International Journal of the Physical Sciences Vol. 6(25), pp. 6025-6035, 23 October, 2011
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS11.1105
ISSN 1992 - 1950 ©2011 Academic Journals

Full Length Research Paper

Performance improvement for pipe breakage prediction modeling using regression method

Abdelwahab M. Bubtiena, Ahmed H. ElShafie* and Othman Jaafar

Department of Civil and Structural Engineering, Faculty of Engineering, National University of Malaysia, 43600 UKM, Bangi, Selangor, Malaysia.

Accepted 23 September, 2011

In water distribution systems, the challenge was always to develop reliable models for predicting the failures for each individual pipe in their lifetime to keep the system reliable. The ability of predicting the pipe failure is one of the most important fundamentals for the effective rehabilitation strategy. Statistical methods are the most used techniques in this field. The prediction accuracy reflects the reliability of the proactive rehabilitation strategies. This article introduces a simple technique to improve and enhance the accuracy of non-linear multiple regression prediction models. The method proposed was applied to predict the number of pipe breaks using 7 predictors for actual water distribution system, where a performance improvement of 4.6% was yielded to the prediction model.

Key words: Pipe breakage, multiple regression analysis, residuals, water distribution systems.

INTRODUCTION

Reliable water distribution systems (WDSs) should be able to provide the consumers with the amount of flow under a pressure not less than as designed. Otherwise, the system is unreliable. To keep and improve the performance or reliability of WDSs, notably meeting pressure, flow requirements and water quality standard-sare a major asset for the well-maintained WDSs by continuous monitoring and maintenance. However, WDSs deteriorate over time. The deterioration or failure of pipes in urban WDSs presents a major challenge to water utilities throughout the world, which may result in a reduction in the water carrying capacity of pipes and lead to substantial repair costs.

To date, several studies on modeling of deterioration have been reported to model the failure of water pipes and the effects of the factors that control deterioration. Each of them considered different techniques and parameters in searching for more realistic deterioration or failure rate predictions (Kleiner et al., 2010). The intent has always been the provision of much-accuracy on piping failures prediction to operators of water distribution networks so that they can arrive at intelligent “repair-or-replace” decisions to keep the system reliable (Tabesh et al., 2009; Christodoulou and Deligianni, 2010).

Pipe deterioration usually results from a combination of several factors as shown in Figure 1. Each of these factors has a certain influence on the likelihood of pipe failure. The most influential factors in pipe failure can be classified as: (1) pipe characteristics including type of pipe material, diameter, length, roughness and age, (2) environmental characteristics consisting of type of soil and climate conditions, (3) operating characteristics of the network including pressure variations and (4) manufacturing and installations codes. The influences of these factors are location specific. Consequently, themodeling of this event becomes very complex due to the high variability in failure patterns between different water distribution networks, and also among the pipes of a given network, in addition to weak correlation between some influential factors, non-linearity and difficulty to express some of the parametric relationship of these factors mathematically.

Prediction in general is completely dependent on inspection of the existing status of the system and analyzing of past records. Considering the complicated composition and different characteristics of pipes in WDSs, direct inspection or physical modelingis often prohibitively expensive for all, since it requires a substantial amount of data to represent specific conditions and environments, especially if every pipe in a system

*Corresponding author. E-mail: elshafie@vlsi.eng.ukm.my.

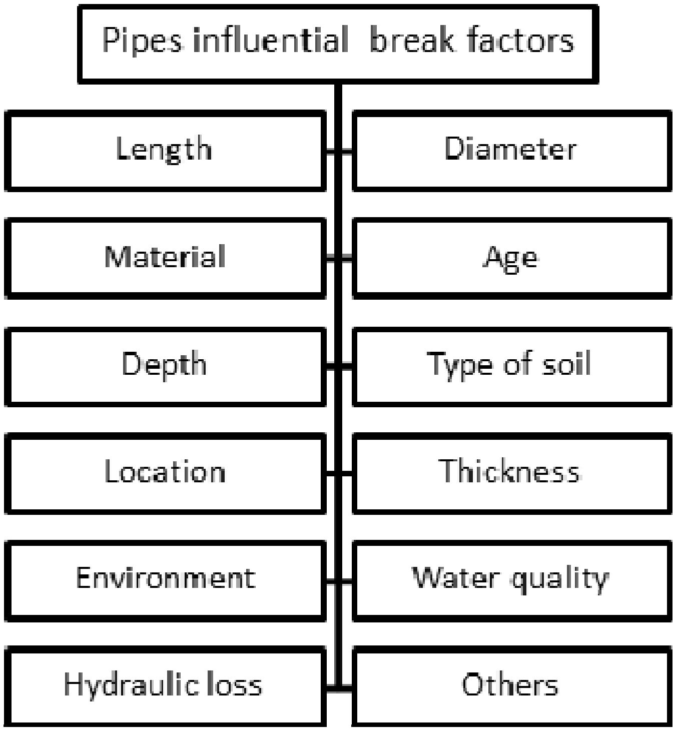

Figure 1. The major influential factors of pipes break.

is aimed(Lawalet al., 2011; Kai and Hua, 2011; Mukhlisin et al., 2011; Ömer, 2011). These data are either unavailable or very costly to obtain for even a modest portion of a distribution network (Kleiner et al., 2009).

Using the assumption that any historical patterns are very likely to continue in the future (if there is no much variety in the general conditions); the predictive models of failure are usually based on identifying breakage pattern using statistical techniques. Identification of breakage patterns over time is an effective and inexpensive alternative to physical modeling. In this case, historical records and information related to each failure event become a suitable failure rate function used in modeling the failure pattern. Therefore, statistical models can develop empirical relationships between the pipe, its exposure to the external and operational environments and its observed failure frequency, and these statistically derived models can be applied with various levels of input data and may thus, be useful for minor water mains for which there are few data available (Park et al., 2008; Ekinci and Konak, 2009; Kleiner et al., 2010).

Statistically, there are many methods can be used for prediction. These methods can be classified into deterministic and probabilistic regression techniques according to the number of predictors that can be modeled, how data is managed in individual cases or in groups and thenature of occurrence of the event being modeled, uniformly or randomly distributed (Bubtiena et al., 2011). The problem can be modeled deterministically when the relationship of the parameters being modeled are known, ignoring any probability of different event sequences and the models are constructed for a condition of assumed certainty, and each input is determined exactly (Mahdi et al., 2011). In the deterministic models constructed for pipe breakage, limited number of break influential factors is usually considered, in addition, the pipes are classified into homogenous groups according to the considered influential factors and the breaks are implicitly assumed uniformly distributed along all the water pipes. Therefore the results of these kinds of models were for pipes group levels (Kleiner at el., 2001).

The regression based models used in modeling the pipe breakage are exponential regression analysis and linear regression analysis. This kind of analysis was first used in pipe breakage modeling (Shamir and Howard, 1979). They used exponential regression analysis to quantify the effects of pipe age upon breaking rate. In addition to the age, Walski and Pelliccia (1982) added the pipe casting and diameter, and they discriminated between first break and subsequent breaks to the model in Shamir and Howard (1979) study. Clark et al. (1982) used two phase model, multivariate linear regression to predict the time to the first break and multivariate exponential regression to predict the number of subsequent breaks.

In linear regression, they used diameter, absolute pressure, length and pipe's material in addition to the industrial and residential development overlaying pipes, as independent variables. In exponential regression, they used pipe's age, surface area and length of pipe in low and moderately corrosivity soil and surface area of pipe in highly corrosive soil. More recently, Park et al. (2008) used a log-linear and power law process (Weibull process) to model failure rates and estimate the economically optimal replacement time of individual pipes in a water distribution system. Ekinci and Konak (2009) presented and compared applications of the log-linear method and the power law process for modeling pipe failure rates. Kleiner et al. (2010) used Non-homogeneous Poisson process (NHPP)-power lawto predict the breakage pattern of individual pipes considering three dynamic factors: the freezing index, the cumulative rain deficit and the snapshot rain deficit.

In linear regression, the relationship between two variables is modeled by minimizing the sum of the squared errors to fit a straight line to a set of data points. One or morevariable is considered to be an explanatory variableor predictor and the other is considered to be a dependent variable or predicted, time-linear regression modeling have been used in modeling the water pipes breakage, e.g. Kettler and Goulter (1985) used pipes age as independent variable to predict the pipes break, they found linear increment in pipe breaks with time. Jacobs and Karney (1994) used pipe lengths and age as independent predictor to find the probability of a day with no breaks in a linear relationship. They adopted clustering concept, so they applied their modeling on three homogenous age based groups of cast iron pipes. Despite the fact that regression models are standard models for statistical time series methods for forecasting and prediction (El Shafie et al., 2009a); nevertheless, all those models were suffering from low correlation between the prediction values and the observed values.

Regression based predictive models used widely in plans for development as well as in strategies for rehabilitation are used for example, to estimate the population, demands and the resources to allow plans to be made about possible developments. In literature, Arayesh (2011) used a multi-variable regression by means of backward method to evaluate the cumulative effect of people participation in protecting, revival, development and use of natural resources. Meanwhile, Malakmohammadi (2011) evaluated the educational applicability and behavioral research to develop realistic research outcomes based on regression techniques. Regression analysis in quality evaluation and performance monitoring has been conducted (Shariati, 2010; Mugisha, 2008). Other techniques, such as notably artificial neural networks (ANN), use regression tool in their prediction (Razavi et al., 2011).

Thisarticle aimed to improve the prediction of water pipe failure using the regression modeling to sustain the reliability of WDS, while the aforementioned efforts used a limited number of predictors, in this article, a multiple regression analysis for seven predictors has been used to predict the pipe breakage. Although, they are intrinsiccally non-linear parameter estimation problem, they are converted into a linear relationship using variable transformation technique, where the parameters entered into the formula as simple multipliers of terms that are added together.The results showed the ability of the multiple regressions to include this number of predictors, and then a technique to improve the prediction is introduced as well.

Multiple regression modeling

Multiple regressions are a general fitting procedure to model the relationship between a dependent (*f*), and an independent variables *(*X_1, X_2...X_n*)* are used for prediction. The general form of these models is:

$$f = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots \beta_n X_n, \qquad (1)$$

where β_0,β_1, β_2,...,β_n are the regression coefficients. The strength of the model lies in the regression coefficients; they represent the weight of contribution of each independent variable (X_n) in the prediction of the dependent variable (*f*), and they are estimated using loss functions. Least squares estimation is one of loss functions aims at minimizing the sum of squared variations of the actual values of the dependent variable from those predicted by the model. However, no perfect prediction can be obtained and usually there is substantial difference between the actual *(y)* and the predicted values(*f*). The limited prediction accuracy is attributed to the large number of uncertain and interrelated parameters that affect the prediction process and the need for a relatively large database to establish a reliable model (El Shafie et al., 2009b). This variation or prediction error is known as the “residual”. The fitness of the model is dependent on this residual and is measured using coefficient of determination (R^2). One form of R^2 is:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}, \qquad (2)$$

In which

$$SS_{tot} = \Sigma_i (y_i - y^-)^2, \qquad (3)$$

$$SS_{err} = \Sigma_i (y_i - f_i)^2, \qquad (4)$$

$$y^- = \frac{1}{n} \Sigma_i^n y_i \quad , \qquad (5)$$

where SS_{tot} is the total sum of squares and SS_{err} is the sum of squares of residuals. y_i and f_i are the observed or actual and the predicted values, respectively. *n* is the number of observations values. $y^-$is the mean of the observed values.

Least squares estimations assume the residual variance around the regression line is the same across all values of the independent variable. In some applications, this assumption is unrealistic. In this case, weighted least squares estimation is used instead. In some applications, e.g. failure prediction, where the goal is to compute the probability of the specific dependent variable to occur, in this case, maximize/minimize likelihood or log-likelihood function is the suitable technique to estimate the model fitness. The larger the likelihood, the larger is the probability of the dependent value to occur and the better the fit of the model.

In many cases, the regression model refuses to be fit to the data and the iterative procedure fails to converge. In such cases, there are many algorithms and criteria that can be used for minimizing the loss functionsto find the best fitting set of coefficients. Specifying some start values, initial step sizes and a criterion for convergence is one method. Other algorithm is quasi-Newton that approximates the second derivatives of the loss function to guide search for the minimum. Some methods use penalty functions and constraining parameters, for instance, 0 values for logistic regression. Simplex procedure is another algorithm that relies on the evaluation of the loss function at each of the iteration. Hooke-Jeeves pattern moves is a simple algorithm usually used when the quasi-Newton and Simplex methods fail to produce reasonable estimates. Rosenbrock pattern search method often succeeds when other methods fail (Draper et al., 1998). This method works by rotating the coordinates of the coefficients space and align one axis with a ridge and the other axes remain orthogonal to this axis. Other methods use the second order derivatives, such as Hessian matrix and standard errors. In this method, the coefficient is estimated according to the value of the second order derivative.

Finally and after estimating the regression coefficients, it becomes necessary to examine the fitness of the overall model. In this regard, there are many methods. Plotting actual values versus predicted values is one way to inspect the appropriateness of the model. Correlation coefficient, mean square error (MSE) and coefficient of determination (R^2) are used widely as performance evaluation of the predictive model fitness (El-Shafie, 2011a). Testing the normality of the residuals by plotting them on probability paper indicates fitness of the model. Finally, plotting the fitted model using the final coefficients estimates is a useful way to examine the models involving two or three independent variables.

In this article, we introduce a simple technique to improve the pipe breakage prediction that resulted from a multiple regression model regardless of the type of regression model and the loss function used to estimate the model fitness. This technique has been applied on a real water distribution system of the city of Benghazi, where a considerable prediction improvement has been shown.

METHODOLOGY

Many non-linear relationships, such as hyperbolas, exponential, power functions, logarithmic functions, polynomial, exponential models with a polynomial exponent and other special functions can be converted into linear relationship using variable transformation technique (Kutner et al., 2004). However, there are some other functions that cannot be linearized. They are intrinsically non-linear parameter estimation problem. The case study is a real example of linearizable function. The origin logarithmic function is:

$$f = a + b_1 \, ln\, x_1 + b_2 \, ln\, x_2 + b_3 \, ln x_3 + \ \dots + b_n \qquad (6)$$

However, the parameters entered into the formula as simple multipliers of terms that are added together in linear form as:

$$F = \beta_0 + \beta_1 \, X_1 + \beta_2 \, X_2 + \beta_3 \, X_3 + \cdots \beta_n X_n, \qquad (7)$$

where

$$F = f, X = lnx, \beta_0 = a \text{and} \beta_n = b_n.$$

The fitness of the predictive model (high R^2) started with the proper selecting to the quantity and quality of the predictor variables. However, all statistical softwares provide automated variable selection procedures in this regard. Among these procedures are backward elimination, forward selection and stepwise regression. Nevertheless, this automated selection should be experience based assessed. Stochastic models are always established based on correlation analysis (El-Shafie et al., 2008). Cross validation is one way to select the best model, studying the cross-correlation sequences, provides information about the mutual correlation between two consecutive time series (El Shafie et al., 2011b), where this method describes how well each observation or actual is predicted when all the observations except the one that is used to fit the model. In summary, the best fitting model is the one with the smallest value of the predicted residual sum of squares. Therefore, examining the normality assumption of residuals is important to evaluate the model fitness. Eliminating the outliers is one way to improve the normality of the residuals and hence, the model fitness. However, the normality assumption is not as important as the assumption that the model provides a good approximation for the true relationship between the predictors and the mean of *f* (Strobach, 1990).

The proposed method is applied after the prediction values have been obtained from the best fit model (Equation 7). Then, the improved predicted values can be obtained from:

$$F_{imp} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots \beta_n X_n + \mathrm{R} \qquad (8)$$

$$R = \mu f * + \gamma. \qquad (9)$$

in which, the residual modeled is added with thepredicted values, where*r*is the modeled or regressed residual, F_{imp} is the improved prediction for each case, μ and γ are respectively the slope and the intercept of the fit line and *f* is the predicted value.

Method application

The method proposed is applied on the real data of WDS. In this case, a pipe breakage is aimed to be predicted. WDS consists of 418 segments of pipes with a total length of 373.147 km and diameters varying from 150 to 2,500 mm. A total of 36.4% of the pipes are 300 mm in diameter and 27.4% are 400 mm in diameter; the other diameters are distributed as shown in Figure 2. In terms of materials, about 34% of the pipes are made of uncoated steel and ductile iron pipes account for about 56% of the total length; meanwhile, concrete pipes make up about 10%. About 25% of the system is more than 36 years old, about 20% is 24 years old and 30% is 5 years old; the rest of the system is about 27 years old on average. The system is supplied from two different sources with different water quality. The soil type in the area of study varied between clay and sandy clay. The corrosion is the main problem for these pipes and the other components of the system. The degradation of the network has made it unable to provide safe potable water for domestic use, adequate quantities of water at sufficient pressure for fire protection or water for industrial use, which has resulted in major environmental, socioeconomic and health problems in the city. Thus, there habilitation has become a national obligation.

The effective rehabilitation, in turn, depends mainly on predictive modeling of pipe breakage. Predicting the future number of breaks is a platform for any rehabilitation strategy. Moreover, it is important to enhance and improve the predictive model to the possible extent to assure maximum reliability of the system and saving cost as well.

Seven predictors: pipe length, diameter, age, pipe material, depth, type of soil and water quality are used to predict breakage rate of the pipes in the study system. The total number of breaks recorded from 2005 until 2009 were considered. Table 1 shows the number of breaks observed for each year. The average number of breaks per year is 133, with an estimated average break rate of about 0.326 break/year/km.

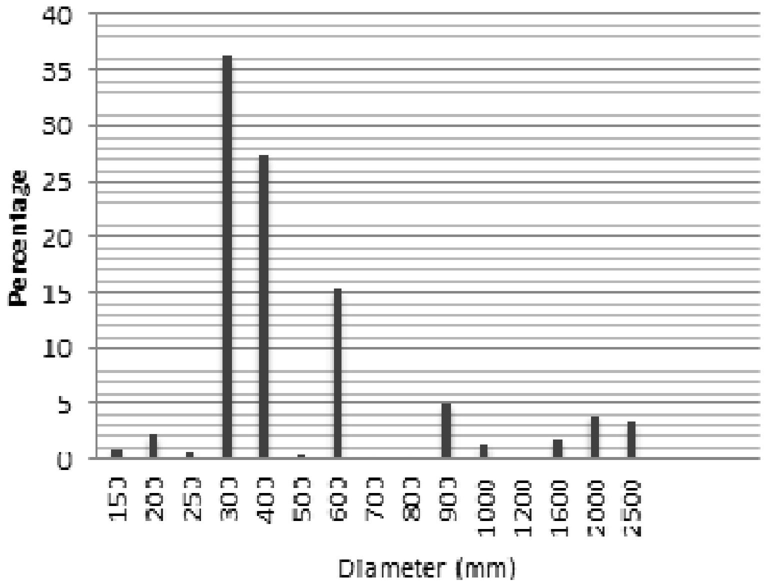

Figure 2. The distribution of diameters in the study system.

Table 1. Number of breaks observed per year.

Year	2005	2006	2007	2008	2009	2010	Total	Average	Breakage rate
Number of breaks	90	70	120	130	150	103	663	133	0.326 br/km/yearr

Table 2. Coding of soil, material and water quality in modeling.

Soil		Pipe material			Water quality		
Severe	Non-severe	Ductile	Uncoated	Concrete	Non-severe	Medium	Severe
1	2	1	2	3	1	2	3

Statistical analysis and correlations between predictors and the predicted variable

Once an appropriate historical data set has been selected and prepared, it was fully characterized and subjected to a comprehensive statistical analysis. Data characterization involved a qualitative assessment of seasonal trends of each potential model parameter. Non-quantitative predictors, such as type of soil, water quality and pipe material are coded as given in Table 2 to facilitate the modeling process. Water quality is estimated in each pipe by performing source tracing command in the used hydraulic solver (EPANET) and all pipes are tagged accordingly. The statistical analysis involved the determination of measures of central tendency, measures of variation, percentile analysis and identification of outliers, erroneous entries and non-entries for each data parameter. The parameter estimates, as given in Table 3, implied us to eliminate few outliers as given in Table 4, from which we could understand the nature of the relationships and correlations between the independent variables and the dependent variable. These correlations are summarized in Table 5. Table 6 shows how the correlations have been improved by eliminating the outliers.

RESULTS

Multiple regression analysis

Applying Equation 1, the best combination of the 7 predictor variables for explaining the variance of a dependent variable (pipe breakage) found using OpenStat software is:

$$Br_{pred} = -0.15547 + 0.34819 \times x1 - 0.09071 \times x2 - 0.39864 \times x3 + 0.30425 \times x4 + 0.26425 \times x5 + 0.14131 \times x6 + 0.22553 \times x7 \quad (10)$$

where Br_{pred} is the predicted number of breaks and $x_{1,}$ $x_{2,}$ $x_{3,}$ $x_{4,}$ $x_{5,}$ x_6 and x_7 are length (m), diameter (mm), depth (m), age (yr), material, type of soil and water quality, respectively.

For all the individual data points observed, their predicted values, the residual, the standard error of estimate of the predicted score and the 95% confidence

Table 3. Distribution parameter estimates.

S/N	Covariates	Variables	Mean	Variance	Standard deviation	Median	Minimum	Maximum	Skewness	kurtosis
1	Length, L (m)	418	892.7	1950007.5	1396.43	633.42	8.67	15020	6.5	53.2
2	Diameter, D (mm)	418	445.7	76300	276.23	400	150	2500	4.3	24.8
3	Depth (m)	418	0.575	0.035	0.187	0.6	0.4	1.00	0.852	-0.205
4	Age (year)	418	26.3	145.52	12.1	32	5	40	-0.904	-0.708
5	Material	418	1.581	0.278	0.527	2	1	3	0.017	-1.247
6	Soil type	418	1.289	0.206	0.454	1	1	2	0.932	-1.137
7	Water quality	418	2.299	0.412	0.642	2	1	3	-0.366	-0.701
8	Number of breaks, Br	418	1.587	14.479	3.805	3.8	0	66	12.76	207

Table 4. Eliminating outliers from some variables data.

S/N	Covariates	Variables	Mean	Variance	Standard deviation	Median	Minimum	Maximum	Skewness	kurtosis
1	Length (m)	403	90.179	285660.76	534.472	600	8.67	2711	1.098	1.365
2	Diameter (mm)	410	416.585	28917.407	170.051	400	150	1000	1.817	3.118
3	Number of breaks	415	1.337	2.506	1.583	0.8269	0	10.677	2.612	9.594

Table 5. Correlation matrix for row data.

	Length (m)	Diameter (mm)	Depth (m)	Age (year)	Material	Type of soil	Water quality	Breakage rate
Breakage rate	0.083	-0.491	-0.644	0.437	-0.224	0.198	0.412	1.000

Table 6. Correlation matrix without outliers values of length, diameter and breakage rate.

	Length (m)	Diameter(mm)	Depth (m)	Age (year)	Material	Type of soil	Water quality	Breakage rate
Breakage rate	0.233	-0.605	-0.649	0.419	-0.212	0.196	0.423	1.000

interval of the predicted score are all calculated and the fitness of the selected model can be examined from the coefficient of determination (R^2 = 0.737).

These results can be represented graphically by drawing the actual values versus the predicted values as given in Figure 3, from which we can see how quite well the fit between the observed or actual and predicted values. The parameters of the actual and predicted values are given in Table 7 and the parameters of the actual versus predicted plot are given in Table 8.

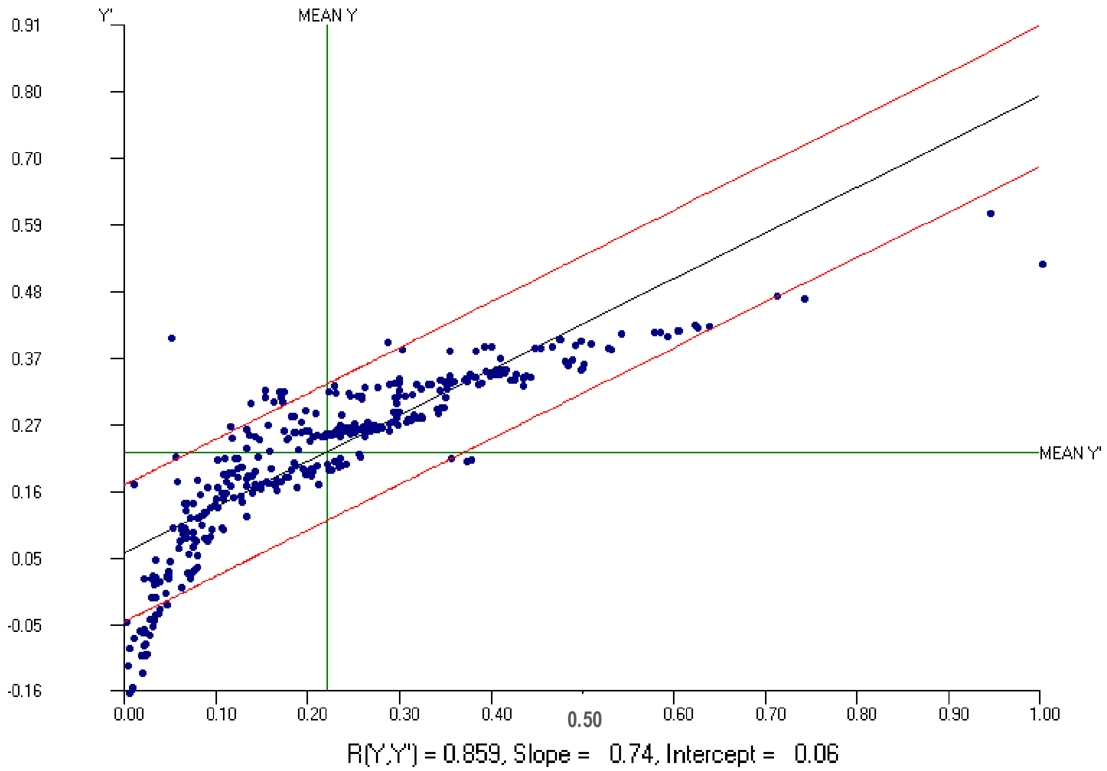

Figure 3. Modeling the actual (*y*) and predicted (*y*´).

Table 7. Data parameters of the actual and the predicted values.

Variable	Mean	Variance	Standard deviation
Actual	0.22	0.02	0.15
Predicted	0.22	0.02	0.13

Table 8. The parameters of the actual versus predicted plot.

Correlation	Slope	Intercept	Standard error of estimate	Number of good cases
0.8587	0.74	0.06	0.07	418

Improvement of the prediction model

It was known that for multiple regression models, plotting the residuals against the predicted values or against each independent variable also helps to check for potential problems. If the residuals appear to fluctuate randomly about 0 with no obvious trend or change in variation as the values of a particular X_n increase, then no violation of assumptions is indicated (Strobach,1990). Now, we introduce that if the residual and the predicted are modeled and plotted, the prediction model can be improved. In our case study, the correlation between the residual and the predicted values can be seen in Figure 4.

Thus, using Equation 8, Equation 10 becomes:

$$Br_{Imp} = \begin{bmatrix} -0.15547 + 0.34819 * x1 - 0.09071 * x2 - 0.39864 * x3 + 0.30425 * \\ x4 + 0.26425 * x5 + 0.14131 * x6 + 0.22553 * x7 \end{bmatrix} + R \qquad (11)$$

where

$$R = -0.26 \times Br_{pred} + 0.06 \qquad (12)$$

Br_{Imp} = Improved predicted breaks.

The coefficient of determination (R^2) of the model

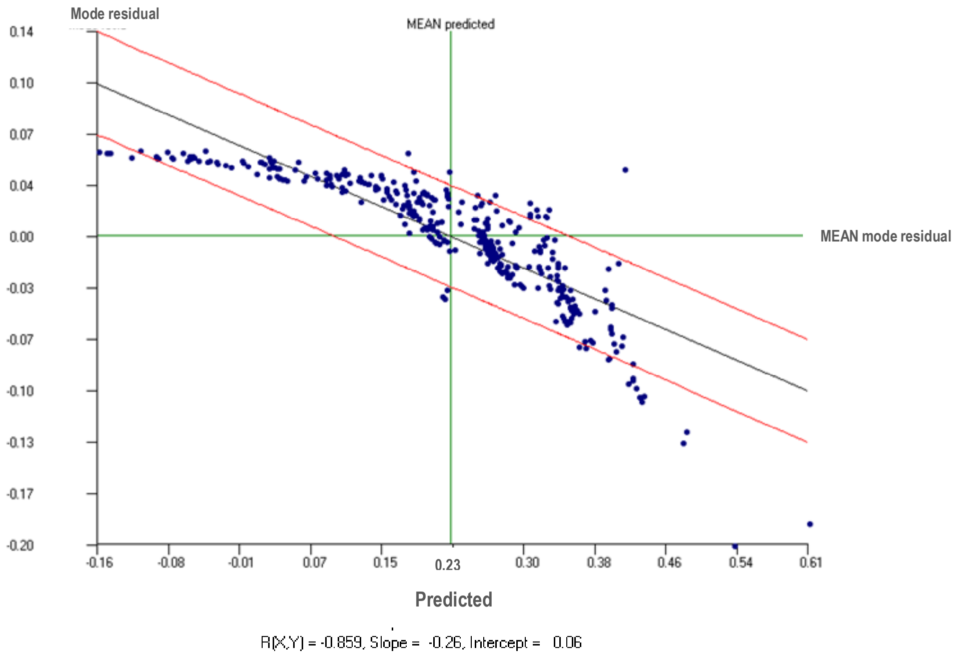

Figure 4. Modeling the predicted values and the modeled residual.

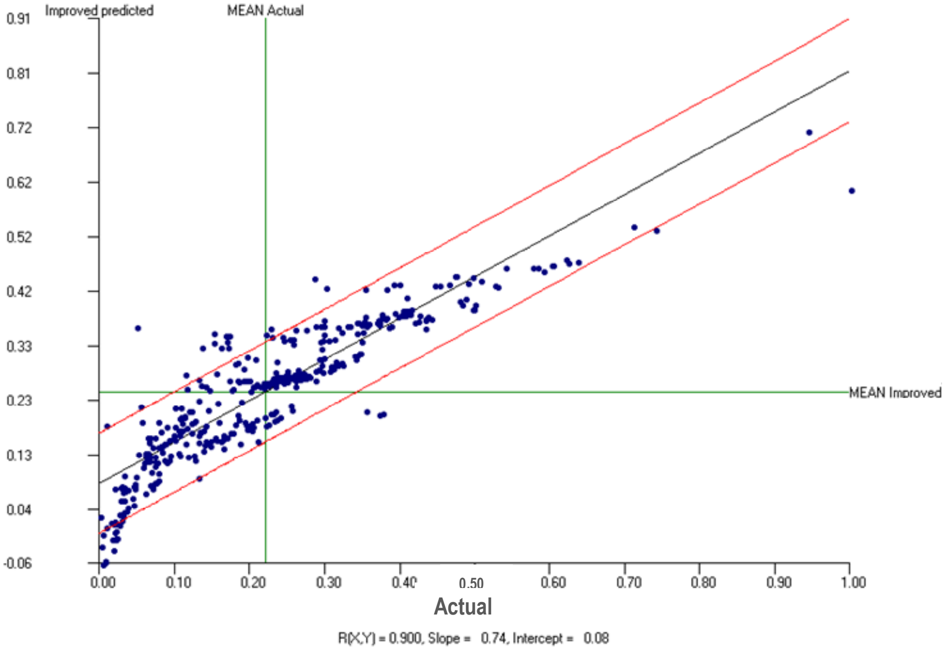

Figure 5. Modeling the improved predicted values and actual values.

becomes 0.9, increment of about 4.6% over the origin model is achieved. Figure 5 represents the relationship and the correlation for the improved prediction model.

Figure 6 and Table 9 show a comparison between the predicted values obtained from the regression model, depicted as predicted 1, the predicted values obtained from the improved prediction model, depicted as predicted 2 and the actual values. Table 10 presents correlation matrix of the actual values, predicted, residual, modeled residual, new residual and the

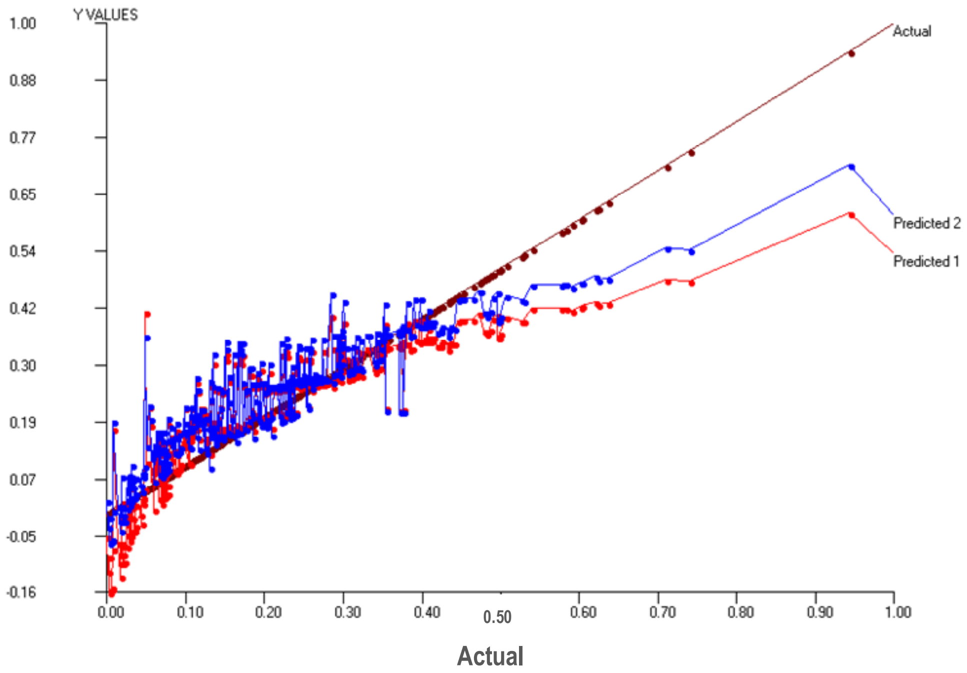

Figure 6. Comparison between the predicted and the actual values.

Table 9. Comparison of the parameters and correlation of the predictions and the actual.

Correlation (with the actual)	**Predicted 1**	**Predicted 2**	**Actual**
	0.859	**0.900**	**1.00**
Means	0.223	0.246	0.223
Standard deviations	0.130	0.124	0.152

Table 10. Correlation matrix of the actual, predicted, new residual, modeled residual and improved predicted.

	Actual	**predicted**	**Residual**	**New residual**	**Modeled residual**	**Improved predicted**
Actual	1.000	0.859	-0.512	-0.859	-1.000	0.900
predicted	0.859	1.000	-0.000	-1.000	-0.859	0.982
Residual	-0.512	-0.000	1.000	0.000	0.512	-0.111
new residual	-0.859	-1.000	0.000	1.000	0.859	-0.982
Modeled residual	-1.000	-0.859	0.512	0.859	1.000	-0.900
Improved predicted	0.900	0.982	-0.111	-0.982	-0.900	1.000

obtained improved predicted values.

DISCUSSION

In combination, the data characterization and statistical analysis assisted to identify the boundaries of the study domain as well as potential deficiencies in the data set. In order to identify the most relevant variables, variety of graphical and statistical analysis are implemented and from which many causes and reasons can be highlighted:

1. The breaks increase as length of pipes increases. Only in this relationship log transforming of both parameters were implemented as shown in Figure 7, otherwise the nature relationship would not be obvious.
2. The majority of numbers of breaks are found in 300 mm pipes (415 breaks) and 400 mm diameter pipes (149 breaks). This is because 63% of the total network length is made of these two sizes. Number of breaks decreases

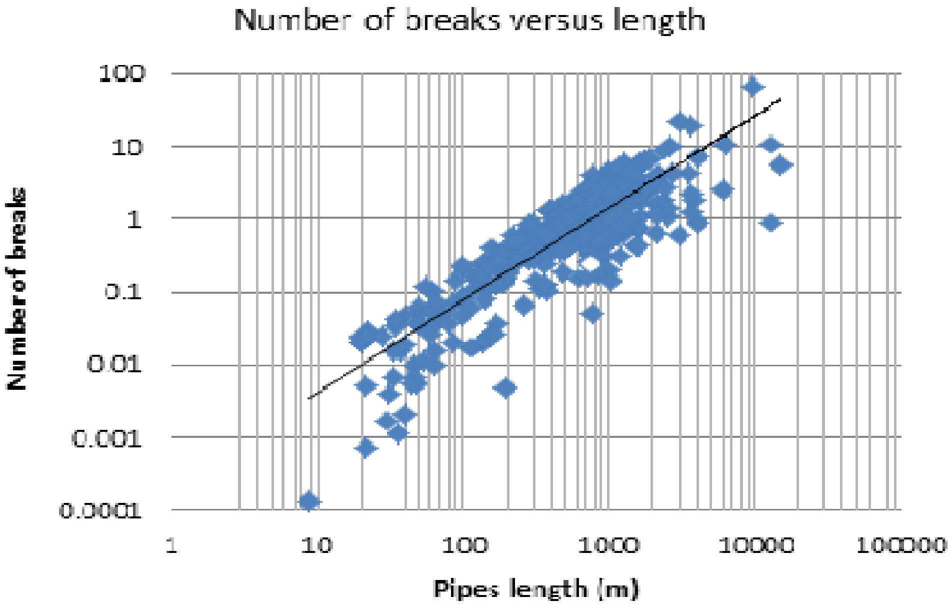

Figure 7. Relationship between pipes length and number of breaks.

as the diameter increases. The correlation between these two parameters is -0.491.

3. The breakage rate increases as the depth of the trench where the pipes are laid decreases, 67% of the failure occurs in pipes at depth of 0.4 m. The correlation between failure and depth found is equal to -0.644, which is the highest among all the other parameters. This is a result of considerable quantity of the smaller sizes pipes found in this depth (43% of pipes) which might be close to the proximity to traffic loads.
4. 67% of the failures occur in age between 32 and 40 years. Although, there are breaks in all ages, starting from 5 years old pipes. The correlation between these two parameters is 0.437.
5. Comparing the different pipes material, 58% of failure occurs in ductile iron; meanwhile 41% of failure occurs in uncoated pipes. This is because ductile iron pipes form about 56% of the total length of the pipes. The correlation between breakage rate and type of material is -0.224.
6. 69% of the failure occurs in zones of non aggressive soil, where about 71% of the pipes are laid. Therefore, the correlation between these two parameters is 0.198.
7. 89% of the pipes convey moderate to aggressive water; therefore, 97% of the failure occurs in those pipes. The correlation between breakage rate and water quality found is 0.412.

Conclusion

This article addressed the prediction of water pipe failure using regression analysis modeling and howit can be improved. A simple method to improve the pipe breakage prediction resulted from multiple non-linear regression models and was introduced regardless of the type of regression model and the loss function used to estimate the model fitness. This method depended on modeling the residual with the predicted data, then adding this new residual to the regression model. This technique was applied on actual water distribution system to predict pipe breakage. Seven predictors were modeled and a considerable performance improvement of the prediction was shown.

REFERENCES

Arayesh B (2011). Regression analysis of effective factor on people participation in protecting and revitalizing of pastures and forests in Ilam province from the view of users BagherArayesh.Afr. J. Agric. Res., 6(2): 416-422.

Bubtiena AM, Elshafie A, Jafaar O (2011). Review on statistical based methods of measuring the reliability of water pipes. J. Adv. Mater. Res., 230-232(5): 1327-1331.

Christodoulou S, Deligianni A (2010). A neurofuzzy decision framework for the management of water distribution networks. J. Water Resource Manage., 24:139-156.

Clark RM, Stafford C, Goodrich J, O'Day D (1982). Water distribution systems: a spatial and cost evaluation. J. Water Resources Plann. Manage., 108(12): 243-256.

Draper NR, Smith H (1998). Applied Regression Analysis", 3rd ed., Wiley Holzer, C. E., III (1977).

Ekinci O, Konak H (2009). An optimization strategy for water distribution networks. J. Water Resources Manage., 23:169-185.

El-Shafie A, Abdelazim T, NoureldinA (2009b).Neural network modeling of time-dependent creep deformations in masonry structures. J. Neural Comput . Applic. DOI 10.1007/s00521-009-0318-3.

El-Shafie A, Abdin A, Noureldin A, Taha M (2009a). Enhancing Inflow Forecasting Model at Aswan High Dam Utilizing Radial Basis Neural Network and Upstream Monitoring Stations Measurements. J. Water Resour Manage., 23(12):2289-2315.

El-shafie A, Mukhlisin M, Najah A, Taha M (2011a). Performance of artificial neural network and regression techniques for rainfall-runoff prediction.Inter. J. the Phys. Sci., 6(8):1997-2003.

El-Shafie A, Noureldin A, Taha, R, Basri H (2008). Neural Network Model For Nile River Inflow Forecasting Based On Correlation Analysis Of Historical Inflow Data. J. Appl. Sci., 8(24): 4487-4499.

El-Shafie A, Noureldin A (2011b). Generalized versus non-generalized neural network model for multi-lead inflow forecasting at Aswan High Dam.Hydrol. Earth Syst. Sci., 15: 841-858

Jacobs P, Karney B (1994). GIS development with application to cast iron water main brekagerates.In Proc. 2nd International Conference on Water Pipeline Systems, Edinburgh. Mechanical Engineering Publication Ltd, London.

Kai C, Hua Z (2011). An experimental study and model validation of pressure in liquid needle-free injection, Int. J. Phys. Sci., 6(7):1552-1562.

Kettler AJ, Goulter I (1985). An analysis of pipe breakage in urban water distribution networks. Can. J. Civ. Eng., 12(2): 286-293:

Kleiner Y, Rajani B (2010). I-WARP: Individual Water Main Renewal Planner, NRCC-53221, Drinking Water Eng. Sci. Discussion, 3(1): 25-41.

Kleiner Y, Rajani B (2001). Comprehensive review of structural deterioration of Water Mains: Statistical Models. J. Urban Water, 3(3):131-150.

Kleiner Y, Rajani B, Sadiq R (2009). Drinking water infrastructure assessment: The National Research Council of Canada perspectiveNRCC-51298June 25, 2008World Environmental and Water Resources Congress 2009.Environmental and Water Resources Institute (EWRI) of the American Society of Civil Engineers, Kansas City, Missouri.

Kutner MH, Nachtsheim C, Neter J (2004). Applied Linear Regression Models, 4th ed. McGraw-Hill/Irwin.

Lawal T, Latiff1 AA, Tjahjanto D, Akib S (2011). The effectiveness of groundwater recharges well to mitigate flood, Int. J. the Phys. Sci. 6(1): 8-14

Malakmohammadi I (2011). Statistical Mix: Sequential statistical analysis approach to legitimate statistical techniques in agricultural extension, education and rural development. Afr. J. Agric. Res., 6(2): 423-431.

Mugisha S (2011). Infrastructure optimization and performance monitoring: empirical findings from the water sector in Uganda. Afr. J. Bus. Manage.. 2(1): 013-025.

Mukhlisin M, Ilyias I, Wan ZY, El-Shafie A, Taha MR (2011). Soil slope deformation behavior in relation to soil water interaction based on centrifuge physical modelling, Int. J. Phys. Sci., 6(13): 3126-3133.

Ömer K (2011). Distribution of turbulence statistics in open-channel flow, Int. J. Phys. Sci., 6(14): 3426-3436

Park S. (2008). Identifying the hazard characteristics of pipes in water distribution systems by using the proportional hazards model theory. KSCE J. Civil Eng., 8(6):663-668.

Razavi S, Jumaat V, El-Shafie A, Mohammadi P (2011). General regression neural network (GRNN) for the first crack analysis prediction of strengthened RC one-way slab by CFRP., Int. J.Phys. Sci., 6(10): 2439-2446.

Shamir U, Howard C (1979). An analytic approach to scheduling pipe replacement. J. AWWA, 71(5): 248-258.

Shariati M, Ramli-Sulong N, Arabnejad M, Shafigh P, Sinaei H (2011). Assessing the strength of reinforced concretestructures through Ultrasonic Pulse Velocity and Schmidt Rebound Hammer tests. Sci. Res. Essays, 6(1): 213-220.

Strobach P (1990). Linear prediction theory, Springer-Verlag, Berlin.

Tabesh, M, Soltani J, Farmani R, Savic D (2009). Assessing Pipe Failure Rate and Mechanical Reliability of Water Distribution Networks Using Data Driven Modeling. J. Hydroinformatics, 11(1): 1-17.

Walski TM, Pelliccia A (1982). Economic analysis of water main breaks. J. AWWA, 74 (3): 140-147.

William GM (2008). Statistics and Measurement Using Open Stat.