Prediction of Breaks in Municipal Drinking Water Linear Assets

Farzad Karimian[1], Khalid Kaddoura[2], Tarek Zayed[3], and Alaa Hawari[4], Osama Moselhi[5]

[1]Project Controller, Crosslinx Transit Solutions, Toronto, Ontario, Canada. farzad.karimian@gmail.com

[2] Asset Management Specialist, Water Department, AECOM Canada, Mississauga, Ontario, Canada. khalid.kaddoura@aecom.com. **(Corrosponding Author)**

[3]Professor, Dept. of Construction and Real Estate, Hong Kong Polytechnic University, Hung Hom, Hong Kong. tarek.zayed@polyu.edu.hk

[4]Associate Professor, Department of Civil and Architectural Engineering, Qatar University, P.O. Box 2713, Doha, Qatar. a.hawari@qu.edu.qa

[5] Professor, Dept. of Building, Civil and Environmental Engineering, Concordia University, Montreal, Quebec, Canada. moselhi@encs.concordia.ca

ABSTRACT

Improper asset management practices increase the probability of water main failures due to inactive intervention actions. The annual number of breaks of each pipe segment is known as one of the most important criteria in the condition assessment of water pipelines. This metric is also considered one of the major performance measures in Levels of Service (LoS) studies. In an effort to maximize the benefits of historical data, this research utilized the Evolutionary Polynomial Regression (EPR) in determining the best mathematical expression for predicting water pipeline failures. The prediction model was trained and tested on the City of Montreal water network. After determining the best independent variables through the Best Subset Regression, pipelines were clustered based on their attributes (length, diameter, age, and material). The majority of the models provided high R^2 values, but the highest performing model's R^2 was

89.35%. Further, a sensitivity analysis was also performed and showed that the most sensitive parameter was the diameter, and the most sensitive material type to age was ferrous material. The tools and stages performed in this research showed promising results in predicting the expected water main failures using four different asset attributes. Therefore, this research can be implemented in asset management best practices and in LoS performance measures to predict the number of water pipeline failures. To further improve the prediction model, additional explanatory variables could be considered along with leveraging multiple artificial intelligence tools.

Keywords: water pipelines, asset management, levels of service, prediction, evolutionary polynomial regression (EPR).

INTRODUCTION

Water infrastructure consists of intensive capital assets, preserved through operation and maintenance, to meet customers' expectations and limit future failures (Kaddoura et al. 2019). Several municipalities are required to meet the minimum expected levels of service (LoS) in order to align with the strategic goals of the councils or local agencies. According to the American Society of Civil Engineers (ASCE) (2017), the condition grade of the US drinking water infrastructure is D and that around 240,000 water main breaks occur each year. In Canada , roughly 30% of the linear drinking water system is in fair to very poor conditions (Canadian Infrastructure Report Card [CIRC] 2019). These assets are prone to continuous deterioration due to physical, environmental, and operational factors.

Many of the municipalities lack comprehensive data regarding their existing infrastructure, which can be accomplished in proper asset management plans (CIRC 2019). The International Organization for Standardization (ISO) 55000 (2014) stated that inventory information and data collection of infrastructure conditions are primary inputs in asset management framework.

Inventory data, in linear water infrastructure, can be in the form of a list of water pipes that includes certain attributes such as age, material type, diameter size, length, pipeline classes, etc. As per ISO 55000 (2014), condition assessment in asset management is accomplished by condition field inspections and condition monitoring.

Currently, there are several field inspection techniques utilized in assessing linear water infrastructure. Generally, tools and techniques that provide higher accuracy outputs, in detecting anomalies, are costly when compared to external acoustic techniques. Due to the expensive direct and indirect costs associated with advanced field inspections, many utilities and municipalities adopt predictive models as a screening tool for condition data collection (American Water Works Association [AWWA] 2019). Therefore, it is vital to rely on robust and reliable predictive tools to avoid misleading information. Cost-effective and accurate predictive tools are capable of predicting future breaks of water pipes by considering historical and available records to lessen field condition assessment inspections. Since many studies rely on expert-based attributes and sub-attributes to rate the condition of water pipes, it is of a great significance to develop reliable data-driven forecasting tools to predict water breaks (Robles-Velasco *et al.*, 2020) considering distinct attribute data.

In an effort to respond to this arising need, this paper utilized the Evolutionary Polynomial Regression (EPR) method to predict the number of water pipe breaks using multiple explanatory variables. Although Kakoudakis et al. (2017) utilized an integrated K-Means clustering and the EPR method in predicting water pipeline failure rates, the authors only considered one material type and concentrated on small diameter sizes (50 – 300 mm), which impacted the holistic view in determinin pipe failures in water networks that include multiple material types and diameter sizes.

This paper, however, provides improved scalable predictive water main asset management tool by 1) performing a Best Subset Regression to understand the most critical explanatory variables in predicting failures; 2) developing a holistic prediction model that considers small, medium, large water mains and different material types; and 3) studying the impact of different variables on the predicted number of breaks through detailed sensitivity analysis. The approach used historical failures and asset data such as the pipe diameter, material type, age, and length. Since this type of information is commonly found in as-built drawings and geodatabases in municipalities, the application of this model shall be scalable. Thus, this research will help municipalities to conclude rational decisions about future interventions by understanding the state of linear water infrastructure and therefore, enhance existing LoS.

BACKGROUND

Parameters Used in Watermain Failure Prediction

The literature focused on 1) factors utilized in predicting the failure rate of water pipelines; and 2) models used for predicting failure rates. The factors utilized in predicting the failure rate of water pipelines were classified into two groups based on 1) whether these factors are static or dynamic through the lifecycle of water pipelines; and 2) whether these factors are physical, environmental, or operational. The failure rate models are classified into four groups: deterministic, statistical, probabilistic, and artificial intelligence. Stone et al. (2002) categorized factors contributing to the failure of water pipelines into two groups: static factors and dynamic factors. Static parameters include the diameter, length, soil type, pipe material, etc. On the other hand, the age, cumulative number of breaks, soil corrosivity and water pressure are examples of dynamic factors affecting the pipe failure rate.

InfraGuide (2003) classified the factors contributing to the deterioration of water pipes to three main categories:

- Physical (i.e. pipe material, pipe wall thickness, pipe age, pipe vintage, pipe diameter, etc.);
- Environmental (i.e. pipe bedding, trench backfill, soil type, groundwater, climate, pipe location, etc.); and
- Operational (i.e. internal water pressure, transient pressure, leakage, water quality, etc.).

The impact of physical factors was examined by several researchers (Berardi et al. 2008, Wang et al. 2009, Xu et al. 2011, Aydogdu and Firat 2014, Arsénio et al. 2014, Jenkins et al. 2014, Kutyłowska 2015, and Lin and Yuan 2019). Moliga et al. (2008) and Shirzad et al. (2014) added additional parameters from various categories as the independent variables to improve the reliability of their models. Additional efforts were implemented to assess the impact of physical and environmental factors on the failure rate prediction models of water mains (Asnaashari et al. 2013, Nishiyama and Filion 2014, Francis et al. 2014, Kabir et al. 2015a, Kimutai et al. 2015, Kabir et al. 2015b, Balekelayi and Tesfamariam 2019, and Snider and McBean 2020). Some others included all three parameters to improve the effectiveness and robustness of the failure rate prediction models (Jafar et al. 2010, Wang et al. 2010, Kabir et al. 2015c, and Zhang et al. 2018).

The summary of all the aforementioned studies is shown in Table 1. The table display the frequency of parameters which were used in the 26 different previous works, including industry and academia, for each category (physical, environmental and operational). Typically, the development of a prediction model is highly dependent on the available information, significance, and reliability of the data. In general, physical factors are the most commonly available

information observed in the databases of the agencies and municipalities. These parameters are mostly extracted from design or shop drawings and are stored in geodatabases. Even though the researchers adopted one or combined multiple factors in developing a prediction model, all researchers used the physical factors as their primary independent variables. By examining these results closely, 12 adopted at least one environmental factor, and nine studies utilized at a minimum one operational factor.

Kimutai et al. (2015) confirmed that physical factors are more critical in estimating the failure rate than environmental factors. Based on the table, the majority of the considered parameters used in this type of prediction were the physical parameters followed by the environmental factors and operational factors, respectively. In specific, the age parameter was significantly considered in the reviewed literature as it demonstrates the time of exposure of the asset, which best explains its deterioration (Folkman 2018). The other most frequent factors utilized in previous studies were the diameter, length, soil type, and pipe material. Berardi et al. (2008) stated that the diameter and length are the most important variables in describing water pipe failure occurrence. Also, Wang et al. (2009) concluded that the length has a great impact on water pipe's failure. Thus, in this study, the major physical factors like age, diameter, length, and pipe material were considered as the independent variables to predict the number of breaks of water pipelines.

[Table 1 near here]

[Figure 1 near here]

Models Utilized in Watermain Failure Prediction

Many researchers developed different models to predict the failure rate of water pipes to understand the state of the infrastructure. These failure prediction models are classified into four

categories; deterministic, statistical, probabilistic, and artificial intelligence (AI) models. A summary of the reviewed models is shown in Table 2. Deterministic models are usually used in cases where the relationship between inputs and outputs is explicit (Bakry et al. 2016a, and Bakry et al. 2016b). The deterministic models can be applied in two approaches: empirical and mechanistic. The empirical approach tries to find the relation between failure rates as the output and attributes of a group of pipes as the inputs. On the contrary, the mechanistic approach can only forecast the remaining useful life of an individual asset.

Statistical models are typically used to predict the useful life or time to failure of infrastructure assets (Lawless 1983). This type of models is applied to homogeneous groups of pipes or other infrastructure assets and require historical failures or data regarding the asset's condition. In regression, the dependent variable is related to at least one of the independent variables. Probabilistic models, however, analyze the probability of an event occurring (Wilson et al. 2017). The probability of occurrence is one and the probability of the event that cannot happen is zero (Mitrani 1998). The information about asset conditions and attributes are required to develop a probabilistic model. The output would be a range of values instead of a specific number. These models need extensive data (Clair and Sinha 2014) which will increase the computational complexity of the models (Moglia 2007 and Wilson et al. 2017).

In AI, the artificial neural network (ANN) is a method that is also utilized to predict pipe failure and deterioration of infrastructure, especially buried pipes. This technique is able to process the information even under large, complex, and uncertain environment. The high-quality database is needed for supervised training and forecasting the future condition of the pipes. ANN is considered a black-box technique and is unable to provide insight into the relationship between the dependent

and independents variables (Clair and Sinha 2014; Moselhi and Hegazy 1993, Atef et al. 2015, and Shirzad et al. 2014, Wilson et al. 2017).

Based on the literature, many of the available tools are either complex or significantly simple. Additionally, some models relied on a specific era of construction and focused on distribution mains (pipelines smaller than 400 mm). Transmission mains are known of their importance and in many cases, their redundancy is significantly minimal in water networks. Also, their cost of failure is significantly high when compared to distribution mains. Therefore, a more generic prediction model is needed which includes the majority of rigid and flexible materials with different diameters. By including a variety of pipe sizes, this study ensured that the developed EPR model is able to predict distribution and transmission main failures.

[Table 2 near here]

Evolutionary Polynomial Regression

EPR technique was first presented by Giustolisi and Savic (2009). This technique utilizes the huge potential of conventional numerical regression techniques and the strength of the Genetic Algorithm (GA) in solving optimization problems (Xu et al. 2011). Later, this approach was used by other researchers in several engineering fields. Savic et al. (2006) and Ugarelli et al. (2009) used EPR to model sewer pipe failures. Rezania et al. (2008) utilized the EPR methodology to evaluate the uplift capacity of suction caissons and shear strength of reinforced concrete deep beams. Elshorbagy and El-Baroudy (2009) compared the EPR and Genetic Programming to develop the prediction model of soil moisture response. Further, Guistolisi and Savic (2009) tested the EPR-Multi Objective Genetic Algorithm (MOGA) (an improved EPR) to develop a groundwater level prediction model based on monthly rainfall. El-Baroudy et al. (2010) utilized the EPR to develop the evapotranspiration process then compared the efficiency of EPR to

Artificial Neural Networks (ANNs) and Genetic Programming (GP). Markus et al. (2010) applied EPR, ANNs and the naive Bayes model to forecast weekly nitrate-N concentrations at a gauging station. Ahangar-Asr et al. (2011) applied EPR to predict mechanical properties of rubber concrete. Fiore et al. (2012) used EPR to provide the predicting torsional strength model of reinforced concrete beams. Costa et al. (2020) implemented the EPR to model the flow duration curves and to enhance the water resrouces planning and management in ungauged sites.

EPR is data-driven technique and can be classified as a grey-box method according to the color coding classification system, which categorizes mathematical models based on the existence of necessary information into three groups; white-box models, black-box models and grey-box models (Giustolisi 2004). The process of establishing the symbolic expressions contains two stages. In the first stage, the EPR tries to find the best model structure using MOGA. Then, the appropriate values for the constants are estimated using the Least-Squares (LS) optimization (Berardi et al. 2008). In EPR-MOGA, seven assumed structures of expression are available and the best case, according to the prior knowledge about the nature of the output, can be selected by the user. The seven structures are as follows (Berardi et al. 2008):

$$Y = a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \dots (X_k)^{ES(j,k)} \cdot f((X_1)^{ES(j,k+1)} \dots (X_k)^{ES(j,2k)}) \quad (1)$$

$$Y = a_0 + \sum_{j=1}^{m} a_j \cdot f((X_1)^{ES(j,1)} \dots (X_k)^{ES(j,k)}) \quad (2)$$

$$Y = a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \dots (X_k)^{ES(j,k)} \cdot f((X_1)^{ES(j,k+1)}) \dots f((X_k)^{ES(j,2k)}) \quad (3)$$

$$Y = \log\left(a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \dots (X_k)^{ES(j,k)}\right) \quad (4)$$

$$Y = \exp\left(a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \dots (X_k)^{ES(j,k)}\right) \quad (5)$$

$$Y = \sin\left(a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \ldots (X_k)^{ES(j,k)}\right) \quad (6)$$

$$Y = \tan\left(a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \ldots (X_k)^{ES(j,k)}\right) \quad (7)$$

Where X is the explanatory variable and $j = [1,\ldots,k]$; *ES* is the matrix of unknown exponents to be defined by the user; *f* is the inner function selected by the user (can be no function, logarithm, exponential, tangent hyperbolic, or secant hyperbolic); a_j represents the unknown polynomial coefficients; *m* is the number of polynomial terms; and a_0 is the bias term. It should be noted that the zero value should be considered in the matrix of the exponent to make the EPR able to remove some variables, which are not powerful enough to predict the output, from the returned expressions.

During the modelling phase, EPR tries to return several equations based on the accuracy and parsimony of the models. The model parsimony can be implemented by optimizing the number of terms, the number of independent variables, or both strategies, where each of these options can be selected by the user. Furthermore, the user can force EPR to generate the equations with an only positive values of constant coefficients ($a_j > 0$). Also, the maximum number of terms in every equation in each run can be specified by the user. In addition, the normalization (if required) can be accomplished by EPR; therefore, the user needs to specify the range in which the inputs or output should be scaled. The EPR can develop a model to forecast the output based on either one input or several inputs. In other words, it can construct Multi Input Single Output (MISO) and/or Single Input Single Output (SISO) models. It should be noted that the limited missing data point can be assumed using linear interpolation by the model (Costa et al. 2020); thus, the model can be developed with an incomplete historical database. During the generation of symbolic expressions, if the EPR cannot find appropriate combination of terms containing $f(x)$ (as an inner function), it

deselects this function. The accuracy of the model is measured using several statistical metrics including the Sum of Squared Error (SSE), Bayesian Information Criterion (BIC), Mean Squared Error (MSE), Final Prediction Error of Akaike (FPE), Akaike's Information Theoretic (AIC), and the Coefficient of Determination (CoD or R^2) as per the following equations:

$$SSE = \sum_{i=1}^{N} (\hat{y}_i - avg(y))^2 \quad (8)$$

$$MSE = \frac{1}{N}\text{SSE} \quad (9)$$

$$BIC = -2\ln L + d\ln(N) \quad (10)$$

$$FPE = \left(\frac{1+(d/N)}{1-(d/N)}\right) SSE \quad (11)$$

$$AIC = -2\ln L + 2d \quad (12)$$

$$CoD/R^2 = 1 - \frac{\Sigma_n(\hat{y} - y_{exp})^2}{\Sigma_n(y_{exp} - avg(y_{exp}))^2} = 1 - \frac{N}{\Sigma_n(y_{exp} - avg(y_{exp}))^2}\ SSE \quad (13)$$

Where N is the number of samples, $\hat{y}$ is the predicted value from the model; y_{exp} is the actual value from the historical data; d is the dimension of the vector of parameters; L is the maximized value of the likelihood function of the estimated model. Table 3 shows the definition of each parametric. In general, a model is selected based on a maximum R^2 and minimum values of the other measures.

[Table 3 near here]

METHODOLOGY

The main objectives of this research were to develop a model capable of predicting the number of breaks in water pipelines and to conduct a sensitivity analysis using the selected prediction model. The overall flow of this research is summarized in Fig. 1. The research commenced by reviewing the literature and exploring the available prediction models developed in infrastructure assets, and

248 more specifically for water pipelines. The models reviewed were classified into statistical,
249 probabilistic, and AI. Further, the study screened the independent variables used by previous
250 researchers in predicting water pipeline failures. The variables were classified into physical,
251 environmental, and operational. Based on the literature and the completeness of the data collected
252 from the City of Montreal, the Best Subset Regression analysis was performed to determine the
253 variables to be included in the prediction model. The segments were classified into homogenous
254 clusters based on the segments' attributes. The data points were split into 80% for training the EPR
255 models and 20% for testing. Several statistical metrics were used to examine the accuracy of the
256 models to select the most reliable one. A sensitivity analysis was then performed by studying the
257 impact of altering the material types, diameter sizes, and lengths on the number of breaks.

258 [Fig. 1 near here]

259 **DATA COLLECTION**

260 The model was developed and deployed on the City of Montréal's water infrastructure network.
261 The City has a population of 1.8 million, and its land area is around 365.1 km^2. There are six
262 water treatment plants and 14 reservoirs used to produce and supply water that is transported over
263 10 pressure zones (Fig. 2). The City owns 5,045 kilometers of water linear assets, containing
264 4,305 km distribution pipes and 740 km transmission pipes. The original dataset of the City
265 included different asset attributes but internal operating pressures of the system were not
266 available. However, it is expected that the operating pressures are between 21 m (30 psi) and 70
267 m (100 psi) (Ghorbanian et al. 2016). The average age of the installed pipelines was
268 approximately 55 years and the oldest was roughly 162 years. Most of the large pipelines were
269 installed using pressure concrete cylinders.

270 The dataset consisted of 56.55% Cast Iron (CI), 26.61% Ductile Iron (DI), 10.47% Cementitious
271 (Asbestos Cement and Concrete Cylinder), 5.54% Plastic pipes (Polyvinyl Chloride [PVC] and

Polyethylene), 0.77% Steel, 0.05% Copper, and 0.01% Galvanized Iron (GI). According to the dataset, CI pipes were firstly used in the 1860s and DI pipes were mostly installed in the 1960's.. Although operational changes may have occurred in the City during the different periods, only deterioration-related failures were considered in building the prediction model.

The City started to perform a systematic recording of pipe failures since 1972, where the dataset contained a total of 22,735 pipe breaks. Fig. 3 shows the number of breaks in the water distribution network of pipelines installed between 1861 and 2015. By examining the data, the number of breaks for CI pipes has steadily increased since 1986 and reached the peak in 2001-2005 interval, before falling slightly during the recent 15 years. Also, Fig. 3 illustrates that the number of breaks of DI pipes installed between 2001 and 2015 was significant compared to other cohorts. Ferrous pipelines have seen major transitions from CI to DI pipelines. Through this transition, the mechanical and physical properties were drastically changed, which affected the overall performance of the pipes. One of these major changes is related to the wall thickness dimensions. In DI pipes, Thickness Class design is thicker as opposed to Pressure Class design for the same diameter (Ductile Iron Pipe Research Association [DIPRA] 2016). As Pressure Class pipes have a thinner wall thickness, it is expected to deteriorate at a higher rate when compared to thicker pipes of the same size. Although the Pressure Class standard was introduced to the industry before 2000s, the industry required a period of time to incorporate the newer type of DI material. Therefore, not all agencies immediately installed DI Pressure Class after the introduction of the same standard.

In the City, most of the small plastic pipelines were installed using PVC but were not significantly popular in the City until the early 1990s. PVC pipelines' installation increased significantly after the 2000s; yet, the inventory of this type dominates only 5.54% of the water infrastructure. Folkman (2018) reported that the majority of PVC failures (35%), in North America, occurred in

pipes installed in the 1980s. The author also concluded that the failures observed in PVC pipes installed in the 1990s, 2000s, and 2010s were 23%, 21%, and 9%, respectively. Given the minimal installations of PVC pipes in the City of Montreal prior to the 2000's and that higher failure percentages were observed post-2000s in North America, the City's PVC failures would roughly match the findings reported in Folkman (2018).

[Fig. 2 & 3 near here]

MODEL IMPLEMENTATION

Best Subset regression is implemented to recognize the most critical factors for predicting a number of breaks for water pipelines from the collected data. The best subset regression was deployed by using Minitab 17 statistical package. The results of the best subset regression were evaluated using common statistical parameters. These parameters were the coefficient of determination (R^2), square root of mean square error (S), and Mallows' coefficient (Cp) (Bakry et al. 2015a; Bakry et al. 2015b). The R^2 values provided insights regarding the capabilities of the model in fitting the data, which ranges between 0 and 1. Models that have significant capability in fitting the data are closer to 1 (Lee and Derrible 2020). The values of S, however, measure the distance between actual values and the response of the developed model. Values that are closer to 0 indicate minimal errors and would describe a reliable model. In order to determine the best balance of the number of predictors in the model, Cp values are used. Models that generate Cp values closer or equal to the number of independent variables plus one are good candidates for selection (Bakry et al. 2015a and Bakry et al. 2015b).

In this research, the Best Subset Regression considered four different variables: length, diameter, age, and material types. These independent factors were the mostly considered variables in predicting failure rates (refer to Table 1). Although the soil type utilization's frequency surpassed

the material type, soil type information was unavailable in the collected data and therefore, was not considered in the model development. In case additional parameters (such as soil type and pressures) to be included in the analysis, the Best Subset Regression should be completed to determine the most critical variables to be considered in the EPR implementation.

As per Table 4, models 5 and 7 had the highest value of R^2, adjusted R^2, and predicted R^2 (68.9%, 68.9%, and 68.1%). The values of S for model 5 and 7 were 24.286 and 24.289, respectively. Also, the values of Cp for the same two models were 3 and 5, respectively. By comparing the two models' Cp values, model 7 was selected as the value was equal to the number of variables plus one. Although the S value of model 7 was higher, the difference was minimal.

After performing the best subset regression and selecting the independent variables, the dataset was further classified into several homogeneous groups, based on age, diameter, and pipe material. The objective of this classification was clustering pipe segments into classes with the same attributes. There were several categories within the same age, diameter, and material for each dataset. Age was selected to take the indirect effect of time-varying solicitation on water mains, since from an engineering point of view, the higher the time of exposure, the higher the chemical and mechanical effects on pipes (Kaddoura et al. 2019*)*. These effects can be caused by several factors such as soil condition, traffic loads, ground movements, etc. (Kaddoura et al. 2019).

After identifying the most critical variables, the EPR model was processed. The EPR generated twelve symbolic expressions, which were used to predict the number of breaks for water pipes in the City of Montréal. The dataset was randomly divided into two subsets: 1950 (80%) samples were used for training and 486 (20%) samples were used for testing. Table 5 shows these expressions and their related R^2 values. In the table, L, D, A, and M represent the length, diameter,

age, and material of the water pipes, respectively. The left side shows the output which was the number of breaks. Among all generated symbolic expressions, the best model should be chosen based on the fitness to the historical data and the parsimony of the equation. In this study, model 10 was selected as it provided the highest R^2 (89.35%).

The other statistical metrics such as SSE, BIC, MSE, FPE, AIC, and GCV of all models are shown in Table 5. These metrics were used to confirm the selection of the prediction model. The model that generated the minimum values of all metrics, in Table 5, was selected. By comparing the models' values, the minimum values corresponded to model 10. Along with the R^2 value, model 10 metrics confirmed that it was the best in predicting the number of breaks. Fig. 4 shows the Pareto graph of the expressions that were generated based on the dataset. In this graph, each point represented a generated symbolic expression. In EPR, the Pareto frontier includes the best combination of the number of constants and variables and the complexity of the model. The selected model (model 10) was specified by the black arrow. The horizontal axis shows the value of ($1-R^2$) or (1-CoD) while the vertical axis shows the percentage of the considered vectors in each model (Giustolisi and Savic 2006). As the complexity decrease, the accuracy of the model reduces. Although model 10 was not the least complex, the accuracy was the highest which is the main concern of decision-makers (Kaddoura et al. 2018).

[Table 5 near here]

[Fig. 4 near here]

DISCUSSION OF RESULTS

The Best Subset Regression offered an opportunity to understand the critical variables in predicting the number of breaks in water mains. Further, one EPR model showed higher statistical accuracy when compared to the others. The analysis demonstrated that the age, diameter, length and pipeline

365 material are important factors in establishing a reliable prediction model for water pipeline failure.
366 While this aligns with many researchers' conclusions (refer to Table 1), this research considered
367 small and large diameters as well as different material types that could be found in many different
368 North American Municipalities.
369 A sensitivity analysis was performed to better understand the impact of the variables on the number
370 of breaks and to further discuss the results. Fig. 5 shows the effect of diameter, length, and pipe
371 material on the number of breaks, respectively, as the pipe ages. In each figure, a factor was
372 changed while the rest remained constant. In these graphs, the vertical axis shows the number of
373 breaks and the horizontal axis represents the age of the pipe. The three graphs demonstrate the
374 increase in the number of future breaks with the increase of age, which confirms with the research
375 conducted by Berardi et al. (2008), Wang et al. (2009), Xu et al. (2011) and many others.
376 According to Fig. 5a, smaller diameter pipelines tend to have a higher number of breaks when
377 compared to larger pipelines. This resulted due to the enormous number of breaks observed in the
378 distribution mains of the data collected. A significant number of breaks were reported between
379 100 mm and 375 mm pipelines. In the water pipeline industry, wall thicknesses of smaller pipes
380 are thinner than larger ones. Due to ageing and minimal maintenance activities, the soil-pipeline
381 interaction in a corrosive environment will have a direct effect on reducing the pipelines'
382 thicknesses and the residual factor of safety (Kaddoura et al. 2019). Additionally, Fig. 5b shows
383 that the number of breaks for longer pipes is higher than shorter ones, as longer pipes will be
384 subjected to bending stresses more than shorter pipes. Given these results, the length and diameter
385 sensitivity analyses align with the previous findings concluded by Berardi et al. (2008) but with a
386 generic model that includes distribution and transmission mains.
387

Fig. 5c shows the sensitivity analysis for different pipe types. The pipes were divided into six clusters and were represented numerically from 1 to 6. The clusters were determined based on the failure rate history. Group 1 was assigned to pipes with the lowest historical failure rate, while Group 6 represented pipes with the highest historical failure rates. According to the figure, the most sensitive pipeline material to age was the Steel pipe, while the lowest one was PVC. PVC pipelines, in water distribution mains, showed the lowest failures during their service life (Folkman 2014). Since they are thermoplastic material, these pipelines are not subjected to corrosion as ferrous materials.

Ferrous materials tend to have extensive failures during their service lives when compared to other types. When pipelines are exposed to deterioration mechanisms such as corrosion, the probability of failure increases. Ferrous pipelines predominate the majority of pipelines' breaks in North America (Folkman 2018). Breaks in ferrous pipes occur due to many factors including soil corrosivity and reduced wall thicknesses (Barton et al. 2019). The corrosivity of soil relies on factors such as the particle size, acidity, moisture content, and electrical conductivity (Ferreira et al. 2007 and Barton et al. 2019). Pipelines that have reduced intervention activities during their service lives will be prone to deterioration and reduced wall thicknesses. The reduction in wall thicknesses will decrease the factor of safety values and hence, cause failures (Kaddoura et al. 2019).

AC pipelines predicted number of breaks were also observed to be in the third cluster (the third highest), as these pipelines have lower vulnerabilities to corrosion when compared to ferrous material (Barton et al. 2019). The results also produced lower predicted number of breaks due to the following 1) AC pipeline material was no longer utilized in the North American water infrastructure in the 1980s (Folkman 2018); and 2) the failure records of AC pipelines were interrupted due to many replacement initiatives being considered in many jurisdictions in North

America, given the health-related consequences of these pipes (Ma et al. 2017). However, the number of break frequency pattern of these mains, in the City of Montreal, confirms with the findings of Folkman (2018) latest survey (e.g. the frequency of AC failures increases with age).

[Fig. 5 near here]

Furthermore, the sensitivity of the input factors was studied and shown in Fig. 6. The figure shows the effect of changing the input factors on the number of breaks of water pipelines. This graph was developed to understand the interrelationship between the number of breaks and its input factors and to determine the most sensitive independent variable. The vertical axis represents the number of breaks (logarithmic scale), whereas the horizontal axis represents the normalized value of each factor. However, the actual values of each factor were listed below the normalized values for a better representation. The corresponding values of the number of breaks were also mentioned in the same figure. The results shows a direct relationship between the age and length with the number of breaks (e.g. the number of breaks increases when the age and length of the pipe increase). Also, the figure shows an inverse relationship between the pipe's diameter and the number of breaks of the pipes (e.g. the number of breaks increases when the pipe diameter decreases). By comparing the impact of different parameters on water main breaks, this research showed that breaks are mostly sensitive to pipeline diameter followed by material, length, and age.

[Fig. 6 near here]

CONCLUSIONS

Water pipelines are an essential component as they supply potable water to customers. Due to the external and internal environment, these assets are subjected to deterioration. In many of the North American cities, the number of breaks is one of the most decisive LoS performance measures. As utilities are concerned with maintaining minimum LoS, reducing the number of breaks is a major

duty for municipalities. This research developed a prediction model by integrating the Best Subset Regression and EPR methods to predict the number of water pipeline breaks considering small to large pipelines (100 mm to 2000 mm) and different material types (thermoplastic, concrete pressure pipes, and ferrous). Based on the findings, both methods are applicable in predicting failures for asset management practices.

The tool relied on data collected from the City of Montreal. After using the Best Subset Regression, four parameters (age, diameter, length, and material type) were selected to predict the watermain breaks. Pipelines were classified based on clusters related to the asset attributes. Based on the generated models, model 10 provided higher accuracies, where the supplied R^2 (89.35) of the same model was the highest. According to the findings, ferrous pipelines were one of the most sensitive materials to age. In addition, the research showed that larger diameters experience a reduced number of breaks when compared to smaller diameter pipelines. By comparing the length, material and diameter parameters, the latter showed the highest sensitivity.

CONTRIBUTIONS TO THE BODY OF KNOWLEDGE

Based on this study, this research offers the following contribution to the body of knowledge:

- Integrating the Best Subset Regression and EPR methods in developing a prediction model.
- Developing a prediction model for water pipelines by considering small, medium, and large pipeline diameter sizes and networks.
- Considering different material types that are expected to represent the inventory of the water network such as thermoplastic, ferrous, and concrete pressure material types.
- Performing statistical analysis through the Best Subset Regression to understand the most critical attributes in predicting water pipeline failures. The research highlighted the

importance of keeping comprehensive asset attribute information by reducing gaps in age, diameter, and material type information of the pipelines.

- Establishing a sensitivity analysis of the utilized explanatory variables to understand the expected changes in the predicted number of breaks. The analysis helped relating the material variations and standards with the sensitivity analysis outcomes.

After successfully adopting the EPR method in predicting the watermain failures, this research is expected to aid decision-makers in planning for future interventions to reduce future predicted breaks and hence, increase the LoS measures in water infrastructure by relying on physical factors.

RECOMMENDATIONS FOR FUTURE RESEARCH

Although the present model supplied reliable prediction model, its quality and reliability is expected to enhance by:

- Considering other excluded factors (soil corrosivity, pressures, external loadings, etc.). However, including additional factors will highly depend on the availability of data and their statistical significance.
- Testing this model on different water networks to investigate its capabilities in case operational activities and climate differ.
- Developing a failure prediction model for rehabilitated water mains using trenchless technology applications.
- Utilizing multiple artificial intelligence models and comparing their performance to conclude the most reliable tool.

DATA AVAILABILITY STATEMENT

Some or all data, models, or code used during the study were provided by a third party. Direct requests for these materials may be made to the provider as indicated in the Acknowledgements. The data is related to water pipelinew' attributes along with historical failures.

ACKNOWLEDGMENTS

The authors would like to thank the City of Montreal for providing the data which was utilized in developing the prediction model.

REFERENCES

Ahangar-Asr, A., Faramarzi, A., Javadi, A. A., & Giustolisi, O. (2011). Modelling mechanical behaviour of rubber concrete using evolutionary polynomial regression. Engineering Computations, 28(4), 492-507. doi:10.1108/02644401111131902

Akaike, H. (1974). "A new look at the statistical model identification." In Selected Papers of Hirotugu Akaike, 215-222, Springer, New York, NY.

Allen, D. M. (1971). "Mean square error of prediction as a criterion for selecting variables." Technometrics, 13(3), 469-475.

Arsénio, A. M., Dheenathayalan, P., Hanssen, R., Vreeburg, J., & Rietveld, L. (2014). Pipe failure predictions in drinking water systems using satellite observations. Structure and Infrastructure Engineering, 1-10.

ASCE. (2017). 2047 Infrastructure Report Card. <https://www.infrastructurereportcard.org/> (Oct 24, 2019)

Asnaashari, A., McBean, E. A., Gharabaghi, B., & Tutt, D. (2013). Forecasting watermain failure using artificial neural network modelling. Canadian Water Resources Journal, 38(1), 24-33.

Atef, A. (2015). Asset management tools for municipal infrastructure considering interdependency and vulnerability (Degree of Doctor of Philosophy).

AWWA. (2019). M77 Condition Assessment of Water Mains. AWWA.

Aydogdu, M., & Firat, M. (2014). Estimation of failure rate in water distribution network using fuzzy clustering and LS-SVM methods. Water Resources Management, 29(5), 1575-1590.

Bakry, I., Alzraiee, H., Kaddoura, K., El Masry, M., and Zayed, T. (2016a). "Condition prediction for chemical grouting rehabilitation of sewer networks." Journal of Performance of Constructed Facilities, 30(6), 04016042.

Bakry, I., Alzraiee, H., Masry, M. E., Kaddoura, K., and Zayed, T. (2016). "Condition prediction for cured-in-place pipe rehabilitation of sewer mains." Journal of Performance of Constructed Facilities, 30(5), 04016016.

Balekelayi, N., & Tesfamariam, S. (2019). Geoadditive Bayesian regression models for water mains failure rate prediction. Proceeding of 13th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP13

Seoul, South Korea.

Berardi, L., Kapelan, Z., Giustolisi, O., & Savic, D. (2008). Development of pipe deterioration models for water distribution systems using EPR. Journal of Hydroinformatics, 10(2), 113-126.

CIRC. (2019). Informing the Future: Assessing the Health of Our Communities Infrastructure. <http://canadianinfrastructure.ca/en/index.html> (Nov 15, 2019)

Clair, A. M. S., & Sinha, S. (2014). Development of a standard data structure for predicting the remaining physical life and consequence of failure of water pipes. Journal of Performance of Constructed Facilities, DOI: 10.1061/(ASCE)CF.19435509.0000384.

Costa, V., Fernandes, W., & Starick, Â. (2020). Identifying Regional Models for Flow Duration Curves with Evolutionary Polynomial Regression: Application for Intermittent Streams. Journal of Hydrologic Engineering, 25(1), 04019059.

Creighton, J. H. (2012). A first course in probability models and statistical inference Springer Science & Business Media, QA273.C847.

Demissie, G., Tesfamariam, S., & 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 Engineering, 3(4), 04017017.

DIPRA. (2016). "Design of Ductile Iron Pipe". Birmingham, AL.

El-Baroudy, I., Elshorbagy, A., Carey, S., Giustolisi, O., & Savic, D. (2010). Comparison of three data-driven techniques in modelling the evapotranspiration process. Journal of Hydroinformatics, 12(4), 365-379.

Elshorbagy, A., & El-Baroudy, I. (2009). Investigating the capabilities of evolutionary data-driven techniques using the challenging estimation of soil moisture content. Journal of Hydroinformatics, 11(3-4), 237-251.

Ferreira, C. A. M., Ponciano, J. A., Vaitsman, D. S., & Pérez, D. V. (2007). Evaluation of the corrosivity of the soil through its chemical composition. Science of the total environment, 388(1-3), 250-255

Fiore, A., Berardi, L., & Marano, G. C. (2012). Predicting torsional strength of RC beams by using evolutionary polynomial regression. Advances in Engineering Software, 47(1), 178-187.

Folkman, S. (2014). PVC Pipe Longevity Report: Affordability and the 100+ Year Benchmark Standard.

Folkman, S. (2018). "Water main break rates in the USA and Canada: A comprehensive study". Utah State University. <https://digitalcommons.usu.edu/mae_facpub/174/> (Nov 15, 2019)

Francis, R. A., Guikema, S. D., & Henneman, L. (2014). Bayesian belief networks for predicting drinking water distribution system pipe breaks. Reliability Engineering & System Safety, 130, 1-11.

Ghorbanian, V., Karney, B., & Guo, Y. (2016). Pressure standards in water distribution systems: reflection on current practice with consideration of some unresolved issues. Journal of Water Resources Planning and Management, 142(8).

Giustolisi, O. (2004). Using genetic programming to determine chezy resistance coefficient in corrugated channels. Journal of Hydroinformatics, 6, 157-173.

Giustolisi, O., & Savic, D. (2009). Advances in data-driven analyses and modelling using EPR-MOGA. Journal of Hydroinformatics, 11(3-4), 225-236.

InfraGuide. (2003). Sustainable Municipal Infrastructure Principles and Guidelines. <http:// www.infraguide.ca> (Jan 13, 2016)

ISO. (2014). Asset management — Overview, principles and terminology.

Jafar, R., Shahrour, I., & Juran, I. (2010). Application of artificial neural networks (ANN) to model the failure of urban water mains. Mathematical and Computer Modelling, 51(9), 1170-1180.

Jenkins, L., Gokhale, S., & McDonald, M. (2014). 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.

Kabir, G., Demissie, G., Sadiq, R., & Tesfamariam, S. (2015a). Integrating failure prediction models for water mains: Bayesian belief network based data fusion. Knowledge-Based Systems, 85 (2015) 159–169.

Kabir, G., Tesfamariam, S., & Sadiq, R. (2015c). Predicting water main failures using bayesian model averaging and survival modelling approach. Reliability Engineering & System Safety, 142, 498-514.

Kabir, G., Tesfamariam, S., Francisque, A., & Sadiq, R. (2015b). Evaluating risk of water mains failure using a bayesian belief network model. European Journal of Operational Research, 240(1), 220-234.

Kaddoura, K., Mady, R., & Lalonde, A. (2019). "ASSESSMENT APPROACH TO EVALUATE THE CONDITIONS OF DUCTILE IRON (DI) WATER DISTRIBUTION PIPELINES." Proc. In CSCE 2019 Annual Conference, Canada.

Kaddoura, K., Zayed, T., & Hawari, A. H. (2018). Multiattribute utility theory deployment in sewer defects assessment. *Journal of Computing in Civil Engineer*ing, 32(2), 04017074.

Kaddoura, K., & Zayed, T. (2018). An integrated assessment approach to prevent risk of sewer exfiltration. *Sustainable cities and society*, *41*, 576-586.

Kakoudakis, K., Behzadian, K., Farmani, R., & Butler, D. (2017). Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K-means clustering. *Urban Water Journal*, *14*(7), 737-742.

Kimutai, E., Betrie, G., Brander, R., Sadiq, R., & Tesfamariam, S. (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, DOI: 10.1061/(ASCE)PS.1949-1204.0000196.

Kuroki, Y., & Matsui, T. (2009). "An approximation algorithm for multidimensional assignment problems minimizing the sum of squared errors." Discrete Applied Mathematics, 157(9), 2124-2135.

Kutyłowska, M. (2015). Neural network approach for failure rate prediction. Engineering Failure Analysis, 47, 41-48.

Lamarre, J. (2018). Drinking Water Network Optimization of Montreal. Proceedings of WDSA/CCWI Joint Conference Proceedings, 1, Kingston, ON.

Lawless, J. (1983). Statistical methods in reliability. Technometrics, 25(4), 305-316.

Lee, D., & Derrible, S. (2020). Predicting Residential Water Demand with Machine-Based Statistical Learning. Journal of Water Resources Planning and Management, 146(1), 04019067.

Lin, P., & Yuan, X. X. (2019). A two-time-scale point process model of water main breaks for infrastructure asset management. Water research, 150, 296-309.

Ljung, L. (1999). System Identification: Theory for the User, Upper Saddle River, NJ, Prentice-Hall PTR.

Ma, C. J., & Kang, G. U. (2017). Actual Situation of Asbestos in Tract Drinking-Water in Korean and Japanese Local Cities. Water, Air, & Soil Pollution, 228(1), 50.

Markus, M., Hejazi, M., Bajcsy, P., Giustolisi, O., & Savic, D. (2010). Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in illinois. Journal of Hydroinformatics, 12(3), 251-261.

Mitrani, I. (1998). Probabilistic modelling Cambridge University Press.

Moglia, M., Davis, P., & Burn, S. (2008). Strong exploration of a cast iron pipe failure model. Reliability Engineering & System Safety, 93(6), 885-896.

Moselhi, O., & Hegazy, T. (1993). Markup estimation using neural network methodology. Computing Systems in Engineering, 4(2), 135-145.

Motiee, H., & Ghasemnejad, S. (2019). Prediction of pipe failure rate in Tehran water distribution networks by applying regression models. Water Supply, 19(3), 695-702.

Nishiyama, M., and Filion, Y. (2014). "Forecasting breaks in cast iron water mains in the city of Kingston with an artificial neural network model." Canadian Journal of Civil Engineering, 41(10), 918-923.

Rezania, M., Javadi, A. A., & Giustolisi, O. (2008). An evolutionary-based data mining technique for assessment of civil engineering systems. Engineering Computations, 25(6), 500-517. doi:10.1108/02644400810891526

Robles-Velasco, A., Cortés, P., Muñuzuri, J., & Onieva, L. (2020). Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliability Engineering & System Safety, 196, 106754.

Savic, D., Giustolisi, O., Berardi, L., Shepherd, W., Djordjevic, S., & Saul, A. (2006). Modelling sewer failure by evolutionary computing. Proceedings of the ICE-Water Management, 159(2), 111-118.

Shi, F., Liu, Y., Liu, Z., & Li, E. (2018). Prediction of pipe performance with stacking ensemble learning based approaches. Journal of Intelligent & Fuzzy Systems, 34(6), 3845-3855.

Shirzad, A., Tabesh, M., & Farmani, R. (2014). A comparison between performance of support vector regression and artificial neural network in prediction of pipe burst rate in water distribution networks. KSCE Journal of Civil Engineering, 18(4), 941-948.

Snider, B., & 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.

Stone, M. (1979). "Comments on Model Selection Criteria of Akaike and Schwarz." Journal of the Royal Statistical Society, Series B (Methodological). 41 (2), pp. 276-278.

Stone, S. L., Dzuray, E. J., Meisegeier, D., Dahlborg, A., Erickson, M., & Tafuri, A. N. (2002). Decision-support tools for predicting the performance of water distribution and wastewater collection systems US Environmental Protection Agency, Office of Research and Development. EPA/600/R-02/029.

Ugarelli, R., Kristensen, S. M., Røstum, J., Sægrov, S., & Di Federico, V. (2009). Statistical analysis and definition of blockages-prediction formulae for the wastewater network of oslo by evolutionary computing. Water Science and Technology, 59(8), 1457-1470.

Wang, C., Niu, Z., Jia, H., & Zhang, H. (2010). An assessment model of water pipe condition using bayesian inference. Journal of Zhejiang University SCIENCE A, 11(7), 495-504.

Wang, Y., Zayed, T., & Moselhi, O. (2009). Prediction models for annual break rates of water mains. Journal of Performance of Constructed Facilities, 23(1), 47-54.

Wilson, D., Filion, Y., and Moore, I. (2017). State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. Urban Water Journal, 14(2), 173-184.

Xu, Q., Chen, Q., Li, W., & Ma, J. (2011). Pipe break prediction based on evolutionary data-driven methods with brief recorded data. Reliability Engineering & System Safety, 96(8), 942-948.

Zhang, B., Guo, T., Zhang, L., Lin, P., Wang, Y., Zhou, J., & Chen, F. (2018). Water pipe failure prediction: A machine learning approach enhanced by domain knowledge. In Human and Machine Learning (pp. 363-383). Springer, Cham.

488

Table 1. Physical, Environmental, and Operational Parameters

Factors	Physical Factors									Environmental Factors									Operational Factors					
Researcher(s)	Pipe Material	Pipe Wall Thickness	Pipe Age	Pipe Length	Pipe Vintage	Pipe Diameter	Type of Joint/Number of Thrust Restraint/Connectio	Pipe Lining and	Depth	Bedding	Trench Backfill	Soil Type and Soil Corrosivity Factors	Climate	Pipe Location	Thawing index	Freezing Index	Rain Deficit	Cathodic Protection	Traffic and Loading	Seismic Activity	Internal Pressure and Transient	Leakage	Water Quality	O&M Practices
Moglia et al. (2007)		X	X	X		X															X			
Berardi et al. (2008)			X	X		X																		
Wang et al. (2009)	X		X	X		X			X															
Jafar et al. (2010)	X	X	X	X		X						X		X							X			
Wang et al. (2010)	X		X			X		X		X	X	X							X	X		X		
Xu et al. (2011)			X	X		X																		
Ansaashari et al. (2013)	X		X	X		X		X				X												
Arsenio et al. (2014)			X																					
Shirzad et al. (2014)			X	X		X			X												X			
Aydogdu and Firat (2014)			X	X		X																		
Nishiyama and Filion (2014)			X	X		X																		
Kabir et al. (2014)		X	X	X		X						X							X	X		X		
Jenkins et al. (2014)			X			X																		
Francis et al. (2014)			X									X	X	X										
Kutylowska (2015)	X		X	X		X																		
Kabir et al. (2015a)			X	X	X	X						X		X										
Kimutai et al. (2015)	X			X		X						X	X											
Kabir et al. (2015b)			X	X	X	X						X												
Demmissie et al. (2017)	X	X	X									X												
Kakoudakis et al. (2017)			X	X	X																			
Shi et al. (2018)		X										X												
Zhang et al. (2018)	X		X	X		X			X			X							X					
Lin and Yuan (2019)			X	X		X																		

Balekelayi and Tesfamariam (2019)	X	X	X	X	X		X	X	X	X	X	X		
Motiee and Ghasemnejad (2019)	X	X	X	X									X	
Robles-Velasco et al. (2020)	X	X	X	X	X								X	X
Snider and McBean (2020)			X	X		X	X	X				X	X	

Table 2. Prediction Models of Water Distribution Networks

Authors (Year)	Model Classification	Methodology	Output Type
Moglia et al. (2007)	Probabilistic	Monte-Carlo Simulation Framework	Probability of Failure for CI Pipes
Wang et al. (2009)	Statistical	Five Multiple Regression Models	Annual Break Rates
Li et al. (2009)	Probabilistic	Monte-Carlo Simulation	Remaining Useful Life
Jafar et al. (2010)	Artificial Intelligence	Six ANN Models	Failure Rate
Wang et al. (2010)	Statistical	Bayesian Inference	Deterioration Rate
Xu et al. (2011)	Statistical	Genetic Programming	Deterioration Rate
Osman and Bainbridge (2011)	Statistical	Rate of Failure (ROF) and Transition State (TS)	Deterioration Rate
Asnaashari et al. (2013)	Artificial Intelligence	ANN and Multi Linear Regression	Failure Rate
Arsénio et al. (2014)	Statistical	Ground Movement Estimated by Radar Satellite Data	replacement-prioritization plan
Shirzad et al. (2014)	Artificial Intelligence	ANN and Support Vector Regression (SVR)	Pipe Burst
Aydogdu and Firat (2014)	Artificial Intelligence	Fuzzy Clustering and Least Squares Support Vector Machine (LS-SVM)	Failure Rate
Nishiyama and Filion (2014)	Artificial Intelligence	ANN	Pipe Breaks
Kabir et al. (2014)	Probabilistic	Bayesian Belief Networks (BBN)	Risk of Failure
Jenkins et al. (2014)	Probabilistic	Weibull Hazard	Failure Rate
Francis et al. (2014)	Probabilistic	Bayesian Belief Networks (BBN)	Pipe Breaks
Kutyłowska (2014)	Artificial Intelligence	ANN	Failure Rate
Kabir et al. (2015a)	Statistical	Bayesian Weibull Proportional Hazard Model (BWPHM)	Failure Rate
Kimutai et al. (2015)	Statistical	Weibull proportional hazard model (WPHM), the Cox proportional hazard model (Cox-PHM), and the Poisson model (PM)	Pipe Failure
Kabir et al. (2015b)	Probabilistic	Bayesian Belief Networks (BBN)	Failure Rate
Demissie et al. (2017)	Statisitical	Nayesian Belief Network	Remaining Service Life
Kakoudakis et al. (2017)	Artificial Intelligence		

Shi et al. (2018)	Artificial Intelligence and Statisitical	Multiple Linear Regression, Artificial Neural Network, Random Forest, Support Vector Machine, and Stacking ensemble learning	Pipe Performance
Zhang et al. (2018)	Statisitical	Bayseian nonparameteric	Pipe Failures
Lin and Yuan, (2019)	Probablistic	Markov Chain Monte Carlo	Number of breaks
Balekelayi and Tesfamariam (2019)	Statsitical	Polynomial P-splines and Geoadditive Bayesian Regression	Breakage rate
Motiee and Ghasemnejad (2019)	Statsitical and Probablistic	Linear regression, exponential regression, poisson generalized linear, and logistic generalized linear	Failure Rate
Robles-Velasco et al. (2020)	Probablistic and Artificial Intellegince	Support vector machine and logestic regression	Pipe Failures
Snider and McBean (2020)	Probablistic and Artificial Intellegince	Weibull analysis and Extreme gradient boosting	Time to break

Table 3. Definitions of Statistical Metrics

Metric	Definition	Value Definition	Reference
AIC	The AIC is a metric that can be used to estimate the probability of the model to estimate the predicted value.	Lower values mean a good and accurate prediction model.	(Akaike 1974)
BIC	This metric measures the complexity and fit of the model in predicting values	Lower values mean a Good prediction and less complex model	(Stone 1979)
FPE	This metric measures the trade-off between the quality and the complexity of the model.	Lower values mean an accurate model	(Ljung 1999)
SSE	This metric measures the squared difference between the predicted and the actual values	Lower values mean lower errors generated and hence, it is a reliable model	(Kuroki and Matsui 2009)
MSE	This metric measures the average amount of the squared error between the predicted and the actual value	Lower values mean lower errors generated and hence, it is a reliable model	Allen (1971)
CoD or R^2	This metric measures the percetnage of the variance of predicting the dependent variable	Higher values provides an indication of a fit model.	Bakry et al. 2015a; Bakry et al. 2015b)

Table 4. Best Subset Regression for City of Montréal

Model Number	Variables	R^2	R^2 (adj)	R^2 (pred)	Mallows Cp	S	Length (km)	Diameter (mm)	Material	Age (years)
1	1	68.6	68.6	67.8	22.3	24.39	X			
2	1	1.6	1.5	1.4	5268.7	43.203		X		
3	2	68.8	68.8	68	8.5	24.316	X		X	
4	2	68.8	68.7	68	14.2	24.345	X			X
5	3	68.9	68.9	68.1	3	24.284	X		X	X
6	3	68.8	68.8	68	10.4	24.321	X	X	X	
7	4	68.9	68.9	68.1	5	24.289	X	X	X	X

Table 5. Symbolic Expressions for Montréal dataset and Related Statistics

#	*Expressions*	R^2 (%)	*SSE*	*BIC*	*MSE*	*FPE*	*AIC*
1	$No.of\ Breaks = 3.4446 \times 10^{-5} \times L^{1.5}$	76.90	476.4	478.2	476.6	476.9	476.9
2	$No.of\ Breaks = 1.3197\ \frac{L^{1.5}}{D^2}$	82.51	360.6	362	360.8	361	361
3	$No.of\ Breaks = 0.08546\ \frac{L^{1.5}\ M^2}{D^2}$	85.04	308.5	309.7	308.7	308.8	308.8
4	$No.of\ Breaks = 1.8835\ \frac{L^{1.5}}{D^2}\ \ln\left(\frac{M^2}{A^{0.5}}\right)$	85.37	301.8	303	302	302.1	302.1
5	$No.of\ Breaks = 0.24999\ \frac{L^{1.5}\ A^{0.5}}{D^2}\ ln\left(\frac{M^2}{A^{0.5}}\right)$	86.40	280.5	281.5	280.6	280.7	280.7
6	$No.of\ Breaks = 0.092319\ \times L^{0.5} + 0.23417\frac{L^{1.5}\ A^{0.5}}{D^2}\ \ln\left(\frac{M^2}{A^{0.5}}\right)$	87.04	267.4	269.4	267.6	267.9	267.9
7	$No.of\ Breaks = 0.12036\ \frac{L^{1.5}\ M^2}{D^2} + 4.8297 \times 10^{-7}\ \frac{L^2\ A^{1.5}}{D^2}\ \ln\left(\frac{1}{L}\right)$	88.03	246.9	248.8	247.1	247.4	247.4
8	$No.of\ Breaks = 0.008929\ \frac{L^{1.5}\ A}{D^2}\ \ln\left(\frac{1}{L^{0.5}}\right) + 0.069455\ \frac{L^{1.5}\ M^{1.5}\ A^{0.5}}{D^2}$	88.72	232.7	234.5	232.9	233.2	233.2
9	$No.of\ Breaks = 0.086502\ L^{0.5} + 0.00051089\ \frac{L^{1.5}\ A^{1.5}}{D^2}\ \ln\left(\frac{1}{L^{0.5}}\right) + 0.021313\ \frac{L^{1.5}\ M^2\ A^{0.5}}{D^2}$	88.86	229.7	232.4	230.1	230.4	230.4
10	$No.of\ Breaks = 0.017785\frac{L^{1.5}\ M^2\ A^{0.5}}{D^2} + 6.1833 \times 10^{-6}\ \frac{L^{1.5}\ A^2}{D\ M^2}\ \ln\left(\frac{D^{1.5}}{L}\right)$	89.35	219.7	221.4	219.9	220.2	220.2
11	$No.of\ Breaks = 0.00044077\ L + 0.017413\frac{L^{1.5}\ M^2\ A^{0.5}}{D^2} + 8.8604 \times 10^{-5}\ \frac{L^{1.5}\ A^2}{D^{1.5}\ M^2}\ \ln\left(\frac{D^{1.5}}{L}\right)$	89.21	222.6	225.2	223	223.3	223.3
12	$No.of\ Breaks = 0.00057323L + 0.049651\ \frac{L^{1.5}\ M^{1.5}\ A^{0.5}}{D^2}\ \ln(M^{0.5}) + 8.8156 \times 10^{-5}\ \frac{L^{1.5}\ A^2}{D^{1.5}\ M^2}\ \ln\left(\frac{D^{1.5}}{L}\right)$	89.30	220.8	223.4	221.1	221.5	221.5