FORECASTING BREAKAGE RATE IN WATER DISTRIBUTION NETWORKS USING EVOLUTIONARY POLYNOMIAL REGRESSION

F. Karimian[1,3], H. Elsawah[1], T. Zayed[1], O. Moselhi[1], A. AlHawari[2]
[1] Department of Building, Civil and Environmental Engineering, Concordia University, Canada.
[2] Department of Civil and Architectural Engineering, Qatar University, Doha, Qatar.
[3] farzad.karimian@gmail.com

Abstract: The economic, social and environmental impact of water main failures impose great pressure on utility managers and municipalities to develop reliable rehabilitation/replacement plans. The annual number of breaks or breakage rate of each pipe segment is known as one of the most important criteria in condition assessment of these pipelines. A model is developed in this research to predict the annual number of breaks in water pipes. The developed model utilizes Evolutionary Polynomial Regression (EPR), which is intuitive data mining technique. The model is applied to a case study to test its effectiveness. The case considers the water distribution networks of in the cities of Doha in Qatar; Montréal, Moncton and Hamilton in Canada. The results indicated that the developed models successfully estimated the breakage rate for the city of Montréal and the number of breaks for the city of Doha with a maximum coefficient of determination of 88.51% and 96.27% respectively. This demonstrates the accuracy and robustness of the developed models in forecasting the number of breaks and breakage rate in water distribution networks.

1 INTRODUCTION:

The Canadian Infrastructure Report Card 2012 (CIRC 2012) shows municipal drinking-water networks ranked “Good: Adequate for now”. Despite this overall rating, 15.4% of water distribution systems in Canada were ranked “fair” to “very poor” that can cause a total cost of $25.9 billion for replacement of the pipes. The “fair”, “poor” and “very poor” condition would be interpreted as deterioration beginning to be reflected, nearing the end of useful life and no residual life expectancy respectively (CIRC 2012). According to this report, 86 Canadian municipalities recorded a total of 719,630 km of water pipelines containing distribution pipes (≤350 mm diameter) and transmission pipes (>350 mm diameter). Based on the expertise’s viewpoint one of the most popular techniques in finding the leakage location in the water pipes is joining the pole to pipes. Therefore the approximate failure place can be discovered by hearing the leak sound. Obviously, this method is insufficient, inaccurate and sometimes cost consuming. According to the CIRC 2012, a considerable percentage of municipalities in Canada do not have complete data of buried infrastructures including water and sewer pipes. While, it is clear that testing, inspection and evaluating of the physical specification of pipes requires large amount of financial reserves, it can be more profitable if the prediction models develop based on the limited historical data instead of real physical data. Recently, a data-mining technique titled Evolutionary Polynomial Regression (EPR) was developed by Giustolisi and Savic. The EPR can be categorized as Grey Box Technique and is performed in two stages: 1) search for the best model using Multi-Objective Genetic Algorithm (MOGA), and 2) parameter estimation for the model using Least Square Method. This type of regression generates several symbolic expressions that are understandable by specialists and professionals, based on various independent variables. The main objectives of this study are: 1) estimate the number of breaks of water mains based on the pipe age to reconstruct the historical databases using regression; 2) develop failure rate prediction models using Evolutionary Polynomial Regression; and 3) develop deterioration curve of water mains with respect to the pipe age using sensitivity analysis.

2 PREVIOUS STUDIES:

Several researchers considered many factors that affect the deterioration of water mains and pipe failure. Stone et al. (2002) categorized these factors in two groups, namely; static and dynamic. The

characteristics of static parameters (e.g., diameter, length, soil type, pipe material, etc.) do not depend on the time, but dynamic factors (e.g., age, cumulative number of breaks, soil corrosivity, water pressure, etc.) change over time. Osman and Bainbridge (2011) tried to identify the effect of time-dependent variables like pipe age, temperature and soil moisture on the deterioration of water pipes. In addition, InfraGuide. (2003) classified the deterioration factors of water pipes into three main categories, namely; physical (e.g., pipe material, pipe wall thickness, pipe age, pipe diameter, type of joints, pipe installation, etc.), environmental (pipe bedding, trench backfill, soil type, pipe location, etc.) and operational (internal water pressure, leakage, water quality, O&M practices, etc.). Several researchers such as Berardi et al. (2008), Xu et al. (2011), Arsénio et al. (2014) and Kutyłowska (2015) examined the impact of physical parameters on the water pipes failure. The most frequent explanatory variables in these studies are age, diameter, length and material of the pipe. Some others tried to add more parameters from various categories (environmental and physical) as the independent variables in order to improve reliability of the models. For example, Wang et al. (2009) and Shirzad et al. (2014) took into the consideration the operational factors like hydraulic pressure and burial depth in addition to physical factors. Asnaashari et al. (2013) considered soil type as the environmental factors, as well as the physical ones. Jafar et al. (2010) employed the pipe characteristics (physical), hydraulic pressure (operational), soil type and pipe location (environmental) as the inputs. Moglia et al. (2008) utilized corrosion rate, wall thickness, internal pressure and external loads as the inputs; while, typically corrosion rate is the output in the most of the previous studies.

During the last three decades, researchers introduced different models to predict the condition of the water pipes for a reliable infrastructure management system using various methodologies. The deterioration and water pipe failure prediction models could be classified into six categories; namely, deterministic, statistical, probabilistic and some advanced mathematical models such as artificial neural networks (ANN), fuzzy logic and heuristic (St. Clair and Sinha, 2012). Recently, several efforts have been made to develop deterioration and pipe failure prediction models. A summary of the most prominent ones is shown in Table 1. The EPR does not require large data sets for training and unlike ANN, it enables recognition of correlations among dependent and independent variables. As such it is not a "Black-Box" technique, but it is classified as a "Grey-Box" technique that can provide insight into the relationship between inputs and output. The process of development and selection of EPR contains engineering knowledge that allows the user to understand the generated equations and the correlation between variables involved.

Table 1 Prediction Models of Water Distribution Networks

Authors (Year)	Model Classification	Methodology	Output Type
Berardi et al. (2008)	Statistical	Evolutionary Polynomial Regression	Pipe Deterioration
Moglia et al. (2008)	Probabilistic	Monte-Carlo Simulation Framework	Cast Iron Pipe Failure
Wang et al. (2009)	Statistical	Five Multiple Regression Models	Annual Break Rates
Li et al. (2009)	Probabilistic	Monte-Carlo Simulation	Remaining Useful Life
Wang et al. (2010)	Statistical	Bayesian Inference	Deterioration Rate
Jafar et al. (2010)	Artificial Neural Networks	Six ANN Models	Failure Rate
Xu et al. (2011)	Statistical	Genetic Programming and Evolutionary Polynomial Regression	Deterioration Rate
Osman and Bainbridge (2011)	Statistical	Rate of Failure (ROF) and Transition State (TS)	Deterioration Rate
Asnaashari et al. (2013)	Artificial Neural Networks	ANN and Multi Linear Regression	Failure Rate
Shirzad et al. (2014)	Artificial Neural Networks	ANN and Support Vector Regression (SVR)	Pipe Burst
Arsénio et al. (2014)	Statistical	Ground Movement Estimated by Radar Satellite Data	replacement-prioritization plan
Kutyłowska (2015)	Artificial Neural Networks	ANN	Failure Rate

2.1 EVOLUTIONARY POLYNOMIAL REGRESSION:

The Evolutionary Polynomial Regression (EPR) technique was first presented by Giustolisi and Savic (2006). The technique utilizes the huge potential of conventional numerical regression techniques and

the strength of Genetic Algorithm in solving optimization problems (Xu et al. 2011). Later, this approach was used by other researchers in several fields. Savic et al. (2006) and Ugarelli et al. (2008) used EPR to model the sewer pipe failures. Several researches were conducted using EPR in different engineering fields, Berardi et al. (2008) and Xu et al. (2011) applied the EPR to develop deterioration models for water distribution networks. 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. Guistolisi and Savic (2009) tested the EPR-MOGA (An improved EPR) to develop groundwater level prediction model based on monthly rainfall. El-Baroudy et al. (2010) utilized the EPR to develop the evapotranspiration process then compared efficiency of Evolutionary Polynomial Regression 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.

Evolutionary Polynomial Regression is data-driven technique and can be classified as a grey box method according to the color coding classification system 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 creating the symbolic expressions contains two stages; in the first stage the EPR tries to find the best model structure using Multi-Objective Genetic Algorithm (MOGA). Then, the appropriate values for constant are estimated using Least-Squares optimization (LS) (Berardi et al. 2008). In the 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. In this study the following equation was chosen:

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

Where, X_k is the kth explanatory variable, ES is the matrix of unknown exponents that should be defined by the user, f is inner function selected by the user (can be no function, logarithm, exponential, tangent hyperbolic, secant hyperbolic), a_j are unknown polynomial coefficients, m is the number of polynomial terms and a_0 is the bias term. It must be remarked that, the zero value should be considered in the matrix of 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 by EPR, it tries to return several equations based on 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. Each of these options can be selected by the user. Furthermore, the user can force EPR to generate the equations with only positive value 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 just needs to identify the range in which the inputs or output should be scaled. The EPR can develop 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 recreated using linear interpolation by EPR; thus, the model can be developed with an incomplete historical database. During the generating symbolic expressions, if the EPR cannot find appropriate combination of terms containing $f(x)$ (as an inner function), it deselects this function. The effectiveness of fitness of developed models was measured by Coefficient of Determination (CoD) equation as follows (Berardi et al. 2008):

$$[2] \quad CoD = 1 - \frac{\sum_n \left(\hat{y} - y_{exp}\right)^2}{\sum_n \left(y_{exp} - avg\left(y_{exp}\right)\right)^2} = 1 - \frac{n}{\sum_n \left(y_{exp} - avg\left(y_{exp}\right)\right)^2} \cdot SSE$$

Where n is the number of samples, $\hat{y}$ is the value that predicted by the model, y_{exp} is the actual amount of the historical data and SSE is the sum of squared errors.

3 RESEARCH METHODOLOGY:

Fig. 1 shows the developed research methodology. It started by presenting a comprehensive literature review to investigate the deterioration and pipe failure prediction models. In the second step, the relevant data were collected from two different municipalities; City of Doha in Qatar and City of

Montréal in Canada. The database of the city of Montréal contains an inventory of pipes' information and the related bursts. However, the number of breaks was not available in the database of the city of Doha. Lack of such data prevents working with EPR because this technique takes into account the number of breaks or breakage rate as a dependent variable in order to develop a pipe failure prediction model. Therefore, it was necessary to estimate the number of breaks for the city of Doha from similar infrastructure databases. Thus, the number of breaks for the city of Doha was estimated using two databases such as: Moncton and Hamilton. Several attempts were carried out using different regression models to predict the number of breaks of water mains based on the pipe's age. The model that provides large number of data points as well as the best R-Square was used to predict the pipe bursts for city of Doha. The third step, the model implementation, the databases were cleaned and classified into homogeneous groups based on age and diameter of the pipes. It means that all pipe segments of each class have the same value of age and diameter. Also, all missing data points were removed from the databases. At this point, data sets of Montréal and Doha are ready to be analyzed using EPR_MOGA_XL. In this study, two different scenarios were implemented; the first one considered the number of breaks as an output and the other one assumed that the dependent variable is the breakage rate (Breaks/Length (Km)/Age (Yr)). Finally, the sensitivity analysis was carried out for both cases to develop different deterioration curve and to identify the effect of each variable on the breakage rate.

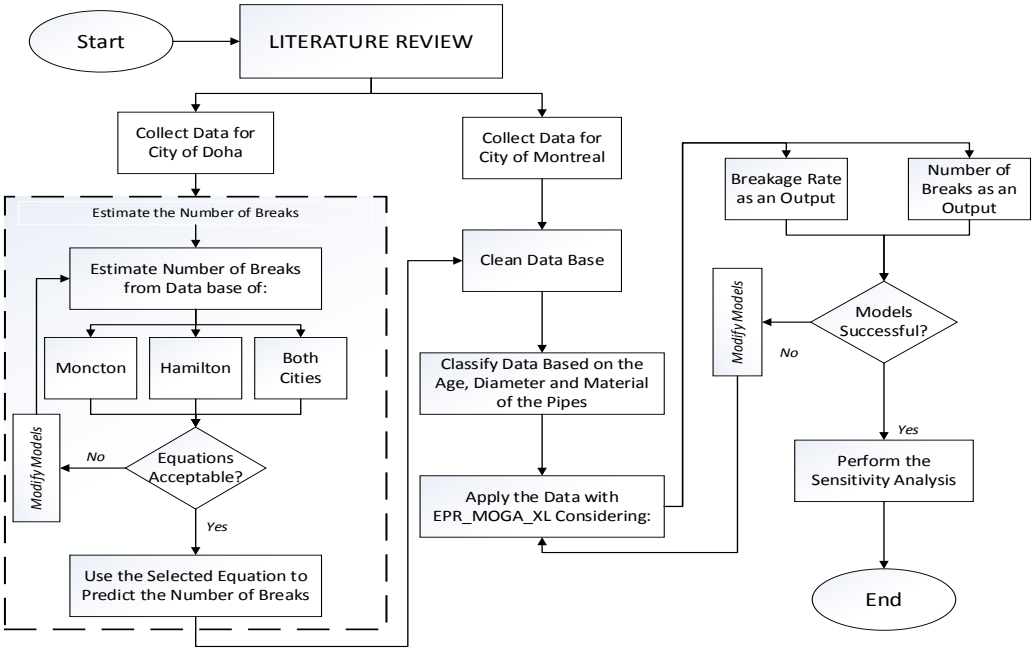

Figure 1 Research Methodology

4 DATA COLLECTION:

In this study, four sets of data from various municipalities were used, namely; city of Moncton, city of Hamilton and city of Montréal in Canada and city of Doha in Qatar. The city of Moncton, Hamilton, Montréal and Doha own 517 km, 1,891 km, 70 km and 4,682 km of water distribution networks respectively (SOIR, 2005, KAHRAMAA Report). The physical characteristics of water pipes in different databases are generic. In fact the results obtained using the Hamilton and Moncton data were very close. In view of this finding and the insufficient data collected from Doha, it was required to use the model developed based on Hamilton and Moncton to predict the number of breaks in Doha. Estimation of number of breaks was implemented by applying regression analysis using data sets of Moncton and Hamilton. Three different models were developed. In the first two models the data of each city was used separately, while, in the third one the combined data for both cities was utilized. In each model the data was clustered into different groups based on the pipe age. The breaks per length (m) was calculated for each age-class by computing the average of number of bursts. Several attempts were conducted to reach the best model using different types of databases. Since, in the database of Doha, there are not pipes older than 33 years, it was not necessary to keep the pipes with the age of 34 and more, therefore they could be deleted in the new inventories. Finally, the model

which provides large number of data points and gives the best performance based on the Coefficient of Determination (R^2), was chosen to estimate the number of breaks for the city of Doha.

Fig. 2 shows the result of regression (based on the No. of Breaks per Length (m)) of Moncton, Hamilton and mixing of both cities, respectively. The equation of each inventory and determined R-Square (R^2) are shown in Table 2. It can be seen that the developed models of Moncton and both Cities are acceptable; while, the one that belongs to city of Hamilton is not promising enough to be used in Doha. Then, number of breaks per length that obtained from these equations should be multiplied by length of related pipe segments to calculate the estimated number of breaks of Doha's database.

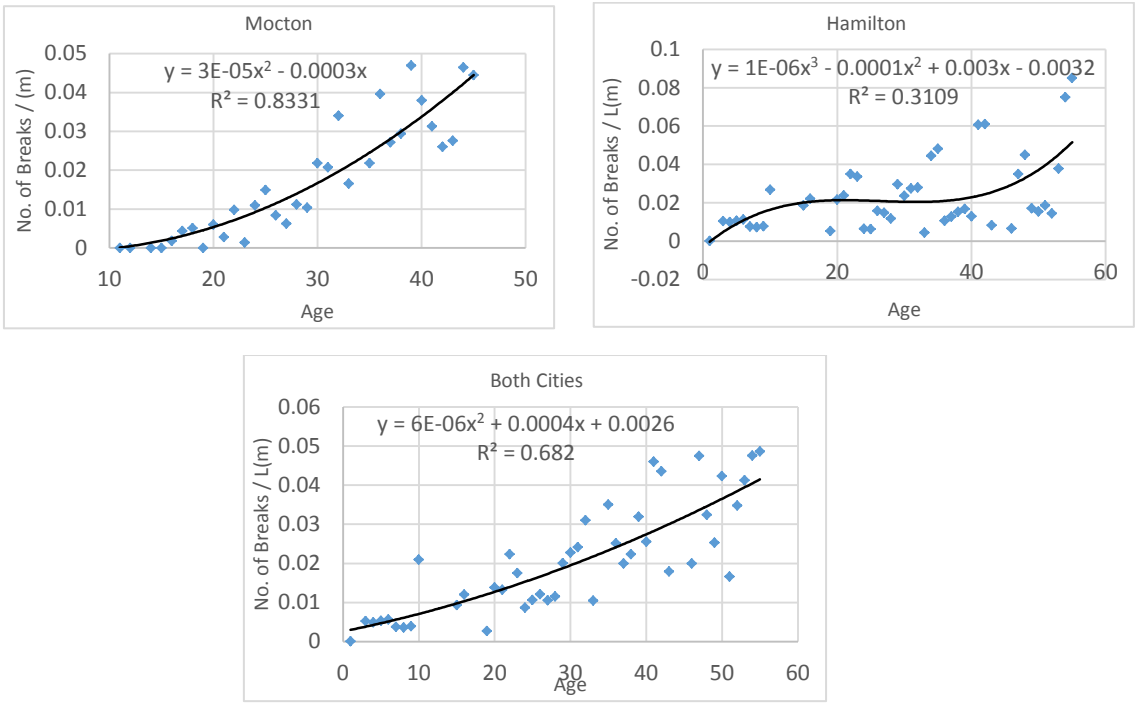

Figure 2 Scatter Plots of No. of Breaks versus Length (m) based on the different databases

Table 2 Equations and related R-Squares

Different Databases	Equations	R-Squared (%)
Moncton	$y = 3E\text{-}05\ x^2 - 0.0003\ x$	83.31
Hamilton	$y = 1E{-}06x^3 - 0.0001x^2 + 0.003x - 0.0032$	31.09
Mixing of Both Cities	$y = 6E\text{-}06\ x^2 + 0.0004\ x + 0.0026$	68.20

The completed database of Doha, including the estimated number of breaks, and Montréal data set was filtered before applying with EPR. Each of them comprises several factors: Table 3 shows the collected variables for both inventories. The units of age, length, diameter, depth of laid and pipe elevation are year, Km, mm, m and m respectively in collected data. For cleaning and organizing databases, several steps performed. Some segments with having either the missing pipe information or inconsistent data were totally removed from databases. The qualitative variables such as pipe type and soil type were translated to numerical ones. For example, if there are four different types of soil, each number from 1 to 4 was assigned to a specific soil type. It is noteworthy mentioning that the maximum number was assigned to the hardest pipe type; in other words, the harder the material, the larger the allocated number and the vice versa. As it was mentioned before, there are two different scenarios in considering the output in the current study; number of breaks and breakage rate. The breakage rate of each segment should be calculated by dividing number of breaks by age (or observation years) and length (km) of the pipe. Finally, two databases were classified into homogeneous groups based on age and diameter of the pipe. A detailed discussion about classification is presented in model implementation section.

5 MODEL IMPLEMENTATION:

5.1 EPR SETTINGS:

In the database of Montréal, the following factors were considered as independent variables; pipe length, diameter, pipe type and age of the pipe. Whereas, inputs in Doha data set were pipe length, diameter, thickness, pipe type, age, pipe elevation and depth laid. Also, the settings which are followed were same for both cases. Nomination of exponents were limited to [-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2] in which the positive and negative value represent the direct and inverse relationship between dependent and independent variables and their amounts show how significant the inputs are. The maximum number of terms in each expression was three and the bias term was considered as zero. In the Regression Method tab, the Least Square parameter estimation was forced to search for just positive value of a_j as a coefficient. The exponential was selected as the inner function. EPR rounded the output to the nearest integer number if the classification is selected as the Modelling Type. Thus, in the scenarios in which the breakage rate was assumed as a dependent variable, Statical Regression should be chosen for considering the output as a real number. The range of variables was constrained to [0.01, 1] for scaling all inputs and output data. The "GA" is the number of generation and depends on several attribute such as number of independent and dependent variables, number of terms and exponents, and was selected as 40 based on the previous experience. There are three different choices in Optimization Strategy; namely, $Min(a_j, SSE), Min(X_i, SSE)$ and $Min(a_j, X_i, SSE)$. Among them the $Min(X_i, SSE)$ was selected. Maximization of the parsimony and the accuracy of the model caused by minimization of a_j, X_i and SSE, respectively.

Table 3 Collected Data for Each City

Factors	Montréal	Doha
Pipe Length	Yes	Yes
Diameter	Yes	Yes
Thickness	No	Yes
No. of Breaks	Yes	Yes*
Pipe Type	Yes	Yes
Age	Yes	Yes
Observation Period	Yes	No
Functional Impairments	Yes	No
Pipe Elevation	No	Yes
Depth Laid	No	Yes

* Estimated by Other Databases

5.2 CLASSIFICATION:

Once, the process of cleaning and reconstruction was conducted, pipe segments were classified into several homogeneous groups based on age, diameter and pipe material. In the database of Doha there were only three nonmetallic pipes out of 1599 pipes, which were removed. Thus the classification of Doha was based on only age and diameter. The following equations were used to group the data:

$$[3] \quad A_{class} = \frac{\sum_{class}(L_p \cdot A_p)}{L_{TA}} \quad , \quad [4] \quad D_{class} = \frac{\sum_{class}(L_p \cdot D_p)}{L_{TD}}$$

Where, L_{TA} and L_{TD} are the total length of pipes with the same age and diameter respectively. Also L_p, A_p and D_p are length, age and diameter of each segment in the group. Thus, there are several categories with the same class of age, diameter and material for each inventory. It should be mentioned that the other features of pipe (such as thickness, length, etc.) can be utilized as grouping criteria in different studies. But in this research age was selected for classification to take into account the indirect effect of time-varying solicitation on water mains. Because as an engineering point of view, the higher the duration of solicitation, the higher the chemical and mechanical harmful effects (caused by soil condition, traffic loads, etc.) on pipes (Berardi et al. 2008). Furthermore, the other equivalent attributes in each database can be calculated by different mathematical functions (such as

sum, average, etc.). In the database of Montréal, the length and the number of breaks of each class were computed by summing corresponding ones of each pipe segment. Likewise, in the Doha database, the same calculations were performed for the length and the number of breaks; while, the other variables such as pipe elevation and burial depth were calculated by computing the average of related features of pipes in that group.

6 ANALYSIS OF RESULTS:

In this study, two databases; the city of Montréal and the city of Doha were utilized to forecast the number of breaks (for the city of Doha) and the breakage rate (for the city of Montréal) using EPR. Among all symbolic expressions that were developed by EPR, the best one was chosen based on the fitness to the historical data and parsimony of the equation. Tables 4 and 5 show different pipe burst prediction models for the city of Montréal and the city of Doha with their corresponding R-Square respectively. It can be seen that in both cases the expressions accuracy is increased when equations get more complicated. It can be understood from Tables 4 and 5 that the age of the pipe has a direct impact on the breakage rate as well as the number of breaks. In addition, the equations describing the number of breaks for the city of Doha has higher coefficient of determination than the equations describing the breakage rate for the city of Montréal.

Table 4 shows twelve breakage rate prediction models that were generated by applying the EPR with database of Montréal. It can be seen that equations 5 considers the age, length and diameter of pipes and has a coefficient of determination of 79.58%. In equation 6, the material of pipes was added to the model which improves the coefficient of determination by 5.71%. While, the coefficient of determination increased from 85.29% to 88.51% from equation 6 to 12, the models became more complicated. Thus, in this study, equation 6 was selected to be the best one to describe the pipe breakage rate as it has the acceptable coefficient of determination with reasonable number of variables.

Table 5 shows the twelve symbolic expressions that were generated by EPR to estimate the number of breaks for the City of Doha. It can be seen that age, length and diameter of the pipes are the most commonly used variables for estimating the number of breaks while depth laid and pipe elevation has been introduced in only the last five equations. It can be notice that as the age of pipes increases the breakage rate increases. However, when the diameter of the pipes increases the breakage rate decreases. It can be remarked that equation 2 has a coefficient of determination of 94.33%. However, it only considers the length and the age of the pipe. In order to test the sensitivity of the prediction model by changing the inputs, model number 9 was chosen as it considers all variables that has a significant effect on the pipe bursts.

Figure 3 and 4 show the Pareto graphs of different models based on Montréal and Doha result respectively. In these graphs each point represents one returned equation and vertical axes shows the number of selected inputs while the horizontal axes demonstrate the fitness of models (1-CoD).

7 SENSITIVITY ANALYSIS:

In this study, the sensitivity analysis was performed to identify the effect of each variable on the pipe bursts when the water pipes get older. Figures 5 to 9 show the deterioration curves for the city of Montréal and the city of Doha. Figures 5 and 6 illustrate the breakage rate for the city of Montréal for different pipe diameters and pipe lengths respectively while keeping all other inputs as constants. The value of these constants were considered as the average of that variables in the database. Figure 5 and 6 show the effect of pipe diameter and pipe length on the breakage rate. The breakage rate of pipes with different diameter has increasing trend as the pipe become older. It is clear from the Figure 5 that the bigger the diameter of the pipes, the higher the value of breakage rate. In addition, Figure 6 shows the breakage rate of pipes with various length are increased when the pipes get older. The same analysis was carried out for city of Doha. As it was mentioned before the equation 9 of database of Doha was used to develop graphs including different deterioration curves. Figure 7, 8 and 9 show the sensitivity analysis for length, depth laid and pipe elevation respectively. It is confirmed in these graphs that, the number of breaks is increasing while the pipes are aging.

Table 4 Produced Equations by EPR for Montréal database and related CoD

	Symbolic Expressions	CoD
1-	$BR=0.0052101\frac{1}{D}$	29.54
2-	$BR=0.017429\frac{L^2}{D^2}$	43.33
3-	$BR=0.011799\frac{1}{D^{1.5}}+0.0025988\frac{1}{D^{1.5}}\ln(L^{1.5})$	66.83
4-	$BR=0.011236\frac{1}{D^{1.5}}+0.0024744\frac{1}{D^{1.5}}\ln(L^{1.5})+0.086518\,A^2$	70.92
5-	$BR=0.0079362\frac{A^2}{L}+0.012017\frac{1}{D^{1.5}}+0.0026574\frac{1}{D^{1.5}}\ln(L^{1.5})$	79.58
6-	$BR=0.028312\frac{M^2A^2}{L}+0.012306\frac{1}{D^{1.5}}+0.0027426\frac{1}{D^{1.5}}\ln(L^{1.5})$	85.29
7-	$BR=0.015285\frac{D\,M^2A^2}{L^{1.5}}+0.01217\frac{1}{D^{1.5}}+0.0026855\frac{1}{D^{1.5}}\ln(L^{1.5})$	85.67
8-	$BR=0.0033716\frac{M^2A^2}{L^{1.5}}\ln\left(\frac{A^{1.5}}{L^{0.5}}\right)+0.012315\frac{1}{D^{1.5}}+0.0027226\frac{1}{D^{1.5}}\ln(L^{1.5})$	87.74
9-	$BR=0.020404\frac{M^2A^2}{L}\ln\left(\frac{A^{1.5}}{L^{0.5}}\right)+0.010121\frac{1}{D^{1.5}}+0.0032494\frac{A^{0.5}}{D^{1.5}}\ln(L^{1.5})$	87.38
10-	$BR=0.20994\frac{D^{0.5}M^2A^2}{L^{0.5}}\ln\left(\frac{A^{1.5}}{L^{0.5}}\right)+0.0083997\frac{1}{D^{1.5}}+0.0038976\frac{A}{D^{1.5}}\ln(L^{1.5})$	88.51
11-	$BR=0.020234\frac{D^{0.5}M^2A^2}{L^{0.5}}\ln\left(\frac{A^{1.5}}{L^{0.5}}\right)+0.0069416\frac{1}{D^{1.5}}+0.0029473\frac{A^{0.5}}{D^{1.5}}\ln\left(\frac{L^{1.5}}{A^{1.5}}\right)$	88.51
12-	$BR=0.21408\frac{D^{0.5}M^2A^2}{L^{0.5}}\ln\left(\frac{A^{1.5}}{L^{0.5}}\right)+0.008365\frac{A^{0.5}}{D^{1.5}}+0.0063721\frac{A}{D^{1.5}}\ln\left(\frac{L}{A^{1.5}}\right)$	87.6

Table 5 Produced Equations by EPR for Doha database and related CoD

	Symbolic Expressions	CoD
1-	$Breaks=1.1493\,A^{0.5}\,L$	90.66
2-	$Breaks=1.8169\times10^{-6}\frac{1}{L^2}+1.1471\,A^{0.5}\,L$	94.33
3-	$Breaks=3.3795\times10^{-6}\frac{A^{0.5}}{L^2}+1.1466\,A^{0.5}\,L$	94.99
4-	$Breaks=1.146\,A^{0.5}\,L+1.7317\times10^{-5}\frac{A^2}{L^2}\ln\left(\frac{1}{A^{1.5}}\right)$	95.21
5-	$Breaks=0.0066601\ln(D^{0.5})+1.1721\,A^{0.5}\,L+1.6884\times10^{-5}\frac{A^2}{L^2}\ln\left(\frac{1}{A^2}\right)$	95.87
6-	$Breaks=1.4624\times10^{-5}\frac{1}{L^{1.5}}\ln(D^{0.5})+1.1467A^{0.5}L+2.0317\times10^{-5}\frac{A^2}{L^2}\ln\left(\frac{1}{A^2}\right)$	95.58
7-	$Breaks=3.4001\times10^{-5}\frac{A^{0.5}}{L^{1.5}}\ln(D^{0.5})+1.1473A^{0.5}L+2.4526\times10^{-5}\frac{A^2}{L^2}\ln\left(\frac{1}{A^2}\right)$	95.7
8-	$Breaks=3.2006\times10^{-5}\frac{A^{0.5}}{L^{1.5}}\ln(D^{0.5}PE^{0.5})+1.1477A^{0.5}L+2.8169\times10^{-5}\frac{A^2}{L^2}\ln\left(\frac{1}{A^2}\right)$	95.91
9-	$Breaks=3.4191\times10^{-7}\frac{A^{0.5}}{L^2D^{0.5}}\ln(PE^{1.5}Del^{0.5})+1.1463A^{0.5}L+3.5493\times10^{-5}\frac{A^2}{L^2}\ln\left(\frac{1}{A^{1.5}}\right)$	96.02
10-	$Breaks=3.4559\times10^{-7}\frac{A^{0.5}}{L^2D^{0.5}}\ln(PE^{1.5}DeL^{0.5})+1.1464A^{0.5}L+3.5792\times10^{-6}\frac{A^2}{L^2DeL^{0.5}}\ln\left(\frac{1}{A^{1.5}}\right)$	96.1
11-	$Breaks=3.817\times10^{-7}\frac{A^{0.5}}{L^2D^{0.5}}\ln\left(\frac{PE^{1.5}DeL^{0.5}}{A^{0.5}}\right)+1.1464A^{0.5}L+3.5748\times10^{-6}\frac{A^2}{L^2DeL^{0.5}}\ln\left(\frac{1}{A^{1.5}}\right)$	96.27
12-	$Breaks=3.822\times10^{-7}\frac{A^{0.5}DeL^{0.5}}{L^2D^{0.5}}\ln\left(\frac{PE^{1.5}DeL^{0.5}}{A^{0.5}}\right)+1.1465A^{0.5}L+3.5775\times10^{-6}\frac{A^2}{L^2DeL^{0.5}}\ln\left(\frac{1}{A^{1.5}}\right)$	96.05

8 CONCLUSION:

The process and specifications of Evolutionary Polynomial Regression technique were described in this paper. The setting of EPR_MOGA software was explained in details. Four different aggregated data sets were utilized in different parts including city of Doha, Montréal, Moncton and Hamilton. In the database of Doha, the number of breaks of pipelines was not available. Therefore a regression model developed to predict the number of breaks based on the pipe age. This was performed using city of Moncton, city of Hamilton and mixing of both cities. The EPR was applied with databases of Doha and Montréal. For each attempt several symbolic expressions were returned by EPR and their fitness was measured by Coefficients of Deterministic. Then, the results were analyzed and sensitivity analysis was implemented to identify the effect of each factor on the number of breaks and breakage rate of the water pipes. As it was confirmed by previous studies, the most critical factors that affect deterioration of water pipes are age, length, diameter and material. Finally, for each data set, one

predicting model was selected to be used by utility managers and decision makers of Water Distribution Networks.

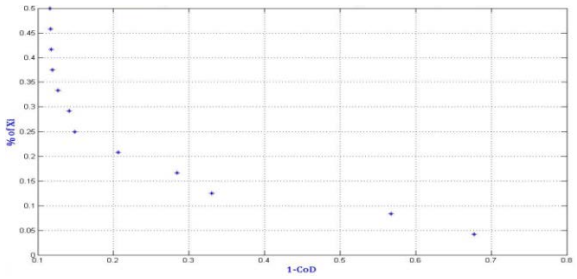

Figure 3 Pareto of Montréal, Fitness vs. Number of Variables

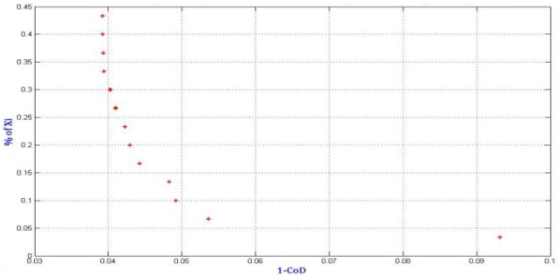

Figure 4 Pareto of Doha, Fitness vs. Number of Variables

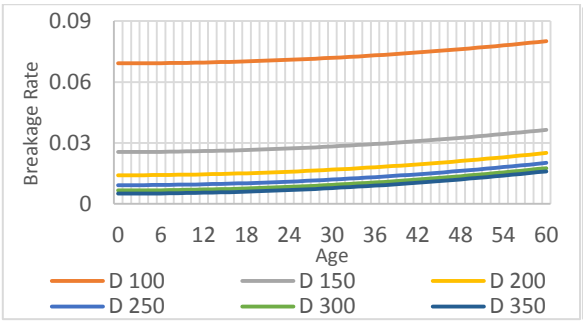

Figure 5 Breakage Rate for Different Pipe Diameters (Montréal)

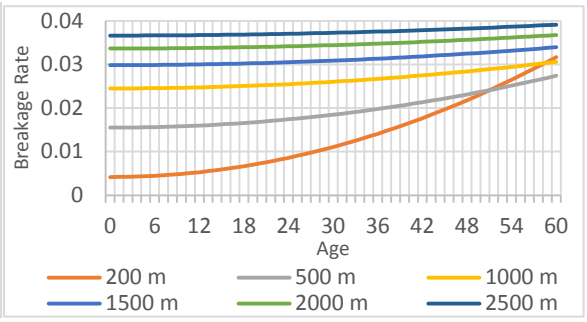

Figure 6 Breakage Rate with Different Pipe Length (Montréal)

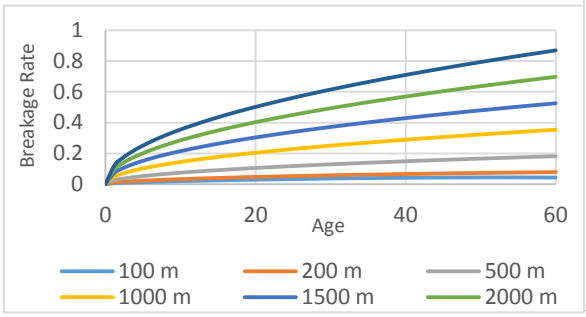

Figure 7 Breakage Rate with Different Pipe Length (Doha)

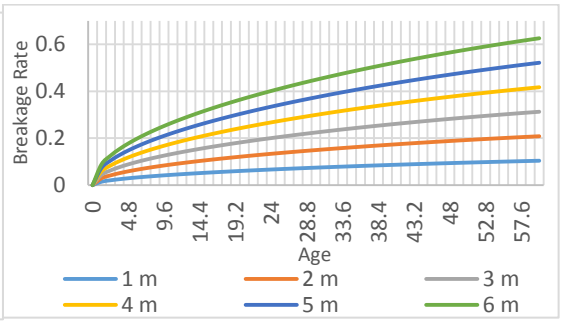

Figure 8 Breakage Rate with Different Depth Laid (Doha)

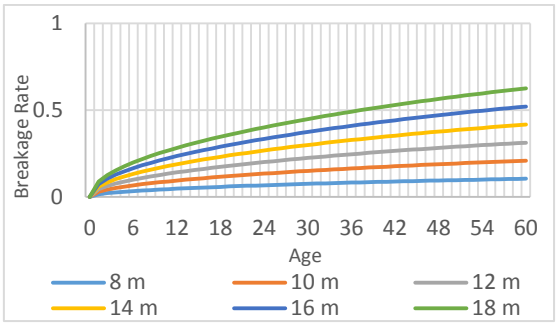

Figure 9 Breakage Rate with Different Pipe Elevation (Doha)

Acknowledgement:

This publication was made possible by NPRP grant # (NPRP-5-165-2-055) from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References:

Achim, D., Ghotb, F. and McManus, K. 2007. Prediction of water pipe asset life using neural networks. *Journal of Infrastructure Systems,*13(1):26-30.

Ahangar-Asr, A., Faramarzi, A., Javadi, A. and Giustolisi, O. 2011. Modelling mechanical behaviour of rubber concrete using evolutionary polynomial regression. *Eng. Computations,*28(4):492-507.
Arsénio, A.M., Dheenathayalan, p., Hanssen, R., Vreeburg, J. and Rietveld, L. 2014. Pipe failure predictions in drinking water systems using satellite observations. *Structure and Infrastructure Engineering,*1-10.
Asnaashari, A., McBean, E.A., Gharabaghi, B. and Tutt, D. 2013. Forecasting watermain failure using artificial neural network modelling. *Canadian Water Resources Journal,* 38 (1): 24-33.
Berardi, L., Kapelan, Z., Giustolisi, O. and Savic, D. 2008. Development of pipe deterioration models for water distribution systems using EPR. *Journal of Hydroinformatics,* 10 (2): 113-26.
CIRC. 2012. Canadian Infrastructure Report Card, Volume 1: 2012, Municipal Roads and Water Systems, www.canadainfrastructure.ca/downloads/Canadian_Infrastructure_Report_Card_EN.pdf
El-Baroudy, I., Elshorbagy, A. Carey, S. Giustolisi, O. and Savic, D. 2010. Comparison of three data-driven techniques in modelling the evapotranspiration process. *Journal of Hydroinformatics,,*12 (4):365-79.
Elshorbagy, A. and 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):237-51.
Fiore, A., Berardi, L. and Marano, G.C. 2012. Predicting torsional strength of RC beams by using evolutionary polynomial regression. *Advances in Engineering Software,* 47 (1): 178-87.
Giustolisi, O., and Savic, D. 2009. Advances in data-driven analyses and modelling using EPR-MOGA. *Journal of Hydroinformatics,* 11 (3-4): 225-36.
Giustolisi, O. 2004. Using genetic programming to determine chezy resistance coefficient in corrugated channels. *Journal of Hydroinformatics,* 6 : 157-73.
Infraguide. 2003. Deterioration and Inspection of Water Distribution Systems. https://www.fcm.ca/Documents/reports/Infraguide/Deterioration_and_Inspection_of_Water_Distribution_Systems_EN.pdf
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): 1170-80.
Kutyłowska, M. 2015. Neural network approach for failure rate prediction. *Engineering Failure Analysis,* 47 : 41-8.
Li, S., Yu, S., Zeng, H. and Liang, R. 2009. Predicting corrosion remaining life of underground pipelines with a mechanically-based probabilistic model. *Petroleum Science and Eng.* 65(3):162-6.
Markus, M, Hejazi, M., Bajcsy, P., Giustolisi, O. and Savic, D. 2010. Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in illinois. *Journal of Hydroinformatics,* 12(3): 251-61.
Moglia, M., Davis, P. and Burn, S. 2008. Strong exploration of a cast iron pipe failure model. *Reliability Engineering & System Safety,* 93 (6): 885-96.
Osman, H., and Bainbridge, K. 2011. Comparison of statistical deterioration models for water distribution networks. *Journal of Performance of Constructed Facilities,* 25 (3): 259-66.
KAHRAMAA Report, QGEWC. Qatar General Electricity and Water Coporation. Available Online at: http://www.km.com.qa/AboutUs/Pages/WaterSector.aspx
Rezania, M., Javadi, A.A. and Giustolisi, O. 2008. An evolutionary-based data mining technique for assessment of civil engineering systems. *Engineering Computations,* 25 (6) : 500-17.
Savic, D., Giustolisi, O., Berardi, L., Shepherd, W., Djordjevic, S., and Saul, ,A. 2006. Modelling sewer failure by evolutionary computing. *Proceedings of the ICE-Water Management,* 159 (2): 111-8.
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): 941-8.
St. Clair, A.M. and Sinha, S. 2012. State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models. *Urban Water Journal,* 9 (2): 85-112.
Stone, S.L., Dzuray, E.J., Meisegeier, D., Dahlborg, A., Erickson, M. and. 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.
Ugarelli, R., Kristensen, S.M., Røstum, J., Sægrov, S. and Di Federico, V. 2008. Statistical analysis and definition of blockages-prediction formulae for the wastewater network of oslo by evolutionary computing. *Water Science and Technology,* 59 (8): 1457-70.
Wang, C., Niu, Z., Jia, H. and Zhang, H. 2010. An assessment model of water pipe condition using bayesian inference. *J. of Zhejiang University SCIENCE A,* 11 (7): 495-504.
Wang, Y., Zayed, T. and Moselhi, O. 2009. Prediction models for annual break rates of water mains. *Journal of Performance of Constructed Facilities,* 23 (1): 47-54.
Xu, Q., Chen, Q., Li, W. and Ma, J. 2011. Pipe break prediction based on evolutionary data-driven methods with brief recorded data. *Reliability Engineering & System Safety,* 96 (8): 942-8.