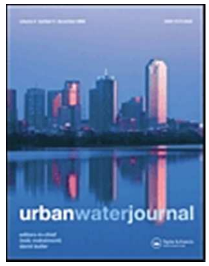

Pipeline Failure Prediction in Water Distribution Networks using Evolutionary Polynomial Regression combined with K-means clustering

Journal:	*Urban Water Journal*
Manuscript ID	NURW-2016-0011.R2
Manuscript Type:	Research Article
Date Submitted by the Author:	18-Aug-2016
Complete List of Authors:	Kakoudakis, Konstantinos; University of Exeter, Centre for Water Systems Behzadian, Kourosh; School of Computing and Engineering, University of West London Farmani, Razieh; University of Exeter, Centre for Water Systems Butler, David; University of Exeter, Centre for Water Systems
Keywords:	k-means clustering, pipe failure predictions, water distribution networks, Evolutionary Polynomial Regression

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, UK*

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

[*]Corresponding author: Konstantinos Kakoudakis (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 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 created groups were divided into training and test datasets using the cross-validation technique for calibration and validation purposes respectively. The *K*-means clustering is employed to partition the training data into a number of clusters for individual EPR models. The proposed approach was demonstrated by application to the cast iron pipes of a water distribution network in the UK. 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 failures in water distribution systems, development of a reliable and accurate prediction model of pipe failure is of paramount importance. The failure is 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). Despite their reasonable accuracy, physical models compared to other methods have significant input data demands because 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, material deterioration (e.g. external and/or internal corrosion) and historical level of pressure (Martínez-Codina *et al.* 2015). 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). Statistical models can cope with the lack of sufficient knowledge related to the complex physical mechanisms that lead to pipe failure although they have some limitations such as requirement for some assumptions (e.g. selection of probability distribution function) that should be substantiated by some knowledge of the phenomenon, which is not always available. In order to overcome the complexity of failure patterns observed in water networks, data-driven methods (Fayyad *et al*. 1996) 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 failures using a non-linear learning process and with no assumption of the form of the relationship between the variables.

EPR (Giustolisi and Savic 2006) is another data-driven method that can be used for prediction of mains pipe breaks (Giustolisi and Savic 2006; Berardi *et al.* 2008, Giustolisi and Berardi 2009). EPR provides a range of statistical equations of pipes failure prediction in a trade-off between training model accuracy and number of polynomial terms. 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 proposed methodology. A description of the case study employed to demonstrate the methodology is given in Section 3. The results are presented and discussed in Section 4 with key findings final remarks are given in the conclusions.

2. Methodology

The proposed methodology consists of the following steps:

- Create pipe groups by aggregating individual pipes using diameter, age and soil type
- Partition the created groups into training and test datasets using the cross-validation technique
- Split the training dataset into *k* clusters using the *K*-means clustering method
- Develop *k* EPR models each associated with the training data of relevant cluster
- Identify the suitable cluster of each test data sample based on the diameter and age of pipe groups

- Use the EPR model corresponding to the associated cluster for each test data sample to calculate the number of failures
- Calculate the performance indicators for the train and test data samples using the observed data

The clusters are created using the KMEANS function in MATLAB (® R2014b) while EPR-MOGA-XL vr.1 (Giustolisi and Savic 2009; Giustolisi *et al*. 2009) is employed to develop the EPR models.

Initially, 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 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. Note that both failed and non-failed pipes are considered here. 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.

The created homogenous groups are split into training and test datasets using the cross-validation technique (Grossman *et al*. 2010) for calibration and validation purposes respectively. The training dataset is partitioned into k clusters based on the age and the diameter. Then, one specific EPR model is developed for each data cluster. The 'explanatory variables' of the EPR models are the total length (L), diameter (D) and age (A) and are the only available explanatory factors for this case study while the 'dependent variable' is the total number of failures (Y).

Finally, the performance of the developed models is evaluated by using the test data. 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 used to predict the number of pipe failures. 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 Evolutionary Polynomial Regression (section 1.1), *K*-means algorithm (section 1.2) and the cross-validation technique (section 1.3) are provided in the supplementary material.

2.1 Model performance assessment

One common way to assess the model prediction ability is the so-called hold-out validation based on a single split of the data, i.e. dividing the entire dataset into two subsets for training and test. However, the model performance derived by this approach would depend significantly on the selection of the training and test datasets. 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. To overcome this drawback the cross-validation method is used (Figure A.1 in supplementary material) for assessing the predictive models. The performance indicators used here are the Coefficient of Determination (R^2) and the Root Mean Square Error (RMSE). Their mathematical relationships are expressed as follows (Moriasi *et al*. 2007):

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

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

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, $\cancel{\bar{y}}$ $\bar{y}_o$= mean value of measurements and *n* = the number of test data samples.

3. Case study

The proposed methodology is demonstrated for prediction of pipe failures in a case study located in part of a water distribution network of a UK city (Table A.1 in supplementary 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. 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~~1). The values shown in Figure 1 are the average values of the 10 iterations of the cross-validation technique. 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 2) 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~~ 1 shows the two performance indicators (R^2 and RMSE) of the predictive models for both training and test data. 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 selected models for both the EPR and the six-clustered EPR approaches show an inverse relationship between the diameter and the number of failures. This relationship is confirmed in the literature (Boxall *et al*. 2007; Berardi *et al*. 2008; Xu *et al*. 2011). On the contrary, the relationship between failure and age shows some complexity. Four selected models with the six-clustered EPR approach corresponding to clusters 1, 2, 4, and 5 show a direct relationship whereas the remaining two models corresponding to clusters 3 and 6 show an inverse relationship. As shown in Figure 2, clusters 3 and 6 entail the oldest pipes. The single model obtained with the EPR approach indicates an inverse relationship.

The main reason for the counterintuitive relationship between pipe failure rate and age in the case study is probably due to the fact that the age of many pipes and particularly the oldest ones is much larger than the time period their failures were systematically recorded since the examined pipe dataset is left truncated. The left truncation occurs when the pipes were installed before their failures were systematically recorded and the number of failures between the installation year and the beginning of the monitoring period is unknown (Scheidegger *et al*. 2015). Hence, the contradiction can be attributed to the lack of pipe failure data collection duration of the monitoring period which is much shorter than the period that the majority of pipes have been in use. Several water authorities have also a brief recorded failure dataset (Pelletier *et al*. 2003; Watson *et al*. 2004). Another possible factor can be that only measurable variables are included in the models. Several explanatory variables, such as design and construction practice, the quality and strength of the material, are not measured and their variation can lead to considerable changes in the subsequent performance of pipes from one age group to another (Boxall *et al*. 2007).

Boxall *et al*. (2007) has also observed a discrepancy in the association between age and pipe failure. Xu *et al*. (2011) examined a brief recorded pipe breakage dataset. They partitioned the pipe database into two clusters of those installed before the beginning of monitoring period and these after the monitoring period. The models they obtained show an inverse relationship between pipe failure and age for the older pipes.

4.1 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. The improvement achieved can linked 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 is compared 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. 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.

4.2 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) (Figure A.2 in supplementary materials). 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. 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.

Figure 6 shows the spatial distribution of predictions of pipe failure rates in the wrong ranges for the six-clustered EPR. The clustered EPR model shows a high accuracy by correctly identifying 85% of the failure rates overall. The achieved accuracy is significantly higher compared to the EPR model which correctly identified 55% of the failure rates (Figure A.3 in supplementary materials).

5. Conclusions

This study presents a new model to predict failures of cast iron pipes in a water distribution networks by combining Evolutionary Polynomial Regression and *K*-means clustering. Individual pipes were *aggregated* using their attributes of age, diameter and soil type to create homogenous groups of pipes. The created homogenous groups were divided into training and test datasets using the cross-validation technique. The training data was partitioned 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 approach here was only applied to cast iron pipes due to the highest failure rate in the network. However, it can be implemented to other pipe materials. 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.

Acknowledgments

The work reported is supported by the UK Engineering & Physical Sciences Research Council (EPSRC) project Safe &SuRe (EP/K006924/1).

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.

Boxall, J. B., O'Hagan A., Pooladsaz, S., Saul A. J. and Unwin D. M., 2007. Estimation of burst rates in water distribution mains. *Institution of Civil Engineers Water Management,* 160 (2), 73–82.

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.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996. From data mining to knowledge discovery in databases. *Al magazine*, 17, 37-54.

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

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 Berardi, L., 2009. Prioritizing Pipe Replacement: From Multiobjective Genetic Algorithms to Operational Decision Support. *Journal of Water Resources Planning and Management, ASCE, USA*, 135 (6), 484–492.

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

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

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.

Martínez-Codina, Á., Castillo, M., González-Zeas, D., and Garrote, L., 2015. Pressure as a predictor of occurrence of pipe breaks in water distribution networks. *Urban Water Journal*, 1-11.

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.

Pelletier, G., A. Mailhot and Villeneuve J. P., 2003. Modeling water pipe breaks-three case studies. *Journal of Water Resources Planning and Management,* 129 (2), 115-123.

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.

Watson, T. G., Christian, C. D., Mason, A. J., Smith, M. H. and Meyer, R., 2004. Bayesian-based pipe failure model. *Journal of Hydroinformatics*, 6 (4), 259-264.

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.

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:

Table 1. Obtained formulas for six-clustered EPR and EPR models

Six-clustered EPR	**Non-clustered EPR**
Cluster 1: $Y=0.513(L^{0.5}A^{0.5}D^{-1})$	$Y=0.01724(LA^{-1}D^{-0.5})$
Cluster 2: $Y=2.206(LAD^{-1})$	
Cluster 3: $Y=0.131(LA^{-1})$	
Cluster 4: $Y=0.219(L^{0.5}A^{0.5}D^{-1})$	
Cluster 5: $Y=2.197(L^{0.5}A^{0.5}D^{-2})$	
Cluster 6: $Y=0.921(LA^{-0.5}D^{-1})$	

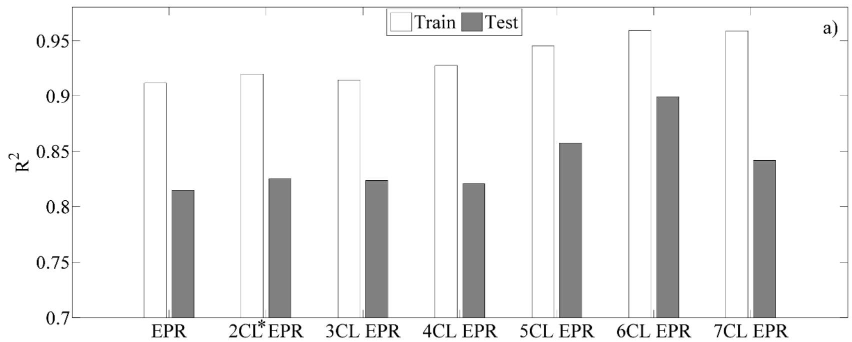

Figure 1a. Performance indicators of the predictive models in terms of R2
*CL=abbreviation for 'clustered' (e.g. 2CL=two-clustered)

508x193mm (300 x 300 DPI)

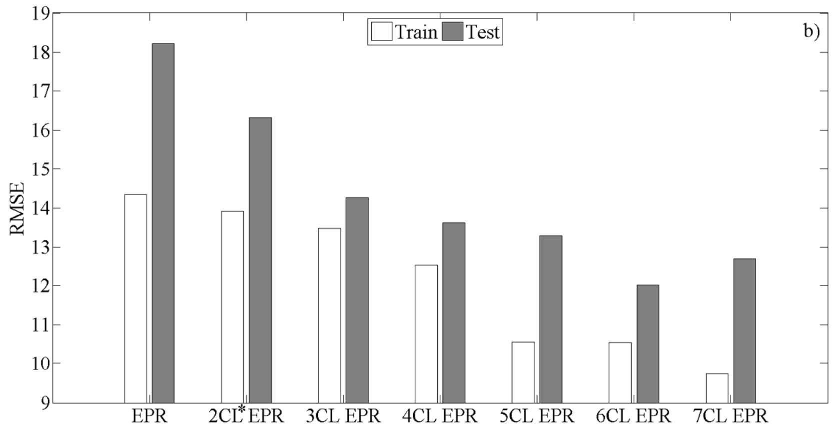

Figure 1b. Performance indicators of the predictive models in terms of RMSE
*CL=abbreviation for ‘clustered’ (e.g. 2CL=two-clustered)

508x243mm (96 x 96 DPI)

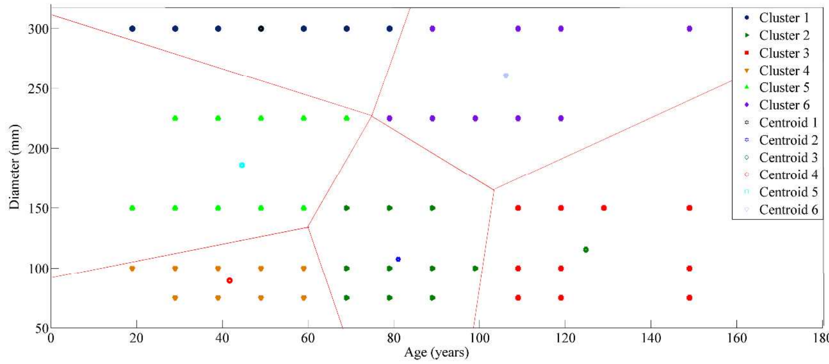

Figure 2. Input data clustering with the six clusters and the corresponding centroids

508x200mm (300 x 300 DPI)

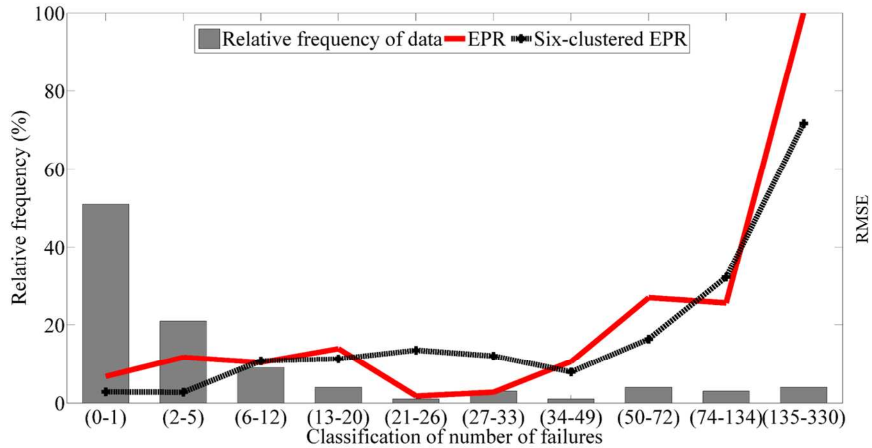

Figure 3. Prediction model error for different intervals of number of failures

279x133mm (150 x 150 DPI)

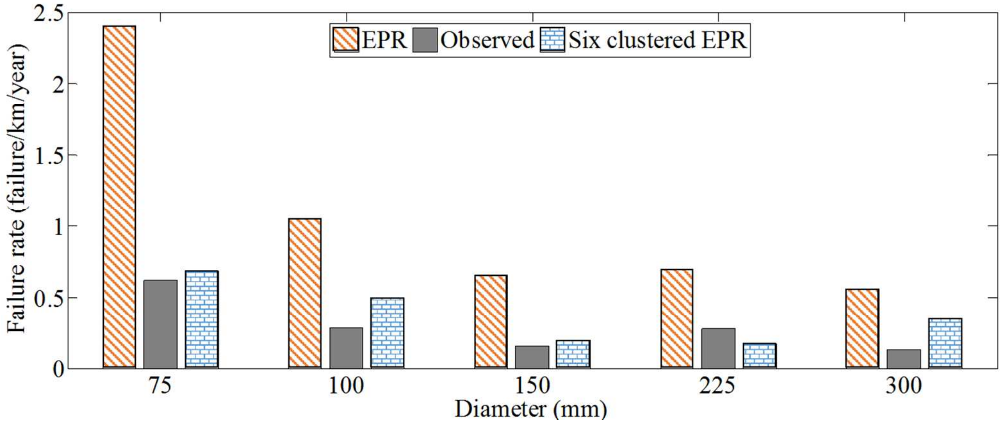

Figure 4a. Average predictions and observations of pipe failure rates based on diameter and

232x97mm (150 x 150 DPI)

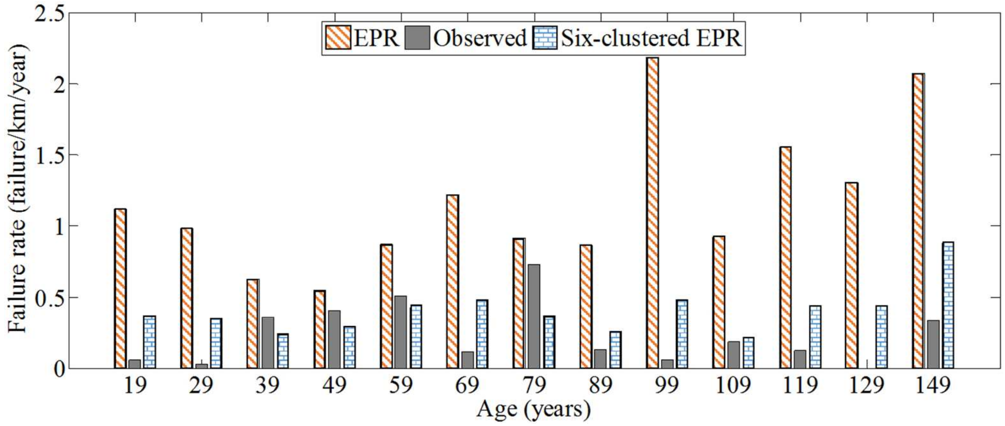

Figure 4b. Average predictions and observations of pipe failure rates based on age

232x98mm (150 x 150 DPI)

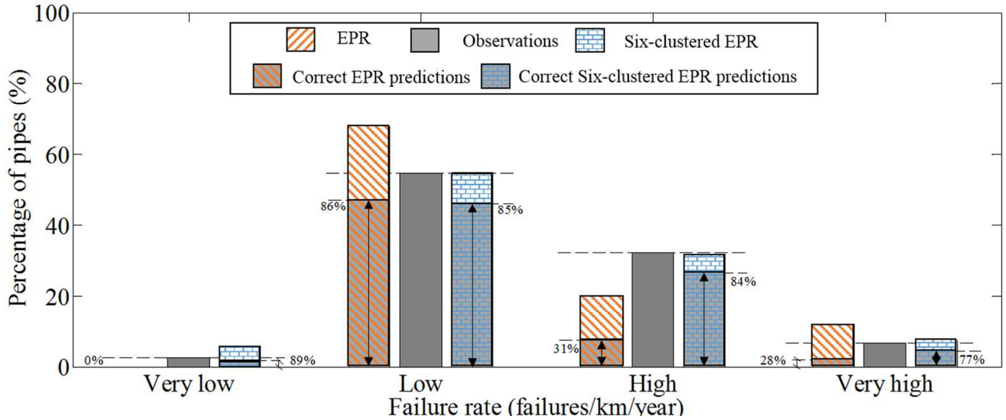

Figure 5. Percentage of pipe failure rates for predictions and observations in different ranges; note that the percentage next to the shaded bars of each predictive model indicates the percentage of correct predictions relative to total observations in each range

233x97mm (150 x 150 DPI)

Figure 6. Six-clustered EPR predictions of pipe failure rate in wrong ranges (black pipes)

140x144mm (300 x 300 DPI)

1. Supplementary materials

1.1 Evolutionary Polynomial Regression

Evolutionary Polynomial Regression (Giustolisi and Savic 2006, Giustolisi and Berardi 2009) 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:

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

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):

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

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 candidate values considered for exponents (E_{ij}) in Eq. (A.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.

1.2 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). 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):

$$\mathrm{J}=\sum_{j=1}^{k}\sum_{i=1}^{n}\left|x_i^{(j)}-c_j\right|^2 \qquad \text{(A.3)}$$

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

1.3 Cross-validation method

Cross-validation method has the advantage that the entire dataset participates in the evaluation of the test set. Other advantage is that each data sample is used for model testing exactly once whereas even in repeated random sub-sampling, some of the original data may be selected more than once in the test dataset and some others may not be selected at all (Gandhi *et al*. 2011). The *m*-fold cross-validation method (Kohavi 1995) is used here. The *m*-fold cross-validation method is an extension of the conventional single-split method in which the data are divided into *m* subsets of (nearly) even size. One subset is taken as the test set (shaded cells in Figure A.1 for a 10-fold instance) and the union of the remaining *m-1* subsets forms the training set. This process is repeated with a new subset of the training/test data and finally the model performance is evaluated *m* times each using a completely different subset of test data. The overall performance is calculated by averaging performance indicators applied to all data in *m* individual performance assessments. In this work, m=10 is used as suggested by Kohavi (1995), in which the union of 9 subsets (i.e. 90% of data) is allocated for training and the one remaining subset (i.e. 10% of data) is retained for test.

1st iteration	1	2	3	4	5	6	7	8	9	10	Training subfolders: 2-10 Test subfolder: 1
2nd iteration	1	2	3	4	5	6	7	8	9	10	Training subfolders: 1, 3-10 Test subfolder: 2
3rd iteration	1	2	3	4	5	6	7	8	9	10	Training subfolders: 1-2, 4-10 Test subfolder: 3
4th iteration	1	2	3	4	5	6	7	8	9	10	Training subfolders: 1-3, 5-10 Test subfolder: 4
5th iteration	1	2	3	4	5	6	7	8	9	10	Training subfolders: 1-4, 6-10 Test subfolder: 5
6th iteration	1	2	3	4	5	6	7	8	9	10	Training subfolders: 1-5, 7-10 Test subfolder: 6
7th iteration	1	2	3	4	5	6	7	8	9	10	Training subfolders: 1-6, 8-10 Test subfolder: 7
8th iteration	1	2	3	4	5	6	7	8	9	10	Training subfolders: 1-7, 9-10 Test subfolder: 8
9th iteration	1	2	3	4	5	6	7	8	9	10	Training subfolders: 1-8, 10 Test subfolder: 9
10th iteration	1	2	3	4	5	6	7	8	9	10	Training subfolders: 1-9 Test subfolder: 10

Figure A.1 10-folds cross-validation technique

1.4 CI pipes of the case study

Table A.1 The main features of the Cast Iron pipes in the case study

Feature	**Value/range**
Installation year	1865-1995
Diameter range	75-300 mm
Total length	814.48 km
Number of pipes	23997
Number of failed pipes	1830
Number of failures	2414

1.5 Comparison of spatial representation of pipe failure rates

Observed failure rates (expressed as number of failures/km/year) of individual pipes can be shown in Figure A.2 by dividing 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]. Figure A.3 shows the spatial distribution of predictions of pipe failure rates in the wrong ranges for the EPR. The EPR model results in a large number of wrong predictions throughout the network especially for 'High' and 'Very high' rates which are the most critical for decision makers.

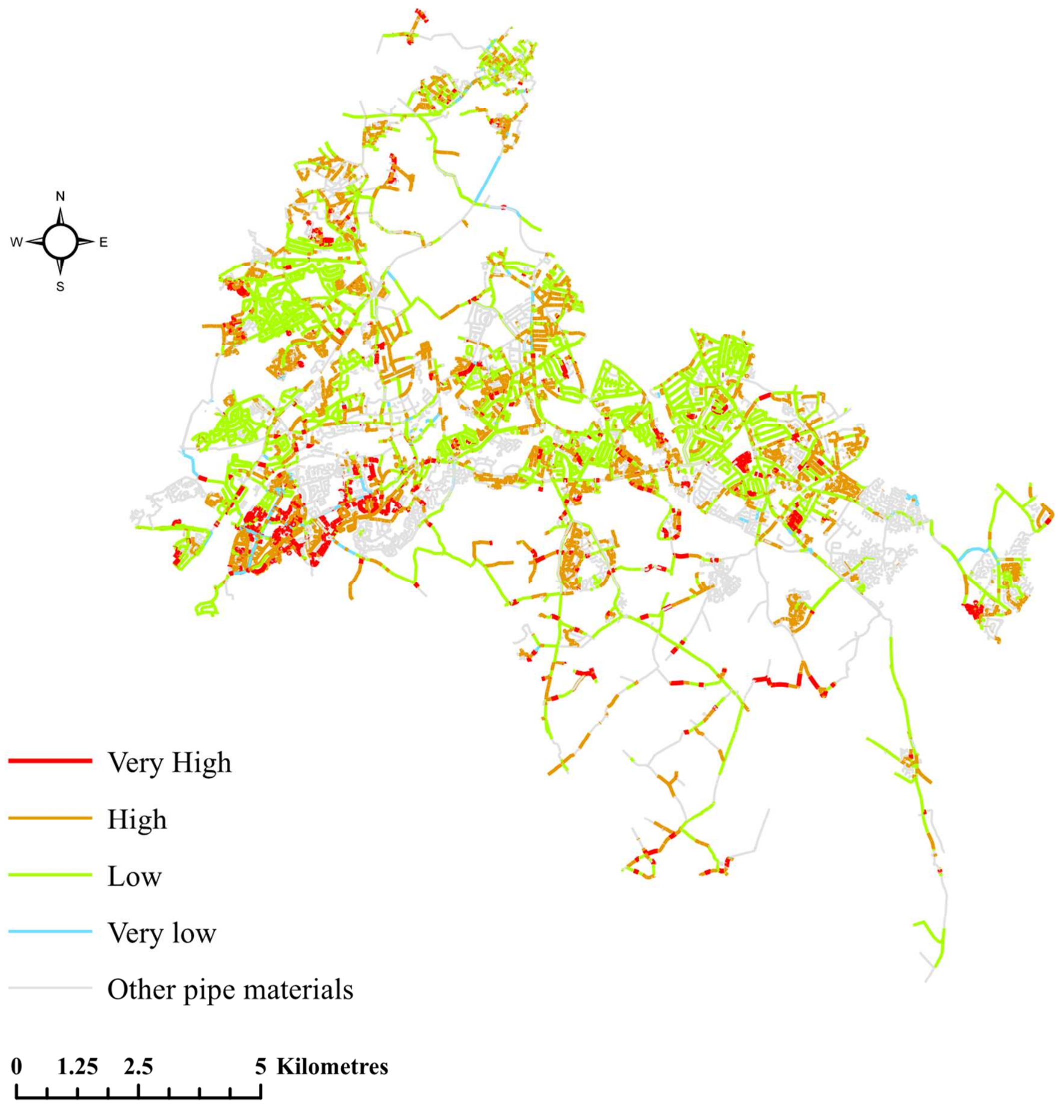

Figure A.2 Observed pipe failure rates of CI pipes

Figure A.3 EPR predictions of pipe failure rate in wrong ranges (black pipes)

2. References

Gandhi, T., Panigrahi, B. K. and Anand. S., 2011. A comparative study of wavelet families for EEG signal classification. *Neurocomputing*, 74 (17), 3051-3057.

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.

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

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 2 (12), 1137-1143.

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.

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.