ARTICLE TEMPLATE

Predictive Analytics for Water Main Breaks Using Spatiotemporal Data

Babak Aslani[a], Shima Mohebbi[a] and Hana Axthelm[b]

[a]Department of Systems Engineering and Operations Research, George Mason University, Fairfax, VA, USA; [b]Data Sceince and Analytics Institute, University of Oklahoma, Norman, OK, USA

ARTICLE HISTORY
Compiled January 9, 2021

ABSTRACT
Water main breaks are a common recurring problem in water distribution networks, resulting in cascading effects in the whole system and the interconnected infrastructures such as transportation. Having integrated the physical features of pipes such as diameter and environmental factors like precipitation, we propose predictive models based on spatiotemporal data and machine learning methods. In this study, the dataset is the main breaks recorded from 2015 to 2020 in the city of Tampa, Florida. First, a spatial clustering is conducted to identify vulnerable areas to breaks. A time series analysis is also carried out for the temporal data. The result of these analyses informed the machine learning algorithms as independent variables. We then compared the predictive models based on information-based and rank-based criteria. Obtained results indicated that Boosted Regression Tree (BRT) model was superior to the others. Finally, we present predicted normalized failure rates for the water distribution network to inform rehabilitation and fortification decisions at the municipality level.

KEYWORDS
Water infrastructure, Machine learning, Spatial clustering, Time series analysis.

1. Introduction

Water distribution networks (WDNs) are among the most essential and expensive municipal infrastructure assets since modern societies are much dependent on them for their regular and routine activities. Any disruptions in this system can affect the water distribution network as well as other existing nearby infrastructures such as sewer, stormwater, transportation, and gas pipes that may lead to catastrophic failures (Kabir et al. 2015).

Water main break is a major concern for every water utility as they disrupt customer service, result in water and revenue loss, and create the potential for contaminants to enter the water distribution system. The total cost of water loss due to water main or pipe breaks is estimated to be 3.8 USD billion per year in North America (Snider and McBean 2020). Moreover, this value increases dramatically when including indirect costs, such as interruption to service, and health impacts (Renzetti, D.Dupont, and D.P.Dupon 2013).

CONTACT Shima Mohebbi Email: smohebbi@gmu.edu

Infrastructure physical features such as pipe diameter, age, length, and material are one contributing factor to water main breaks. Environmental factors like soil factors, precipitation, and seasonal climate variations also play a key role in the occurrence of failures in the water network. Operational features such as hydraulic pressure and water velocity are other critical players in the WDNs affecting the functionality of pipes. However, only some of these factors, such as pipe age, diameter, and temperature, are measurable and available for the establishment of predictive models (Kabir et al. (2015);Shirzad and Safari (2019)).

Water main breaks can have multifaceted consequences. First, they can disturb the redundancy/vulnerability of the network. Second, they can impose economic pressure in terms of water loss, rehabilitation cost, and the cost of damage caused by water main failure. Finally, the main breaks can directly have an impact on public safety and security (Phan et al. 2019). Focusing on the economic impact, there have been more than 2 million breaks in Canada and the United States since January 2000, with an average of 700 water main breaks every day, costing more than CAD 10 billions/year (Kabir et al. 2015). A main break costs $42,000 on average based on a survey by the Water Research Foundation (Chen et al. 2019).

However, the impacts of water main breaks are not confined to economic and social parts. These events may have harmful effects on public health due to a deterioration of water quality (Martinez-Codina et al. 2016). In fact, the potable water system has been identified as a significant factor in waterborne disease outbreaks. The low and negative pressure resulting from water breaks potentially allows contamination of drinking water from adjacent soils (Shortridge and Guikema 2019).

The water infrastructure in North America is old and deteriorating. Therefore, water mains breaks are creating floods and service disruptions daily. The rates of water main break soared by 27% from 11.0 to 14.0 breaks/ (100 miles)/year Between 2012 and 2018. As a concerning fact, the break rates of cast iron and asbestos cement pipe, composing 41% of the installed water mains in the US and Canada, have increased by more than 40% over six years (Folkman 2018).

Risk assessment for maintenance prioritization of pipes and other components of the water distribution network has gained increasing attention from municipalities and other decision-makers toward more effective management of water main breaks. In this approach, critical points in the WDN are identified through the assessment of risks based on the likelihood of failure events (Phan et al. 2019). Prediction models can help utilities reduce future breaks by identifying which pipes are most likely to break, and when. Utilities can use these predictions to develop more effective asset management plans and replace pipes before major breaks occur (Snider and McBean 2020).

In addition to the likelihood of failure, a broad range of situations, characterized by uncertainties and emergence, can be incorporated in order to have a holistic picture of risk associated with pipe failure. Those situations require different approaches to capture the comprehensive nature of risk in this context (Aven 2016). For example, pipes are subject to two types of deterioration: (1) structural deterioration, which diminishes the pipe's structural resilience and the ability to bear external stresses, and (2) deterioration of internal surfaces, which results in diminished hydraulic capacity, degradation of water quality, and reduced structural resilience in cases of severe internal corrosion. Both types of deterioration harm the reliability of the water distribution network (Kleiner and Rajani 2001). While including these aspects improve the rehabilitation and fortification decisions regarding water mains, they are not in the scope of the current study.

Predicting the future failures based on analyzing historical data of water main breaks is a useful tool for the fortification of the vulnerable components to postpone (and even eliminate) the possible propagating failures (Barton et al. 2019). Machine-learning algorithms have been adopted as effective methods in a range of applications to develop accurate models that are able to predict results, one of which is the prediction of pipe failures in a water distribution system. By employing these algorithms, the goal is to identify the time of next break for a pipe and to mitigate the ramifications of this disruption (Snider and McBean 2020).

In this study, we developed several predictive models based on the spatiotemporal data for the following purposes: (a) we aim to understand, investigate, and determine which features provide the most contribution to the model by quantifying variable importance, (b) to develop data-driven models to find how the most critical elements influence the magnitude of water main breaks, and (c) to develop an accurate prediction model for predicting main water breakage. The main contribution of the current study is to present an integrated framework for prediction of water main breaks. We adopted the set of variables in our model based on studies incorporating the physical variables of the water infrastructure network and the environmental factors like (see Yamijala, Guikema, and Brumbelow (2009)) to develop a reliable and comprehensive predictive model. While considering spatial clustering is recently addressed in the literature (see Chen and Guikema (2020)), our contribution is integrating time series analyses to extract the underlying failure patterns so that the accuracy of the predictive model is enriched. Informing the machine learning models from the spatial clustering (hotspot analysis) and the time series analysis using the concept of data fusion is another feature of the proposed framework. Fig. 1 shows the details of the proposed data-driven framework in the paper.

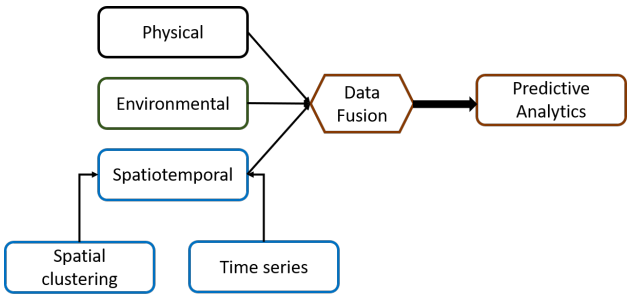


Figure 1. The proposed data-driven framework.

The remainder of the paper is structured as follows: Section 2 includes the relevant literature and the identified gaps. Section 4 explains the case study and the data used in the study. Section 3 presents a detailed explanation of the predictive analytics developed in the paper. Section 5 contains the results of the implementation for the case study in details. Finally, the discussion and future directions are provided in section 6.

2. Literature Review

There is a rich body of literature for pipe break prediction, which can be categorized into three main streams: (a) Physical models, (b) Purely statistical models, and (c) Data Mining or Machine Learning-based models.

2.1. *Physical Models*

Physical models attempt to describe the mechanisms causing breaks by analyzing the loads that a pipe is subjected to and the capacity of the pipe to handle these loads. Various physical models are developed in the literature (See Rajani and Kleiner (2001) for a comprehensive review). The main advantage of physical models is that they do not require large amounts of historical data to develop.

This approach is focused on developing mathematical models based on the fundamental physics related to pipe breaks to provide insights about breaks. While these models are fundamentally compelling, the data requirements for physical models often require in-field inspections/surveys and are time-consuming and expensive(Francis, Guikema, and Henneman 2014). Furthermore, due to structural differences among various systems, extrapolating the results to other mains is very difficult and inaccurate. For these reasons, physical models are often only applicable for large transmission mains or critical infrastructure (Wilson, Filion, and Moore 2017).

2.2. *Statistical Inferential Models*

Statistical inferential models implement statistical techniques to historical break data to detect patterns and make inferences about the pipe breaks. Compared with physical models, statistical models are less expensive and less time-consuming. These models aim to evaluate the existing data for formulating trends based on statistical measures such as correlation and covariance. The goal of inferential models is to improve understanding and not to make accurate predictions.

In the past, most studies on pipeline breakage models focused only on static factors such as pipe material type, diameter, and soil type, which could lead to biased results. Kleiner and Rajani (2000) realized the importance of time-dependent factors such as the age of the pipe, water temperature, and soil temperature. They first grouped the water mains into different partitions that were uniform and homogeneous regarding their response to deterioration and stress-inducing mechanisms. They then applied a generalized multi-variate exponential model and a multi-variate power model where the model input was a vector of time-dependent covariates of environmental and/or operational factors. The main limitation found with this approach was that only time-dependent factors were considered. While other studies have also recognized the benefits of time-dependent factors, they have not entirely removed the static factors.

Vanrenterghem-Raven (2007) aimed to develop a proportional hazards model for a water distribution piping network located in a complex urban area to identify key risk factors in the failure of pipes. They considered both inherent risk factors such as pipe material, diameter, length, date of installation, break history, and environmental factors like traffic, water zone, proximity to subways, highways in their model. They tested the proposed framework on a case study of Long Island City, with 220 miles of pipes and 20 years of break data from 1982 to 2002. They also examined the applicability of the model in the stratification of the data based on material, break frequency, and history.

Wang et al. (2010) suggested a new approach based on Bayesian configuration against pipe condition to find factor weights. They considered a range of factors such as size, age, inner coating, outer coating, soil condition, bedding condition, trench depth, electrical recharge, the number of road lanes, material, and operational pressure. They concluded that the factors with smaller weight values or with weights having relatively stable posterior means and narrow uncertainty bounds have less influence on pipe con-

ditions. The model was the most sensitive to variations of pipe age.

Duchesne, Toumbou, and Villeneuve (2016) compared the techniques of linear regression, Weibull-Exponential-Exponential (WEE), and Weibull-Exponential-Exponential-Exponential (WEEE) to study the number of breaks in a water main network. These models were calibrated using least squares and maximum likelihood methods. They only considered pipe age as an explanatory variable in their models, no other covariates such as pipe diameter or material were included.

Demissie, Tesfamariam, and Sadiq (2017) proposed a Dynamic Bayesian Network (DBN) considering both static and time-dependent factors for the problem. A DBN is an extension of a Bayesian Belief Network (BBN) but can apply to time-dependent factors as well as static ones. They concluded that a BBN by itself would not suffice and more model flexibility was required.

Statistical models can be more flexible than physical models as they can be applied with various types of input data. However,Yamijala, Guikema, and Brumbelow (2009) compared the predictive accuracy of several statistical regression models in the literature for estimating the probability or number of pipe breaks and/or leaks on individual pipe segments. The results showed that future research should be focused on improving the accuracy of predictive models for pipe break.

2.3. *Machine Learning Approaches*

Data mining and machine learning techniques are becoming more popular and useful in solving real-world problems such as predicting pipeline breakage. Adopting machine-learning algorithms helps to overcome the proportional hazard and linear covariate relationship assumptions common with many statistical models. The supervised machine-learning models can identify and model the complex relations between pipe predictor variables and pipe breakage.

Wood and Lence (2009) developed an approach for identifying key data asset management, predicting pipe breaks, and selecting appropriate models. The method was applied to the District of Maple Ridge, B.C., Canada, to identify the current and future magnitude of a utilitys pipe burst. The goal of their study was to enhance the development of pipe replacement priorities based on the predicted breaks. They also identified critical data to collect in future data acquisition programs.

Tabesh et al. (2009) presented two data-driven modeling techniques (Artificial Neural Network (ANN) and neuro-fuzzy systems) to have a more comprehensive and more accurate predictive model for pipe failure rate and to have an improved assessment of the reliability of pipes. They considered parameters like pressure and pipe depth, as well as the common factors such as diameter and length of pipes. The proposed models were applied to a real case in Iran. The result showed that the ANN model was more realistic and accurate in the prediction of pipe failure rates.

Jafar, Shahrour, and Juran (2010) employed ANN for estimation of the failure rate and the optimal replacement time for the individual pipes. They used a 14-year data set of a water distribution system in a northern city of France. They used six ANN-based models for the prediction of water mains failure and the determination of the benefit index to optimize the investment for the rehabilitation and maintenance of urban water mains.

Wang et al. (2013) attempted to solve a bipartite ranking problem to determine which pipes have the highest risks of breakage using static and time-dependent factors. They compared 5 data mining algorithms, using the area under the curve score as a

comparison method. They examined RankBoost.B, ANN, Cox, Nave Bayes Classifier, and Logistic Regression. They found that RankBoost.B performed the best, followed by logistic regression.

Shirzad, Tabesh, and Farmani (2014) compared the performance of ANN and Support Vector Regression (SVR) methods in predicting the Pipe Burst Rate (PBR). They also studied the impact of hydraulic pressure on the accuracy of the data-driven pipe burst prediction model. They used two case studies for their analyses. Results revealed that in both case studies, ANN is a better (universal) predictor than SVR but cannot be generalized since it is not consistent with the physical behavior observed.

Kakoudakis et al. (2017) used Evolutionary Polynomial Regression (EPR) based on pipe length, diameter, and age. Individual pipes were aggregated into homogeneous groups based on age, diameter, and soil type. These groups were divided into training and test sets and cross-validated. The method of k-means clustering was used to partition the training data into clusters for individual EPR models. Then, these models were able to calculate the failure rate for individual pipes.

Kumar et al. (2018) attempted to predict which city blocks in Syracuse, NY were most likely to have a water main break in the next three years. They developed various machine learning classification methods to solve the problem and examine the relationships between predictive factors.

Snider and McBean (2018) applied a state-of-the-art gradient boosting machine learning algorithm (xgboost) to a large ductile iron pipe failure dataset. The model was designed to predict the time to next failure for individual ductile iron pipes. The overall root-mean-square error for the xgboost model was 5.81, a 1.2% improvement over the Random Forest (RF) model, and a 25.9% improvement over the ANN model. The results suggested that xgboost algorithm is a reliable option for the industry to predict time to pipe failure.

Sattar et al. (2019) proposed a novel failure rate prediction model by the extreme learning machine (ELM) to provide the required information for optimum ongoing maintenance/rehabilitation of a water network. The model was trained by more than 9500 instances of pipe failure in the Greater Toronto Area, Canada, from 1920 to 2005. The data included pipe attributes, including length, diameter, material, and previously recorded failures. The model had a superior prediction accuracy compared to other machine learning algorithms, such as feed-forward ANN, support vector regression, and non-linear regression.

Shirzad and Safari (2019) used RF technique and Multivariate Adaptive Regression Splines (MARS) to predict pipe failure rate using pipe diameter, length, installation depth, age, and average hydraulic pressure as input variables. The RF technique performed better than MARS, but MARS was chosen for implementation since it provided explicit equations that could be used as practical tools.

Robles-Velasco et al. (2020) also examined predicting pipe failure by using logistic regression and support vector classification. They studied the relationships between factors unlike some previously mentioned. Their results showed that the logistic regression model performed slightly better than the support vector classification model and the city was able to avoid 30% of the breakage by replacing 3% of the pipes.

Considering the limitations of physical and statistical models, which are comprehensively reviewed in the literature, adopting machine learning algorithms seems a more promising approach in the prediction of water main breaks. Moreover, including spatial methods such as hotspot analysis is not present in the majority of previous works. Combining the spatial patterns and the temporal data provides a solid foundation for a holistic model for the goal of failure prediction. In addition, it appears

that including a time-series analysis in the predictive models is a neglected approach in the literature.

3. Predictive Models Development

In this section, the methods implemented in this study will be discussed. First, spatial clustering and hotspot analysis will be explained. Following that, the process of developing the water distribution network, estimating some physical variables, and the source of environmental factors will be presented. Finally, machine learning algorithms will be explained in detail.

3.1. *Spatial Clustering*

Hotspot analysis (Getis-Ord G_i^* statistic) is a well-established method of spatial clustering for analyzing the features of spatial data (points or areas) (Esri 2016; Ord and Getis 1995). This method is an extension of the General G-statistic method for quantifying the spatial autocorrelation over an area. However, the G_i^* statistic calculates a measure of spatial autocorrelation variation for each point (polygon) in the area, instead of an overall index in the general framework. This method evaluates the similarity degree for high or low values of a feature (number of water main breaks in our problem) within a specified geographical distance (neighborhood). This index is calculated by equation 1:

$$G_i^*(d) = \frac{\sum\limits_j w_{ij}(d)x_j}{\sum\limits_j x_j} \tag{1}$$

Where x_j is the number of breaks for each region, w_{ij} is the spatial weight for the pair neighbors of i and j, and n is the total number of samples in the dataset. The geographical distances from each feature to its neighboring features are calculated by the Euclidean method. The spatial weight matrix is an $n \times n$ matrix, in which each value is a weight that reflects the relationship between a pair of features in the study region. The threshold parameter d defines the distance within which locations i and j are considered as neighbors ($w_{ij} = 1$ in the weight matrix), and beyond that distance, the locations are no longer considered as neighbors ($w_{ij} = 0$ in the weight matrix).

The statistical significance of clustering is evaluated based on a confidence level and a normalized z-score. The standardized $G_i^*(d)$ as a z-score is calculated based on equation 2:

$$G_i^*(d) = \frac{\sum\limits_j w_{ij}(d)x_j - \bar{X}\sum\limits_{(j)} w_{ij}(d)}{s\sqrt{\frac{n\sum\limits_j w_{ij}^2 - (\sum\limits_j w_{ij}(d))^2}{n-1}}}) \Longrightarrow \bar{X} = \frac{\sum\limits_j x_j}{n}, s = \sqrt{\frac{\sum\limits_j x_j^2}{n} - (\bar{X})^2} \tag{2}$$

To interpret the results of hotspot analysis, two maps should be analyzed concurrently. One map indicating the location of spatial clusters in the study area. Positive values of G_i^* statistic indicate spatial dependence among high values and negative values show spatial dependence for low values. The second map reflects the statis-

tical significance of each polygon compared to its neighbors by providing a p-value (Peeters et al. 2015). These z-scores and p-values maps together are used to label an area as a hotspot (spatial cluster of high data values), cold-spot (spatial cluster of low data values), or an outlier (a high value surrounded by low values or vice versa). In this study, as we are focusing only on hotspots, we labeled polygons with positive z-scores and statistically significant p-values as high/medium and low hotspot levels. We assigned a none hotspot level in our modeling approach for polygons with negative z-score and significant p-values, which indicate cold-spots. Evidently, all other areas with insignificant p-values were also categorized in the none hotspot level. The results of this analysis are provided in section 5.

3.2. *Physical Network and Environmental Features*

To construct the water distribution network, we first simplified the distribution system shapefile from a network consisting of 76,000 pipes to just over 1,200 pipes, using the skelebrator tool in WaterGems software. The skelebrator combines pipes that are in series or run parallel to each other to one equivalent pipe. In addition, pipes with a diameter less than 8 inches were eliminated from the model for further simplification. Afterward, the average water consumption at each junction was incorporated into the hydraulic model in ArcGIS.

For the material, we assigned the material of each pipe based on the nearest distance to the original network. The material of the original network is composed of 85% ductile iron pipes, 9% cast iron, 3% galvanized iron, 2% HDPE, and 1% PVC by length. For the age variable, inspired by Santana (2015), we used the parcel-level land use data publicly available from Hillsborough County Property Appraiser (HCPA). After finding the midpoints of pipes, an estimation of age was assigned based on the located census block group. However, in the case of not having an age estimation, we considered 30 years ago as a baseline age.

Average temperature and total precipitation for each month a break occurred were collected from the National Oceanic and Atmospheric Administration (NOAA). For each month a break occurred, the number of days in that month where the temperature was greater than or equal to 70 and 90 degrees Fahrenheit, the number of days where precipitation was over 0.1 inches, the total precipitation, the average temperature, the average maximum temperature, and the average minimum temperature. Table 1 summarizes the variables featured in the dataset, which can be classified into three different categories: physical, environmental, and others.

Table 1. Dataset Variable Categories

Physical	Environmental	Others
Longitude	Available land	Pipe ID
Latitude	Available water	Hotspot level
Length	Total precipitation	Date of break
Diameter	Number of days precipitation greater than or equal to 0.1 in.	Trend
Material	Number of days greater than or equal to 70 degrees Fahrenheit	Year built
	Number of days greater than or equal to 90 degrees Fahrenheit	Census tract
	Average temperature	Contract Type
	Average max temperature	
	Average min temperature	

3.3. *Machine Learning Methods*

Due to the limitations of physical and statistical methods for the prediction of water main breaks, we approached the problem from a modeling method combining machine learning and statistical techniques by four well-established methods in the literature, Random Forest (RF), Boosted Regression Tree BRT, Multivariate Adaptive Regression Splines (MARS), and Artificial Neural Networks (ANN). The following sections will go further in-depth on each of these modeling techniques.

3.3.1. Random Forest

The random forest method is a well-known machine learning technique for classification and regression analysis based on combinations of several decision trees. It is a decision tree-based method that consists of a large number of individual decision trees that operate together to produce better results. The output, in this case, is the mean prediction (regression) of the individual trees. This algorithm uses bagging and feature randomness when building individual trees to create an uncorrelated forest of trees with a more accurate grouped prediction. The first requirement of this algorithm is to provide some predictive power for the input variables, meaning that the model built with these variables will perform better than random guessing. The second requirement is that the predictions and errors of the individual trees must have low correlation values with each other. To this end, the algorithm uses a technique called bagging, which trains the model on different sets of data and uses various features to make decisions and to creates individual uncorrelated trees that buffer against each other (Shirzad and Safari 2019).

3.3.2. Boosted Regression Tree

The boosted regression tree algorithm, also known as gradient boosting, is similar to the RF algorithm in that it is a decision tree method that uses a combination of individual decision trees to provide better results. Similar to RF, BRT also uses a random subset of the data with a replacement to build each individual tree. The main difference between these algorithms is that while RF implements the bagging technique, BRT uses the boosting technique. The boosting technique involves weighing each individual tree so that they are applied in such a way that poorly modeled data by a previous tree has a higher probability of being selected for a future tree. After the first tree is fitted, the model will take into account that trees error when fitting the next tree and so on. The model continuously tries to increase its results using this sequential approach. To name some advantages this modeling technique provides are that it is robust to outliers, the best fit is automatically detected, and it is stochastic, which improves predictive performance results (Chen, Beekman, and Guikema 2017).

3.3.3. Multivariate Adaptive Regression Splines

Multivariate adaptive regression splines algorithm creates a piecewise linear model that provides an intuitive stepping block into nonlinearity. MARS can capture the nonlinearity aspect of polynomial regression by assessing cutpoints, also called knots, like a step function. The procedure for this algorithm determines each data point for each predictor as a knot and creates a linear regression model with the selected feature(s). The process can continue until many knots are found, producing a highly nonlinear pattern. Once the full set of knots has been found, knots that do not sig-

nificantly contribute to predictive accuracy can be sequentially removed or pruned to find the optimal number of knots (Shirzad and Safari 2019).

3.3.4. Artificial Neural Networks

Artificial Neural Networks(ANN) is a computational approach inspired by the biological nervous systems process. ANNs are adaptive and capable of handling complex systems, which can identify patterns and learn from their interactions with the environment. The architecture of ANN includes several nodes (neurons) organized in input and output layers as well as several hidden layers. ANNs are flexible and learn in an iterative process of adjusting the weights of inputs and biases. The most common learning is supervised learning, which provides a response value is predicted for a set of input values. The difference between the predicted response and the actual target values is defined as the error value. The network weights are adjusted iteratively based on the error value by a back-propagation technique (BPNN) in order to minimize the error (Jafar, Shahrour, and Juran 2010). In this paper, we used a single hidden layer network with weight decay.

4. Case Study

In this study, a dataset containing the temporal data of water main breaks in the water distribution network of the city of Tampa, Florida, is utilized. The original dataset only provided information about the spatial location of failures, the time and date the main break reported and documented, and the contractor type (incomplete) of water main breaks from 2015 to early 2020 in the predefined area. This dataset is being updated continuously with recent incidents. More than 3336 events are considered as our database for analysis. As pipe specific features of each break are not available in the original database, we then used estimation techniques to add the associated failed pipe to each break, the census tract, and the available land and water around the breaks to the raw data. We also included pipe-related features like length, diameter, year built, and material to the water main break dataset based on the assigned pipe to each incident. Fig. 2 depicts the distribution of water main breaks over the selected period.

The final dataset of water main breaks is summarized in Fig. 3. The breaks are first separated based on the occurrence years and then categorized in different classes of diameter, length, age, material, and hotspot level. The rates are normalized based on the total number of each category in the network (for example the total number of pipes older than 80 years) to make a clearer image about their contributions to the total breaks.

5. Results and Discussion

5.1. *Spatial Clustering*

We discretized the study area, boundaries of Tampa, Florida, based on census tracts to conduct the spatial clustering analysis. In specific, we used the census tracts data of Hillsborough county of Florida in 2019. The first step of the spatial clustering phase is calculating Getis-Ord G_i^* statistic for all polygons inside our study region. Fig. 4 shows

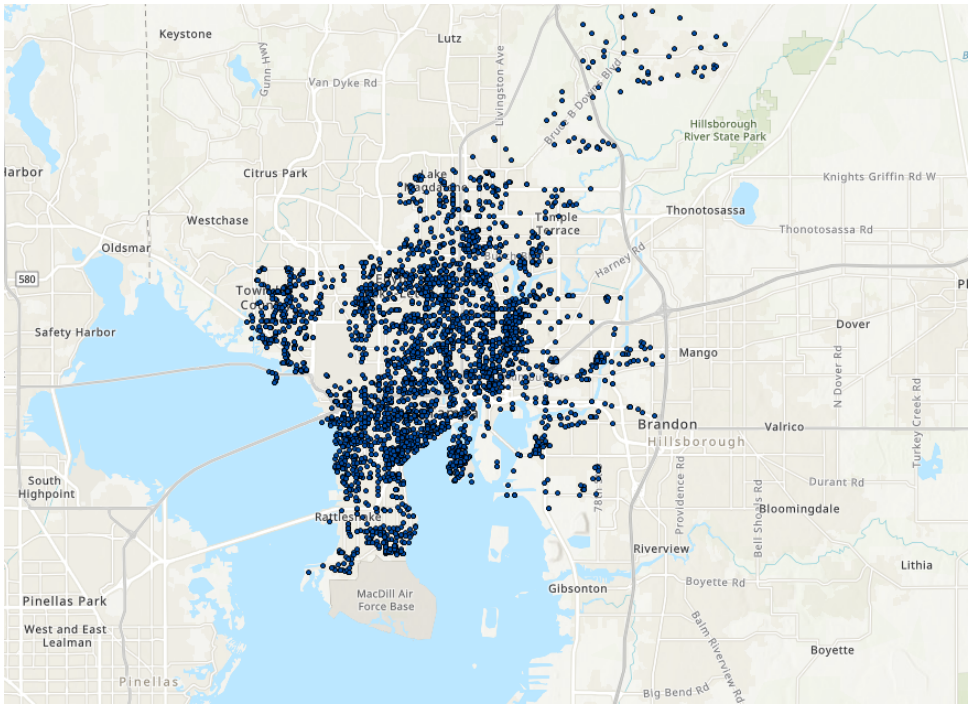


Figure 2. Water Main Breaks.

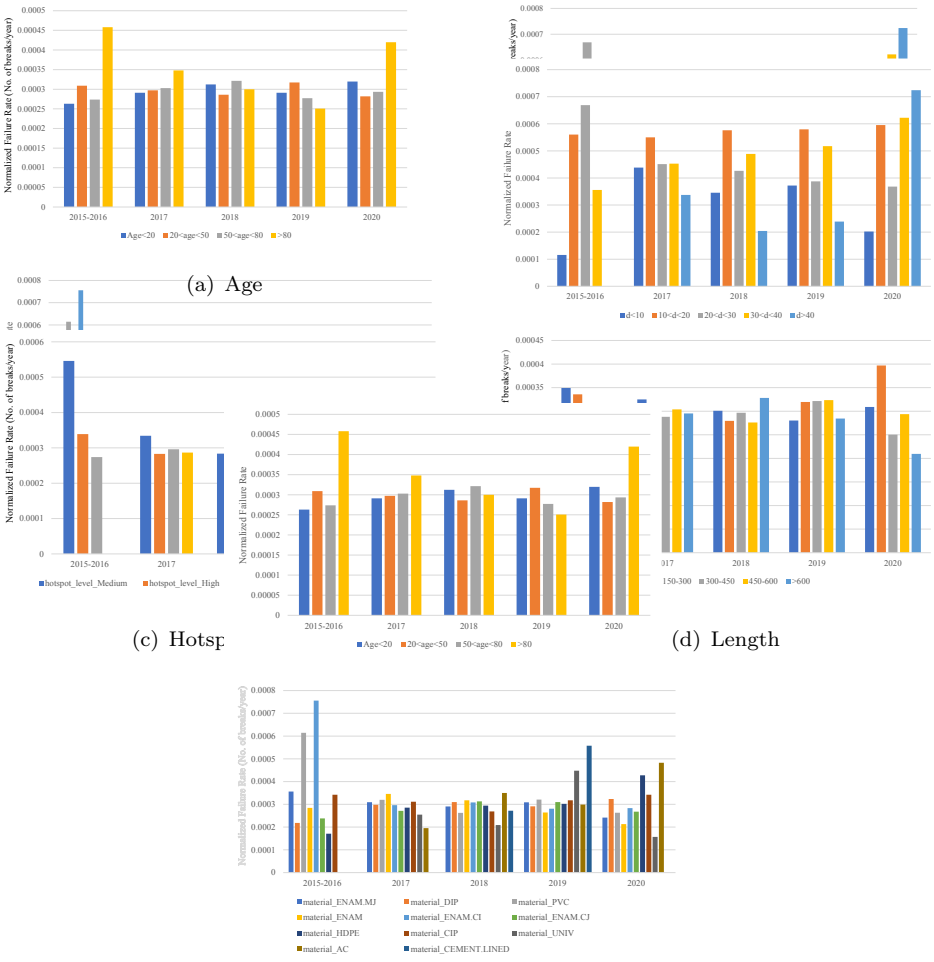
the distribution of this statistic over the city of Tampa. The northern and central regions have the lowest and highest value, respectively. The higher value of this measure indicates that in the neighborhood there is a similar pattern for a higher frequency of water main breaks. On the other hand, the negative values reflect higher clustering among neighboring regions with lower frequencies of breaks. Finally, by mapping the p-values (normalized z-values of Getis-Ord statistics), we can make a clearer interpretation of the hotspots. Fig. 5 illustrates the result of this mapping, in which we can see some statistically significant patterns. By comparing the mapped p-values by the Getis-Ord statistic map, we can detect several hotspots in the southwestern and central sections of the city. There is also one noticeable cold spot located in the northern area of Tampa. To incorporate this analysis in our machine learning algorithms, we defined a categorical variable named hotspot level with three levels of high/medium, low, and none based on the obtained p-values and z-scores. As we mentioned in section 3.1, we only labeled polygons with positive z-score and significant p-values as the hotpost. We incorporated areas with insignificant p-values and ones with significant p-values for negative z-scores in the none level for this categorical variable.

5.2. *Machine Learning Algorithms*

The procedure of developing machine learning models consists of three sections: Data Understanding, Data Preparation, and Data Modeling. This approach is also called the CRoss-Industry Standard Process for Data Mining (CRISP-DM).

5.2.1. Data Pre-processing

In this paper, the data pre-processing procedure begins with determining if any variables are highly correlated. Typically, variables with a correlation value of 0.7 or greater



(a) Age

(c) Hotsp

(d) Length

(e) Material

Figure 3. The summarized water main breaks of Tampa based on different variables

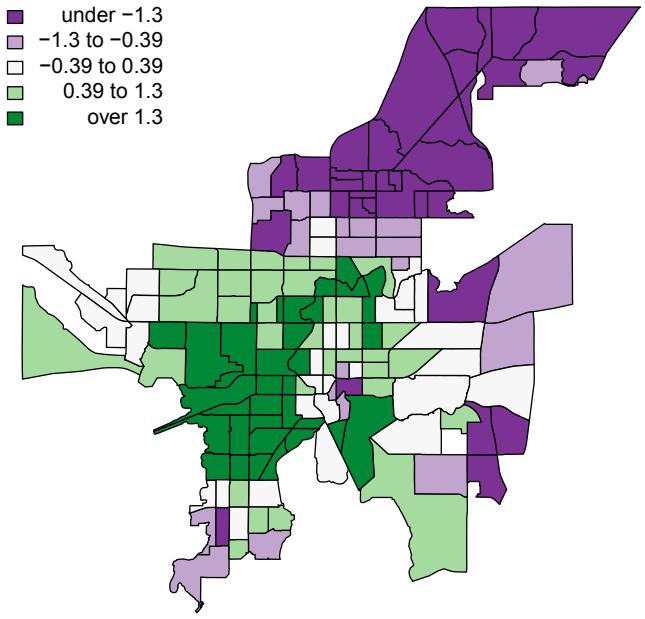


Figure 4. Spatial pattern of Getis-Ord statistic.

are considered to be too highly correlated. A correlation matrix using all of the variables was generated and based on the correlation values, it was determined that not all of the variables are necessary. The variables that were dropped include the number of days the temperature was greater than or equal to 90 degrees Fahrenheit, the number of days the temperature was greater than or equal to 70 degrees Fahrenheit, the average max temperature, the average min temperature, and the number of days the precipitation was greater than or equal to 0.1 inches. This set of dropped variables accounts for almost all of the environmental variables except for average temperature and total precipitation. Fig. 6 shows only the statistically significant correlations for each pair of variables in the dataset.

The next step in this section was to examine the time series analysis for this temporal data. To this end, seasonality was looked for in the number of breaks by month Fig. 7 and by month-year Fig. 8.

The autocorrelation and partial autocorrelation plots (Fig. 9 and Fig. 10, respectively) for the monthly data were the foundations of time-series analysis. From these plots, a trend, not specific seasonality, can be observed. Hence, the trend variable, representing each month in an increasing order, was added to extract the underlying time series behavior from the data. For example, 1 corresponds to the first month in the dataset, 2 represents the following month, and so on.

The final aspect of the Data Understanding section is to examine the missingness in the data. The variables contract type, pipe material, and the year the pipe was built were the only variables that exhibited missing values. Contract type had the most missingness with over 71% of the data missing. Since this variable was not important to stakeholders, it was dropped from our analysis. Pipe material showed 6.8% missingness and the year built had 1.2% missingness. The missing values were accounted for in the

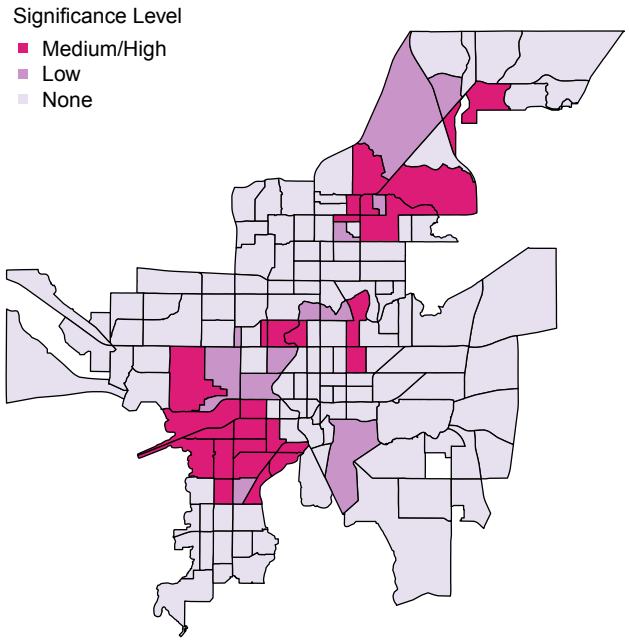


Figure 5. Spatial pattern of P-values of Getis-Ord statistics.

next section.

5.2.2. Data Preparation

This section begins by examining outliers in the dataset. Outliers were determined using Tukeys rule for finding outliers based on the interquartile range. Based on Tukeys rule, the outliers are the values more than 1.5 times the interquartile range from the quartiles, either below $Q1 - 1.5 \times IQR$, or above $Q3 + 1.5 \times IQR$. The outliers were then replaced with blank values.

The next step in the Data Preparation section was to complete some feature engineering. The first aspect was to resolve the missing values found in examining the missingness of the dataset in the previous section and the ones created by examining outliers. The technique of multiple imputation was used to resolve these missing values. We then selected one of the complete datasets created by multiple imputation with no missing values for the next phases. Since the hotspot level and material were categorical variables in the dataset, the next feature engineering aspect was to convert them to numerical variables. We used the well-known machine learning one-hot-encoding technique for this purpose. In this procedure, each level of the original categorical variable is converted into a new column and a binary value (notation for true/false) is assigned to the column.

Considering the pipe failure rate (pfr) to predict when a pipe will break as the response variable is a common approach in the literature (Sattar et al. 2019; Shirzad and Safari 2019; Tabesh et al. 2009). Here, we define a variable, pipe failure rate, as the dependent variable. Since the trend and weather data were broken down by month, the pipe failure rate was also defined based on a monthly time scale. Equation

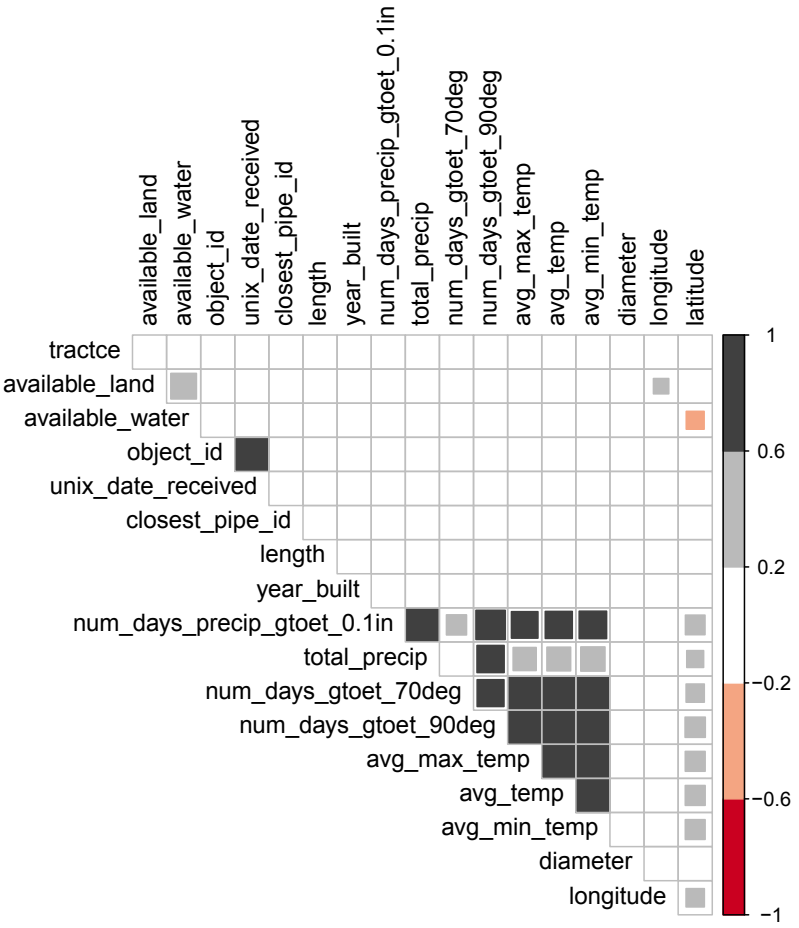


Figure 6. Correlation Heatmap.

3 defines this variable as:

$$pfr_i = \frac{number\ of\ breaks\ for\ pipe\ i}{Total\ number\ of\ months\ in\ dataset} \tag{3}$$

Please note that since our dataset contains 50 month of breaks recorded, the denominator of Equation 3 is considered 50 in our calculations, but the formula can be generalized to any time scale. Another aspect of the Data Preparation section was to normalize the data to eliminate the scale differences between variables. It is important to note that doing this usually increases accuracy (Shen et al. 2016; Singh and Singh 2019). The final aspect was to split the data into separate datasets to train and test the models on. The training data set is comprised of 70% of the data and the testing data set is comprised of the remaining 30% portion. The datasets were selected through random sampling.

5.2.3. Data Modeling

At the beginning of the Data Modeling step, a baseline linear regression model was formed to be able to compare other models. For all of the models, the response (dependent) variable was the pipe failure rate. The baseline linear regression model included all of the variables. In order to determine which variables provided the most signif-

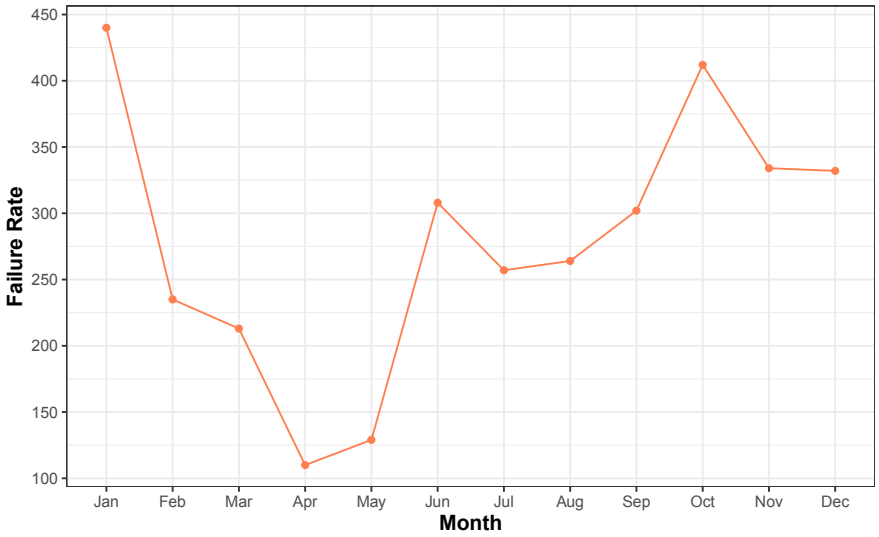


Figure 7. Number of Breaks by Month.

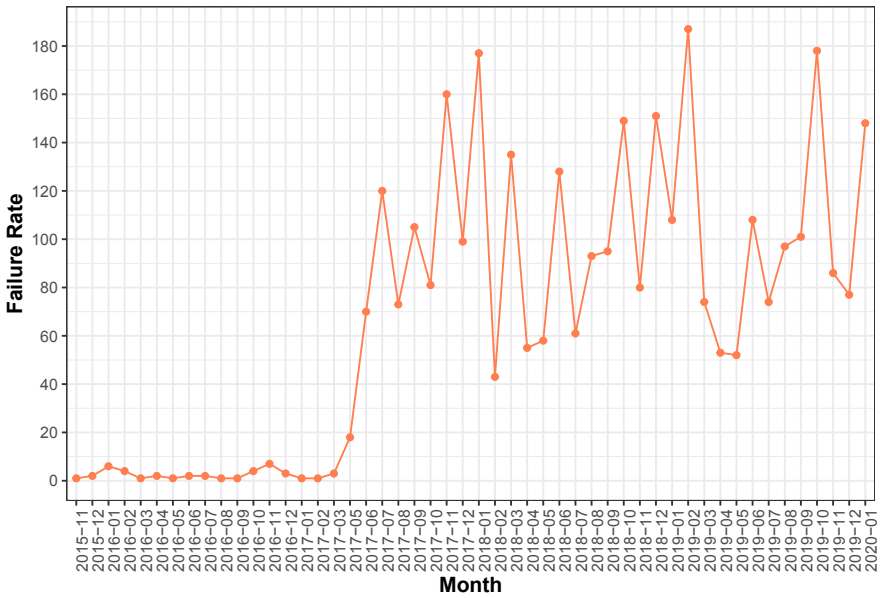


Figure 8. Number of Breaks by Month-Year.

icance, stepwise logistic regression going forward and backward was completed. The final model formula, including the final features and their units used in the training and test datasets, is presented in Table 2 with the pipe failure rate as the response variable.

The developed models were cross-validated with five folds. The idea for k-fold cross-validation is that the dataset is divided into k subsets and one of the k subsets is used

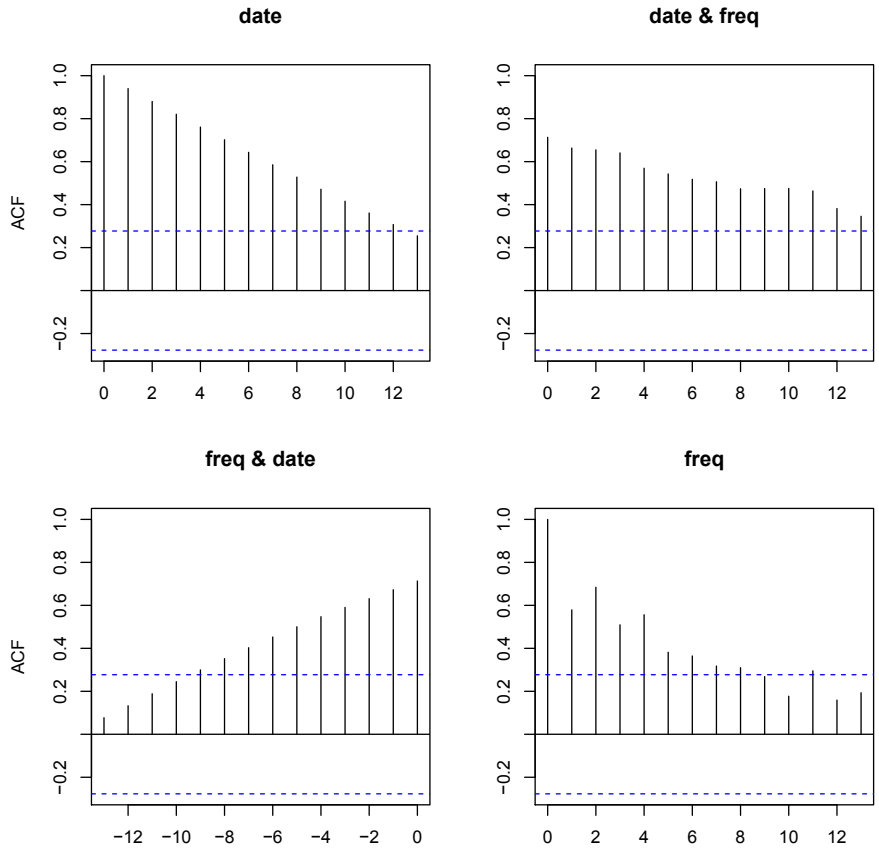


Figure 9. Auto correlation function plots.

as the test set while the remaining are used as the training set. This process is repeated k times. The error estimation is average overall k trials to get the total effectiveness of the model. Performing k-fold cross-validation significantly reduces bias and variance. As a general rule, $k = 5$ is the most common setting for validation. The models were also tuned using various grids to see if performance results could increase.

5.2.4. Information-base Performance Measures

We compared the performance of predictive models based on the measures defined in two separated sections. First, we use some common performance measures such as the root mean square error (RMSE) value for the training and testing sets and the mean absolute error (MAE) value for the training and testing sets. The testing set RMSE value was given the highest precedence when comparing model performance. The RMSE value is an indicator of how close the models predicted values are to the actual values in the dataset. The MAE value measures the average magnitude of the errors in a prediction without considering their direction. Essentially, it is the average over the sample of the absolute differences between the predicted and actual values. Even though it is argued in the literature that using RMSE is not a good option in comparing predictive models for zero-inflated data of water main breaks, we still report this measure. However, we also calculate the information-based criteria (refer to Mohebbi et al. (2019)) Akaikes classic Information Criterion (AIC), Information Complexity (ICOMP), and Consistent Akaikes AIC (CAIC) and use these measures as well as the rank-based measure to interpret the performance of models. Table 3 shows the results from the models.

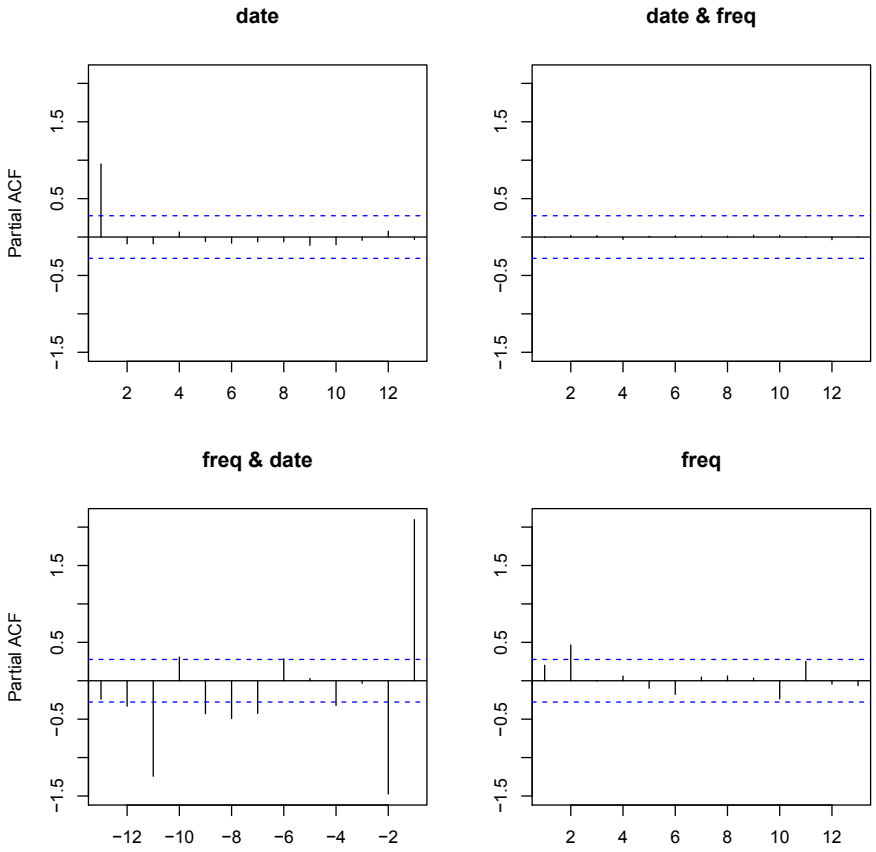


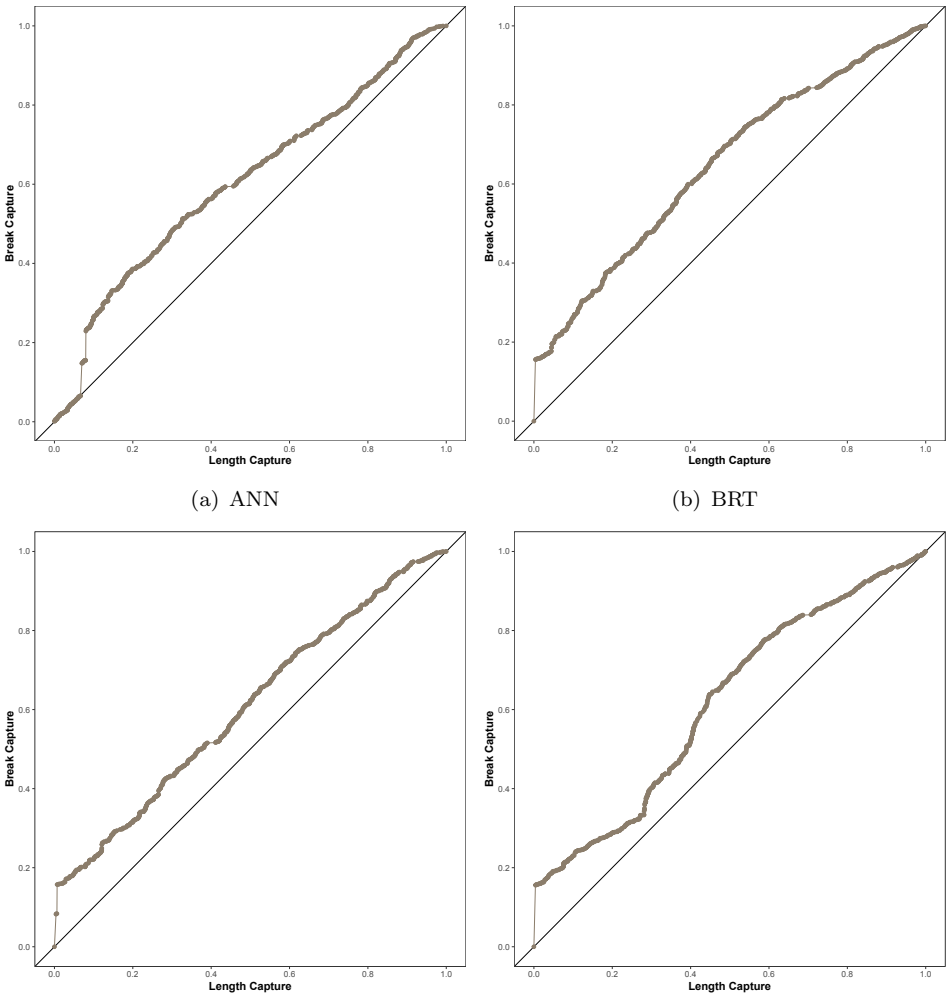
Figure 10. Partial auto correlation function plots.

5.2.5. Rank-base Performance Measures

While using performance measures such as the mean square error (MSE) and the root mean square error (RMSE) is a common approach in the literature to compare predictive models, it is argued that the data related to infrastructure are zero-inflated. In other words, there is a noticeable share of instances with zero breaks. As a result, using MSE and RMSE criteria will skew the accuracy evaluation towards the zero break instances (Chen et al. 2019). Inspired by Chen and Guikema (2020) and Choi et al. (2017), we implement a rank-based evaluation to have a more useful analysis for our case study. This method's output will provide decision-makers at the municipality level with valuable information to better prioritize limited budgets and resources.

The idea behind this rank aggregation approach is to evaluate the predictive models based on their accuracy in predicting the higher rate of breakage. In this method, a model is better if it can accurately assign a higher number of predicted breaks on the pipes with a higher failure rate and predict fewer breaks on those with lower rates. We can quantify this measure of goodness for prioritization by counting the number of potential breaks avoided when sorting the predicted breaks in a descending manner (high to the low number of predicted breaks) (Chen et al. 2019). The output of this method can be visualized by using a rank-ordered break capture curve, and the area under this curve reflects the goodness of fit for the predictive model. Fig. 11 and Table 4 show the Ranked Ordered Curve for the predictive models developed in this paper as well as the associated area under the curve.

As we mentioned, infrastructures' administrators are more interested in allocating



(a) ANN (b) BRT

(c) MARS (d) RF

Figure 11. Ranked Order Plot, Break Proportion Capture vs. Length Proportion Capture

Table 2. Significance level and units for final features used in the training and testing datasets

Variable	Unit	Coefficient	P-value
pfr (Response Variable)	breaks per month	NA	NA
trend	NA	-0.0002	0.0418 *
available land	square feet	0.0804	<2e-16 ***
available water	cubic foot	-0.2654	<2e-16 ***
total-precip	millimeters	-0.0056	0.00175 *
avg-Temp	degrees Fahrenheit	0.01514	3.95e-14 ***
Length	ft	-0.4062	< 2e-16 ***
Diameter	in	-0.9872	< 2e-16 ***
Year built	NA	-0.9991	< 2e-16 ***
Hotspot level none	NA	0.0162	<2e-16 ***
Hotspot level low	NA	0.0219	<2e-16 ***
Hotspot level medium-High	NA	-0.0168	<2e-16 ***
Material ENAM.MJ	NA	0.4352	<2e-16 ***
Material DIP	NA	0.9824	< 2e-16***
Material PVC	NA	0.4401	< 2e-16***
Material ENAM	NA	0.3537	<2e-16 ***
Material ENAM.CI	NA	0.1873	2.23e-14***
Material ENAM.CJ	NA	0.3846	<2e-16 ***
Material HDPE	NA	0.4691	<2e-16 ***
Material UNIV	NA	0.2498	<2e-16 ***
Year-break	NA	0.0614	3.49e-14 ***
avgpfr	breaks per month	0.1382	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
NA: Not Applicable for categorical variables

Table 3. Results of model in the performance measures

Model	RMSE Train	MAE Train	RMSE Test	MAE Test	ICOMP	AIC	CAIC
ANN	0.084	0.021	0.083	0.014	2565.783	2610.787	2744.063
BRT	0.008	0.001	0.007	0.001	2455.655	2500.658	2633.933
MARS	0.017	0.003	0.015	0.002	2462.381	2507.384	2640.660
RF	0.030	0.012	0.028	0.012	2466.510	2511.513	2644.788

the limited budget and resources to the highest priority needs. To this end, we stratified the data based on the length and the number of breaks. We calculated the area under the Ranked Ordered Curve by considering only the top 20%, 40%, and 60% of top-ranked number of breaks and length of the system (For instance, in 20% case for length, we only considered the portion of data to the point that covers 20% of the total length of the network). Table 5 shows the area under the curve for each of these scenarios.

By having the results of two information-based and rank-based criteria, we can now have a comprehensive interpretation regarding the performance of predictive models. Based on Table 3, BRT outperforms other models with a lower MAE measure (We do not interpret RMSE measure). The AIC, ICOMP, and CAIC measure also certify that BRT has a better performance. By focusing on this section, MARS is the second-best

Table 4. Area under the Ranked Ordered Curve

Predictive Model	ANN	BRT	MARS	RF
Area Under Curve	0.605	**0.647**	0.604	0.615

Table 5. Area under the Ranked Ordered Curve at top segments for length and number of breaks

Predictive Model		ANN	BRT	MARS	RF
Length	20%	0.041	**0.051**	0.045	0.044
	40%	0.136	**0.145**	0.125	0.121
	60%	0.262	**0.285**	0.252	0.256
Number of breaks	20%	0.004	0.008	0.010	**0.012**
	40%	0.052	0.058	0.069	**0.075**
	60%	0.172	0.145	**0.174**	0.141

one, and RF is the next best model. By incorporating the rank-based criteria, results presented in Table 4, we can infer that BRT is still the best model by covering the highest area under the curve compared to other competitors. However, this metric was higher than 0.6 for all four developed models. Finally, the results of top-ranked failures in Table 5 indicate that BRT covers the highest length of the system in all cases. RF and MARS have a better performance in terms of the number of breaks, though. ANN fails to predict well in the top 20% of breaks in both length and number of breaks sections, while its performance improves markedly for higher layers. To sum up, we can conclude that BRT reflects a reliable performance in both information and rank-based performance measures. Therefore, BRT is the recommended model to the stakeholders and decision-makers to apply to real water networks of Tampa for water main break prediction. If the networks administrators are only interested in top-rank layers of failures, BRT is still the best option for length and RF can predict the highest failure rates more accurately.

We highlighted the first goal of the study as to quantify and interpret the most important variables in the predictive models. To achieve this goal, we focus on the variable importance measure provided by the developed models. We used a model-independent metric for the developed models by tracking the changes in model's statistics, such as generalized cross-validation, for each predictor and accumulating the reduction in the statistic when each feature of the predictor is added to the model. We used this total reduction as the measurement of variable importance. Fig. 12 shows this importance criterion for the most essential variables in each model. Starting with BRT as the best model, DIP material is the variable with the most contribution, which is expected as 85% percent of the network is composed of this material option. The age (Year-built), diameter, and length of pipes are the next crucial variables. This also verifies the role of aging water infrastructure in experiencing higher rates of water main breaks in North America. Interestingly, in the developed MARS model, only these three variables are important. We can see an almost similar pattern in two other predictive models in terms of variable importance. However, the Trend (time series variable) is the second important variable in ANN, while two types of material are present in RF.

Finally, To better aid municipalities in rehabilitation and fortification decisions, the model that provided the best performance, i.e. BRT, was used to predict the failure rate for the entire water distribution network in the City of Tampa. Fig. 13 illustrates the normalized failure rates for all pipelines in the present situation and the predicted ones by the BRT in the next five years. This figure can be used to identify the areas with higher failure rates in order to mitigate the consequences of water main breaks

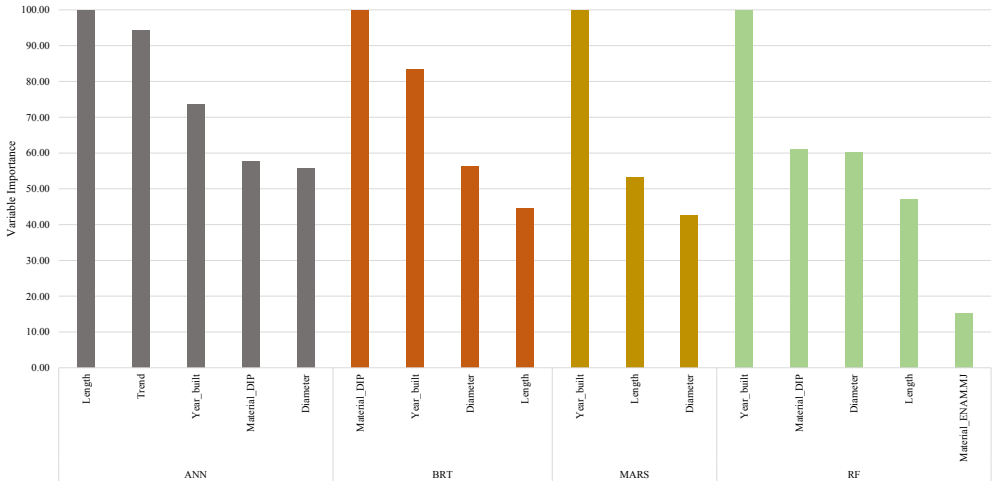


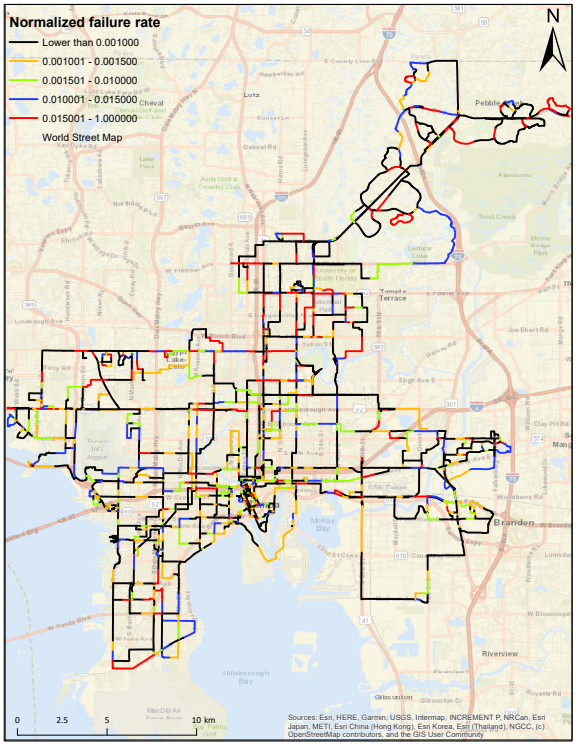
Figure 12. Variable Importance of predictive models

in terms of financial burdens and service disruptions.

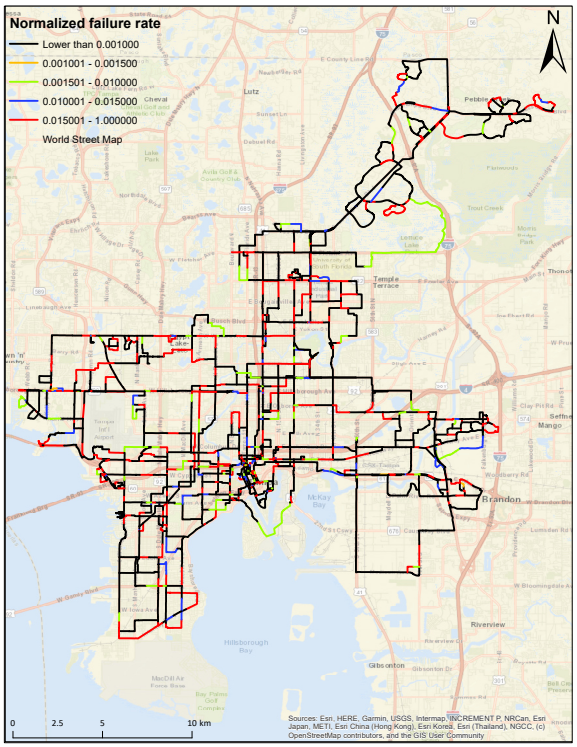
6. Conclusions

In this paper, we aimed to develop predictive models for water main breaks using spatiotemporal data. We used the dataset of recorded water main breaks in Tampa, Florida, as our case study. To improve the accuracy of our models, we incorporated several variables to the initial database, including the material of pipes, (approximate) age of pipes, average temperature, and total precipitation. We first identified the most vulnerable areas in the city by performing a spatial clustering analysis. The result of this analysis was translated into a categorical independent variable named hotspot level in the modeling phase. We then performed a time-series analysis to address the temporal pattern of our data. The output of this analysis, variable trend, and the known average failure rate were also incorporated into the predictive models as another independent variable. We defined the pipe failure rate as the response variable in this study. The results of stepwise logistic regression showed that both spatial and temporal variables, in addition to physical and environmental factors, are significant. Afterward, we developed four predictive models based on various machine learning algorithms. We compared the performance of these models by several performance measures. The results indicated that BRT was superior to others.

From a practical point of view, decision-makers can use the results of this study in the following ways. First, they can investigate the hotspots to identify noticeable technical and physical differences compared to non-significant places. The results will also help them design more proactive maintenance schemes to prevent (eliminate) future failure-related problems. Renovating the equipment and components (fortification) in the pipes with a higher predicted failure rate located in hotspots to decrease their vulnerability can be another worthy resulting strategy from the information provided by predictive models. Moreover, the output of top-ranked failures from the proposed rank-based performance measure can help the decision-makers allocate scarce and expensive resources only to the network's highest priorities in an efficient manner. The



(a) Present Situation



(b) 5-years Predicted Situation

Figure 13. The present and 5-years predicted normalized failure rate for WDN in Tampa

final output of the developed framework can be improving the overall resilience of the water network to perform better in unpredictable situations and enhancing the performance of interdependent infrastructures (transportation) by eliminating the Achilles' heel of the water network.

For future studies, there are some possible promising directions. Including the cause of water main breaks in the modeling phase can potentially improve predictive and vulnerability analyses. A high percentage of breaks are usually because of physical problems inside the network. While some breaks are due to natural disasters and are somehow inevitable, a part of breaks can be a direct result of human actions such as carelessly excavating activities. Thus, making a distinction among these reasons can provide more insightful information for decision-makers.

Acknowledgment

This work was supported by the US National Science Foundation under Grant Number 1638301. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. In addition, we are grateful to the anonymous referees for their comments and suggestions resulting in an improved presentation.

Disclosure Statement

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.

References

Aven, T. (2016). Risk assessment and risk management: Review of recent advances on their foundation. *European Journal of Operational Research*, 253(1): 1-13.

Barton, N. A., Farewell, T. S., Hallett, S. H., and Acland, T. F. (2019). Improving pipe failure predictions: Factors effecting pipe failure in drinking water networks. *Water research*, 114926.

Chen, T. Y. J., Beekman, J. A., and Guikema, S. D. (2017). Drinking water distribution systems asset management: Statistical modelling of pipe breaks. *In Pipelines 2017*, pp. 173-186.

Chen, T. Y. J., & Guikema, S. D. (2020). Prediction of water main failures with the spatial clustering of breaks. *Reliability Engineering & System Safety*, 203: 107108.

Chen, T. Y. J., Beekman, J. A., David Guikema, S., and Shashaani, S. (2019). Statistical Modeling in Absence of System Specific Data: Exploratory Empirical Analysis for Prediction of Water Main Breaks. *Journal of Infrastructure Systems*, 25(2): 04019009.

Choi, G. B., Kim, J. W., Suh, J. C., Jang, K. H., & Lee, J. M. (2017). A prioritization method for replacement of water mains using rank aggregation.*Korean Journal of Chemical Engineering*, 34(10): 2584-2590.

Demissie, G., Tesfamariam, S., and Sadiq, R. (2017). Prediction of pipe failure by considering time-dependent factors: Dynamic Bayesian belief network model. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineerin*, 3(4): 04017017.

Duchesne, S., Toumbou, B., and Villeneuve, J.-P. (2016). Validation and comparison of different statistical models for the prediction of water main pipe breaks in a municipal network in Qubec, Canada. *Revue des sciences de leau/Journal of Water Science*, 29(1): 11-24.

Esri. (2016). How emerging hot spot analysis works.
Folkman, S. (2018). Water main break rates in the USA and Canada: A comprehensive study.
Francis, R. A., Guikema, S. D., and Henneman, L. (2014). Bayesian belief networks for predicting drinking water distribution system pipe breaks. *Reliability Engineering & System Safety*, 130: 1-11.
Jafar, R., Shahrour, I., and Juran, I. (2010). Application of Artificial Neural Networks (ANN) to model the failure of urban water mains. *Mathematical and Computer Modelling*, 51(9-10) :1170-1180.
Kabir, G., Tesfamariam, S., Francisque, A., and Sadiq, R. (2015). Evaluating risk of water mains failure using a Bayesian belief network model. *Revue des sciences de leau/Journal of Water Science*, 240(1): 220-234.
Kakoudakis, K., Behzadian, K., Farmani, R., and Butler, D. (2017). Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K-means clustering. *Urban Water Journal*, 14(7): 737-742.
Kleiner, Y. , and Rajani, B. (2000). Considering time-dependent factors in the statistical prediction of water main breaks. *Paper presented at the Proc., American Water Works Association Infrastructure Conference.*
Kleiner, Y. , and Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: statistical models. *Urban water* 3(3): 131-150.
Kumar, A., Rizvi, S. A. A., Brooks, B., Vanderveld, R. A., Wilson, K. H., Kenney, C., and Zuckerbraun, J. (2018). Using machine learning to assess the risk of and prevent water main breaks. *Paper presented at the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.*
Martinez-Codina, A., Castillo, M., Gonzalez-Zeas, D., and Garrote, L. (2016). Pressure as a predictor of occurrence of pipe breaks in water distribution networks. *Urban Water Journal* 13(7): 676-686.
Mohebbi, S., Pamukcu, E., & Bozdogan, H. (2019). A new data adaptive elastic net predictive model using hybridized smoothed covariance estimators with information complexity. *Journal of Statistical Computation and Simulation* 89(6): 1060-1089.
Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis* 27(4): 286-306.
Peeters, A., Zude, M., Kthner, J., nl, M., Kanber, R., Hetzroni, A., and Ben-Gal, A. (2015). GetisOrds hot-and cold-spot statistics as a basis for multivariate spatial clustering of orchard tree data. *Computers and Electronics in Agriculture* 111: 140-150.
Phan, H. C., Dhar, A. S., Hu, G., and Sadiq, R. (2019). Managing water main breaks in distribution networksA risk-based decision making. *Reliability Engineering & System Safety* 191: 106581.
Rajani, B., and Kleiner, Y. (2001). Comprehensive review of structural deterioration of water mains: physically based models.*Urban water*, 3(3): 151-164.
Renzetti, S., Dupont, D., and Dupont, D. P. (2013). Buried treasure: the economics of leak detection and water loss prevention in Ontario. *Rep. No. ESRC-2013* 1.
Robles-Velasco, A., Corts, P., Muuzuri, J., and Onieva, L. (2020). Prediction of pipe failures in water supply networks using logistic regression and support vector classification. *Reliability Engineering & System Safety* 196: 106754.
Santana, M. V. E. (2015). The Effect of Urbanization on the Embodied Energy of Drinking Water in Tampa, Florida.
Sattar, A. M., Erturul, . F., Gharabaghi, B., McBean, E. A., and Cao, J. (2019). Extreme learning machine model for water network management. *Neural Computing and Applications* 31(1): 157-169.
Shen, X., Gong, X., Cai, Y., Guo, Y., Tu, J., Li, H., ... & Zhu, Z. J. (2016). Normalization and integration of large-scale metabolomics data using support vector regression. . *Metabolomics* 12(5): 89.
Shirzad, A., and Safari, M. J. S. (2019). Pipe failure rate prediction in water distribution networks using multivariate adaptive regression splines and random forest techniques. *Urban*

Water Journal 16(9): 653-661.
Shirzad, A., Tabesh, M., and Farmani, R. (2014). A comparison between performance of support vector regression and artificial neural network in prediction of pipe burst rate in water distribution networks.*KSCE Journal of Civil Engineering* 18(4): 941-948.
Shortridge, J. E. , and Guikema, S. D. (2014). Public health and pipe breaks in water distribution systems: analysis with internet search volume as a proxy. *Water research*
Singh, D., & Singh, B. (2019). Investigating the impact of data normalization on classification performance. *Applied Soft Computing* 105524.
Snider, B., and McBean, E. A. (2018). Improving time to failure predictions for water distribution systems using extreme gradient boosting algorithm. *In WDSA/CCWI Joint Conference Proceedings (Vol. 1).*
Snider, B., and McBean, E. A. (2020). Improving Urban Water Security through Pipe-Break Prediction Models: Machine Learning or Survival Analysis.*Journal of Environmental Engineering*146(3): 04019129.
Snider, B., and McBean, E. A. (2020). Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions. *Urban Water Journal*, 1-14.
Tabesh, M., Soltani, J., Farmani, R., and Savic, D. (2009). Assessing pipe failure rate and mechanical reliability of water distribution networks using data-driven modeling. *Journal of Hydroinformatics* 11(1): 1-17.
Vanrenterghem-Raven, A. (2007). Risk factors of structural degradation of an urban water distribution system. *RJournal of infrastructure systems* 13(1): 55-64.
Wang, C. W., Niu, Z. G., Jia, H., and Zhang, H. W. (2010). An assessment model of water pipe condition using Bayesian inference. *Journal of Zhejiang University-SCIENCE A* 11(7): 495-504.
Wang, R., Dong, W., Wang, Y., Tang, K., and Yao, X. (2013). Pipe failure prediction: A data mining method. *Paper presented at the 2013 IEEE 29th International Conference on Data Engineering (ICDE).*
Wilson, D., Y. Filion, and I. Moore. 2017. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. *Urban Water Journal* 14(2): 173184.
Wood, A., and Lence, B. J. (2009). Using water main break data to improve asset management for small and medium utilities: District of Maple Ridge, BC. *Journal of Infrastructure Systems* 15(2): 111-119.
Yamijala, S., Guikema, S. D., and Brumbelow, K. (2009). Statistical models for the analysis of water distribution system pipe break data. *Reliability Engineering & System Safety* 94(2): 282-293.