Urban Water Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/nurw20

Integrating spatial clustering with predictive modeling of pipe failures in water distribution systems

Ahmed A. Abokifa & Lina Sela

To cite this article: Ahmed A. Abokifa & Lina Sela (2023): Integrating spatial clustering with predictive modeling of pipe failures in water distribution systems, Urban Water Journal, DOI: 10.1080/1573062X.2023.2180393

To link to this article: https://doi.org/10.1080/1573062X.2023.2180393

URBAN WATER JOURNAL
https://doi.org/10.1080/1573062X.2023.2180393

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Integrating spatial clustering with predictive modeling of pipe failures in water distribution systems

Ahmed A. Abokifa[a] and Lina Sela[b]

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

ABSTRACT

Pipe failures in water distribution infrastructure (WDI) have significant economic, environmental and public health impacts. To alleviate these impacts, repair and replacement decisions need to be prioritized to effectively reduce failure rates. In this study, a computational framework is proposed for WDI asset management that couples spatial clustering analysis with predictive modeling of pipe failures. First, hotspot/coldspot clusters of statistically significant high/low failure rates are identified using local indicators of spatial association. Second, the predictive abilities of eight statistical learning techniques are systematically tested, and the best-performing method is implemented to forecast failure rates, (breaks/(km.year)) within different sectors of the WDI. Third, the framework is implemented to compare the impact of adopting proactive instead of reactive pipe replacement strategies. Applying the framework to a real-life, large-scale WDI revealed that spatial clustering of pipe failures improves the accuracy of the prediction models.

ARTICLE HISTORY
Received 15 July 2022
Revised 05 January 2023
Accepted 09 February 2023

KEYWORDS
Asset management; pipe failure; pipe replacement; spatial clustering; spatial regression; water distribution systems

Introduction

Aging water supply systems across Europe and North America face increasing pressure to satisfy the demands of the rapidly growing urban population (Hering et al. 2013). In the United States, the aging water distribution infrastructure (WDI) incurs nearly 0.24 million water main breaks every year, wasting over two trillion gallons of treated drinking water (American Society of Civil Engineers 2017). Fiscal investments needed to rehabilitate and upgrade the WDI are immense (American Water Works Association 2012). Yet, the gigantic scale of water distribution systems, together with the fact that most pipes are buried and inaccessible for routine physical inspection (Kleiner and Rajani 2001), poses significant challenges toward prioritizing pipe maintenance decisions. To improve the efficacy of asset management programs, computational tools have been proposed to guide pipe repair-or-replacement (RoR) decisions in order to reduce the rates of water main failures (Folkman 2018; Stone et al. 2002). The support provided by such computational tools can help alleviate the substantial economic and environmental impacts of water main failures due to water and energy losses, as well as the social impacts represented by service interruptions and road closures.

In addition to the economic, environmental and social impacts, the deteriorated physical condition of WDIs can pose significant implications for public health due to the potential for contaminant intrusion. Previous epidemiological studies observed a strong association between increased cases of gastrointestinal illness and concurrent breaches of the physical and hydraulic integrities of WDIs (Ercumen, Gruber, and Colford 2014). Such lapses in the hydraulic integrity of the WDI, represented by substantial pressure drops, may occur either due to routine operational procedures, such as pump and valve operations, or as a result of extreme events, such as transmission main bursts and sudden changes in water demands. In many cases, disinfectant residual concentrations are not sufficient (or even nonexistent) to rapidly inactivate extraneous pathogenic contaminants entering the system, which puts the consumers at risk of infection with waterborne diseases.

To alleviate the above-mentioned impacts, significant research efforts have aimed to develop computational tools for pipe failure prediction (Rifaai, Abokifa, and Sela 2022; Scheidegger, Leitão, and Scholten 2015; Shirzad and Safari 2020; St. Clair and Sinha 2012; Wilson, Filion, and Moore 2015). Such predictive models constitute a key component of multi-criteria decision support frameworks that can be used to inform pipe rehabilitation decisions (Barton, Hallett, and Jude 2022). The majority of these works focused on forecasting pipe failures using either physical- or statistical-based models (Alizadeh et al. 2019; Kleiner and Rajani 2001; Konstantinou and Stoianov 2020; Rajani and Kleiner 2001). Physical-based models aim to simulate the physical mechanisms of pipe failure, and hence require a significant number of parameters that are specific to the pipe under study. The application of physical-based models is hence mostly limited to pipe failure prediction in major transmission mains where the impacts of pipe failure are most significant.

In addition to physical-based models, various statistical models have been proposed in the literature (Nishiyama and Filion 2013; Yamijala, Guikema, and Brumbelow 2009), which can generally be classified into either deterministic or

CONTACT Ahmed A. Abokifa abokifa@uic.edu
Supplemental data for this article can be accessed https://doi.org/10.1080/1573062X.2023.2180393

probabilistic models (Kleiner and Rajani 2001), but can also be classified based on the modeled entity (individual pipes vs. the entire pipe network) and modeled events (occurrence of failures vs. end of pipe lifetime) (Scheidegger, Leitão, and Scholten 2015). Recent works have also developed models for forecasting the occurrence probability of each type of pipe failure (Shin et al. 2016). Previously developed statistical models extensively relied on a wide array of survival analysis methods, which include proportional hazards models (PHMs) such as Cox-PHM and Weibull-PHM (Jenkins, Gokhale, and McDonald 2015; Kimutai et al. 2015). These models aim to predict the time-to-failure by estimating the probability that a break will occur at some time in the future for individual pipe segments.

While survival PHMs can technically be applied with any level of data availability (Kleiner and Rajani 2001), extensive failure records collected over long periods of time are crucial for properly inferring pipe break probabilities (Yamijala, Guikema, and Brumbelow 2009). Many utilities have only recently started collecting and curating pipe break data in a consistent and extensive manner. Such short-term datasets can still provide useful information through the application of data-driven methods that characterize pipe failures and the contributing factors in an aggregate manner depending on varying levels of available information. These include multiple linear and nonlinear regression models (Wang, Zayed, and Moselhi 2009) that predict annual failure rates as a function of different covariates (e.g. pipe material, diameter, age and length). Regression models have also been extended by relaxing the normality assumption to produce generalized linear models (Yamijala, Guikema, and Brumbelow 2009), and by incorporating the uncertainty in the model parameters using Bayesian regression (Kabir et al. 2015). In addition, recent works have implemented non- and semi-parametric statistical learning methods for pipe failure prediction. For instance, Berardi et al. (2008) introduced the implementation of evolutionary polynomial regression (EPR) for pipe failure prediction. EPR was also implemented by Xu et al. (2011) to develop pipe break models for the water distribution system of the city of Beijing, and by Laucelli et al. (2014) in examining the relationship between climate-related predictors and pipe failure. Kakoudakis, Farmani, and Butler (2018) used EPR and artificial neural networks (ANNs) to examine the influence of weather conditions on pipe failure. Tabesh et al. (2009) found that ANN models gave more accurate pipe failure predictions compared to neuro-fuzzy and multivariate regression models. Almheiri, Meguid, and Zayed (2021) developed a deep neural network framework to predict the risk index of pipe failure considering the effects of different factors including seasonal variation, chlorine content and traffic conditions. Fan et al. (2022) examined the performance of five different machine-learning algorithms in predicting pipe failures, including LightGBM, ANNs, Logistic Regression, K-Nearest Neighbors (KNNs) and Support Vector Classification for pipe failure prediction. Other methods examined in the literature include graph convolutional neural network-integrated deep reinforcement learning (Fan, Zhang, and Infrastructure 2022).

In addition to predictive modeling, a few studies focused on the exploratory analysis of pipe failure data with the aim of identifying unusual (i.e. non-random) patterns of pipe failure (Christodoulou et al. 2012; de Oliveira et al. 2011b; de Oliveira, Garrett, and Soibelman 2011a; Oliveira, Garrett, and Soibelman 2009). These studies focused primarily on spatial clustering analysis to reveal regions within the WDI characterized by particularly high/low failure rates. Although such analysis enables examining the dependence of the failure rates within these critical regions on the local characteristics of the WDI, limited attempts have been made in previous literature to leverage the useful outcomes provided by clustering analysis in the development of failure prediction models. In a recent study, Chen and Guikema (2020) explored whether the use of spatial clusters as an explanatory variable can improve the accuracy of pipe break machine learning models. In this study, results of the clustering analysis were added as one of the explanatory variables in the machine learning models, which overall lead to improving the accuracy of the models. A similar approach was adopted by Aslani, Mohebbi, and Axthelm (2021), where the results of spatial clustering were added as independent variables to improve the predictions of machine learning failure models. Additionally, Kakoudakis et al. (2017) implemented K-means clustering to partition the training data for EPR pipe failure models.

While these attempts have shown the value of leveraging the outcomes of spatial clustering analysis in improving the accuracy of failure prediction models, a few important questions remain unanswered. First, in most of these studies, the results of the clustering analysis were included in the set of explanatory variables used to develop the models. However, a different way of implementing the results of clustering analysis is by developing separate models for different clusters. The rationale for this is that failure patterns within different clusters are driven by factors that may potentially be different from those driving the failures in other clusters. Hence, better failure prediction can be achieved by using different sets of explanatory variables to predict failure rates in different clusters. Second, the characteristics of the WDI that serve as failure predictors (e.g. pipe material and age) may themselves exhibit unique spatial patterns. These patterns are inherently attributed to the way WDIs evolve to accommodate population growth and cities' expansion. Such spatial patterns exhibited by failure predictors have generally been ignored by previous studies, which can potentially bias the estimation of the predictive models by over- or understating the importance of the predictors (Chi and Zhu 2008).

In a recent study, the use of spatial autocorrelation analysis (SAA) was applied for the identification of pipe failure patterns by revealing the locations and statistical significance of hot- and cold-spot clusters of pipe failures (Abokifa and Sela 2019). Building upon this recent work, this study proposes an integrated approach that couples spatial clustering analysis with predictive pipe failure modeling. The contributions of this study are (1) proposing a framework for developing cluster-specific models that are locally tailored to incorporate different sets of predictors for different clusters, (2) developing a novel approach for explicitly accounting for spatial patterns exhibited by failure rate predictors (e.g. pipe age and material) in the development of the failure prediction models through the inclusion of spatially lagged predictors (SLPs) and (3) presenting an integrated framework through which pipe-failure

records collected over a short time period (3 years) can be leveraged in extracting useful information that can aid in asset management operations of a large metropolitan water utility.

Methodology

Overview

Given information about the layout and characteristics of the pipe network, and the locations of historical pipe failures, the proposed approach involves three main steps. First, local indicators of spatial association (LISA) are employed to identify hotspot and coldspot clusters of pipe failure and to verify their statistical significance. Second, predictive models are constructed to develop relationships between annual failure rates (AFRs) and pipe characteristics in each of the identified hotspot/coldspot clusters, while explicitly accounting for spatial patterns exhibited by failure predictors. To this end, the performance of eight different statistical learning methods that represent a wide array of linear and nonlinear multi-parameter functions of different complexities are compared, and the best-performing model is selected. Third, cluster-specific prediction models are used to assess the impact of different pipe RoR strategies on reducing the AFR.

Given the short time span of the failure dataset, individual pipe failures are aggregated, and the prediction models are constructed for small groups of pipes. This is done by first dividing the domain of the WDI into a number of *zones* that can either follow the layout of a regularly spaced grid (e.g. square/rectangular cells), or have irregularly shaped boundaries based on pre-defined pressure/service zones or zip codes. The AFR for each zone is calculated by dividing the total number of reported failures within the zone boundaries by the total length of pipes by the time period of the study (i.e. breaks/(km.year)). Pipe characteristics, including age, diameter, length and materials of pipes are extracted from the network GIS files, and their mean values are computed for each zone to serve as the set of candidate explanatory variables in the predictive models. Depending on the size of the failure dataset, and the time span over which it was collected, the proposed framework can be flexibly applied at any desired level of spatial resolution for making pipe RoR decisions.

Spatial clustering analysis

Local Indicators of Spatial Association

Spatial association (autocorrelation) analysis (SAA) examines the degree to which a specific process of interest is correlated to itself in space by assessing the relationship between the observed value of the phenomenon at any location and the values of the same phenomenon at adjacent locations (Legendre 1993). Here, local indicators of spatial association (LISA), based on Local Moran's *I* index (Anselin 1995), are employed to reveal spatial clusters of pipe failures based on the observed AFR in each zone. For any zone *i*, the I_i index is calculated as (Anselin 1995)

$$I_i = \frac{y_i - \bar{y}}{s^2} \sum_{j=1, j\neq i}^{N} w_{ij}\left(y_j - \bar{y}\right) \quad (1)$$

where N is the number of zones; y_i and y_j are the observed values of the AFR at zone i, and its neighboring zone j, respectively; $\bar{y}$ and s are the mean and standard deviation of the observed AFR across all zones; and w_{ij} is the spatial weight assigned to the connection between zones i and j, which can take any value in the range [0,1]. Zones i and j are considered 'neighbors' if the Euclidian distance between their centroids is less than a selected threshold distance (d_{thr}). For neighboring zones, the pairwise weight is positive, while ($w_{ij} = 0$) for non-neighboring zones. More discussion on how the spatial weights are formulated is provided in section S1 of the supporting information (SI).

The value of I_i can range anywhere between [−1, 1]. For zones where $I_i > 0$ (i.e. positive autocorrelation), neighboring zones have similarly high or low AFR, and hence zone i is considered part of a cluster. For zones where $I_i < 0$ (i.e. negative autocorrelation), zone i is considered an outlier since neighboring zones have dissimilar AFRs. Zones belonging to clusters are further examined to reveal whether the observed AFRs within the zone itself and within neighboring zones are above or below the mean AFR across all zones. Accordingly, cluster zones are classified into either high-high (HH) zones, which are zones with high AFR in a high AFR neighborhood, or low-low (LL) zones which are the exact opposite. Finally, hotspots are defined as clusters of neighboring HH zones, while coldspots are clusters of neighboring LL zones.

Statistical significance testing

To test the statistical significance of the identified clusters/outliers, a *p*-value is computed as a test statistic of the null hypothesis that the observed spatial pattern is simply the outcome of spatial randomness (i.e. the pipe failure occurs randomly across the study domain). In order to compute the *p*-value for any zone (p_i), the distribution of I_i at the zone under the null hypothesis needs to be known. To avoid making any assumptions about the distribution (e.g. normal distribution), a set of r random permutations is generated. For each permutation, the observed AFR values for all zones (except zone i) are randomly shuffled across the domain, and the I_i index is recalculated. This process generates a distribution of I_i values that represent the null hypothesis of spatial randomness for each zone. The *p*-value is calculated as $p_i = (m+1)/(r+1)$, where m is the number of instances from the generated distribution that are greater than the observed I_i index. The smaller the value of p_i, the higher the statistical significance of the identified cluster/outlier for zone i. To prevent the potential inflation of false-positive rates due to multiple comparisons, the *p*-values are corrected by means of the False Discovery Rate (FDR) method of Benjamini and Hochberg (1995). A significance level is then imposed by selecting a cutoff *p*-value above which the identified clusters/outliers are deemed non-significant. The results reported in this work consider a significance level of 0.01 using a set of $r = 999$ random permutations.

Predictive modeling and analysis

Statistical learning methods

A wide range of supervised statistical learning methods exists with different degrees of complexity and interpretability (Obringer and Nateghi 2018). These methods can be broadly classified into parametric models and non-parametric models depending on whether or not the relationship between the response variable and the predictors is assumed to follow a specific function (James et al. 2013). Here, the capabilities of eight different models are tested to compare the performance of a wide variety of parametric and non-parametric methods for the prediction of the AFR. The tested models include four parametric-linear and four data-driven non-parametric learning methods as listed in Table 1.

The first class of models tested herein comprises four parametric-linear models, which all define the relationship between the AFR and the predictors as a multiple linear regression (MLR) function but use different procedures for estimating the regression parameters, namely the non-regularized ordinary least square (OLS) estimation, the *l* 1-regularized least absolute shrinkage and selection operator (LASSO) estimation (Tibshirani 1996), the *l* 2-regularized ridge (RD) estimation (Hoerl and Kennard 1970) and the mixed *l* 1,*l*2-regularized elastic-net (EN) estimation (Zou and Hastie 2005). The second class of models tested in this study comprises four non-parametric models, namely support vector regression (SVR) (Smola and Schölkopf 2004), random forest (RF) regression (Liaw and Wiener 2002), artificial neural network (ANN) regression (Specht 1991) and K-nearest neighbor (KNN) regression (Altman 1992). Details on the mathematical formulations of the examined models can be found in section S2 in the SI.

For all eight models, the response variable (y_i) is the observed AFR at any zone *i*, and is considered a function of a set of *M* predictors $x_i = (x_{i,1}, \ldots, x_{i,M})$ that represent the characteristics of the WDI within the zone (e.g. mean age of pipes, the fraction of certain pipe materials, etc.). The entire dataset comprising the AFR and its predictors in all the zones is lumped as (y, X), where y is the $N \times 1$ vector of the observed AFR in all zones: $y = \{y_1, \ldots, y_N\}$; and $\mathbf{X}$ is the $N \times M$ matrix of predictors for all zones: $\mathbf{X} = [x_1, \ldots, x_N]$. The values for the AFR and each of the predictors are first standardized to have a zero mean and a standard deviation of 1 before developing the predictive models.

The rationale for testing parametric-linear models is that they are generally easy to construct and that they seamlessly lend themselves to statistical inferencing (Obringer and Nateghi 2018). Nevertheless, since the dependencies in real data are rarely of a linear nature, such linear models possess limited flexibility as they often fail to fully capture the complexity of the true relationships. On the other hand, non-parametric models offer a great deal of flexibility in representing non-linear relationships since they directly harness the available data to approximate the relationships. Yet, data-driven non-parametric learning methods are particularly data-intensive and are generally more prone to overfitting than parametric models (James et al. 2013).

Model performance assessment and tuning

The full dataset (y, X) is randomly split into 70% for training and 30% for validation of all the aforementioned models. A two-sample Kolmogorov–Smirnov (K–S) test is conducted to verify whether the training and validation sets follow similar distributions. Performance assessment of all the models was based on two metrics, namely the mean absolute error (MAE) and the root mean squared error (RMSE). These error metrics were calculated for both the in-sample (training) data and out-of-sample (validation) data to asses both the explanative ability and the predictive accuracy of the models, respectively.

Except for the OLS model, each of the predictive models tested herein comprises one or more hyper-parameter that requires tuning. For the parametric-linear models LASSO and RD, the hyper-parameters are either the λ_1 or λ_2 regularization parameters, while for EN, both regularization parameters are assumed to be equal (in this study) thus yielding one hyper-parameter ($\lambda_1 = \lambda_2 = \lambda$). Non-parametric models typically comprise more than one hyper-parameter, which makes tuning them a rather complex task. Herein, only one hyper-parameter is tuned for each of the non-parametric models, namely the tolerance margin (ε), the number of trees (T), the number of neurons in the hidden layers (B) and the number of nearest neighbors (K), for the SVR, RF, ANN and KNN models, respectively. Other hyper-parameters (e.g. number of hidden layers in ANN, tree depth in RF) are kept constant as explained in section S2 in the SI. To tune the hyper-parameters and reduce overfitting of all the models, a cross-validation process is implemented (James et al. 2013). First, a range of hyper-parameter settings is generated for each model by enumeration. Then, using 10-fold cross-validation, each hyper-parameter setting was iteratively trained with 90% of the training data and then tested on the remaining 10%, and the average mean squared error (MSE) of the testing was calculated for the 10 trials. The hyper-parameter setting that resulted in the minimum average MSE was selected.

Table 1. Statistical learning methods examined for the prediction of the AFR.

Class	Model	Full name	Description
Parametric-linear	OLS	Ordinary least squares	Multivariate linear regression model
	LASSO	Least absolute shrinkage and selection operator	*l*1-norm extended OLS model to induce sparsity and avoid overfitting
	RD	Ridge	*l*2-norm extended OLS model to induce sparsity and avoid overfitting
	EN	Elastic net	Joint *l*1, *l*2-norm extended OLS model to induce sparsity and avoid overfitting
Non-parametric	SVR	Support Vector Regression	A kernel-based regression method where the cost function ignores the errors within a specific tolerance margin
	RF	Random forest	An ensemble-based model averaging the output of multiple bootstrapped regression trees
	KNN	K-nearest neighbor	A similarity-based regression model that returns the average of the K nearest neighbors from the training space
	ANN	Artificial neural network	A multilayer perceptron regression model where each neuron computes a non-linear function on the weighted average of neurons in the previous layer

Incorporating spatial influence

The effects of spatial dependence are incorporated into the predictive models by including spatially lagged predictors (SLPs) for all zones. The SLPs are computed as the weighted sum of the corresponding predictors observed at neighboring zones. For instance, if the predictor of interest is the mean age of pipes in the zone, the corresponding SLP is calculated as the weighted sum of the mean ages of pipes within neighboring zones using the spatial weights (w_{ij}) as described in section S1. The intuition behind including SLPs is that the observed AFR at any zone is not expected to be exclusively dependent on the pipe characteristics within the zone, but will also depend to some degree on the characteristics of pipes within adjacent zones. Section S3 in the SI provides additional mathematical details on the special case of incorporating SLPs in the parametric-linear models represented by the MLR function, which gives the spatial cross-regressive model (Anselin 2002; Florax and Folmer 1992).

Cluster-specific prediction models

Finally, the outcomes of the clustering analysis are leveraged to further improve the predictive accuracy of the prediction models by developing separate cluster-specific models for the hotspot and non-hotspot zones of the WDI. The rationale here is that each model can better predict the AFRs within its specific regions by accounting for the local characteristics of each cluster. To develop the cluster-specific prediction models, the dataset (y, X) is divided into two subsets representing the zones that belong to the hotspot cluster and those that do not, and then train two separate models for the hotspot zones using the corresponding (y^{HS}, X^{HS}) subset, and the non-hotspot zones using the (y^{NHS}, X^{NHS}) subset. Using the same breakdown of data (70% for training and 30% for validation), the explanative ability and predictive accuracy of the cluster-specific models are compared to those of a single network-wide model using the same performance metrics, i.e. MAE, and RMSE.

Asset management decision-making

After fitting the cluster-specific models, the developed framework is implemented to assess the potential advantage of adopting proactive instead of reactive pipe repair or replacement (RoR) strategies. A reactive strategy is simulated by a scenario in which the utility maintains the status quo of targeting pipes with previously reported breaks for RoR. Alternatively, a proactive strategy that leverages the outcomes of the framework presented herein to target certain pipes within the critical hotspot regions for RoR is proposed. The developed prediction models are then used to assess the outcomes of each of the proposed strategies by examining the expected reduction in the AFR across the WDI for each strategy.

Results and Discussion

Model application

The proposed framework is demonstrated on a dataset of pipe failures retrieved from the routine maintenance records of a large metropolitan water utility in the U.S. supplying approximately 150 MGD of treated drinking water to over one million consumers. The WDI under study comprises over 8,625 km of pipes and spans 1,400 km^2. A detailed description of the physical pipe characteristics, including the length, diameter, approximate age and material of the pipes, is available in section S4 in the SI. The failure dataset consists of 5,506 records of pipe failures that were repaired by the utility over a period of 36 months from September 2016–2019. Each failure record comprises the time of initial report and repair completion, and the geographic location of the failure. The spatial distribution of the pipe failures and the layout of the case study WDI is depicted in Figure S1.

As mentioned above, the domain of the WDI is first divided into a set of zones to overcome the short time-period of the failures dataset. Herein, a regular grid of square-shaped zones with the size of 500 m × 500 m (0.25 km^2) was implemented. This resulted in breaking down the WDI domain into 2655 zones with at least 1 km of pipes/zone, of which 1,400 zones had at least one failure event during the 3-year period. The number of failures per zone ranged from 1 to 30 failures/zone, and the AFR ranged from 0.04 to 1.98 failures/(km.year). The distribution of the AFR across the zones of the studied WDI is featured in Figure 1(a).

Pairwise spatial weights (w_{ij}) constitute a key component of both the clustering analysis and predictive modeling conducted in this study. As explained above, these weights are a function of the threshold distance (d_{thr}) that determines the span of spatial effect for the different zones that is being accounted for in the analysis. Hence, d_{thr} needs to be carefully selected in order to properly capture the spatial patterns exhibited by the AFR. Different values for d_{thr} in the range of 2–16 km were examined, and a $d_{thr} = 10$km was found to capture the spatial pattern in the AFR with the highest statistical significance. Section S5 in the SI provides a detailed analysis of the selection of the threshold distance.

Spatial clustering analysis

The first step of the proposed approach involves identifying clusters of exceptionally high and low failure rates. Following the approach described above, local Moran's I indices are calculated according to Eq. 1 to reveal hotspot/coldspot clusters of pipe failures. Out of the 2,655 zones, the calculated I_i is positive for 1,726 zones (65%) and negative for 929 zones (35%), which indicates that on a global level, the failure data is spatially clustered (i.e. neighboring zones have similarly high or low AFR) instead of dispersed. Testing for statistical significance (with $r = 999$ random permutations and a significance level of 0.01 on the FDR corrected p-values), 492 zones were found to belong to statistically significant hotspot clusters and 449 zones belonged to statistically significant coldspot clusters (of which, only 62 have a non-zero AFR). Figure 1(b) depicts the identified hotspot and coldspot clusters. A cluster of hotspot zones is identified at the central/north-central part of the WDI and a relatively smaller cluster of coldspot zones is identified at the southeastern part of the WDI.

On average, hotspot zones have an AFR of 0.58 breaks/(km.year), which is ~1.53X the average AFR observed across all zones of the WDI (0.38 breaks/(km.year)), while those of the

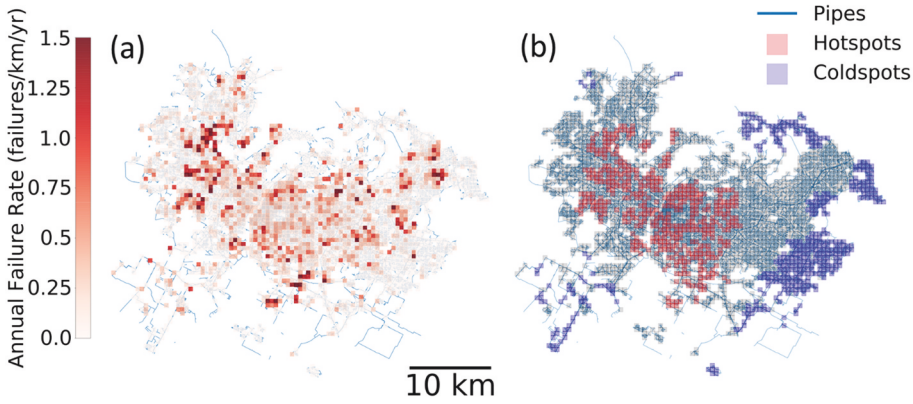

Figure 1. (a) Distribution of the AFR for a zone size of 0.25 km^2; and (b) Hotspot (red) and coldspot (blue) clusters of pipe failures identified by LISA. Gray zones represent areas that are neither statistically significant hotspot nor coldspot zones. Blue lines represent the pipes.

coldspot clusters have an average of 0.12 breaks/(km.year) (zero AFR zones excluded). It is important to note that such patterns cannot be easily detected by simply visualizing the AFR in each zone (Figure 1(a)). This shows the necessity for employing a spatial clustering approach, like the LISA implemented herein, to reveal the locations of the clusters, and more importantly, to test their statistical significance. Furthermore, the significant variability in the AFR observed across the zones of the different clusters suggests that different factors might be influencing pipe failure in each of the clusters. Hence, to accurately predict failure rates, different models need to be independently developed for each cluster in order to account for the best set of local predictors. This is particularly important for the zones of the hotspot cluster experiencing the highest failure rates across the WDI, which makes them natural candidates for proactive pipe RoR programs.

Predictive modeling of the failure rates

Non-spatial models

First, all eight models listed in Table 1 are trained to predict the AFR in all 1,400 zones using a set of seven predictors, specifically mean age (T_{avg}), mean diameter (D_{avg}), total length (L_{tot}) and the fraction of materials by pipe length in each zone, including cast iron CI (f_{CI}), ductile iron DI (f_{DI}), polyvinyl chloride PVC (f_{PVC}) and asbestos cement AC (f_{AC}). The entire dataset comprising the AFR and its predictors in all 1,400 zones (zero AFR zones excluded) is randomly split into 980 zones for training (70%) and 420 for validation (30%). A two-sample K-S test statistic of 0.05, with a p-value of 0.44, is obtained for the distributions of the AFR in the training and validation samples, which indicates that the distributions of both samples are similar. This can also be seen in Figure S2 in the SI, which shows that both samples exhibit similar cumulative distribution functions.

Using the training data, the hyper-parameters for all eight models were tuned by means of a 10-fold cross-validation test. The average MSE across the tested ranges for each of the hyper-parameters is depicted in Figure S3, where the hyper-parameter settings that yielded the minimum average MSE are selected for further analysis. Table 2 lists the two performance assessment metrics, namely MAE, and RMSE, for all eight models for both the training and validation runs. With the exception of the RF model, both the parametric-linear and non-parametric models gave comparable training results. All models, with the exception of RF, do not display a strong fit to the AFR data, with the non-parametric models showing a slightly better fit than the linear-parametric models as indicated by the slightly smaller values for the MSE and RMSE. On the other hand, RF is the best-performing model for the training run.

Table 2. Performance assessment of the network-wide non-spatial models.

		Training		Validation	
Model		MAE	RMSE	MAE	RMSE
Parametric-linear	OLS	0.628	0.892	0.666	0.899
	LASSO	0.628	0.892	0.666	0.899
	RD	0.628	0.892	0.666	0.899
	EN	0.628	0.892	0.666	0.899
Non-parametric	RF	0.492	0.678	0.642	0.906
	ANN	0.623	0.886	0.659	0.898
	SVR	0.568	0.834	0.632	0.894
	KNN	0.604	0.860	0.663	0.905

Despite the slightly better training performance shown by the non-parametric models, their performance significantly worsens when they are used to predict the out-of-sample AFRs. The RF model shows the biggest drop in performance between the training and validation runs, followed by the SVR model and the KNN model. On the other hand, linear-parametric models show a relatively more consistent performance across the training and validation runs, which indicates that they are less prone to overfitting compared to non-parametric models. Taken together, the results listed in Table 2 indicate that, despite their flexibility in fitting complex functions, complex data-driven models are not necessarily more accurate when predicting the out-of-sample AFR. On the other hand, parametric-linear models can still display a comparable predictive capability for out-of-sample data despite their intrinsic simplicity.

Spatial models

The distribution of pipe material and age in the studied WDI exhibits noticeable clustering, which can be seen from Figure 2 that depicts the pipe material of the highest fraction and the mean age of pipes in each zone. Figure 2 shows that the old CI pipes are primarily concentrated in the central part of the

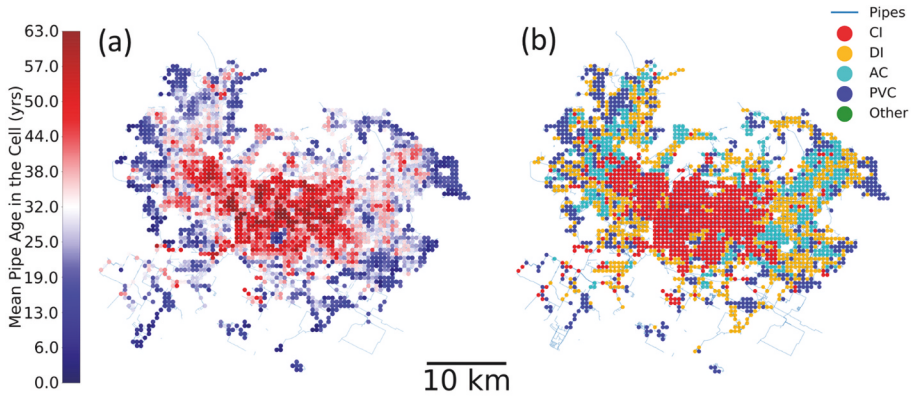

Figure 2. (a) Distribution of the mean pipe age in each zone; (b) Pipe material with the highest fraction in each zone.

network, while newer PVC pipes are more distributed across the boundaries of the WDI. Such spatial structure in the predictors needs to be accounted for in the predictive models to avoid biasing the estimation of their parameters and thus compromise their accuracy and predictive ability.

To examine the significance of incorporating the spatial structure of the data into the development of the prediction models, seven additional predictors representing the spatial lags of each of the original predictors were added to the seven original predictors mentioned in the previous section. Following the same procedure described in the previous section, all eight spatial models were trained using 70% of the data, tuned by means of a 10-fold cross-validation test, and then tested using the remaining 30%.

Table 3 lists the two performance assessment metrics for all eight spatial models. By comparing the results to those of the non-spatial models (Table 2), it can be seen that the inclusion of the SLPs slightly enhances the performance of all eight learning methods for both the training and validation runs. This can be seen from the consistent, albeit small, decrease in the MAE and RMSE before and after the inclusion of SLPs. Furthermore, it is important to note that the addition of the SLPs does not result in overfitting as evidenced by the consistent decrease in the errors for the validation runs in Table 3 compared to those in Table 2. Overall, the results listed in Table 3 indicate that controlling for the spatial structure in the AFR predictors improves the accuracy of predictive models. This becomes particularly important when these models are implemented for making pipe RoR decisions as displayed in later sections. The results also assert that parametric-linear models are generally less prone to overfitting than non-parametric models and that non-parametric show only a slightly better prediction accuracy for the out-of-sample data despite their superior performance of the non-parametric models with the in-sample data.

Cluster-specific models

Having included the spatial effects in the formulation of the predictive models, the outcomes of the clustering analysis we used to further improve the explanatory and predictive abilities of the prediction models. To this end, the regression dataset is divided into two subsets representing the zones that belong to hotspot and non-hotspot clusters, and then develop two independent sets of spatial models for each dataset separately. Table 4 lists the three performance assessment metrics for all eight cluster-specific models. By comparing the results to those of the network-wide spatial models (Table 3), it can be clearly seen that the incorporation of the clustering analysis outcomes in the development of prediction models enhances the performance of all eight learning methods for both the training and validation runs. On average, the training MAE and RMSE of the cluster-specific models are ~20% and ~16% lower than the network-wide spatial models, respectively. A similar enhancement is also observed for the validation run, where the cluster-specific models show an ~18% and ~12% reduction in the MAE and RMSE, respectively, compared to the network-wide spatial models. Taken together, the results in Table 4 indicate that the performance of the cluster-specific models is noticeably higher than that of the network-wide models, which can be attributed to the fact that cluster-specific models are more capable of capturing the proper influence of various predictors in the different sections of the WDI as explained in the following subsection.

Table 3. Performance assessment of the network-wide spatial models.

		Training		Validation	
Model		MAE	RMSE	MAE	RMSE
Parametric-linear	OLS	0.625	0.887	0.660	0.897
	LASSO	0.624	0.884	0.663	0.897
	RD	0.624	0.885	0.662	0.897
	EN	0.620	0.882	0.659	0.895
Non-parametric	RF	0.484	0.658	0.623	0.882
	ANN	0.610	0.869	0.650	0.894
	SVR	0.566	0.829	0.641	0.894
	KNN	0.587	0.852	0.643	0.893

Table 4. Performance assessment of the cluster-specific spatial models.

		Training		Validation	
Model		MAE	RMSE	MAE	RMSE
Parametric-linear	OLS	0.528	0.772	0.562	0.797
	LASSO	0.528	0.772	0.561	0.797
	RD	0.529	0.774	0.562	0.797
	EN	0.527	0.770	0.560	0.796
Non-parametric	RF	0.374	0.515	0.536	0.796
	ANN	0.512	0.758	0.545	0.798
	SVR	0.461	0.716	0.535	0.795
	KNN	0.516	0.763	0.549	0.808

Model Selection and Predictor Importance

In addition to knowing which model performs the best, it is crucial to understand the relative importance of the predictors on the model predictions. Since the data for the predictor variables is standardized, the magnitudes of the estimated coefficients by the parametric-linear models provide a direct means for interpreting the relative significance of each of the predictors. This is not the case for non-parametric models that generally require extrinsic tests to examine the relative importance of the predictors (e.g. by examining the decrease of accuracy in predictions on the out-of-bag samples when a given predictor is excluded (James et al. 2013)). Furthermore, both LASSO and EN models are equipped with the capability of predictor selection thanks to the *l*1 regularization component that modifies the estimation to achieve sparsity. As the value of λ_1 increases, the sparsity objective becomes more stringent and hence the estimated regression coefficients shrink until reaching zero at different λ_1 values, at which the corresponding predictors are deselected from the model. As previously described, the minimum set of predictors that best explain the AFR can be obtained by setting the *l*1 regularization parameter (λ_1) using a 10-fold cross-validation test.

To evaluate the overall performance of the examined forecasting approaches, prediction accuracy should not be adopted as the sole evaluation criteria. Instead, careful attention should also be given to the structural stability and complexity of the forecasting model (Boland, Baumann, and Dziegielewski 1981). The consistent drop in the performance of the non-parametric models for the out-of-sample data compared to the training data as observed in Tables 2–4, signals a concerning deficiency in the stability of these models compared to parametric-linear models. Furthermore, the slightly better in-sample performance exhibited by the non-parametric models does not make up for the loss of comprehensibility and credibility that accompanies their significant complexity. Hence, given its interpretability, stability and simplicity, the cluster-specific EN model is used for the remainder of the analysis in this study despite its slightly lower in-sample accuracy compared to the cluster-specific non-parametric models. Figure 3 depicts the observed versus modeled AFR for the cluster-specific-EN model fitted separately to the hotspot (HS) zones (red triangles) and non-hotspot (NHS) zones (blue circles) of the WDI.

Table 5. Regression coefficients estimated by the EN model.

	EN coefficient		
Predictor	All zones	Hotspots	Non-hotspots
f_{CI}	+0.09	+0.22	0.00
f_{PVC}	0.00	0.00	0.00
f_{DI}	0.00	0.00	0.00
f_{AC}	+0.33	+0.54	+0.28
L_{tot}	−0.17	−0.16	−0.18
T_{avg}	+0.15	+0.12	+0.09
D_{avg}	−0.12	−0.17	−0.06

Table 5 lists the regression coefficients determined by the network-wide spatial-EN model fitted to all the zones in the studied WDI together with the coefficients of the cluster-specific EN models fitted separately to the hotspot and non-hotspot zones. For the network-wide model, the fraction of AC pipes appears to have the largest positive coefficient among the covariates (+0.33), followed by the mean age of pipes (+0.15) and the fraction of CI pipes (+0.09). The total length of pipes has the most negative coefficient (−0.17) followed by the average diameter (−0.12). The spatial-EN estimation renders zero coefficients for two out of the seven predictors, namely the fraction of PVC pipes (f_{PVC}) and fraction of DI pipes (f_{DI}). One possible explanation for such outcome is collinearity among the predictors, which happens whenever two or more predictors are linearly dependent (Dormann et al. 2013). In the presence of strong collinearity between predictors, the *l*1 regularization parameter (λ_1) in the EN model leads to the selection of only a subset of the collinear predictors, typically the ones whose absolute coefficients are the largest, and discards the others. Multiple metrics can be used to check whether the set of exogenous variables exhibit some form of collinearity (Dormann et al. 2013). Herein, the pairwise Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRCC) are calculated for each pair of predictors, and the results are depicted in Figure S4 of the SI. The value of PCC and SRCC can range anywhere from +1 for perfect positive correlation to −1 for perfect negative correlation, while a zero value indicates

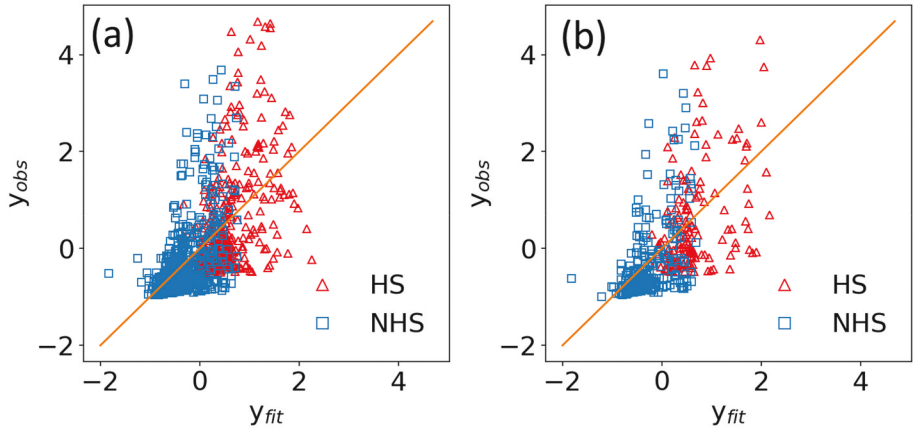

Figure 3. Observed versus predicted AFR (standardized) for the cluster-specific EN model fitted separately to the hotspot zones (red triangles) and non-hotspot zones (blue circles) for the (a) training and (b) validation datasets.

no correlation. As can be seen from Figure S4, T_{avg} exhibits strong negative correlation with f_{PVC} (PCC = −0.57, SRCC = −0.52), and with f_{DI} (PCC = −0.44, SRCC = −0.46). Furthermore, both f_{PVC} and f_{DI} exhibit negative correlation with f_{CI}. Nonetheless, it is important to note that this result is system-specific and is the mere outcome of the high degree of cross-correlation between pipe age and material in the tested WDI, which leads the spatial-EN algorithm to exclude f_{PVC} and f_{DI} from the model. While this is expected since CI and AC pipes have traditionally been used in WDI before the introduction of DI and PVC pipes, this might not necessarily be the case in other WDIs.

The enhanced performance of the cluster-specific models compared to the network-wide spatial models can be attributed to the fact that the contribution of various predictors varies between different zones of the WDI as seen from the estimated EN coefficients listed in Table 5. For the HS model, the highest two coefficients correspond to the fractions of AC and CI pipes, which implies that these two variables affect pipe failure in HS zones the most. On the other hand, zero coefficients are rendered for the fraction of CI pipes in the NHS model, which implies that replacing CI pipes in the non-hotspot zones will not have as a significant effect as would replace the same amount of CI pipes in HS zones. The following section further demonstrates how water utilities can leverage such insights in designing proactive pipe RoR programs to effectively reduce pipe failure rates in the WDI.

Impact on Infrastructure Asset Management

The proposed unified framework can help water utilities better manage their assets by targeting pipe RoR decisions. By identifying HS zones with elevated failure rates, the utility can direct its limited resources into inspecting the pipes in these zones. Moreover, the results of the prediction models indicate the relative importance of the different factors driving pipe failures in different zones. To further demonstrate this, three hypothetical scenarios for pipe replacement are simulated. In the first scenario, the utility is assumed to preserve the status-quo of making replacement decisions in a reactive manner by replacing pipes that experience failure events. In the second scenario, the utility is assumed to have an understanding of the spatial clustering exhibited by pipe failures and to also possess a prediction model for the AFR in the different clusters, i.e. HS and NHS. Based on this knowledge, the utility pursues a preventive pipe replacement program by proactively replacing the pipes in the areas experiencing high failure rates. Furthermore, since the fraction of AC and CI pipes has the highest variable importance in the cluster-specific EN model for the HS model (Table 5), the utility uses this knowledge to replace only AC and CI pipes within the hotspot zones. To elucidate that the impact of replacing the pipes within the HS zones is not merely an outcome of the fact that old AC and CI pipes are being replaced but also the location of these pipes, a third scenario in which the same length of AC and CI pipes is replaced in the NHS zones is considered.

In the first scenario (reactive strategy), the utility hypothetically replaces 610 km of pipes that have incurred at least one failure event during the study period of 36 months. In the second scenario (proactive strategy), the utility replaces an equivalent 610 km of AC and CI pipes exclusively from the HS zones. In the third scenario, the utility replaces an equivalent pipe length of AC and CI pipes exclusively from the NHS zones. In all scenarios, the pipes are replaced with new PVC pipes of the same length. To assess the impact of all scenarios on reducing the AFR, the new pipe characteristics are recalculated for all the zones, and the cluster-specific EN models fitted previously to the HS and NHS zone models are used to predict the new AFR for each scenario.

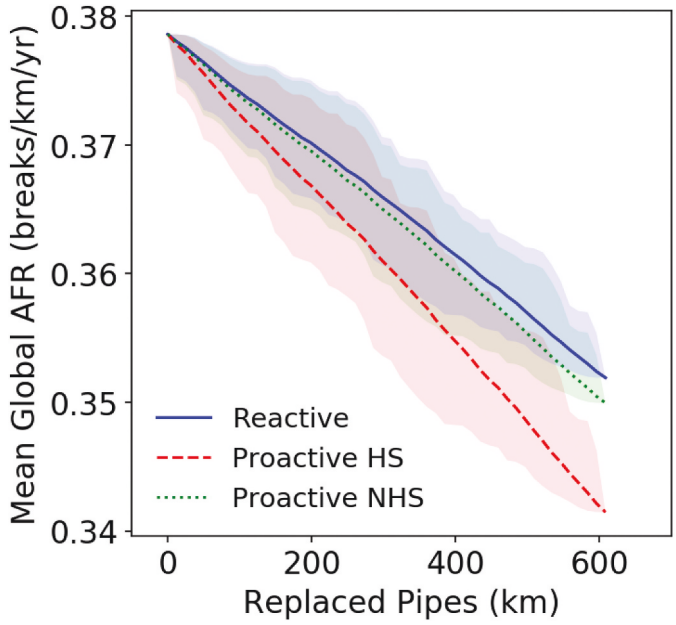

Figure 4. Reduction in the mean global AFR attained by targeted replacement of AC and CI pipes in the hotspot zones (red-dashed) and non-hotspot zones (green-dotted), and reactive pipe replacement (blue-solid) in the entire WDI; shaded envelopes represent the min-max range of 100 different realizations.

Figure 4 depicts the gradual reduction in the AFR attained by the three pipe replacement strategies, where line plots and shaded envelopes represent the mean and the min-max range of 100 different realizations for replacing the pipes in a different order, respectively. By replacing only the pipes that have incurred previous failures (blue line), the mean AFR for the entire network drops to 0.354 breaks/(km.year) (−7%), while the proactive pipe replacement approach (red line) would reduce the mean global AFR to 0.340 breaks/(km.year) (−10%). The third scenario (green line) shows almost no improvement for proactively replacing AC and CI pipes in the non-hotspot zones, which asserts the fact that the spatial location of the replaced pipes plays as significant role as their age and material, and hence emphasizes the importance of spatial clustering as an integral component of the presented approach.

Data Limitations and Recommendations for Future Development

Although the failure dataset implemented herein is of a short time-horizon (3 years), the proposed approach was still capable

of providing useful information that can potentially aid in asset management operations. Nevertheless, with larger failure datasets collected over longer time-periods, the presented approach can be enhanced as follows:

(1) The spatial resolution can be enhanced by developing failure prediction models for individual pipes instead of zones or regions if failure records are available for longer time-periods. For instance, failure data collected over multiple decades would typically comprise numerous pipes that incurred more than one failure event and can hence be used in developing survival models (e.g. PHMs) at the pipe-level. Following the approach presented herein, the outcomes of the clustering analysis can be leveraged in developing such pipe-level models by constructing separate models for pipes in HS zones and NHS zones. Furthermore, spatially lagged predictors can also be included in these models, which can potentially enhance their performance as demonstrated herein.
(2) While the out-of-sample performance of the non-parametric models tested herein was noticeably worse than their in-sample performance, this can be primarily attributed to the small size of the failure dataset used to train these data-driven models. Furthermore, the performance of non-parametric models can be potentially enhanced by including more covariates (e.g. soil conditions, hydraulic parameters, traffic loadings and land use), and thus their strong capability to model complex non-linear relationships can be better harnessed.
(3) By using a dataset of failure records collected over a longer time-period, the temporal dimension can be introduced to the failure prediction models to account for the dynamic changes in the failure rates. Such temporal variations in the failure rate may stem from climatic changes in weather conditions and other time-dependent covariates. Furthermore, longer periods of data collection would enable the use of failure rates with higher temporal resolution. For instance, instead of aggregating failures at the yearly level as was done in this study, failure rates can be aggregated at the monthly/quarterly level. This would enable including important seasonal variables, such as temperature and precipitation, that are known to influence pipe failure (Almheiri, Meguid, and Zayed 2020).

Conclusions

Water main failures in dense urban areas pose significant economic, social and environmental consequences as well as severe implications for public health. This study proposes an integrated computational framework that integrates spatial clustering analysis with predictive pipe failure modeling. The proposed approach integrates two main components: First, spatial autocorrelation analysis, based on the local index of Moran's *I*, is implemented for identifying statistically significant hotspot and coldspot clusters of pipe failures. Second, statistical learning methods are developed and tested for the prediction of pipe failure rates within the clusters based on the local characteristics of the infrastructure, while simultaneously accounting for the spatial patterns exhibited by these characteristics. Finally, the integrated approach is used for comparing different pipe replacement strategies in order to improve the efficacy of asset management decisions.

The framework is demonstrated on a short-term (36 months) dataset of pipe failures retrieved from the maintenance records of a real-life, full-scale metropolitan water utility in the United States. For the studied infrastructure, pipe failures were found to be significantly clustered, and the locations of pipe failure hotspot and coldspot clusters were successfully revealed. A strong degree of clustering was also exhibited by the characteristics of the pipe infrastructure, specifically pipe age and material. Key insights revealed by this study are (1) failing to account for the spatial patterns in the failure predictors reduces the accuracy of the non-spatial predictive models compared to their spatial counterparts, (2) the explanatory and predictive abilities of the spatial models further improved when the outcomes of the clustering analysis were leveraged to tailor these models to appropriately account for the local predictors in each cluster and (3) both linear-parametric and data-driven non-parametric models showed similar prediction accuracy for the out-of-sample data despite the better training performance exhibited by the more complex non-parametric models.

Although the proposed approach was demonstrated on a short-term failure dataset, it was still capable of providing useful information that can aid in guiding pipe rehabilitation decisions. As confidence in decision-making highly depends on the accuracy of failure prediction models, this work emphasizes the importance of reporting, collection and storage of pipe failure records for enhancing the efficacy of asset management operations. Future work should aim to apply the proposed approach to a larger pipe failure dataset to enhance the spatiotemporal resolution of pipe failure predictions. Furthermore, the incorporation of additional covariates, particularly time-dependent and dynamic variables, may enhance the predictive accuracy of the proposed approach.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the University of Texas at Austin Startup Grant and by Cooperative Agreement No. 83595001 awarded by the U.S. Environmental Protection Agency to The University of Texas at Austin. This work has not been formally reviewed by EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication. Partial funding by the National Science Foundation (NSF) awards No. 2015603 and 2015658 is gratefully acknowledged.

Data availability

Some or all data, models, or code generated or used during the study are proprietary or confidential in nature and may only be provided with restrictions (e.g. anonymized data). Pipe failure data and network GIS records were provided by the water utility under a confidentiality agreement between the water utility and the second author. The data are available by request from the authors and may only be provided after

obtaining the utility's approval and undergoing potential anonymization. The computational framework developed in this study was implemented using the Python programming language. The spatial autocorrelation analysis is conducted using the Python Spatial Analysis Library (PySAL) (Rey and Anselin 2010). Predictive modeling is conducted by scikit-learn (Pedregosa et al. 2011). Other Python libraries used for data analysis and visualization include numpy, pandas, geopandas and matplotlib (Hunter 2007; McKinney 2010; Van Der Walt, Colbert, and Varoquaux 2011).

References

Abokifa, A.A.A., and L. Sela. 2019. "Identification of Spatial Patterns in Water Distribution Pipe Failure Data Using Spatial Autocorrelation Analysis." *Journal of Water Resources Planning and Management* 145 (12): 04019057. doi:10.1061/(ASCE)WR.1943-5452.0001135.

Alizadeh, Z., J. Yazdi, S. Mohammadiun, K. Hewage, and R. Sadiq. 2019. "Evaluation of Data Driven Models for Pipe Burst Prediction in Urban Water Distribution Systems." *Urban Water Journal* 16 (2): 136–145. doi:10.1080/1573062X.2019.1637004.

Almheiri, Z., M. Meguid, and T. Zayed. 2020. "An Approach to Predict the Failure of Water Mains under Climatic Variations." *International Journal of Geosynthetics and Ground Engineering* 6 (4): 1–16.

Almheiri, Z., M. Meguid, and T. Zayed. 2021. "Failure Modeling of Water Distribution Pipelines Using meta-learning Algorithms." *Water Research* 205: 117680. doi:10.1016/j.watres.2021.117680.

Altman, N.S. 1992. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression." *American Statistician* 46 (3): 175–185.

American Society of Civil Engineers. 2017. *Infrastructure Report Card: Drinking Water*. Reston, VA: American Society of Civil Engineers (ASCE) Publisher.

American Water Works Association. 2012. *Buried No Longer: Confronting America's Water Infrastructure Challenge*. Denver, CO: American Water Works Association (AWWA) Publisher.

Anselin, L. 1995. "Local Indicators of Spatial Association—LISA." *Geographical Analysis* 27 (2): 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x.

Anselin, L. 2002. "Under the Hood Issues in the Specification and Interpretation of Spatial Regression Models." *Agricultural Economics* 27 (3): 247–267. doi:10.1111/j.1574-0862.2002.tb00120.x.

Aslani, B., S. Mohebbi, and H. Axthelm. 2021. "Predictive Analytics for Water Main Breaks Using Spatiotemporal Data". *Urban Water Journal*. 18 (6): 433–448. doi:10.1080/1573062X.2021.1893363.

Barton, N.A., S.H. Hallett, and S.R. Jude. 2022. "The Challenges of Predicting Pipe Failures in Clean Water Networks: A View from Current Practice." *Water Supply* 22 (1): 527–541. doi:10.2166/ws.2021.255.

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 57 (1): 289–300.

Berardi, L., O. Giustolisi, Z. Kapelan, and D.A. Savic. 2008. "Development of Pipe Deterioration Models for Water Distribution Systems Using EPR." *Journal of Hydroinformatics* 10 (2): 113. doi:10.2166/hydro.2008.012.

Boland, J., D.D. Baumann, and B. Dziegielewski. 1981. *An Assessment of Municipal and Industrial Water Use Forecasting Approaches*. Carbondale IL: Planning And Management Consultants Ltd.

Chen, T.Y.J., and S.D. Guikema. 2020. "Prediction of Water Main Failures with the Spatial Clustering of Breaks." *Reliability Engineering and System Safety*. 203: 107108.

Chi, G., and J. Zhu. 2008. "Spatial Regression Models for Demographic Analysis." *Population Research and Policy Review* 27 (1): 17–42. doi:10.1007/s11113-007-9051-8.

Christodoulou, S., A. Gagatsis, A. Agathokleous, S. Xanthos, and S. Kranioti. 2012. "Urban Water Distribution Network Asset Management Using Spatio-Temporal Analysis of Pipe-Failure Data." *14th International Conference on Computing in Civil and Building Engineering* 27: 29.

de Oliveira, D.P., J.H. Garrett, and L. Soibelman. 2011a. "A density-based Spatial Clustering Approach for Defining Local Indicators of Drinking Water Distribution Pipe Breakage." *Advanced Engineering Informatics* 25 (2): 380–389. doi:10.1016/j.aei.2010.09.001.

de Oliveira, D.P., D.B. Neill, J.H. Garrett, and L. Soibelman. 2011b. "Detection of Patterns in Water Distribution Pipe Breakage Using Spatial Scan Statistics for Point Events in a Physical Network." *Journal of Computing in Civil Engineering* 25 (1): 21–30. doi:10.1061/(ASCE)CP.1943-5487.0000079.

Dormann, C.F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, Jaime R. García. Marquéz, et al. 2013. "Collinearity: A Review of Methods to Deal with It and A Simulation Study Evaluating Their Performance." *Ecography (Cop.)* 36 (1): 27–46. doi:10.1111/j.1600-0587.2012.07348.x.

Ercumen, A., J.S. Gruber Jr, and J M. Colford. 2014. "Water Distribution System Deficiencies and Gastrointestinal Illness: A Systematic Review and Meta-Analysis." *Environmental Health Perspectives* 122 (7): 651–661. doi:10.1289/ehp.1306912.

Fan, Xudong, Xiaowei Wang, Xijin Zhang, and Xiong Bill Yu. 2022. "Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors". *Reliability Engineering & System Safety*. 219: doi:10.1016/j.ress.2021.108185.

Fan, Xudong, Xijin Zhang, and Xiong Yu. 2022. "A graph convolution network–deep reinforcement learning model for resilient water distribution network repair decisions". *Computer–Aided Civil and Infrastructure Engineering*. 37 (12): 1547–1565.

Florax, R., and H. Folmer. 1992. "Specification and Estimation of Spatial Linear Regression Models: Monte Carlo Evaluation of pre-test Estimators." *Regional Science and Urban Economics* 22 (3): 405–432. doi:10.1016/0166-0462(92)90037-2.

Folkman, S. 2018. *Water Main Break Rates in the USA and Canada: A Comprehensive Study*. Logan, Utah: Utah State Univ.

Hering, J.G., T.D. Waite, R.G. Luthy, J.E. Drewes, and D.L. Sedlak. 2013. "A Changing Framework for Urban Water Systems." *Environmental Science & Technology* 47 (19): 10721–10726. doi:10.1021/es4007096.

Hoerl, A.E., and R.W. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12 (1): 55–67. doi:10.1080/00401706.1970.10488634.

Hunter, J.D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90. doi:10.1109/MCSE.2007.55.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning*. New York: Springer.

Jenkins, L., S. Gokhale, and M. McDonald. 2015. "Comparison of Pipeline Failure Prediction Models for Water Distribution Networks with Uncertain and Limited Data." *Journal of Pipeline Systems Engineering and Practice* 6 (2): 04014012. doi:10.1061/(ASCE)PS.1949-1204.0000181.

Kabir, G., S. Tesfamariam, J. Loeppky, and R. Sadiq. 2015. "Integrating Bayesian Linear Regression with Ordered Weighted Averaging: Uncertainty Analysis for Predicting Water Main Failures." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 1 (3): 04015007. doi:10.1061/AJRUA6.0000820.

Kakoudakis, K., K. Behzadian, R. Farmani, and D. Butler. 2017. "Pipeline Failure Prediction in Water Distribution Networks Using Evolutionary Polynomial Regression Combined with K -means Clustering." *Urban Water Journal* 14 (7): 737–742. doi:10.1080/1573062X.2016.1253755.

Kakoudakis, K., R. Farmani, and D. Butler. 2018. "Pipeline Failure Prediction in Water Distribution Networks Using Weather Conditions as Explanatory Factors." *Journal of Hydroinformatics* 20 (5): 1191–1200. doi:10.2166/hydro.2018.152.

Kimutai, E., G. Betrie, R. Brander, R. Sadiq, and S. Tesfamariam. 2015. "Comparison of Statistical Models for Predicting Pipe Failures: Illustrative Example with the City of Calgary Water Main Failure." *Journal of Pipeline Systems Engineering and Practice* 6 (4): 04015005. doi:10.1061/(ASCE)PS.1949-1204.0000196.

Kleiner, Y., and B. Rajani. 2001. "Comprehensive Review of Structural Deterioration of Water Mains: Statistical Models." *Urban Water* 3 (3): 131–150. doi:10.1016/S1462-0758(01)00033-4.

Konstantinou, C., and I. Stoianov. 2020. "A Comparative Study of Statistical and Machine Learning Methods to Infer Causes of Pipe Breaks in Water Supply Networks." *Urban Water Journal* 17 (6): 534–548. doi:10.1080/1573062X.2020.1800758.

Laucelli, D., B. Rajani, Y. Kleiner, and O. Giustolisi. 2014. "Study on Relationships between climate-related Covariates and Pipe Bursts Using evolutionary-based Modelling." *Journal of Hydroinformatics* 16 (4): 743–757. doi:10.2166/hydro.2013.082.

Legendre, P. 1993. "Spatial Autocorrelation: Trouble or New Paradigm?" *Ecology* 74 (6): 1659–1673. doi:10.2307/1939924.

Liaw, A., and M. Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22.

McKinney, W. 2010. "Data Structures for Statistical Computing in Python." Proceedings of the 9th Python in Science Conference. Austin, TX, pp. 51–56.

Nishiyama, M., and Y. Filion. 2013. "Review of Statistical Water Main Break Prediction Models." *Canadian Journal of Civil Engineering* 40 (10): 972–979. doi:10.1139/cjce-2012-0424.

Obringer, R., and R. Nateghi. 2018. "Predicting Urban Reservoir Levels Using Statistical Learning Techniques." *Scientific Reports* 8 (1): 1–9. doi:10.1038/s41598-018-23509-w.

Oliveira, D., J.H. Garrett, and L. Soibelman. 2009. "Spatial Clustering Analysis of Water Main Break Events." In *Computing in Civil Engineering*. 338–347. Reston, VA: American Society of Civil Engineers (ASCE).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (Oct): 2825–2830.

Rajani, B., and Y. Kleiner. 2001. "Comprehensive Review of Structural Deterioration of Water Mains: Physically Based Models." *Urban Water* 3 (3): 151–164. doi:10.1016/S1462-0758(01)00032-2.

Rey, S.J., and L. Anselin. 2010. *PySAL: A Python Library of Spatial Analytical Methods, In: Handbook of Applied Spatial Analysis*, 175–193. Berlin/Heidelberg, Germany: Springer.

Rifaai, T.M., A.A. Abokifa, and L. Sela. 2022. "Integrated Approach for Pipe Failure Prediction and Condition Scoring in Water Infrastructure Systems." *Reliability Engineering & System Safety* 220: 108271. doi:10.1016/j.ress.2021.108271.

Scheidegger, A., J.P. Leitão, and L. Scholten. 2015. "Statistical Failure Models for Water Distribution Pipes - A Review from A Unified Perspective." *Water Research* 83: 237–247. doi:10.1016/j.watres.2015.06.027.

Shin, H., K. Kobayashi, J. Koo, and M. Do. 2016. "Estimating Burst Probability of Water Pipelines with a Competing Hazard Model." *Journal of Hydroinformatics* 18 (1): 126–135. doi:10.2166/hydro.2015.016.

Shirzad, A., and M.J.S. Safari. 2020. "Pipe Failure Rate Prediction in Water Distribution Networks Using Multivariate Adaptive Regression Splines and Random Forest Techniques." *Urban Water Journal* 16 (9): 653–661. doi:10.1080/1573062X.2020.1713384.

Smola, A.J., and B. Schölkopf. 2004. "A Tutorial on Support Vector Regression." *Statistics and Computing* 14 (3): 199–222. doi:10.1023/B:STCO.0000035301.49549.88.

Specht, D.F. 1991. "A General Regression Neural Network." *IEEE Transactions on Neural Networks* 2 (6): 568–576. doi:10.1109/72.97934.

St. Clair, A.M., and S. Sinha. 2012. "State-of-the-technology Review on Water Pipe Condition, Deterioration and Failure Rate Prediction Models!" *Urban Water Journal* 9 (2): 85–112. doi:10.1080/1573062X.2011.644566.

Stone, S.L., E.J. Dzuray, D. Meisegeier, A.S. Dahlborg, M. Erickson, and A. N. Tafuri. 2002. "Decision-support Tools for Predicting the Performance of Water Distribution and Wastewater Collection Systems." *U.S. Environmental Protection Agency, Office of Research and Development*.

Tabesh, M., J. Soltani, R. Farmani, and D. Savic. 2009. "Assessing Pipe Failure Rate and Mechanical Reliability of Water Distribution Networks Using data-driven Modeling." *Journal of Hydroinformatics* 11 (1): 1–17. doi:10.2166/hydro.2009.008.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 58 (1): 267–288.

Van Der Walt, S., S.C. Colbert, and G. Varoquaux. 2011. "The NumPy Array: A Structure for Efficient Numerical Computation." *Computing in Science & Engineering* 13 (2): 22. doi:10.1109/MCSE.2011.37.

Wang, Y., T. Zayed, and O. Moselhi. 2009. "Prediction Models for Annual Break Rates of Water Mains." *Journal of Performance of Constructed Facilities* 23 (1): 47–54. doi:10.1061/(ASCE)0887-3828(2009)23:1(47).

Wilson, D., Y. Filion, and I. Moore. 2015. "State-of-the-art Review of Water Pipe Failure Prediction Models and Applicability to large-diameter Mains." *Urban Water Journal* 14 (2): 173–184. doi:10.1080/1573062X.2015.1080848.

Xu, Q., Q. Chen, W. Li, and J. Ma. 2011. "Pipe Break Prediction Based on Evolutionary data-driven Methods with Brief Recorded Data." *Reliability Engineering & System Safety* 96 (8): 942–948. doi:10.1016/j.ress.2011.03.010.

Yamijala, S., S.D. Guikema, and K. Brumbelow. 2009. "Statistical Models for the Analysis of Water Distribution System Pipe Break Data." *Reliability Engineering & System Safety* 94 (2): 282–293. doi:10.1016/j.ress.2008.03.011.

Zou, H., and T. Hastie. 2005. "Regression and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–320. doi:10.1111/j.1467-9868.2005.00503.x.