Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Integrated approach for pipe failure prediction and condition scoring in water infrastructure systems

Talha M. Rifaai [a], Ahmed A. Abokifa [b], Lina Sela [a,*]

[a] Department of Civil, Architectural and Environmental Engineering, University of Texas, Austin, United States
[b] Department of Civil, Materials and Environmental Engineering, University of Illinois, Chicago, United States

ARTICLE INFO

Keywords:
Pipe failure prediction
Logistic regression
Mean time to failure
Risk scoring

ABSTRACT

Pipe failures in water distribution infrastructure have significant economic, environmental, and public health impacts. To alleviate these impacts, pipe deterioration modeling has been increasingly implemented to characterize and predict pipe failure patterns with the aim of prioritizing repair and replacement decisions. Logistic regression has been recognized in recent literature as a strong candidate for failure prediction modeling. However, previous studies have often been limited to demonstrating the application of logistic regression for estimating failure probabilities. This study builds on previous efforts by proposing an approach for implementing logistic regression into a holistic framework for asset management decision-making. This framework incorporates logistic regression modeling, with a flexible time-interval, into a practical condition scoring methodology that accounts for the attitude of water utilities towards risk. The developed framework is demonstrated and tested on a 20-year pipe failure dataset of a large metropolitan US city. The logistic regression model displayed high accuracy in estimating the probability of failure within different time intervals, and the scoring method showed a reasonable ability to predict the criticality of repair decisions for pipes based on their condition.

1. Introduction

In the 2017 Infrastructure Report Card, the American Society of Civil Engineers rated the drinking water infrastructure in the United States as poor and at risk. Water infrastructure in the US incurs 240,000 main breaks wasting more than two trillion gallons of treated drinking water every year [1]. In addition to water losses, structural deterioration can undermine water quality and hydraulic integrity of the water distribution systems [2]. Today, deteriorated water distribution systems lead to a daunting $1 trillion worth of overdue repairs for the next two to three decades. Beyond the prohibitive economic costs, pipe deterioration also has health and safety, social, and environmental impacts including leached chemicals, undermined fire-fighting flows, interrupted services, a decreased quality of life, and wasted resources [3,4]. In the face of such deterioration, the replacement rate by water utilities has not exceeded an average of 0.5% per year, which would take an estimated 200 years to completely renew the current infrastructure [1]. As water utilities struggle to keep pace with pipe repair orders with little financial means and support at hand, asset managers need to rationalize those resources to anticipate and prioritize the rehabilitation of deteriorated pipes.

Pipe deterioration refers to the process through which a pipe's structural capacity is compromised, thus eventually leading to a failure point when a pipe can no longer withstand internal or external pressure. Pipe deterioration involves complex mechanical and chemical processes that researchers are still working on thoroughly understanding. Modeling these deterioration processes often requires investigating a complex set of factors and their relationship with the deterioration outcome. A wide range of factors has been identified in the literature that contribute to pipe deterioration. These factors can be classified into pipe-intrinsic, environmental, or operational factors [5]. Pipe-intrinsic factors that have been extensively investigated in the literature include pipe material, age, diameter, and length. Other deterioration factors can be either operational, such as internal pressure properties, network operations, and previous failures, or environmental, such as weather conditions, soil hazards, or hydrogeological conditions [5–7]. By using a set of these deterioration factors, previous works have investigated their influence on the deterioration process and aimed to identify how they dictate failure trends.

In recent decades, pipe deterioration models have been extensively developed to characterize pipe deterioration processes, evince failure

[*] Corresponding author.
E-mail address: linasela@utexas.edu (L. Sela).

https://doi.org/10.1016/j.ress.2021.108271
Received 12 March 2021; Received in revised form 10 December 2021; Accepted 11 December 2021
Available online 14 December 2021
0951-8320/

patterns, and anticipate failure events. As described by Kleiner and Rajani [2,8], these models can be classified into either physical or statistical in nature. Physical models study the mechanical properties of pipes and their environment to determine the nature of the influence and how it impacts a pipe's service life. To characterize structural stress and identify failure points, these physical models generally require detailed pipe-level information. Some models include pipe intrinsic properties related to coating, joint types, and wall thickness. Operational information might also be needed to evaluate how a pipe reacts to pressure transients, chemical water properties, or hydraulic pressure. A pipe environment brings another set of components to be considered including pipe burial depth, soil properties, temperature, etc. Ideally, a model that thoroughly characterizes the physical deterioration process would need to include all such information that influences the process. However, in practice, acquiring detailed information about each individual pipe and its environment is a costly endeavor. Most water utilities typically only have general information about their pipe network, which drastically limits the potential of physical models. Nevertheless, physical models can be very useful when detailed analysis is needed for a critical portion of the network or select pipes to understand and model a deterioration behavior of specific interest.

On the other hand, statistical models evaluate deterioration patterns on a larger scale by analyzing population-wide relationships between deterioration and pipe attributes and can hence assess the overall condition of the entire distribution system [9,10]. In 2001, Kleiner and Rajani [2] presented an overview of statistical models, which they classified into three categories: deterministic, probabilistic single-variate, and probabilistic multi-variate models. This review was later updated in 2012 to include another level of classification related to the type of deterioration indicating whether the statistical model was interested in breakage frequency, survival rate, or condition rating [11]. According to the review, deterministic models have generally used grouped data and included time exponential and time linear approaches to estimate the number of breaks or the age at failure. In contrast, probabilistic models have incorporated uncertainty in determining model parameters to analyze the probability of failure, life expectancy, or failure clustering.

In their works, Kleiner and Rajani presented a comprehensive review of statistical models and underlying assumptions, but a unified perspective was still needed for model comparison. To that end, Scheidegger et al. [12] presented a review of statistical models on a comparable basis by formulating the models into their failure rates representation. According to this review, models have used different statistical distributions to characterize failure rates. Exponential, Gamma, or Weibull distributions have been typically used to model the increase in failure rates over time due to aging [13–17]. In these models, the impact of past failures has been incorporated in different terms. Some models identify separate failure rate expressions past different levels of past failures [13–15]. Nevertheless, these models generally include numerous parameters and require large datasets for calibration. Others incorporate the number of past failures as a covariate to account for the effect of repetitive repairs [18,19]. Alternately, other models used more simplistic failure rate expressions; either assuming a continuous increase of the failure rate [16,20,21] or a constant rate [22].

One advantageous set of statistical models used in both exploratory and predictive studies of pipe failure is logistic regression [23–25]. Common advantages recognized in the literature include the relatively simple mathematical framework, avoiding the need to make assumptions regarding the distribution of covariates, and its ability to discern the relative importance of the covariates on pipe deterioration [23]. In the context of pipe failure, logistic regression can be used to model the relationship between a set of covariates consisting of pipe attributes and a probability of a pipe failure. By selecting an appropriate probability threshold, the continuous response variable can be transformed into a binary response variable indicating whether a pipe is deficient (value equal to 1) or non-deficient (value equal to 0).

Several studies have used logistic regression to explore the influence of various factors on pipe failures [26–28]. Previous work using logistic regression has commonly identified age, diameter, pipe material, and length as significant factors influencing pipe deterioration [23,24,26–28]. However, studies seem to yield conflicting results and no consensus has yet been reached as to what set of factors best explains deterioration trends. In particular, Ana et al. [23] applied a logistic regression model on a Belgian sewer pipe network to model the probability of a pipe having a good condition. The authors used a backward stepwise elimination method to select covariates and identified age, non-concrete material, and length as significant covariates. When compared to two other similar studies using logistic regression [24,25], there was no agreement on the significance of any of the selected covariates. Although the underlying causes of pipe failures in pressurized water distribution systems could differ from those responsible for pipe failures in gravity-driven sewer systems, such conflictual results are also common in water distribution system models. In fact, pipe deterioration is often considered a system-specific phenomenon because differences in construction practices, quality standards, and local conditions can highly influence deterioration patterns in a pipe network.

Logistic regression has also been used in several cases to predict pipe failures, and different methods to assess a model's predictive performance have been used [7,29,30]. Ariaratnam et al. [24] specified a logistic regression model to calculate the probability of having a deficient pipe condition which the authors defined as the two lowest condition ratings. Hypothesis testing on three portions of the dataset was used to conclude that the model was stable for condition prediction. A more common measure of the logistic predictive performance has been through classification metrics such as specificity, precision, accuracy, and the Receiver Operating Characteristic (ROC) curve. Salman and Salem [27] compared an ordinal regression, a multinomial logistic regression, and a binary logistic regression model based on their ability to predict sewer pipe condition states according to the pipeline assessment and certification program framework [31]. Their results showed that binary logistic regression performed slightly better with precision scores of 78% for the non-deficient condition and 45.8% for the deficient condition.

When compared with other statistical models for pipe condition assessment, logistic regression has been reported to produce comparable performance. Kleiner and Rajani [32] used several datasets of individual mains from Canadian water utilities for comparing logistic regression, Naïve Bayesian Classification (NBC), Non-Homogeneous Poisson Process, and heuristic models. The authors assessed model performance based on the number of correct predictions and concluded that the logistic regression model had a similar performance to the NBC in the training phase, and no one model showed superior performance. In some cases, binary logistic regression performed better than other commonly used models. In particular, Debón et al. [33] applied a Cox proportional hazard model and a logistic regression model to a medium-sized Spanish city pipe network to predict future failure events. The authors concluded that the logistic model showed a superior fit based on simulated ROC curves.

Further uses of logistic models have included ranking pipes per their likelihood of failure [26,32], and associating the likelihood of failure with information on consequences to assign risk scores to individual pipes as a tool to help water utilities prioritize pipe rehabilitation [27,28]. In fact, pipe condition scoring is often a practical input for asset management. Several methodologies have been developed in recent decades to encode collected information on defects into condition scores. A five-step comprehensive framework for infrastructure asset management and planning long-term investments by integrating various operational and convenience factors including pipe adjacency and group replacement considerations was proposed in [34]. A decision support system for designing intervention programs for water infrastructure was proosed in [35]. The approach first groups water supply and sewer pipes into practical and efficient replacement projects based on their proximity and priority of renewal. Then, a multi-objective genetic algorithm optimizes the work program while integrating the water company's strategic policy into a multi-objective function. Pipe failure probability was also integraed with consequence of failure based the impact of topological changes in the water network [36]. Kley et al. [37] reviewed available condition scoring methodologies in the literature and categorized them into priority-based and substance-based methodologies. Priority-based methodologies assign a score representing the urgency for rehabilitation, and often incorporate information on the level of severity and the density of defects in a pipe. On the other hand, substance-based methodologies are more purpose-driven in the sense that pipes are ranked per the level of rehabilitation required. For example, a defect requiring a no-dig fix will rank better than a defect needing open trench excavation. Scoring methodologies can also be classified based on whether they use ratio or condition assessment methods [38]. Ratio methods incorporate a cost-benefit component by comparing the replacement value to the repair cost (facility condition index, asset condition index). Condition assessment methods, such as subjective grading, defect weighting, and statistical methods, assign scores based on an evaluation of the defects. Because traditional methods often provide a limited interpretation of a pipe's condition, Opila and Attoh-Okine [38] suggested a methodology for calculating the Mean Time to Failure (MTF) based on different pipe failure models to capture how a condition reflects on a pipe's service life. The authors further converted MTF estimates into condition scores according to a flexible scale. Ultimately, the obtained scores incorporated various factors in addition to a water utility's risk level.

Studies using logistic regression have typically reduced the prediction time interval to a period small enough to include at most one failure [39]. However, a given failure may result from unusual stress and not reflect a deteriorating trend, and thus it is important to observe failure history in a more flexible manner to cover longer deterioration trends. Additionally, estimating pipe failure probabilities does not provide a direct measure of criticality that reflects the utility's attitude towards risk, which can assist a water utility in defining maintenance priorities for a specific planning period. To address this research gap, this study proposes a novel decision-making framework that is adaptive to water utilities planning constraints as well as their attitude towards risk. The proposed framework incorporates a condition scoring methodology [38] with a logistic regression model to generate an estimate of the expected remaining service life as a metric for condition assessment of pipelines. Furthermore, the framework allows for a flexible choice of the time interval of the prediction and discount rate, which reflects the utility's propensity to delay or promote rehabilitation efforts, and thus makes it readily usable by water utilities. The intended contribution of this paper is threefold: (1) assess the performance of a logistic regression model with a flexible time-interval choice, (2) provide a measure of the MTF based on a developed logistic regression model, and (3) assign and evaluate pipe condition scores using the MTF measure.

2. Methodology

The proposed approach for pipe failure prediction and condition scoring involves five steps: (1) data collection and processing, (2) developing a logistic regression model to estimate the probability of a pipe failure in a given time period, (3) estimating the mean time to failure for each pipe, (4) assigning scores to each pipe according to a condition scoring method reflecting a water utility's attitude towards risk, and (5) evaluating model performance. The main steps of the proposed approach are illustrated in Fig. 1. The proposed approach was applied on a dataset of 4153 water distribution pipes with 6769 failure events covering a time period from 2000 to 2019 (a detailed description of the data dataset used in this work is provided in Section 3.1).

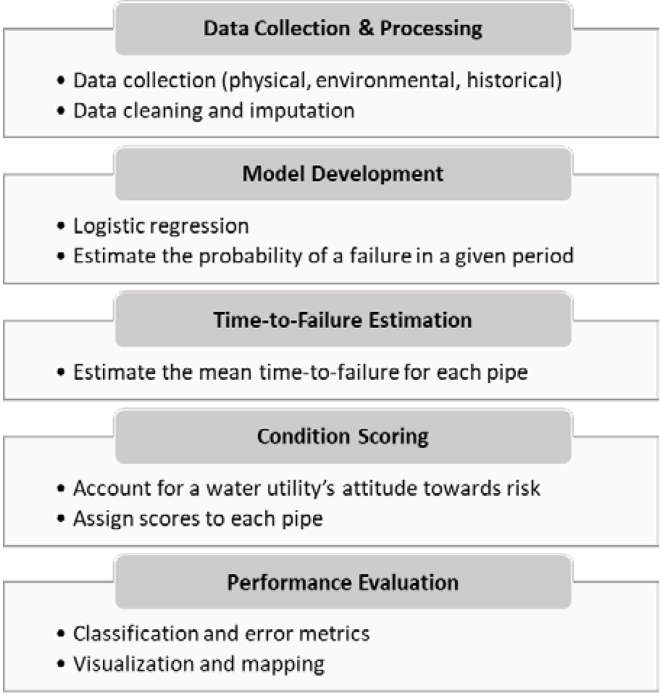
Data Collection & Processing
• Data collection (physical, environmental, historical)
• Data cleaning and imputation
Model Development
• Logistic regression
• Estimate the probability of a failure in a given period
Time-to-Failure Estimation
• Estimate the mean time-to-failure for each pipe
Condition Scoring
• Account for a water utility's attitude towards risk
• Assign scores to each pipe
Performance Evaluation
• Classification and error metrics
• Visualization and mapping

Fig. 1. Research methodology.

2.1. Logistic regression model

2.1.1. Model formulation

The first step of the modeling approach is to estimate the probability of pipe failure using physical, environmental, and historical information of individual pipes. To estimate failure probability, the proposed approach relies on a logistic regression model with a flexible prediction time interval, T. One to several years can be chosen as a T-year period for predicting the probability of pipe failure depending on the resulting model performance and the water utility's preference that reflects its planning horizon.

For an individual pipe i, the logistic regression model estimates the probability of the pipe failure event Y_{ij} occurring in a jth T-year period given a set of pipe covariates represented using a vector X_{ij}, where $Y_{ij} = 1$ indicates that pipe i failed at least once and $Y_{ij} = 0$ otherwise. The covariates include the characteristics of an individual pipe i measured at the beginning of the jth T-year period with $j = 1, 2, \ldots, n_i$, where n_i is the total number of T-year periods covered by a single pipe i's timeline. These covariates can include pipe attributes, environmental, and operational conditions, depending on the available information. In Section 3.1, a detailed description of the pipe information used in this work is provided. Fig. 2 schematically illustrates the timeline of an individual pipe, its covariates, and the prediction time-period T. Each covariate influences the probability of failure according to the regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$. Eq. (1) represents a mean probability p_f of a failure event Y_{ij} for a single pipe i in the jth T-year period given pipe covariates, X_{ij}.

$$\mu_{ij} = p_f\left(Y_{ij} = 1 \middle| X_{ij}\right) = \frac{1}{1 + e^{-X_{ij}^T \boldsymbol{\beta}}} \tag{1}$$

To specify the relationship between covariates and the response variable, regression coefficients $\boldsymbol{\beta}$ are to be estimated. For deterministic models, regression coefficients are typically estimated by maximizing a likelihood function which assumes observations to be independent [40]. However, since the dataset included observations of the same pipe but for different periods, the longitudinal nature of observations could potentially induce correlation across pipe failure responses, thus

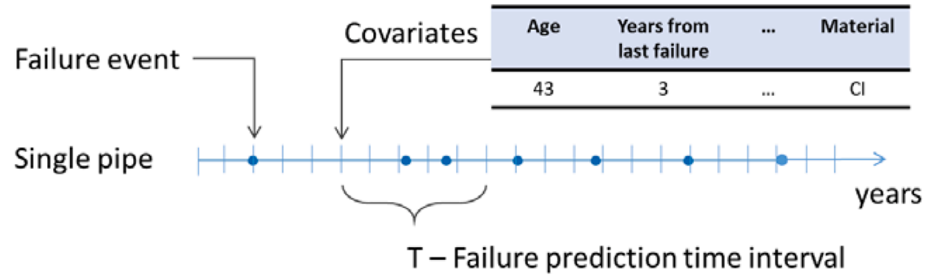

Fig. 2. Diagram of the failure timeline and prediction covariates for an individual pipe.

violating the sensitive assumption of independence.

This limitation can be observed from the way the raw data is restructured in the form of non-overlapping periods of T years. Hence, the same pipe appears as a data point j times in the model. These repeated measures might create a correlation in the dataset that is similar to the standard autocorrelation often exhibited in many time series (e.g., hydrological time series and spatial environmental data) where samples are not spaced enough in time or space. Correlated samples might not provide an accurate representation of the population. However, in statistical models, the purpose is to characterize a population when only a sample of the population is available. Hence, measurements collected from the population - pipe failures in the present study - need to be a reliable representation of the population. It follows that there is a need to account for a potential correlation between samples.

To account for a possible correlation between outcomes for each individual pipe, a Generalized Estimating Equations (GEE) method was used to estimate regression coefficients by incorporating within-cluster effects through the population average [41]. A cluster in the present dataset refers to a single pipe with multiple observations. The GEE method also avoids the need to explicitly specify a probability model of the correlation structure.

Let $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ \ldots, Y_{in_i})^T$ represent the response vector of the ith pipe consisting of n_i observations and $\boldsymbol{\mu}_i = \ (\mu_{i1}, \mu_{i2}, \ \ldots, \mu_{in_i})^T$ refers to the mean vector of failure probability for pipe $\boldsymbol{i}$. Let V_i be the variance-covariance matrix for $\boldsymbol{Y}_i$ defined as $\boldsymbol{V}_i \ = \phi \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i(\alpha) \boldsymbol{A}_i^{\frac{1}{2}}$, where $\boldsymbol{A}_i$is the diagonal matrix of the variances of $\boldsymbol{Y}_i$, $\boldsymbol{A}_i \ = diag\{var(\mu_{i1}), \ldots, var(\mu_{in_i})\}$, $\boldsymbol{R}_i(\alpha)$ is known as the working correlation structure, and ϕ is the error variance. The error variance can be estimated as $\phi \ = \frac{1}{N-p} \sum_{j=1}^{n_i} \sum_{i=1}^{p} e_{ij}^2$, where $e_{ij}$is the response residuals defined as $e_{ij} = (Y_{ij} - \mu_{ij}) \ / \sqrt{var(\mu_{ij})}$ using the current values of the $\boldsymbol{\beta}$ coefficients [42]. $\boldsymbol{R}_i(\alpha)$ is a square matrix of elements $Corr(Y_{ij}, Y_{ik})$ and size $n_i \times n_i$ and is defined based on one of several commonly used types of covariance structures. $\boldsymbol{R}_i(\alpha)$ also depends on a parameter α, which is estimated iteratively based on the number of covariates, p, and response residuals e_{ij}. The parameter α represents the correlation between observations for the same pipe at different T-year periods. Table 1 details the matrix elements and parameter estimation for the independent, exchangeable, and autoregressive correlation structures used in this study. The exchangeable structure assumes the same correlation coefficient across observations for the same pipe, i.e., α defines the correlation strength between each pair of the observations. The autoregressive structure assumes a stronger correlation between failure events that are closer to each other in time. In this case, a higher α will lead to a higher correlation between consecutive observations compared to observations that are farther apart.

Table 1
$\boldsymbol{R}_i(\boldsymbol{\alpha})$ Matrix elements for common working correlation structures.

Correlation structure	$Corr(Y_{ij}, Y_{ik})$	Parameter estimator
Independent	$Corr(Y_{ij}, Y_{ik}) = \{ \begin{smallmatrix} 1\ j = k \\ 0\ j \ \neq k \end{smallmatrix}$	–
Exchangeable	$Corr(Y_{ij}, Y_{ik}) = \{ \begin{smallmatrix} 1\ j = k \\ \alpha\ j \ \neq k \end{smallmatrix}$	$\widehat{\alpha} \ = \frac{1}{N' - p} \sum_{i=1}^{K} \sum_{j \neq k} e_{ij} e_{ik}$ $N' \ = \sum_{i=1}^{K} n_i (n_i \ - 1)$
Autoregressive AR(1)	$Corr(Y_{ij}, Y_{i,j+m}) \ = \alpha^m$, $m \ = 0, 1, \ldots, \ n_i - j$	$\widehat{\alpha} \ = \frac{1}{K_1 - p} \sum_{i=1}^{K} \sum_{j \leq n_i - 1} e_{ij} e_{i,j+1}$ $K_1 \ = \sum_{i=1}^{K} (n_i \ - 1)$

Despite the existing difference among correlation structures, estimates of the regression coefficients are asymptotically consistent even in the event of a misspecification of the correlation structure [43]. For K pipes and p covariates, regression coefficients $\boldsymbol{\beta}$ can be estimated by solving the GEE in Eq. (2):

$$\sum_{i=1}^{K} \frac{\partial \mu_i}{\partial \beta_j} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0 \ j = 1, \ldots, p \tag{2}$$

To decide upon the goodness of fit of a logistic model based on a specified correlation structure that accounts for potential correlation from multiple observations from each pipe, the Quasi-likelihood under the Independence model Criterion (QIC) was used [44]. Unlike likelihood-based methods such as the Maximum-Likelihood (ML), GEE-based models do not explicitly specify a likelihood function. However, the QIC metric provides an alternative to the commonly used Akaike Information Criterion (AIC) metric to compare the goodness of fit for different GEE models, such that a GEE model with a lower QIC value fits better the dataset.

2.1.2. Covariate selection

An important step in the procedure of developing a logistic regression model is the selection of covariates. Covariate selection can improve a model's interpretability, filter out covariates with low relevance without compromising model accuracy, avoid overfitting and improve prediction performance for new observations. In this study, covariate selection is carried out in two steps. First, Least Absolute Shrinkage and Selection Operator (LASSO) regression is used to reduce the number of covariates based on their contribution to the performance of the logistic regression model [45]. Secondly, a Recursive Feature Elimination (RFE) method is performed to further reduce the number of covariates [46].

LASSO regression is a statistical tool that performs variable selection by shrinking less significant regression coefficients to zero [45]. Coefficient shrinkage is possible by integrating an additional term to the error minimization, such that the goal of LASSO regression is to solve:

$$\min_{\beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} - Y_{ij} log(\mu_{ij}) - (1 - Y_{ij}) log(1 - \mu_{ij}) + \lambda \| \boldsymbol{\beta} \|_1 \right\} \tag{3}$$

where μ_{ij} is the predicted probability and Y_{ij} is a failure event for a single pipe i in the jth T-year period given pipe covariates, and λ is a regularization parameter that balances between two objectives: minimizing the error between the predicted failure probability and observed failures (first term) and regularization (second term). The $l1$ norm is defined as

$\| \boldsymbol{\beta} \|_1 = \sum_{i=1}^{p} |\beta_i|$ penalizes a model with many covariates. The rationale for including the $l1$ penalty is that it achieves sparsity by eliminating the predictors that explain the response variable the least. By cross-validating over λ values, the value that yields the best objective function is selected for a given dataset.

An additional step of covariate selection was performed using RFE. The goal of RFE is to select covariates by recursively considering a decreasing number of covariates [46]. First, a logistic regression model is trained on the set of covariates selected after the LASSO regularization step, and the statistical significance of each covariate is obtained using p-values for each covariate's coefficient. The covariate with the highest p-value is eliminated from the current set of covariates and the procedure is repeated on the resulting subsets until the highest p-value is below a specified cutoff (0.05 in this study). The final subset of covariates was then used to develop the final logistic regression model that estimates pipe failure probability for a given T-year period.

The outcome of the logistic regression model provides an estimate of the probability of a pipe failure in a T-year period by integrating the effects of the correlation structure and selected physical, environmental, and historical information. Then, a discrete decision about the state of the pipe can be made by setting a discrimination threshold on a given failure probability of a pipe. If the failure probability exceeds this threshold value, a pipe is expected to fail in the next T-year period, i.e., the failure outcome is equal to 1. If the estimated probability is below the designated threshold value, the pipe is expected to survive, i.e., the failure outcome is equal to 0.

2.2. Estimating mean time to failure

The developed logistic regression model estimates failure probabilities for each pipe, which provides a measure of criticality for a given T-year period. While such a measure can assist a water utility in defining maintenance priorities for a planning period, it does not provide a direct measure of the expected time to failure. To estimate the remaining time to pipe failure, the proposed approach relies on calculating the Mean Time to Failure (MTF). MTF is a reliability parameter typically used to account for the expected life expectancy in the design of products [47]. For repairable systems, MTF refers to the time between failures, i.e., inter-failure time, and can be estimated as the arithmetic mean of the survival probability over time:

$$MTF = \int_{t_0}^{\infty} P_s(t) dt \tag{4}$$

where t_0 denotes the pipe's repair time and $P_s(t)$ is the survival function defined as the probability that a pipe will survive past a time t. For a given pipe, with a number of n T-year periods, Eq. (4) can be approximated as:

$$MTF \approx T \sum_{n=0}^{\infty} P_s(n) \tag{5}$$

where $P_s(n)$ is the probability of survival past time $t = t_o + nT$. The probability that a pipe survives past a time t is approximated by the product of the probabilities that the pipe survives during each of the successive T-year periods leading to time t, with each T-year survival event being conditional on the pipe surviving up to the beginning of the T-year period. Thus, $P_s(n)$ can be approximated as $P_s(n) = \prod_{k=0}^{n} p_s(k)$, where $p_s(k)$ is the conditional probability that a pipe survives during the period from $t_0 + kT$ to $t_0 + (k+1)T$ (i.e., it survived the k-th T-year period with $k = 1, 2, \ldots, n$) given that it survived in all previous intervals for $k > 0$. Thus, the MTF can be approximated as:

$$MTF \approx T \sum_{n=0}^{\infty} \prod_{k=0}^{n} p_s(k) \tag{6}$$

Since the event "at least one failure" is the complement of a survival event, i.e., $p_s(k) = 1 - p_f(k)$, the probability of failure in a T-year period p_f, as estimated by the developed logistic regression model in Eq. (1), can be used to calculate the MTF as follows:

$$MTF \approx T \sum_{n=0}^{\infty} \prod_{k=0}^{n} \left(1 - p_f(k)\right) \approx T \sum_{n=0}^{\infty} \prod_{k=0}^{n} \frac{1}{1 + e^{X(k)^T \beta}} \tag{7}$$

where $X(k)$ represents the vector of covariates measured at the beginning of the kth T-year period. Only time dependent covariates, e.g., pipe age, vary across k values. Therefore, this method converts the failure probabilities in a limited time interval to a measure of the expected time to the next failure. The MTF is a direct measure that can be used by water utilities to decide whether to include pipes in repair and improvement projects.

2.3. Condition scoring

The first outcome of the proposed approach is a T-year probability of pipe failure, and the second outcome is an estimate of the mean time to the next failure. The third step assigns pipe condition scores to facilitate the water utility's risk assessment and prioritize maintenance, replacement, and decide on project scope. Furthermore, the scoring approach is flexible to the utility's risk attitude and the granularity of scores it desires.

The condition scoring method [38], uses the economic concept of discount rate to assign condition scores to pipes based on the MTF estimates. According to its economic interpretation, a discount rate typically implies the extent to which future benefits are valued, where a higher discount rate implies a lower present value of money accrued in the future and a lower discount rate implies a higher present value of money. In this study, a discount rate d is a factor that reflects the utility's attitude towards risk in the condition scoring of pipes, where a higher discount rate reflects a tendency to delay rehabilitation efforts. Given a maximum desired criticality score S_{max}, a discount rate d, and the MTF of a pipe, a pipe's condition score can be determined as:

$$S = \frac{S_{max}}{(1 + d)^{MTF}} \tag{7}$$

This condition scoring method assigns a single score to a pipe, which lumps the impact of various environmental and physical covariates and pipe failure history (as reflected in the MTF), as well as a utility's attitude towards risk (d) and decision scale (S_{max}). Higher scores indicate higher criticality, and higher discount rates suggest that fewer pipes will have high scores for a given MTF, thus reflecting a lower level of rehabilitation priority [38].

Fig. 3 illustrates the continuous condition scoring proposed in Eq. (7) as a function of the calculated MTF proposed in Eq. (6). Based on the curve, scores can be assigned to pipes on either a continuous (solid line in Fig. 3) or a discrete (dashed line in Fig. 3) scale. The stepwise condition scoring can be obtained by rounding the continuous scores to integer values, e.g., a pipe with a continuous score of $3.5 \leq S < 4.5$ will be assigned a discrete score of 4. In the present study, scores were assigned using a discrete scale, which allowed to categorize pipes into a finite number of groups that can serve as a practical input for asset management.

An advantage of using pipe scores is the ability to capture the likelihood of failure as inferred from the dataset without specifically estimating time to failure. In fact, this scoring method incorporates pipes covariates, probability of failure, as well as utilities' preferences, in a simple and easily interpretable single metric that can be used to rank pipes and prioritize rehabilitation efforts.

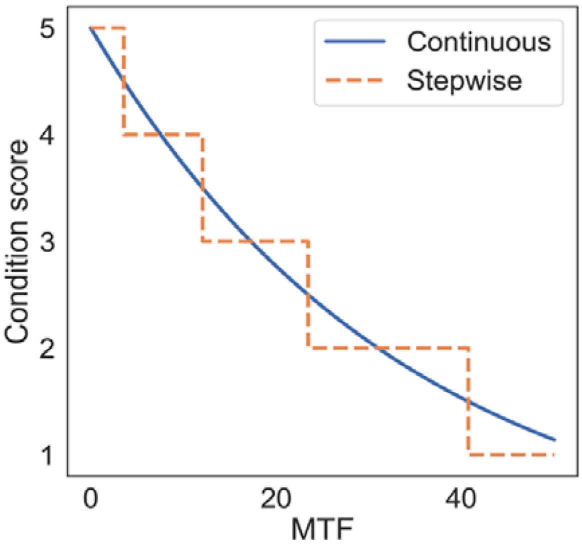

Fig. 3. Condition scoring curve.

2.4. Model evaluation

The proposed framework includes a logistic regression model for pipe failure prediction based on estimated failure probabilities and a condition scoring method using the concept of MTF. In order to evaluate the accuracy of the proposed framework, several classification and error metrics were employed. To evaluate the performance of the logistic regression model, a confusion matrix, which summarizes the performance of a classification model by showing both discrepancy and agreement between true labels and predicted labels, is used [48]. Before computing the confusion matrix, predictions are obtained by converting failure probabilities to a binary outcome (i.e., pipe failure or no pipe failure) by setting a probability threshold. Following the confusion matrix terminology, correctly predicted labels are either True Positives (TP) or True Negatives (TN), and incorrectly predicted labels are either False Positives (FP) or False Negatives (FN). Here, a positive represents a pipe failure, and a negative represents a pipe survival. Additionally, based on the confusion matrix, several performance metrics are calculated. Table 2 shows the calculated metrics and their definitions [49].

The accuracy metric measures the correctly predicted fraction of all pipe failure events. While accuracy treats failure and survival events equally, precision, also called the positive predictive value, measures the correctly predicted fraction of predictions. Ideally, higher values of precision are targeted. However, higher precision is only useful if correct failure predictions make up a higher fraction of all true failure events in the dataset. This latter fraction translates as recall. All performance metrics, including accuracy, recall, and precision, range between 0 and 1, where 1 indicates a perfect classification model, and 0 indicates the opposite. Matthews Correlation Coefficient (MCC) provides an alternative metric that is unaffected by unbalanced datasets. A dataset is called unbalanced if the ratio of true failure events to true survival events in the dataset is significantly low. MCC yields a high score if the model correctly predicts both the majority of failure and survival events. An MCC equal to 1 reflects a perfect prediction, a 0 value represents a random prediction, and −1 reflects an inverse prediction.

Table 2
Model evaluation metrics.

Classification metric	Definition
Precision	$\frac{TP}{TP+FP}$
Recall or True Positive Rate	$\frac{TP}{TP+FN}$
False Positive Rate	$\frac{FP}{TN+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Matthews Correlation Coefficient	$\frac{TP\cdot TN-FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Since predictions are made based on a chosen probability threshold, the defined classification metrics can only be comprehensively interpreted if a threshold value is justified. To decide upon the choice of a probability threshold, Receiving Operating Characteristic (ROC) and Precision-Recall curves are common tools to analyze the impact of a varying threshold on model performance [33,48,50]. A ROC curve is a graphical tool that plots True Positive Rate (TPR) values versus False Positive Rate (FPR) values for a varying threshold. A high TPR indicates the rate of correctly predicted pipes that are expected to fail, and a low FPR indicates the rate of pipes whose failure was incorrectly predicted by the model. Hence, the goal is to achieve a high TPR and a low FPR. A performance metric associated with a ROC curve is the Area Under the Curve (AUC). The closer AUC is to 1, the better the model is at correctly predicting the true events and simultaneously minimizing false predictions.

While the ROC curve allows visualizing how well a classifier captures true labels, ROC curves can be influenced by imbalanced true and positive events. When the number of negative events is much greater than the number of positive events (as typically occurs for pipe failure data where a majority of pipes do not exhibit failures), the FPR can be artificially suppressed making it more difficult to assess the model performance. Instead, the Precision-Recall curve performs better for imbalanced datasets, where precision indicates the fraction of pipes identified by the model to be expected to fail that indeed experience failure, and recall indicates the sensitivity of model prediction [50]. A tradeoff applies between precision and recall as the probability threshold varies. When the probability threshold is low, the number of unidentified failure events is expected to decrease, thus having higher recall values. However, the number of events incorrectly classified as failures will increase as well, thus decreasing the model’s precision. As the probability threshold increases, fewer relevant events will be identified (i.e. lower recall), however, the confidence (i.e., precision) of correctly identified events will be greater. It is useful to plot precision and recall curves against the threshold settings, thus visualizing how different threshold levels specifically influence both curves. Visualizing the precision and recall tradeoff curves allows the water utility to directly set the probability threshold to achieve a desired level of performance.

Classification metrics listed in Table 2 and ROC and Precision-Recall curves are useful to improve failure predictability and, in turn, the MTF and condition scoring by determining the probability threshold. For MTF calculation and condition scoring, results can be evaluated against the observations by comparing the MTF to the actual time to failure for pipes that failed more than once in the observation period by using qualitative and quantitative measures such as histograms, boxplots, and the Root Mean Square Error (RMSE).

3. Application and results

The proposed framework is demonstrated using the information provided by the City of Austin, which included data about pipe characteristics, locations, and failure history. All models developed in this work were implemented in Python 3.7, and preliminary data processing was executed in ArcGIS Pro 2.4.0.

3.1. Data description and preprocessing

The studied drinking water distribution system consists of 244,830 pipe segments with a total network length of 5202.1 miles. Out of the total number of pipes, only 4425 pipes incurred failures that were recorded in the utility’s database. These repaired pipes account for a total of 6989 recorded repair events spanning from 2000 to 2019. A repair event is typically triggered by a reported leak and refers to an intervention from a utility's maintenance team to restore a pipe into service. Prior to considering pipe attributes, the dataset was screened for duplicates and other inconsistencies, including failure events that were not stored in a readable format. After omitting duplicate and inconsistent entries, the dataset analyzed in this study comprised of 6769 failure events from 4153 pipes that had a total length of 336.48 miles representing 6.5% of the entire network length.

Fig. 4. Shows the annual failure rate per unit length for the entire network from 2001 to 2018. The first year 2000 and last year 2019 were excluded from this figure as failure data collection may not have been complete. Across the 2001 to 2018 period, pipes had 5.03 failures per 100 km per year on average with a standard deviation of 2.36. Break rates mostly fluctuated between 4 and 8 failures per 100 km per year. A 2018 survey of water utilities in the USA and Canada reported an average failure rate of 8.7 breaks per 100 km per year, which was compared to other sources reporting failure rates ranging from 13 to 19 breaks per 100 km per year [51]. This report also refers to typical industry targets of 7 to 10 breaks per 100 km per year. This suggests that the failure rate calculated based on the dataset provided by the city of Austin was low. The failure records in the dataset only consisted of pipes representing 6.5% of the entire network, and another portion of the network must have suffered past failures, that however were not recorded. Also, as can be seen in Fig. 4, unusually low failure rates were recorded in 2001 and 2002 with no provided explanation. Similarly, unusually high failure rates were recorded in 2011, which can be partially attributed to the exceptional drought experienced by Texas during 2011. Despite years with unusual rates, the entire pipe failure dataset was considered in the analysis. Excluding outliers was not warranted since individual events could not be directly associated with any identified variability in trends. Also, rejecting some events might influence potential correlations across the pipe network since a pipe failure might have an impact on adjacent pipes or other parts of the network.

Relevant attributes that were provided with the dataset included pipe length, diameter, age, material, and pressure zone. Physical, environmental, and historical information used in this analysis is briefly summarized below.

Pipe material. The majority of pipes consist of cast iron (CI) pipes (71% of pipe length) followed by ductile iron (DI) (6.1%), Polymerizing Vinyl Chloride (PVC) (5%), and Asbestos Cement (AC) (13.7%). Other pipe materials included concrete steel cylinder, polybutylene, and copper, which comprised less than 4%. More than half of the pipes had only one past failure and 77.3% had either one or two past failures in the 20 years observation period.

Pressure zones. Pipe attributes included six main pressure zones, North (NO), Central North (CN), North West (NW), South (SO), Central South (CS), South West (SW), and Others. CS, CN, and N pressure zones included 73.6% of the pipes with recorded past failures.

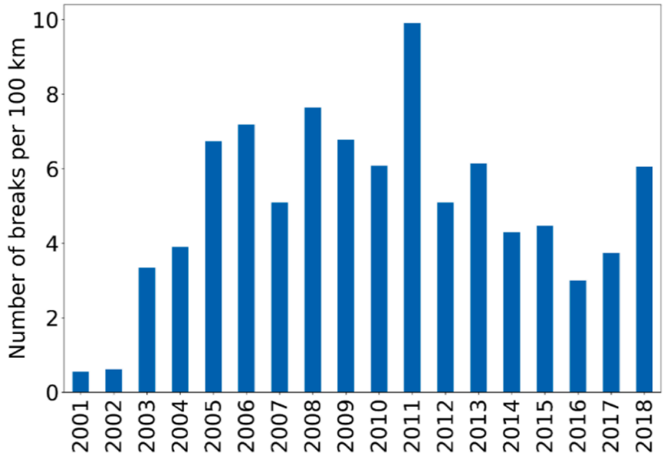

Fig. 4. Failure rate per year.

Pipe age. Fig. 5 shows the distribution of pipe ages by pipe length and material. Newer pipes consist mostly of DI and PVC, and older pipes consist mostly of CI and AC pipes. As common with pipe records, approximately 12% of pipes were missing pipe age. The age of the pipes was approximated using spatial interpolation based on radial basis function [52]. The age of CI pipes was further adjusted based on our discussions with the water utility following the changes in installation practices. As suggested by the water utility, CI pipe installation ceased in the early 1980 s. A cutoff was therefore defined such that estimated installation dates for CI pipes that were dated after 1980 (approximately 3% of all the pipes) were instead approximated by assigning an age value from the geographically nearest pipe that was installed before 1980. This approximation assumed that those CI pipes were installed in the same year as the nearest pipes that were installed before 1980. Such an assumption is reasonable considering that rehabilitation efforts typically target several pipes in a given geographical area for cost considerations.

Soil and land use. Soil information was extracted from the Soil Survey Geographic (SSURGO) Database as provided by the National Cooperative Soil Survey. The database is made publicly available by the United States Department of Agriculture (USDA) [53]. Soil attributes included the dominant soil order, which is defined in accordance with USDA soil taxonomy [54]. The dominant soil order refers to a soil classification that lumps soil properties like depth, structure, and moisture. Additionally, land attributes were assigned to pipes with information on road type and land use as potential covariates [55]. Pipe elevation information was extracted from the 2-ft contour elevations map published by the City of Austin in 2012 [56]. Annual precipitation was also considered as a model covariate and was provided as an average rainfall associated with soil information.

Table 3 summarizes the primary characteristics of the main covariates considered in the model. Overall, 15 different continuous and categorical covariates were considered in the regression model. Note that the actual number of implemented covariates is greater due to dummy coding of categorical variables [57]. All data was standardized by removing the mean and scaling to unit variance before proceeding with the regression analysis. Thus, the values of the regression coefficients reflect the relative importance of the standardized covariates in determining the dependent variable of the regression model.

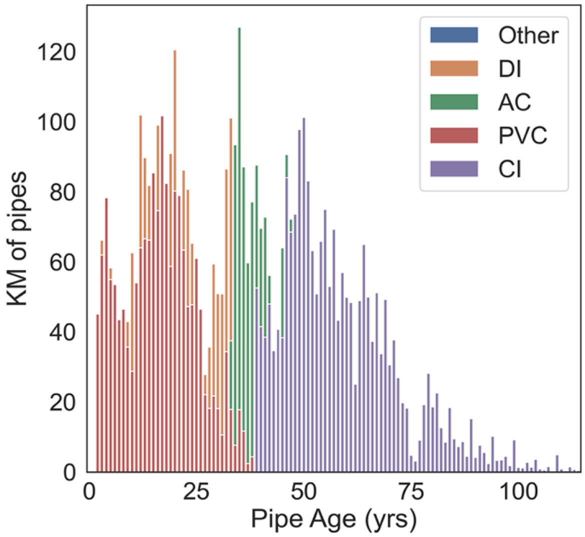

Fig. 5. Distribution of pipe age and material.

Table 3
List of covariates per category.

Category	Covariate	Unit	Main characteristics
Pipe characteristic	Length	ft	Mean: 419.18; Std*: 424.18
	Diameter	inch	Mean: 7.13; Std: 4.64
	Age	years	Mean: 45.17; Std: 18.74
	Material	–	CI; DI; AC; PVC; Other
Failure history	Number of past failures	–	Mode: 0; Mean: 0.51
	Years from last failure	years	Mean: 5.58; Std: 3.80
Soil attribute	Elevation	ft	Mean: 611.41; Std:103.34
	Concrete corrosion potential	–	Low/Moderate/High
	Steel corrosion potential	–	Low/Moderate/High
	Saturated hydraulic conductivity	inch/hr	Mean: 20.75; Std: 30.44
	Dominant soil order	–	Mollisols/Vertisols/Entisols/…
Land attribute	Land use	–	Commercial/Residential/Office
	Road type	–	Interstate/Minor arterials/Private Road
Weather	Mean annual precipitation	mm	Mean: 876.81; Std: 79.70
Operational	Pressure zone	–	CN; NO; CS; SO; NW; SW; Other

* standard deviation (Std).

3.2. Logistic regression results

3.2.1. Model selection

The first step towards estimating the mean time to failure, is selecting the appropriate planning period, i.e., T. Several T-year periods ($T = 1, \ldots, 6$) were applied, trained, and evaluated based on the five performance metrics mentioned previously. For model training, approximately 75% of the dataset with observations from 2000 to 2012 or 2015 (depending on the T-year period) was selected and used to train the logistic regression models. The remaining records were held out for testing and validation. The time periods selected for training and testing of each model are listed in Table S1 in the Supporting Information (SI). Table 4 summarizes the performance of the trained regression models for each T-year period when applied to the test data set. In this study, the optimal period was selected based on three criteria: (a) a high resulting performance across the majority of scores, (b) a period that offers practical implementation for the utility's asset management, and (c) a period that reduces imbalanced classification [58].

Table 4 shows that a period of $T = 1$, results in low precision, recall, AUC, and MCC scores, and although good performance is achieved based on the accuracy and FPR scores, these are mostly attributed to the imbalanced classification of the observations, with less than 6% of failure events versus non-failure events in the dataset. An important issue with imbalanced data is that there may not be sufficient observations belonging to the minority class (i.e., pipes with failures) to adequately represent both distributions. Similar results are observed for $T = 2$ and 3, with low precision, recall, and MCC scores. As the T-year period increases, the performance generally improves, trading off increasing FPR and temporal resolution of predictions.

A period of 5 years was chosen as a T-year response window in this study. In other words, the output of the regression model estimates the failure probability of a pipe in the next 5 years. First, it achieves good performance across all metrics, with the highest MCC scores, second-highest precision, recall, and AUC, while maintaining high accuracy and low FPR, compared to other T-year periods. Additionally, the water utility's Capital Improvement Program follows a 5-year planning window, according to which a budget is allocated for pipe rehabilitation. It follows that a measure of pipe failure risk that covers the allocation period (i.e., $T = 5$) ensures a coherent approach to rehabilitation. In terms of data imbalance (i.e., the number of failure events versus the non-failure events), the shorter the T-year period is, the more imbalanced the dataset becomes. Preprocessing the dataset with a 5-year response variable yielded 32% failure events versus 68% non-failure events, which considerably reduced class imbalance. Consequently, a 5-year period was chosen for its practical application and the higher predictive accuracy it provided. The remaining results shown in the paper refer to a 5-year period; however, the proposed approach generalizes to different planning periods, and can hence be adjusted accordingly. Figs. S1–S3 in the SI show similar results for other response periods.

Table 4
Performance scores for T-year time interval selection.

T	AUC	Precision	Recall	FPR	Accuracy	MCC
1	0.51	0.28	0.28	0.05	0.90	0.23
2	0.67	0.40	0.40	0.09	0.84	0.31
3	0.69	0.46	0.46	0.15	0.77	0.31
4	0.68	0.61	0.61	0.14	0.79	0.46
5	0.68	0.67	0.67	0.14	0.80	0.53
6	0.64	0.70	0.70	0.20	0.76	0.50

To estimate the effects of covariates, the logistic regression model used the GEE with an independent covariance structure. In fact, when compared to an exchangeable correlation (QIC = 13,721.76), the independent structure provided a better fit (QIC = 13,859.32), whereas the model failed to converge with an autoregressive covariance structure. The goodness of fit with an independent covariance structure suggests that failure events across pipes do not display a significant correlation in the present dataset. Additionally, estimates of covariates effects are still consistent despite possible misspecification of the correlation structure [43]. Therefore, the final model estimated coefficients and failure probabilities based on an independent covariance structure.

3.2.2. Effects of covariates

The initial set of covariates was included in the LASSO regression model that was cross-validated across a range of continuous values for the regularization parameter λ. LASSO regression reached an optimum at $\lambda = 0.03$, thus filtering out 22 continuous and categorical covariates. The 25 retained covariates were recursively modeled into a GEE logistic regression model with an independent covariance structure, and variables with the highest p-value were filtered out until the highest p-value of a subset was below a 0.05 cutoff. As an exception, despite its low statistical significance in the dataset, pipe age was retained considering its proven importance in the literature [5,59,60]. The resulting subset of covariates and their corresponding coefficients are shown in Table 5.

For pipe material, only the CI type was retained, which suggests that other material types did not provide sufficient statistical significance to count towards the final subset of covariates. In fact, over 70% of the studied dataset consisted of CI pipes. The consideration of a larger representation of other materials should allow for their analysis with more certainty in terms of impact on failure. Also, despite an expected high influence of steel and concrete corrosivity covariates, their values were only available for a portion of the dataset, which might have led to

Table 5
Logistic regression model coefficients.

Covariate (Alias)	Description	Coefficient
Intercept	Intercept	−0.83
upTime	Years from last failure	0.87
pipeLength	Pipe length	0.20
NOPF	Number of past failures	0.15
pipeMaterial_CI	CI pipe material	0.08
soilOrder_Vertisols	Soil order: Vertisols	0.08
landUse_residential	Residential land use	0.07
pipeAge	Pipe age	0.04
pipeDiameter	Pipe diameter	−0.07
pressureZone_NW	North-West pressure zone	−0.09

their exclusion from significant covariates. When coefficients are ranked from most to least influential, as in Table 5, covariates related to failure history show some of the highest contributions to pipe failure. The number of years from the last failure (upTime) appears as the most influential attribute, thus suggesting that the more time elapses from a previous break, the more likely a pipe is to fail within the next 5-year period. This correlation is also illustrated in Fig. 6. A possible explanation for this effect is that a longer period without failure might indicate a longer exposure to internal and external factors affecting the structural integrity of a pipe. This interpretation supports the “in-usage” and “wear-out” phases of the bathtub failure rate curve assumption where the failure rate is expected to increase until a failure occurs [2].

Additionally, the more total previous breaks are recorded at a pipe level, as integrated by the number of previous failures (NOPF) covariate, the higher the pipe failure probability is. This observation also matches the conceptual failure rate “bathtub” model, in that the failure rate increases as the number of previous failures increases [12]. A rich failure history of a pipe could suggest a structural integrity issue that has been further undermined by repeated repairs. In terms of pipe characteristics, covariates’ importance was generally consistent with previous research findings. Pipe length has been associated with higher failure probability [61–63]. Beyond an additional exposure directly correlated to pipe length, longer pipes could be more exposed to varying environmental conditions and more sensitive to effects like pressure transients [62]. Also in consistence with the literature findings, smaller pipes inversely affect failure probability such that pipes with small diameters are associated with thinner walls which translates into a lower structural strength [59-61,64].

When comparing the logistic regression models having different T-year prediction periods, there was an overall agreement in terms of the most influential covariates and their magnitudes. Table S2 in the SI list the range of the coefficients of the most significant covariates in the logistic regression models. In all the models, the number of years from last failure was the most influential covariate, followed by the number of previous failures and pipe length. Other variables, such as pipe characteristics (i.e., material and diameter) and environmental impacts (i.e., land-use, pressure zone, and soil order) were an order of magnitude less influential, although still significant. Fig. 7 shows the median of failure probabilities versus the time from last failure for different T-periods. As expected, for shorter T-year prediction periods, the probability of failure is lower compared to longer T-year periods. For example, the probability of a pipe failure in the next 3-years is lower compared to its probability to fail in the next 6-years. Hence, the proposed model can be adjusted to the desired prediction period, based on utility’s planning periods, as long as the performance of the models is accounted for, as summarized in Table 4 and discussed previously.

3.2.3. Model performance evaluation

In order to define a discrimination threshold for the developed logistic regression model and make predictions, the ROC curve is first generated for the test data, as shown in Fig. 8. The corresponding AUC is 0.68, thus suggesting a reasonable discrimination strength for predicting pipe failures. By setting a discrimination threshold, the model can be positioned at a specific point along the ROC curve. For example, setting the discrimination threshold at 0.75 would result in 60% TPR (knee point in Fig. 8) just before the slope is sharply reduced. However, as mentioned previously, while the ROC curve evinces the discrimination strength of the model, it is insensitive to the imbalance of the dataset and gives no measure of precision. It might be tempting to seek an additional 10% of TPR by conceding 20% of FPR (by adjusting the probability threshold from 0.69 to 0.53), but a marginal increase in the FPR, which is twice the marginal increase in the TPR, could result in a number of false alarms that is much higher than twice the additional number of correct predictions.

To account for the model’s precision, the precision-recall versus discrimination threshold curves are plotted in Fig. 9. The precision-recall curves can be visually used to control for the correct proportion of total predictions based on threshold values. While the objective is to maximize both precision and recall, the two metrics are conflicting, and a level of compromise needs to be determined. A choice of a discrimination threshold should be determined based on an acceptable level of performance for each metric. Acceptable levels may be determined per the priorities of the water utility. For example, a water utility might want to account for the fact that missing a true failure event is worse than having a false alarm. In fact, because the loss in recall is typically more costly than a similar loss in precision, setting a recall level that is higher than precision could be warranted. In this study, no such preference was expressed by the utility, hence the chosen probability discrimination threshold (0.69) was determined as the intercept of precision and recall such that both metrics are at 67%. By defining such a threshold, 67% of true failure events were correctly predicted by the model, and 67% of predicted failures corresponded to true failure events.

Using the designed discrimination threshold of 0.69, the confusion matrix is computed for the test set in Table 6. According to this confusion matrix, the model accuracy was calculated at 80%, the FPR at 14%, and

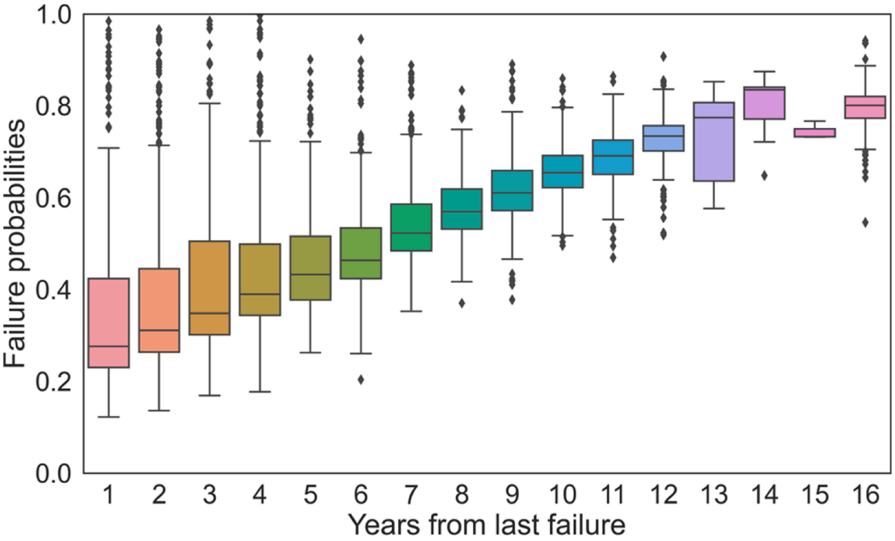

Fig. 6. Failure probabilities versus the time from last failure. Whiskers are 1.5 times the interquartile range, any data point beyond is considered an outlier.

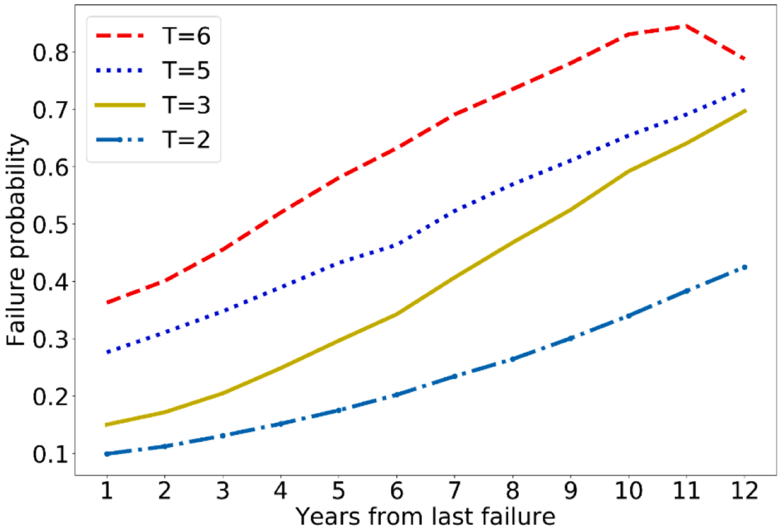

Fig. 7. Median failure probabilities versus the time from last failure for different T-year periods.

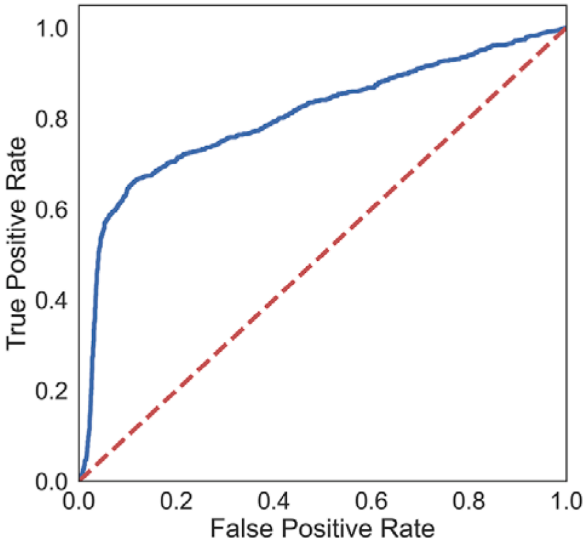

Fig. 8. ROC curve.

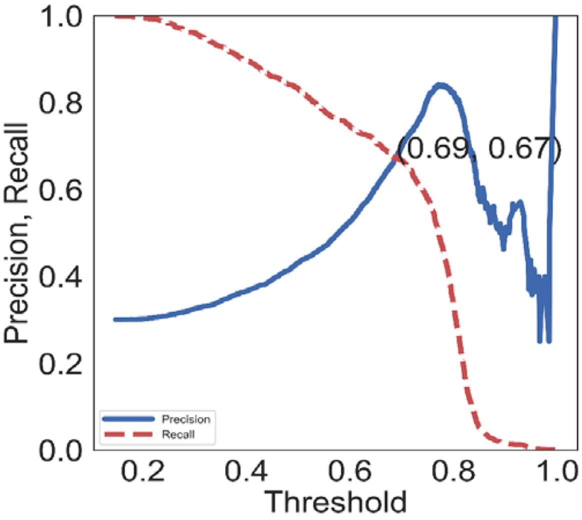

Fig. 9. Precision (solid line) and recall (dashed line) versus discrimination threshold.

the MCC was equal to 0.53. Making a direct comparison with other papers is difficult since the performance metrics are heavily dependent on the problem formulation, the choice of method, and the used datasets. Additionally, previous studies reported different metrics. For instance, [65] reported average MCC around 0.25, precision around 0.15 and recall around 0.45; [29] reported a maximum recall of 0.036 and AUC 0.773; [7] reported average AUC around 0.75, precision between 0.3 and 0.5 and recall between 0.4 and 0.55; and [66] reported 0.58 precision. Overall, compared to the metrics typically reported in the literature [7,29,65,66] our model's predictive strength was deemed satisfactory. Figs. S1–S3 in the SI show the failure probability curves, precision and recall, and the confusion matrices for the regression models with $T = 3$ and 6.

3.3. Mean time to failure and condition scoring

Logistic regression provided failure probabilities for limited time intervals. The MTF equation allowed us to further use these probabilities to compute the expected times to failure given the selected covariates of each pipe (as listed in Table 5). Fig. 10 shows how the obtained values evolve over time from the previous failure for the entire data set. As can be seen, the expected time to failure is shorter as the time from last failure increases. Also, the MTF average values decrease from around 6 years to below 1 year with decreasing standard deviations. Low uncertainty associated with shorter MTF values for longer elapsed times since last failure reflects the pipes with a higher failure probability. It is noteworthy to mention that MTF values do not exceed 12 years, which is induced by a high failure rate in the dataset. In fact, the dataset that was used to calculate MTF consisted of only pipes with at least 1 failure event in a 20-year observation period. Consequently, MTF calculations do not reflect the normally expected pipe life expectancies in the entire network, but instead, they give an expected time between failures for pipes with characteristics and a failure history similar to those in the observed dataset.

Table 6
Confusion matrix with the 0.69 probability threshold.

	Predicted non-failure	Predicted failure
True non-failure	2526	411
True failure	411	845

To validate the estimates, an error was measured as the difference

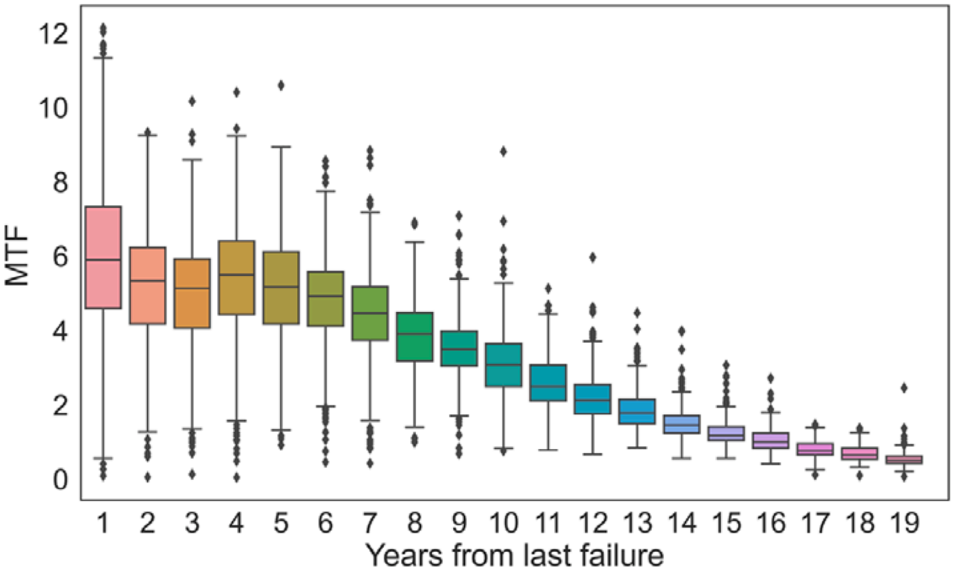

Fig. 10. Mean time to failure versus time from last failure. Whiskers are 1.5 times the interquartile range, any data point beyond is considered an outlier.

between MTF values for both training and test data and the actual time between failures. The error was calculated for 1222 pipes that failed in at least two different years, such that the actual time between failures could be measured. As a result, the error had a near-normal distribution with a mean equal to 0.54 years and a standard deviation of 3.10 years. The root mean square deviation associated with the error was equal to 3.29 years. Although an MTF value was on average off by more than 3 years, the near-zero mean suggests a tendency towards correct predictions. A larger sample could potentially reduce the deviation and lead to more accurate MTF estimates.

The final step of obtaining a pipe score was conducted on the entire 20-year dataset utilizing the scoring equation (Eq. (7)). The scoring scale of 1 to 5, with 5 indicating high criticality, was chosen to match the water utility's existing scoring scale. Based on this scale, the scoring curve was charted for different discount rates, as shown in Fig. 11.

The choice of the discount rate should reflect a water utility's attitude towards risk, and maintenance and replacement strategy. As can be seen, a higher discount rate leads to a decreased condition score for a given MTF, thus reflecting a propensity to delay rehabilitation efforts by increasing the portion of pipes with low scores. Intuitively, these score curves can be viewed as the "present value" of "future" pipe failure. In other words, a pipe failure that is expected to occur in the near future (i.e., lower MTF) is valued more highly (or critically) by the utility (i.e., has a higher score) as a candidate for replacement/rehabilitation. Similarly, a pipe failure that is expected to occur farther in the future (i.e., higher MTF), does not require urgent replacement, and thus will get a lower score. The slopes of the curves are controlled by the rate, d, which represents the utility's attitude towards risk. The diminishing slope in the score curves represents the diminishing value of failures that will occur further into the future. For example, an extremely conservative and risk-averse utility will have very low rate values, e.g., with $d = 0$, all pipes will get the maximum score Smax, regardless of their expected time to failure. On the other hand, a less conservative utility, e.g., with $d = 0.5$, will assign a score greater than 3 only to pipes with a 1-year or less MTF, i.e., that are expected to fail in the next year.

A discount rate of 0.2 was selected to reflect a conservative maintenance approach, and a stepwise scoring curve with discrete values was utilized for practical implementation by the water utility (Fig. 12). As in MTF calculations, assigned scores were also updated each time an annual failure was recorded. By using the last assigned scores, a water

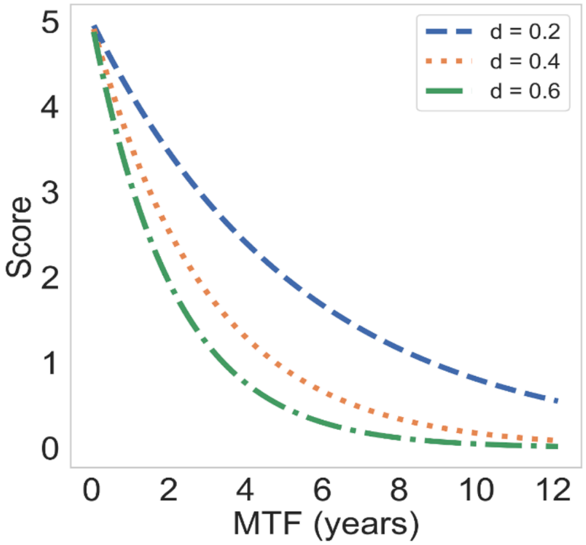

Fig. 11. Condition scores as a function of MTF for different discount rates.

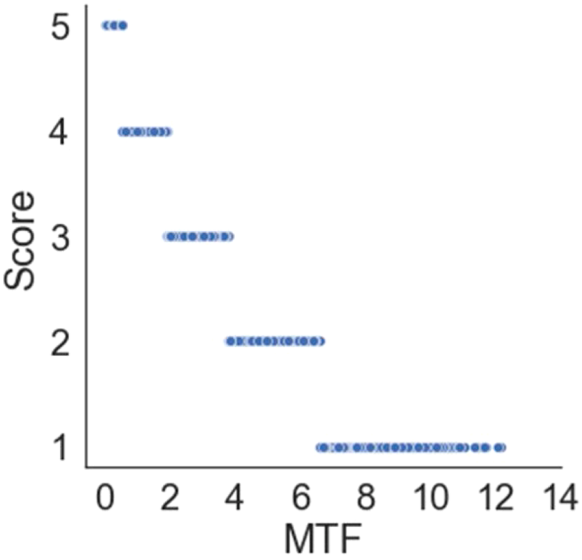

Fig. 12. Stepwise scoring curve using a 0.2 discount rate.

utility can analyze the latest criticality of its pipes.

Fig. 13 displays a map of a portion of the city's water distribution network given the last assigned scores. This condition scoring map can be easily integrated in any spatial software, e.g., ArcGIS, which can be used as a communication tool to share the results among the different divisions involved in pipe condition assessment, including operations, planning, and asset management. Although beyond the scope of this study, future research could further incorporate spatial correlation into the condition score assessment [6,9].

To evaluate the scoring method, scores were assessed against actual time to failure. Pipes with lower condition scores, in general, took more time to fail again. This result suggests that assigned condition scores can give a plausible measure of the criticality for the pipes' condition. By analyzing the proportions of network length per condition score, it is noted that 8.4% of the studied pipe network's length has a score of 5, 29.6% has a score of 4, 28.2% has a score of 3, 29.5% with a score of 2, and 4.4% has a score of 1. Out of the portion of the network having a score of 5, 88.9% of the length consisted of pipes with 15 to 19 years elapsed from last failure. This proportion is consistent with the inferred covariates' effects which suggested that a longer time from last failure leads to higher failure probability. In terms of pipe diameter, 89.4% of pipes with the worst score of 5 had diameters less than 8 inches, as opposed to 86.6% for all scores. As demonstrated in the logistic model, the covariate for pipe diameter had an effect equal to −0.07 which led to the 2.8% difference. Although minor, this result is consistent with literature findings which support that pipes with smaller diameters (less than 8 inches) tend to have a higher likelihood to fail and thus a worse condition score. Additionally, pipes older than 60 years made up 28.9% of pipes scoring a 5 or 4 as opposed to 26.5% for all scores. This difference suggests the tendency for older pipes to have a worse condition score, and the small percentage is due to the low effect of the pipe age covariate.

While condition scores incorporate how deterioration factors influence failure probability for each pipe, they do not provide a measure of the consequence of failure. Risk assessment methods typically include both criticality and consequence scores when prioritizing asset management [36]. Yet, to assign an integrated risk score, an advantage of the described condition scoring method is its linear scale [38], since condition scores are considered a present value of a future failure event based on a chosen discount rate. For example, a pipe with a condition score of 4 is twice as critical as a pipe with a score of 2. A risk score can thus be simply obtained by multiplying the assigned condition score by a consequence score. The resulting risk score can eventually be used to rank pipes per risk level [38].

4. Practical implications and limitations

The key limitations of the proposed approach primarily stem from data restrictions. The present study focuses exclusively on the pipes that experienced past failures by trying to estimate condition scores based on the expected time to failure, i.e., MTF. This approach is appropriate for the regression-based modeling method implemented in this study. Unlike the majority of previous regression-based models, the main outcome of the proposed approach is based on estimating the MTF as opposed to a binary decision. Regression-based models cannot explicitly account for censoring in the failure dataset (i.e., pipes that did not experience any failures during the observation period). One way to get around this limitation is by setting the MTF for pipes that did not experience any failures to be equal to the total observation period. Nevertheless, such practice could result in significant errors since the actual time to failure may be significantly longer than the observation period. Additionally, although a model for predicting MTF for pipes that did not experience failures can be developed, this model cannot be validated, and will hence have limited usability for water utilities. Hence, the MTF calculations do not reflect normal life expectancies for all the pipes in the entire network, but instead, MTF gives an expected time between failures for pipes with characteristics and a failure history similar to those in the observed dataset.

Our modeling approach was primarily motivated by the practical need and current practice of the water utility to be able to plan capital

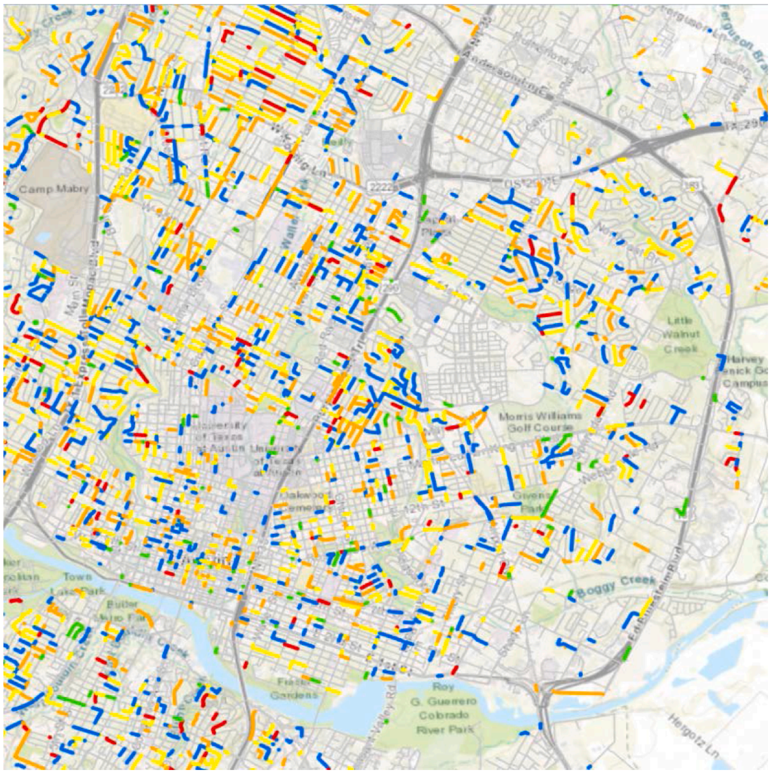
Camp Mabry
Little Walnut Creek
Morris Williams Golf Course
Boggy Creek
Roy G. Guerrero Colorado River Park
University of Texas at Austin
Oakwood Cemetery
Town Lake Park

Fig. 13. Predicted pipe condition scores (1-green, 2-blue, 3-yellow, 4-orange, 5-red).

investment projects. Involving the water utility during this project was a critical step towards practical implementation and for the utilization of more advanced methods in their current decision-making processes. Hence, although our model is limited to pipes that had at least one failure, it still provides insights and is practically useful for assessing the condition of this critical subset of pipes, which included over 4000 pipes for the utility under consideration. In sum, the pipe failure prediction problem, although extensively studied in the literature, has been shown to be challenging in terms of generalizing conclusions across different systems. Pipe deterioration and the results of prediction models are heavily dependent on the specific system characteristic, local conditions, and the available datasets [7,9,26].

It is worth noting that the academic-utility partnership provided various advantages to this study. This included gaining direct access to data as well as institutional knowledge and expertise. The utility provided ample time for meetings, discussions, sharing institutional knowledge, and data with the research team. The development of the models required dedicated personnel with expertise in hydraulic engineering, statistical modeling, and programming, in addition to wider support from the utility's network of practitioners to synthesize and incorporate the data and knowledge into the modeling and analysis. The developed models and methods (including logistic regression, mean time to failure estimation, selection of the T-year prediction period, performance metrics, discrete scoring, rate, and visualization) were discussed with the utility throughout the study during multiple meetings between the research team and utility personnel. The outcomes of this study (data, models, and codes) were shared with the utility. The flexibility to adjust to different time horizons and the simplicity of the outcome of the scoring method, while relying on sound theoretical principles, was of key importance for the utility. Although the research team prioritized using open-source software (Python) for model development and analysis, commercial GIS-based software was also used extensively to communicate the results with the water utility and to synthesize with current data management practices. Undoubtedly, further work is needed to incorporate research outcomes in the current decision-making process for assessing the state of the pipes in the distribution network, which includes personnel training, integration with current data management practices, long-term validation, as well as other technical and organizational considerations.

This study enjoyed the support of a proactive and forward-looking water utility. Nonetheless, this study still encountered several challenges associated with data collection, quantity, and quality, which are symptomatic of the broader water sector [67,68]. While the statistical approaches allowed handling some of the uncertainties associated with the recorded data, they heavily relied on historical data recorded by the utilities over a long period of time in order to properly infer pipe break probabilities. However, most utilities do not have detailed and ample enough records of their infrastructure and pipe break data. Moreover, the advantages mentioned above could be barriers for smaller and budget-constrained utilities with limited accessibility to skilled personnel, data, and software. Overall, this study contributes to the body of studies that highlight the need for academic-utility partnerships for sharing resources and expertise for a successful knowledge transfer to advance water infrastructure management [69].

5. Conclusions

This paper proposed and tested a systematic approach to capture the criticality of pipes in a water supply system using GEE logistic regression, and to assign practical condition scores for asset management prioritization. A pipe network dataset was first preprocessed to define a T-year failure outcome variable and extract features that provide information on soil, traffic, land use, failure history, and operational attributes. A GEE logistic regression model was then specified with reasonable accuracy in estimating the probability of recording at least one failure in a 5-year time interval. Beyond a measure of a period-specific criticality for pipes as provided by the logistic classifier, the MTF metric estimated the expected inter-failure times. The estimates were then used to apply a flexible scoring approach to discriminate pipes based on their criticality. The pipe scoring provided condition metrics with a reasonable ability to predict poor conditions.

The promising results would still need further validation with larger datasets. An accuracy of 80% was achieved by the logistic classifier, but specifying the model on failure records covering a period longer than 20 years might mitigate the uncertainty related to the described performance metrics. Additionally, the MTF calculations have a fundamental assumption that past trend perpetuates. Because failure history is used only from the last 20 years, the model does not provide a full simulation of a pipe's life cycle. Therefore, accuracy is bound to decline as predictions are made farther into the future. Also, uncertainty underlying the logistic regression model is accumulated as the MTF calculations integrate probabilities infinitely into the future. The choice of the time-interval in the logistic model is also a factor that influences this uncertainty. It follows that failure probabilities generated by the logistic regression model are theoretically provided with higher performance compared to pipe scores. However, failure probabilities only provide information on a time-interval specific condition, whereas pipe scores attempt to additionally capture a practical measure of the service life. These limitations in the application of this methodology might justify for a water utility to choose between using probability outcomes or pipe scores depending on the need. For example, a water utility that prepares a 5-year rehabilitation plan could use 5-year probabilities as a measure of risk, whereas using 5-year failure probabilities might not suffice in integrating risk in a long-term rehabilitation strategy.

The suggested framework demonstrates that useful results can be inferred using a GEE logistic model on a dataset covering a limited time interval and suffering potential censorship. Overall, the proposed methods provided two practical outcomes: (1) a predictive logistic regression model to help prioritize rehabilitation for a specific time interval that is determined based on the quality of the dataset and on the utility's preference, and (2) an integrated condition scoring model to estimate pipe criticality. Future research could further assess the performance of the presented model by using larger and high-quality datasets as they become available. Also comparing the logistic regression model to other statistical and data-driven models could provide further analysis of the performance [70]. Beyond a classical performance evaluation, this paper intended to provide a flexible framework that can adapt to real-world complexity that water utilities have to contend with. Research has shown that deterioration modeling can be region- and system-specific, and results may differ per local conditions. Hence, developing models that not only deliver good performance but also allow for flexible use is of key importance for water utilities to be able to use and rely on model predictions.

CRediT authorship contribution statement

Talha M. Rifaai: Methodology, Software, Visualization, Data curation, Writing – original draft. **Ahmed A. Abokifa:** Writing – review & editing. **Lina Sela:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgments

The authors would like to thank Austin Water for providing the data, guidance, and support throughout this study. This work was supported by the Fulbright Program and partially by the National Science

Foundation under Grants 1943428 and 1953206.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ress.2021.108271.

References

[1] ASCE (American Society of Civil Engineers). "Infrastructure Report Card." Reston, VA, USA: ASCE, 2017.
[2] Kleiner Y, Rajani B. Comprehensive review of structural deterioration of water mains: statistical models. UrbanWater 2001;3(3):131–50. https://doi.org/10.1016/S1462-0758(01)00033-4.
[3] Jensen HA, Jerez DJ. A stochastic framework for reliability and sensitivity analysis of large scale water distribution networks. Reliab Eng Syst Saf 2018;176:80–92. https://doi.org/10.1016/j.ress.2018.04.001. March.
[4] Kamali B, Ziaei AN, Beheshti A, Farmani R. An open-source toolbox for investigating functional resilience in sewer networks based on global resilience analysis. Reliab Eng Syst Saf 2021;218:108201. https://doi.org/10.1016/j.ress.2021.108201. PB.
[5] Barton NA, Farewell TS, Hallett SH, Acland TF. Improving pipe failure predictions: factors effecting pipe failure in drinking water networks. Water Res 2019;164: 114926. https://doi.org/10.1016/j.watres.2019.114926. Elsevier LtdNov.
[6] Abokifa AA, Sela L. Identification of spatial patterns in water distribution pipe failure data using spatial autocorrelation analysis. J Water Resour Plan Manag 2019;145(12):1–12. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001135.
[7] Jara-Arriagada C, Stoianov I. Pipe breaks and estimating the impact of pressure control in water supply networks. Reliab Eng Syst Saf 2021;210:107525. https://doi.org/10.1016/j.ress.2021.107525. May 2020.
[8] Rajani B, Kleiner Y. Comprehensive review of structural deterioration of water mains: physically based models. UrbanWater 2001;3:151–64. https://doi.org/10.1016/S1462-0758(01)00032-2. 3.
[9] Chen TYJ, Guikema SD. Prediction of water main failures with the spatial clustering of breaks. Reliab Eng Syst Saf 2020;203:107108. https://doi.org/10.1016/j.ress.2020.107108. May 2019.
[10] Tang K, Parsons DJ, Jude S. Comparison of automatic and guided learning for Bayesian networks to analyse pipe failures in the water distribution system. Reliab Eng Syst Saf 2019;186:24–36. https://doi.org/10.1016/j.ress.2019.02.001. January.
[11] Liu Z, Kleiner Y, Rajani B. “Condition assessment technologies for water transmission and distribution systems”. United States Environmental Protection Agency (EPA), National Service Center for Environmental Publications (NSCEP); March 2012, EPA/600/r-12/017.
[12] Scheidegger A, Leitão JP, Scholten L. Statistical failure models for water distribution pipes -A review from a unified perspective. Water Res 2015;83: 237–47. https://doi.org/10.1016/j.watres.2015.06.027. Elsevier LtdOct.
[13] P. Eisenbeis, “Modelisation statistique de la prevision des defaillances des conduites d’eau potable,” Doctoral dissertation; Université Louis Pasteur (Strasbourg). 1994.
[14] Gustafson JM, Clancy DV. Modeling the occurrence of breaks in cast iron water mains using methods of survival analysis. In: Proceedings of AWWA Annual Conference. Denver: American Water Works Association; 1999.
[15] Pelletier, Geneviève. "Impact du remplacement des conduites d’aqueduc sur le nombre annuel de bris." PhD diss., Université du Québec, Institut national de la recherche scientifique, 2000.
[16] Economou T, Kapelan Z, Bailey T. A zero-inflated Bayesian model for the prediction of water pipe bursts. In: Proceedings of the 10th annual water distribution systems analysis conference, WDSA 2008; 2009. p. 724–34. https://doi.org/10.1061/41024(340)61.
[17] Scheidegger A, Scholten L, Maurer M, Reichert P. Extension of pipe failure models to consider the absence of data from replaced pipes. Water Res 2013;47(11): 3696–705. https://doi.org/10.1016/j.watres.2013.04.017.
[18] Kleiner Y, Rajani B. I-WARP: individual water mAin renewal planner. Drink Water Eng Sci 2010;3(1):71–7. https://doi.org/10.5194/dwes-3-71-2010. May.
[19] Le Gat Y. Une extension du processus de Yule pour la modélisation stochastique des événements récurrents : application aux défaillances de canalisations d’eau sous pression,”. Doctoral dissertation. AgroParisTech - Sciences de l’eau (Option Statistique) 2009.
[20] Røstum J. Statistical modelling of Pipe failures in water networks. Nor Univ Sci Technol 2000:1–132. February.
[21] Francis RA, Guikema SD, Henneman L. Bayesian belief networks for predicting drinking water distribution system pipe breaks. Reliab Eng Syst Saf 2014;130:1–11. https://doi.org/10.1016/j.ress.2014.04.024.
[22] Watson TG, Christian CD, Mason AJ, Smith MH, Meyer R. Bayesian-based pipe failure model. J Hydroinform 2004;6(4):259–64. https://doi.org/10.2166/hydro.2004.0019. IWA PublishingOct.
[23] Ana E, et al. An investigation of the factors influencing sewer structural deterioration. Urban Water J 2009;6(4):303–12. https://doi.org/10.1080/15730620902810902. Aug.
[24] Ariaratnam ST, El-Assaly A, Yang Y. Assessment of infrastructure inspection needs using logistic models. J Infrastruct Syst 2001;7(4):160–5. https://doi.org/10.1061/(ASCE)1076-0342(2001)7:4(160). Dec.
[25] Davies JP, Clarke BA, Whiter JT, Cunningham RJ, Leidi A. The structural condition of rigid sewer pipes: a statistical investigation. Urban Water 2001;3(4):277–86. https://doi.org/10.1016/S1462-0758(01)00036-X.
[26] Yamijala S, Guikema SD, Brumbelow K. Statistical models for the analysis of water distribution system pipe break data. Reliab Eng Syst Saf 2009;94(2):282–93. https://doi.org/10.1016/j.ress.2008.03.011. Feb.
[27] Salman B, Salem O. Modeling failure of wastewater collection lines using various section-level regression models. J Infrastruct Syst 2012;18(2):146–54. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000075. May.
[28] Cooper NR, Blakey G, Sherwin C, Ta T, Whiter JT, Woodward CA. The use of GIS to develop a probability-based trunk mains burst risk model. Urban Water 2000;2(2): 97–103. https://doi.org/10.1016/S1462-0758(00)00047-9.
[29] Robles-Velasco A, Cortés P, Muñuzuri J, Onieva L. Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliab Eng Syst Saf 2020;196(November 2019):106754. https://doi.org/10.1016/j.ress.2019.106754.
[30] Robles-Velasco A, Cortés P, Muñuzuri J, Onieva L. Estimation of a logistic regression model by a genetic algorithm to predict pipe failures in sewer networks. OR Spectr 2021;43(3):759–76. https://doi.org/10.1007/s00291-020-00614-9.
[31] Khaleghian H, Shan Y, Lewis P. Development of a quality assurance process for sewer pipeline assessment and certification program (PACP) inspection data. Pipelines. ASCE: Phoenix, Arizona; 2017. p. 360–9. August 6-9.
[32] Kleiner Y, Rajani B. Comparison of four models to rank failure likelihood of individual pipes. J Hydroinform 2012;14(3):659–81. https://doi.org/10.2166/hydro.2011.029.
[33] Debón A, Carrión A, Cabrera E, Solano H. Comparing risk of failure models in water supply networks using ROC curves. Reliab Eng Syst Saf 2010;95(1):43–8. https://doi.org/10.1016/j.ress.2009.07.004. Jan.
[34] Ramos-Salgado C, Muñuzuri J, Aparicio-Ruiz P, Onieva L. A comprehensive framework to efficiently plan short and long-term investments in water supply and sewer networks. Reliab Eng Syst Saf 2022;219. https://doi.org/10.1016/j.ress.2021.108248.
[35] Ramos-Salgado C, Muñuzuri J, Aparicio-Ruiz P, Onieva L. A decision support system to design water supply and sewer pipes replacement intervention programs. Reliab Eng Syst Saf 2021;216. https://doi.org/10.1016/j.ress.2021.107967.
[36] Phan HC, Dhar AS, Hu G, Sadiq R. Managing water main breaks in distribution networks-–A risk-based decision making. Reliab Eng Syst Saf 2019;191(July): 106581. https://doi.org/10.1016/j.ress.2019.106581.
[37] Kley G, Kropp I, Schmidt T, Caradot N. Review of available technologies and methodologies for sewer condition evaluation,” Project sema. Kompetenzzentrum Wasser Berlin gGmbH 2013.
[38] Opila MC, Attoh-Okine N. Novel approach in pipe condition scoring. J Pipeline Syst Eng Pract 2011;2(3):82–90. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000081. Aug.
[39] Wilson D, Filion Y, Moore I. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. Urban Water J 2017;14(2): 173–84. https://doi.org/10.1080/1573062X.2015.1080848. Taylor and Francis Ltd.Feb.
[40] Myung IJ. Tutorial on maximum likelihood estimation. J Math Psychol 2003;47(1): 90–100. https://doi.org/10.1016/S0022-2496(02)00028-7. Feb.
[41] Hardin JW, Hilbe JM. Generalized estimating equations. 2nd Ed. Chapman and Hall/CRC; 2012.
[42] Wang M. Generalized estimating equations in longitudinal data analysis: a review and recent developments. Advances in Statistics 2014:303728. https://doi.org/10.1155/2014/303728.
[43] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13–22. https://doi.org/10.1093/biomet/73.1.13.
[44] Pan W. Akaike’s information criterion in generalized estimating equations. Biometrics 2001;57(1):120–5.
[45] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 1996;58(1):267–88.
[46] Louw N, Steel SJ. Variable selection in kernel Fisher discriminant analysis by means of recursive feature elimination. Comput Stat Data Anal 2006;51:2043–55. https://doi.org/10.1016/j.csda.2005.12.018.
[47] Birolini A. Reliability engineering: theory and practice. 7th ed. Berlin Heidelberg: Springer; 2014.
[48] Giraldo-González MM, Rodríguez JP. Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks. Water 2020;12(4):1153. https://doi.org/10.3390/W12041153. SwitzerlandApr.
[49] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics 2020;21(1):1–13. https://doi.org/10.1186/s12864-019-6413-7.
[50] Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on machine learning; 2006. https://doi.org/10.1145/1143844.
[51] Folkman S. Water main break rates in the USA and Canada: a comprehensive study. Mechanical and Aerospace Engineering Faculty Publications; 2018. p. 1–49. March.
[52] G.B. Wright, “Radial basis function interpolation: numerical and analytical developments,” Ph.D. Dissertation. University of Colorado Boulder, Department of Applied Mathematics. 2003.
[53] USDA, “Web soil survey,” Natural Resources Conservation Service, United States Department of Agriculture. p. http://websoilsurvey.nrcs.usda.gov/, 2015.
[54] Staff SS. Soil taxonomy: a basic system of soil classification for making and interpreting soil surveys. U.S. department of agriculture handbook 436. 2nd editio. United States Dept. of Agriculture, Naturel Resources Conservation Service; 1999.

[55] Texas Department of Transportation “TxDOT open data portal.” [Online]. Available: https://gis-txdot.opendata.arcgis.com/. [Accessed: 03-Feb-2021].

[56] City of Austin, “2012 2-foot contours,” [Online]. Available: https://data.austintexas.gov/Geodata/2012-2-foot-Contours/bxuc-pk4k. [Accessed: 03-Feb-2021].

[57] Venkataramana M, Subbarayudu M, Rajani M, Sreenivasulu KN. Regression analysis with categorical regressor variables. Int J Stat Syst 2016;11(2):8. https://doi.org/10.2307/2987272.

[58] Japkowicz N, Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intelligent data analysis 2002;6(5):429–49.

[59] Mackey T, Cashman A, Cumberbatch R. Identification of factors contributing to the deterioration and losses in the water distribution system in barbados. Paris: UNESCO; 2014. Google Search.

[60] Wang Y, Zayed T, Moselhi O. Prediction models for annual break rates of water mains. J Perform Constr Facil 2009;23(1):40–6. https://doi.org/10.1061/(ASCE)0887-3828(2009)23:1(47). Jan.

[61] Berardi L, Giustolisi O, Kapelan Z, Savic DA. Development of pipe deterioration models for water distribution systems using EPR. J. Hydroinform 2008;10(2):113–26. https://doi.org/10.2166/hydro.2008.012. Mar.

[62] Boulos PF, Karney BW, Wood DJ, Lingireddy S. Hydraulic transient guidelines for protecting water distribution systems. J Am Water Works Assoc 2005;97(5):111–24. https://doi.org/10.1002/j.1551-8833.2005.tb10892.x. May.

[63] Sattar AMA, Ertuğrul ÖF, Gharabaghi B, McBean EA, Cao J. Extreme learning machine model for water network management. Neural Comput Appl 2019;31(1):157–69. https://doi.org/10.1007/s00521-017-2987-7. Jan.

[64] Jun HJ, Park JK, Bae CH. Factors affecting steel water-transmission pipe failure and pipe-failure mechanisms. J Environ Eng 2020;146(6):04020034. https://doi.org/10.1061/(asce)ee.1943-7870.0001692. Jun.

[65] M. Rahbaralam, D. Modesto, A. Abdollahi, F.M. Cucchietti, and D. Barcelona, “Predictive analytics for water asset management: machine learning and survival analysis,” arXiv Preprint. arXiv2007.03744, pp. 1–19, 2020.

[66] Kumar A, et al. Using machine learning to assess the risk of and prevent water main breaks. In: Proceeding of the ACM SIGKDD international conference on knowledge discovery data min; 2018. p. 472–80. https://doi.org/10.1145/3219819.3219835.

[67] Flancher D. 2019 State of the Water Industry: A Rising Tide? Journal-American Water Works Association 2019;111(7):70–7.

[68] R. Kadiyala and C. Macintosh, “Leveraging other industries - big data management (Phase I),” The Water Research Foundation (WRF) 2018. Project: SENG7R16/4836.

[69] Keck JC, Lee J. A new model for industry-university partnerships. J Am Water Works Assoc 2015;107(11):84–90. https://doi.org/10.5942/jawwa.2015.107.0161.

[70] Fan X, Wang X, Zhang X, Yu X. Machine learning based water pipe failure prediction: the effects of engineering, geology, climate and socio-economic factors. Reliab Eng Syst Saf 2021:108185. https://doi.org/10.1016/j.ress.2021.108185.