Water Supply

IWA PUBLISHING

Water Supply Vol 22 No 1, 527 doi: 10.2166/ws.2021.255

The challenges of predicting pipe failures in clean water networks: a view from current practice

N. A. Barton, S. H. Hallett* and S. R. Jude

School of Water, Energy and Environment, Cranfield University, Bedfordshire MK43 0AL, United Kingdom
*Corresponding author. E-mail: s.hallett@cranfield.ac.uk

NAB, 0000-0002-1822-3590; SHH, 0000-0002-8776-7049

ABSTRACT

Pipe failure models can aid proactive management decisions and help target pipes in need of preventative repair or replacement. Yet, there are several uncertainties and challenges that arise when developing models, resulting in discord between failure predictions and those observed in the field. This paper aims to raise awareness of the main challenges, uncertainties, and potential advances discussed in key themes, supported by a series of semi-structured interviews undertaken with water professionals. The main discussion topics include data management, data limitations, pre-processing difficulties, model scalability and future opportunities and challenges. Improving data quality and quantity is key in improving pipe failure models. Technological advances in the collection of continuous real-time data from ubiquitous smart networks offer opportunities to improve data collection, whilst machine learning and data analytics methods offer a chance to improve future predictions. In some instances, technological approaches may provide better solutions to tackling short term proactive management. Yet, there remains an opportunity for pipe failure models to provide valuable insights for long-term rehabilitation and replacement planning.

Key words: clean water, data analytics, infrastructure planning, smart networks, water supply

HIGHLIGHTS

- A variety of pipe failure model challenges and uncertainties are discussed.
- Industry professional interviews provide deeper understanding of the issues.
- Data management, quality, pre-processing and scalability influence modelling.
- Future opportunities include the benefits of real-time data, data analytics and machine learning.

1. INTRODUCTION

Pipe failures result in billions of litres lost from water networks each day, which wastes water, causes damage to infrastructure, and interrupts continuous service. Considerable effort is being concentrated on minimising water loss and improving services through performance commitments aimed to reduce pipe failures (Ramirez *et al.* 2020; Robles-Velasco *et al.* 2020). There has been a recent focus on innovation, partly driven by advances in technology and through regulator incentives (Ofwat 2020) as a way of catalysing change in the industry. In consequence, water companies are moving towards data analytics and statistical models (referred to hereafter as pipe failure models) to provide insights for proactive management decisions. Pipe failure models represent a distinct field of data analytics, and although commencing with Shamir & Howard (1979), they are still considered innovative today with the emergence and application of machine learning approaches and big data.

Pipe failure models provide a means to understand future network performance, discerning patterns from historical data, describing the response as either the probability of failure (probability), time-to-failure (survival analysis), pipe failure rates (regression) or whether or not a failure will occur (classification) (Economou *et al.* 2012). Many studies report useful insights that inform proactive (also known as preventative) management decisions. Nonetheless, uncertainties are an integral part of prediction modelling, and underlying many pipe failure studies are issues surrounding data quality and quantity and model

development. As such, there remains a degree of scepticism of pipe failure models (Kleiner & Rajani 2012; Konstantinou & Stoianov 2020) and the discord between failure predictions and those observed in the field, potentially hindering effective decision making.

Water networks have complex failure histories, and studies have often reported poor data quality (Mailhot *et al.* 2000; Kabir *et al.* 2016; Tang *et al.* 2019). In particular, the timing and location of pipe failures (Wu & Liu 2017), the limited failure history and absence of decommissioned pipe data, and limited variables leading to a lack of understanding about the various factors influencing pipe failures (Scheidegger *et al.* 2015). Model scalability is perhaps the most complex challenge in pipe failure assessment, a theme widely discussed in the literature due to the gravitas of its inherent effect on predictive accuracy (Kleiner & Rajani 2012). Due to these challenges, pre-processing steps are often required, including relocating of pipe failures, handling of missing and inaccurate data, and scaling the model appropriately by, employing pipe grouping methods at the network or census level (Chen *et al.* 2019), k-means clustering (Kakoudakis *et al.* 2018), or recutting pipes to improve model accuracy (Winkler *et al.* 2018). Understanding these challenges is vital for both managers and professionals involved in developing pipe failure models, yet whilst literature has acknowledged some difficulties, to our knowledge, there are few studies explicitly discussing these challenges holistically. Furthermore, it is warranted that these issues should be further explored from the perspective of industry professionals, which to our knowledge has not been documented. This perspective offers a more in-depth understanding of the challenges.

While general study limitations and uncertainties are reported in the literature, it is important to consider the issues faced, from data collection and management to specific data processing solutions, all of which affect the predicted outcome. Thus, this study discusses a holistic view of the various challenges providing further understanding and insight from current practices in water companies through a series of semi-structured interviews. Accordingly, the study is discussed in five key themes: (i) challenges of data management, (ii) limitations of pipe failure data, (iii) complexities of pre-processing, (iv) understanding model scalability, and (v) future opportunities and challenges.

2. METHODOLOGY

Water professionals from UK water companies were interviewed using a semi-structured approach to reveal views on the selected themes. The aim was to identify key issues and challenges faced to complement and support the topic discussions formed from the literature and an understanding of pipe failure models. The semi-structured interviews were based on the practical guidelines detailed in Adams (2015). The interview guide was pre-defined based on previous knowledge of the subject area (Barton *et al.* 2019, 2020) and the desire to explore topics based on these findings. In this respect, the methodological approach adopted is similar to deductive qualitative content analysis, whereby the structure of the analysis was based on previous knowledge and the interview guide designed to understand and characterise extant themes and situations. The authors use manifest analysis, whereby quotes of participants are used to provide further detailed understanding on a theme of interest (Bengtsson 2016). The semi-structured nature allowed participants, by design, the freedom to diverge and elaborate beyond a certain point, revealing further information of relevance (Adams 2015). Consent was obtained from all participants, and anonymity provided. A pilot semi-structured interview based on the questions was completed on two water industry professionals before commencement on the primary cohort of participants. The semi-structured interviews were undertaken between January 2021 and April 2021 and involved eight participants drawn from different UK companies. Each interview was restricted to approximately one hour to reduce participant and interviewer fatigue (participant details are provided in Table 1). The interviews were recorded and transcribed so relevant information for each of the themes could be included.

The number of participants in semi-structured interviews is often discussed, but it is common for studies to have between one and 30 participants (Bengtsson 2016). Galvin (2015) reviewed several engineering articles and found the average number of participants was between eight and 17, but further found several examples where the number of participants was lower, as low as two in some cases. For homogenous groups where all participants had a familiarity within the study area, Guest *et al.* (2006) found a lower number of participants was sufficient. Noting this, the number of participants should be determined based on conclusive answers to the research questions with sufficient confidence, to the point of information saturation, which occurs when no new information is being observed (Ritchie *et al.* 2003; Fusch & Ness 2015; Bengtsson 2016). Many professional practitioners were asked to participate in this study. The participants selected had no previous involvement in the project and were recruited through 'snowball' methods. Eight identified themselves as topic experts. The mix of data scientists, asset managers and leakage analysts constitute enough participants to collect the desired information,

Table 1 | Participants of the semi-structured interview

ID	Area of expertise in clean water network management	Sector	Interview duration
P1	Data scientist	Water utility	1:07:23
P2	Network modeller	Water utility	59:58
P3	Leakage analyst	Water utility	52:09
P4	Network modeller	Water utility	55:53
P5	Asset manager	Consultant	57:12
P6	Leakage analyst	Consultant	58:41
P7	Data scientist	Water utility	1:01:02
P8	Network manager	Water utility	45:14

mainly due to the distinct topics, the participants' deep expertise providing a rich quality and quantity of information, and the length and design of the interviews.

3. RESULTS AND DISCUSSION

The results are structured based on the key themes outlined in the introduction, divided further by areas of interest derived from the interviews. The themes and illustrative quotes identify the challenges and uncertainty seen in pipe failure models from the perspective of professionals. Figure 1 shows a systematic map of the themes and sub-themes discussed hereafter.

3.1. Challenges of managing data

Many Water Distribution Networks (WDN) consist of an antiquated asset base nearing the end of their engineering life span (Snider & McBean 2019). Over the past century, upgrades, repairs, and industry regulation and management changes have

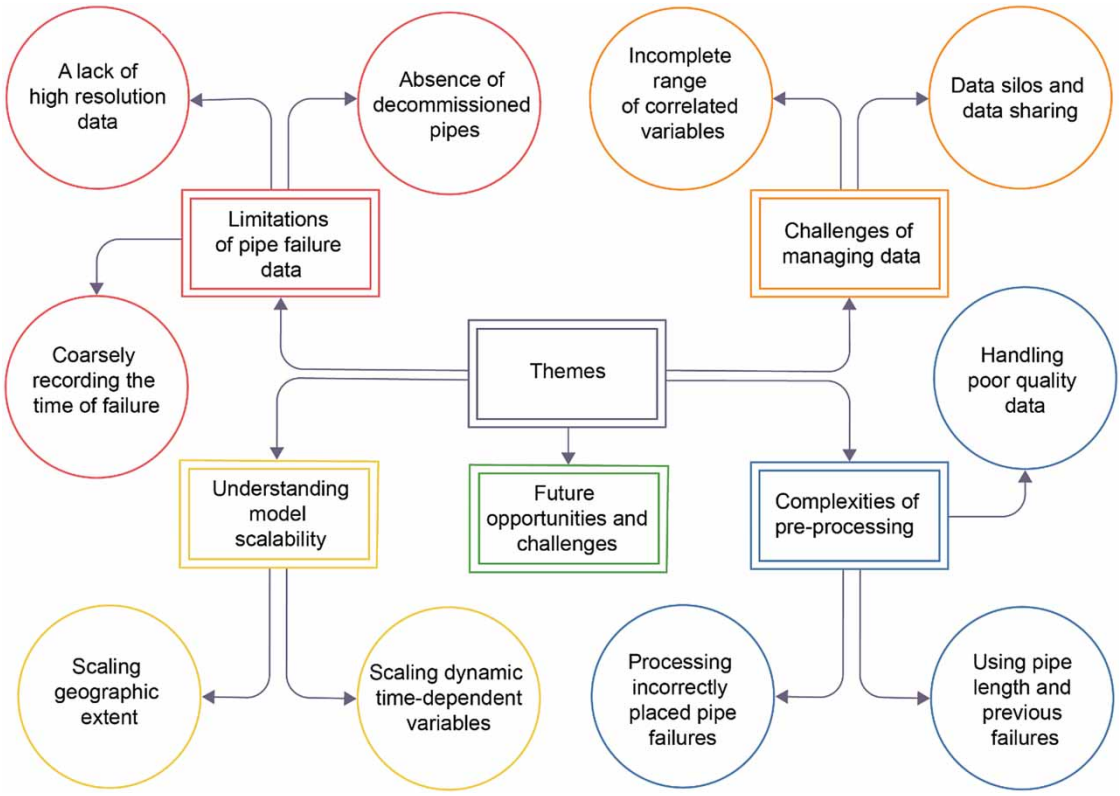

Figure 1 | Systematic map of the themes and sub-themes explored in this study.

led to a complex non-linear network comprised of discontinuous pipe materials, requiring careful documentation of asset status information. Over time, data collected on the constantly evolving WDNs have grown organically with the needs of utility companies, from the original paper asset records to more widely captured contemporary data using advanced real-time sensors and digital storage. Whilst current data is rigorously checked and is generally reliable, much of the information on the historical network construction remains limited, often in analogue form, or of poor quality and is often highly variable in terms of data accuracy, coverage, scale, and format, which significantly limits its quality, accessibility, and usability (Likhari *et al.* 2017). All participants acknowledged historical data quality as an issue, with one suggesting *'The actual quality of the* [historical] *data will be limited.'* [P4]. The unreliable nature of the data creates two apparent problems. Firstly, the lack of historical failure data results in imbalanced data that is problematic to model. Due to the lack of historical data, attempts have been made at pooling data from different water companies to increase the size of data sets, such as the UKWIR National Mains Failure Database (NMFD) (UKWIR 2020). Yet much effort would need to be expended to clean, process, and harmonise the disparate databases, often collected through different formats and processes unique to each water company. Previous attempts to model using this data have shown poor results (Tang *et al.* 2019). Renaud *et al.* (2007) tried a similar approach by combining data from water companies across France to overcome the scarce data issue and found no marked improvement. Secondly, there is a question of geodetic reliability and accuracy of pipe locations recorded before using Geographical Information Systems (GIS) and Global Navigation Satellite System (GNSS). Participants highlighted the issues with historical pipe locations, with one highlighting the problem with finding pipes when they failed, *'Trying to locate this old pipe, even with ground penetrating radar, we had no chance of finding it at all.'* [P6]. Where pipes are accurately located, there is speculation over the level of accuracy in the information, such as the recorded material or the pipe diameter with one participant suggesting '*What material will be laid won't be accurate, water companies will do their best, but it won't all be right.'* [P4]. Similar observations were made on the network data, which is constantly evolving and updating, making it difficult to maintain the new asset data and pipe replacements, meaning failure models based on older network data will not be applied to the most recent network and may be identifying pipes already replaced or rehabilitated, *'The GIS system should have accurate existing data on an asset, but there are multiple versions across departments where work is constantly being done.'* [P2]. However, it was suggested by the respondents that considerable progress has been made in this respect, and water companies are making concerted efforts to improve network data, mainly aided by advances in GIS.

The accuracy of historical data is essential to modern analytical approaches. While problems persist, it is evident most water companies have directed resources into ensuring the best quality possible. Yet the quality of historical data will remain an issue: *'The problem is the [historical] data can't change, and it's very difficult to change it retrospectively. So, you're only really talking about improving the way things get done in the future and not improving past data.*' [P5]. Most importantly, all participants had commenced concerted efforts to capture future data more accurately, acknowledging the importance of data accuracy for gaining valuable insights.

In the UK, the water sector was privatised and in the process government regulators were introduced, which led to significant improvements in data collection and quality (Ofwat 2006) but in turn presented new challenges. The introduction of Asset Management Plan (AMP) cycles was created as a means of permit regulatory price reviews and performance assessment. Regulatory changes between AMP cycles and the need to deliver business value and efficiency whilst reporting to the regulatory authority have resulted in associated variations in data collection and subsequent gaps in data, inhibiting its collective usability (Tang 2018). When discussed with participants, several suggested data was collected foremost to meet regulatory requirements, as noted by several participants:

> *'Regulatory periods can drive data collection, or whatever the business sees as cost saving or improving business efficiency.'* [P5].
>
> *'It will be down to the individual water company what data they collect, but as long as they can report to Ofwat the total number of failures for assets and the length of the assets that is kind of the limit that is imposed.'* [P1].
>
> *'Some managers will say we will no longer collect that data since it's no longer reportable, without understanding the value of it for the business in going forward, which lost several years of information.'* [P6].

Future data management methods should prioritise continuous and consistent data collection across regulatory periods, although one participant noted '*Getting new information relies heavily on budget.'* [P8], which presents another problem.

Nonetheless, ongoing data collection will provide prolonged periods of data collection, which is essential. Some respondents noted that many operatives were not made aware of the importance of data – in which case, there is no incentive to keep collecting it. Communication is the key to ensuring operatives understand the use of the data. Continuous data collection is also achievable by adopting ubiquitous smart sensor technologies, such as multi-sensor platforms or fibre optics (e.g., Distributed Acoustic Sensors (DAS)). Such approaches will result in complete data capture without gaps. Such data characteristics often require data standardisation and processing to be performed more efficiently, so expertise with big data technologies should be considered (Sun & Scanlon 2019).

3.1.1. An incomplete range of correlated variables

Predicting pipe failures requires several variables correlated with the various modes and mechanisms of failure (for further detail see Barton *et al.* (2019, 2020)). Collecting variables is a complex task, and pipe failure models are ordinarily developed based on limited data because variables are not routinely collected, are unnecessary for regulatory reporting requirements or immediate business needs (see section 2.1), or are considered too costly to acquire due to budget restrictions (Ramirez *et al.* 2020). As one participant noted '*In a situation where water companies are squeezed for budget, things that look like a 'nice to have' get squeezed out.*' [P1]. Where data is available, it is often framed for a particular use elsewhere in the business, meaning additional time spent on data cleaning and transformation. The involuntary omission of important variables results in variable bias, preventing the estimator from converging correctly, causing an inaccurate representation of pipe failures (Tang *et al.* 2019; Konstantinou & Stoianov 2020). If pipe failure models are to aid decision makers effectively, water companies must prioritise data collection, ensuring enough data related to pipe failures is readily available and collected in useful formats.

3.1.2. Data silos and data sharing

Data silos still exist widely in water companies due to evolving operations, creating many specialist areas and multiple management layers, leading to barriers inhibiting data sharing. Open data policies and incentives have been shown to lead to profound benefits, primarily using data lakes that store and transfer data more efficiently. However, this approach can create additional complexity, cost, and latency, worsened by increased data volumes (Sun & Scanlon 2019). Participants acknowledged that data silos exist as a means of working, one participant mentioned '*Different teams have different datasets, where the rest of the teams don't have access to it or don't even know it exists.*' [P8]. However, several suggested that data lakes widely exist, offering broader data sharing opportunities. In this case, participants highlighted difficulties accessing the data, especially when using systems that require specialist support, with one remarking '*We have a specialist team who look after the data…we can request the information at any time, but it's difficult because the team is always really busy.*' [P1]. Where specialist teams exist to help with data extraction and transformation, it is evident that difficulties persist due to a lack of professional informatics staff. In these instances, data were requested, kept, and updated locally since this proved easier, but in doing so subsequently creating multiple versions of the data. Issues were also identified with communicating the data lake and its intended use, with one participant noting '*It's misunderstood and not accessed by everyone who could benefit from it*' [P4]. Further problems regarding the misuse of data were noted as a problem: *'I think that you'd have to make sure that there was a good description of what all the data sets were and what they mean. It can be very easy for someone who doesn't know that data to use it in a way it wasn't intended for and be making the wrong interpretations of it.'* [P4].

Despite data silos, most participants suggested that data sharing was not a particular issue (unless regulatory restrictions or GDPR apply). Furthermore, there was an apparent enthusiasm for innovation and change when discussing data sharing and transferring of data. A summary discussion around data ontologies and frameworks for shareable and reusable knowledge provoked a positive reaction from all participants, who were open to the idea of new ways to improve collaborative working and data sharing. More focus should be placed on establishing a data management protocol, where data can be stored and shared easily. Big data techniques can help with these challenges, especially when holding and processing several different data formats, but the approach should be intuitive and easily accessible (Ponce Romero *et al.* 2017).

3.2. Limitations of pipe failure data

3.2.1. Coarse recording the time of failure

Irrespective of the challenges posed from historical data and absent variables, readily available information reveals problems that limit pipe failure models, and of these, the lag time between the failure occurring and the failure being observed is a significant one. Sometimes pipe failures can persist underground unnoticed for some time before being discovered. Kleiner &

Rajani (2012) highlighted the lag between actual failures and the time they are discovered, noting the effect of this on short term predictions due to the momentary time-related conditions being obscured. Participants expressed difficulty with finding pipe failures which occur as small leaks in the network, acknowledging that some can go unnoticed for extended periods:

'Some failures have gone unnoticed for years, but sometimes we don't know how long.' [P7].

'If it were days or weeks people would be very happy, I've known them to go [undetected] for years.' [P6].

Pipe failures can be difficult to detect, especially in certain conditions, for example where sandy soils are present, the water leaking from the pipe is less likely to reach the surface. Likewise, WDNs with a smaller percentage of metallic pipes are more difficult to listen for leaks in, because plastic pipes are not as receptive to the acoustic loggers used to detect failures.

The uncertainty of failure timing is inherent within most pipe failure datasets and requires a concerted effort by the water company to record the time of failure accurately. The authors have worked with a dataset that abates this problem to a certain extent by defining pipe failures as either 'reactive' or 'proactive'. Reactive pipe failures signify water emergence at the ground surface, posing a potential impact on continuous water supply, eliciting a quick response by the repair team (typically responding no later than 72 hrs). By this definition, failure time recorded is likely to be more accurate. However, it is also worth noting that water emerging at the surface may also be delayed by some time or even go completely unnoticed in rural areas. Conversely, proactive failures represent low-level leaks that may not immediately affect the network's operation, and as water remains below the surface, it often goes undetected. Proactive failures are actively sought using detection technicians and can be discovered long after the original failure incident.

The failure discovery lag time is a discernible problem, but summarising data by extended prediction intervals (e.g., annual, or greater) is one way to reduce the impact. However, if short-term predictions are required, then recording the time of failure accurately is necessary, and yet acknowledged as a considerable challenge. Acoustic sensors record real-time data and are used for detecting leakage more accurately. Using these data can help to accurately identify the time of failure (Water & Wastewater Treatment 2017).

3.2.2. A lack of high-resolution data

Typically, data correlated with pipe failures originate from various sources at various scales (Kabir *et al.* 2016). Environmental data, such as weather or soil, are just a few examples of secondary data. These data are estimated by area and, for weather, by time intervals under certain assumptions (Robles-Velasco *et al.* 2020), and therefore, do not capture the nature of the exacting conditions causing a failure event. For example, the Met Office Meteorological Office Rainfall and Evaporation Calculation Systems (MORECS) is apportioned over a 40×40 grid and summarised weekly. This, coupled with the lag times of changing weather conditions to the repercussions in soil movement and poorly located pipe failures (see section 3.3.2), can obfuscate the exact conditions of failure (Kleiner & Rajani 2012).

Another example of high-resolution data are localised soil conditions that are particularly important and often altered during excavation and construction, this localised change in soil is not typically recorded. PVC pipes depend on lateral support from undisturbed soil (soil unmoved during the excavation, laying and backfilling of the pipe), on either side of the pipe to carry the backfill material load. Otherwise, the pipe can deform and flatten into an oval shape, reducing its resilience (Saadeldin *et al.* 2015). Since limited information is available regarding localised soil conditions, the number of potentially flattened pipes cannot be quantified. Kleiner & Rajani (2012) noted a general lack of such geographically related data, just as with the experience of some participants:

'If you've got a national soil map on a large scale, the named ground you're going to get in an urban environment, particularly one where it's congested with other utilities, and you've got imported backfill mixed in with your topsoil, it can be hard to estimate the soil conditions around the pipe. This affects corrosion rates and shrink well etc.' [P4].

More localised data is required to improve failure predictions and understand the impact of localised weather and soil conditions. With emerging smart sensor technologies, detailed data collected at a higher spatial (pipe level) and temporal (real-time) resolution for changes in soil conditions or weather conditions is possible and should intuitively improve failure predictions. Other high-resolution data may also include pressure data, water temperature or water consumption for example.

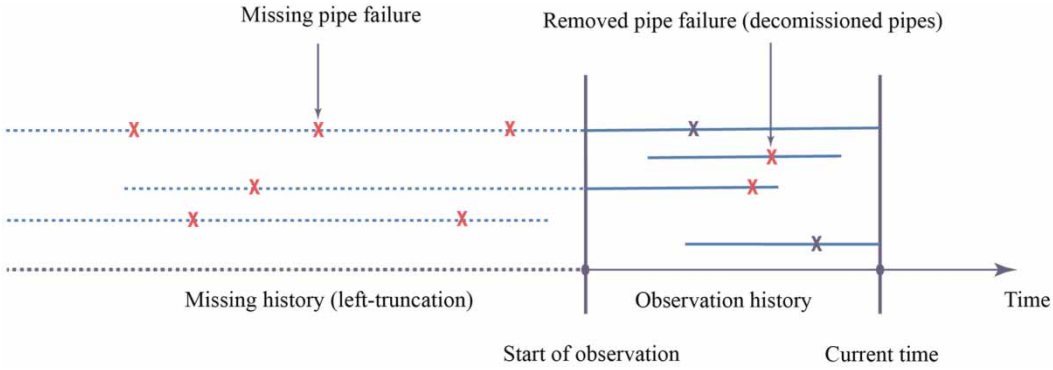

Figure 2 | Depiction of pipe failure data with missing historical data (left truncation), removed pipe failures (decommissioned pipes), and reported failures.

Yet, conditions for pipe installations should be documented in more detail to help understand localised soil conditions. These data, if available, may sometimes help explain variations in pipe failures.

3.2.3. The absence of decommissioned pipes

Pipe failure records are ordinarily limited compared to the life of the pipe, unless those assets are relatively modern (Economou *et al.* 2012). Consequently, pipes installed before failure records began may have failed numerous times but were never recorded, resulting in left-truncated data. Such data is present in most historical datasets as acknowledged by all participants, with one noting '*You've got the failure records which will be incomplete and do not go back to the start of the asset in most cases*' [P1] (Figure 2). It is important to emphasise that left-truncated data is not associated with left-censored data, a term sometimes used in pipe failure literature. Left-censoring only occurs if the number of breaks before pipe failure records is known, but when they happened is unknown, this is fundamentally different (Mailhot *et al.* 2000). Data collected during the failure record period often remove discontinued pipe information to reflect latter network conditions (Figure 2), since as one participant noted, '*Decommissioned data is unnecessary for day-to-day operations.*' [P2].

Subsequently, helpful information about pipe failures is removed from the historical training dataset, information that could potentially improve the power of the model during training by increasing the failure sample size (Scheidegger *et al.* 2015). Nishiyama & Filion (2013) suggest models utilising small data sets and few pipe failures regularly lead to low coefficients of determination and poor accuracy during a review of several studies. Many pipe failure models, such as deterministic models, do not account for these data characteristics (Scheidegger *et al.* 2015). Better accuracy has been reported when datasets contain more pipe failures and span some $\geq$30 years (Liu *et al.* 2012; Winkler *et al.* 2018; Snider & McBean 2019). Decommissioned data should be retained and included in training datasets, yet data collection periods can only increase with time, which provides some optimism for improving future model accuracy.

3.3. Complexities of pre-processing

3.3.1. Handling poor quality data

Pipe failure records are collected by repair teams who are under pressure to repair pipes quickly. Interruptions due to data collection can lead to further costs and delays to customer service, and as one participant suggested, '*It's not always important to point out a poorly recorded pipe age when the primary focus is to stop water from pouring out of the ground.*' [P3]. As a result, data collection can be inconsistent, incomplete or inaccurate, as one participant acknowledged, '*You've got the failure records which will be incomplete and do not go back to the start of the asset in most cases. The records that we do have will be incomplete and have inaccuracies.*' [P4]. Poor data also results from equipment failure such as old, uncalibrated sensors creating signal noise (see Wu & Liu (2017, p. 977) for further descriptions), from historical data and through human data processing, resulting in duplicate or erroneous data. Water companies should encourage complete data capture in the first instance and, where possible, should move to repair equipment failure quickly so as to ensure continuous data collection. This could be achieved by preparing data collection protocols and training staff on the importance of accurate and complete data collection, with one participant noting that '*Data collection is the biggest issue in that it is not done with the full understanding of how valuable the data is.*' [P6].

Despite attempts to improve data quality, some data quality issues will be present, therefore important pre-processing steps are required. Pre-processing approaches to address missing data are important since this can pose a negative impact on the use of data-driven approaches, leading to insufficient historical data to implement the pipe failure method (Wu & Liu 2017). In instances where missing data is present, a list-wise deletion (removing missing values) is often the easiest path to creating a workable dataset (Winkler *et al.* 2018; Snider & McBean 2019). Yet discarding data results in a loss of information, introduction of bias, unreliable parameter estimates, underpowered models and poor convergence (Tang *et al.* 2019). Therefore, it is essential to retain as much data as possible. Alternative approaches that maintain data include making inferences from event narratives or similarities with other events, especially for qualitative information such as material type or failure type. Here, errors in the interpretation can represent another source of uncertainty, for example, *'Typically, you would record either 'cut out' or 'repair', and you could infer a failure type from that, but it would certainly be unreliable…you won't necessarily know why it's failed. In fact, you probably don't know why it's failed.'* [P4]. Alternatively, missing data can be handled through a range of approaches, imputing values based on probability functions or distributions from available datasets (Sattar *et al.* 2016), substituting mean values or standard deviation or through K-nearest neighbour (KNN) techniques, that calculate a value based on similarity to values within close proximity (Levinas *et al.* 2021). Machine learning methods also provide further solution, where some methods such as gradient boosting decision trees can accommodate missing data by imputing averages at the leaf node (Hastie *et al.* 2009). Yet approximations can obfuscate the exact localised conditions associated with pipe failure, introducing further uncertainty, and if a large amount of data is missing (typically >25%), then imputing can lead to incorrect results. Whilst these techniques aim to improve the data, pre-processing techniques are also available to reduce data complexity.

Data engineering processes such as feature extraction techniques use original data to obtain a new set of less redundant variables, optimising the model in the process (García *et al.* 2016). Further developments include feature engineering, a means of deriving new variables from domain knowledge. Recently, deep feature synthesis has focused on automating new variables from existing relational data, providing new variables that perform well regardless of the model used (Rahbaralam *et al.* 2020). Data fusion techniques aggregate data from multiple sources to achieve improved, higher quality and more relevant information (Castanedo 2013). Other popular feature engineering approaches exist such as one-hot-encoding, a procedure that converts original categorical data into new binary values that are easier to interpret and are even necessary for some machine learning models (Aslani *et al.* 2021). Finally, many models have a high sensitivity to variables on different scales, especially so with some machine learning approaches. Standardisation can then be used to scale variables and reduce the effect of outliers that are often observed (Robles Velasco *et al.* 2021).

3.3.2. Processing incorrectly placed pipe failures

Pipe failures are sometimes incorrectly located and appear some distance from any known pipe. Moreover, it is not uncommon for correctly located pipe failures to contain contrasting information with that already present in the existing pipe database (pipe diameter or material type are commonly incorrect).

> *'You rely on the contractor providing the right information back, and normally they can get the information right. But in all the mud and swamp of digging out the ground, they'll not always get the pipe material or diameter size right because they are plastered, so it'll be done afterwards.'* [P6].

> *'Often guys in the field are under pressure to get the repair finished, especially if it is disrupting customers, so collecting data is not always a priority.'* [P1].

> *'They don't have enough people, and when they call incidents that need to get out and sort them, [data collection] it's definitely not a priority.'* [P8].

Since repair technicians train to understand pipe characteristics, contrasting information is presumed to be a result of incorrectly located pipe failures, where data has been updated after the repair event, or where data has been poorly recorded:

> *'Historically we won't have included which particular pipe asset has failed [or] what the address was nearby, so if you got multiple assets in a street, you won't know which one failed.'* [P4].

'Location information from the original call is not always updated, especially if the inspection team are familiar with the area.' [P6].

Poorly located pipe failures are ordinarily handled during the data processing by either listwise deletion or relocation. As noted, listwise deletion can cause excessive data removal and subsequent problems; therefore, relocation is preferred. The authors have previously worked with data where some of the pipe failures required relocating to the nearest point on a pipe with similar characteristics, following a matching protocol firstly within 3 m, then 30 m, and then up to 1 km (see Figure 3). For pipe failures where information contrasts the pipe database, this method provides a valuable solution yet presumes that both sets of data are correct. Conceivably, this may not be the case, and if not, the error persists and can thus mostly pass undetected.

Understandably, pipe failures are difficult to accurately predict when the location of several pipe failures remains in question. Water companies must ensure data collection in the field is accurate by checking details against those records within the existing pipe database. Thus, by highlighting any contrasting information during excavation, there is a chance to resolve any data issues before the repair has finished, an approach likely to reduce the number of failures that need relocating. Likewise, pipe failures should be correctly located with the pipe network where the failure occurred. One participant noted that creating easy-to-use software such as drop-down boxes with specific responses is not always beneficial: '*We used dropdown boxes to make it easier, but the guys tend to always go for the first option regardless. The same thing was observed by many water companies.*' [P6]. Incentivising and encouraging data capture, highlighting its importance, or auditing data capture may be more appropriate. However, organisational silos where teams work specifically towards their own business needs can be a challenge. It is important to encourage communication and collaborative working between teams so that the wider company goals are achieved.

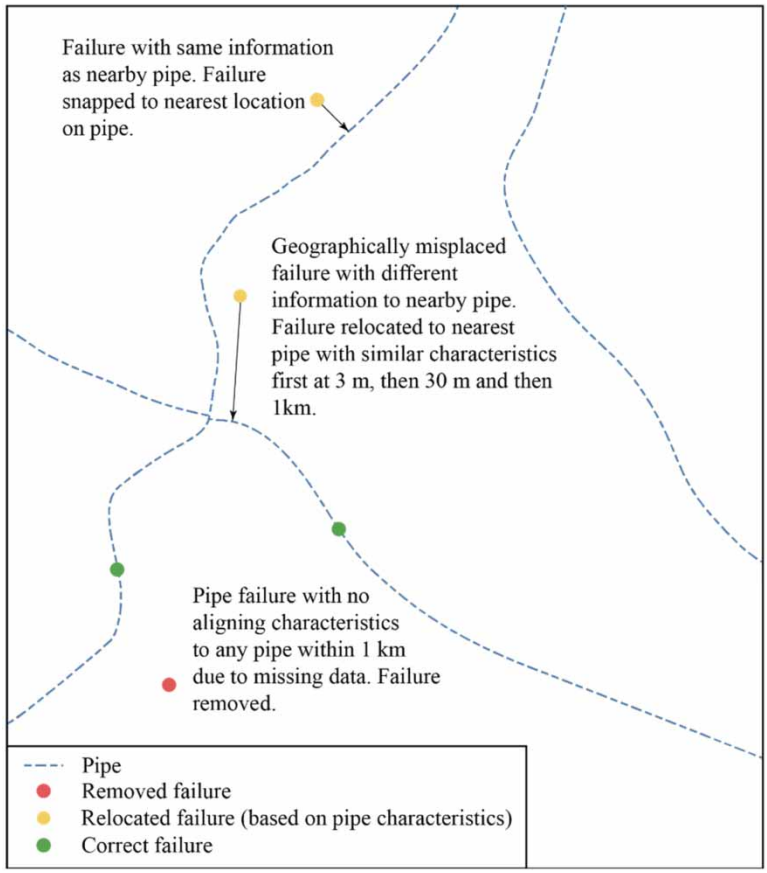

Figure 3 | Illustrative example of a water distribution network with methods for the relocation of pipe failure data. Green – correctly geographically located pipe failures snapped to nearest pipe. Yellow – geographically misplaced pipe failure moved to the nearest pipe within 1 km. Red – failure shares no characteristics of any pipe within 1 km.

3.3.3. Using pipe length and previous failures

Pipe length is a proxy for exposure to failure, where longer lengths denote a higher risk, and the number of failures is directly proportional to the pipe length (Kleiner & Rajani 2012; Robles-Velasco *et al.* 2020). In studies of failure rate, the pipe length, used as a log offset, normalises the failure rate and implies a uniformly distributed number of failures along the length of the pipe (Kleiner & Rajani 2012). Other studies have used pipe length as a covariate to explain failures, believing this is a more appropriate approach (Shamir & Howard 1979; Winkler *et al.* 2018) or even as a unit for determining other variable levels (Robles-Velasco *et al.* 2020). Kleiner & Rajani (2012) discussed the relationship between pipe failures and pipe length, reporting a significant element of randomness. However, a relationship was present when summarising groups of pipes with similar characteristics, accumulating enough pipe failures and length for this randomness to reduce. Pipe length can also have a statistical meaning for other related factors, such as traffic loading, operational stress or bedding conditions, yet this is impossible to quantify (Berardi *et al.* 2008). Pipe length is important, yet it can be problematic since, in many networks, considerably disparate lengths of pipes have been identified, making them difficult to model. In this regard, some authors have re-cut the lengths of pipe by street to get more accurate predictions (Winkler *et al.* 2018). Alternatively, Berardi *et al.* (2008) suggest that all the variables related to pipe length should be represented by length related mean, whilst others propose logarithmically transforming the data (Robles Velasco *et al.* 2021). In models where pipes are grouped, they are typically done so by available soil characteristics and fail to represent similar conditions for traffic loading and operational stress, amongst others.

Failure models use pipe failure history explicitly or implicitly. Implicitly, pipe failure history is used as a surrogate for deterioration and the performance of the pipe, and is represented as either the total number of previous failures or transformed and used as a covariate or stratification criterion. Kleiner & Rajani (2012) used recency (a measure of how recent the last failure occurred) as an expression of pipe failure history, which considers whether historical breaks have occurred more at the beginning or end of the observation period. However, as previously discussed, most pipe failure datasets are prone to left truncation, where failures, before the records started, are absent, resulting in an unreliable estimate of deterioration for a given pipe. Pipe age is also used to measure deterioration, suggesting that older pipes fail more often since they have been subject to deterioration over a more extended period. Again, many participants identified several occasions where older pipes outperformed new pipe installations, with one participant adding, '*Age is different to performance and shouldn't be considered because old pipes often outperform new pipes.*' [P2]. The type of model used should dictate how and if failure history can be used as a covariate, but in the absence of knowledge on the pipe condition, the number of previous failures could be a valuable surrogate for deterioration and resilience to failure.

3.4. Understanding model scalability

3.4.1. Scaling geographic extent

Typically, water companies would prefer failure models to report on individual assets, yet at the pipe level, failures are infrequent. Consequently, the combination of infrequent failures and short failure records mean datasets for pipe level models are imbalanced (more non-failures than failures present with a ratio >9:1 ratio). Three common ways to address data imbalances reported from the literature include the following:

1. Scaling the model to either use longer prediction intervals (greater than annual failures) or grouping pipes by similar characteristics to accumulate more failures by the length of pipe.
2. For classification models, adopting the approach of over and under-sampling, class weighting (Winkler *et al.* 2018) or Synthetic Minority Oversampling Technique (SMOTE) (Chawla *et al.* 2002). For regression models, the use of zero-inflated models has been reported as effective (Economou *et al.* 2012). However, there are limitations to sampling methods and zero-inflation models, and the results are not always favourable, nor do they necessarily improve the model accuracy (Kleiner & Rajani 2012). These techniques are not always necessary for probabilistic modelling.
3. Increasing failure records and ensuring decommissioned data remains in the data set, '*The best way to tackle this problem is always to collect more data.*' (Chicco 2017, p. 7).

Predicting failures at the asset level is seen as desirable since many management decisions are made for individual pipes, yet these models have typically resulted in poor accuracy (Liu *et al.* 2012). This was also observed by participants who have experience of industry working failure models, with one saying, '*It would be lovely to have a model running at the asset level, but I don't think there's any company that has sufficient granularity and accuracy in their data to deliver that.*' [P6]. A desirable and common precursor to pipe failure analysis is partitioning pipes into homogenous groups based on similar characteristics,

or in some recent studies, by using k-means clustering (Kleiner & Rajani 2012). Grouping pipes by similar characteristics whilst summing the number of pipe failures and pipe length increases the number of failures and reduces data imbalance for the model to train on, resulting in greater accuracy (Kleiner & Rajani 2012). Kleiner & Rajani (2012) concluded that inherent uncertainty and a lack of data makes pipe level failure models unreliable, which some participants also suggested, '*For smaller diameter pipes, we found that creating cohorts is necessary to make any appreciable progress.*' [P4]. However, grouping can amass pipe lengths several km long, which does not always aid localised management decisions, but can highlight potential areas for further investigation. Grouping pipes also assumes that all pipes within the group share similar failure rates, which is not often the case. In particular, one pipe may have a far greater number of failures than others within the group, consequently increasing the failure rate for all pipes unnecessarily.

3.4.2. Scaling dynamic time-dependent variables

Early failure models traditionally represented the steady deterioration of pipes. As models evolved, researchers started to include dynamic time-dependent variables, yet mathematically, the statistical analysis is substantially challenging. Dynamic time-dependent variables can be variable or cyclical, and they can often mask the effects of pipe deterioration. The effectiveness of using these variables relies upon the time interval used in forecasting them accurately into the future. To accommodate the fluctuations in seasonal variation, prediction intervals should be less than annual, typically interannual or monthly, yet decreasing the time interval increases the number of observations in a dataset to a point where the data imbalance naturally becomes difficult to compute with any accuracy. Some participants were more sceptical regarding short term failures due to the experience gained within their own companies, with one participate stating '*I certainly wouldn't be placing much faith on these models on a weekly or monthly basis.*' [P4]. The use of weather covariates is typically within the training model only and cannot be used in forecasting unless weather forecast data is included. Forecasting weather data accurately is a complex task, especially with global temperature change, and would add another dimension of uncertainty to the final predictions. However, weather covariates are used since their inclusion is likely to reveal the '*True background ageing rates*' through showing increased failure frequency over time (Kleiner & Rajani 2012, p. 671).

3.5. Future opportunities and challenges

Informed proactive management decisions seek to determine the right place and time for active intervention. Big data collected through smart sensors provides an opportunity to improve the accuracy of predictions by providing continuous high-resolution real-time data capture (such as water pressure data) and a wider variety of variables associated with the mechanisms of failure. This was discussed by participants, one suggesting '*We are on a journey, and at some point, in the near future data will be available to improve models.*' [P5]. With technology driving big data collection, actionable information will start to emerge, and it is feasible that water companies will find alternative means to reduce pipe failures. For example, ontologies using DAS deployed across the network records real-time data that can detect unique failure characteristics using a statistical model to detect specific acoustic signatures, alerting users to a leakage (Levinas *et al.* 2021). This approach can detect a leak within 10 metres of its location after only 30 litres of water has been lost (Water Industry Journal 2018), providing a real-time accuracy that statistical models alone have not yet achieved.

Recent focus has been placed on analytics, where automated decisions are built based on real-time data, reducing the amount of water lost by detecting pipe failures and reacting with an appropriate course of action (e.g., reducing pipe pressure) (Lin *et al.* 2012). One participant noted, '*I've seen a lot of great work from our data science team, and I think we are on the cusp of understanding and realising a world of real-time modelling.*' [P7]. Digital twin approaches extend real-time modelling by providing a virtual representation of an asset, spanning its life cycle, using real-time data, machine learning and simulation to help decision-making. Digital twins are being explored for leakage, energy efficiency, water quality and maintenance planning to name a few, but will be an essential support system for future management (Conejos Fuertes *et al.* 2020). Using smart sensors, operation teams can respond quickly to failures, reducing service interruptions (Water & Wastewater Treatment 2017). It is unlikely that failure models alone will compete with smart networks for operations teams in the near future, especially when short term predictions for short sections of a pipe typically reveal poor accuracy (Kleiner & Rajani 2012). However, failure models can still be used to inform long-term asset management decisions, identifying areas for rehabilitation or replacement. By shifting focus towards proactive management, water utilities can reduce cost and inefficiencies and build future resilience into the network. Yet failure models need to provide a level of accuracy that gives confidence to the predictions: '*It's always challenging, and whatever you predict is never going to be exactly right…you will always have a problem convincing people*

that your model is right because it probably isn't.' [P4]. In this respect, it is perhaps more helpful to predict the probability of failure or contribute risk and vulnerability maps that can rank pipes against each other in relative terms instead of predicting the exact number of failures. Such approaches can even provide helpful insights at an engineering level to take constructive action towards extending life expectancy and building future resilience into the network. Targeting areas of the network this way can also help to facilitate teams actively searching for leaks that are difficult to find, locating optimal areas in which to deploy hydrophones and ground microphones.'

There is a widespread belief that machine learning provides an opportunity to augment more traditional statistical models and improve predictions since they can extract complex relationships (Snider & McBean 2019). However, for limited structured data, often seen in pipe failure datasets, the improvements are often minor after data pre-processing is conducted. Given the additional development and computational time, the most straightforward approaches may prove the most effective. Simple models offer an easier way to explore the predicted outputs and provide interpretability, albeit only in the hands of professionals with domain-specific knowledge (Rudin 2019), with one participant noting, *'As far as possible we try not to have a completely black-box approach that relies heavily on a kind of machine learning where you can't easily work out how you got from your start point to your endpoint.'* [P4]. Yet there is wide interest in machine learning, and many models have shown improved accuracy. However, these more advanced methods cannot significantly improve pipe failure predictions based on insufficient data. One participant noted *'The challenge with this is that the insight is only ever going to be*

Table 2 | Summary of key findings including challenges and priorities for future efforts to improve pipe failure models

Challenge	Findings
Challenges of Managing Data	• Collect continuous data across regulatory periods. • Collect more data correlated with pipe failures in useable formats. • Remove data silos and promote the use of ontological data sharing. • Establish data management protocols, and harmonise the disparate databases, providing data polls that are initiative.
Limitations of Pipe Failure Data	• Summarise prediction intervals (e.g., annual, or greater) to reduce the effects of coarsely recorded failure times. • Use real-time data collection if available to estimate the time of failure (e.g., pressure or flow data). • Determine reactive and proactive failures. • Use high-resolution data using smart sensors. • Collect information on localised conditions during excavation operations to repair and lay new pipes. • Use decommissioned pipe data for training data sets.
Complexities of Pre-processing	• Communicate the importance of data to encourage accurate and complete repair data capture – especially location. • Establish standardised formal data collection protocols. • Use advanced pre-processing methods such as imputing, KNN, feature engineering or data fusion. • Devise a suitable means to place pipe failures accurately, linking site records to pipe records. • Consider dissimilar asset lengths by grouping, re-cutting, length related mean or logarithmic transformation. • The type of model used should dictate how and if failure history can be used as a covariate.
Understanding model scalability	• The desired output should determine the model scalability, although short term predictions at pipe levels perform poorly. • At low spatial and temporal scales, data imbalanced must be carefully considered, perhaps using sampling techniques or zero-inflated models.
Future opportunities and challenges	• Collection of improved quality data must be prioritised. • Advanced data-driven techniques such as machine learning, used in combination with continuous data offers more accurate predictions. • The use of digital twins, smart sensors and real time data offers more accuracy than short-term pipe failure models for operations teams. • Failure models can be used for long term strategy management.

as good as the data.' [P3]. Therefore, it is likely that a combination of improved data and more advanced models can help support the rationale for an improved decision-making process. Large complex datasets with many variables and complex relationships are where machine learning models perform best, but this also depends on the ability of the researcher to create accurate, interpretable models (Rudin 2019).

3.6. Key findings

To highlight important aspects of the discussion, the key findings are summarised and presented in Table 2, separated by each of the five themes.

3.7. Study limitations

There are several difficulties with conducting semi-structured interviews. Ensuring respondents understand the question can be problematic, primarily when broad and open-ended (Oltmann 2016). Addressing this, during the interviews, the initial questions were well understood as only practitioner participants with a depth of professional knowledge were selected, their experience, leading to further interesting discussions. Also, responses can be biased since they are the opinions of the individuals, and whilst the professionals involved provided thorough responses, there may have been useful information that participants were not privy to or held back on. Bias can also be introduced by the interviewers verbal and non-verbal cues (Ritchie *et al.* 2003). Yet, questions were asked in a manner intended to reduce bias by refraining from becoming too involved in the discussion and offering comments (Adams 2015). Face-to-face interviews are particularly prone to reactivity problems, where respondents provide socially acceptable responses. However, given the nature of the questions, which were not sensitive or controversial and the open ended approach, the likelihood of reactivity in responses is limited (Oltmann 2016). This was further limited by turning the camera off during the interview, becoming a method of a phone call as suggested by Oltmann (2016). The information gathered is restricted in geographical scope, and in this respect, it is not generalisable.

4. CONCLUSION

This study has sought to discuss the key challenges and uncertainty in pipe failure models holistically, drawing from the literature and complementing the findings with further understanding from current practices in water companies. The findings reveal that uncertainty stems from poor data quality and quantity, resulting from challenges presented when collecting and managing the data. There needs to be a focus on data collection, in particular communication as to the importance of the data, continuous data collection, sharing relevant data and collecting continuous data. Improving data in the first instance will ensure reliable pipe failure models. Further challenges arise from the infrequent number of failures in pipe failure networks that present mathematical probable to model. The combination of this and poor data quality and quantity makes short term-modelling (weekly, or monthly) for individual assets difficult, and is unlikely to provide an accuracy that guides management decisions.

Modelling pipe failure is an essential component in the proactive management of assets, enabling the identification of areas in the WDN most at risk of failure, guiding replacement strategies, rehabilitation and maintenance schemes or helping target areas of the network for further leakage investigation. With the advent of smart networks and improving analytical techniques, water companies can be more proactive in managing pipe failures, using real-time modelling and insight. The collection of vast data and potentially using modern analytics and machine learning methods offers information to improve pipe failure models, helping water companies make better decisions, especially considering future water demands and a changing climate.

ACKNOWLEDGEMENTS

This work was supported by the UK Natural Environment Research Council [NERC Ref: NE/M009009/1] and Anglian Water plc, who had no role in this study. The work was also supported by the participants. The authors are grateful for their contribution.

COMPETING INTERESTS

The authors declare no conflict of interest. The founding sponsors had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, and in the decision to publish.

AUTHOR CONTRIBUTIONS

Neal Barton: Conceptualisation, Methodology, Investigation, Data Curation, Writing – Original Draft, Writing – Review and Editing, Visualisation and Project administration. Stephen Hallett: Writing – Review and Editing, Supervision, Funding Acquisition. Simon Jude: Review and Editing, Supervision, Funding Acquisition.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Adams, W. C. 2015 Conducting semi-structured interviews. In: *Handbook of Practical Program Evaluation* (Newcomer, K. E., Hatry, H. P. & Wholey, J. S., eds.). John Wiley & Sons, Inc, Hoboken, NJ, USA, pp. 492–505. https://doi.org/10.1002/9781119171386.ch19.

Aslani, B., Mohebbi, S. & Axthelm, H. 2021 Predictive analytics for water main breaks using spatiotemporal data. *Urban Water J.* **00**, 1–16. https://doi.org/10.1080/1573062X.2021.1893363.

Barton, N. A., Farewell, T. S., Hallett, S. H. & Acland, T. F. 2019 Improving pipe failure predictions: factors affecting pipe failure in drinking water networks. *Water Res.* **164**, 114926. https://doi.org/10.1016/j.watres.2019.114926.

Barton, N. A., Farewell, T. S. & Hallett, S. H. 2020 Using generalized additive models to investigate the environmental effects on pipe failure in clean water networks. *npj Clean Water* **3**, 31. https://doi.org/10.1038/s41545-020-0077-3.

Bengtsson, M. 2016 How to plan and perform a qualitative study using content analysis. *NursingPlus Open* **2**, 8–14. https://doi.org/10.1016/j.npls.2016.01.001.

Berardi, L., Giustolisi, O., Kapelan, Z. & Savic, D. A. 2008 Development of pipe deterioration models for water distribution systems using EPR. *J. Hydroinform.* **10**, 113–126. https://doi.org/10.2166/hydro.2008.012.

Castanedo, F. 2013 A review of data fusion techniques. *Sci. World J.* **2013**, 704504. https://doi.org/10.1155/2013/704504.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. 2002 SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357. https://doi.org/10.1613/jair.953.

Chen, T. Y.-J., Beekman, J. A., David Guikema, S. & Shashaani, S. 2019 Statistical modeling in absence of system specific data: exploratory empirical analysis for prediction of water main breaks. *J. Infrastruct. Syst.* **25**, 04019009. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000482.

Chicco, D. 2017 Ten quick tips for machine learning in computational biology. *BioData Min.* **10**, 35. https://doi.org/10.1186/s13040-017-0155-3.

Conejos Fuertes, P., Martínez Alzamora, F., Hervás Carot, M. & Alonso Campos, J. C. 2020 Building and exploiting a Digital Twin for the management of drinking water distribution networks. *Urban Water J.* **17**, 704–713. https://doi.org/10.1080/1573062X.2020.1771382.

Economou, T., Kapelan, Z. & Bailey, T. C. 2012 On the prediction of underground water pipe failures: zero inflation and pipe-specific effects. *J. Hydroinform.* **14**, 872–883. https://doi.org/10.2166/hydro.2012.144.

Fusch, P. & Ness, L. 2015 Are we there yet? data saturation in qualitative research. *Qual. Rep.* **20**, 1408–1416. https://doi.org/10.46743/2160-3715/2015.2281.

Galvin, R. 2015 How many interviews are enough? do qualitative interviews in building energy consumption research produce reliable knowledge? *J. Build. Eng.* **1**, 2–12. https://doi.org/10.1016/j.jobe.2014.12.001.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M. & Herrera, F. 2016 Big data preprocessing: methods and prospects. *Big Data Anal.* **1**, 1–22. https://doi.org/10.1186/s41044-016-0014-0.

Guest, G., Bunce, A. & Johnson, L. 2006 How many interviews are enough? *Field Methods* **18**, 59–82. https://doi.org/10.1177/1525822X05279903.

Hastie, T., Tibshirani, R. & Friedman, J. 2009 *The Elements of Statistical Learning*, 2nd edn. Springer, Dordrecht, the Netherlands. https://link.springer.com/book/10.1007/978-0-387-84858-7.

Kabir, G., Tesfamariam, S., Loeppky, J. & Sadiq, R. 2016 Predicting water main failures: a Bayesian model updating approach. *Knowledge-Based Syst.* **110**, 144–156. https://doi.org/10.1016/j.knosys.2016.07.024.

Kakoudakis, K., Farmani, R. & Butler, D. 2018 Pipeline failure prediction in water distribution networks using weather conditions as explanatory factors. *J. Hydroinform.* **20**, 1191–1200. https://doi.org/10.2166/hydro.2018.152.

Kleiner, Y. & Rajani, B. 2012 Comparison of four models to rank failure likelihood of individual pipes. *J. Hydroinform.* **14**, 659–681. https://doi.org/10.2166/hydro.2011.029.

Konstantinou, C. & Stoianov, I. 2020 A comparative study of statistical and machine learning methods to infer causes of pipe breaks in water supply networks. *Urban Water J.* **17**, 534–548. https://doi.org/10.1080/1573062X.2020.1800758.

Levinas, D., Perelman, G. & Ostfeld, A. 2021 Water leak localization using high-resolution pressure sensors. *Water* **13**, 591. https://doi.org/10.3390/w13050591.

Likhari, R., Webb, S., Bricker, S., Bonsor, H., Frith, N. & Catchpole, S. 2017 *Mapping Underground Assets in the UK 11–12*.

Lin, J., Sedigh, S. & Hurson, A. R. 2012 Ontologies and decision support for failure mitigation in intelligent water distribution networks. In *2012 45th Hawaii International Conference on System Sciences*. IEEE, pp. 1187–1196. https://doi.org/10.1109/HICSS.2012.458.

Liu, Z., Kleiner, Y., Rajani, B., Wang, L. & Condit, W. 2012 *Condition Assessment Technologies for Water Transmission and Distribution Systems*. United States Environmental Protection Agency.

Mailhot, A., Pelletier, G., Noël, J.-F. & Villeneuve, J.-P. 2000 Modeling the evolution of the structural state of water pipe networks with brief recorded pipe break histories: methodology and application. *Water Resour. Res.* **36**, 3053–3062. https://doi.org/10.1029/2000WR900185.

Nishiyama, M. & Filion, Y. 2013 Review of statistical water main break prediction models. *Can. J. Civ. Eng.* **40**, 972–979. https://doi.org/10.1139/cjce-2012-0424.

Ofwat 2006 *The Development of the Water Industry in England and Wales*. Ofwat, Birmingham.

Ofwat 2020 *Innovation in the Water Sector*. Available from: https://www.ofwat.gov.uk/regulated-companies/innovation-in-the-water-sector/.

Oltmann, S. M. 2016 Qualitative interviews: a methodological discussion of the interviewer and respondent contexts. *Forum Qual. Sozialforsch.* **17**. https://doi.org/10.17169/fqs-17.2.2551.

Ponce Romero, J., Hallett, S. & Jude, S. 2017 Leveraging big data tools and technologies: addressing the challenges of the water quality sector. *Sustainability* **9**, 2160. https://doi.org/10.3390/su9122160.

Rahbaralam, M., Modesto, D., Cardús, J., Abdollahi, A. & Cucchietti, F. M. 2020 *Predictive Analytics for Water Asset Management: Machine Learning and Survival Analysis 1–19*.

Ramirez, R., Torres, D., López-Jimenez, P. A. & Cobacho, R. 2020 A front-line and cost-effective model for the assessment of service life of network pipes. *Water (Switzerland)* **12**, 1–23. https://doi.org/10.3390/w12030667.

Renaud, E., De Massiac, J. C., Brémond, B. & Laplaud, C. 2007 SIROCO, a decision support system for rehabilitation adapted for small and medium size water distribution companies. In *LESAM 2007 - 2nd Leading Edge Conference on Strategic Asset Management*. International Water Association, Lisbon, Portugal, pp. 1–15.

Ritchie, J., Lewis, J. & Elam, G. 2003 Designing and selecting samples. In: *Qualitative Research Practice* (Ritchie, J. & Lewis, J., eds). SAGE Publications Inc, Trowbridge, Wiltshire, pp. 77–108.

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

Robles Velasco, A., Muñuzuri, J., Onieva, L. & Rodríguez Palero, M. 2021 Trends and applications of machine learning in water supply networks management. *J. Ind. Eng. Manage.* **14**, 45. https://doi.org/10.3926/jiem.3280.

Rudin, C. 2019 Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215. https://doi.org/10.1038/s42256-019-0048-x.

Saadeldin, R., Hu, Y. & Henni, A. 2015 Numerical analysis of buried pipes under field geo-environmental conditions. *Int. J. Geo-Eng.* **6**. https://doi.org/10.1186/s40703-015-0005-4.

Sattar, A. M. A., Gharabaghi, B. & McBean, E. A. 2016 Prediction of timing of watermain failure using gene expression models. *Water Resour. Manage.* **30**, 1635–1651. https://doi.org/10.1007/s11269-016-1241-x.

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

Shamir, U. & Howard, C. D. D. 1979 An analytic approach to scheduling pipe replacement. *J. Am. Water Works Assoc.* **71**, 248–258. https://doi.org/10.1002/j.1551-8833.1979.tb04345.x.

Snider, B. & McBean, E. A. 2019 Improving urban water security through pipe-break prediction models: machine learning or survival analysis. *J. Environ. Eng.* **146**, 04019129. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001657.

Sun, A. Y. & Scanlon, B. R. 2019 How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. *Environ. Res. Lett.* **14**, 073001. https://doi.org/10.1088/1748-9326/ab1b7d.

Tang, K. 2018 Statistical modelling approaches to improve risk assessments for United Kingdom's water infrastructure. PhD thesis (unpublished), Cranfield University, Cranfield, Bedfordshire, UK.

Tang, K., Parsons, D. J. & Jude, S. 2019 Comparison of automatic and guided learning for Bayesian networks to analyse pipe failures in the water distribution system. *Reliab. Eng. Syst. Saf.* **186**, 24–36. https://doi.org/10.1016/j.ress.2019.02.001.

UKWIR 2020 *The National Failure Database*. Available from: https://ukwir.org/the-national-failure-database (accessed 28 February 2021).

Water Industry Journal 2018 *Reducing Leaks: the Rise of Leak Detection Technology in Water Pipes*. Available from: https://www.waterindustryjournal.co.uk/reducing-leaks-the-rise-of-leak-detection-technology-in-water-pipes (accessed 7 October 2020).

Water & Wastewater Treatment 2017 *Leakage: Acoustic Loggers Strike the Right Note for Affinity Water*. Available from: https://wwtonline.co.uk/features/southern-water-s-leakage-reduction-drive (accessed 6 October 2020).

Winkler, D., Haltmeier, M., Kleidorfer, M., Rauch, W. & Tscheikner-Gratl, F. 2018 Pipe failure modelling for water distribution networks using boosted decision trees. *Struct. Infrastruct. Eng.* **14**, 1402–1411. https://doi.org/10.1080/15732479.2018.1443145.

Wu, Y. & Liu, S. 2017 A review of data-driven approaches for burst detection in water distribution systems. *Urban Water J.* **14**, 972–983. https://doi.org/10.1080/1573062X.2017.1279191.

First received 31 May 2021; accepted in revised form 24 July 2021. Available online 9 August 2021