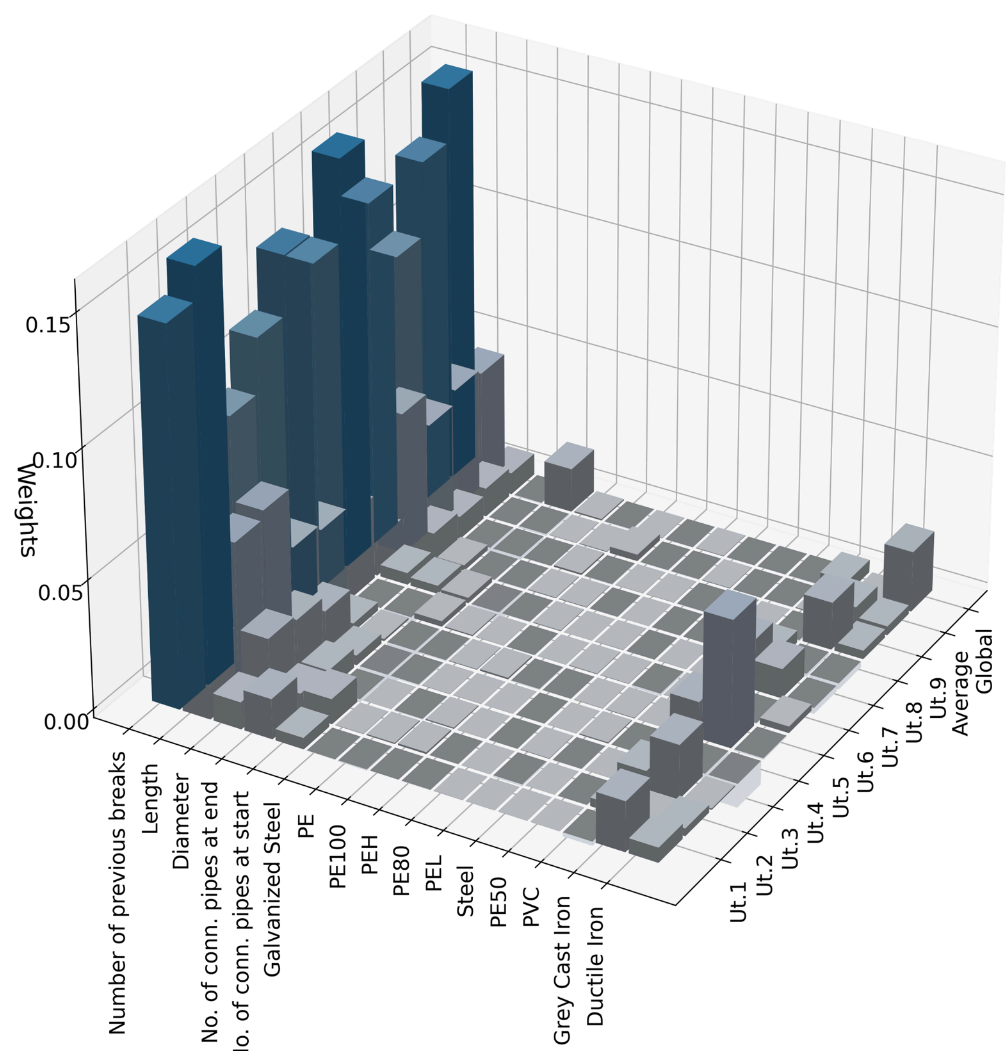


Fig. 8. Permutation-based variable importance.

breaks as a variable, the most influential variables were maximum pressure and pressure range, followed by material, diameter and length. A previous break probably weakens the structural strength of pipes that they cannot withstand higher loads anymore. Length is correlated with more exposed surface area. The more exposure to hazards the more risk of failure is expected. Length is reported by many literature as one of the most influential variables in pipe break predictions [39,55,79]. Pipe material represents different characteristics of pipes (e.g., structural strength, corrosion resistance) and many literatures trained separate models for each material type (e.g., [63]). In our study we trained one model for all material types and then their importance was highlighted in variable importance analysis. It should be noted that some material types (e.g., Galvanized Steel and some PE variations) have an insignificant number of samples in the dataset and hence do not contribute to the predictions. Furthermore, one might notice the absence of the age of pipe as a variable. In RSF, age is part of the structure of the model and is provided as a response variable. As the model uses Kaplan-Meier estimator to predict survival curves, without age RSF cannot be run. Hence, age is not only an important variable in RSF but also the required variable of the model. Age is reported by many literature [11,30,33,39,73] as the most or one of the most important variables.

We also trained a model based on utility 9′s full dataset (without dropping variables) which had gathered the widest range of variables (e.g., maximum pressure, annual average daily traffic, existence of quick clay, pipe depth, interior and exterior protection) and conducted a variable importance analysis to verify if the significance of the number of previous breaks still holds. The results (in Appendix) indicate that the performance of this model did not significantly improve, and the number of previous breaks still remains by far the most influential variable. However, the analysis was inconclusive in determining the second important variable. In fact, the importance of pipe length, number of pipes connected at ending manhole and material type, especially with respect to ductile iron pipes, was found to be relatively equal and were consistently competing for the second rank each time the model was run.

As the results show that the number of previous breaks is the most important variable, we considered removing it from the model to see how the models performs without the previous break information to test the applicability of the models without this information. Fig. 9 shows the performance of local and global models without previous break information. Comparing Fig. 9 with Fig. 4, the average values of C-index and C-index-ipcw dropped by around 0.15 and 0.06, respectively (it should be noted that the built-in function for the weights shown in Fig. 8 is based on the drop of C-index only). Hence, removing previous break information impacts the model performance considerably. This result further solidifies the importance of historical break data.

Historical break record data is defined and considered in different ways in literature. Snider and McBean [63] and Snider and McBean [62] included the number of previous breaks and also the age at the last, second last and third last breaks in their model. However, due to the high correlation between the number of previous breaks and the age at last break, we decided to drop the latter from our models. Fan et al. [20] included the number of previous breaks and the interval time to last break and found that the latter was the most important variable and the importance of the former was mediocre. However, they did not evaluate how the model performs without the interval time to last break as an explanatory variable. In our study, due to a high correlation between the number of previous breaks and the time to last break, it is difficult to determine which variable is more important. Including both variables in the analysis may be considered overfitting. The Pearson correlation coefficient between the two was 0.65. Additional statistical analysis revealed that pipes with no previous breaks typically break on average in 33 years. However, pipes with one, two, or three previous breaks are expected to break again within 13, 9, and 6 years, respectively.

There are other variables which are not considered in this study but are deemed important in literature. Specifically, this study lacks the environmental factors such as soil type (especially corrosive soils), weather-related variables (e.g., temperature, soil moisture, seasonality) [24,52,70,74] which are deemed to be important [4]. The existing literature lacks a holistic understanding of all the variables, along with their respective rankings, that influence the failure probability of pipes. Therefore, further research is merited in this area. The objective of this study is to investigate whether small municipalities, despite lacking many of the factors considered important in the literature, could still benefit from utility models and data available to them. Therefore, we have chosen to use only variables that are available to all datasets in the study.

5. Conclusion

To aid water utilities in their asset management approaches, this

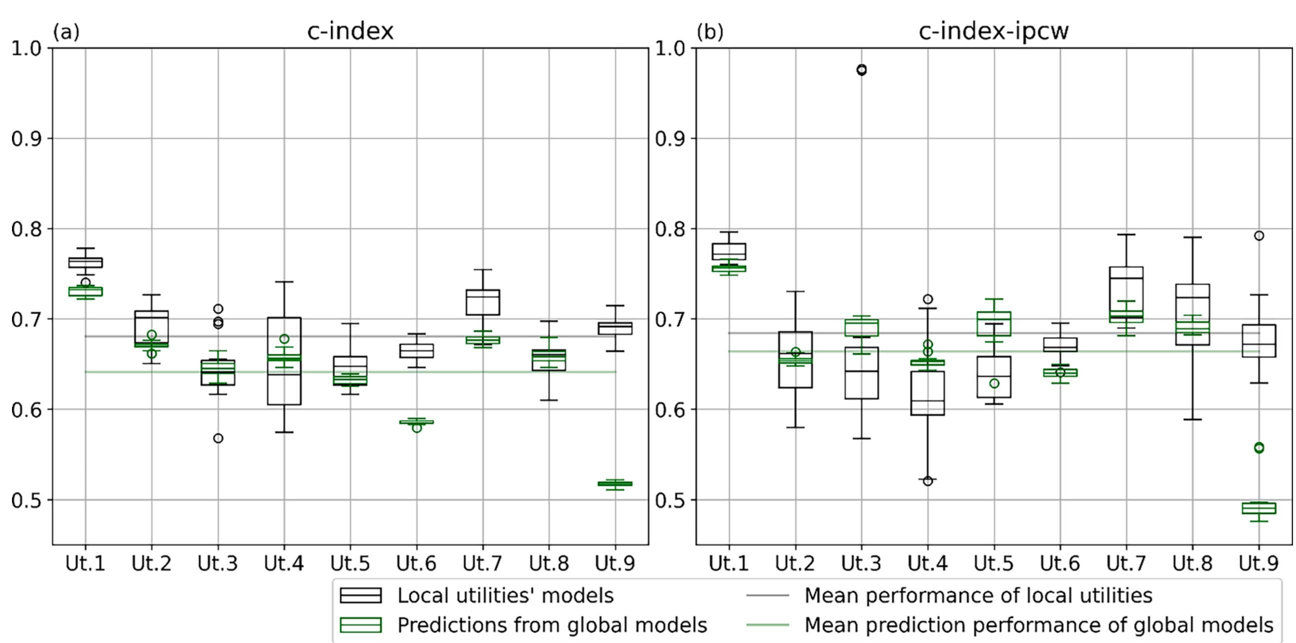


Fig. 9. Models' performance without the number of previous breaks.

study evaluated whether global and local deterioration models for water distribution pipes are generalizable and transferable between utilities, and their ability to predict pipe break probabilities for other utilities. Prior to training the models, several data pre-processing techniques were introduced to boost the performance of the models. The techniques introduced include addressing the problems of left-truncation and break data scarcity. We used a few explanatory variables (length, diameter, material type, previous break, number of connected pipes at the starting manhole and at the end) and age with break status as response variables. Our goal was to examine whether small municipalities with access to other utility models could benefit from them, despite the lack of many variables which are deemed important in the literature. Using RSF, nine local and nine global models were trained in a way that each global model excludes one utility's dataset from its training and then predicts the excluded utility's pipe breaks (time to next break). Moreover, the three largest utilities among the nine utilities were chosen as reference utilities to predict the pipe failures of the other utilities. The performance of the global models, local models and reference models were presented comparatively. Furthermore, survival curves for certain groups of pipes were drawn to see how survival probabilities change from one group to another and identify homogeneous groups of pipes. Finally, an analysis of variable importance was carried out to identify the most important variables so that utilities focus on the reliability of the most important variables while collecting data for them. From this study the following conclusions can be drawn:

1) The global models predict the pipe failures of other utilities with almost as good performance as individual models of the local utilities predicting their own pipe failures. However, the global models have not been checked with other countries' datasets (in cases where the material composition, history and local/climatic conditions differ) to test the full boundaries of usability. If the aforementioned conditions are similar, it can be a good idea to use a globally trained model and predict the pipe failures of small utilities where they do not have enough data to make a model on their own.
2) We can also use local reference models to predict other utilities, but their transferability between the utilities depends heavily on their size and degree of representativeness. The results of this study imply that the reference models perform better in predicting other utilities when they possess the largest database and are geographically close to the predicted utility. The use of the global model is safer in terms of performance and the geographical distance becomes less relevant. It is therefore a good idea for utilities to pool their data resources to establish global models for smaller utilities.
3) Pipes with and without previous breaks behave extremely different. Especially the first break is a turning point for the reliability of pipes. Material type (GCI, DI and PVC) has little influence on the next break after the first break happens. In general, the results show that the occurrence of previous breaks is a very important predictor. This means that the accuracy of survival curves can be improved considerably by considering previous breaks in addition to just the pipe material. From a practical point of view, the influence of previous breaks can also be used as a motivation for smaller utilities to collect data on failures and digitize old failure records – when they are aware that the previous breaks have such a big influence, and that data about this will help them being much more efficient in targeted rehabilitation of their systems, they should be inclined to invest resource in getting that data.
4) Ensuring the reliability of datasets before modeling is crucial. Most of the big errors in the datasets are identifiable with reasonable efforts. Modelers need to check the quality and reliability of data with different methods and look at it with different perspectives. Special care of the data of the most important variables (i.e., break data, age, length, and material) should be taken. They should be cleaned of errors and mistakes as much as possible.

Declaration of generative ai and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Shamsuddin Daulat: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Marius Møller Rokstad:** Conceptualization, Data curation, Methodology, Supervision, Writing – review & editing. **Stian Bruaset:** Conceptualization, Funding acquisition, Investigation, Writing – review & editing. **Jeroen Langeveld:** Supervision, Writing – review & editing. **Franz Tscheikner-Gratl:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgement

This work was supported by the Norwegian Research Council under the project B for VA-net (project number 309604).

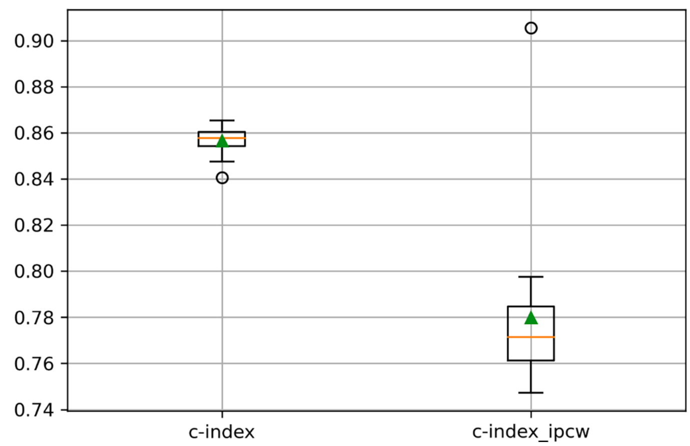
Appendices

Available data on variables in each utility

Variables	Variable type	Ut.1	Ut.2	Ut.3	Ut.4	Ut.5	Ut.6	Ut.7	Ut.8	Ut.9
Network type	Categorical	☒	☒	☐	☒	☒	☐	☒	☒	☒
Length	Numerical	☒	☒	☒	☒	☒	☒	☒	☒	☒
Diameter	Numerical	☒	☒	☒	☒	☒	☒	☒	☒	☒
Cross section shape	Categorical	☐	☐	☒	☐	☐	☐	☐	☐	☐
Joint type	Categorical	☐	☐	☒	☐	☐	☐	☐	☐	☒
Interior protection	Categorical	☐	☒	☐	☐	☐	☐	☐	☐	☒
Exterior protection	Categorical	☐	☐	☐	☐	☐	☐	☐	☐	☒
Quick clay	Numerical	☐	☒	☒	☒	☒	☒	☐	☒	☒
Nominal pressure	Categorical	☐	☐	☐	☒	☐	☐	☐	☐	☐
Ground surface material	Categorical	☐	☐	☐	☐	☐	☐	☐	☐	☒
Exterior mass	Categorical	☐	☐	☐	☐	☐	☐	☐	☐	☒
Zone	Categorical	☒	☐	☐	☐	☒	☒	☒	☒	☒
Max. Pressure	Numerical	☐	☐	☐	☐	☐	☐	☐	☐	☒
Pipe depth	Numerical	☐	☐	☐	☐	☐	☐	☐	☐	☒
No. of buildings in the vicinity	Numerical	☐	☐	☐	☐	☐	☐	☐	☐	☒
Traffic (AADT)	Numerical	☐	☐	☐	☐	☐	☐	☐	☐	☒
Material type	Categorical	☒	☒	☒	☒	☒	☒	☒	☒	☒
No. of conn. at start	Numerical	☒	☒	☒	☒	☒	☒	☒	☒	☒
No. of conn. at end	Numerical	☒	☒	☒	☒	☒	☒	☒	☒	☒
Number of previous breaks	Numerical	☒	☒	☒	☒	☒	☒	☒	☒	☒
Survival/break age	Numerical	☒	☒	☒	☒	☒	☒	☒	☒	☒
Status	Categorical	☒	☒	☒	☒	☒	☒	☒	☒	☒

Notes: (1) Any variable with more than 50% missing values are considered as unavailable in this table. (2) Categorical variables with 100% identical values are considered as unavailable, as the reliability of those values is questionable.

Utility 9′s model performance predicting parts of its own network, trained on all available variables in that utility



Permutation-based variable importance analysis of utility 9′s model

Rank	Explanatory variable	Weight
1	Number of previous breaks	0.115
2	No. of conn. pipes at end	0.033
3	Length	0.024
4	Max. pressure	0.018
5	Ductile iron	0.013
6	Interior protection	0.011
7	Traffic (AADT)	0.007
8	Quick clay	0.006
9	GSM=LE	0.006
10	Diameter	0.004
11	PVC	0.004
12	Exterior mass=unknown	0.003
13	Grey cast iron	0.003
14	Exterior protection	0.002
15	Pipe depth	0.002
16	No. of build in the vicinity	0.002
17	No. of conn. pipes at start	0.001
18	Joint type=SP	0.001
19	SONE	0.001
20	Joint type=Unknown Joint	0.000
21	Joint type =IM	0.000
22	Material=Others	0.000
23	Joint type =SE	0.000
24	GSM=FJ	0.000
25	GSM =unknown	0.000
26	Joint type =SV	0.000
27	Exterior mass =PM	0.000
28	PEH	0.000
29	PE50	0.000
30	Exterior mass =PF	0.000
31	Joint type =SS	0.000
32	Network type=*S*	0.000
33	Joint type =MUF	0.000
34	Joint type =MM	0.000
35	Joint type =IMU	0.000
36	Joint type =GJ	0.000
37	Joint type =FS	0.000
38	Joint type =ST	0.000
39	Material=Unknown	0.000
40	Joint type =TD	0.000
41	Joint type =TY	0.000
42	PEL	0.000
43	PE80	0.000
44	PE	0.000
45	Exterior mass =`	0.000
46	Exterior mass =TM	0.000
47	Exterior mass =SU	0.000

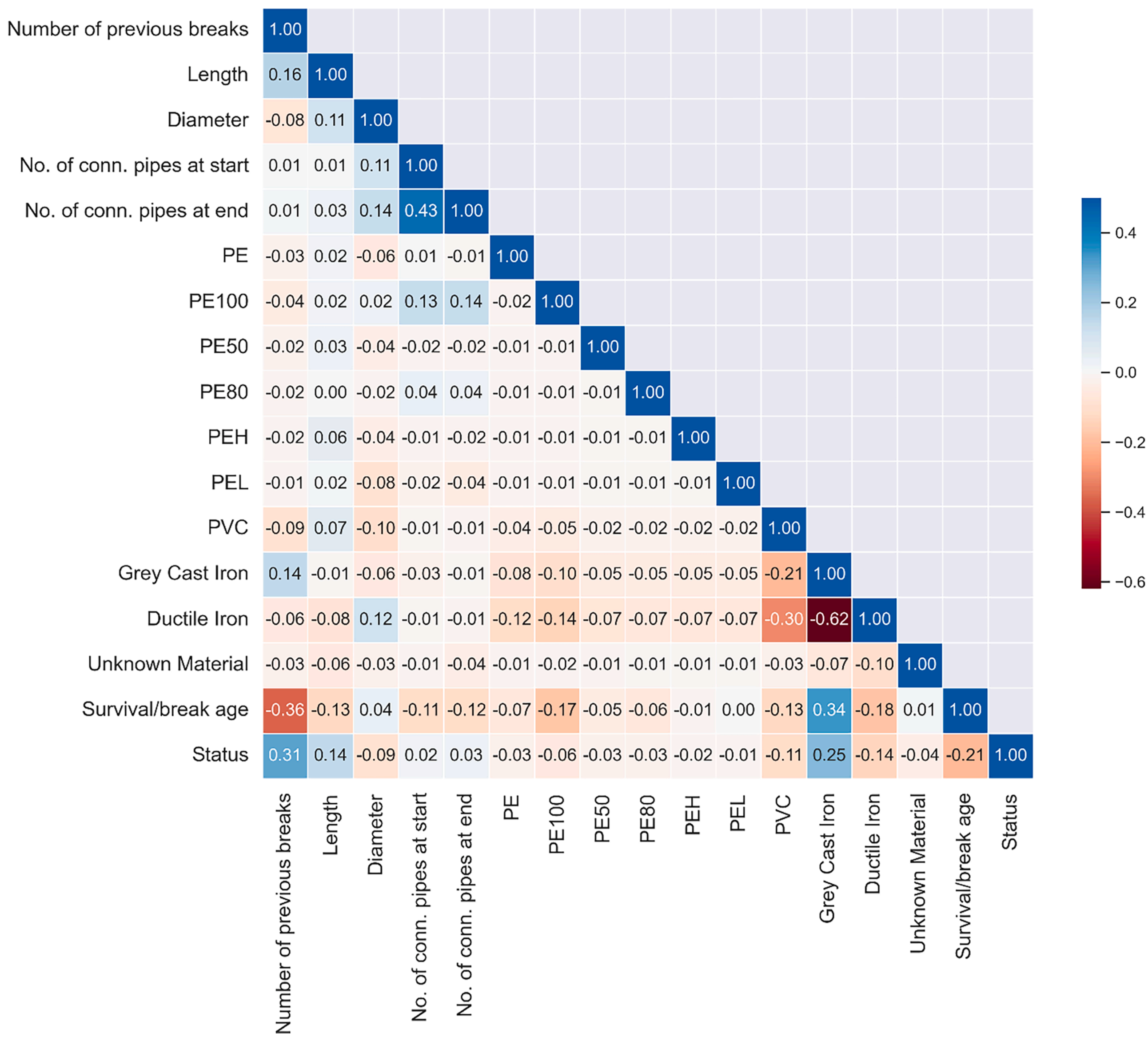
(*continued on next page*)

(*continued*)

Rank	Explanatory variable	Weight
48	Exterior mass =SA	0.000
49	Exterior mass =PU	0.000
50	Exterior mass =PG	0.000
51	Exterior mass =NN	0.000
52	GSM =*L*	0.000
53	Exterior mass =*L*	0.000
54	GSM =VA	0.000
55	GSM =TO	0.000
56	GSM =TM	0.000
57	GSM =SG	0.000
58	GSM =*S*	0.000
59	GSM =NN	0.000
60	GSM =MO	0.000
61	GSM =FS	0.000
62	PE100	0.000
63	GSM =OM	0.000
64	Galvanized Steel	0.000
65	Joint type =SM	0.000
66	Network type =*H*	−0.001

Abbreviations of the table.
AADT=Annual average daily traffic.
GSM=Ground surface material:.
TM=Peat/Swamp, VA=water, FA=Mountain, FJ=Mountain, FS=Fine Sand/Silt, LE=Clay, NN=Unknown, OM=Filled Mass, SG=Sand/Gravel, SJ=Sea.
Joint type:.
ST=Tensile joint, BL= Lead sleeve, BM=Bolt sleeve, FK=Fiber putty joint, FS=Flange joint, FA=False, GJ=The gang, IM=Push-in sleeve, LI=Glue joint, MM=Metal clamp sleeve, MØ=Mortar, MUF=Sleeve unspecified, SA=Cast asphalt, SM=Screw-in sleeve, SS=Cement packed, SV=Welding connection, SE=Welding Electrosocket, SP=Weld Mirror, TY=Tobacco joint, ÅP=Open.
Exterior mass:.
AV=Waste, GM=Gravel masses, GR=Gravel/single, NN=Unknown, PF= Fine gravel, PG=Coarse gravel, PM= Medium gravel, PU=Gravel, SA= Sandy, TM= Local masses.
Network type:.
F=Common pipes, *H*=Main pipes, *O*=Transmission pipes, *S*=connection pipes.

Pearson correlation between each two mutual variables for the combined dataset



References

[1] Aalen O. Nonparametric inference for a family of counting processes. Ann Stat 1978;6:701–26.

[2] Almheiri Z, Meguid M, Zayed T. Failure modeling of water distribution pipelines using meta-learning algorithms. Water Res 2021;205:117680. https://doi.org/10.1016/j.watres.2021.117680.

[3] ASCE. 2021 report card for America's infrastructure: a comprehensive assessment of america's infrastructurE. American Society of Civil Engineers (ASCE); 2021.

[4] Barton NA, Farewell TS, Hallett SH, Acland TF. Improving pipe failure predictions: factors effecting pipe failure in drinking water networks. Water Res 2019;164:114926. https://doi.org/10.1016/j.watres.2019.114926.

[5] Berardi L, Giustolisi O, Kapelan Z, Savic DA. Development of pipe deterioration models for water distribution systems using EPR. J Hydroinf 2008;10:113–26. https://doi.org/10.2166/hydro.2008.012.

[6] Beuken R, Eijkman J, Savic D, Hummelen A, Blokker M. Twenty years of asset management research for Dutch drinking water utilities. Water Sci Technol: Water Supply 2020;20:2941–50. https://doi.org/10.2166/ws.2020.179.

[7] Breiman L. Random Forests. Machine Learning 2001;45:5–32.

[8] Bruaset S. Long-term sustainable management of the urban water and wastewater pipe networks. doctoral dissertation. Norwegian University of Science and Technology; 2019.

[9] Caradot N, Riechel M, Fesneau M, Hernandez N, Torres A, Sonnenberg H, Eckert E, Lengemann N, Waschnewski J, Rouault P. Practical benchmarking of statistical and machine learning models for predicting the condition of sewer pipes in Berlin, Germany. J Hydroinf 2018;20:1131–47. https://doi.org/10.2166/hydro.2018.217.

[10] Carrión A, Solano H, Gamiz ML, Debón A. Evaluation of the Reliability of a Water Supply Network from Right-Censored and Left-Truncated Break Data. Water Resour Manage 2010;24:2917–35. https://doi.org/10.1007/s11269-010-9587-y.

[11] Carvalho G, Amado C, Brito RS, Coelho ST, Leitão JP. Analysing the importance of variables for sewer failure prediction. Urban Water J 2018;15:338–45. https://doi.org/10.1080/1573062X.2018.1459748.

[12] Chen TY-J, Guikema SD. Prediction of water main failures with the spatial clustering of breaks. Reliab Eng Syst Safety 2020;203:107108. https://doi.org/10.1016/j.ress.2020.107108.

[13] Chen TY-J, Vladeanu G, Yazdekhasti S, Daly CM. Performance Evaluation of Pipe Break Machine Learning Models Using Datasets from Multiple Utilities. J Infrastruct Syst 2022;28. https://doi.org/10.1061/(asce)is.1943-555x.0000683.

[14] Christodoulou S, Deligianni A, Aslani P, Agathokleous A. Risk-based asset management of water piping networks using neurofuzzy systems. Computers, Environ Urban Syst 2009;33:138–49. https://doi.org/10.1016/j.compenvurbsys.2008.12.001.

[15] Clark RM, Stafford CL, Goodrich JA. Water distribution systems: a spatial and cost evaluation. J Water Resour Plann Manage Division 1982;108(3):243–56.

[16] Dandy GC, Engelhardt M. V Aluation and O Ptimal S Trategies of. J Water Resour Plann Manage 2001;127:214–23.

[17] Daulat S, Rokstad MM, Klein-Paste A, Langeveld J, Tscheikner-Gratl F. Challenges of integrated multi-infrastructure asset management: a review of pavement, sewer, and water distribution networks. Struct Infrastruct Eng 2022;0:1–20. https://doi.org/10.1080/15732479.2022.2119480.

[18] Dawood T, Elwakil E, Novoa HM, Delgado JFG. Water pipe failure prediction and risk models: state-of-the-art review. Can J Civ Eng 2020;47:1117–27. https://doi.org/10.1139/cjce-2019-0481.

[19] Dawood T, Elwakil E, Novoa HM, Delgado JFG. Artificial intelligence for the modeling of water pipes deterioration mechanisms. Autom Constr 2020;120: 103398. https://doi.org/10.1016/j.autcon.2020.103398.

[20] Fan X, Wang X, Zhang X, , ASCE, Yu Xiong (Bill), P.E.F.. Machine learning based water pipe failure prediction: the effects of engineering, geology, climate and socio-economic factors. Reliab Eng Syst Safety 2022;219:108185. https://doi.org/10.1016/j.ress.2021.108185.

[21] Fan X, Zhang X, Yu X. Uncertainty quantification of a deep learning model for failure rate prediction of water distribution networks. Reliab Eng Syst Saf 2023: 109088. https://doi.org/10.1016/j.ress.2023.109088.

[22] Fares H, Zayed T. Hierarchical Fuzzy Expert System for Risk of Failure of Water Mains. J Pipeline Syst Eng Pract 2010;1:53–62. https://doi.org/10.1061/(asce)ps.1949-1204.0000037.

[23] Francisque A, Tesfamariam S, Kabir G, Haider H, Reeder A, Sadiq R. Water mains renewal planning framework for small to medium sized water utilities: a life cycle cost analysis approach. Urban Water J 2017;14:493–501. https://doi.org/10.1080/1573062X.2016.1223321.

[24] Gould SJF, Boulaire FA, Burn S, Zhao XL, Kodikara JK. Seasonal factors influencing the failure of buried water reticulation pipes. Water Sci Technol 2011;63:2692–9. https://doi.org/10.2166/wst.2011.507.

[25] Haider H, Sadiq R, Tesfamariam S. Intra-utility performance management model (In-UPM) for the sustainability of small to medium sized water utilities: conceptualization to development. J Cleaner Prod 2016;133:777–94. https://doi.org/10.1016/j.jclepro.2016.06.003.

[26] Haider H, Sadiq R, Tesfamariam S. Performance indicators for small-and medium-sized water supply systems: a review. Environ Rev 2014;22:1–40. https://doi.org/10.1139/er-2013-0013.

[27] Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361:: AID-SIM168>3.0.CO;2-4.

[28] Harvey R, McBean EA, Gharabaghi B. Predicting the Timing of Water Main Failure Using Artificial Neural Networks. J Water Resour Plann Manage 2014;140:425–34. https://doi.org/10.1061/(asce)wr.1943-5452.0000354.

[29] Iannacone L, Gardoni P. Physics-based repair rate curves for segmented pipelines subject to seismic excitations. Sustain Resilient Infrastruct 2023;8:121–41. https://doi.org/10.1080/23789689.2021.2000146.

[30] Iannacone L, Sharma N, Tabandeh A, Gardoni P. Modeling Time-varying Reliability and Resilience of Deteriorating Infrastructure. Reliab Eng Syst Safety 2022;217: 108074. https://doi.org/10.1016/j.ress.2021.108074.

[31] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008;2:841–60. https://doi.org/10.1214/08-AOAS169.

[32] Jara-Arriagada C, Stoianov I. Pipe breaks and estimating the impact of pressure control in water supply networks. Reliab Eng Syst Safety 2021;210:107525. https://doi.org/10.1016/j.ress.2021.107525.

[33] Kabir G, Balek NBC, Tesfamariam S. Sewer Structural Condition Prediction Integrating Bayesian Model Averaging with Logistic Regression. J Perform Constr Facil 2018;32. https://doi.org/10.1061/(ASCE)CF.1943-5509.0001162. 04010019-1-04010019–10.

[34] Kabir G, Tesfamariam S, Hemsing J, Sadiq R. Handling incomplete and missing data in water network database using imputation methods. Sustain Resilient Infrastruct 2020;5:365–77. https://doi.org/10.1080/23789689.2019.1600960.

[35] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Statist Assoc 1958;53:457–81.

[36] Kleinbaum DG, Klein M. Survival analysis a self-learning text. New York: Springer; 2012.

[37] Kleiner Y, Rajani B. Comprehensive review of structural deterioration of water mains: statistical models Rajani, B.B.; Kleiner, Y. NRCC-43722. Urban Water 2001; 3:151–64.

[38] Kleiner Y, Rajani B, Wang S. Consideration of static and dynamic effects to plan water main renewal. In: Middle East Water 2007, 4th International Exhibition and Conference for Water Technology; 2007. p. 1–13.

[39] Konstantinou C, Stoianov I. A comparative study of statistical and machine learning methods to infer causes of pipe breaks in water supply networks. Urban Water J 2020;17:534–48. https://doi.org/10.1080/1573062X.2020.1800758.

[40] Laakso T, Kokkonen T, Mellin I, Vahala R. Sewer life span prediction: comparison of methods and assessment of the sample impact on the results. Water (Switzerland) 2019;11:1–14. https://doi.org/10.3390/W11122657.

[41] Le Gat Y. Extending the Yule process to model recurrent pipe failures in water supply networks. Urban Water J 2014;11:617–30. https://doi.org/10.1080/1573062X.2013.783088.

[42] Longato E, Vettoretti M, Di Camillo B. A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. J Biomed Inform 2020;108:103496. https://doi.org/10.1016/j.jbi.2020.103496.

[43] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Rep 1966;50:163–70.

[44] Marlow DR, Moglia M, Cook S, Beale DJ. Towards sustainable urban water management: a critical reassessment. Water Res 2013;47:7150–61. https://doi.org/10.1016/j.watres.2013.07.046.

[45] Chapter 1 - The History of Text Mining. In: Miner G, Delen D, Elder J, Fast A, Hill T, Nisbet RA, editors. Practical text mining and statistical analysis for non-structured text data applications. Boston: Academic Press; 2012. p. 3–27. https://doi.org/10.1016/B978-0-12-386979-1.00001-3.

[46] Nelson W. Theory and Applications of Hazard Plotting for Censored Failure Data. Technometrics 1972;14:945–66. https://doi.org/10.1080/00401706.1972.10488991.

[47] Nishiyama M, Filion Y. Review of statistical water main break prediction models. Can J Civ Eng 2013;40:972–9. https://doi.org/10.1139/cjce-2012-0424.

[48] OECD. Financing water supply, sanitation and flood protection: challenges in eu member states and policy options. Paris: OECD Studies on Water, OECD Publishing; 2020. https://doi.org/10.1787/6893cdac-en.

[49] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine Learning in Python. J Mach Learn Res 2011;12:2825–30.

[50] Piratla KR, Yerri SR, Yazdekhasti S, Cho J, Koo D, Matthews JC. Empirical Analysis of Water-Main Failure Consequences. Procedia Eng 2015;118:727–34. https://doi.org/10.1016/j.proeng.2015.08.507.

[51] Pölsterl S. Scikit-survival: a library for time-to-event analysis built on top of scikit-learn. J Mach Learn Res 2020;21:1–6.

[52] Rajani B, Kleiner Y, Sink J-E. Exploration of the relationship between water main breaks and temperature covariates. Urban Water J 2012;9:67–84. https://doi.org/10.1080/1573062X.2011.630093.

[53] Ramos-Salgado C, Muñuzuri J, Aparicio-Ruiz P, Onieva L. A comprehensive framework to efficiently plan short and long-term investments in water supply and sewer networks. Reliab Eng Syst Safety 2022;219. https://doi.org/10.1016/j.ress.2021.108248.

[54] Rifaai TM, Abokifa AA, Sela L. Integrated approach for pipe failure prediction and condition scoring in water infrastructure systems. Reliab Eng Syst Safety 2022;220: 108271. https://doi.org/10.1016/j.ress.2021.108271.

[55] Robles-Velasco A, Cortés P, Muñuzuri J, Onieva L. Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliab Eng Syst Safety 2020;196:106754. https://doi.org/10.1016/j.ress.2019.106754.

[56] Røstum J. Statistical modelling of pipe failures in water networks. Norwegian University of Science and Technology. Norwegian University of Science and Technology; 2000.

[57] Scheidegger A, Leitão JP, Scholten L. Statistical failure models for water distribution pipes - A review from a unified perspective. Water Res 2015;83: 237–47. https://doi.org/10.1016/j.watres.2015.06.027.

[58] Scheidegger A, Scholten L, Maurer M, Reichert P. Extension of pipe failure models to consider the absence ofdata from replaced pipes. Water Res 2013;47:3696–705. https://doi.org/10.1016/j.watres.2013.04.017.

[59] Shaban IA, Eltoukhy AEE, Zayed T. Systematic and scientometric analyses of predictors for modelling water pipes deterioration. Autom Constr 2023;149: 104710. https://doi.org/10.1016/j.autcon.2022.104710.

[60] Shamir U, Howard CDD. An analytic approach to scheduling pipe replacement. Am. Water Works Assoc. J. 1979;71:248–58. https://doi.org/10.1002/j.1551-8833.1979.tb04345.x.

[61] Shortridge JE, Guikema SD. Public health and pipe breaks in water distribution systems: analysis with internet search volume as a proxy. Water Res 2014;53: 26–34. https://doi.org/10.1016/j.watres.2014.01.013.

[62] Snider B, McBean EA. Assessing the Impact of Pipe Rehabilitation on Decreasing Watermain Break Rates Using Random Survival Forest Models. J Water Resour Plann Manage 2022;148. https://doi.org/10.1061/(asce)wr.1943-5452.0001579.

[63] Snider B, McBean EA. Combining Machine Learning and Survival Statistics to Predict Remaining Service Life of Watermains. J Infrastruct Syst 2021;27:1–14. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000629.

[64] SSB Norway. Municipal water supply [WWW document]. 2022. URL, https://www.ssb.no/en/natur-og-miljo/vann-og-avlop/statistikk/kommunal-vannforsyning. accessed 8.1.22.

[65] St. Clair AM, Sinha S. State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models! Urban Water J 2012;9:85–112. https://doi.org/10.1080/1573062X.2011.644566.

[66] Tabesh M, Soltani J, Farmani R, Savic D. Assessing pipe failure rate and mechanical reliability of water distribution networks using data-driven modeling. J Hydroinf 2009;11:1–17. https://doi.org/10.2166/hydro.2009.008.

[67] Tscheikner-Gratl F, Sitzenfrei R, Rauch W, Kleidorfer M. Integrated rehabilitation planning of urban infrastructure systems using a street section priority model. Urban Water J 2016;13:28–40. https://doi.org/10.1080/1573062X.2015.1057174.

[68] Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 2011;30:1105–17. https://doi.org/10.1002/sim.4154.

[69] Walski TM, Pelliccia A. Economic Analysis of Water Main Breaks. J /Am Water Works Assoc 1982;74:140–7. https://doi.org/10.1002/j.1551-8833.1982.tb04874.x.

[70] Wasim M, Shoaib S, Mubarak NM, Inamuddin, Asiri AM. Factors influencing corrosion of metal pipes in soils. Environ Chem Lett 2018;16:861–79. https://doi.org/10.1007/s10311-018-0731-x.

[71] Wiesenfeld BM, Aphinyanaphongs Y, Nov O. AI model transferability in healthcare: a sociotechnical perspective. Nat Mach Intell 2022;4:807–9. https://doi.org/10.1038/s42256-022-00544-x.
[72] Wilson D, Filion Y, Moore I. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. Urban Water J 2017;14:173–84. https://doi.org/10.1080/1573062X.2015.1080848.
[73] Winkler D, Haltmeier M, Kleidorfer M, Rauch W, Tscheikner-Gratl F. Pipe failure modelling for water distribution networks using boosted decision trees. Struct Infrastruct Eng 2018;14:1402–11. https://doi.org/10.1080/15732479.2018.1443145.
[74] Wols BA, Thienen Pvan. Impact of weather conditions on pipe failure: a statistical analysis. J Water Supply : Res Technol - AQUA 2014;63:212–23. https://doi.org/10.2166/aqua.2013.088.
[75] Wood A, Lence BJ. Using Water Main Break Data to Improve Asset Management for Small and Medium Utilities: district of Maple Ridge, B.C. J Infrastruct Syst 2009; 15:111–9. https://doi.org/10.1061/(asce)1076-0342(2009)15:2(111).
[76] Wood A, Lence BJ. Using Water Main Break Data to Improve Asset Management for Small and Medium Utilities: district of Maple Ridge, B.C. J Infrastruct Syst 2009; 15:111–9. https://doi.org/10.1061/(asce)1076-0342(2009)15:2(111).
[77] Xu H, Sinha SK. Modeling Pipe Break Data Using Survival Analysis with Machine Learning Imputation Methods. J Perform Constr Facil 2021;35. https://doi.org/10.1061/(asce)cf.1943-5509.0001649.
[78] Xu H, Sinha SK. Applying Survival Analysis to Pipeline Data: gaps and Challenges. Pipelines 2020. ASCE; 2020. p. 148–58.
[79] Yamijala S, Guikema SD, Brumbelow K. Statistical models for the analysis of water distribution system pipe break data. Reliab Eng Syst Saf 2009;94:282–93. https://doi.org/10.1016/j.ress.2008.03.011.
[80] Yuan X. Principles and guidelines of deterioration modelling for water and waste water assets. Infrastruct Asset Manage 2016;4:19–35. https://doi.org/10.1680/jinam.16.00017.