International Journal of Architectural, Civil and Construction Sciences
ISSN: 2415-1734
Vol:11, No:7, 2017

Deterioration Assessment Models for Water Pipelines

L. Parvizsedghy, I. Gkountis, A. Senouci, T. Zayed, M. Alsharqawi, H. El Chanati, M. El-Abbasy, F. Mosleh

***Abstract*—**The aging and deterioration of water pipelines in cities worldwide result in more frequent water main breaks, water service disruptions, and flooding damage. Therefore, there is an urgent need for undertaking proper maintenance procedures to avoid breaks and disastrous failures. However, due to budget limitations, the maintenance of water pipeline networks needs to be prioritized through efficient deterioration assessment models. Previous studies focused on the development of structural or physical deterioration assessment models, which require expensive inspection data. But, this paper aims at developing deterioration assessment models for water pipelines using statistical techniques. Several deterioration models were developed based on pipeline size, material type, and soil type using linear regression analysis. The categorical nature of some variables affecting pipeline deterioration was considered through developing several categorical models. The developed models were validated with an average validity percentage greater than 95%. Moreover, sensitivity analysis was carried out against different classifications and it displayed higher importance of age of pipes compared to other factors. The developed models will be helpful for the water municipalities and asset managers to assess the condition of their pipes and prioritize them for maintenance and inspection purposes.

***Keywords*—**Water pipelines, deterioration assessment models, regression analysis.

I. Introduction

THE deterioration of water distribution network leads to a compromised water quality, increased breakage and leakage rates, and reduced hydraulic capacity. The 2017 ASCE report card [1] rated the performance of the US water distribution infrastructure a poor grade of "D". Moreover, ASCE reported that most of the US water pipelines are over 100 years old. Thus, with the increasing number of deteriorated pipelines in the US, Canada and around the globe, deterioration assessment modeling is of paramount importance. The American Water Works Association [2] estimated an investment of over one trillion US dollar to replace all water pipelines in the US, out of which $384.2 billion is needed alone for the maintenance of water infrastructure in the next 20 years. The Canadian Infrastructure Report Card [3] displayed the Canadian water infrastructure in a good condition. However, there are ongoing concerns due to high number of failures and pipe breaks.

L. Parvizsedghy, I. Gkountis, H. El Chanati, M. El-Abbasy, and F. Mosleh were with Dept. of Building, Civil and Environmental Engineering, Concordia University, Montréal, Canada.

A. Senouci is with Dept. of Construction Management, University of Houston, Houston, Texas, USA.

T. Zayed is with Dept. of Building, Civil and Environmental Engineering, Concordia University, Montréal, Canada.

M. Alsharqawi is with Dept. of Building, Civil and Environmental Engineering, Concordia University, Montréal, Canada (corresponding author; e-mail: mohammed.alsharqawi@concordia.ca).

Water infrastructure failures have negative monetary, health, and safety consequences. Therefore, there is a need to increase the reliability of water distribution networks and reduce their maintenance costs. Deterioration assessment tools are helpful and efficient in prioritizing the maintenance of pipelines, especially when available budget is limited. Previous deterioration assessment models were either inspection-based or did not consider the interdependency between the factors affecting pipeline deterioration. Structural/ physical deterioration assessment using inspection-based methods is very costly, given the limited inspection budget. Moreover, the inspection of all pipelines in a large water distribution network is not manageable. Alternatively, statistical models can accurately assess the pipeline deterioration and help the municipalities in prioritizing the maintenance of their water pipelines.

In order to address the above-mentioned limitations, the aim of this research is to develop statistical pipeline deterioration models using regression analysis. Water pipeline historical data, which are used to develop the proposed models, were collected from Canadian municipalities. The variables that contribute to the deterioration of water pipelines were identified. Regression models were developed using the pipeline condition indices obtained from the model developed by El Chanati et al. [4] and the values of the identified factors that were collected from Canadian municipalities. Categorical and non-categorical regression models were developed for several pipeline classifications. The categorical regression models can be used to compute the condition of pipelines in case of any missing input data.

II. Research Objectives

The main objectives of the present study are as follows:

- Identify the contributory factors affecting the deterioration of water pipelines.
- Develop deterioration assessment models for water pipelines.
- Develop deterioration index and breakage rate forecasting models.

III. Background

Previous Studies

A distribution network is the most expensive component of a water supply system [5]. Its total expenditure accounts for more than 80% of the entire water supply system [6]. The National Guide to Sustainable Municipal Infrastructure best practice [7] emphasized on the importance of a planned inspection program for water distribution systems to ensure a safe, cost-effective, reliable, and sustainable water supply. Water pipeline deterioration leads to impaired water quality,

increased breakage rate, reduced hydraulic capacity, and high leakage rate. Therefore, the deterioration assessment of water pipelines is essential to assist municipalities in planning their inspection and rehabilitation actions. The deterioration assessment of water pipelines is usually conducted through two methods, namely, physical-based (i.e. direct inspection) and statistical-based approaches. The first method studies the physical mechanisms underlying pipeline failures. However, this method requires costly data [8]. Consequently, physical models are only justified for major transmission water pipelines because of their potential failure costs. Contrarily, the second method can be used for the majority of water pipelines because its input data is less costly and easy to obtain.

Several studies have been carried out to assess the condition or performance of water pipelines. Yan and Vairavamoorthy [9] used fuzzy Multi-Criteria Decision-Making (MCDM) technique for the condition assessment of water pipelines. Geem [10] used Artificial Neural Network (ANN) while developing a Decision Support System (DSS) to assess the condition of water pipelines. Al-Barqawi and Zayed [11], [12] developed condition assessment models for water mains using Analytic Hierarchy Process (AHP) and ANN methods, respectively. Geem et al. [13] applied Multiple Linear Regression (MLP) and ANN techniques to develop water pipeline condition assessment models. The results of the study presented the outperformance of the ANN technique as it resulted in a higher coefficient of determination (R2). Al-Barqawi and Zayed [14] developed an integrated AHP/ANN based condition assessment model for the water mains. Wang et al. [15] developed multiple regression models to predict annual break rates of water mains. Zhou et al. [16] developed a condition assessment model for water pipelines using fuzzy Preference Ranking Organization METHod for Enrichment Evaluation (PROMETHEE II) MCDM technique. Fares and Zayed [17] designed a framework to assess the failure risk of water mains using hierarchical fuzzy expert system. Wang et al. [18] used Bayesian inference to evaluate the condition of water pipelines. Clair and Sinha [19] applied a weighted factor and fuzzy inference methodology to forecast the performance index of metallic water pipelines. Although the above-mentioned models produced satisfactory results, they did not consider the interdependency of model variables/factors. Moreover, none of the developed models considered the categorical nature of variables such as; pipeline material type, pipe diameter size and soil type.

Recently, El Chanati et al. [4] used ANP and fuzzy inference techniques to develop a model that forecasts the condition of water pipelines. The model considers the interdependency and uncertainty of the factors. It developed indices to calculate pipeline conditions based on age, diameter, material type, size, installation quality, surface type, ground water depth and quality, soil type, C-factor, and breakage rate. The model did not provide a mathematical function to facilitate the calculation process. Moreover, it did not predict the deterioration of the pipelines during their service lives. The National Guide to Sustainable Municipal Infrastructure [7] classified the variables that affect the deterioration of pipes into three categories: physical, environmental, and operational. Yan and Vairavamoorthy [9] developed their condition rating model using physical and environmental factors only. The model considered pipe age, diameter, and material as physical factors, and road loading, soil condition, and surroundings as environmental factors. Furthermore, it was limited to one soil type. Geem [10] developed another condition rating model that included seven physical and environmental factors, namely, pipe age, material and diameter, bedding condition, corrosion, temperature, and trench width. However, the model development relied on randomly-generated data. Al Barqawi and Zayed [11] considered soil type, road surface, pipe depth, diameter, material, age, number of breaks, and C-factor while assessing the condition of pipelines. But, the developed model did not account for the interdependency and uncertainty of the factors.

Regression analysis was used herein to generate water pipeline deterioration models for various classifications. Robust deterioration assessment models using regression analysis were developed for several infrastructure types [12], [20]-[22]. Regression models represent the mathematical best-fit representation of a database given several constraints [23]. Firstly, the errors around the best fit are independent from the predictor variables. Secondly, the errors around the best fit are constant for all variables. Finally, the errors are normally distributed around the best fit. The independent variables can be either quantitative (i.e. numerical) or categorical (i.e. classification). The variables that include pipeline characteristics (e.g. material type) and surrounding environment (e.g. soil type) can be considered as quantitative (numerical). However, they lose their significance in the statistical tests of regression analysis. Consequently, it is helpful to consider such variables as categorical.

IV. Research Methodology

Fig. 1 summarizes the developed deterioration assessment methodology. Firstly, an extensive literature review was conducted to compile previous studies on the deterioration assessment of water pipelines and determine the main factors affecting the deterioration of water pipelines. Secondly, historical data on water pipeline networks was gathered from several Canadian cities. The collected data was then used to compute the pipeline deterioration indices using the model developed by El Chanati et al. [4]. The results obtained created an initial regression analysis platform.

The historical data and computed deterioration indices were used as input in the regression analysis. Two linear regression analysis methods were used to develop the deterioration assessment models. Several input variables (e.g. installation quality, soil type, material type, and ground water depth) were qualitatively valued where; all the qualitative input variables were numerically coded as shown in Table I. Those values were obtained from the effect values reported by El Chanati et al. [4]. Regression models were developed to compute the deterioration index of several pipeline categories. The models needed the values of all input variables to compute the output.

However, the values of few input variables were missing in the database. As a result, categorical regression models were developed to address this limitation. Input variables, such as installation quality, soil type, material type, and ground water depth, were qualitatively described in categorical regression models. Given the fact that each categorical variable should have at least one equation, choosing numerous qualitative input variables in one model results in large number of equations. Consequently, this research has limited the number of categorical variables through comparing the results obtained using several combinations of these qualitative variables. The developed regression models were validated using a dataset representing 20% of the collected data. A sensitivity analysis was carried out to quantify the importance of each factor in the developed regression models. The pipeline age was found to be the highest impacting factor on the pipeline deterioration.

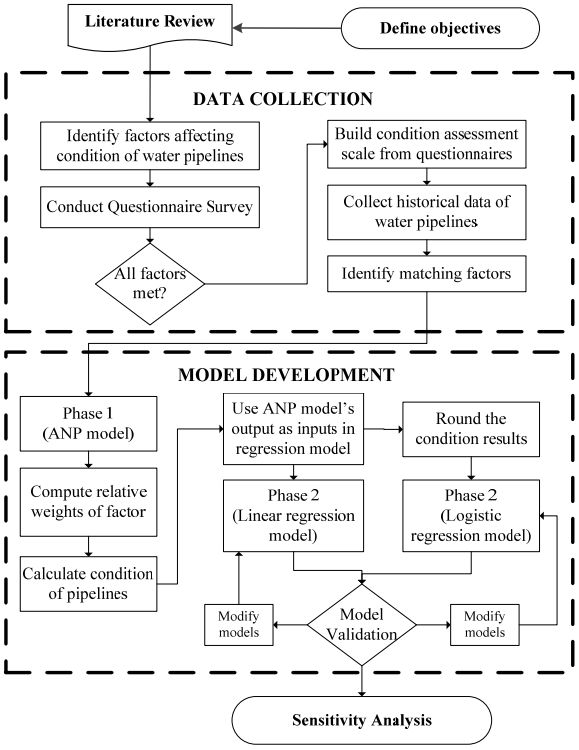

Fig. 1 Research methodology

TABLE I
NUMERICAL VALUES USED TO CODE QUALITATIVE VARIABLES

Variable	Characteristic	Numerical Code
Installation Quality (IQ)	Poor	2
	Fair	6
	Good	10
Ground Water (GW)	Shallow	2
	Moderate	5
	Deep	10
Soil Type (ST)	Aggressive	2
	Moderate	5
	Non-Aggressive	10

V. DATA COLLECTION

The literature review was used to identify the factors affecting the deterioration of water pipelines. The data collection included two main steps. Firstly, a set of data was collected from a Canadian municipality. The data included the following pipe characteristics: age, material type, size, breakage rate, C-factor, water quality, and surface type. The performance indices of the pipes in the collected database were computed using the model developed by El Chanati et al. [4]. The contributing factors' data and the estimated pipe deterioration indices were the inputs of the regression models. The factors affecting the deterioration of water pipelines were identified through an extensive literature review. As shown in Fig. 2, the factors were grouped into three main categories: Physical, Environmental, and Operational [11]. The pipeline physical factors included material type, age, size (i.e. diameter), and installation quality. The Environmental category includes ground water depth, soil type (i.e. aggressive or non-aggressive), and location. Finally, the Operational category includes the flow velocity or C-factor, breakage rate, and water quality.

VI. MODEL DEVELOPMENT

A. Best Subset Analysis

In order to develop each model, training and testing datasets were randomly prepared. The training and testing datasets represented 80 and 20% of the database, respectively. Several variable combinations were considered during the model development. Combinations of variables are determined through the subset analysis and with/without considering their diameter sizes. The analysis of this classification was based on the failure modes prediction guide for water pipes [7]. The best variable combinations were determined using the best subset analysis. The best subset of the variables was determined using four main criteria: coefficient of determination (R^2), adjusted R^2, mean square error (S or MSE), and Mallow C_p. The R^2 value varies between zero and one where; an R^2 value close to one indicates a higher efficiency of the model in fitting the data. The other indicator used to determine the best subset is the Mallow C_p where; a smaller Mallow C_p normally indicates that the model can predict future outcomes in an unbiased manner. Among the tested subset models, the one with the number of selected variables and constants closest to the Mallow C_p value is the most precise. The results of the best subset analysis for the overall regression model are shown in Table II. Although the coefficient of determination value for most of the subsets is acceptable, the last subset of variables was selected herein. The Mallow C_p's value for this subset is exactly equal to the number of variables (i.e. seven) plus one (i.e. number of constants). The difference of these two values (i.e. Mallow C_p's value and number of variables plus one) for the variables of the sixth subset, except the breakage rate, is 0.8, which is satisfactory. However, the breakage rate variable is one of the most important indicators that affect the pipeline deterioration. In other subsets, the Mallow C_p's value is significantly larger

than the number of variables plus one, which signifies the incapability of those models to accurately predict the outcome in an unbiased manner. Therefore, the last subset was selected for the model development. The selected subset includes the following variables: age, pipe size, C-factor, installation quality, ground water depth, soil type, and breakage rate. The model has an R^2 equal to 96.5%. After performing a correlation analysis, the Age and C-factor variables displayed a high Pearson correlation value of 0.969, which implies the interdependency between them and thus, they cannot be used together in the same regression model. Consequently, the variable of Age and C-factor were not included in the same model.

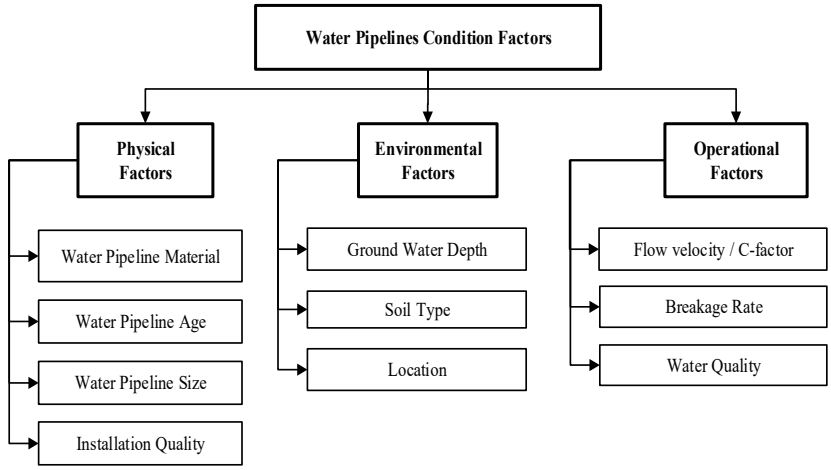

Fig. 2 Factors affecting water pipelines condition

TABLE II
BEST SUBSET ANALYSIS

Variables	R-Sq	R-Sq (adj)	Mallows C_p	S	Age	Size	C factor	I.Q.	Ground Water	Soil Type	Breakage rate
1	94.9	94.9	382.9	0.41	*						
1	88.4	88.4	2004.7	0.63			*				
2	95.6	95.6	227.6	0.39	*				*		
2	95.2	95.2	309.3	0.40	*		*				
3	95.9	95.9	309.3	0.37	*	*			*		
3	95.9	95.9	139.2	0.37	*			*	*		
4	96.3	96.2	57.5	0.36	*	*		*	*		
4	96.1	96.1	92.2	0.36	*	*	*		*		
5	96.4	96.4	20.5	0.35	*	*	*	*	*		
5	96.4	96.3	33.7	0.35	*	*		*	*	*	
6	96.5	96.5	6.2	0.35	*	*	*	*	*	*	
6	96.4	96.4	21.9	0.35	*	*	*	*	*		*
7	**96.5**	**96.5**	**8**	**0.35**	*	*	*	*	*	*	*

B. Traditional Regression Models

Since age and C-factor cannot be included in the same model as deduced previously, two separate water pipeline deterioration assessment models were developed for each category. The first category of the model was called "Overall Model", which included the pipe diameter as one of the variables. Two other categories, namely large and small size pipes, were used based on the pipeline diameter size (i.e. large > 300 mm and small < 300 mm). Separate models were developed for each category. Table III shows six common regression models with quantitative variables and four categorical regression models. The first and second models included the Age and C-factor, respectively. They were developed for pipelines without any diameter classification. The third model included the age. Due to the fact that the P-value of soil type was not satisfactory, this variable was removed from the third model. On the other hand, the fourth model included four variables besides C-factor. The fifth model included the installation quality, ground water depth, age, and soil type. Finally, the sixth model included the breakage rate and C-factor and excluded the age.

C. Categorical Regression Models

The last four models (i.e. 7 to 10) shown in Table III were developed by considering few variables as categorical. Various combinations of categorical variables were tested during the model development process. The variables that did not yield logical results were removed from the developed models. The variables, which did not significantly change the deterioration of various categories, were also removed from the developed models. The first developed categorical model considered the pipeline material type as a category. All material types were initially used to develop categorical models. However, asbestos, concrete, and cast iron pipes

generated similar results. Therefore, they were grouped into one category. Finally, four deterioration assessment equations were developed based on the pipeline material type category. The only difference between various categorical regression equations is their constant values as shown in Table IV. It is obvious that the first material group yielded lower pipeline deterioration. On the other hand, the fourth material group yielded the highest pipeline deterioration. Table V summarizes the material type categories and their descriptions.

TABLE III
CONDITION ASSESSMENT MODELS

Models		Equation
Common Regression Models	1: Overall Water Pipelines (With age)	**CI** = 7.54 - 4.39***AG** + 0.225***DI** + 0.437***IQ** + 0.542***GW** + 0.329***ST** - 0.309***BR**
	2: Overall Water Pipelines (With C-factor)	**CI** = 3.28 + 2.97***CF** + 0.274***DI** + 0.848***IQ** + 0.811***GW** + 0.785***ST** - 0.608***BR**
	3: Small Pipelines (With age, removing soil type)	**CI** = 7.93 - 4.27***AG** + 0.328***IQ** + 0.303***GW** - 0.872***BR**
	4: Small Pipelines (With C-factor)	**CI** = 3.91 + 2.87***CF** + 0.670***IQ** + 0.485***GW** + 0.409***ST** - 1.23***BR**
	5: Large Pipelines (With age)	**CI** = 7.64 - 4.58***AG** + 0.412***IQ** + 0.620***GW** + 0.528***ST**
	6: Large Pipelines (With C-factor)	**CI** = 3.23 + 3.15***CF** + 0.797***IQ** + 0.899***GW** + 1.06***ST** - 0.498***BR**
Categorical Regression Models	7: Material Type	**CI**= C_1 - 4.68884***AG** + 0.358854***DI**+ 0.439668***IQ** + 0.065772***GW** + 0.240868*** ST** - 0.137076***BR**
	8: Soil Type	**CI**= C_2 - 5.03383***AG** + 0.30422***DI**+ 0.356489***IQ** + 0.058199***GW** - 0.171302***BR**
	9: Size and Material Type	**CI**= C3 - 4.57835***AG** + 0.424003***IQ** + 0.502984***GW** + 0.23131***ST** - 0.203788***BR**
	10: Size, Material Type and GWD	**CI**= C4 - 4.54795***AG** + 0.41312***IQ** + 0.228297***ST** -0.315282*** BR**

CI = Condition Index, AG = Age, CF = C factor, DI = Diameter, IQ = Installation Quality, GW = Ground Water Depth, ST = Soil Type, BR = Breakage Rate

TABLE IV
MODELS' CONSTANT VALUES

Constant Values	Model No. 1	2	3	4	5	6
C1:	**7.8055**	7.5677	7.7095	**7.5030**	-	-
C2:	**7.9366**	8.0983	**8.1039**	-	-	-
C3:	**8.2773**	8.0885	8.2257	7.9908	8.0150	7.8262
C4:	**8.8106**	8.3663	8.3542	8.6525	8.2082	8.1961
C4 (+12):	8.5241	8.0798	8.0677	8.3660	7.9217	7.9096
C4 (+24):	8.4168	7.9725	7.9604	8.2587	7.8144	7.8023

Constant Values	Model No. 7	8	9	10	11	12
C1:	-	-	-	-	-	-
C2:	-	-	-	-	-	-
C3:	7.9635	7.7286	7.8824	7.6936	7.8309	**7.5960**
C4:	8.7661	8.3218	8.3097	8.5162	8.0719	8.0598
C4 (+12):	8.4796	8.0352	8.0231	8.2297	7.7853	7.7732
C4 (+24):	8.3723	7.9280	7.9158	8.1224	7.6780	**7.6659**

The first material type category included polyethylene pipelines. The fourth material type included asbestos, concrete, and cast iron pipelines. The eighth model included different condition assessment equations using the soil type as a category. As shown in Table IV, the difference between the highest and the lowest conditions was around 0.2 (i.e. 8.1-7.9 = 0.2) condition units. As shown, the highest condition was obtained for non-aggressive soil types. On the other hand, the lowest condition was obtained for aggressive soil type because of its deteriorating nature. The ninth model combined the effect of two categorical variables: size and material type. Three pipeline size groups, large, medium, and small, were used to develop the model. The size classification was selected based on the results of testing the closeness of various size groups. The model included 12 deterioration assessment equations. The highest condition value was obtained for polyethylene pipelines. On the other hand, the lowest condition was obtained for small asbestos, concrete, and cast iron pipelines. The ground water depth was considered as a categorical variable for the tenth model. 36 deterioration assessment equations were developed based on pipeline size, material, and ground water depth categories. The ground water depth included three categories: deep, moderate, and shallow. The large polyethylene pipelines buried in deep ground water locations resulted in the highest condition values. On the other hand, small pipelines from asbestos, concrete and cast iron located in shallow ground water resulted in the lowest condition values.

TABLE V
CATEGORIES' DEFINITION

Constant Values	Category No. 1	2	3	4	5	6
C1: MT	M1	M2	M3	M4	-	-
C2: ST	S1	S2	S3	-	-	-
C3: DI, MT	D1, M1	D1, M2	D1, M3	D1, M4	D2, M1	D2, M2
C4: DI, MT, GWD	D1, M1,G1	D1, M1,G2	D1, M1,G3	D1, M2,G1	D1, M2,G2	D1, M2,G3
C4 (+12):	D2, M1,G1	D2, M1,G2	D2, M1,G3	D2, M2,G1	D2, M2,G2	D2, M2,G3
C4 (+24):	D3, M1,G1	D3, M1,G2	D3, M1,G3	D3, M2,G1	D3, M2,G2	D3, M2,G3

Constant Values	Category No. 7	8	9	10	11	12
C1: MT	-	-	-	-	-	-
C2: ST	-	-	-	-	-	-
C3: DI, MT	D2, M3	D2, M4	D3, M1	D3, M2	D3, M3	D3, M4
C4: DI, MT, GWD	D1, M3,G1	D1, M3,G2	D1, M3,G3	D1, M4,G1	D1, M4,G2	D1, M4,G3
C4 (+12):	D2, M3,G1	D2, M3,G2	D2, M3,G3	D2, M4,G1	D2, M4,G2	D2, M4,G3
C4 (+24):	D3, M3,G1	D3, M3,G2	D3, M3,G3	D3, M4,G1	D3, M4,G2	D3, M4,G3

MT: Material Type, M1: Polyethylene, M2: PVC, M3: Ductile Iron, M4: Concrete, Asbestos, and Cast Iron, ST: Soil Type, S1: Aggressive, S2: Moderate, S3: Non-Aggressive, DI: Pipes' Diameter, D1: Pipes larger than 450 mm, D2: Pipes between 250 & 350 mm, D3: Pipes smaller than 250 mm, GWD: Ground Water Depth, G1: Deep, G2: Moderate and G3: Shallow ground water.

D.Statistical Tests

Statistical tests were carried out to validate the models where F-test and t-test were conducted to study the significance of the parameters. In the F-test, the null Hypothesis (H_0) assumes that all coefficients (i.e. $\beta_0, \beta_{1,...,} \beta_{p-1}$) are equal to zero. The alternate hypothesis (H_a) assumes that at least one variable has a non-zero coefficient (β_k). P (F) shows the results of the F-test. If P (F) is lower than the confidence interval, the null Hypothesis is rejected, which means that at least one coefficient is non-zero. The confidence interval of the test (α) was assumed as 0.05. All models achieved P-values equal to zero, implying their validity. The coefficient of multiple determination (R^2) is another diagnostic measure of the model that checks the variation of data around the fitted model. A higher correlation shows that there is a little variation around the fitted model. All models displayed correlation coefficients greater than 94%, with little variance around the fitted line.

The t-test was performed for each variable separately to check their significance. The Null Hypothesis (H_0) of the test assumed that the coefficient of the selected variable was equal to zero, while the alternate Hypothesis (H_a) assumed that it non-zero. If the P-value was less than the confidence interval, the Null Hypothesis is rejected and the variable is significant to the model. ANOVA results for all models are shown in Table VI, which includes the variable coefficients, SE coefficients, T-values, and P-values. Most of the variables had P-values equal to zero. This means that the variables were significant to the models and were therefore accepted. The only exception was the breakage rate variable in the seventh and eighth models, which showed a negligible non-zero P-value. Variables were normalized to be in the similar range as shown in Table VII. For example, the actual age was divided by 90 to obtain a normalized age value.

TABLE VI
ANOVA RESULTS FOR MODELS' COEFFICIENTS

Model	Predictor	Constant	AG	DI	CF	IQ	GWD	ST	BR
No. 1	Coef.	7.538	-4.394	0.225	-	0.437	0.542	0.329	-0.309
	SE Coef.	0.096	0.08	0.026	-	0.052	0.041	0.048	0.066
	T	78.82	-54.59	8.59	-	8.47	13.27	6.9	-4.71
	P	0	0	0	-	0	0	0	0
No. 2	Coef.	3.285	-	2.966	0.274	0.848	0.811	0.785	-0.608
	SE Coef.	0.057	-	0.087	0.036	0.069	0.055	0.062	0.091
	T	57.14	-	33.99	7.55	12.36	14.75	12.67	-6.66
	P	0	-	0	0	0	0	0	0
No. 3	Coef.	7.93069	-4.2739	-	-	0.32803	0.303	-	-0.87
	SE Coef.	0.06571	0.05868	-	-	0.04651	0.03776	-	0.05464
	T	120.7	-72.83	-	-	7.05	8.03	-	-15.96
	P	0	0	-	-	0	0	-	0
No. 4	Coef.	3.909	-	-	2.867	0.67	0.485	0.40914	-1.2265
	SE Coef.	0.06574	-	-	0.09057	0.074	0.06	0.06957	0.08999
	T	59.47	-	-	31.66	9.03	8.04	5.88	-13.63
	P	0	-		0	0	0	0	0
No. 5	Coef.	7.6398	-4.5812	-	-	0.41198	0.62047	0.52786	-
	SE Coef.	0.1475	0.1327	-	-	0.08203	0.06951	0.0762	-
	T	51.79	-34.52	-	-	5.02	8.93	6.93	-
	P	0	0	-	-	0	0	0	-
No. 6	Coef.	3.22519	-	-	3.147	0.7972	0.89923	1.06224	-0.5
	SE Coef.	0.08624	-	-	0.14	0.111	0.0905	0.09329	0.193
	T	37.4	-	-	22.48	7.18	9.94	11.39	-2.58
	P	0	-	-	0	0	0	0	0.01
No. 7	Coef.	C1	-4.6888	0.3589	-	0.4397	0.0658	0.2409	-0.1371
	SE Coef.	D1	0.0776	0.0313	-	0.0378	0.0039	0.0324	0.0715
	F	92.6028	-60.3972	11.4717	-	11.634	16.9169	7.4407	-1.9175
	P	0	0	0	-	0	0	0	0.056
No. 8	Coef.	C_2	-5.03383	0.30422	-	0.35649	0.0582	-	-0.1713
	SE Coef.	D2	0.071558	0.031639	-	0.037928	0.003953	-	0.075098
	T	103.145	-70.346	9.615	-	9.399	14.722	-	-2.281
	P	0	0	0	-	0	0	-	0.023
No. 9	Coef.	C_3	-4.57835	-	-	0.424	0.50298	0.23131	-0.20379
	SE Coef.	D3	0.077679	-	-	0.037816	0.030907	0.034292	0.072873
	T	110.556	-58.939	-	-	11.212	16.274	6.745	-2.796
	P	0	0	-	-	0	0	0	0.005
No. 10	Coef.	C_4	-4.54795	-	-	0.41312	-	0.2283	-0.31528
	SE Coef.	D4	0.076361	-	-	0.037134	-	0.033635	0.073994
	T	129.243	-59.559	-	-	11.125	-	6.787	-4.261
	P	0	0	-	-	0	-	0	0

E. Residual Analysis

After the primary statistical tests, the residuals of the models were checked for normality error, homoscedasticity, and independence of error. For the normality error test, the

residuals' distribution was compared with the normal probability distribution (NPD). A normal distribution of the residuals showed that the errors were normally distributed, which validated the efficiency of the models. A small deviation from the normal distribution is usually tolerated. The deviations determine the possibility of the presence of outliers. Removing the outliers resulted in a higher correlation. However, the outliers were saved as they might represent important data patterns. Moreover, the distribution of the residuals around the fitted values was investigated. According to Kutner et al. [24], the symmetry of the distribution of the residuals around the fitted value is due to the consistency of the variance around the fitted values. The plots of residuals in all models proved the reliability and soundness of the models.

F. Model Validation

The generated models were validated using the testing dataset. In this step, the generated outputs were compared with the actual ones. The efficiency of the models was also measured through various indicators as computed in (1)-(4):

$$AIP = \left\{\sum_{i=1}^{n}\left|1-\left(\frac{E_i}{C_i}\right)\right|\right\}\times\frac{100}{n} \quad (1)$$

$$AVP = 100 - AIP \quad (2)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}(C_i - E_i)^2/n} \quad (3)$$

$$MAE = \frac{\sum_{i=1}^{n}|C_i - E_i|}{n} \quad (4)$$

where: AIP = Average Invalidity Percent; AVP = Average Validity Percent; $RMSE$ = Root Mean Squared Error; MAE = Mean Absolute Error; E_i = estimated value; C_i = actual value; and n = number of events.

TABLE VII
FACTORS' NORMALIZATION METHOD

Predictor		Age	Diameter	C-factor	Installation Quality	Ground Water	Soil Type	Breakage Rate
Normalization	Unit	Years	Inches	NA	NA	Meters	NA	Failures/ kilometer/ year
	Method	AG/ 90	(DI-150)/ 300	(CF-20)/ 105	(IQ-2)/ 10	(GW-2)/ 8	(ST-2)/8	BR/4

TABLE VIII
MODELS' VALIDATION RESULTS

Models	Validation Technique					
	P-Value	R^2 (%)	AIP (%)	AVP (%)	RMSE	MAE
No. 1	0.000	97.2	5.21	94.79	0.01	0.32
No. 2	0.000	94.6	3.58	96.42	0.01	0.23
No. 3	0.000	98.4	2.17	97.83	0.01	0.12
No. 4	0.000	95.8	3.86	96.14	0.01	0.21
No. 5	0.000	96.6	4.25	95.75	0.02	0.29
No. 6	0.000	94.1	5.26	94.74	0.02	0.35
No. 7	0.000	96.7	4.06	95.94	0.02	0.24
No. 8	0.000	96.4	4.23	95.77	0.22	0.27
No. 9	0.000	96.6	4.12	95.88	0.01	0.26
No. 10	0.000	96.8	3.83	96.16	0.01	0.24

The validation results and the models' correlation coefficients are shown in Table VIII. The Average Invalidity Percentage (AIP) values of the models were mostly less than 5%, which implies the validity of the developed models. The third model obtained the lowest AIP value of 2.17% and the highest Average Validity Percentage (AVP) value of 97.83%. The Root Mean Squared Error (RMSE) was 0.01 for most of the models except for the fifth and sixth models, which displayed an RMSE of 0.02, which was satisfactory and showed the reliability of the produced models. All models achieved Mean Absolute Error (MAE) values less than 0.35, which also proved the validity of the developed models. The third model obtained the lowest MAE value. The validation plots of the actual and generated conditions for the whole dataset are shown in Fig. 3. All models showed satisfactory results as there is a little discrepancy between the lines representing the pipeline actual and predicted deterioration indices.

G. Discussion on the Application of Developed Models

This section discusses the advantages of the different models and their application in an actual water pipeline network case study. In order to predict the deterioration and condition of pipelines, the developed models can be applied, depending on the availability of the data. When the data for all variables are available, the overall water pipeline model should be preferably used. However, data on some variables (e.g. pipe diameter) might be missing from historical databases of large water distribution networks. The missing data do not allow the user to use the overall models (i.e. first and second models) while computing the condition of the pipes. The traditional regression models for small and large pipes overcome this limitation and enable the user to compute the condition of the pipes by only knowing the pipe size category. For large diameter category (e.g. most of water mains), the model for large diameter pipes will be used. After choosing the model (Table III), the condition of the pipes is calculated by substituting the value of each variable in the regression model. For example, let us consider a pipe with the following data: 1) AG:5, 2) DI:250, 3) IQ: Good, 4) GW: Deep depth, 5) ST: Moderate, and 6) BR: 0.01. The equivalent numerical values for the qualitative variables are obtained using Table I. The first model (with age) is selected (i.e. **CI** = 7.54 - 4.39***AG** + 0.225***DI** + 0.437***IQ** + 0.542***GW** + 0.329***ST** - 0.309***BR**). Then, the actual values are normalized using the equations in Table VII. Finally, the variables are substituted with their actual values and the condition index is computed, which was estimated, in this case, at 8.55.

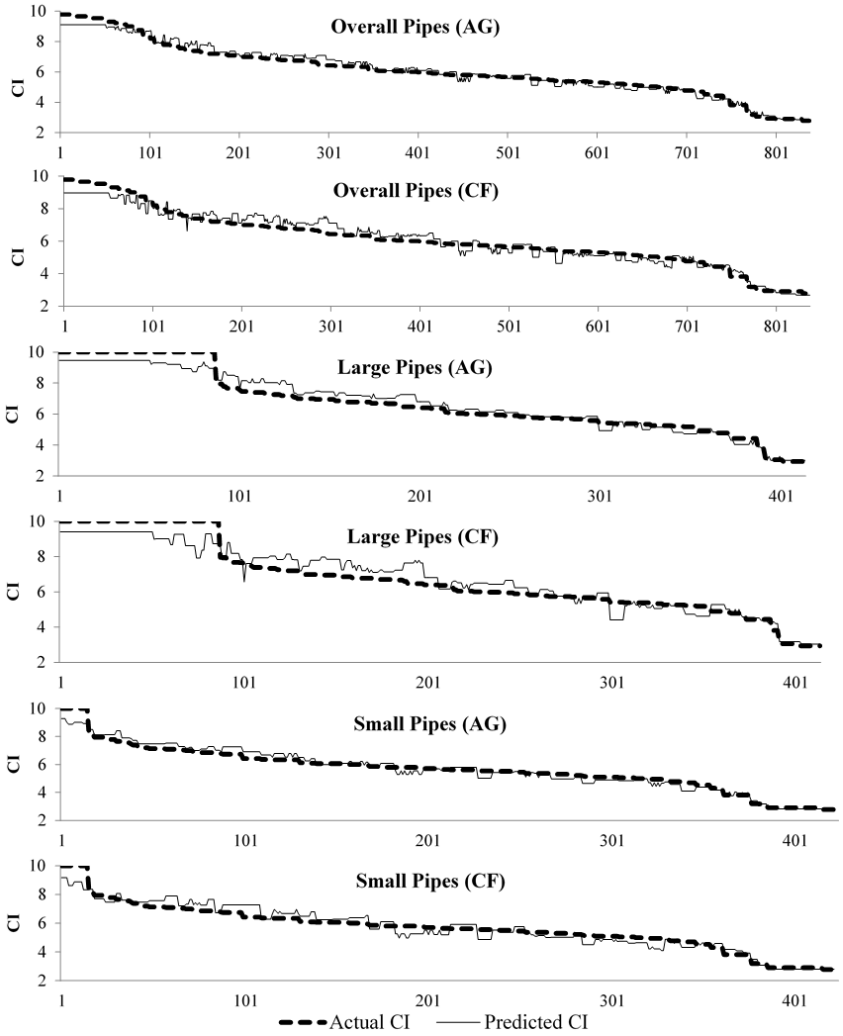

Fig. 3 Models' validation results

Traditional regression models require data for all input variables including age, diameter, installation quality, ground water depth, soil type, and breakage rate. In a large pipeline network of a city, data for certain numerical and categorical variables might be missing. For example, municipalities might not have enough information regarding the soil type and water depth around the pipes in different parts of the city. In categorical models, there are different models that can be used when the data for one or more variables are not available. For example, if the soil type in various pipeline beddings is not available, the eighth model may be used. In the presence of an aggressive soil type, the constant value is equal to 7.9366 as shown in Table IV according to the first model of the eighth model group. If the soil type is considered as non-aggressive, the constant value will be 8.1039. Consequently, the estimated value of the pipeline's actual deterioration will vary with a range of 0.2 condition units. If material type and ground water depth data are missing, the tenth model might be used to forecast the condition of the pipes. If the pipeline size is known, the number of categories is decreased to 12. The maximum variation between various classifications of the unknown variables is about 0.8 condition units. As a result, the categorical regression models can be used to compute the deterioration of pipeline sections in a large city network when part of the required data is not available. Unlike traditional regression models, categorical models consider pipeline material types, which are important in the deterioration assessment.

Three main steps are used to compute the deterioration index of pipelines using categorical regression models: 1) select the category of the pipe from Table V; 2) locate related constant values from Table IV; and 3) insert the constant values into the equations of Table III. For example, let us consider a 450 mm-diameter pipeline made of polyethylene and buried in a shallow ground water location. According to Table V, the pipeline that belongs to the category of "D3, M1, G3" is related to group no. 27. The associated constant value of 7.9604, which is obtained from Table IV, is inserted into the deterioration assessment equation for the eleventh regression model. Finally, the deterioration of the pipe is

estimated by replacing the values of age, installation quality, soil type, and breakage rate.

VII. Deterioration and Breakage Rate Forecasting Model

Deterioration profiles can be used to estimate the deterioration of water pipelines during their service lives and optimize the maintenance and rehabilitation of water pipelines during their life cycle. In this study, the collected data were used to build a model that predicts water pipeline deterioration based on its age. A cubic regression line was the best-fit model with an R^2 of 98.8%, as shown in Fig. 4. Furthermore, the correlation of water pipeline deterioration with its breakage rate was investigated to find a model that forecasts the breakage rate of water pipelines based on their deterioration condition. A model was developed to estimate the breakage rate of pipelines by considering first, second, and third order regression equations of their deterioration indices. The best regression equation was found to be of a third order as presented in Fig. 5. The developed model yielded a determination coefficient equal to 94.6%, which proves the efficiency of the model. Equation (5) predicts the condition of the pipe during its service life, while (6) correlates pipe breakage rate with its condition. The model predicts the breakage rate using the pipeline actual deterioration index. The investigation of pipeline deterioration index versus the age did not result in an efficient model and the correlation coefficient was too small to report.

$$\boldsymbol{CI} = 9.787 - 0.1717 \times \boldsymbol{AG} + 0.002883 \times \boldsymbol{AG^2} - 0.00002 \times \boldsymbol{AG^3} \quad (5)$$

$$\boldsymbol{BR} = 13.96 - 5.543 \times \boldsymbol{CI} + 0.7256 \times \boldsymbol{CI^2} - 0.03113 \times \boldsymbol{CI^3} \quad (6)$$

where: CI = pipe condition index, AG = pipe age, and BR= pipe breakage rate.

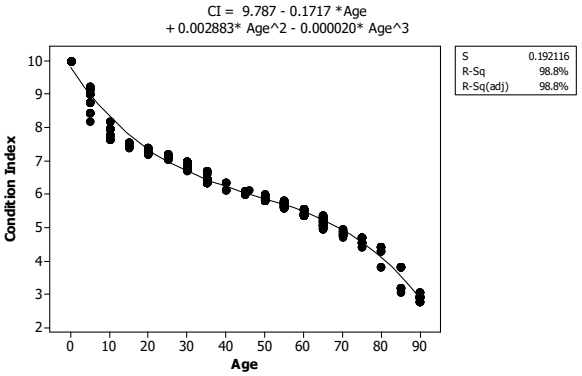

Fig. 4 Model to forecast breakage rate through condition index (CI)

The developed models can estimate the overall condition and breakage rate of the pipes during their service lives without any additional data. The models are developed in a generic format to predict the values of the pipe condition and its breakage rate during its service life. To forecast the overall condition index and breakage rate of a pipe, its condition is first estimated by replacing the age value in (5). Then, the computed condition index is inserted into (6) to find the associated breakage rate of the pipe. For example, let us assume a user wants to compute the tenth-year breakage rate of a pipe. First, 10 is substituted with the age factor in (5) and the Condition Index (CI) is calculated; which results a value of 8.3383 in this case. Then, the computed CI is inserted into (6) to compute the Breakage Rate (BR), which was estimated at 0.142 failures/year/km in this case. However, this is a very generic application of these models. A condition assessment model would be more accurate in predicting the breakage rate.

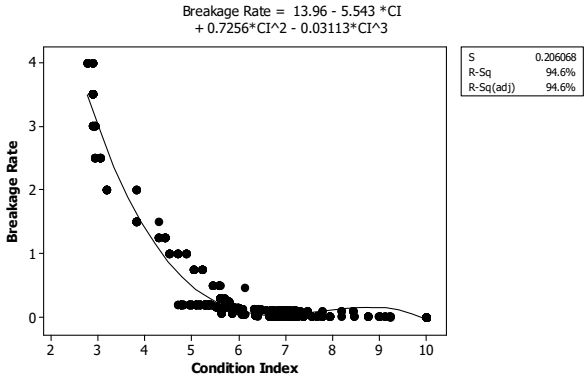

Fig. 5 Model to forecast breakage rate via condition index (CI)

VIII. Conclusion

This study developed deterioration assessment models for water pipelines. The factors contributing to the deterioration of the pipelines were identified through an extensive literature review. These factors were classified into three groups: physical, environmental, and operational. Water pipeline historical data were gathered from a Canadian municipality. The database was divided into training (80%) and testing (20%) datasets. Several combinations of the variables were considered. Six regression models were developed assuming numerical variables where; two of the models were developed for water pipelines, regardless of their diameter size; and the other two models were developed for small and large pipelines, respectively.

Categorical regression models were also developed considering different categories of pipeline material, diameter size, ground water depth, and soil type. The regression equations of the categorical models differed only in the constant value. The developed models were validated, using the testing dataset, and displayed an AVP of 95%. The correlation coefficients of the models were estimated to be greater than 96%. Sensitivity analysis, which was performed for non-categorical regression model, visualized the importance of age compared to the other variables. The categorical models yielded higher conditions for larger and polyethylene pipelines. The pipelines buried in non-aggressive soils and deep ground waters were subject to less deteriorating conditions. The developed models will help water pipeline operators and municipalities in the deterioration assessment of

their water network, even in the absence of some missing data. Finally, two more models were developed to predict the deterioration of pipelines during their service lives. These models can be used to forecast the condition profile and breakage rate of pipelines during their life cycles. Such models will assist the municipalities in prioritizing the maintenance and replacement of their water pipelines.

Acknowledgments

The authors gratefully acknowledge the support provided by Qatar National Research Fund (QNRF) for this research project under award No. QNRF-NPRP 4-529-2-193.

References

[1] American Society of Civil Engineers (2017). "Report Card for America's Infrastructure", <http://www.infrastructurereportcard.org>, (March 2017).

[2] American Water Works Association (AWWA) (2012). "Buried No Longer: Confronting America's Water Infrastructure Challenge.", <http://www.awwa.org/Portals/0/files/legreg/documents/BuriedNoLonger.pdf>, (May 2015).

[3] Canadian Infrastructure Report Card (2016). "Informing the Future", <http://www.canadainfrastructure.ca/en/index.html>, (March 2017).

[4] El Chanati, H., El-Abbasy, M.S., Mosleh F., Senouci, A., Abouhamad, M., Gkountis, I., Zayed, T., and Al-Derham, H. (2015). "Multi-Criteria Decision Making Models for Water Pipelines", ASCE, Journal of Performance of Constructed Facilities, 30(4), 04015090.

[5] Giustolisi, O., Laucelli, D., & Dragan, A. S. (2006). "Development of rehabilitation plans for water mains replacement considering risk and cost-benefit assessment." J. Civil Engineering and Environmental Systems, 23(3), 175-190.

[6] Kleiner, Y., & Rajani, B. (2000). "Considering time-dependent factors in the statistical prediction of water main breaks." Proc. Infrastructure Conference, AWWA, 1-12.

[7] NRC. (2003). "Deterioration and Inspection of Water Distribution Systems" National guide to sustainable municipal infrastructure, Issue No. 1.1, Ottawa, Ontario, Canada.

[8] Kleiner, Y., & Rajani, B. (2001). "Comprehensive review of structural deterioration of water mains: Physical models." J. Urban Water, 3(3), 151-164.

[9] Yan, J.M., and Vairavamoorthy, K. (2003). "Fuzzy Approach for Pipe Condition Assessment", ASCE, Pipeline Engineering and Construction International Conference, Baltimore, Maryland, USA.

[10] Geem, Z.W. (2003). "Window-Based Decision Support System for the Water Pipe Condition Assessment using Artificial Neural Network", ASCE, World Water and Environmental Resources Congress, Philadelphia, Pennsylvania, USA.

[11] Al-Barqawi, H., and Zayed, T. (2006). "Condition Rating Model for Underground Infrastructure Sustainable Water Mains", ASCE, Journal of Performance of Constructed Facilities, Vol. 20, No. 2, pp. 126-135.

[12] Al-Barqawi, H., and Zayed, T. (2006). "Assessment Model of Water Main Conditions", ASCE, Pipeline Division Specialty Conference, Chicago, Illinois, USA.

[13] Geem, Z. W., Tseng, C.L., Kim, J., and Bae, C. (2007). "Trenchless Water Pipe Condition Assessment using Artificial Neural Network", ASCE, International Conference on Pipeline Engineering and Construction, Boston, Massachusetts, USA.

[14] Al-Barqawi, H., and Zayed, T. (2008). "Infrastructure Management: Integrated AHP/ANN Model to Evaluate Municipal Water Mains' Performance", ASCE, Journal of Infrastructure Systems, Vol. 14, No. 4, pp. 305-318.

[15] Wang, Y., Zayed, T., and Moselhi, O. (2009). "Prediction Models for Annual Break Rates of Water Mains", ASCE, Journal of Performance of Constructed Facilities, Vol. 23, No. 1, pp. 47-54.

[16] Zhou, Y., Vairavamoorthy, K., and Grimshaw, F. (2009). "Development of a Fuzzy Based Pipe Condition Assessment Model using PROMETHEE", ASCE, The 29th World Environmental and Water Resources Congress, Kansas City, Missouri, USA.

[17] Fares, H., and Zayed, T. (2010). "Hierarchical Fuzzy Expert System for Risk of Failure of Water Mains", ASCE, Journal of Pipeline Systems Engineering and Practice, Vol. 1, No. 1, pp. 53-62.

[18] Wang, C.W., Niu, Z.G., Jia, H., and Zhang, H.W. (2010). "An Assessment Model of Water Pipe Condition using Bayesian Inference", Journal of Zhejiang University Science A, Vol. 11, No. 7, pp. 495-504.

[19] Clair, A.M.S., and Sinha, S.K. (2011). "Development and the Comparison of a Weighted Factor and Fuzzy Inference Model for Performance Prediction of Metallic Water Pipelines", ASCE, Proceedings of the Pipelines 2011 Conference, Seattle, Washington, USA.

[20] Elhag, T. and Wang, Y. (2007). "Risk Assessment for Bridge Maintenance Projects: Neural Networks versus Regression Techniques", Journal of Computing in Civil Engineering, Volume 21, pp. 402-409.

[21] Chughtai, F., & Zayed, T. (2009). "Infrastructure Condition Prediction Models for Sustainable Sewer Pipelines." Journal of performnce of constructed facilities, 333-341.

[22] El-Abbasy, M., Senouci, A., Zayed, T., Mirahadi, F., and Parvizsedghy, L. (2014). "Condition Prediction Models for Oil and Gas Pipelines Using Regression Analysis." J. Constr. Eng. Manage., 140(6), 04014013.

[23] Levine, D., Stephanm, D., Krehbiel, T., Berenson, M., and Bliss, J. (2002). Statistics for Managers - Using Microsoft Excel, 3rd Edition, Prentice Hall, Upper Saddle River, NJ.

[24] Kutner, M. H., Nachtsheim, C.J., Neter, J., & Li, W. (2004). Applied Linear Statistical Models. 5th Edition, McGraw-Hill, New York, NY.