REVIEW

Open Access

Asset management analytics for urban water mains: a literature review

Atefeh Delnaz[1], Fuzhan Nasiri[1*] and S. Samuel Li[1]

Abstract

This study presents a review of the state-of-the-art literature on water pipe failure predictions, assessment of water losses risk, optimal pipe maintenance plans, and maintenance coordination strategies. In addition, it provides a categorization of water main (WM) failures as well as a taxonomy of WM maintenance strategies. In particular, predictive and prescriptive analytics are highlighted with the investigation of their contributions and drawbacks from methodological and application perspectives. This review aims at providing a review of failure analytics developed recently in water mains domain either for prediction of failure or identification of optimal maintenance strategies conjointly. Future research directions and challenges are elaborated in advancing the understanding about the mechanisms leading to failures. The existing gaps between theory and practice in managing assets across water distribution networks ensuring cost-effectiveness and reliability are discussed. As knowledge about the state of the water mains and related areas is crucial, thus, this review provides an state-of-the-art update from recent studies, and accordingly, presents and discusses avenues for future research.

Keywords Urban water distribution networks, Water mains, Maintenance, Asset management, Data Analytics, Failure analysis, Reliability

Introduction

A water distribution network (WDN) carries freshwater from one or more sources to municipalities for essential human consumption, economic development, and social activities. The WDN is a complex system. It consists of main and booster pumps, water mains typically buried underground, branching pipes, elevated water towers, and interconnected sub-networks for individual neighborhoods. The system or part of it will fail when one or more of its key components, in particular the water mains (WMs), break. There have been numerous cases of water main failures globally, with severe consequences. Examples of consequences include high replacement costs, revenue losses, water damages and contamination, traffic disruptions, and consumer service interruptions (Fares and Zayed 2010; Besner et al. 2011; Malm et al. 2015; Kakoudakis et al. 2017; Liang et al. 2018; Vishwakarma and Sinha 2020; Dawood et al. 2020a).

Proactive interventions for reducing failure risks are necessary and cost-effective, particularly true in the context of aging WMs. Take major urban centers in Canada as example. Statistics Canada (Trudeau 2020) reported fair to very poor conditions for a significant portion of the WDNs. One simple reason is that these WMs have reached or are reaching the end of expected service life. Another reason is that the nature of water mains buried underground makes it complicated and costly to maintain and replace. As a result, there has been an increasing rate of failures over time (Asnaashari et al. 2013; Sattar et al. 2016; Folkman 2018; Snider and McBean 2018, 2021). Over the past decade, studies of the WM problems have made a significant progress, taking advantages of constantly advancing data-driven techniques. The studies aimed

*Correspondence:
Fuzhan Nasiri
fuzhan.nasiri@concordia.ca
[1] Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, QC H3G 1M8, Canada

to detect water main breaks, analyze failure risks, and optimize maintenance.

This study aims to scrutinize the recent literature of water main failure prediction models, failure consequences, failure risk, optimal WM maintenance strategies, and optimal coordinated maintenance strategies. An awareness of the state of the water mains and related areas is crucial. The existing review articles have mostly focused on water main prediction models (St. Clair and Sinha 2012; Dawood et al. 2020b), the effect of availability and quantity of the database on WM failure models (Snider and McBean 2020a; Chen et al. 2022), the effect of the limited, uncertain dataset on WM failure models (Jenkins et al. 2014), and the effect of combined datasets from different utilities on the performance of machine learning models for predicting future breaks (Chen et al. 2022). Other reviews have discussed different approaches to optimizing rehabilitation and maintenance strategies for WMs and integrated infrastructures (Abusamra 2018; Ghobadi et al. 2021; Ramos-Salgado et al. 2022; Shahata et al. 2022; Barton et al. 2022). This review aims to address issues of the failure prediction models developed recently and optimal maintenance strategies conjointly. This review also addresses future research directions in predictive and prescriptive analytics considering cold-region climatic variables.

It is important to consider water main maintenance in coordination with other infrastructures. It is also important to pay close attention to indirect costs and consequences, but the challenge lies in quantifying the indirect consequences in a monetary value (Muhlbauer 2004); to the best of the authors' knowledge, the literature in this domain is scarce (Atef 2010; Yerri et al. 2017). Water main failure consequences are categorized into direct and indirect costs. Loss of production, repair and return to service and pipeline replacement are direct costs. Travel delay, supply outage and substitution, health risk, property damage, customer dissatisfaction and environmental damages are examples of indirect costs (Fares and Zayed 2010; Besner et al. 2011; Malm et al. 2015; Kakoudakis et al. 2017; Vishwakarma and Sinha 2020; Dawood et al. 2020a). Considering indirect costs would make a major difference to WM maintenance plans (Yerri et al. 2017).

Issues exist in either the models or water main datasets themselves or both. They need to be discussed in detail. Thus, the purpose of this review is to provide an update of the knowledge from the recent studies and, more importantly, to explore ways to address the issues in future studies. This would help generate new ideas to improve failure prediction and risk analysis, and thus to reduce the costs in WM asset management planning, rehabilitation and renewal.

In the forthcoming review, the selection of literature is guided by the quest for answers to key questions pertinent to WMs. Some examples of such questions are given below:

- Different methods and techniques have been proposed for locating and managing leaks in WMs (Misiunas 2005; Hamilton and Charalambous 2013; Zyoud and Fuchs-Hanusch 2019, 2020; Karimian et al. 2021). What are the requirements of these approaches? What are the pros and cons?
- Failure models of various types have been used to analyze water main datasets (Economou et al. 2012; American Water Works Association 2019; Snider and McBean 2020a; Snider 2021; Barton et al. 2022). To what extent have the models met the expectation to predict the probability of future failures, time to next failure, and failure rate of pipe, or to predict whether or not a break will happen?
- Failures reportedly could result from a large variety of factors: physical factors (e.g., pipe age, diameter, material, length, and wall thickness), environmental factors (e.g., soil type, climate, freeze/thaw properties, pipe bedding, trench backfill, traffic, and groundwater), and operational factors (e.g., number of pervious failures, water quality, internal water pressure, transient pressure, and leakage) (Stamou et al. 2000; Wang et al. 2009; Arsénio et al. 2015; Lin and Yuan 2019; Karimian et al. 2021). What are the main issues in data acquisition and quality? Are there factors with dominant influence on failures? Will these dominant factors change over time?

Availability of sensory and clouding systems has led to production of vast digital data from WMs. It is very crucial to take advantage of the available data to support short-term and long-term plans of asset management. The use of analytics has shown a rising trend. This review provides a timely update of the existing models for predicting failures and for management planning. Critical pipes are to be identified using the predictive models and then are further included in maintenance plans. The maintenance plans are efficiently optimized to save time, costs and resources.

Predictive analytics

Predictive models of water main failures and pipeline deteriorations (Kleiner and Rajani 2001; Rajani and Kleiner 2001; El-Abbasy et al. 2019; Robles-Velasco et al. 2020; Dawood et al. 2020a) may be classified into two main types: a physical law-based model and a data-driven model (Rajani and Kleiner 2001; Snider and McBean 2020a). The first type of model requires significant amounts of input data to analyze physical behaviors leading to a failure. The analysis involves comparing the resistance capacity of a pipeline to expected loads. The data includes an extensive list of parameters and needs to be collected from the field. Therefore, it is costly and time consuming to use physical law-based models (Rajani and Kleiner 2001). The implementation of the models should be limited to critical pipelines (Wilson et al. 2017). The second type of model uses historical data to discern patterns between historical values of some relevant parameters and breakage rates of pipelines. This type of model is much less expensive to use, compared to a physical law-based model. Thus, it is suitable to implement a data-driven model to all pipelines, as long as historical data exists (El-Abbasy et al. 2019; Snider and McBean 2020a).

The data-driven models may be subdivided into a deterministic model, a probabilistic model, and an artificial intelligence model:

- The deterministic model relies on regression techniques to predict *time to next break* of pipe or break rate and often assumes uniform breaks in water main groups. This assumption rules out uncertainties within a dataset.
- The probabilistic model (e.g., a survival analysis model) uses historical data to predict the probability of water main failure. The model deals with inherent randomness that is expected to be within a dataset of pipe breaks.
- The artificial intelligence model adopts a learning approach to recognizing complicated relationships between input and output data, without calculating the covariate relationships like the deterministic and probabilistic models. Using the artificial intelligence model has the potential to significantly reduce the number of field inspections needed, provide timely warning of break risks and thus avoid a large number of breaks as well as their consequences (Fu et al. 2013; Marzouk and Osama 2017; Kakoudakis et al. 2017; Snider and McBean 2018; Ghobadi et al. 2021).

The classification of data-driven models and their subcategories are shown in Figure 1.

WDNs are a complicated system consisting of interconnected pipes and hydraulic control elements in order to transport potable water to urban populations (Ostfeld 2015). The water infrastructures are aging and deteriorating drastically throughout the major urban centers, which leads to WM failures. They are major problems for municipalities due to high costs for replacement/repair and consequences such as the disruption of services, health issues resulting from contaminated water, and revenue losses (Snider and McBean 2020a). Models should be used to predict

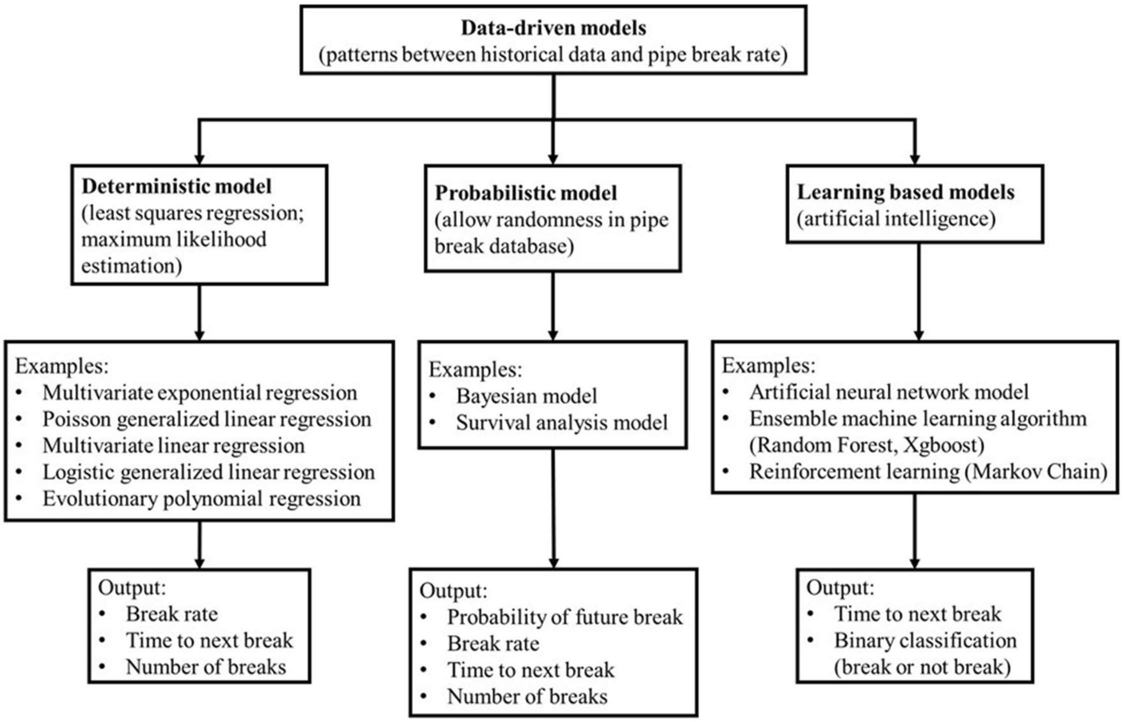
Data-driven models
(patterns between historical data and pipe break rate)
Deterministic model
(least squares regression; maximum likelihood estimation)
Probabilistic model
(allow randomness in pipe break database)
Learning based models
(artificial intelligence)
Examples:
• Multivariate exponential regression
• Poisson generalized linear regression
• Multivariate linear regression
• Logistic generalized linear regression
• Evolutionary polynomial regression
Examples:
• Bayesian model
• Survival analysis model
Examples:
• Artificial neural network model
• Ensemble machine learning algorithm (Random Forest, Xgboost)
• Reinforcement learning (Markov Chain)
Output:
• Break rate
• Time to next break
• Number of breaks
Output:
• Probability of future break
• Break rate
• Time to next break
• Number of breaks
Output:
• Time to next break
• Binary classification (break or not break)

Fig. 1 Classification of water main predictive models

breaks ahead of their occurrences and to plan rehabilitation and replacement. This would promote sustainable infrastructures and save costs.

This review focuses on the data-driven models, considering practicality, data requirement, and liableness. It provides a comprehensive overview of the past 15 years of literature to investigate the relationship and synergy between predictive analytics and prescriptive analytics for water mains. A structured survey of the literature was performed using such keywords as "water main deterioration", "deterioration models", "prediction models", "probabilistic prediction models", "asset management", "water infrastructure", "failure consequence", "water main risk analysis", "integrated municipal infrastructure" and "infrastructure optimization". In total, over 60 articles were reviewed in their entirety.

The strengths and limitations of identified publications were analyzed. Several researchers have used regression, probabilistic and machine learning models for water main failure prediction, as explained above in detail. The application of these models highly depends on the availability of the dataset and desired output. The output of water main failure predictive models can be break rate, number of breaks, probability of future breaks, and time to next break. One of the great advantages of putting these studies together is that it will help municipalities to select and use the most appropriate model, depending on the availability of dataset. Therefore, by using these models, the key question of when and where the break will happen would be answered.

Wang et al. (2009) developed deterioration models, using data of water main breaks to predict annual break rates. These were multiple regression models involving pipe diameter, length, age and material, and identifying the length as having the greatest impact. This is possibly because their model output is breaks per kilometer length per year. They claimed that the models helped analyze break trends. Bruaset and Sægrov (2018) developed a linear regression model that correlated the failure rate of water main to frost heave of the ground due to the air temperature in a cold region. They found the failure rate increasing during the winter months and gray cast iron pipes (usually laid in trenches) being more vulnerable to fail. This implies that the failure rate would decrease under climate warming.

Xu and Sinha (2020) discussed some challenges and gaps in the use of survival analysis models. The use may give failure rate, number of failures, and time to next break (which can be interpreted as either the useful life span of a pipe or estimated remaining useful life). One challenge is the treatment of left truncation. In the literature, left truncation has not been addressed in most survival analyses of water pipeline failure. This would cause a bias in the results. The issue of left truncation needs attention and solutions.

An analysis of pipeline networks is costly and time consuming due to the complexity and large scale of the networks. Therefore, a failure analysis of pipelines is crucial for the efficient management of networks. There is a trend of increasing use of machine learning algorithms to predict the failure rate. Zakikhani et al. (2021) provided a review of failure prediction models, including machine learning models for oil and gas pipelines. Malek Mohammadi et al. (2021) used K-Nearest Neighbor (KNN) to predict the condition of sewer pipes. Also, machine learning algorithms have been recently used in prediction models of infrastructure failure. For example, Marcelino et al. (2021) used general machine learning to predict pavement performance.

Karimian et al. (2021) used an Evolutionary Polynomial Regression model to predict pipeline breaks. They clustered pipelines based on pipe age, diameter, length and material, and showed that pipelines of smaller diameter were more prone to failure. The occurrence of breaks was most sensitive to pipe diameter. For predicting the time to next break of ductile iron pipes, Snider and McBean (2018) made a comparison among a gradient-boosting algorithm model, an Artificial Neural Network (ANN) model and a Random Forest algorithm, suggesting the first one outperformed the other two. This is because gradient-boosting is an ensemble algorithm or a combination of multiple learning algorithms (usually decision trees) that form a stronger predictive model with better performance.

Al-Ali et al. (2019) reported a Logistic Regression (LR) model, aiming to find the most proper parameters for predicting the probability of water main failure, and leading to prioritizing pipes and planning an annual renewal. Dawood et al. (2020a) suggested considering soil type, traffic loads, trenchless method of construction, contractor experience and other influential factors, for improved results of pipe deterioration model. They recommended fuzzy-based assessments to reduce the risks of failure incidents.

In the study of pipeline failures in Colombia's WDN, Giraldo-González and Rodríguez (2020) assessed three regression models and four machine learning models. The regression models were Linear Regression, Poisson Regression (PR) and Evolutionary Polynomial Regression, and the machine learning models were ANN, Bayes, Support Vector Machine and Gradient-Boosted Tree (GBT). The machine learning models used physical factors (age, diameter and length), environmental factors (moisture content, soil contraction, expansion potential, precipitation and land use), and operational factors (valve, hydrant, and previous failure) as predictors. The study used confusion matrices, accuracy and Receiver Operating Characteristic (ROC) curves as an evaluation criterion. The study concluded that PR outperformed the other regression models and GBT outperformed the other machine learning models.

Rahbaralam et al. (2020) employed two machine learning algorithms (LR and extreme gradient boosting) and one survival analysis model (Cox proportional hazard model) to predict Barcelona's water main failures. The algorithms were fed with data after being resampled for feature selection, feature engineering and balancing. The algorithms were evaluated using accuracy, F1 score, recall, precision, Area Under the ROC Curve (AUC) and Matthew's Correlation Coefficient (MCC). The extreme gradient boosting technique was the best.

Water main breaks interrupt services and cause revenue losses (Snider and McBean 2020a). Predictive models of break expected in the future help sustain WDNs and reduce costs. In Snider and McBean (2020b), the gradient boosting decision tree machine learning (xgboost) was compared with Weibull proportional hazard survival analysis, in terms of the effect of censored events on time to next break of cast iron pipes. The xgboost model combines multiple decision trees, which strengthens the performance.

Snider and McBean (2020b) reported that the xgboost model underpredicted time to next break because of the inability to include censored events. Removing censored events from a training dataset is not desirable for long-term planning of asset management. For this reason, they concluded that the model was adequate only for short-term planning of asset management. The Weibull proportional hazard survival analysis could learn from longer censored events in a training dataset; it frequently over-predicts break times (i.e., longer time to break) and therefore is appropriate for use for long-term planning. The analysis can give insights about pipe conditions by using historical data of pipe breaks. Note that unlike inspection data, historical data can easily be found in many water utilities (Xu and Sinha 2020, 2021).

Aslani et al. (2021) used machine learning models to predict water pipeline breaks, with input of spatiotemporal data. Vulnerable locations were identified by conducting a spatial clustering. They converted the results of the clustering analysis to an independent feature called hotspot level for subsequent use in the modeling process. They suggested that the results were useful for municipalities to locate hotspots and mitigate the vulnerability by pipe component renovations.

Robles-Velasco et al. (2020) used LR and Support Vector Classification (SVC) to predict whether a pipe will break or not. LR performed slightly better than SVC. The model output was between 0 and 1. This can be interpreted as the probability of failure, which is highly desirable nowadays. The probability of failure could be used by municipalities to optimally manage their annual rehabilitation plans. Many studies apply machine learning models to pipes that have had breaks (Harvey et al. 2013; Shirzad et al. 2014; Sattar et al. 2016; Kutyłowska 2017; Kerwin and Adey 2018). Robles-Velasco et al. (2020) considered all pipes rather than just those which had experienced breaks. They used three homogenized models with respect to the types of material and then a global model. A correlation analysis identified the covariance between standardized variables. They reported that replacing only 3% of pipelines could prevent around 30% of failures.

Chen et al. (2022) investigated the effect of combined datasets from different utilities on the performance of machine learning models for predicting future breaks. They combined datasets belonging to six utilities in three ways: using the dataset of only one utility, using a stratified sampling of all utilities and using a combined data of all utilities. The results showed that having a large quantity of data does not result in a better prediction model, but instead a sufficient amount of high-quality data such as historical breaks gives a better prediction model.

The examination of the above studies shows that in the case where only a limited amount of input dataset is available and where the purpose is to interpret break trends, regression models could be the best choice. Although survival analysis models are more suitable for long-term management plans, they over-predict break time and cannot handle the complexity that exists in water main dataset. On the other hand, machine learning models are more appropriate for water mains with good amounts of dataset as the models can treat complex relationships between input and output variables. However, these models are suitable only for short-term management planning. Also, it seems that physical parameters which are more accessible in water main dataset and widely used throughout the literature, have more impact on the output of the models. However, the effect of other parameters has yet to be discovered. In the following subsections, the problems existing in either the models or water main datasets itself explained in detail.

Data preparation for modelling

Most machine learning algorithms require data preparations: standardization, encoding, and feature transformation. Standardization rescales all factors. Some machine learning algorithms do not need standardization, however it improves model convergence. Standardization also improves model accuracy (Buntine et al. 2009; Shen et al. 2016). Consider a support vector classifier (SVC). This algorithm works based on maximizing the distance between the separating plane (hyperplane) and the support vectors (data points closer to the hyperplane). When the algorithm calculates the distances, without standardization, features with larger values will dominate features with smaller values. Therefore, standardization is required to reduce the dominancy effect between features and improve the model convergence (Lokman et al. 2019). Consequently, depending on the type of machine learning model selected, standardization might help improve accuracy.

Encoding categorical attributes yields numerical values for use in SVC and LR. The two widely used coding systems: one-hot-encoding, and dummy coding, convert categorical data into binary values (Cohen et al. 2014; Rahbaralam et al. 2020; Aslani et al. 2021). The integer encoding assigns an integer to categorical attributes based on failure rate per unit length (Robles-Velasco et al. 2020). The first two coding systems have the limitation that there is a significant increase in predictors when there are a large number of categories in the categorical attributes. Therefore, depending on the amount of dataset, a suitable coding system should be selected.

In many WMs prediction models, some attributes are difficult to model because of their disparity. Consider pipe length for instance. Disparate lengths of pipes exist in a dataset. Therefore, despite the fact that length is an important predictor, it is problematic. Some authors recut the length by street (Winkler et al. 2018), some used feature transformation and logarithms of length rather than the actual length and improved the accuracy noticeably (Robles-Velasco et al. 2020), and others used mean values for all variables related to length (Berardi et al. 2008). Therefore, the length of water mains needs attention and preparation before being fed into the model for better accuracy.

In some machine learning algorithms, tuning hyperparameters is an important issue which is difficult to properly address (Liu and Zio 2019; Fujiwara et al. 2020). This is because only a few hyperparameters need to be calibrated, which is not enough to capture all the variations in the model. When there are extensive variations in a model but insufficient parameters to capture the variations, an overfitting may occur (Ahmadi et al. 2015). Thus, overfitting should constantly be checked and avoided.

Missing data

In WM dataset, the issue of missing data is common (Osman et al. 2018). Handling missing data in the preprocessing is crucial. Missing data leads to losing some valuable information and causing data insufficiency (Wu and Liu 2017; Winkler et al. 2018). Consequently, removing missing values from a dataset can result in negative effects on data-driven models, unreliable parameter predictions, loss of valuable information, bias, and poor models (Tang et al. 2019). Therefore, it is necessary to keep as much information as possible (Barton et al. 2022).

Alternatively, there are several imputation techniques to handle this issue, e.g., traditional methods such as simple ways of substituting missing data with mean, median and constant values, or more advanced methods such as imputation using machine learning algorithms (for example, substituting missing data with the mean values from KNN in the training dataset) (Levinas et al. 2021; Xu and Sinha 2021). Advanced imputation methods are often better than simple imputation methods (Osman and Bainbridge 2011; Kabir et al. 2019). It is concluded that prior to developing a prediction model, one must have clean data and ensure minimal missing data.

Imbalanced dataset

Imbalanced data, censoring, and left truncation are three important issues associated with predictions of water main failures (Scheidegger et al. 2015; Xu and Sinha 2020). In water supply networks majority of pipelines never suffered from a failure. If the majority of pipelines in a dataset have not experience a break (one class) and a minority of them have experienced at least one break (another class), the dataset is considered as imbalanced, also known as unbalanced (Robles-Velasco et al. 2020) and as censored (Li et al. 2016; Snider and McBean 2020a)). Figure 2 depicts imbalanced data belonging to the City of Kitchener water main break dataset.

Dealing with imbalanced datasets is a challenging topic in data mining, receiving extensive research attention (Zhang and Wang 2013; Ribeiro and Reynoso-Meza 2020). Resampling may be implemented to an imbalanced dataset through random under-sampling, random

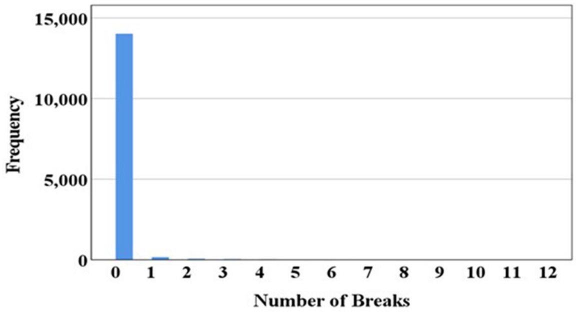

Fig. 2 Frequency of number of breaks in a WDN (imbalanced dataset) (Data source: https://open-kitchenergis.opendata.arcgis.com/datasets/water-main-breaks/explore?location=43.459288%2C-80.434081%2C12.12 and https://open-kitchenergis.opendata.arcgis.com/datasets/water-mains/explore?location=43.434199%2C-80.474206%2C12.58)

over-sampling, and Synthetic Minority Over-sampling Technique (SMOTE) in classification models (He and Garcia 2009). Random under-sampling is a well-known method that removes examples of the majority class. Although this method decreases the computational time, it is at the expense of losing some valuable information (Japkowicz 2000; Seiffert et al. 2009).

Random over-sampling, on the other hand, randomly replicates the existing minority examples to make the dataset balanced. This technique also has its own limitations such as increasing the size of the dataset and causing the model to be overfitted. Thus, it is not applicable in the case of having an extensive dataset (García-Pedrajas et al. 2012). SMOTE randomly generates synthetic minority examples based on nearest neighbors and therefore it is a better way for balancing the dataset; it improves model performance (Fujiwara et al. 2020; Rahbaralam et al. 2020). Nevertheless, depending on the nature of dataset, one of the techniques might work better than the others.

An imbalanced dataset is also an issue in other fields, e.g., medical diagnostic and credit card fraud detection problems (Verhein and Chawla 2007). In such cases, the classification problem becomes very difficult since the main goal in imbalanced datasets is to predict the minority class (Huang et al. 2006). The models in question cannot be properly trained in the training phase and thus cannot correctly predict the minority class (Liu and Zio 2019). A naïve model could predict all data as the majority class and will likely achieve an accuracy of 99%. However, such models are useless in many cases. To evaluate the goodness of a model, accuracy serves a common metric measurement. However, accuracy alone is not considered as a suitable evaluation measurement in the case of an imbalanced data and might cause misinterpretation. Therefore, other metric measurements (e.g., F-measure) are very much demanded (Huang et al. 2006; Harvey and McBean 2014).

The confusion matrix is a good way of evaluation in the case of an imbalanced dataset. Accuracy and Recall are two metrics derived from the matrix. Accuracy gives the percentage of correctly predicted pipes while Recall measures the accuracy of true failures. However, higher Recall is at the expense of misclassification. AUC is another metric measurement that shows the capability of the model to avoid misclassification and can be computed from the ROC curve.

Censored events

Censoring happens when no pipe breaks are observed within a limited period of time, and this is the case in most water utilities datasets. Figure 3 illustrates an example of data censorship in water main break dataset. There are a large number of pipes in service, which have never experienced a break. Censored events can be handled by a traditional survival analysis (e.g., Cox proportional hazard models). On the contrary, many machine learning models are not capable of handling censored events. Although machine learning models are more capable of interpreting complex relationships that exist in a water main dataset, when using machine learning models, censoring is a concern.

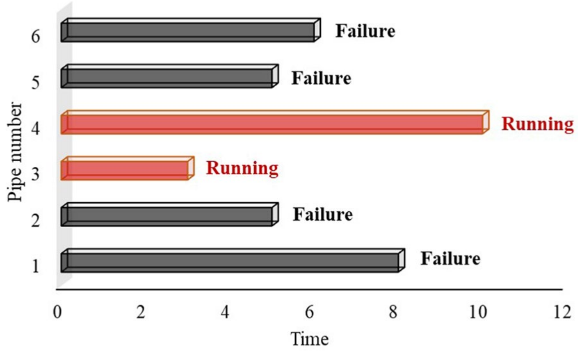

Fig. 3 Censored data (Modified from Snider and McBean (2020a))

Censoring is almost the case for all WM datasets. Although survival models (e.g., Weibull proportional hazard survival analysis) can cope with censored data (Wang et al. 2019; Almheiri et al. 2021) and are good for long-term management planning (Snider and McBean 2020b), they are not suitable for modeling complex relationships between variables. Machine learning algorithms, on the other hand, are very efficient to model complex relationships between variables, but they are good only for short-term management planning (Snider and McBean 2020b). For example, xgboost has been found to surpass other single machine learning models such as Random Forest and ANN (Zhang et al. 2017; Snider and McBean 2018). The problem with xgboost is that it is not programmatically structured to deal with censored data. In fact, it removes censored data at the training stage so it cannot learn from the censored data, therefore it is constantly underpredict time to failure (Snider and McBean 2020b).

Machine learning models are more desirable for use to predict WM failures. To cope with the problem of censoring, a survival machine learning model (a combination of machine learning with a survival statistics) can be used, one exampling being Random Survival Forest, which is relatively new. These models not only incorporate censored data but also utilize data-driven approaches to model complex relationship between input and output variables (Snider and McBean 2021).

Left truncation

Left truncation occurs when the records of pipe failures before collecting data are missing. Like censoring, this is also always the case in a water main dataset and it is acknowledged widely (Barton et al. 2022). The effects of left truncation have been overlooked in many studies (Snider and McBean 2020b) even though this issue causes a systematic bias, especially for survival analysis models (Scheidegger et al. 2015; Xu and Sinha 2019). Instead, they assume the first recorded failure is the first real failure. This assumption will lead to bias and inaccurate predictions (Xu and Sinha 2020; Hawari et al. 2020). The scale and shape of the survival curve can be severely biased due to left truncation, which results in a change in estimates of the mean time to failure (MTTF) (Xu and Sinha 2021).

There are several ways to tackle the left truncation issue. One way is to revise the probability function (Mailhot et al. 2000; Scheidegger et al. 2013). Xu and Sinha (2021) proposed an integration of ANN imputation method with Weibull proportional hazard survival analysis to calibrate the survival curve and reduce MTTF estimation bias caused by left truncation. They showed a drop of bias from 14.3% to 2.1% by applying the method.

Correlation analysis

A correlation between predictors reduces the accuracy and increases computing time for most machine learning algorithms (Hall 1999; Kumar and Chong 2018), except tree-based algorithms which can handle correlations (Eisler and Holmes 2021). The issue of correlations between independent attributes has serious impacts, but it was not addressed (Snider and McBean 2018; Roccetti et al. 2019; Giraldo-González and Rodríguez 2020; Weeraddana et al. 2020; Rahbaralam et al. 2020; Dawood et al. 2020a; Amini and Dziedzic 2021). There are different methods to investigate the correlation, e.g., the t-test, ANOVA, MANOVA, Chi Squared and Pearson's correlation analysis. Depending on the nature of the dataset (i.e., being numerical or categorical), the above-mentioned methods are useful. Pearson's correlation analysis is one of the most widely used methods, but it is useful only for identifying the correlation between numerical variables (Zhang et al. 2014).

Prescriptive analytics

The literature in the domain of water distribution networks can be divided into different categories in many ways. In this review, literature is divided into two main categories: predictive analytics and prescriptive analytics. In the following subsections, literature of prescriptive analytics is explained in more details.

Failure consequences assessment and risk analysis

This paper reviewed the existing literature related to identifying risk, criticality index and failure consequences for WMs. Fares and Zayed (2010) utilized a hierarchy fuzzy expert system to evaluate the risk of WM failure. Their considered 16 risk factors. According to their study, risk factors can be divided to factors that lead to failure (deterioration factors) and factors which result from failure (consequence factors). They demonstrated that the most significant influences on failure risk are pipe age, pipe material, and pipe breakage rate, respectively. Kabir et al. (2015) proposed a Bayesian Belief Network model to prioritize metallic WMs and evaluate the risk of WMs failure. They used structural integrity, hydraulic capacity, water quality and consequence factors in their model, and they claimed that any other factors could also be included in their model. They showed that the model can visualize the most vulnerable, sensitive and the highest risk pipes within a WDN.

Mugume et al. (2015) simulated a simplified synthetic water distribution system in EPANET and a synthetic urban drainage system in the Storm Water Management Model. They investigated the system performance under the condition of pipe failure. They focused on minimizing failure consequences to improve resilience in urban water systems. They also investigated the effect of rehabilitation strategies including pipe replacement on resilience. They showed that if failure scenarios are considered during urban water systems design, the loss of system functionality could be minimized.

Al-Zahrani et al. (2016) identified the vulnerable locations in a WDN using a fuzzy-based decision support system. These vulnerable locations experience more structural failures as well as failures in supplying water at the target quality. Their model was applied to a case where a risk index was developed to show both the probability of failures and their impacts. They showed that the model helped utilities to prioritize pipes within the system based on the overall failure risk.

Vishwakarma and Sinha (2020) used the fuzzy inference method for developing the consequence of failure. They proposed a quantitative risk matrix for risk visualization, that compared to semi-quantitative and qualitative risk matrix, reduce subjectivity in the design process. Their model framework covers different types of the consequence of failure assessment such as economic, environmental and social impacts, as well as operational intelligence and complexity of renewal activities. They improve previously developed techniques of assessing failure consequences by using a quantitative risk matrix. Utilizing risk assessment has multiple advantages for management programs, such as supporting pipes renewal prioritization decisions and moving from reactive maintenance plans to proactive plans.

Phan et al. (2019) used a risk assessment framework in a case study of WM in a WDN. The calculation of the probability of failure used Weibull distribution. They used a fuzzy inference system to aggregate failure consequences because unifying different types of consequences into one outcome is difficult. Consequences consist of impacts on the redundancy/vulnerability of the network, water loss and rehabilitation costs and of impacts on public health. They used the diameter to quantify the volume of water loss and algebraic connectivity to consider the topological consequence. The topological consequence is useful for redundancy reduction. In order to prioritize water mains, a risk map is developed for use by decision-makers.

Balekelayi and Tesfamariam (2021) performed a hydrodynamic assessment for the wastewater system of Calgary using ordered weighted averaging technique to identify the criticality index of the wastewater pipes. A dynamic deterioration model was combined with the proposed criticality index to determine the operational risk of the wastewater pipes. This technique helps municipalities to prioritize the inspection and replacement of sewer pipes. They showed that the technique can successfully identify the criticality index of wastewater pipes when hydrodynamic data are not available. Using the information, hydraulic models can be regularly updated and thus wastewater pipe inspection plans can be prioritized. The results of the study can also be used for water mains.

Risk is a multiplication of the probability of failure (POF) and consequences of failure (COF). The probability of WM failures can be derived from the prediction models explained earlier in predictive analytics section. However, in order to achieve a good maintenance plan, POF is not the only factor that matters, and COF is another important factor. This is because some pipes might have the least POF but the highest COF in the network, which might be overlooked in the prioritization plan. Therefore, assessing COF is also of relevance. The failure consequences can be economic, environmental and social impacts. The indirect costs of failure should also be taken into consideration. Thus, the determination of failure consequences and hence the risk are difficult, because of uncertainties and many factors involved. Often, fuzzy techniques are used to deal with uncertainties and to quantify failure consequences and risk factors.

Maintenance planning, scheduling and prioritization

In this section, papers in regard to water loss minimization and asset management plans have been collected. The deterioration of assets is inevitable due to aging. Thus, an efficient asset management becomes crucial for assets to continue delivering an adequate level of service. There are efficient asset management plans in various infrastructures sectors such as road networks, urban railways and metro systems, buildings, wastewater and drainage systems (Mohammadi et al. 2018, 2019, 2020; Dziedzic et al. 2021). However, there is less progress in case of water systems asset management.

Kleiner et al. (2010) developed a non-homogeneous Poisson model for the analysis and forecast of breakage patterns in individual water mains, considering both static and dynamic factors. Their case study was for a water utility in Eastern Ontario. Different costs associated with each pipe were considered, including the costs of pipe replacement and repair, the costs of water loss due to failure, and cost-saving due to roadwork coordination. They used the results of pipe break predictions for the water main renewal schedule plan, utilizing a multi-objective genetic algorithm.

Malm et al. (2015) developed a Cost-Benefit Analysis for leakage reduction. They compared the costs and benefits for each alternative over time. They also considered uncertainty analysis. The results show that considering uncertainty analysis improved the results of the Cost-Benefit Analysis. They considered four different alternatives to reduce leakage in their case study of Gothenburg. It was found that reactively repairing, despite a high leakage rate, is more cost effective, compared to proactively pipe replacement.

Zyoud and Fuchs-Hanusch (2019, 2020) applied different techniques to a real water supply system in Palestine. They compared the traditional Multi Criteria Decision Making approach and the Analytic Hierarchy Process (AHP) method for a water loss management problem. Although AHP is easy to implement and has strong potential in structuring and decomposing complex decision problems, it cannot handle uncertainties. Therefore, Fuzzy AHP has been used to deal with uncertainty and incomplete information.

Barton et al. (2022) revealed that the quantity and quality of data have an important impact on the accuracy of WM failure models, and poor data results in low accuracy of models. They suggested that there should be increased focus on data collection since poor quality data makes it hard for utilities to manage WMs rehabilitation plan. They show that long term management plans for water mains remain a challenging issue and require further attention.

Ghobadi et al. (2021) proposed a pipe replacement scheduling method based on a life cycle cost assessment. In order to obtain an optimal replacement plan, a multi-objective nondominated sorting genetic algorithm (NSGA-II) is used. The proposed replacement plan avoids investment peaks and smooth the investment time series based on life cycle cost. Unlike many other studies, they considered that limitations exist in the annual budget in their model. They show that by using online monitoring and recording failure data, the accuracy of the pipe failure rate is improved, and the annual replacement plans can be updated. The scheduling plan becomes near optimal.

Decision making software tools and methodologies would help municipalities to perform their water infrastructure maintenance plans more efficiently. These plans usually consider a set of predefined alternatives. However, more practical replacement plans which affect several pipes simultaneously rather than just replacing individual pipes haven't been considered in these methodologies. Ramos-Salgado et al. (2022) scheduled a sustainable water supply replacement plan with a five-step infrastructure asset management framework. (1) As the first step, a replacement priority index for every network asset has been obtained. (2) Despite the previous maintenance strategies (considering individual pipelines), they used street sections as the operational replacement unit to reduce the social consequences related to each intervention. (3) They considered the replacement plan of two adjacent pipes at the same time even with having different priority of replacement for the sake of operational and convenience criteria, since it is more acceptable by utilities and more aligned with their policies. Also, a fair budget allocation performed in their study based on social and geographic criteria to ensure a decent investment distribution between districts and towns. (4) After specifying the replacement priority of the network assets, a short-term, mid-term and long-term replacement plan is required. In this regard, a set of indicators for performance evaluation of the network is needed which specify the investment level and certain courses of action. A combination of four indicators is used name infrastructure value index (ratio between the value of the infrastructure at the current state and its replacement cost), average network age, average risk index, and the average probability of failure. These indicators are easy to calculate and interpret. They also present various information on the performance of the network. (5) Lastly a mathematical technique is used to calculate the required budget more efficiently.

Maintenance coordination and prioritization

There have been tremendous efforts on maintenance plans as an individual asset management plan. A coordinated asset management plan is much needed to better manage existing infrastructure assets, but the coordination has been neglected by many municipalities. This section gives particular attention to the coordination of interrelated infrastructures, optimum replacement time of them (e.g., roads, water and sewers) and prioritization of their budget allocation. Integrated rehabilitation actions among the co-located infrastructure assets are necessary when developing a renewal plan. This could decrease or avoid unnecessary rework, rehabilitation costs, service disruptions and risks (Halfawy 2008; Abusamra 2018).

Marzouk and Osama (2015) proposed a decision support tool to determine the optimal time of maintenance and replacement of mixed infrastructures simultaneously (i.e., pavement, water pipes, sewer pipes, gas pipes, and electrical cables). This approach could prevent costs associated with the surface layer of pavements to be destroyed multiple times (for example once for sewer pipes replacement and once for water pipes replacement). The useful life of different infrastructures was first identified by simulation, and then depending on the replacement time and costs, a decision was made on the optimal maintenance and replacement time. With regard to uncertainties of models, a fuzzy approach was applied. The key goal of their study was the minimization of the total costs of infrastructure replacement.

Marzouk and Osama (2017) presented a method for the coordinated maintenance of road, water distribution and wastewater distribution networks. First, a deterioration model is developed using a hierarchical fuzzy expert system technique to assess the condition of each infrastructure asset. Then, a risk model is developed using a fuzzy Monte Carlo simulation to calculate POF and AHP to calculate COF. Lastly, a multi-objective optimization using genetic algorithm (GA) is developed, with four objective functions: (1) minimizing the overall risk, (2) maximizing level of service (LOS), (3) maximizing the overall conditions of the assets, and (4) minimizing life cycle cost (LCC). The optimization model considers seven scenarios of actions for: (1) road segment only; (2) water only; (3) sewer only; (4) road and water; (5) road and sewer; (6) sewer and water; (7) road, sewer and water. The optimization constraints were set to meet the minimum requirements of the condition, performance and risk for all infrastructures within the annual budget. The results showed an average integrated risk index of 5.45 over a planning horizon of 20 years. Over 86% of the projects were recommended under integrated scenarios as follows: road, water and sewer at 38%; road and sewer at 24%; road and water at 24%. These maximize cost saving.

Abusamra (2018) pointed out numerous attempts to improve infrastructure maintenance and intervention plans within a limited budget. However, most of them were successful only in developing a plan for short-term planning and a single asset. The author proposed optimization models to help decision makers to identify a coordinated maintenance plan for the co-located infrastructure assets (i.e., roads, water, and sewer). Two multi-objective models were discussed: (1) evolutionary GAs optimization, which used a set of meta-heuristic rules to find a near-optimum solution; (2) linear programming optimization to find an exact solution. The objective function was to maximize an overall improvement and to maximize the network health index. The results showed an overall enhancement (time, cost, efficiency, risk, etc.) of 29% over a planning horizon of 25 years, achieved from coordinating the interventions. Compared to the conventional approach, coordination reduced disruptions and interventions by 67%.

Amador-Jimenez and Mohammadi (2020) considered different budgeting scenarios such as worst-first, silos, and trade-off optimization, to assess the pros and cons of proposed scenarios. They aimed to investigate the prioritization of budget allocation and management plans for different infrastructure assets (i.e., pavements, sanitary sewers, storm sewers and water mains), based on the proposed scenarios, and to select the superior management plan among all. They show that a trade-off optimization analysis improves results, giving the highest priority to water mains and lower priority to pavements and storm pipes in terms of investment management planning.

Very recently, Shahata et al. (2022) proposed a multi-stage integer programming that is capable of optimizing the most suitable, cost-effective renewal action (if any) for road, sewer and water infrastructure assets. The objective function was to maximize risk reduction in a cost-effective manner. Their decision-making approach used risk assessment and a performance rating model. The model also used rehabilitation alternatives, giving priority to integrated renewal actions. They showed that the approach is capable of reducing risk costs by using integrated actions (e.g., road, water and sewer by 36%; road and sewer by 23%; road and water by 25%). They also showed that their integrated model can enhance budget-saving, compared to the conventional silos approach (renewal plan of only each infrastructure). In order to improve the model's practicality, the consequence of each intervention alternative such as the impact on travel delay, noise pollution costs, lost business revenue, etc. should also be considered.

Discussions

This review of water asset management analytics has revealed: a) a need to explore the influence of environmental factors on WM failures; b) a need to consider both direct and indirect costs in optimal mitigation analysis and replacement prioritization. The environmental factors indirectly contribute to failures. The contribution is particularly significant for WMs in cold regions. Failure models should be coupled with costs (direct and indirect) as a constraint in optimal scheduling plans. The coupling renders failure predictions meaningful as the ultimate goals are to update asset management plans and prioritize rehabilitation or replacement.

Further research efforts are needed to reveal new insights about contributing mechanisms of WM failures, to create novel ideas for reliable predictions of failures, and to invent ways for putting theoretical predictions into practical use in managing and maintaining WMs in a cost-effective manner. The mechanisms are more complex in cold regions. More details about potential avenues for future research are discussed below under each category of analytics:

Directions of future research in predictive analytics

In spite of extensive studies of WM failures over the past decades, significant knowledge gaps exist in predictive analytics. Environmental factors (e.g., weather conditions, climate factors and so on) are reportedly less influential than physical factors (e.g., pipe diameter, pipe length and so on). However, the influence of the environmental factors such as climatic variations and freezing in cold regions has received little attention (Kleiner and Rajani 2002; Farmani et al. 2017; Demissie et al. 2017; Almheiri et al. 2020). In the cold regions, pipes are more susceptible to break due to temperature fluctuations. Frozen water inside a pipe expands. Even if the pipe does not break, it can significantly degrade. Freezing temperature fluctuations result in extra stresses on pipes. Moisture on the ground can cause frosts at freezing temperatures and lead to ground movement and hence stresses on the pipes. Cast-iron pipes are more prone to failures at freezing temperatures because of the erosion of soils around them. If they are not lined with protection materials, they begin to corrode from inside and ultimately break. In future research, it would be meaningful to create homogenous groups of pipes based on the environmental factors such as soil type, freezing index, temperature, precipitation and frost depth in order to investigate their influence on WM failures.

The past studies have overlooked issues related to the apparent age of pipes based on their conditions. An application of rehabilitation techniques such as lining and cathodic retrofit to existing pipes causes a change in the conditions of the pipes and thus redefines their ages. Therefore, the influence of applied rehabilitation techniques and the resulting change need to be investigated.

One important step before developing any prediction model is data preprocessing and preparation. The missing gap of data needs to be handled properly. If the available amount of the missing data is not meaningful, they can simply be excluded. Otherwise, an existing missing gap should be filled by predictions using advanced imputation methods. Correlated attributes must be removed as they will decrease modelling efficiency significantly. One needs to pay adequate attention to imbalanced dataset and general data preparation before applying any prediction method because these steps impact modelling reliability significantly. Resampling dataset is a good way to cope with imbalanced dataset.

Unresolved issues of censoring, and left truncation are common with WM datasets. One way to deal with the issues is to use survival machine learning models (a combination of machine learning with a survival statistics). The models can handle both censoring and a complex relationship between input and output variables. Table 1 presents a summary of predictive analytics and the applied techniques published in the past 13 years.

Directions of future research in prescriptive analytics

Previous studies using prescriptive analytics have been limited to consideration of economic costs as the maintenance objective to optimize. The social and environmental costs (indirect costs) associated with a failure are commonly ignored in WM maintenance planning and rehabilitation scheduling. Future research should aim to maximize the system reliability and at the same time minimize the risk index and failure consequences (costs). Beside economic costs, the social and environmental costs can have significant influence on maintenance planning and scheduling, and they should be considered.

The multiplication of probability and COF determines the risk factor; through this link, a risk map can be developed and utilized to develop a maintenance prioritization plan. The coupling of a WM prediction model, probability of WM failure and COF allows us to develop a precise, practical maintenance prioritization plan. This goal can be achieved using an optimization model, together with decision-making methods. The goal should be set in a way to reduce leakage, which in turn decreases expenses (direct and indirect) and increases the expectancy life of assets. However, long-term management plans remain challenging and further attention is needed.

The literature in related to optimization of maintenance/replacement time for infrastructures networks with coordination and prioritization of maintenance activities are rare. In reality, a WM infrastructure is often maintained in association with other infrastructure such as pavement, and thus the asset management impact of one infrastructure on the other is inevitable. The mutual impact remains essentially an under explored area. A prescriptive analysis of interdependent infrastructures would be helpful to prioritize budget allocations and to identify the optimal replacement/maintenance time in a realist setting.

In conclusion, the need to adopt a coordinated maintenance plan for integrated infrastructure assets is extensively acknowledged in industry and academia. When the assets reach an unacceptable LOS, which need some actions and interventions, the optimum decision on how to repair all overlapping assets using the pre-existing and limited budget and without overspending, remains challenging. Therefore, priorities should be set in a way to answer these questions: Which asset is more critical and needs immediate action? What are the actions/interventions (repair, rehabilitation, replace or do nothing)? When is the best time the work should be done? One important requirement for all coordinated maintenance plans is the ability to support long-term planning. In this regard, the life cycles of different infrastructure assets

Table 1 A summary of predictive analytics applications for water pipes failure

Cluster of research	Techniques	Applications	Input parameters
Deterministic model	– Multiple Regression models – Linear Regression – Poisson Regression – Multivariate adaptive regression splines	– Water mains annual break rates prediction – Finding correlation between water main failure rates and input parameters	Physical Factors: DIA, LEN, AGE, MAT
Probabilistic model	– Cox proportional hazard model – Weibull proportional hazard survival analysis	– Water main failures prediction – Time to next break prediction	Physical Factors: DIA, LEN, AGE, MAT, THK Operational Factors: WP, VEL, TRF, RT, WPH Environmental Factors: SR, SPH, MC, FI, SZN
Artificial intelligence model	– ANN – Random Forest – Xgboost – LR – SVC – Evolutionary Polynomial Regression – Boosted regression tree	– Predict the failure rate of pipeline networks – Binary classification which shows whether or not the pipe break	Physical Factors: DIA, LEN, AGE, MAT, YEAR, NC, NT Operational Factors: WP, TRF, NB, BD, BY Environmental Factors: MC, ST, PP, LU, LO

should be considered in these models. Table 2 presents a summary of prescriptive analytics and the applied techniques published in the past 12 years.

Proposition

In the light of the above literature review and after considering the knowledge gaps related to the existing analytics methods and issues associate with water main datasets, establishing an integrated approach for smart water mains asset management is advocated (Figure 4) incorporating the synergy between failure models (predictive analytics) and maintenance strategies (prescriptive analytics). Most WM datasets mainly consist of physical factors of pipes such as age, diameter, length and material. Usually, they do not include operational factors such as annual average daily traffic (AADT), number of breaks, water pressure, and environmental factors such as freezing and thawing index, temperature, precipitation, frost depth, and rain deficit. Therefore, in order to investigate the effects of the environmental factors, this study suggests merging them with WM datasets. After cleansing and careful data pre-processing, dimensionality reduction is useful to reduce dataset dimensions and computing time. To aggregate the efforts for similar regions with similar characteristics, one may perform clustering which is relatively new in this domain. The next step is to select features that contribute the most in failure prediction models. Concretely, with sufficient data, a prediction model can be developed as the ultimate step in predictive analytics.

In regard to prescriptive analytics, depending on the types of prediction model in previous stage, either the

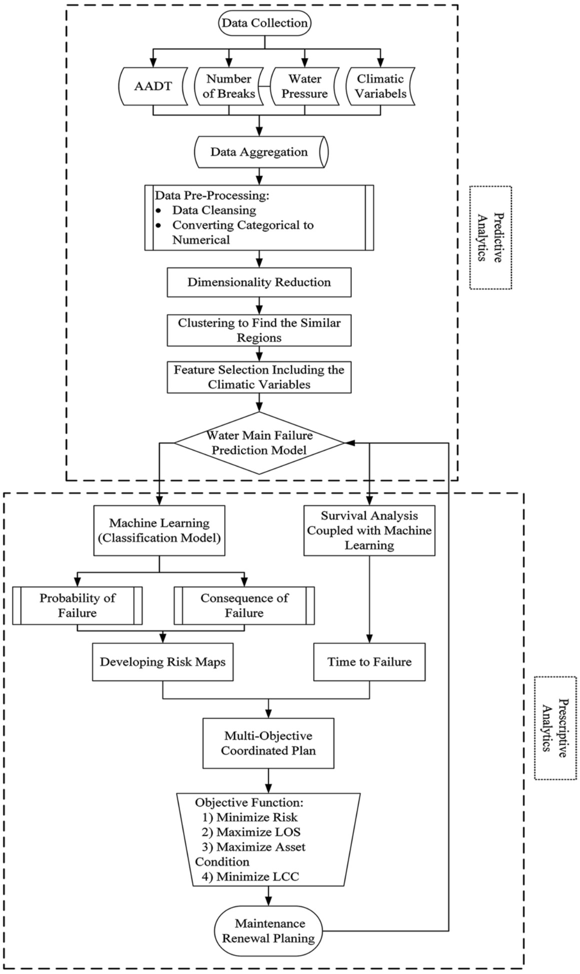
Data Collection
AADT
Number of Breaks
Water Pressure
Climatic Variabels
Data Aggregation
Data Pre-Processing:
• Data Cleansing
• Converting Categorical to Numerical
Dimensionality Reduction
Clustering to Find the Similar Regions
Feature Selection Including the Climatic Variables
Water Main Failure Prediction Model
Predictive Analytics
Machine Learning (Classification Model)
Survival Analysis Coupled with Machine Learning
Probability of Failure
Consequence of Failure
Developing Risk Maps
Time to Failure
Multi-Objective Coordinated Plan
Prescriptive Analytics
Objective Function:
1) Minimize Risk
2) Maximize LOS
3) Maximize Asset Condition
4) Minimize LCC
Maintenance Renewal Planing

Fig. 4 Proposed integrated predictive and prescriptive analytics for smart water main asset management

Table 2 A summary of prescriptive analytics applications for water pipes failure

Cluster of research	Techniques	Applications
Failure consequences assessment and risk analysis	– Hierarchy fuzzy expert system – Bayesian Belief Network, Fuzzy-based decision support system – Ordered weighted averaging technique – EPANET – Fuzzy inference method – Quantitative risk matrix	– Determination of the risk and criticality index – Failure consequences assessment, resilience investigation – Risk visualization
Maintenance planning and scheduling	– Multi Criteria Decision Making approach – Analytic Hierarchy Process – Cost–Benefit Analysis – NSGA-II – Non-homogeneous Poisson model – Multi-objective genetic algorithm	– Water loss minimization and management – Pipe replacement plan considering life cycle cost assessment – Schedule long-term asset management framework
Maintenance coordination and prioritization	– Decision support tool based on a fuzzy approach – Worst-first, silos, and trade-off optimization – GA optimization – Evolutionary GA (based on heuristic rules) – MOSEK linear programming – Integer programming optimization	– Optimum replacement time of integrated infrastructure – Prioritization of budget allocation for different infrastructure assets – Minimizing risk and LCC – Maximizing LOS, asset's overall condition, overall improvements and network health index

time to failure or POF/COF could be mapped across the water mains network. To obtain COF, the indirect costs of failure, such as proximity to environmental/external factors (e.g., rail tracks and transmission gas mains) and the impact on the LOS and the costumer class (e.g., hospital, emergency services, residential) could be considered. By minimizing the risk of failure, the infrastructure maintenance plan can be prioritized accordingly. Using the prediction models resulted from the predictive analytics step, a multi-objective maintenance plan could be developed in coordination with other infrastructures such as roads and sewer pipes. The other optimization objectives could be maximizing LOS, maximizing asset condition, and minimizing LCC. Lastly, the rehabilitation/replacement plan will be scheduled. It is expected that after implementing the rehabilitation/replacement plan in WMs, the condition of the networks will change. So, the predictive models should be updated based on the new information after each prediction-prescription-implementation cycle.

Conclusions

Water pipe failures have increased drastically due to a slow rate of replacements and thus aging of WMs. This issue is difficult to resolve because such networks are complex and are typically buried underground. In many municipalities, most parts of the networks have reached the end of their service life, expediting even more failures in near future. Given that failures incur revenue losses and cause interruptions to service and economic activities, it becomes increasingly urgent to find better solutions. Various financial, societal, and technical constraints make it infeasible to think of replacing aging WMs, which typically serves many residential, commercial, industrial and institutional consumers, and which consists of a vast network of interconnected pipelines, pumps, valves, regulators and tanks. Thus, predicting near-future failures is of economic, social and environmental relevance.

This review provided a comprehensive overview of the methods proposed for predicting and minimizing the failures and their consequences. It has provided new insights into the knowledge gaps identified in the existing studies related to the applications of predictive and prescriptive analytics in water systems asset management. In spite of extensive research efforts over the past decades, the treatment of imbalanced data, censoring and left truncation remains as key research gaps. The other gaps correspond to how to increase sustainability, reliability and resilience of WM systems through the use of predictive models and efficient rehabilitation planning.

Considering the literature and the identified gaps, this study proposed a failure analytics framework for WMs and discussed a number of avenues for future research. It is worthy to highlight that the quality of dataset could have a significant impact on the performance of prediction models. To achieve this goal, this review recommends that municipalities use advanced inspection technologies which result in establishing more accurate prediction models, leading in turn to more precise data-drive prescription analytics that improve the reliability of WMs and create cost efficiency gains.

Abbreviations

AGE	Pipe age (year)
AUC	Area Under the ROC Curve
ANN	Artificial Neural Network
BD	Break density (Breaks/Km^2)
AHP	Analytic Hierarchy Process
BY	Break year (year)
COF	Consequences Of Failure
DIA	Pipe diameter (mm)
FI	Freezing Index (degree days)
GA	Genetic Algorithm
GBT	Gradient-Boosted Tree
KNN	K-Nearest Neighbor
LEN	Pipe length (m)
LOS	Level Of Service
LCC	Life Cycle Cost
LO	Pipe location
LR	Logistic Regression
LU	Land use
MAT	Material
MCC	Matthew's Correlation Coefficient
MC	Moisture content (%)
NB	Number of previous breaks
NC	Number of connections
NT	Network type
PR	Poisson Regression
PP	Precipitation (mm)
POF	Probability Of Failure
ROC	Receiver Operating Characteristic
RT	Road type
SPH	Soil PH
SR	Soil resistivity
ST	Soil type
SVC	Support Vector Classification
SMOTE	Synthetic Minority Over-sampling Technique
SZN	Season
THK	Pipe thickness
TRF	Traffic
WDN	Water Distribution Network
WM	Water Main
WP	Water pressure (KPa)
WPH	Water PH
YEAR	Installation year (year)

Acknowledgements

The authors are very much thankful to four anonymous reviewers whose comments and suggestions were very helpful in improving the quality of this manuscript.

Author contributions

The work has been done under the supervision of FN and SI. The first draft of the manuscript was written by AD. Reviews, edits, and further analysis were performed by FN and SI. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The corresponding author, Fuzhan Nasiri, is an unpaid member of the editorial board of Environmental Systems Research. The authors declare that they have no other competing interests.

Received: 12 January 2023 Accepted: 23 February 2023
Published online: 18 April 2023

References

Abusamra S (2018) Coordination and multi-objective optimization framework for managing municipal infrastructure under performance-based contracts, PhD Thesis, Concordia University, Montreal, QC, Canada

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

Al-Ali AM, Laurent J, Dulot JP (2019) Developing deterioration prediction model for the potable water pipes renewal plan–case of Jubail Industrial City, KSA. Desalin Water Treat 176:324–332. https://doi.org/10.5004/dwt.2020.25539

Al-Zahrani M, Abo-Monasar A, Sadiq R (2016) Risk-based prioritization of water main failure using fuzzy synthetic evaluation technique. J Water Supply Res Technol 65:145–161. https://doi.org/10.2166/AQUA.2015.051

Almheiri Z, Meguid M, Zayed T (2021) Failure modeling of water distribution pipelines using meta-learning algorithms. Water Res 205:117680. https://doi.org/10.1016/J.WATRES.2021.117680

Almheiri Z, Meguid M, Zayed T (2020) An approach to predict the failure of water mains under climatic variations. Int J Geosynth Gr Eng 64(6):1–16. https://doi.org/10.1007/S40891-020-00237-8

Amador-Jimenez L, Mohammadi A (2020) Decision making methods to prioritise asset-management plans for municipal infrastructure. Infrastruct Asset Manag 8:11–24. https://doi.org/10.1680/JINAM.19.00064

American Water Works Association (2019) Condition assessment of water mains. American Water Works Association, Denver

Amini M, Dziedzic R (2021) Comparison of machine learning classifiers for predicting water main failure. Proc Can Sociery Civ Eng Annu Conf. https://www.mdpi.com/2073-4441/12/4/1153/pdf

Arsénio AM, Dheenathayalan P, Hanssen R et al (2015) Pipe failure predictions in drinking water systems using satellite observations. Struct Infrastruct Eng 11:1102–1111. https://doi.org/10.1080/15732479.2014.938660

Aslani B, Mohebbi S, Axthelm H (2021) Predictive analytics for water main breaks using spatiotemporal data. Urban Water J. https://doi.org/10.1080/1573062X.2021.1893363

Asnaashari A, McBean EA, Gharabaghi B, Tutt D (2013) Forecasting watermain failure using artificial neural network modelling. Can Water Resour J 38:24–33. https://doi.org/10.1080/07011784.2013.774153

Atef A (2010) Optimal condition assessment policies for water and sewer infrastructure (Doctoral dissertation, Nile University)

Balekelayi N, Tesfamariam S (2021) Operational risk-based decision making for wastewater pipe management. J Infrastruct Syst 27:04020042. https://doi.org/10.1061/(asce)is.1943-555x.0000586

Barton NA, Hallett SH, Jude SR (2022) The challenges of predicting pipe failures in clean water networks: a view from current practice. Water Supply 22:527–541. https://doi.org/10.2166/WS.2021.255

Berardi L, Giustolisi O, Kapelan Z, Savic DA (2008) Development of pipe deterioration models for water distribution systems using EPR. J Hydroinformatics 10:113–126. https://doi.org/10.2166/HYDRO.2008.012

Besner MC, Prévost M, Regli S (2011) Assessing the public health risk of microbial intrusion events in distribution systems: Conceptual model, available data, and challenges. Water Res 45:961–979. https://doi.org/10.1016/J.WATRES.2010.10.035

Bruaset S, Sægrov S (2018) An analysis of the potential impact of climate change on the structural reliability of drinking water pipes in cold climate regions. Water 10:411. https://doi.org/10.3390/W10040411

Buntine W, Grobelnik M, Mladenić D, Shawe-Taylor J (2009) Machine learning and knowledge discovery in databases. Lect Notes Comput Sci. https://doi.org/10.1007/978-3-642-04174-7

Chen TY-J, Vladeanu G, Yazdekhasti S, Daly CM (2022) Performance evaluation of pipe break machine learning models using datasets from multiple utilities. J Infrastruct Syst 28:05022002. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000683

Cohen P, West SG, Aiken LS (2014) Applied multiple regression/correlation analysis for the behavioral sciences. Psychol Press. https://doi.org/10.4324/9781410606266

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

Dawood T, Elwakil E, Novoa HM, Delgado JFG (2020b) Water pipe failure prediction and risk models: state-of-the-art review. Can J Civ Eng 47:1117–1127. https://doi.org/10.1139/CJCE-2019-0481

Demissie G, Asce SM, Tesfamariam S et al (2017) Prediction of Pipe failure by considering time-dependent factors: dynamic Bayesian Belief Network Model. ASCE-ASME J Risk Uncertain Eng Syst Part A Civ Eng 3:04017017. https://doi.org/10.1061/AJRUA6.0000920

Dziedzic R, Amador L, An C et al (2021) A framework for asset management planning in sustainable and resilient cities. IEEE Int Symp Technol Soc. https://doi.org/10.1109/ISTAS52410.2021.9629158

Economou T, Kapelan Z, Bailey TC (2012) On the prediction of underground water pipe failures: zero inflation and pipe-specific effects. J Hydroinformatics 14:872–883. https://doi.org/10.2166/HYDRO.2012.144

Eisler C, Holmes M (2021) Applying automated machine learning to improve budget estimates for a naval fleet maintenance facility. In ICPRAM. https://doi.org/10.5220/0010302205860593

El-Abbasy MS, Zayed T, El CH et al (2019) Simulation-based deterioration patterns of water pipelines. Struct Infrastruct Eng 15:965–982. https://doi.org/10.1080/15732479.2019.1599965

Fares H, Zayed T (2010) Hierarchical fuzzy expert system for risk of failure of water mains. J Pipeline Syst Eng Pract 1:53–62. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000037

Farmani R, Kakoudakis K, Behzadian K, Butler D (2017) Pipe failure prediction in water distribution systems considering static and dynamic factors. Procedia Eng 186:117–126. https://doi.org/10.1016/J.PROENG.2017.03.217

Folkman S (2018) Water main break rates in the USA and Canada: a comprehensive study. Mech Aerosp Eng Fac Publ

Fu G, Kapelan Z, Kasprzyk JR, Reed P (2013) Optimal design of water distribution systems using many-objective visual analytics. J Water Resour Plan Manag 139:624–633. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000311

Fujiwara K, Huang Y, Hori K et al (2020) Over- and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis. Front Public Heal 8:178. https://doi.org/10.3389/FPUBH.2020.00178

García-Pedrajas N, Peŕez-Rodríguez J, De Haro-Garciá A (2012) OligoIS: Scalable instance selection for class-imbalanced data sets. IEEE Trans Cybern 43:332–346. https://doi.org/10.1109/TSMCB.2012.2206381

Ghobadi F, Jeong G, Kang D (2021) Water pipe replacement scheduling based on life cycle cost assessment and optimization algorithm. Water (switzerland) 13:605. https://doi.org/10.3390/w13050605

Giraldo-González MM, Rodríguez JP (2020) Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks. Water (Switzerland). https://doi.org/10.3390/W12041153

Halfawy MR (2008) Integration of Municipal Infrastructure Asset Management Processes: Challenges and Solutions. J Comput Civ Eng 22:216–229. https://doi.org/10.1061/(ASCE)0887-3801(2008)22:3(216)

Hall M (1999) Correlation-based feature selection for machine learning. Doctoral dissertation, University of Waikato, Dept. of Computer Science

Hamilton S, Charalambous B (2013) Leak detection: technology and implementation. IWA Publishing, London, UK

Harvey R, McBean EA (2014) Comparing the utility of decision trees and support vector machines when planning inspections of linear sewer infrastructure. J Hydroinformatics 16:1265–1279. https://doi.org/10.2166/HYDRO.2014.007

Harvey R, McBean EA, Gharabaghi B (2013) Predicting the timing of water main failure using artificial neural networks. J Water Resour Plan Manag 140:425–434. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000354

Hawari A, Alkadour F, Elmasry M, Zayed T (2020) A state of the art review on condition assessment models developed for sewer pipelines. Eng Appl Artif Intell 93:103721. https://doi.org/10.1016/J.ENGAPPAI.2020.103721

He H, Garcia EA (2009) IEEE Transactions on knowledge and data engineering. IEEE Trans Knowl Data Eng 21:1263–1284. https://doi.org/10.1109/TKDE.2008.239

Huang D-S, Li K, Irwin GW et al (2006) Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. Intell Control Autom. https://doi.org/10.1007/978-3-540-37256-1_89

Japkowicz N (2000) The class imbalance problem: Significance and strategies. In: In Proc. of the Int'l Conf. on Artificial Intelligence. pp 111–117

Jenkins L, Gokhale S, Asce F, Mcdonald M (2014) Comparison of pipeline failure prediction models for water distribution networks with uncertain and limited data. J Pipeline Syst Eng Pract 6:04014012. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000181

Kabir G, Tesfamariam S, Francisque A, Sadiq R (2015) Evaluating risk of water mains failure using a Bayesian belief network model. Eur J Oper Res 240:220–234. https://doi.org/10.1016/J.EJOR.2014.06.033

Kabir G, Tesfamariam S, Hemsing J, Sadiq R (2019) Handling incomplete and missing data in water network database using imputation methods. Sustain Resilient Infrastruct 5:365–377. https://doi.org/10.1080/23789689.2019.1600960

Kakoudakis K, Behzadian K, Farmani R, Butler D (2017) Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K-means clustering. Urban Water J 14:737–742. https://doi.org/10.1080/1573062X.2016.1253755

Karimian F, Kaddoura K, Zayed T et al (2021) Prediction of breaks in municipal drinking water linear assets. J Pipeline Syst Eng Pract 12:04020060. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000511

Kerwin S, Adey BT (2018) Performance comparison for pipe failure prediction using artificial neural networks. In: In Proc., 6th Int. Symp. on Life-Cycle Civil Engineering. pp 1337–1342

Kleiner Y, Nafi A, Rajani B (2010) Planning renewal of water mains while considering deterioration, economies of scale and adjacent infrastructure. Water Sci Technol Water Supply 10:897–906. https://doi.org/10.2166/ws.2010.571

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

Kleiner Y, Rajani B (2002) Forecasting variations and trends in water-main breaks. J Infrastruct Syst 8:122–131. https://doi.org/10.1061/(ASCE)1076-0342(2002)8:4(122)

Kumar S, Chong I (2018) Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. Int J Environ Res Public Health 15:2907. https://doi.org/10.3390/ijerph15122907

Kutyłowska M (2017) Prediction of failure frequency of water-pipe network in the selected city. Period Polytech Civ Eng 61:548–553. https://doi.org/10.3311/PPCI.9997

Levinas D, Perelman G (2021) Ostfeld A (2021) Water leak localization using high-resolution pressure sensors. Water 13:591. https://doi.org/10.3390/W13050591

Li J, Zhou S, Han Y (2016) Advances in battery manufacturing, service, and management systems. Wiley, Hoboken

Liang B, Li Z, Wang Y, Chen F (2018) Long-term RNN: predicting hazard function for proactive maintenance of water mains. Int Conf Inf Knowl Manag Proc 1687–1690. https://doi.org/10.1145/3269206.3269321

Lin P, Yuan XX (2019) A two-time-scale point process model of water main breaks for infrastructure asset management. Water Res 150:296–309. https://doi.org/10.1016/J.WATRES.2018.11.066

Liu J, Zio E (2019) Integration of feature vector selection and support vector machine for classification of imbalanced data. Appl Soft Comput 75:702–711. https://doi.org/10.1016/J.ASOC.2018.11.045

Lokman S-F, Othman AT, Bakar MHA, Musa S (2019) The impact of different feature scaling methods on intrusion detection for in-vehicle controller area network (CAN). Int Conf Adv Cyber Secur. https://doi.org/10.1007/978-981-15-2693-0_14

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

Malek Mohammadi M, Najafi M, Serajiantehrani R et al (2021) Using machine learning to predict condition of sewer pipes. Pipelines 2021:185–195. https://doi.org/10.1061/9780784483602.022

Malm A, Moberg F, Rosén L (2015) Pettersson TJR (2015) Cost-benefit analysis and uncertainty analysis of water loss reduction measures: case study of the gothenburg drinking water distribution system. Water Resour Manag 2915(29):5451–5468. https://doi.org/10.1007/s11269-015-1128-2

Marcelino P, de Lurdes AM, Fortunato E, Gomes MC (2021) Machine learning approach for pavement performance prediction. Int J Pavement Eng 22:341–354. https://doi.org/10.1080/10298436.2019.1609673

Marzouk M, Osama A (2017) Fuzzy-based methodology for integrated infrastructure asset management. Int J Comput Intell Syst 10:745–759. https://doi.org/10.2991/ijcis.2017.10.1.50

Marzouk M, Osama A (2015) Fuzzy approach for optimum replacement time of mixed infrastructures. Civ Eng Environ Syst 32:269–280. https://doi.org/10.1080/10286608.2014.1002715

Misiunas D (2005) Failure monitoring and asset condition assessment in water supply systems. PhD Thesis. Department of Electrical Engineering and Automation, Lund University, Sweden

Mohammadi A, Amador-Jimenez L, Nasiri F (2020) Reliable, effective, and sustainable Urban railways: a model for optimal planning and asset management. J Constr Eng Manag 146:04020057. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001839

Mohammadi A, Amador-Jimenez L, Nasiri F (2019) Review of asset management for metro systems: challenges and opportunities. Transp Rev 39:309–326. https://doi.org/10.1080/01441647.2018.1470119

Mohammadi A, Igwe C, Amador L, Nasiri F (2018) Novel asset management framework for road maintenance. Canadian Society of Civil Engineers (CSCE) Annual Conference, June 13-18, Fredericton, NB, Canada

Mugume SN, Diao K, Astaraie-Imani M et al (2015) Enhancing resilience in urban water systems for future cities. Water Sci Technol Water Supply 15:1343–1352. https://doi.org/10.2166/ws.2015.098

Muhlbauer WK (2004) Pipeline risk management manual : ideas, techniques, and resources. Elsevier, Amsterdam

Osman H, Bainbridge K (2011) Comparison of statistical deterioration models for water distribution networks. J Perform Constr Facil 25:259–266. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000157

Osman MS, Abu-Mahfouz AM, Page PR (2018) A Survey on data imputation techniques: water distribution system as a use case. IEEE Access 6:63279–63291. https://doi.org/10.1109/ACCESS.2018.2877269

Ostfeld A (2015) Water distribution networks. Stud. Comput Intell 565:101–124. https://doi.org/10.1007/978-3-662-44160-2_4

Phan HC, Dhar AS, Hu G, Sadiq R (2019) Managing water main breaks in distribution networks––A risk-based decision making. Reliab Eng Syst Saf 191:106581. https://doi.org/10.1016/J.RESS.2019.106581

Rahbaralam M, Modesto D, Cardús J, Abdollahi A, Cucchietti FM (2020) Predictive analytics for water asset management: machine learning and survival analysis. arXiv preprint arXiv:2007.03744

Rajani B, Kleiner Y (2001) Comprehensive review of structural deterioration of water mains: physically based models. Urban Water 3:151–164. https://doi.org/10.1016/S1462-0758(01)00032-2

Ramos-Salgado C, Muñuzuri J, Aparicio-Ruiz P, Onieva L (2022) A comprehensive framework to efficiently plan short and long-term investments in water supply and sewer networks. Reliab Eng Syst Saf. 219:108248. https://doi.org/10.1016/J.RESS.2021.108248

Ribeiro VHA, Reynoso-Meza G (2020) Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets. Expert Syst Appl 147:113232. https://doi.org/10.1016/J.ESWA.2020.113232

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

Roccetti M, Delnevo G, Casini L, Cappiello G (2019) Is bigger always better? a controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures. J Big Data 6:1–23. https://doi.org/10.1186/S40537-019-0235-Y

Sattar AM, Gharabaghi B, McBean EA (2016) Prediction of timing of watermain failure using gene expression models. Water Resour Manag 30:1635–1651. https://doi.org/10.1007/S11269-016-1241-X

Scheidegger A, Leitão JP, 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

Scheidegger A, Scholten L, Maurer M, Reichert P (2013) Extension of pipe failure models to consider the absence of data from replaced pipes. Water Res 47:3696–3705. https://doi.org/10.1016/J.WATRES.2013.04.017

Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A (2009) RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Trans Syst Man, Cybern A Syst Humans 40:185–197. https://doi.org/10.1109/TSMCA.2009.2029559

Shahata K, El-Zahab S, Zayed T, Alfalah G (2022) Rehabilitation of municipal infrastructure using risk-based performance. Autom Constr 140:104335. https://doi.org/10.1016/J.AUTCON.2022.104335

Shen X, Gong X, Cai Y et al (2016) Normalization and integration of large-scale metabolomics data using support vector regression. Metabolomics 12:1–12. https://doi.org/10.1007/S11306-016-1026-5

Shirzad A, Tabesh M, Farmani R (2014) A comparison between performance of support vector regression and artificial neural network in prediction of pipe burst rate in water distribution networks. KSCE J Civ Eng 18:941–948. https://doi.org/10.1007/S12205-014-0537-8

Snider B (2021) Preparing for the replacement era: understanding north america's aging water distribution systems

Snider B, McBean EA (2018) Improving time to failure predictions for water distribution systems using extreme gradient boosting algorithm. In: proceedings of the 1st international water system distribution analysis (WDSA)/ computing and control for the water industry conference, July 23-25, Kingston, ON, Canada

Snider B, McBean EA (2021) Combining machine learning and survival statistics to predict remaining service life of watermains. J Infrastruct Syst 27:04021019. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000629

Snider B, McBean EA (2020a) Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions. Urban Water J 17:163–176. https://doi.org/10.1080/1573062X.2020.1748664

Snider B, McBean EA (2020b) 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

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

Stamou AI, Latsa M, Assimacopoulos D (2000) Design of two-storey final settling tanks using mathematical models. J Hydroinformatics 2:235–245. https://doi.org/10.2166/HYDRO.2000.0021

Tang K, Parsons DJ, 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

Trudeau MP (2020) SWM and urban water: Smart management for an absurd system? Water Int 45:678–692. https://doi.org/10.1080/02508060.2020.1783063

Verhein F, Chawla S (2007) Using significant, positively associated and relatively class correlated rules for associative classification of imbalanced datasets. Seventh IEEE Int Conf Data Min. https://doi.org/10.1109/ICDM.2007.63

Vishwakarma A, Sinha SK (2020) Development of a consequence of failure model and risk matrix for water pipelines infrastructure systems. Pipelines 2020. VA Am Soc Civ Eng. https://doi.org/10.1061/9780784483213019

Wang P, Li Y, Reddy CK (2019) Machine learning for survival analysis: a survey. ACM Comput Surv 51:1–36. https://doi.org/10.1145/3214306

Wang Y, Zayed T, Moselhi O (2009) Prediction models for annual break rates of water mains. J Perform Constr Facil 23:47–54. https://doi.org/10.1061/(asce)0887-3828(2009)23:1(47)

Weeraddana D, Liang B, Li Z, et al (2020) Utilizing machine learning to prevent water main breaks by understanding pipeline failure drivers. arXiv Prepr arXiv200603385. https://doi.org/10.48550/arXiv.2006.03385

Wilson D, Filion Y, Moore I (2017) State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. Taylor Fr 14:173–184. https://doi.org/10.1080/1573062X.2015.1080848

Winkler D, Haltmeier M, Kleidorfer M et al (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

Xu H, Sinha SK (2020) Applying survival analysis to pipeline data: gaps and challenges Pipelines 2020. VA Am Soc Civ Eng. https://doi.org/10.1061/9780784483213017

Xu H, Sinha SK (2021) Modeling pipe break data using survival analysis with machine learning imputation methods. J Perform Constr Facil 35:04021071. https://doi.org/10.1061/(ASCE)CF.1943-5509.0001649

Xu H, Sinha SK (2019) A framework for statistical analysis of water pipeline field performance data. Pipelines 2019 Multidiscip Top Util Eng Surv. VA Am Soc Civ Eng DOI. https://doi.org/10.1061/9780784482506019

Yerri SR, Piratla KR, Matthews JC et al (2017) Empirical analysis of large diameter water main break consequences. Resour Conserv Recycl 123:242–248. https://doi.org/10.1016/J.RESCONREC.2016.03.015

Zakikhani K, Nasiri F, Zayed T (2021) A failure prediction model for corrosion in gas transmission pipelines: Proc Inst Mech Eng. Part O J Risk Reliab 235:374–390. https://doi.org/10.1177/1748006X20976802

Zhang C, Liu C, Zhang X, Almpanidis G (2017) An up-to-date comparison of state-of-the-art classification algorithms. Expert Syst Appl 82:128–150. https://doi.org/10.1016/J.ESWA.2017.04.003

Zhang Y, Wang D (2013) A cost-sensitive ensemble method for class-imbalanced datasets. Abstr Appl Anal. https://doi.org/10.1155/2013/196256

Zhang Z, McDonnell KT, Zadok E, Mueller K (2014) Visual correlation analysis of numerical and categorical data on the correlation map. IEEE Trans vis Comput Graph 21:289–303. https://doi.org/10.1109/TVCG.2014.2350494

Zyoud SH, Fuchs-Hanusch D (2019) Comparison of several decision-making techniques: a case of water losses management in developing countries. Int J Inf Technol Decis Mak 18:1551–1578. https://doi.org/10.1142/S0219622019500275

Zyoud SH, Fuchs-Hanusch D (2020) An integrated decision-making framework to appraise water losses in municipal water systems. Int J Inf Technol Decis Mak 19:1293–1326. https://doi.org/10.1142/S0219622020500297

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.