https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021

Open Access Drinking Water Engineering and Science Discussions

Implementing and evaluating various machine learning models for pipe burst prediction

Ahmad Ravanbakhsh[1], Mehdi Momeni[2*], Amir Robati[2]

[1] PhD in Water Resources Engineering and Management, Department of Civil Engineering, Azad University of Kerman, Iran
[2]Assistant Professor, Department of Civil Engineering, Azad University of Kerman, Iran

Correspondence to: Mehdi Momeni (zimaraz.pars1387@gmail.com)

Abstract. By accurate predicting of pipe bursts, it is possible to schedule pipe maintenance, rehabilitation and improve level of services in water distribution networks (WDNs). In this study we aimed to implement five artificial intelligence and machine learning regression models such as multivariate adaptive regression splines (MARS), M5' regression tree (M5'), Least square support vector regression (LS-SVR), fuzzy regression based on c-means clustering (FCMR) and regressive convolution neural network with support vector regression (RCNN-SVR) for predicting pipe burst rate and evaluating the performance of these models. The most effective parameters for regression models are pipes age, diameter, depth of installation, length, average and maximum hydraulic pressure. In the present study, collected data include 158 cases for polyethylene (PE) and 124 cases for asbestos cement (AC) pipes during 2012-2019. The results indicate that RCNN-SVR model has a great performance of pipe burst rate (PBR) prediction.

1. Introduction

Water distribution networks (WDNs) are critical infrastructures. The objective of WDNs is to provide water with desirable quantity, quality and pressure for the consumers. However, in case of pipe failure which is the progressive effect of physical, operational and weather-related factors, might fail the WDN to achieve these goals (Kakoudakis, 2019). A pipe bursts when the residual strength of a deteriorated pipe can no longer resist the force inflicted on it (Berardi et al., 2008). Pipe burst prediction helps to prioritize the maintenance, repair, rehabilitation and replacement of pipes after assessing and forecasting pipe propensity to burst. In addition, pipe burst prediction can be used for budget allocation and cost analysis of dynamic or static designing of water distribution networks. In the literature, there are typically two categories consisting of physical and statistical methods for modeling of pipes burst (Grigg, 2007) (Rajani and Kleiner, 2001). Physical models are developed to understand the physical process of pipe deterioration. In this models, the items that may affect the pipes burst, include environmental conditions, quality of manufacturing, installation procedure, internal and external loads, surrounding soil, ground traffic and etc. (Wilson et al., 2015). The physical mechanisms of pipe burst are complex and not well-understood, and there is limited data available on the breakage failure modes due to the inspection difficulty and lack of historical data

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021


Open Access Drinking Water Engineering and Science Discussions

(Rajani and Kleiner, 2001). Statistical methods, model the pipes burst based on historical data. The assumption of these models is that pipes with similar specification and working environment will experience similar deterioration pattern (Kleiner and Rajani, 2010). Since physical models developed for pipe failure prediction are complicated and expensive, they can be used on a limited number of pipes. But statistical models based on historical data are less expensive, and have vast applications.

The goal of any data analysis is to extract accurate estimation from the raw information. One of the most substantial and typical issues is whether there is statistical relationship between a response variable (Y) and explanatory variables (Xi). One way to answer this issue is to employ regression analysis in order to model its relationship (Alexopoulos, 2010). Different studies have been proposed various pipe failure prediction methods such as physical (Randall-Smith et al., 1992), multivariate adaptive regression spline (Kutylowska, 2019), artificial neural networks (Achim et al., 2007) (Kutylowska, 2017), support vector machines (Kutylowska, 2018), fuzzy logic (Rajani and Tesfamariam, 2007), neuro-fuzzy systems (Christodoulou et al., 2004) (Tabesh et al., 2009) and evolutionary polynomial regression (Berardi et al., 2008).

In this research, pipe length (L), diameter (Dim), average hydraulic pressure (Pa), maximum hydraulic pressure (Pm), age (A) and installation depth (ID) are used as input of regression models and pipe burst rate (PBR) obtained as the output. In addition, correlation between these factors and PBR have been investigated. After implementing the various artificial intelligence and machine learning models such as multivariate adaptive regression splines (MARS), M5' regression tree (M5'), Least square support vector regression (LS-SVR), fuzzy regression based on c-means clustering (FCMR) and regressive convolution neural network with support vector regression model (RCNN-SVR) in a real water distribution network, the corresponding predicted PBR values have been evaluated to find the best model-based prediction method.

2. Methodology

In order to implement the regression models, six different input variables consist of pipe diameter, length, age, depth of installation, average and maximum hydraulic pressure have been used. The output of all mentioned prediction models is PBR. PBR values are calculated using the following equation:

$$\mathrm{PBR} = \frac{\text{number of annual pipe bursts}}{\text{pipe length (km)}} \quad (1)$$

The collected data have been split into training and test sets by random sampling. 85% of data have been selected for training and the rest of them have been used to test the models. By using several evaluation indices, the testing dataset evaluate the performance of the models on future unseen data. Further analysis will be performed to investigate the Pearson correlation of the PBR and the variables.

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021


Open Access Drinking Water Engineering and Science Discussions

2.1. Description of regression models

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) variable and independent variable(s) (predictor). In this section five multivariate regression models for pipe failure prediction will be implemented and discussed.

2.1.1. Multivariate adaptive regression spline (MARS)

Multivariate Adaptive Regression Splines (MARS) is a multi-variable non-parametric regression analysis for fitting the relationship between a set of input variables and dependent variables introduced by Friedman (1991). Recently the MARS as a powerful regression technique has been used for modeling of different types of data (Rezaie-balf, 2019) (Heddam and Kisi, 2018) (Safari, 2019) (Emamgolizadeh et al., 2015) (Forghani and Peralta, 2017). In this method the training data sets are partitioned into separate regions, and each one gets its own regression line called basis functions. The break values between the intervals are called knots. Through modelling procedure, forward and backward stages are accomplished. Each stage has certain responsibilities, where important and appropriate variables are designated in the forward stage, while, in backward stage less important variables are eliminated to enhance the model performance (Friedman, 1991) (Sharda et al., 2008). The general MARS model equation is defined as:

$$Y=\beta_0+\sum_{m=1}^{M}\beta_m.h_m(X) \quad (2)$$

Where

$$h_m(X)=\prod_{k=1}^{K_m}\left[S_{k,m}(X_{V(k,m)}-t_{k,m})\right]_+ \quad (3)$$

Where β_0 and β_m are the parameter values and their functions are similar to the regression coefficient of the linear regression model; the $h_m(X)$ is the spline basis function that represents the data in each sub-region; M is the number of sub-regions or the number of basis functions (BFs) in the model, which adjusted at the first step; "+" means the argument that is a truncated power function, K_m is the knot quantity; $S_{k,m}$is +1 or -1 which shows the BF's direction; V(k,m) is the variable label and $t_{k,m}$ is the cut-off point.

The BFs represent the relationship between the knots using the reflected pairs of hockey stick function (f) as follows:

$$f(x_i)=\max\ (0,x-c) \quad (4)$$

Or

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021


Open Access Drinking Water Engineering and Science Discussions

$$f(x_i) = \max\ (0, c-x) \quad (5)$$

Here, c is a threshold value that denotes the knot, where the behavior of the function changes. This model searches over the space of all inputs and predictor values (knots) as well as the interactions between variables. During this search, an increasingly larger number of basis functions are added to the model to minimize a lack-of-fit criterion. As a result of these operations, MARS automatically determines the most important independent variables as well as the most significant interactions among them. It is noted that the search for the best predictor and knot location is performed in an iterative process. The predictors as well as the knot location, having the most contribution to the model, are selected first. Also, at the end of each iteration, the introduction of an interaction is checked for possible model improvements.

The obtained BFs for Joopar WDN for AC and PE pipes are:

AC pipes:

$$\begin{aligned}
&BF1 = \max\ (0, A - Pm)\\
&BF2 = \max\ (0,162.93 - L) \times \max\ (0, Pm - 70.2)\\
&BF3 = \max\ (0,162.93 - L) \times \max\ (0,70.2 - Pm)\\
&BF4 = \max\ (0, Pa - 63.4)\\
&BF5 = \max\ (0,63.4 - Pa)\\
&BF6 = \max\ (0,49.9 - Pa)\\
&BF7 = \max\ (0,46 - A) \times \max\ (0,99.39 - L)
\end{aligned} \quad (6)$$

$$PBR = 1.323 + 0.347 \times BF1 + 0.001 \times BF2 + 0.0002 \times BF3 - 0.057 \times BF4 - 0.0371 \times BF5 + 0.035 \times BF6 + 0.006 \times BF7 \quad (7)$$

PE pipes:

$$\begin{aligned}
&BF1 = \max\ (0, L - 112.05)\\
&BF2 = \max\ (0,112.05 - L)\\
&BF3 = \max\ (0,0.9 - ID) \times \max\ (0,71.3 - Pa) \times \max\ (0, L - 116.74)\\
&BF4 = \max\ (0,0.9 - ID) \times \max\ (0,71.3 - Pa) \times \max\ (0,116.74 - L)\\
&BF5 = BF4 \times \max\ (0, Dim - 58)\\
&BF6 = BF4 \times \max\ (0,58 - Dim)
\end{aligned} \quad (8)$$

$$PBR = 1.338 - 0.004 \times BF1 + 0.0135 \times BF2 + 0.006 \times BF3 + 0.044 \times BF4 - 0.0008 \times BF5 - 0.001 \times BF6 \quad (9)$$

2.1.2. M5' model tree (M5')

M5 tree is a decision tree learner for regression problems introduced by Quinlan (1992). The M5 tree has three main types of nodes; decision nodes, leaf nodes and a root node. A decision node has two or more branches, each representing values for the attributes. Leaf node represents a decision on the numerical target, and the topmost decision node in a tree is called root node. The model is established according to a binary decision tree in which there are linear regression functions in the leaf nodes, which sets a relationship between independent and dependent variables (Rahimikhoob et al., 2013). Wang and Witten

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021


Open Access Drinking Water Engineering and Science Discussions

(1997) expanded the algorithm and introduced a new method called the M5' algorithm, which is structured by trees that graphically represent if-then rules. This algorithm is built top-down from a root node and involves partitioning the input space into many sub-spaces and fitting a linear regression model in each of the subspaces with similar values by calculating standard deviation.

The M5' method builds a tree in three phases; growing phase, pruning phase and smoothing phase. In growing phase, the dataset is split on different attributes. Then the standard deviation for each branch is calculated. The resulting standard deviation is subtracted from the standard deviation before the split. The result is the standard deviation reduction (SDR) which is based on the decrease in standard deviation after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest standard deviation reduction. SDR is represented by Quinlan (1992):

$$SDR=sd(K)-\sum\frac{|K_i|}{|K|}sd(K_i) \quad (10)$$

Where, K represents a set of examples that reaches the node; K_i and sd represent the subset of examples that has the i'th outcome of the potential set and the standard deviation respectively (Wang et al., 2010).

At the end of the first phase, there is a large tree that over fits the data, so a pruning phase must be employed. In this phase, the tree is pruned back from each leaf until an estimate of the expected error that will be experienced at each node cannot be reduced any further (Wang et al., 2010). Finally, the smoothing phase is performed to compensate for the sharp discontinuities that will inevitably occur between adjacent linear models at the leaves of the pruned trees, particularly for some models constructed from a smaller number of training examples (Ditthakit et al., 2012). In this phase, the adjacent linear equations are updated in such a way that the predicted outputs for the neighboring input vectors corresponding to the different equations are becoming close in value. This process substantially increases the accuracy of prediction (Witten and Frank, 2005).

M5' model tree implemented in our case study results as:

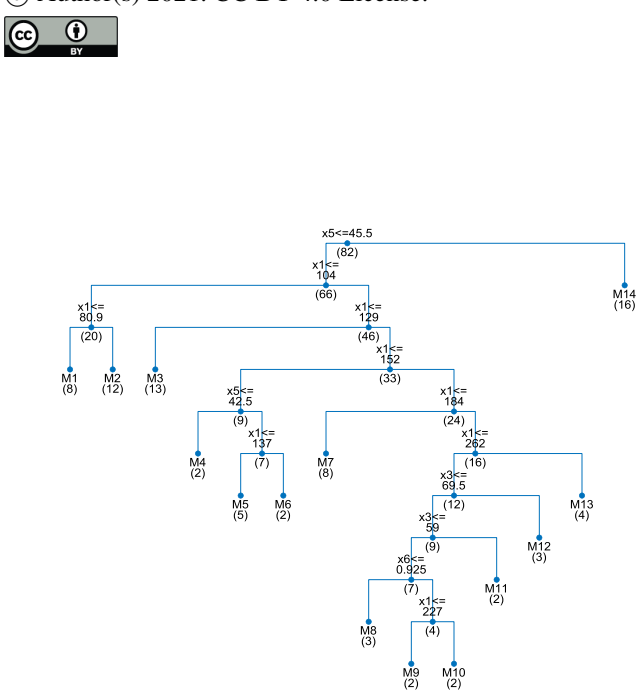
AC pipes model:

M1 = 3.66 -0.0267×L, M2 = 2.7 -0.0145×L, M3 = 2.12 -0.00895×L, M4 = 0.875, M5 = 1.13, M6 = 1.77, M7 = 0.759, M8 = 1.18, M9 = 0.615, M10 = 0.999, M11 = 1.58, M12 = 0.588, M13 = 0.515, M14 = -9.02 -0.00418×L +0.248×A

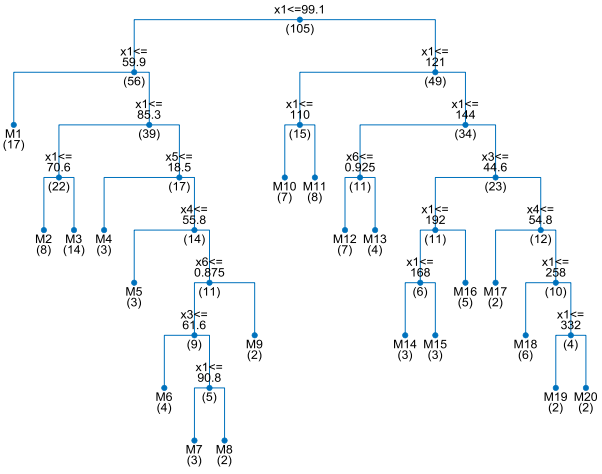
PE pipes model:

M1 = 3.69 -0.0228×L, M2 = 3.87 -0.03×L, M3 = 3.17 -0.0201×L, M4 = 1.34, M5 = 2.66, M6 = 1.65, M7 = 1.92, M8 = 2.66, M9 = 1.36, M10 = 2.4 -0.0115×L, M11 = 2.16 -0.00934×L, M12 = 1.9, M13 = 1.17, M14 = 0.821, M15 = 1.15, M16 = 0.984 -0.00185×L, M17 = 1.97, M18 = 1.18, M19 = 0.432, M20 = 1.02

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021



(a)



(b)

Figure 1. M5' tree graphical model for (a) AC pipes and (b) PE pipes

2.1.3. Fuzzy c-regression (FCR):

Fuzzy c–regression (FCR) model introduced by Hathaway and Bezdek (1993). This method is an extension of fuzzy c-means approach which is one of the most popular clustering method. It performs classification based on the iterative minimization of the following objective function and constraints (Bezdek et al. 1984; Bezdek 1981; Dave 1992):

$$J_q(\mu,V,X)=\sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{i,j})^q D_{i,j}^2 \tag{11}$$

Subject to:

$$0\le\mu_{i,j}\le 1$$

$$\sum_{i=1}^{c}\mu_{i,j}=1 \tag{12}$$

$$0<\sum_{j=1}^{n}\mu_{i,j}<n$$

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021


Open Access Drinking Water Engineering and Science Discussions

Where $i\in\{1,\ldots,c\}$, $j\in\{1,\ldots,n\}$, n is number of data points, c is number of clusters, μ is the fuzzy membership matrix, q is the fuzzifier where $q\geq 0$, V is cluster center vector. X is a data vector and $D_{i,j}$ is the distance between observation x_j and cluster center v_i. By using a Lagrangian multiplier, V and μ can be obtained by optimizing the objective function in (1):

$$\mu_{i,j}=\frac{1}{\sum_{k=1}^{c}\left(\frac{D_{i,j}}{D_{k,j}}\right)^{\frac{2}{q-1}}} \tag{13}$$

$$v_i=\frac{\sum_{j=1}^{n}\left[(\mu_{i,j})^q x_j\right]}{\sum_{j=1}^{n}(\mu_{i,j})^q} \tag{14}$$

The membership values are initialized randomly and both these and the cluster centers are iteratively updated until the maximum change in $\mu_{i,j}$ becomes less than or equal to a specified threshold ε. q is normally set to 2 as this is the best value for the fuzzifier while the membership $\mu_{i,j}$ is randomly initialized. The cluster center v_i and membership values $\mu_{i,j}$ are then iteratively updated using (32) and (33) respectively until either the maximum number of iterations or threshold ε is reached (Ameer et all, 2008). Finally the weighted least square is used for regression model, in which weights are membership values of train data and for each cluster, regression coefficients (β) is calculated:

$$\beta_i=(X^T W_i X)^{-1}.X^T W_i Y \tag{15}$$

Where Y is observed PBR, X is dependent variables and $W_i=\mathrm{diag}\{\mu_i\}$ for all train observations. Then by using calculated v_i the membership values of test data are used for prediction:

$$y_{pre}(j)=\sum_{i=1}^{c}\mu_{i,j}.x_{test}(j).\beta_i \tag{16}$$

Where $x_{test}(j)$ is jth test observation such that:

$$x_{test}(j)=[1\ X_{j1}\ldots X_{jp}] \tag{17}$$

2.1.4. least-squares support vector regression (LSSVR)

Support vector machines (Vapnik, 1995) (Vapnik, 1998a) (Vapnik, 1998b) have been introduced for solving pattern recognition problems. The SVM system used to estimate regression is called Support Vector Regression (SVR) which has been used in various different prediction problems. This method maps data x into a high dimensional feature space using non-linear mapping and performs linear regression in this space.

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021


Open Access Drinking Water Engineering and Science Discussions

$$f(x)=W^T\phi(x)+b \tag{18}$$

In which $b\in R$ and W will be found by minimizing the following objective function (ζ) with constraints:

$$\min\ \zeta(W,\xi)=\frac{1}{2}W^TW+\gamma\frac{1}{2}\xi^T\xi \tag{19}$$

$$\text{s.t.}\quad y_l=Z^TW+b1_l+\xi \tag{20}$$

Where y is observed PBR, l is the number of observations, $Z=(\varphi(x_1),\varphi(x_2),\ldots,\varphi(x_l))$ in which φ is a mapping to some higher (maybe infinite) dimensional Hilbert space (H), $\xi=(\xi_1,\xi_2,\ldots,\xi_l)^T$ is a vector consisting of slack variables, and γ is a positive real regularized parameter.

The Lagrangian function for the optimization problem is:

$$L(W,b,\xi,\alpha)=\zeta(W,\xi)-\alpha^T(Z^TW+b1_l+\xi-y) \tag{21}$$

Where α is a vector consisting of Lagrange multipliers. So we have the following set of linear equations:

$$\begin{cases}\frac{\partial L}{\partial W}=0\ \Rightarrow\ W=Z.\alpha\\ \frac{\partial L}{\partial b}=0\Rightarrow\ \alpha^T.1_l=0\\ \frac{\partial L}{\partial \xi}=0\ \Rightarrow\ \ \alpha=\gamma\xi\\ \frac{\partial L}{\partial \alpha}=0\ \Rightarrow\ Z^TW+b1_l+\xi-y=0\end{cases} \tag{22}$$

By eliminating w and ξ, one can obtain the following linear system:

$$\begin{bmatrix}0 & 1_l^T\\ 1_l & H\end{bmatrix}\begin{bmatrix}b\\ \alpha\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix} \tag{23}$$

Where $H=K+\gamma^{-1}I_l$ and $K=Z^TZ$ which is defined as $K_{i,j}=\varphi(x_i)^T\varphi(x_j)=\kappa(x_i,x_j)$ and $\kappa(0,0)$ is a kernel function. The solution of this problem can be found by the following three steps:

(1) Solve η , ν from $H.\eta=1_l$ and $H.\nu=y$;

(2) Compute $s=1_l^T.\eta$;

(3) Find solution: $b=\eta^T.y/s$, $\alpha=\nu-b.\eta$

So we have:

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021

Open Access Drinking Water Engineering and Science Discussions

$$f(x)=W^{T}\phi(x)+b=\alpha^{T}Z^{T}\phi(x)+b=\sum_{i=1}^{l}\alpha^{T}\phi(x_i)^{T}\phi(x)+b=\sum_{i=1}^{l}\alpha^{T}\kappa(x,x_i)+b \quad (24)$$

2.1.5. Regressive Convolution Neural network and SVR (RCNN-SVR):

Zhang and Li (2018) propose a regressive convolution neural network model and combined the deep neural network with SVR and designed an RCNN-SVR model. This structure has two main step, the feature extraction step in RCNN model, and predicting step in SVR model. The feature extraction is performed by three convolution layers (Conv1, Conv2, Conv3), and three max pooling layer, (Maxpool1, Maxplool2, Maxpool3), one rectified linear units (ReLU) layer, and one normalization (Norm) layer (Deo and Singh, 2107). The prediction step consists of a fully-connected layer and a regression layer. Convolutional layers apply sliding convolutional filters to the input. The layers convolve the input by moving the filters along the input vertically and horizontally and computing the dot product of the weights and the input, and then adding a bias term. A max pooling layer performs down-sampling by dividing the input into rectangular pooling regions, and computing the maximum of each region. A ReLU layer performs a threshold operation to each element of the input, where any value less than zero is set to zero. Finally, a channel-wise local response (cross-channel) normalization layer carries out channel-wise normalization. In prediction stage the support vector regression (SVR) uses features which extracted from RCNN.

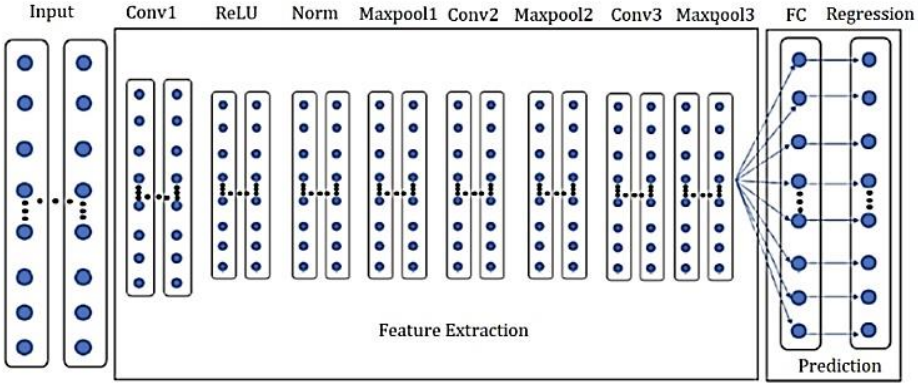


Figure 2. The RCNN-SVR structure. Zhang and Li (2018)

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021


Open Access Drinking Water Engineering and Science Discussions

2.2. Model performance assessment

In this study the performance of models is evaluated by employing Root-Mean Squared Error (RMSE), Normalized Mean Squared Error (NMSE), Normalized Mean Bias Error (NMBE) and Mean Absolute Percentage Error (MAPE) which are defined as below:

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(PBR_i^{obs}-PBR_i^{pre})^2} \quad (25)$$

$$NMSE=\frac{\frac{1}{n}\sum_{i=1}^{n}(PBR_i^{obs}-PBR_i^{pre})^2}{var(PBR^{obs})} \quad (26)$$

$$NMBE=\frac{\frac{1}{n}\sum_{i=1}^{n}(PBR_i^{pre}-PBR_i^{obs})}{\overline{PBR^{obs}}} \quad (27)$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{PBR_i^{pre}-PBR_i^{obs}}{PBR_i^{obs}}\right| \quad (28)$$

Where n is the total number of observed data, PBR^{obs} is the observed value of PBR, PBR^{pre} is the predicted value of PBR, $\overline{PBR^{obs}}$ is mean of the PBR observed values and $var(PBR^{obs})$ is variance of the PBR observed values.

3. Case study: WDN of Joopar city

The WDN of Joopar city is selected as the case study for pipe failure prediction. Joopar with an altitude of 1893 m height above sea level is located in about 25 km south of Kerman, Iran. It has an area of 12 Km^2 and covers 2622 water subscribes with 51.6 km of water distribution pipes (Figure 3 and 4). The network with a lifespan of more than 50 years, was built in the early days with asbestos cement pipes and developed with polyethylene. In this case study, 158 cases of pipe failure for polyethylene (PE) pipes with diameters of 29.4–101.4 mm and 124 cases for asbestos cement (AC) pipes with diameters of 100–200 mm have been used as regression model datasets which have been collected by author during 2012-2019. As mentioned, diameter, length, installation depth, age, maximum and average hydraulic pressure of pipes are considered as the main variables that influence the PBR of pipes. Figure 5 visualizes a graphical representation of these pipe features for burst cases.

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021

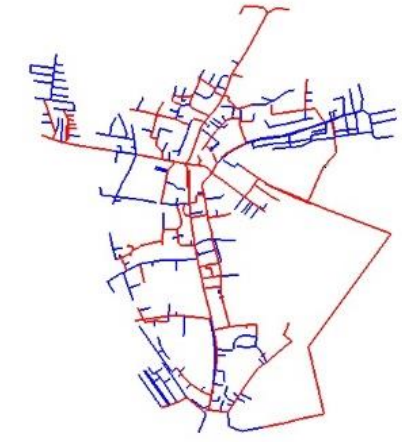
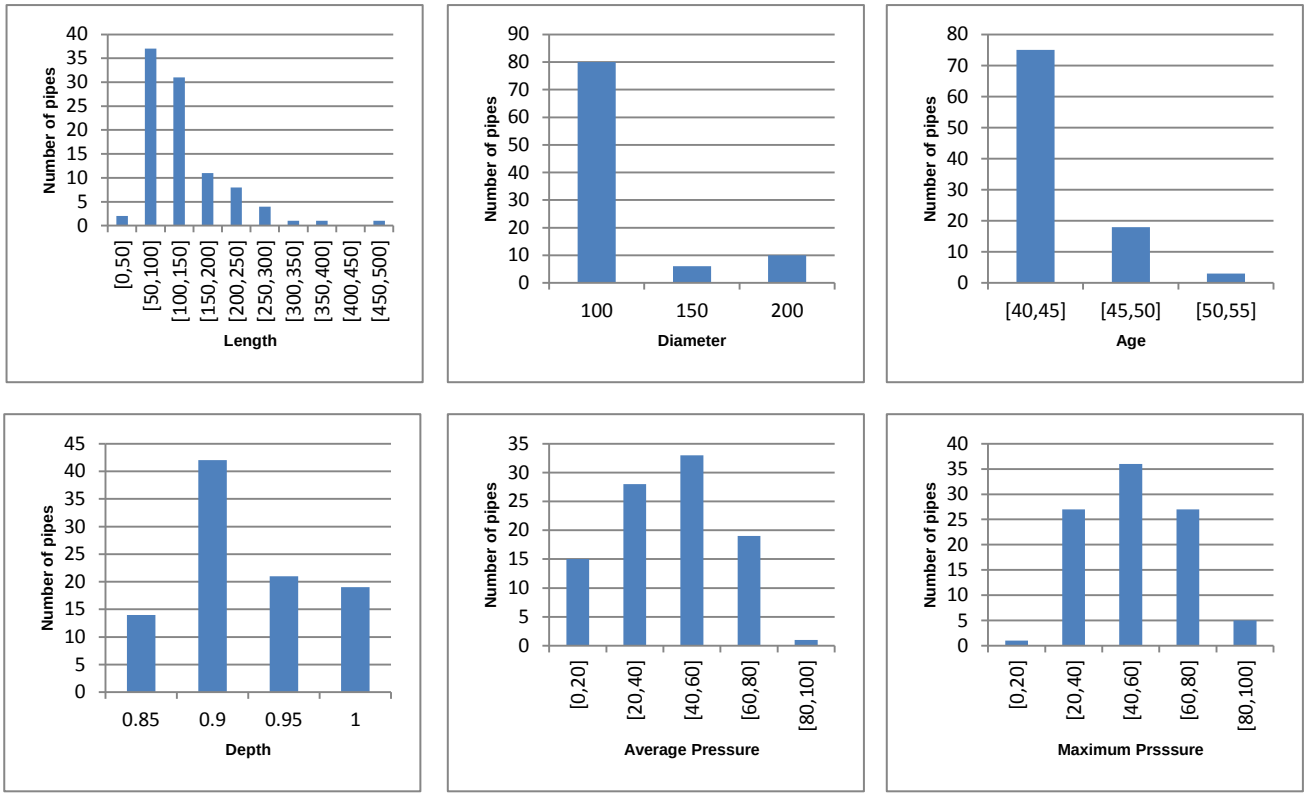


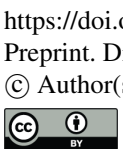
Figure 3. WaterGEMS model of Joopar (Blue lines and red lines represent PE and AC pipes respectively)

Figure 4. WDN of Joopar overlay with the ©Google earth 2021 picture

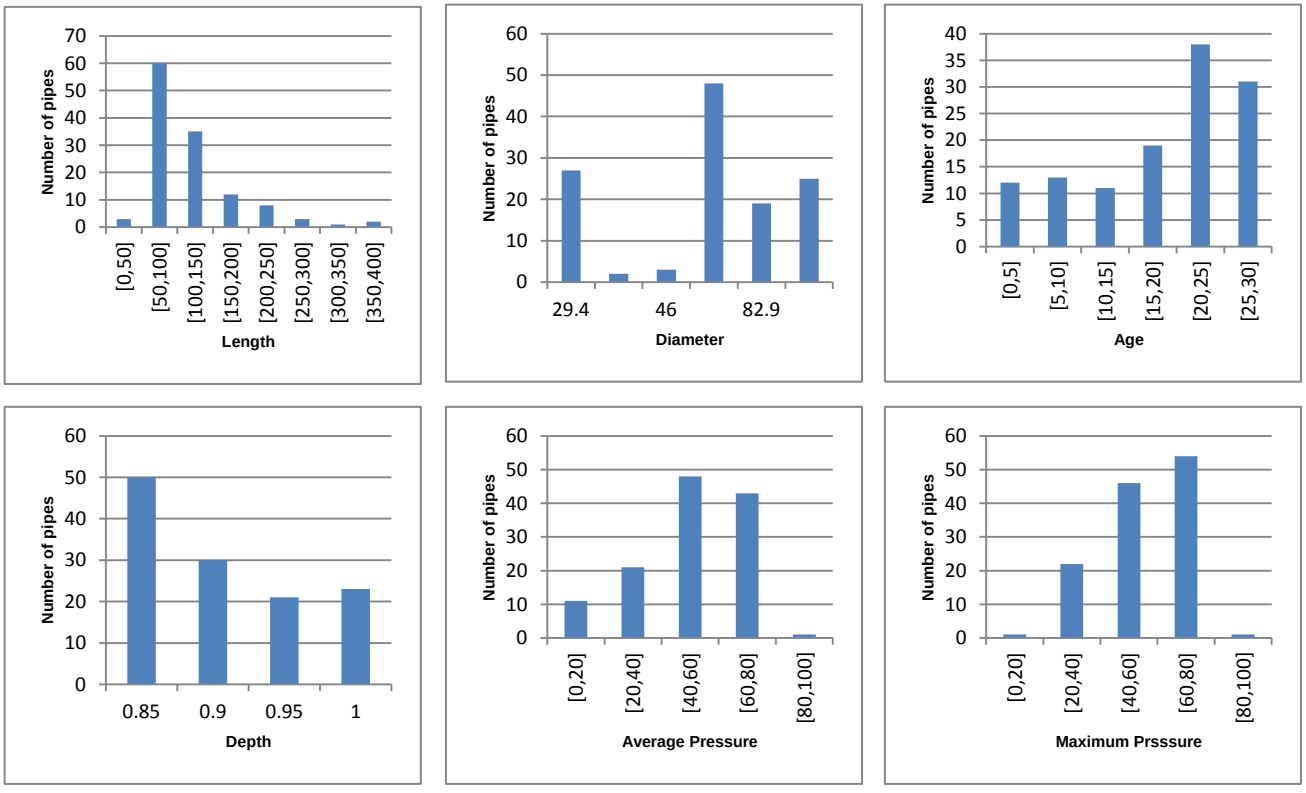


(a)

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021



Open Access Drinking Water Engineering and Science Discussions



(b)

Figure 5. Histograms of each pipe feature (a) asbestos cement and (b) polyethylene

4. Results and discussion

Pearson correlation coefficients between PBR and pipe burst features have been determined to confirm the suggested relationship between the age, diameter, maximum and average pressure of pipes with PBR. Performances of models have been assessed via calculating some error criteria that helps us to find the best regression model.

4.1. Correlation coefficients

The linear relationship of the collected data is measured with the Pearson correlation coefficients. As can be seen from the obtained results listed in table 1, it is evident that there is a correlation between PBR and the pipe burst variables. A higher PBR values obtained with the higher average or maximum pressure of PE pipes. Conversely, increase in the values of length, diameter and depth of PE pipes decrease the corresponding values of PBR.

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021

Open Access Drinking Water Engineering and Science Discussions

It is worth mentioning that the development of the WDN of Joopar city was performed with PE pipes, which have a resistance near existing pressure in the network. So it is expected to have low correlation between age and high correlation between pressure and PBR. On the other hand, as represented results show, there is a negative correlation coefficient between diameter and PBR. Based on local investigations, it has been found that old asbestos cement pipes can bear a pressure more than the present pressure in the network. Findings show there is a strong positive correlation between age and PBR because of aged pipes, verifying that by increasing the age of pipes, PBR will increase. Also it can be seen that there is a positive correlation coefficients between both P_{avg} and P_{max} and PBR .

According to equation (1), PBR has inverse relation with length and because of low variation of failure statistics with length during the investigation period, large negative correlation can be seen in both PE and AC pipes and PBR.

Table 1. Pearson correlation coefficients of the pipe burst rate and the effective parameters for polyethylene and asbestos cement pipes

R(correlation coefficient)	Variables					
	Length	Diameter	Pavg	Pmax	Age	Depth
PBR of PE	-0.561	-0.130	0.241	0.230	0.034	-0.276
PBR of AC	-0.567	0.012	0.162	0.186	0.782	0.065

4.2. Evaluation of regression models performance

According to the mentioned regression techniques, data-driven pipe burst models were set up for the asbestos cement and polyethylene pipes in Joopar WDN. The comparison of five methods results are listed in table 2. According to the calculated values, the RCNN-SVR model, with a relatively long computational time, is the most accurate burst rate predictor of pipes and has the lowest RMSE and MAPE among other methods. According to authors' knowledge, implementing the RCNN-SVR model for PBR prediction has not been reported yet.

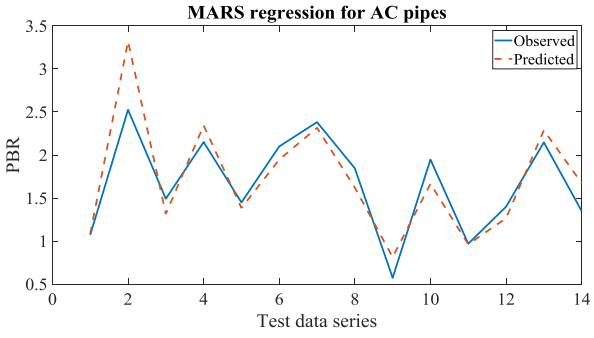
Table 2. Evaluation of the proposed models on the test data in the case study

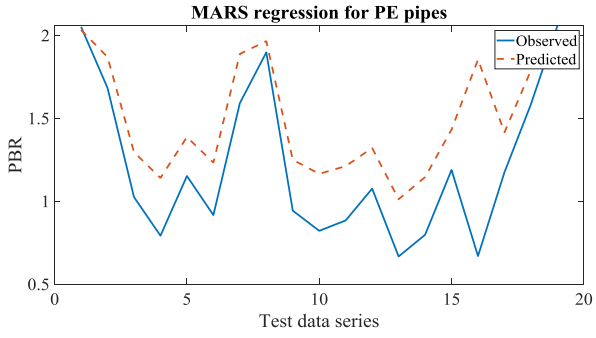
Performance indicator	Models for PE pipes					Models for AC pipes				
	MARS	M5'	FCR	LSSVR	RCNN-SVR	MARS	M5'	FCR	LSSVR	RCNN-SVR
RMSE	0.37	0.30	0.38	0.35	0.052	0.27	0.34	0.20	0.26	0.071
NMSE	0.66	0.41	0.71	0.60	0.13	0.23	0.36	0.13	0.20	0.016
NMBE	-0.24	-0.11	-0.23	-0.25	-0.011	-0.026	-0.11	-0.07	-0.13	-0.013
MAPE	0.33	0.17	0.31	0.30	0.040	0.13	0.14	0.14	0.16	0.04

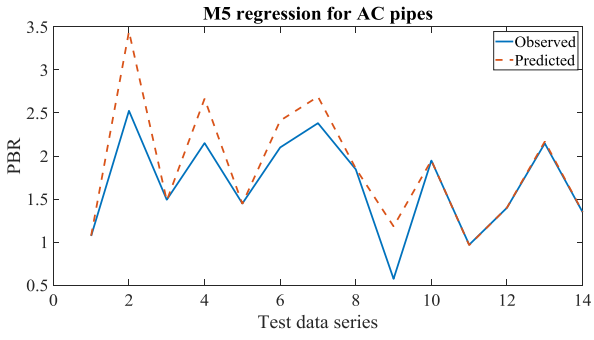
Figure 6 compare the observed PBRs with the values predicted by regression models. Graphs show that the PBR values predicted by the RCNN-SVR model have the best compatibility with the observed PBRs.

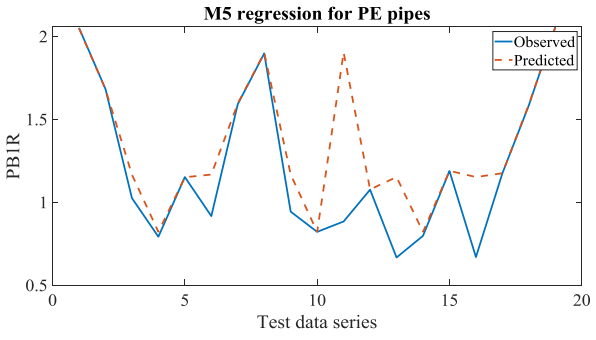
https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021

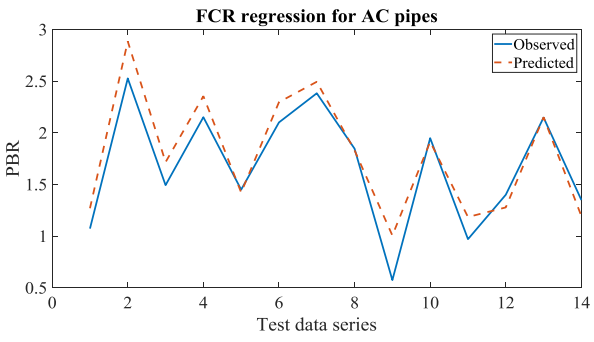
Open Access
Drinking Water
Engineering and Science
Discussions

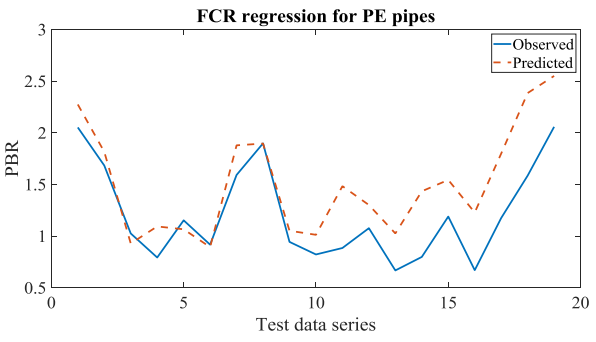


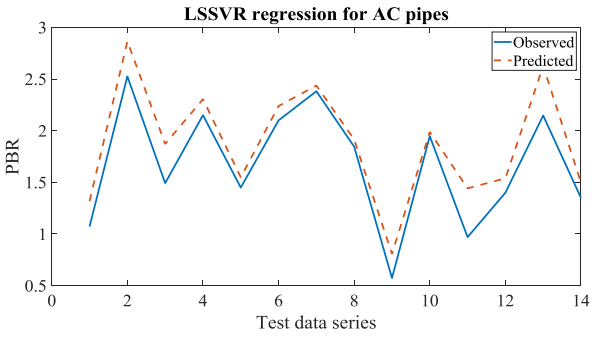


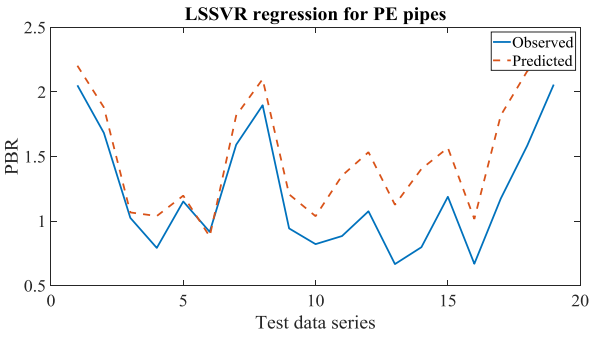












https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021

Open Access
Drinking Water
Engineering and Science
Discussions

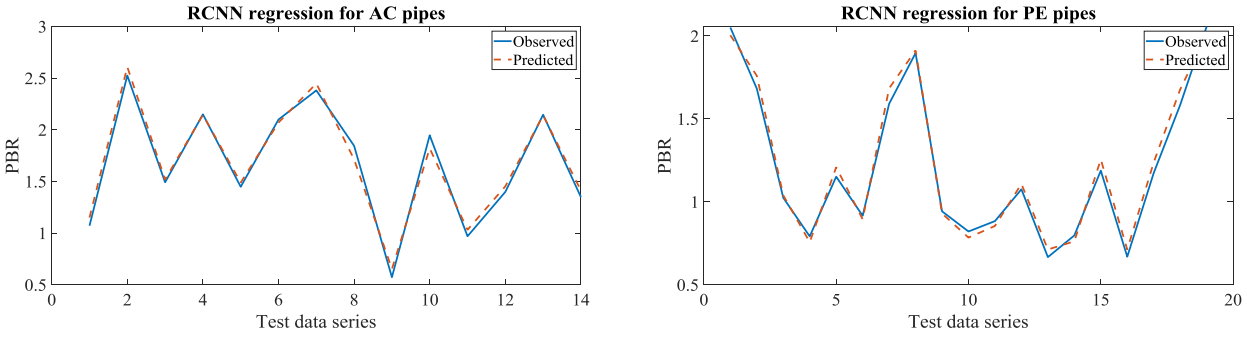


Figure 6. Comparison of the observed PBRs with the values predicted by regression methods for case study

5. Conclusion

Failure of pipes in water distribution networks (WDNs) is an inevitable event that leads numerous issues. Prediction of pipe burst helps to optimize the budget allocation and better utilization programming. This paper compared and evaluated five artificial intelligence and machine learning methods; multivariate adaptive regression splines (MARS), M5' regression tree (M5'), Least square support vector regression (LS-SVR), fuzzy regression based on c-means clustering (FCMR) and regressive convolution neural network with support vector regression (RCNN-SVR) for pipe failure prediction and implemented them in Joopar WDN as a real case study. Pipe failure data were collected during an eight-year-period and consist of 124 cases for asbestos cement (AC) and 158 cases for polyethylene (PE) pipes with 100-200 mm and 29.4-101.4 mm diameter respectively. Models were setup based on pipes age, diameter, length, installation depth, maximum and average hydraulic pressure of pipes as the input variables and pipe burst rate (PBR) as model output. Models performance has been compared with error assessment criteria such as Normalized Mean Bias Error (NMBE), Normalized Mean Squared Error (NMSE), Root-Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Findings show that RCNN-SVR is the most accurate prediction model, which has the lowest values of RMSE, NMSE and MAPE which can effectively predict the burst rate. The positive correlation coefficient between age and PBR is high in approximately 50-year-old AC pipes and low in PE pipes. Also analyses show that there is positive correlation between pressure and PBR for PE and AC pipes. As length is one of the main parameters in PBR formula, the correlation between length and PBR is evident.

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021


Open Access Drinking Water Engineering and Science Discussions

295

References

Achim, D., Ghotb, F., and McManus, K. J., :Prediction of Water Pipe Asset Life Using Neural Networks, Journal of Infrastructure Systems 13 (1): 26–30. doi:10.1061/(ASCE)1076-0342(2007)13:1(26), 2007.

Alexopoulos, E. C., :Introduction to multivariate regression analysis, Hippokratia vol. 14, 23-8, 2010.

300 Ameer, A., Karmakar, M., Gour, C., and Dooley, L. S., :Review on Fuzzy Clustering Algorithms, Journal of Advanced Computations, 2(3) pp. 169–181, 2008.

Berardi, L., Giustolisi, O., Kapelan, Z., and Savic, D. A.: Development of pipe deterioration models for water distribution systems using EPR. Journal of Hydroinformatics, 10(2), 113–126. https://doi.org/10.2166/hydro.2008.012, 2008.

Bezdek, J.C., :Pattern Recognition with Fuzzy Objective Function Algorithm., New York: Plenum Press, 1981.

305 Bezdek, J. C., Ehrlich, R., and Full, W., :FCM: The fuzzy c-means clustering algorithm, Computers & Geosciences, 10(2–3), 191–203, 1984.

Christodoulou, S., Aslani, P., and Vanreterghem, A., :A Risk Analysis Framework for Evaluating Structural Degradation of Water Mains in Urban Settings, Using Neurofuzzy Systems and Statistical Modeling Techniques, In Proceedings of the World Water and Environmental Resources Congress and Related Symposia. Philadelphia, PA: ASCE, 2004.

310 Dave, R.N., :Boundary detection through fuzzy clustering, IEEE International Conference on Fuzzy Systems. pp. 127-134, 1992.

Deo, R. C., Kisi, O., and Singh, V. P., :Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. Atmospheric Research, 184, 149–175, 2017.

Ditthakit, P., and Chinnarasri, C., :Estimation of Pan Coefficient using M5 Model Tree. American Journal of Environmental
315 Sciences. 8. 95-103, 2012.

Emamgholizadeh, S., Bateni, S., Shahsavani, D., Ashrafi, T., and Ghorbani, H., :Estimation of soil cation exchange capacity using Genetic Expression Programming (GEP) and Multivariate Adaptive Regression Splines (MARS). Journal of Hydrology. 529. 10.1016/j.jhydrol.2015.08.025, 2015.

Forghani, A., and Peralta, R., :Transport modeling and multivariate adaptive regression splines for evaluating performance
320 of ASR systems in freshwater aquifers. Journal of Hydrology, 553, pp.540-548, 2017.

Friedman, J. H., :Multivariate Adaptive Regression Splines, Ann. Statist. 19 no. 1, 1-67. doi:10.1214/aos/1176347963, 1991.

Grigg, N., :Main Break Prediction, Prevention and Control, American Water Works Research Foundation, 2007.

Hathaway, R. J., and Bezdek, J. C., :Switching regression models and fuzzy clustering, in IEEE Transactions on Fuzzy Systems, vol. 1, no. 3, pp. 195-204, Aug., doi: 10.1109/91.236552, 1993.

325 Heddam, S., and Kisi, O., :Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J Hydrol 559:499–509, 2018.

Kakoudakis, K., :Pipe failure prediction and impacts assessment in a water distribution network, a thesis for the degree of doctor of philosophy in engineering, the University of Exeter, March 2019.

Kleiner, Y., and Rajani B., :I-warp: Individual water main renewal planner, Drinking Water Engineering and Science, vol. 3,
330 pp. 71–77, 2010.

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021


Open Access Drinking Water Engineering and Science Discussions

Kutylowska, M., :Regression Methods for Predicting Rate and Type of Failures of Water Conduits, Ecological Chemistry and Engineering A 24 (2): 193–205, 2017.

Kutylowska, M., :Forecasting Failure Rate of Water Pipes, Water Supply 19 (1): 264–273. doi:10.2166/ws.2018.078, 2018.

Kutylowska, M., :Application of MARSplines Method for Failure Rate Prediction, Periodica Polytechnica Civil Engineering 63 (1), 87–92, 2019.

Quinlan, J. R., :Learning with continuous classes, in Proceedings of the 5th Australian Joint Conference on Artificial Intelligence (AI '92), pp. 343–348, World Scientific, Singapore, 1992.

Rahimikhoob, A., Asadi, M., and Mashal, M., :A comparison between conventional and M5 model tree methods for converting pan evaporation to reference evapotranspiration for semi-arid region, Water Resources Management, vol. 27, no. 14, pp. 4815–4826, 2013.

Rajani, B. and Kleiner, Y., :Comprehensive review of structural deterioration of water mains: physically based models, Urban Water, vol. 3, no. 3, pp. 151–164, https://doi.org/10.1016/S1462-0758(01)00032-2, 2001.

Rajani, B., and Tesfamariam, S., :Estimating Time to Failure of Cast Iron Water Mains, Proceedings of the ICE - Water Management 160 (2): 83–88, 2007.

Randall-Smith, M., Russell, A., and Oliphant, R., :Guidance Manual for the Structural Condition Assessment of the Trunk Mains, Swindon, UK, Water Research Centre, 1992.

Rezaie-Balf, M., :multivariate adaptive regression splines model for prediction of local scour depth downstream of an apron under 2D horizontal jets. Iran J Sci Technol Trans Civ Eng 43(1):103–115, 2019.

Safari, M. J. S., :Decision tree (DT), generalized regression neural network (GR) and multivariate adaptive regression splines (MARS) models for sediment transport in sewer pipes. Water Science & Technology. 79. 1113-1122. 10.2166/wst.2019.106, 2019.

Sharda, V., Prasher, S., Patel, R., Ojasvi, P. and Prakash, C., :Performance of Multivariate Adaptive Regression Splines (MARS) in predicting runoff in mid-Himalayan micro-watersheds with limited data / Performances de régressions par splines multiples et adaptives (MARS) pour la prévision d'écoulement au sein de micro-bassins versants Himalayens d'altitudes intermédiaires avec peu de données. Hydrological Sciences Journal, 53(6), pp.1165-1175, 2008.

Suykens, J. A. K., :Vandewalle, Joos P. L.; "Least squares support vector machine classifiers, Neural Processing Letters, vol. 9, no. 3, Jun, pp. 293–300, 1999.

Tabesh, M., Soltani, J., Farmani, R., and Savic, D., :Assessing Pipe Failure Rate and Mechanical Reliability of Water Distribution Networks Using Data-driven Modeling, Journal of Hydroinformatics 11 (1): 1–17. doi:10.2166/hydro.2009.008, 2009.

Vapnik, V., :The nature of statistical learning theory, Springer-Verlag, New-York, 1995.

Vapnik, V., :Statistical learning theory, John Wiley, New-York, 1998.

Vapnik, V., :The support vector method of function estimation, In Nonlinear Modeling: advanced black-box techniques, Kluwer Academic Publishers, Boston, pp.55-85, 1998.

Wang, C., Z. Niu, Jia, H., and Zhang, H., :An Assessment Model of Water Pipe Condition Using Bayesian Inference., Journal of Zhejiang University-Science A (Applied Physics and Engineering) 11 (7): 495–504. doi:10.1631/jzus.A0900628, 2010.

Wang, Y., Witten, IH., : Induction of model trees for predicting continuous classes. In: Proceedings of the Poster Papers of the European Conference on Machine Learning, University of Economics, Faculty of Informatics and Statistics, Prague, 1996.

https://doi.org/10.5194/dwes-2021-7
Preprint. Discussion started: 29 March 2021

Open Access Drinking Water Engineering and Science Discussions

Wilson, D., Filion, Y., and Moore, I., :State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains, Urban Water Journal, 14(2), 173–184, 2015.

Witten, I.H. and Frank,E. :Data Mining: Practical Machine Learning Tools and Technique. 3rd Edn., Kaufmann Publishers, San Francisco, USA., ISBN-10: 0120884070, pp: 664, 2005.

375 Zhang Y., Guo L., Li Q., and Li J., :Electricity consumption forecasting method based on mpso-bp neural network model. proceedings of the 2016 4th international conference on electrical electronics engineering and computer science (iceeecs 2016), 50:674-678, 2016.

.