Pipeline Failure Prediction in Water Distribution Networks using Evolutionary Polynomial Regression combined with k-means clustering

Konstantinos Kakoudakis[1*], Kourosh Behzadian[2], Raziyeh Farmani[1] and David Butler[1]

[1]*College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, EX4 4QF, UK*

[2]*School of Computing and Engineering, University of West London, London, W5 5RF, UK*

[*]Corresponding author: Konstantinos Kakoudakis (email kk337@exeter.ac.uk)

Pipeline Failure Prediction in Water Distribution Networks using Evolutionary Polynomial Regression combined with k-means clustering

Abstract

This paper presents a new approach for improving pipeline failure predictions by combining a data-driven statistical model, i.e. Evolutionary Polynomial Regression (EPR), with *k*-means clustering. The EPR is used for prediction of pipe failures in case iron pipes based on length, diameter and age of pipes as explanatory factors. Individual pipes are *aggregated* using their attributes of age, diameter and soil type to create homogenous groups of pipes. The *k*-means clustering is employed to partition input data into a number of clusters for individual EPR models. The proposed approach was demonstrated by application to a water distribution network in the UK. The prediction accuracy was evaluated using a cross-validation technique. Results show the proposed approach is able to significantly reduce the error of pipe failure predictions especially in the case of a large number of failures. The prediction models were used to calculate the failure rate of individual pipes for rehabilitation planning.

Keywords: Evolutionary Polynomial Regression, *K*-means clustering, pipe failure predictions, water distribution networks

1. Introduction

Due to the high economic, environmental and social costs resulting from pipe bursts in water distribution systems, development of a reliable and accurate prediction model of pipe failure is of paramount importance. Failure of water pipes can be the cumulative effect of various pipe-intrinsic, operational and environmental factors. Pipe failure implies a decrease in the service level, resulting in economic, environmental and social costs. Water utilities usually follow one of two rehabilitation strategies: reactive or proactive (Røstum 2000). In a reactive strategy, a pipe will be rehabilitated after failure is detected whereas pipe rehabilitation in a proactive strategy is scheduled in advance after assessing and forecasting pipe propensity to fail. Due to the advantages of taking a proactive approach

(e.g. maintenance/improvement of current level of service), researchers and practitioners have striven to develop predictive models in which the likelihood of pipe failure is identified for future planning of replacement/ rehabilitation.

Predictive models can be classified into physical (Rajani and Kleiner 2001), statistical (Kleiner and Rajani 2001; Scheidegger *et al*. 2015) and data-driven entailing artificial neural network (Clair and Sinsha 2012) and evolutionary polynomial regression (Giustolisi and Savic 2006; Berardi *et al*. 2008). Physical models analyse the loads to which the pipes are subject and the capacity of the pipes to resist these loads in order to predict their propensity to break (Rajani and Kleiner 2001). In spite of having a reasonable accuracy, physical models compared to other methods have significant input data demands due to the fact that they try to simulate the mechanisms that lead to pipe failure whereas the other methods try to identify breakage patterns using historical failure data. These demands involve gaining an understanding of structural behaviour of buried pipes, pipe-soil interaction and knowledge about the quality of installation, internal and external stresses and material deterioration (e.g. external and/or internal corrosion). The relatively high cost of obtaining these data can be justified only for major transmission water mains where the cost of failure is high. In contrast, statistical models are applicable to various levels of input data and capable of linking pipe breakage patterns to various pipe descriptive variables and other environmental and operational factors using regression analysis of historical pipes break data (Kleiner and Rajani 2001). In order to overcome the complexity of failure patterns observed in water networks and capture the non-linear interactions between various parameters, data-driven methods such as Artificial Neural Networks (ANNs) have also been developed (Ahn *et al*. 2005; Achim *et al*. 2007, Tabesh *et al*. 2009). ANNs are data-driven 'black-box' models, able to capture the complex relationship between input and output pipe failures using a non-linear learning process and with no assumption of the form of the relationship between the variables.

Evolutionary Polynomial Regression (EPR) is another data driven method that can be used for prediction of mains pipe breaks (Giustolisi and Savic 2006; Berardi *et al.* 2008). EPR provides a range of statistical equations of pipes failure prediction in a trade-off between training model accuracy and number of polynomial terms (i.e. model complexity). This particular feature can be counted as the main strength of EPR giving a flexible approach to the decision maker to select the most appropriate

polynomial model. However, the single polynomial regression model must capture different failure patterns in the entire database. To overcome this limitation and better understand the patterns of pipes failure, Xu *et al*. (2011) first partitioned the pipe database into two clusters of those installed before the monitoring period and the others after the monitoring period. They then developed two distinctive prediction EPR models, one for each cluster. Although this clustering approach enhanced the failure prediction accuracy to a certain extent, a more precise clustering approach is required to accommodate the high variability of pipes failure patterns and thus improve the accuracy of predictive models. Therefore, this paper presents a novel predictive method by combining an Evolutionary Polynomial Regression model with the *k*-means clustering method (MacQueen 1967) with the aim to achieve more accurate predictions of the expected number of pipe failures. The rest of the paper is organized as follows. The second section describes the details of the proposed methodology. A description of the case study employed to demonstrate the methodology is given in Section 3 followed by the EPR settings in Section 4. The results are presented and discussed in Section 5 with key findings final remarks are given in the conclusions.

2. Methodology

Figure 1 shows the framework of the proposed methodology in three phases (or seven steps) of data preparation, model development and model test. The software used to create the clusters is MATLAB while EPR-MOGA-XL vr.1 (Giustolisi and Savic 2009; Giustolisi *et al*. 2009) is employed to develop the EPR models.

In the first phase, the individual pipes are aggregated into homogenous groups using pipe descriptive variables and environmental factors. This is based on the assumption that pipes with similar specific intrinsic properties such as material, diameter and age are expected to have the same breakage pattern (Kleiner and Rajani 2012). In addition to the pipe characteristics, soil type, as an environmental factor, is used as an aggregation criterion because soil properties have been associated with the corrosion of the metallic pipes (Sadiq *et al*. 2004; Kabir *et al*. 2015) and this is a dominant factor contributing to their failure (Makar 2000; Folkman 2012). Each aggregated homogenous class of pipes (i.e. with

specific age, diameter and soil type) takes a length and a number of failures equal to the total lengths and total number of failures for the individual pipes of the same attributes, respectively. The original dataset containing a large number of individual pipes is converted to a new dataset containing homogenous groups of pipes based on diameter, soil type and age. Homogenous groups are then divided into two subsets to provide the training and test datasets used in phases 2 and 3, respectively. More specifically, the training dataset is used to calibrate the predictive models while the test dataset is used for model validation and performance assessment.

The second phase entails two consecutive steps (2 and 3) to develop the *k* predictive EPR models corresponding to the *k* subsets. This is achieved by first partitioning the training dataset into *k* clusters based on diameter and age of groups using the *k*-means algorithm (step 2). Then, one specific EPR model is developed for each data cluster of the database.

In phase 3, the performance of the developed models is evaluated by using test data in four steps. The Euclidian distance of input variables (i.e. age and diameter) between the test data sample and the counterpart cluster centre values (known as centroids) is calculated to identify the suitable cluster for each test data. The corresponding EPR model associated with the relevant cluster is selected in step 5 to predict the number of pipe failures in step 6. By calculating the number of failures using the *k* EPR models for all test data samples, performance indicators can be evaluated by using the predicted number of failures for the test dataset and the corresponding observations. Various numbers of clusters are tested to identify the optimal number which provides the highest improvement compared to the non-clustered EPR. Further details of the EPR models and the *k*-means algorithm used in this paper are described individually in the following sections.

2.1 Evolutionary Polynomial Regression

Evolutionary Polynomial Regression (Giustolisi and Savic 2006) is a data-driven method based on numerical and symbolic regression that is able to produce series of pseudo-polynomial models. After the user selects the generalised model structure, EPR employs a multi-objective search strategy to estimate unknown constant parameters of the assumed models using the least squares method. As a

result of the multi-objective optimization approach, each single EPR run returns a number of polynomial models on a Pareto optimal front which is a trade-off between accuracy (fitness) and parsimony. The first criterion aims to maximise the model fit to the observed data (or minimise the model error) and the second (parsimony) aims to minimise the number of explanatory variables and/or polynomial terms in the model. Here, the number of polynomial terms is a surrogate for the model parsimony criterion. Its role is to prevent over-fitting of the model to data and thus endeavour to capture underlying general phenomena without replicating noise in data. Finally, the user can select the model of interest with respect to a specified model accuracy and/or parsimony. The general form of polynomial EPR model (Giustolisi and Savic 2006) is expressed as:

$$Y = \sum_{j=1}^{m} F\left(X, f(X), a_j\right) + a_o \quad (1)$$

where Y= estimated output; a_j= unknown polynomial coefficients (i.e. model parameters); F= function finally constructed by the EPR process; X= the matrix of explanatory variables; f= function selected by the user; and m= the maximum number of polynomial terms and a_0= unknown constant.

The specific model structure selected here for analysis of pipe failure is (Giustolisi and Savic 2006):

$$Y = \sum_{j=1}^{m} a_j \left((X_1)^{E_{1j}} \dots (X_i)^{E_{ij}}\right) + a_0 \quad (2)$$

where Y=predicted number of pipe failures, X_i =explanatory variable i, E_{ij} =matrix of unknown exponents. The candidate explanatory variables (X) that we use for pipe failure predictive model are the total group length (L), the diameter (D) and the age (A) of pipes.

The first step in applying the EPR is to establish the inputs and the output used in the process. The 'explanatory variables' considered here are the total length, diameter and age and the 'dependent variable' is the total number of failures within each homogenous class all for cast iron pipes. The 'explanatory variables' considered are the available for the examined network. Note that both failed and non-failed pipes are considered here in the database as the models aim to develop a relationship not only between pipe failure and explanatory variables but between all pipes of the same material and associated explanatory variables.

2.2 K-means clustering

K-means clustering as a data clustering approach is used here to partition dataset of pipeline failure into specific number of clusters (i.e. *k*) based on the available pipelines attributes (i.e. diameter and age of groups) based on the *k*-means algorithm. Generally, data clustering is a data exploration technique that groups objects with similar characteristics together and thus classifies a large number of objects into a small number of clusters in order to facilitate their further processing (Pham *et al*. 2005). The creation of the clusters is based on the principle of maximising the intra cluster similarity and minimising the inter cluster similarity (Wettschereck *et al*. 1997). *K*-means is an unsupervised learning algorithm popular due to its simplicity and efficiency (Kanungo *et al*. 2002). It is based on assigning *n* data samples into *k* clusters such that an objective function of dissimilarity (or distance) is minimised (Jang *et al*. 1997). The search algorithm moves data samples between clusters until the objective function cannot be minimised further. In the case of the dissimilarity measure, minimisation of the Euclidean distance is usually chosen as the objective function as (Kim and Keo 2015):

$$J=\sum_{j=1}^{k}\sum_{i=1}^{n}\left|x_i^{(j)}-c_j\right|^2 \quad (3)$$

where $\left|x_i^{(j)}-c_j\right|^2$ = Euclidean distance of specified criteria between *i*th data sample $x_i^{(j)}$ and *j*th cluster centre c_j; $x_i^{(j)}$= vector of specified criteria for *i*th data sample assigned to *j*th cluster centre; *J*=overall distance indicator for the *n* data samples from their respective cluster centres.

The *k*-means clustering applied here uses the KMEANS function in MATLAB (® R2014b) to partition the training data into a number of specified clusters based on diameter and age. Note that *k*-means' reliance on a predefined number of clusters is regarded as a drawback (Kim and Seo 2015) and therefore various number of clusters are analysed here to identify the optimal number. For any specific number of *k*, the number of EPR models developed is equal to this number of clusters. Each of the *k* developed EPR models was associated with the training data of the relevant cluster. For the test dataset, only the Euclidean distance of the input variables is calculated and compared between each test data sample and all the centroids to identify the suitable cluster according to the smallest distance.

2.3 Model performance assessment

Once a data-driven model is developed, its performance needs to be validated for an 'unseen' dataset and compared with similar benchmark models in fair conditions. One of the most common ways to assess model prediction ability is so-called hold-out validation based on a single split of the data, i.e. dividing the entire dataset into two subsets, for example, 70% for training/calibration and 30% for test (Garthwaite and Jolliffe 2002). However, model performance derived by this approach would depend significantly on the selection of the training and test datasets. In particular, if the data have not been evenly distributed over the training and test datasets, this validation may not be a true representation of model performance (e.g. the error in training data is much larger than the one in the test data). In addition, in the case of an insufficient or small number of available data, model performance in a single split approach would be poor due to the very small portion of data left for validation. To overcome this drawback and given the small number of pipe failure data after pipe grouping, cross-validation method (Grossman *et al*. 2010) is used here for assessment of performance indicators in predictive models. (See supplementary materials for this method). The performance indicators used here are the Coefficient of Determination (R^2) as a measure for correlation between predictions and observations and Root Mean Square Error (RMSE) as a measure for error prediction. The mathematical relationships of these indicators are expressed as follows (Moriasi *et al*. 2007):

$$R^2=\frac{(\sum_{i=1}^{j}(y_{p,i}-\bar{y}_p)(y_{o,i}-\bar{y}_o))^2}{\sum_{i=1}^{j}(y_{p,i}-\bar{y}_p)^2\sum_{i=1}^{j}(y_{o,i}-\bar{y}_o)^2} \quad (4)$$

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{j}(y_{p,i}-y_{o,i})^2}{j}} \quad (5)$$

where $y_{p,i}$ = prediction value for test sample i; $y_{o,i}$ = measurement value for test sample i, $\bar{y}_p$ = mean value of predictions, $\bar{y}_o$ = mean value of measurements and j = the number of test data samples.

3. Case study

The proposed methodology is demonstrated here for prediction of pipe failure in a case study located in part of a water distribution network of a UK city. The database contains a large number of individual pipes made of five different materials. Preliminary analysis showed that the highest pipe failure rate

(number of failures/km/year) is 0.258 for Cast Iron (CI) pipes compared to other material types which are 0.079 for Asbestos Cement (AC) pipes, 0.080 for Ductile Iron (DI) pipes, 0.015 for Polyethylene (PE) pipes and 0.118 for Polyvinyl chloride (PVC) pipes. In addition, pipe records show that 85% of the failed pipes are made of CI pipes which constitute 73% of the network's total length. Based on these findings, it can be concluded that the CI pipes are more prone to failure and therefore only they are considered in this paper for construction of the predictive models.

4. EPR settings

Before running the EPR model, potential values for its parameters in Eqs. (1) and (2) need to be set up. The candidate values considered for exponents (E_{ij}) in Eq. (2) were -2, -1, -0.5, 0, 0.5, 1 and 2 which describe potential square, linear or square root exponents for explanatory variables of the EPR model. The value 0 was chosen to deselect input candidates with no influence on the output, while the positive and negative values were considered to describe potential direct and inverse relationship between the inputs and the output of the model. The maximum number of polynomial terms was set to 3 (i.e. m=3) excluding the constant term (a_0) to ensure the best fit without unnecessary complexity. Unnecessary complexity is defined as the addition of new terms that fit mostly random noise in the raw data rather than the underlying phenomenon. The result of each single EPR run is three regression models corresponding to the maximum number of polynomial terms defined in advance.

5. Results and discussion

Following the procedure described above for the data preparation, grouping of individual pipe failure data resulted in 141 data samples for developing the EPR models. In order to avoid over-fitting and in compliance with the parsimony rules, one polynomial term EPR model was selected from the Pareto front for all model runs analysed in this paper (Berardi *et al*. 2008). The cluster based approach was applied for different numbers of clusters (k) and the most appropriate number of clusters was identified by comparing the performance indicators. The results showed that the two performance indicators are improved by increasing the number of clusters until six clusters when no further improvement is

achieved for both training and test data (Figure 2). The comparison indicates that the most accurate results are achieved with the six-clustered EPR approach. Another limiting factor for increasing number of clusters is the number of data samples assigned to each cluster for model training. The number of samples needs to be equal or greater than the number of parameters to be estimated in the construction phase of the EPR model. With respect to this criterion, the six-clustered EPR was satisfactory as the minimum number of samples in one of the clusters was 7 (Figure A.2 in the supplementary materials) which was greater than the number of parameters to be identified in the EPR (i.e. 4).

For comparative purposes, the results obtained from the cluster-based EPR models are compared here with the non-clustered EPR. Figure 2 shows the two performance indicators (R^2 and RMSE) of the predictive models for both training and test data using the cross-validation technique. The results show that both performance indicators for the clustered EPR models are better than the non-clustered EPR approach for all the different number of clusters and for both training and test data. More specifically, the comparison of the six-clustered EPR with the non-clustered EPR shows a significant improvement especially for the test (i.e. improvement of 34% for RMSE and 10% for R^2). All these can be attributed to the fact that clustering would be beneficial for pipe failure analysis and thus more appropriate EPR models fitted to the clustered data are identified effectively.

Table 1 lists the associated models obtained from developing the six-clustered EPR and non-clustered EPR corresponding to one of the ten iterations of cross-validation. In both models, total number of pipe failures (Y) were selected from one polynomial term comprising of total group length (L), the diameter (D) and the age (A) of pipes with the defined candidates of exponents. Note that one polynomial term prediction model was selected and preferred here for all models in order to avoid possible overfitting of test data.

The available pipe failure data typically show one or more of the following characteristics (Scheidegger *et al*. 2015): right censored observations, left truncation and absence of replaced pipe data. The left truncation which is the case in the examined network occurs when the pipes were installed before their failures were systematically recorded. As a consequence the number of failures before the beginning of the monitoring period is not known. The monitoring period corresponds to a portion of the in-usage period which can be small or big depending on the pipes' installation year and during its duration there

is a not a clear trend between failure rate and age. The counterintuitive (inverse) relationship between the number of failures and age observed in Table 1 is attributed to the absence of data for the entire in-usage period. In order to overcome the left-truncation character of the data the proposed methodology was applied only on pipes installed from 1955 and later since they show a descending failure rate. Table 2 lists the associated models obtained from the two approaches corresponding to one of the ten iterations of cross-validation. The obtained models show a direct relationship between pipe failure and age.

5.1 Comparison between EPR and Six-clustered EPR

Further analysis of this comparison can be seen in Figure 3 where the RMSE of the test data is plotted for both models based on different intervals of the number of pipe failures. This quantifies the initial impression that the clustered EPR is able to decrease prediction errors in most intervals especially giving a substantial error reduction for pipe failure events with a large number (i.e. 135-330 interval). In addition, although the improvements of the RMSE for the intervals with a low number of failures (i.e. 0-1 and 2-5) is small in absolute terms, the overall model accuracy improvement is significant due to impact on over 70% of the database. The model prediction of the clustered EPR is poorer than the EPR only for a few intervals which only accounts for 5% of the database (For further verification, see Figure A.3 in the supplementary materials). All this can be due to the fact that the clustered EPR can better represent the behaviour of pipeline failure by clustering the database of the pipe characteristics (i.e. age and diameter) and dedicating a specific EPR for each cluster.

The accuracy of predictions for pipe failure rates in different pipe characteristics (i.e. diameter and age) is compared here for both models in Figure 4. It is evident that EPR is unable to precisely predict small pipe diameter failure whereas this prediction has substantially improved for the six-clustered EPR (i.e. average failure rates for different pipe diameters in Figure 4a). This is due to the fact that the six-clustered EPR employs a number of models to predict pipe failures of different clusters while the EPR is limited to a single model for all pipe characteristics. High variability of number of failures in small pipe diameters could be another possible explanation for the inaccuracy of the single EPR model and thus large prediction errors. Failure predictions for other pipe diameters have also improved in the

clustered EPR compared to the EPR that tend to highly overestimate true pipe failure rates. The imprecision of the EPR predictions is more apparent for different pipe ages especially for old pipes (Figure 4b). However, the predictions for the six-clustered EPR show its ability to predict true pipe failure rates with a relatively reasonable accuracy in most age groups.

5.2 Spatial variation of pipe failure rate

The predictive models have been used to spatially represent failure rates of individual pipes in the water distribution network and classify them in different ranges to identify more vulnerable regions as also shown by Kabir *et al*. (2015). The observed failure rates (expressed as number of failures/km/year) of individual pipes were classified using the Jenks Natural Breaks method (Jenks, 1963). This method divides the data into four ranges as 'very low' [0-0. 097], 'low' [0.097-0.248], 'high' [0.248-0.4570] and 'very high' [greater than 0.457]. Comparison between the accuracy of the two predictive models can be summarised in the overall percentage of pipe failure rates in different ranges as shown in Figure 5. It is apparent that the overall percentages of pipe failure predictions in the six-clustered EPR relates more closely to observations than the EPR in all ranges. More specifically, the EPR model has either overestimated ('low' and 'very high' ranges) or underestimated 'very low' and 'high' ranges) the percentages of observed pipe failure rates.

Furthermore, the portion of those failure rate predictions which are in the correct observation ranges are shown in Figure 5 as shaded areas in the prediction bars along with a correct predictions percentage of the associated ranges. As it can be seen, the clustered EPR has more correct predictions than the EPR predictions in most ranges (i.e. 'Low', 'High' and 'Very high'). In 'Low' failure rate, although the EPR has been able to predict with a relatively similar performance (86% vs 85%), it has a high proportion of wrong predictions compared to the corresponding range of the clustered model. Even for a small percentage of 'Very low' pipe failure rate, the EPR was unable to predict whereas the clustered EPR model could identify most of true failure rates in this range. Similarly, a large percentage of the EPR predictions in 'High' and 'Very high' rates fail to fall within the correct ranges of pipe failures. All this

can be linked to the fact that the clustered input data are associated with the most relevant clustered models which result in more accurate predictions (see Figure A.4-6 in supplementary materials).

6. Conclusions

This study presents a new model to predict failures of cast iron pipes in water distribution networks by combining Evolutionary Polynomial Regression and *k*-means clustering. This was achieved by partitioning the input data based on pipe characteristics (i.e. diameter and age) into a predefined number of clusters using a *k*-means algorithm and an individual EPR model was developed for each created cluster. Individual EPR models were used to predict the number of failures as functions of pipe diameter, age and length from aggregated homogenous pipe databases. The performance of the clustered EPR model was compared with the non-clustered EPR in a case study using a cross-validation technique. The following can be concluded here:

- Combining *k*-means clustering with the EPR results in a considerable improvement of the prediction accuracy for pipe failures.
- The clustered EPR model can be effectively used to predict and identify individual pipe failure rates with different ranges and a high accuracy.
- The clustered predictive model is specifically capable for prediction of extreme pipe failures (i.e. both small and large number of failures). This could be very useful for water utilities managers to make more informed and precise decisions for future rehabilitation planning.

Although the proposed clustered approach is able to accurately calculate the pipe failures, it may suffer from some shortcomings. The first is the need to specify the number of clusters in advance of developing the model. This necessitates doing sensitivity analysis of a wide range of potential cluster numbers over historic data of pipe failures. The computational effort required for developing the clustered EPR approach is over *k* times the standard EPR model. In addition, the approach here was only applied to cast iron pipes due to the highest failure rate in the network. However, it can be analysed to other pipe materials separately providing sufficient number of failures are available to ensure reasonable model accuracy. Finally, other than the explanatory factors analysed here, there are other

factors affecting pipe failures such as weather data which are worthwhile to be investigated in the future research.

7. References

Achim, D., Ghotb, F., and McManus K. J., 2007. Prediction of water pipe asse life using neural networks. *Journal of infrastructure systems*, 13 (1), 26-30.

Ahn, J., Lee, S., Lee, G., and Koo, J., 2005. Predicting water pipes breaks using neural network. *Water Supply* 5 (3-4), 159-172

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

Clair St, A. M. and Sinha, S., 2012. State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models!. *Urban Water Journal*, 9 (2), 85-112.

Folkman S., 2012. Water Main Break Rates in the USA and Canada: A Comprehensive Study, Utah State University, Buried Structures Laboratory, Logan, UT

Garthwaite, P. H., Jolliffe, I. T. and Jones, B., 2002. Statistical inference. Oxford University Press.

Giustolisi, O. and Savic, D. A., 2006. A symbolic data-driven technique based on evolutionary polynomial regression. *Journal of Hydroinformatics*, 8 (3), 207-222.

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

-Giustolisi, O., Savic, D. and Laucelli, D. 2009. Asset deterioration analysis using multi-utility data and multi-objective data mining. *Journal of Hydroinformatics* 11 (3-4), 211–224.

Grossman R, Seni G, Elder, J, Agarwal, N. and Liu, H., 2010. Ensemble Methods in Data Mining: Improving 522 Accuracy Through Combining Predictions. Morgan & Claypool

Jang, J. S. R., Sun, C. T. and Mizutani, E., 1997. Neuro-fuzzy and soft computing; a computational approach to learning and machine intelligence. Practice Hall.

Jenks, G. F. 1963. Generalization in statistical mapping. *Annals of the Association of American Geographers*, 53 (1), 15-26.

Kabir, G., Demissie, G., Sadiq, R. and Tesfamariam, S., 2015. Integrating failure prediction models for water mains: Bayesian belief network based data fusion. *Knowledge-Based Systems*, http://dx.doi.org/10.1016/j.knosys.2015.05.002

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. and Wu, A. Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence IEEE*, 24 (7), 881-892.

Kim, S.E. and Seo, I. W., 2015. Artificial Neural Network ensemble modeling with conjunctive data clustering for water quality prediction in rivers, *Journal of Hydro-Environment Research*, http://dx.doi.org/10.1016/j.jher.2014.09.006

Kleiner, Y. and Rajani, B., 2001. Comprehensive review of structural deterioration of water mains: statistical models. *Urban water*, 3 (3), 131-150.

Kleiner, Y. and Rajani, B., 2012. Comparison of four models to rank failure likelihood of individual pipes. *Journal of Hydroinformatics*, 14 (3), 659-681.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. *In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, 1 (14), 281-297.

Makar, J. M., 2000. A preliminary analysis of failures in grey cast iron water pipes. *Engineering Failure Analysis*, 7 (1), 43-53.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D. and Veith, T. L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the Asabe*, 50 (3), 885-900.

Pham, D. T., Dimov, S. S. and Nguyen, C. D., 2005. Selection of K in K-means clustering. *In Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219 (1), 103-119.

Rajani, B. and Kleiner, Y., 2001. Comprehensive review of structural deterioration of water mains: physically based models. *Urban water*, 3 (3), 151-164.

Røstum, J., 2000. *Statistical modelling of pipe failures in water networks*, Thesis (PhD). University of Science and Technology, Norway

Sadiq, R., Rajani, B. and Kleiner, Y., 2004. Fuzzy-based method to evaluate soil corrosivity for prediction of water main deterioration. *Journal of Infrastructure Systems*, 10 (4), 149–156.

Scheidegger, A., Leitão, J. P. and Scholten, L., 2015. Statistical failure models for water distribution pipes–A review from a unified perspective. *Water research*, 83, 237-247.

Tabesh, M., Soltani, J., Farmani R. and Savic D. A., 2009. Assessing pipe failure rate and mechanical reliability of water distribution networks using data-driven modelling. *Journal of Hydroinformatics*, 11 (1), 1-17.

Wettschereck, D., Aha, D. W. and Mohri, T., 1997. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, 11 (1-5), 273-314.

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

List of Figures: